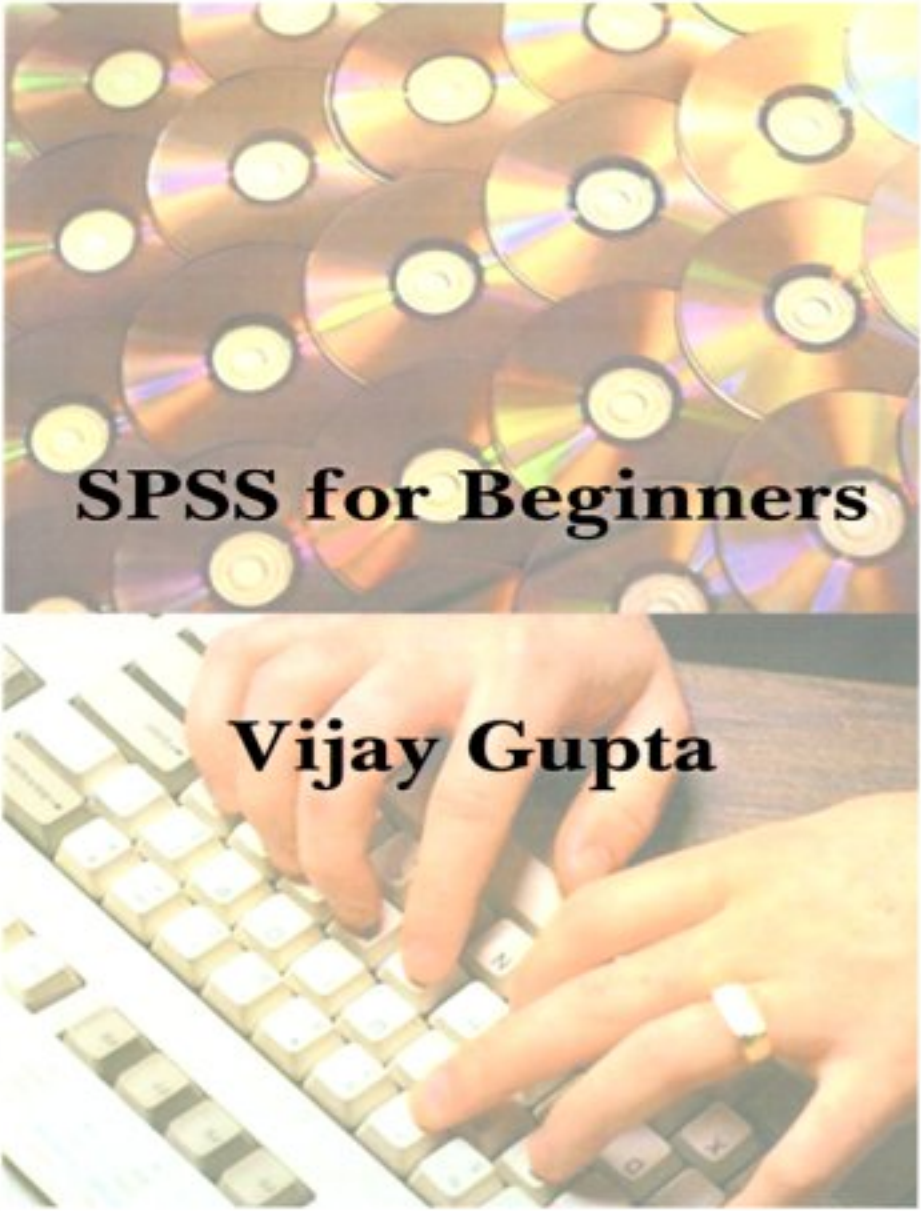

SPSS for Beginners
Vijay Gupta

SPSS for Beginners

Published by VJBooks Inc.

You can reach the author at vgupta1000@aol.com. The author welcomes feedback but will not act as a help desk for the SPSS program.

Library of Congress Catalog No.: Pending
ISBN: Pending
First year of printing: 1999
Date of this copy: Dec 15, 1999

This book is based on SPSS versions 7.x through 10.0. SPSS is a registered trademark of SPSS Inc.

Publisher: VJBooks Inc.
Editor: Vijay Gupta
Author: Vijay Gupta

About the Author

Vijay Gupta has taught statistics, econometrics, SPSS, LIMDEP, STATA, Excel, Word, Access, and SAS to graduate students at Georgetown University. A Georgetown University graduate with a Masters degree in economics, he has a vision of making the tools of econometrics and statistics easily accessible to professionals and graduate students. At the Georgetown Public Policy Institute he received rave reviews for making statistics and SPSS so easy and "non-mathematical." He has also taught statistics to institutions in the US and abroad.

In addition, he has assisted the World Bank and other organizations with econometric analysis, survey design, design of international investments, cost-benefit and sensitivity analysis, development of risk management strategies, database development, information system design and implementation, and training and troubleshooting in several areas. Vijay has worked on capital markets, labor policy design, oil research, trade, currency markets, transportation policy, market research and other topics on The Middle East, Africa, East Asia, Latin America, and the Caribbean. He has worked in Lebanon, Oman, Egypt, India, Zambia, and the U.S.

He is currently working on:

- a package of SPSS Scripts "Making the Formatting of Output Easy"
- a manual on Word
- a manual for Excel
- a tutorial for E-Views
- an Excel add-in titled "Tools for Enriching Excel's Data Analysis Capacity"
- a Graphical User Interface for basic SAS.

Expect them to be available during Early 2000. Early versions can be downloaded from www.vgupta.com.

Acknowledgments

To SPSS Inc, for their permission to use screen shots of SPSS.

To the brave souls who have to learn statistics!

Dedication

To my Grandmother, the late Mrs. Indubala Sukhadia, member of India's Parliament. The greatest person I will ever know. A lady with more fierce courage, radiant dignity, and leadership and mentoring abilities than any other.

Any Feedback is Welcome

You can e-mail Vijay Gupta at vgupta1000@aol.com.

TABLE OF CONTENTS

Contents

Detailed Contents

To take quizzes on topics within each chapter go to
http://www.spss.org/wwwroot/spssquiz.asp

Introduction

Merits of the book

This book is the only user-oriented book on SPSS:

- It uses a series of pictures and simple instructions to teach each procedure. Users can conduct procedures by following the graphically illustrated examples. The book is designed for the novice - even those who are inexperienced with SPSS, statistics, or computers. Though its content leans toward econometric analysis, the book can be used by those in varied fields, such as market research, criminology, public policy, management, business administration, nursing, medicine, psychology, sociology, anthropology, etc.
- Each method is taught in a step-by-step manner.
- An analytical thread is followed throughout the book - the goal of this method is to show users how to combine different procedures to maximize the benefits of using SPSS.
- To ensure simplicity, the book does not get into the details of statistical procedures. Nor does it use mathematical notation or lengthy discussions. Though it does not qualify as a substitute for a statistics text, users may find that the book contains most of the statistics concepts they will need to use.

Organization of the Chapters

The chapters progress naturally, **following the order that one would expect to find in a typical statistics project**.

Chapter 1, “Data Handling," teaches the user how to work with data in SPSS.

> It teaches how to insert data into SPSS, define missing values, label variables, sort data, filter the file (work on sub-sets of the file) and other data steps. Some advanced data procedures, such as reading ASCII text files and merging files, are covered at the end of the book (chapters 12 and 13).

Chapter 2, “Creating New Variables,” shows the user how to create new categorical and continuous variables.

> The new variables are created from transformations applied to the existing variables in the data file and by using standard mathematical, statistical, and logical operators and functions on these variables.

Chapter 3, “Univariate Analysis,” highlights an often-overlooked step - comprehensive analysis of each variable.

Several procedures are addressed - included among these are obtaining information on the distribution of each variable using histograms, Q-Q and P-P plots, descriptives, frequency analysis, and boxplots. The chapter also looks at other univariate analysis procedures, including testing for means, using the T-Test and error bars, and depicting univariate attributes using several types of graphs (bar, line, area, and pie).

Chapter 4, "Comparing Variables," explains how to compare two or more similar variables.

The methods used include comparison of means and graphical evaluations.

Chapter 5, "Patterns Across Variables (Multivariate Statistics)," shows how to conduct basic analysis of patterns across variables.

The procedures taught include bivariate and partial correlations, scatter plots, and the use of stem and leaf graphs, boxplots, extreme value tables, and bar/line/area graphs.

Chapter 6, "Custom Tables," explains how to explore the details of the data using custom tables of statistics and frequencies.

In Chapter 7, "Linear Regression," users will learn linear regression analysis (OLS).

This includes checking for the breakdown of classical assumptions and the implications of each breakdown (heteroskedasticity, mis-specification, measurement errors, collinearity, etc.) in the interpretation of the linear regression. A major drawback of SPSS is its inability to test directly for the breakdown of classical conditions. Each test must be performed step-by-step. For illustration, details are provided for conducting one such test - the White's Test for heteroskedasticity.

Chapter 8, "Correcting for the Breakdown of Classical Assumptions," is a continuation of the analysis of regression from chapter 7. Chapter 8 provides examples of correcting for the breakdown of the classical assumptions.

Procedures taught include WLS and Weight Estimation to correct for heteroskedasticity, creation of an index from several variables to correct for collinearity, 2SLS to correct for simultaneity bias, and model re-specification to correct for mis-specification. This is the most important chapter for econometricians because SPSS does not provide many features that automatically diagnose and correct for the breakdown of classical assumptions.

Chapter 9, "Maximum Likelihood Estimation: Logit, and Non-Linear Estimation," teaches non-linear estimation methods, including non-linear regression and the Logit.

This chapter also suggests briefly how to interpret the output.

Chapter 10 teaches "comparative analysis," a term not found in any SPSS, statistics, or econometrics textbook. In this context, this term means "analyzing and comparing the results of procedures by sub-samples of the data set."

Using this method of analysis, regression and statistical analysis can be explained in greater detail. One can compare results across categories of certain variables, e.g. - gender, race, etc. In our experience, we have found such an analysis to be extremely useful. Moreover, the procedures taught in this chapter will enable users to work more efficiently.

Chapter 11, “Formatting Output,” teaches how to format output.

> This is a SPSS feature ignored by most users. Reviewers of reports will often equate good formatting with thorough analysis. It is therefore recommended that users learn how to properly format output.

Chapters 1-11 form the sequence of most statistics projects. Usually, they will be sufficient for projects/classes of the typical user. Some users may need more advanced data handling and statistical procedures. Chapters 12-18 explore several of these procedures. The ordering of the chapters is based on the relative usage of these procedures in advanced statistical projects and econometric analysis.

Chapter 12, “Reading ASCII Text Data,” and chapter 13 “Adding Data,”deal specifically with reading ASCII text files and merging files.

> The task of reading ASCII text data has become easier in SPSS 9.0 (as compared to all earlier versions). This text teaches the procedure from versions 7.x forward.

Chapter 14, "Non-Parametric Testing," shows the use of some non-parametric methods.

> The exploration of various non-parametric methods, beyond the topic-specific methods included in chapters 3, 4, and 5, are discussed herein.

Chapter 15, "Setting System Options," shows how to set some default settings.

> Users may wish to quickly browse through this brief section before reading Chapter 1.

Chapter 16 shows how to read data from any ODBC source database application/format.

> SPSS 9.0 also has some more database-specific features. Such features are beyond the scope of this book and are therefore not included in this section that deals specifically with ODBC source databases.

Chapter 17 shows time series analysis.

> The chapter includes a simple explanation of the non-stationarity problem and cointegration. It also shows how to correct for non-stationarity, determine the specifications for an ARIMA model, and conduct an ARIMA estimation. Correction for first-order autocorrelation is also demonstrated.

Chapter 18 teaches how to use the two programming languages of SPSS)without having to do any code-writing yourself).

> The languages are:
>
> 1. Syntax -- for programming procedures and data manipulation
> 2. Script -- (mainly) for programming on output tables and charts

Book 2 in this series ("SPSS for Beginners: Advanced Methods") will include chapters on hierarchical cluster analysis, discriminant analysis, factor analysis, optimal scaling, correspondence analysis, reliability analysis, multi-dimensional scaling, general log-linear models, advanced ANOVA and GLM techniques, survival analysis, advanced ranking, using

programming in SPSS syntax, distance (Euclidean and other) measurement, M-estimators, and Probit and seasonal aspects of time series.

As these chapters are produced, they will be available for free download at www.spsss.org. This may be the first interactive book in academic history! Depending on your comments/feedback/requests, we will be making regular changes to the book and the free material on the web site.

The table of contents is exhaustive. Refer to it to find topics of interest.

The index is in two parts - part 1 is a menu-to-chapter (and section) mapping, whereas part 2 is a regular index.

Conventions used in this book

- ⌘ All menu options are in all-caps. For example, the shortened version of: "Click on the menu[1] "Statistics," choose the option "Regression," within that menu, choose the option "Linear regression," will read:
  "Go to STATISTICS / REGRESSION / LINEAR REGRESSION."
- ⌘ Quotation marks identify options in pictures. For example: Select the button "Clustered."
- ⌘ Variable names are usually in italics. For example, *gender*, *wage*, and *fam_id*. Variable names are expanded sometimes within the text. For example, *work_ex* would read *work experience*.
- ⌘ Text and pictures are placed side-by-side. When a paragraph describes some text in a picture, the picture will typically be to the right of the paragraph.
- ⌘ Written instructions are linked to highlighted portions of the picture they describe. The highlighted portions are denoted either by a rectangle or ellipse around the relevant picture-component or by a thick arrow, which should prompt the user to click on the image.
- ⌘ Some terms the user will need to know: a dialog box is the box in any Windows® software program that opens up when a menu option is chosen. A menu option is the list of procedures that the user will find on the top of the computer screen.
- ⌘ Text that is shaded but not boxed is a note, reminder, or tip that digresses a bit from the main text.
- ⌘ Text that is shaded a darker gray and boxed highlights key features.

Data set used in the example followed through this book

One data set is used for most of the illustrations and examples in this book. This allows the user to use the book as a tutorial. Unfortunately, as of present, we cannot find an uncorrupted version of the data file (we had a virus problem). As and when we can obtain such a copy (from an ex-student hopefully) we will place it at http://www.spss.org/wwwroot/spssdown.asp. I created a data set that has the same variable names, sample size and coding as in the corrupted file. The "proxy" data file is provided in the zipped file you downloaded. (Your results will not match

[1] A "menu" is a list of options available from the list on the top of the computer screen. Most software applications have these standard menus: FILE, EDIT, WINDOW, and HELP.

those shown in this book.) The file is called "spssbook.sav." For chapter 17, the data file I used is also included in the zipped file. The data file is called "ch17_data.sav."

The variables in the data set:

1. *Fam_id*: an id number, unique for each family surveyed.
2. *Fam_mem*: the family member responding to the survey. A family (with a unique *fam_id*) may have several family members who answered the survey.
3. *Wage*: the hourly wage of the respondent.
4. *Age*: the age (in years) of the respondent.
5. *Work_ex*: the work experience (in years) of the respondent.
6. *Gender*: a dummy variable taking the value “0” for male respondents and “1” for female respondents.
7. *Pub_sec*: a dummy variable, taking the value “0” if the respondent works in the private sector and “1” if in the public sector.
8. *Educ or educatio*: level of education (in years) of the respondent.

A few more points to note:

- For some examples, new variables are introduced, such as “father’s education” or “mother's education.” For some topics, a totally different data set is used if the example set was not appropriate (e.g. - for time series analysis in chapter 17.)
- The spellings of variables may differ across chapters. This was an oversight by the author. For example, in some chapters the user may note that education level is referred to as *educ* while in others it is referred to as *educatio*.
- With this book, the author hopes to create a marketing trend by raising revenue through advertisements. The author is willing to incorporate, on a limited basis, some community advertisements. The potential world market for this book is 200,000 students each year.

Quick reference and index: Relation between SPSS menu options and the sections in the book

Menu	Sub-Menu	Section that teaches the menu option
FILE	NEW	-
„	OPEN	1.1
„	DATABASE CAPTURE	16
„	READ ASCII DATA	12
„	SAVE	-
„	SAVE AS	-
„	DISPLAY DATA INFO	-
„	APPLY DATA DICTIONARY	-
„	STOP SPSS PROCESSOR	-
EDIT	OPTIONS	15.1

Menu	Sub-Menu	Section that teaches the menu option
„	ALL OTHER SUB-MENUS	-
VIEW	STATUS BAR	15.2
„	TOOLBARS	15.2
„	FONTS	15.2
„	GRID LINES	15.2
„	VALUE LABELS	15.2
DATA	DEFINE VARIABLE	1.2
„	DEFINE DATES	-
„	TEMPLATES	-
„	INSERT VARIABLE	-
„	INSERT CASE, GO TO CASE	-
„	SORT CASES	1.5
„	TRANSPOSE	-
„	MERGE FILES	13
„	AGGREGATE	1.4
„	ORTHOGONAL DESIGN	-
„	SPLIT FILE	10
„	SELECT CASES	1.7
„	WEIGHT CASES	1.3
TRANSFORM	COMPUTE	2.2
„	RANDOM NUMBER SEED	-
„	COUNT	2.4
„	RECODE	2.1
„	RANK CASES	-
„	AUTOMATIC RECODE	2.1
„	CREATE TIME SERIES	17.4
„	REPLACE MISSING VALUES	1.8, 17.4.a
STATISTICS / SUMMARIZE (ANALYZE)	FREQUENCIES	3.2.a
„	DESCRIPTIVES	3.3.a
„	EXPLORE	5.4
„	CROSSTABS	-
„	ALL OTHER	-

Menu	Sub-Menu	Section that teaches the menu option
STATISTICS / CUSTOM TABLES	BASIC TABLES	6.1
„	GENERAL TABLES	2.3 and 6.2 together
„	TABLES OF FREQUENCIES	6.2
STATISTICS / COMPARE MEANS	MEANS	-
„	ONE SAMPLE T-TEST	3.4.b
„	INDEPENDENT SAMPLES T-TEST	5.5.b
„	PAIRED SAMPLES T-TEST	4.3.b
„	ONE-WAY ANOVA	5.5.c
STATISTICS / GENERAL LINEAR MODEL		-
STATISTICS /CORRELATE	BIVARIATE	5.3.a, 5.3.b
„	PARTIAL	5.3.c
„	DISTANCE	-
STATISTICS / REGRESSION	LINEAR	7 (and 8)
„	CURVE ESTIMATION	9.1.a
„	LOGISTIC [LOGIT]	9.1
„	PROBIT	-
„	NON-LINEAR	9.1.b
„	WEIGHT ESTIMATION	8.2.a
„	2-STAGE LEAST SQUARES	8.4
STATISTICS / LOGLINEAR		-
STATISTICS / CLASSIFY	K-MEANS CLUSTER	2.5
„	HIERARCHICAL CLUSTER	-
„	DISCRIMINANT	-
STATISTICS / DATA REDUCTION		-

Menu	Sub-Menu	Section that teaches the menu option
STATISTICS / SCALE		-
STATISTICS / NONPARAMETRIC TESTS	CHI-SQUARE	14.2
„	BINOMIAL	14.1
„	RUNS	14.3
„	1 SAMPLE K-S	3.2.e
„	2 INDEPENDENT SAMPLES	5.5.d
„	K INDEPENDENT SAMPLES	5.5.d
„	2 RELATED SAMPLES	4.3.c
„	K RELATED SAMPLES	4.3.c
STATISTICS / TIME SERIES	EXPONENTIAL SMOOTHING, X11 ARIMA, SEASONAL DECOMPOSITION	-
„	ARIMA	17.5
„	AUTOREGRESSION	17.6
STATISTICS / SURVIVAL		-
STATISTICS / MULTIPLE SETS	DEFINE SETS	2.3
„	FREQUENCIES	2.3 (see 3.1.a also)
„	CROSSTABS	2.3
GRAPHS	BAR	3.1, 4.1, 5.1
„	LINE	3.1, 5.1
„	AREA	3.1, 5.1
„	PIE	3.1, 4.1, 5.1
„	HIGH-LOW, PARETO, CONTROL	-
„	BOXPLOT	3.3.b, 4.2, 5.1.d
„	ERROR BAR	3.4.a, 4.3.a, 5.5.a
„	SCATTER	5.2
„	HISTOGRAM	3.2.a
„	P-P	3.2.b, 3.2.c, 3.2.d
„	Q-Q	3.2.b, 3.2.c, 3.2.d

Menu	Sub-Menu	Section that teaches the menu option
„	SEQUENCE	17.1
„	TIME SERIES/AUTO CORRELATIONS	17.2
„	TIME SERIES/CROSS CORRELATIONS	17.3
„	TIME SERIES/SPECTRAL	-
UTILITIES	VARIABLES	1.2.f
„	FILE INFO	1.2.g
„	DEFINE SETS	1.9
„	USE SETS	1.9
„	RUN SCRIPT	18.1
„	ALL OTHER	-

Ch 1. DATA HANDLING

Before conducting any statistical or graphical analysis, one must have the data in a form amenable to a reliable and organised analysis. In this book, the procedures used to achieve this are termed "Data Handling[2]."[3] SPSS terms them "Data Mining." We desist from using their term because "Data Mining" typically involves more complex data management than that presented in this book and that which will be practical for most users.

The most important procedures are in sections 1.1, 1.2, and 1.7.

In section 1.1, we describe the steps required to read data from three popular formats: spreadsheet (Excel, Lotus and Quattropro), database (Paradox, Dbase, SYLK, DIF), and SPSS and other statistical programs (SAS, STATA, E-VIEWS). See chapter 12 for more information on reading ASCII text data.

Section 1.2 shows the relevance and importance of defining the attributes of each variable in the data. It then shows the method for defining these attributes. You need to perform these steps only once - the first time you read a data set into SPSS (and, as you will learn later in chapters 2 and 14, whenever you merge files or create a new variable). The procedures taught here are necessary for obtaining well-labeled output and avoiding mistakes from the use of incorrect data values or the misreading of a series by SPSS. The usefulness will become clear when you read section 1.2.

Section 1.3 succinctly shows why and how to weigh a data set **if** the providers of the data or another reliable and respectable authority on the data set recommend such weighing.

Sometimes, you may want to analyze the data at a more aggregate level than the data set permits. For example, let's assume you have a data setthat includes data on the 50 states for 30 years (1,500 observations in total). You want to do an analysis of national means over the years. For this, a data set with only 30 observations, each representing an "aggregate" (the national total) for one year, would be ideal. Section 1.4 shows how to create such an "aggregated" data set.

In section 1.5, we describe the steps involved in sorting the data file by numeric and/or alphabetical variables. Sorting is often required prior to conducting other procedures.

---

[2] We can roughly divide these procedures into three sub-categories:

- Data handling procedures essential for any analysis. These include the reading of the data and the defining of each variable's attributes (Sections 1.1, 1.2, and chapters 12 and 16.)
- Data handling procedures deemed essential or important because of the nature of the data set or analysis. These include weighing of the variables, reducing the size of the data set, adding new data to an existing data set, creating data sets aggregated at higher levels, etc. (Sections 1.3, 1.4, 1.6, and chapter 13.)
- Data handling procedures for enhancing/enabling other statistical and graphical procedures. These include the sorting of data, filtering of a Sub-set of the data, and replacing of missing values (Sections 1.5-1.8.)

[3] The "Data Handling" procedures can be found in the menus: FILE and DATA. From the perspective of a beginner or teacher, the biggest drawback of SPSS is the inefficient organisation of menus and sub-menus. Finding the correct menu to conduct a procedure can be highly vexing.

If your data set is too large for ease of calculation, then the size can be reduced in a reliable manner as shown in section 1.6.

Section 1.7 teaches the manners in which the data set can be filtered so that analysis can be restricted to a desired Sub-set of the data set. This procedure is frequently used. For example, you may want to analyze only that portion of the data that is relevant for "Males over 25 years in age."

Replacing missing values is discussed in section 1.8

Creating new variables (e.g. - the square of an existing variable) is addressed in chapter 2.

The most complex data handling technique is "Merging" files. It is discussed in chapter 13.

Another data management technique, "Split File," is presented in chapter 10.

Ch 1. Section 1 Reading (opening) the data set

Data can be obtained in several formats:

- SPSS files (1.1.a)
- Spreadsheet - Excel, Lotus (1.1.b)
- Database - dbase, paradox (1.1.c)
- Files from other statistical programs (1.1.d)
- ASCII text (chapter 12)
- Complex database formats - Oracle, Access (chapter 16)

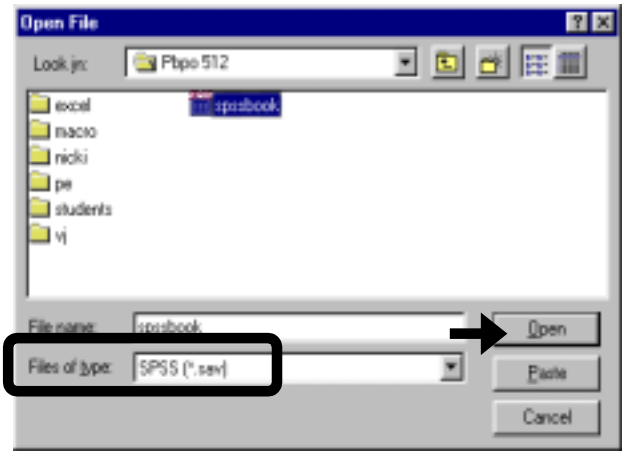
Ch. 1. Section 1.a. Reading SPSS data

In SPSS, go to FILE/OPEN.

Click on the button “Files of Type.”

Select the option “SPSS (*.sav).”

Click on "Open.”



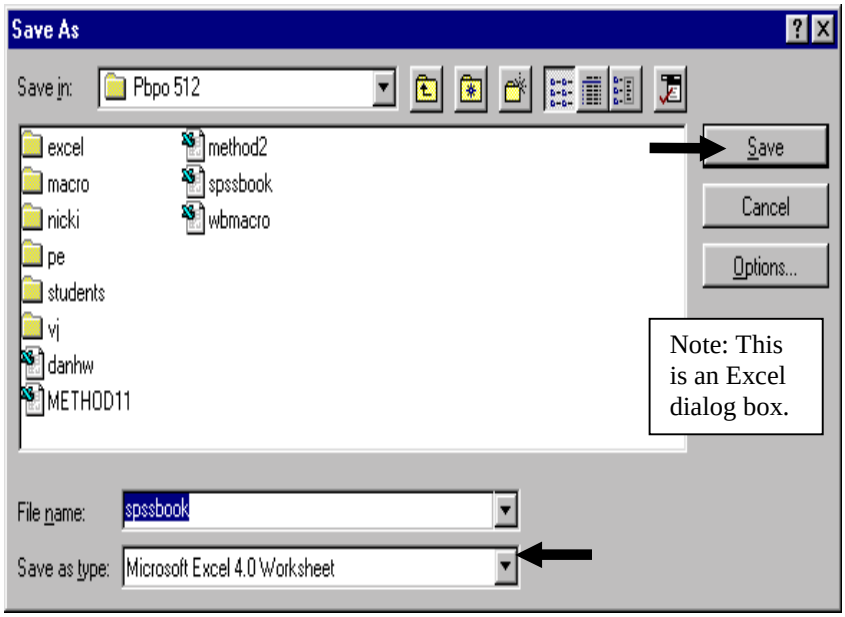
Ch 1. Section 1.b. Reading data from spreadsheet formats - Excel, Lotus 1-2-3

While in Excel, in the first row, type the names of the variables. Each variable name must include no more than eight characters with no spaces[4].

While in Excel, note (on a piece of paper) the range that you want to use[5].

Next, click on the downward arrow in the last line of the dialog box ("Save as type" -see picture on right) and choose the option "Microsoft Excel 4 Worksheet."

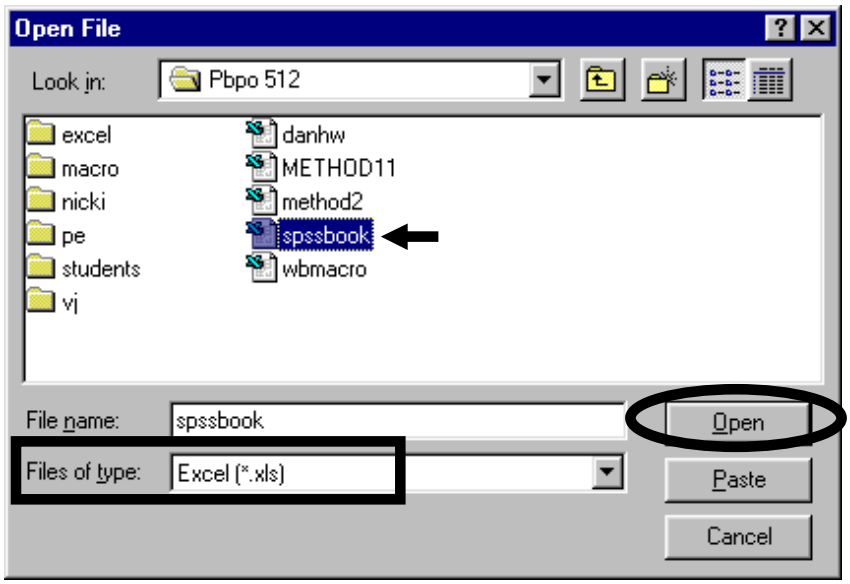
Click on "Save."



In SPSS, go to FILE/OPEN.

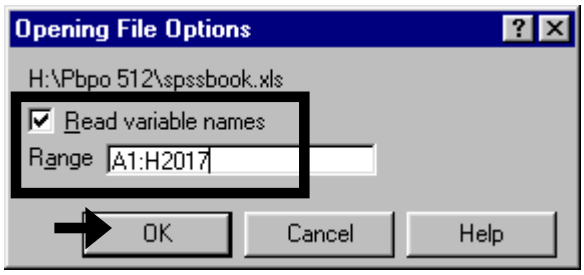
Click on the button "Files of Type." Select the option "Excel (*.xls)."

Select the file, then click on "Open."



SPSS will request the range of the data in Excel and whether to read the variable names. Select to read the variable names and enter the range.

Click on "OK."



[4] Otherwise, SPSS will read the data, but will rename the variable.

[5] Look at the range that has the data you require. In this range, the cell that is on the upper-left extreme corner is the beginning of the range. The cell on the extreme lower right is the end of the range. If the start cell is in row 1 and column "A" and the end cell is in the row 2017 and column "H," then the range is "A1: H2017."

The data within the defined range will be read. Save the opened file as a SPSS file by going to the menu option FILE/ SAVE AS and saving with the extension ".sav."

A similar procedure applies for other spreadsheet formats. Lotus files have the extensions "wk."

Note: the newer versions of SPSS can read files from Excel 5 and higher using methods shown in chapter 16. SPSS will request the name of the spreadsheet that includes the data you wish to use. We advise you to use Excel 4 as the transport format. In Excel, save the file as an Excel 4 file (as shown on the previous page) with a different name than the original Excel file's name (to preclude the possibility of over-writing the original file). Then follow the instructions given on the previous page.

Ch 1. Section 1.c. Reading data from simple database formats - Dbase, Paradox

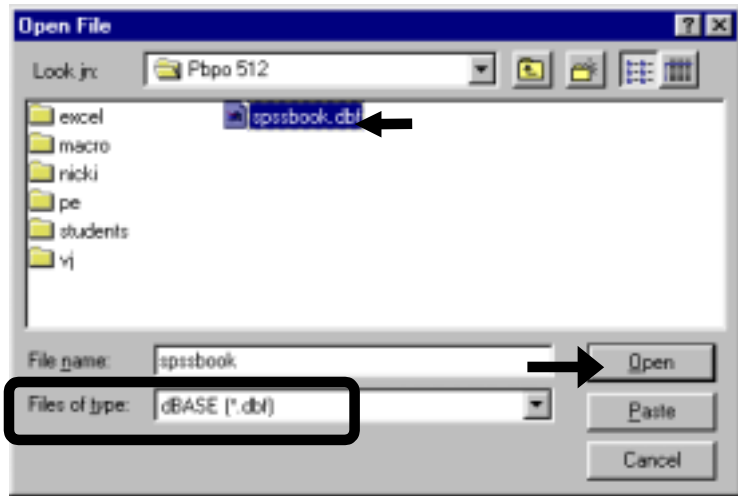
In SPSS, go to FILE/OPEN.

Click on the button "Files of Type." Select the option "dBase (*.dbf)."

Press "Open." The data will be read. Save the data as a SPSS file.

Similar procedures apply to opening data in Paradox, .dif, and .slk formats.

For more complex formats like Oracle, Access, and any other database format, see chapter 16.



Ch 1. Section 1.d. Reading data from other statistical programs (SAS, STATA, etc.)

A data file from SAS, STATA, TSP, E-Views, or other statistical programs **cannot** be opened directly in SPSS.

Rather, while still in the statistical program that contains your data, you must save the file in a format that SPSS can read. Usually these formats are Excel 4.0 (.xls) or Dbase 3 (.dbf). Then follow the instructions given earlier (sections 1.1.b and 1.1.c) for reading data from spreadsheet/database formats.

Another option is to purchase data format conversion software such as "STATTRANSFER" or "DBMSCOPY." This is the preferred option. These software titles can convert between an amazing range of file formats (spreadsheet, database, statistical, ASCII text, etc.) and, moreover, they convert the attributes of all the variables, i.e. - the variable labels, value labels, and data type. (See section 1.2 to understand the importance of these attributes)

Ch 1. Section 2 Defining the attributes of variables

After you have opened the data source, you should assign characteristics to your variables[6]. **These attributes must be clearly defined at the outset before conducting any graphical or statistical procedure**:

1. **Type** (or data type). Data can be of several types, including numeric, text, currency, and others (and can have different types within each of these broad classifications). An incorrect type-definition may not always cause problems, but sometimes does and should therefore be avoided. By defining the type, you are ensuring that SPSS is reading and using the variable correctly and that decimal accuracy is maintained. (See section 1.2.a.)

2. **Variable label**. Defining a label for a variable makes output easier to read but does not have any effect on the actual analysis. For example, the label "Family Identification Number" is easier to understand (especially for a reviewer or reader of your work) than the name of the variable, *fam_id*. (See section 1.2.b.)

   In effect, using variable labels indicates to SPSS that: "When I am using the variable *fam_id*, in any and all output tables and charts produced, use the label "Family Identification Number" rather than the variable name *fam_id*."

   In order to make SPSS display the labels, go to EDIT / OPTIONS. Click on the tab OUTPUT/NAVIGATOR LABELS. Choose the option "Label" for both "Variables" and "Values." This must be done only once for one computer. See chapter for more.

3. **Missing value declaration**. This is essential for an accurate analysis. Failing to define the missing values will lead to SPSS using invalid values of a variable in procedures, thereby biasing results of statistical procedures. (See section 1.2.c.)

4. **Column format** can assist in improving the on-screen viewing of data by using appropriate column sizes (width) and displaying appropriate decimal places (See section 1.2.d.). It does not affect or change the actual stored values.

5. **Value labels** are similar to variable labels. Whereas "variable" labels define the label to use instead of the name of the variable in output, "value" labels enable the use of labels instead of values for specific values of a variable, thereby improving the quality of output. For example, for the variable *gender*, the labels "Male" and "Female" are easier to understand than "0" or "1." (See section 1.2.e.)

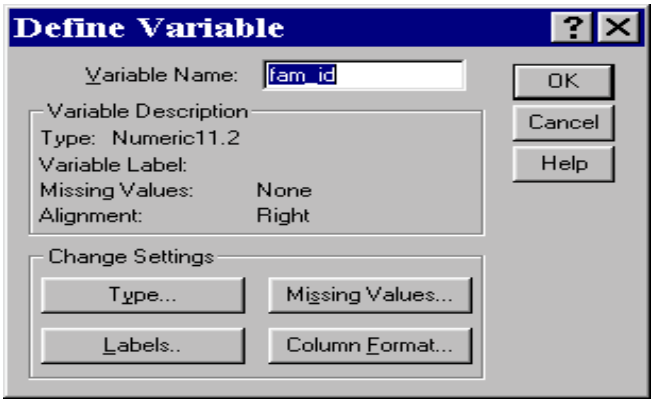
   In effect, using value labels indicates to SPSS that: "When I am using the variable *gender*, in any and all output tables and charts produced, use the label "Male" instead of the value "0" and the label "Female" instead of the value "1"."

[6] If you create a new variable using "compute" or "recode" (see chapter 2) or add variables using "merge" (see chapter 13), you must define the attributes of the variables after the variables have been created/added.

To define the attributes, click on the title of the variable that you wish to define.

Go to DATA/ DEFINE VARIABLE (or double-click on the left mouse).

Sections 1.2.a to 1.2.e describe how to define the five attributes of a variable.



Ch 1. Section 2.a. Variable Type

Choose the **Type** of data that the variable should be stored as. The most common choice is "numeric," which means the variable has a numeric value. The other common choice is "string," which means the variable is in text format. Below is a table showing the data types:

TYPE	EXAMPLE
Numeric	1000.05
Comma	1,000.005
Scientific	1 * e3 (the number means 1 multiplied by 10 raised to the power 3, i.e. (1)*(10^3)
Dollar	$1,000.00
String	Alabama

SPSS usually picks up the format automatically. As a result, you typically need not worry about setting or changing the data type. However, you may wish to change the data type if:

1. Too many or too few decimal points are displayed.
2. The number is too large. If the number is 12323786592, for example, it is difficult to immediately determine its size. Instead, if the data type were made "comma," then the number would read as "12,323,786,592." If the data type was made scientific, then the number would read as "12.32*E9," which can be quickly read as 12 billion. ("E3" is thousands, "E6" is millions, "E9" is billions.)
3. Currency formats are to be displayed.
4. Error messages about variable types are produced when you request that SPSS conduct a procedure[7]. Such a message indicates that the variable may be incorrectly defined.

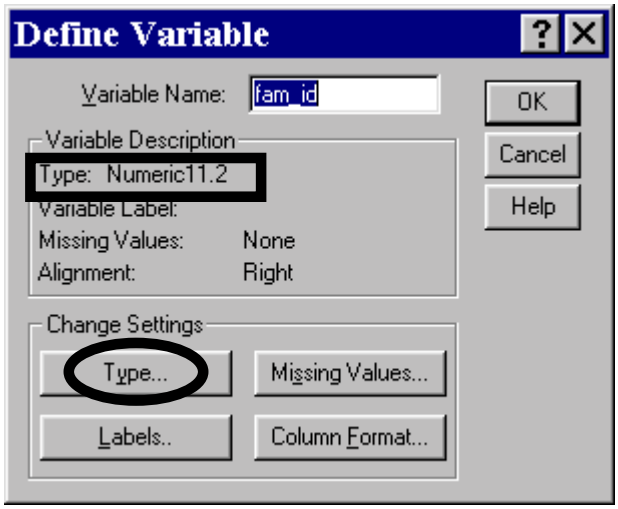
[7] For example, "Variable not numeric, cannot perform requested procedure."

Example 1: Numeric data type

To change the data "Type," click on the relevant variable. Go to DATA/ DEFINE VARIABLE.

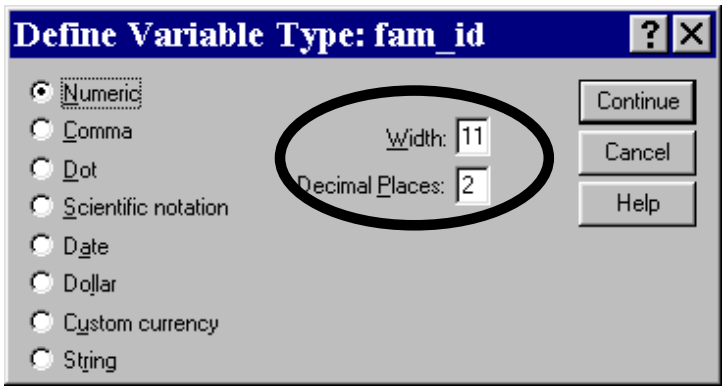
The dialog box shown in the picture on the right will open. In the area "Variable Description," you will see the currently defined data type: "Numeric 11.2" (11 digit wide numeric variable with 2 decimal points). You want to change this.

To do so, click on the button labeled "Type."



The choices are listed in the dialog box.

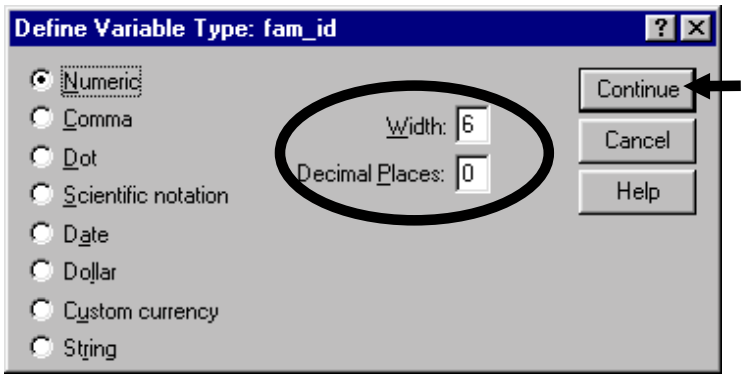
You can see the current specification: a "Width" of 11 with 2 "Decimal Places."



In the "Width" box, specify how many digits of the variable to display and in the "Decimal Places" box specify the number of decimal places to be displayed.

The variable is of maximum width 6[8], so type 6 into the box "Width." Since it is an ID, the number does not have decimal points. You will therefore want to type 0 into the box "Decimal Places."

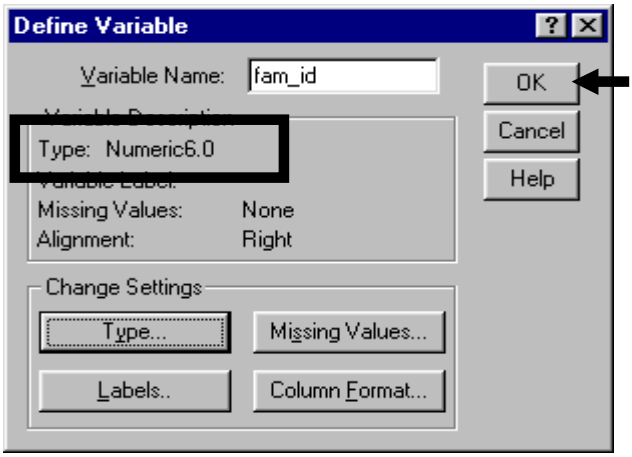
Click on "Continue."



[8] We knew this from information provided by the supplier of the data.

Click on "OK."

The data type of the variable will be changed from "width 11 with 2 decimal places" to "width 6 with no decimal places."



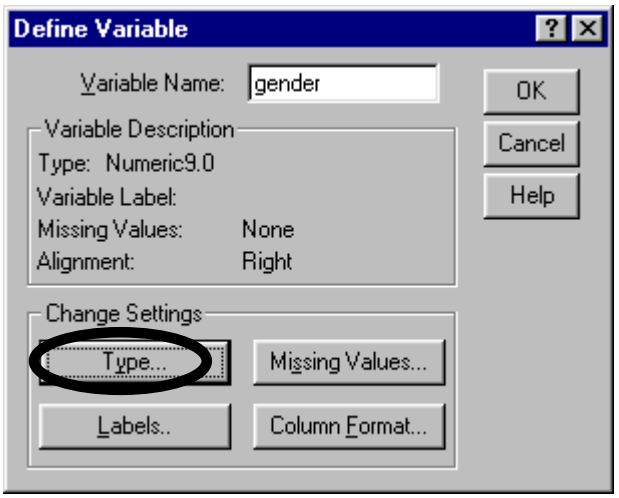
Example 2: Setting the data type for a dummy variable

Gender can take on only two values, 0 or 1, and has no post-decimal values. Therefore, a width above 2 is excessive. Hence, we will make the width 2 and decimal places equal to zero.

Click on the title of the variable *gender* in the data editor.

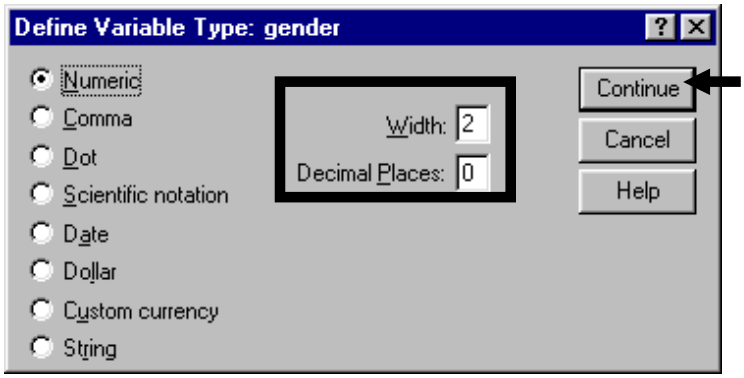
Go to DATA/ DEFINE VARIABLE.

Click on the button "Type."



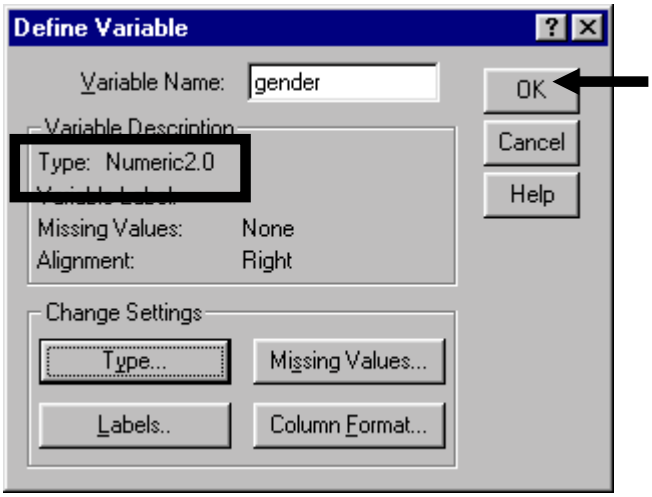
Change width to 2 and decimal places to 0 by typing into the boxes "Width" and "Decimal Places" respectively[9].

Click on "Continue."



[9] A width of 1 would also suffice.

Click on “OK.”

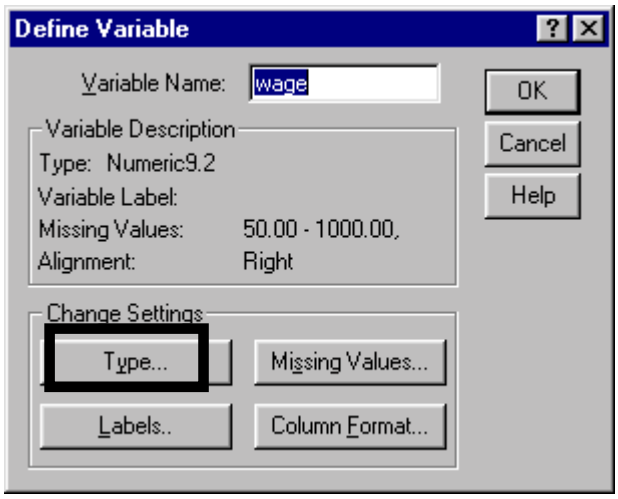


Example 3: Currency type format

We now show an example of the dollar format.

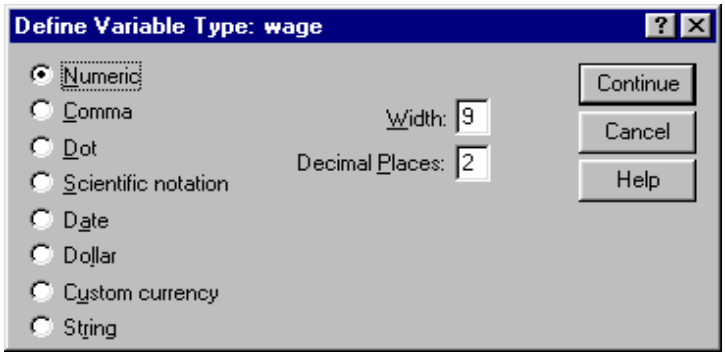
Click on the variable *wage*. Go to DATA/ DEFINE VARIABLE.

Click on the button "Type.”

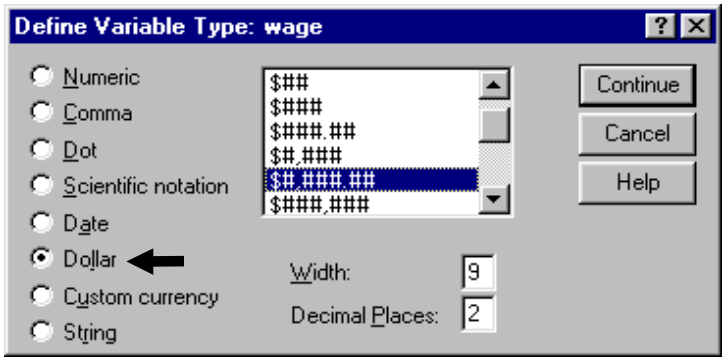


Wage has been given a default data type "numeric, width of 9 and 2 decimal places.”

This data type is not wrong but we would like to be more precise.

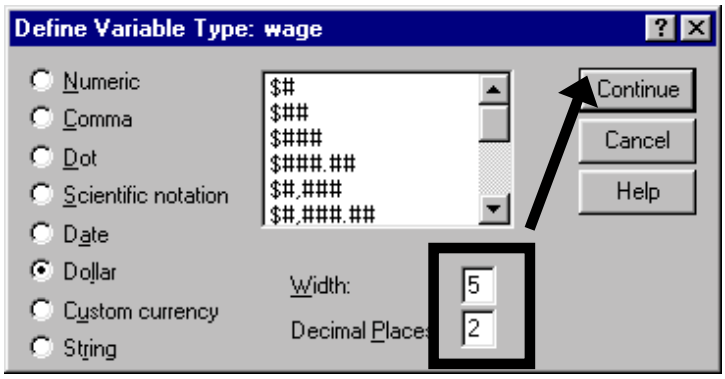


Select the data type "Dollar."



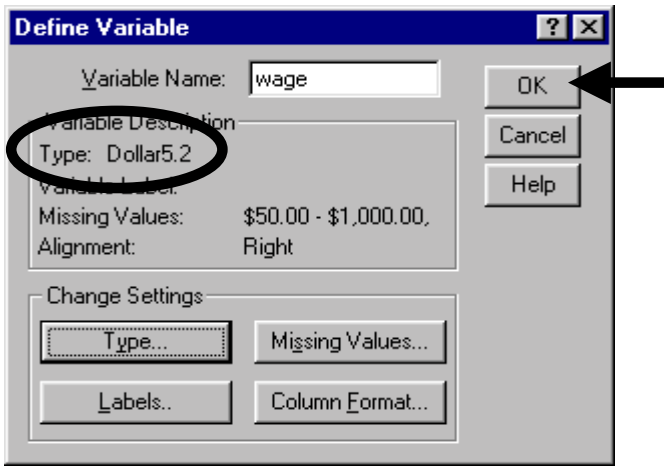
Enter the appropriate width and number of decimal places in the boxes "Width" and "Decimal Places."

Click on "Continue."



Click on "OK."

Now, the variable will be displayed (on-screen and in output) with a dollar sign preceding it.



Ch 1. Section 2.b. Missing Values

It is often the case that agencies that compile and provide data sets assign values like "99" when the response for that observation did not meet the criterion for being considered a valid response. For example, the variable *work_ex* may have been assigned these codes for invalid responses:

- 97 for "No Response"
- 98 for "Not Applicable"
- 99 for "Illegible Answer"

By defining these values as missing, we ensure that SPSS does not use these observations in any procedure involving *work_ex*[10].

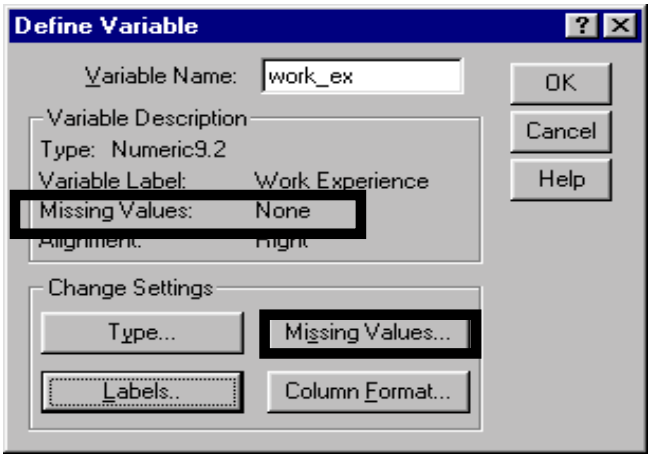
Note: We are instructing SPSS: "Consider 97-99 as blanks for the purpose of any calculation or procedure done using that variable." The numbers 97 through 99 will still be seen on the data sheet but will not be used in any calculations and procedures.

To define the missing values, click on the variable *work_ex*.

Go to DATA/ DEFINE VARIABLE.

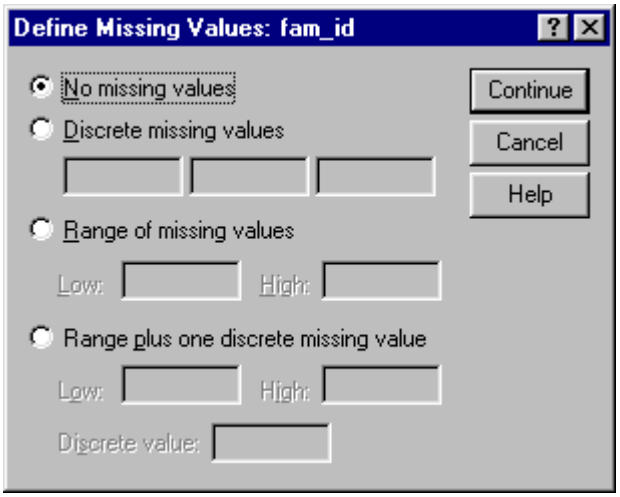
In the area "Variable Description," you can see that no value is defined in the line "Missing Values."

Click on the button "Missing Values."



The following dialog box will open.

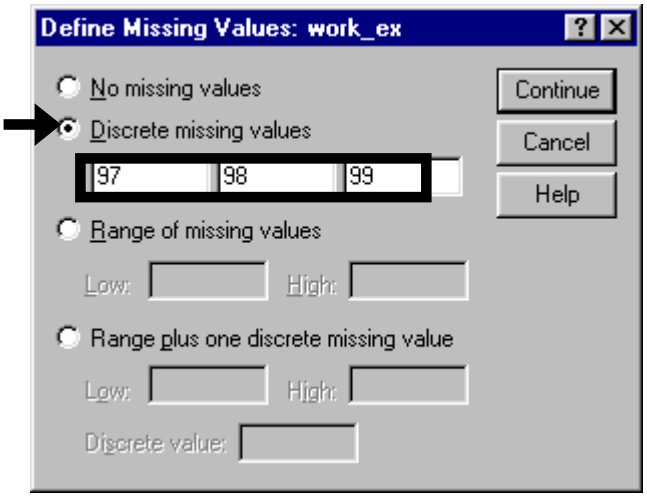
Here, you have to enter the values that are to be considered missing.



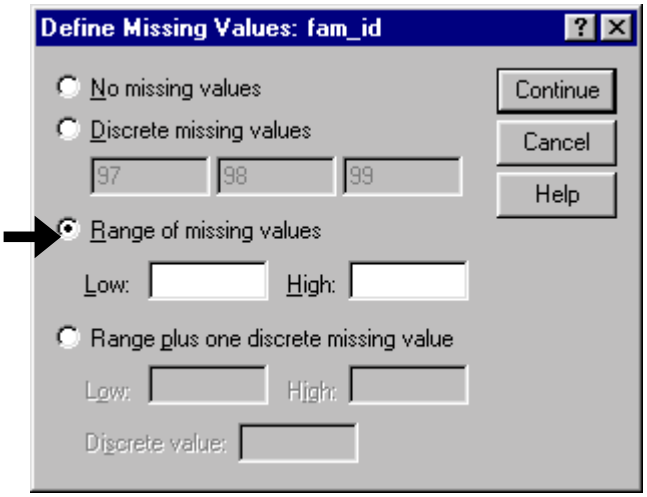
[10] You will still see these numbers on the screen. The values you define as missing are called "User (defined) Missing" in contrast to the "System Missing" value of "null" seen as a period (dot) on the screen.

Click on “Discrete Missing Values.”

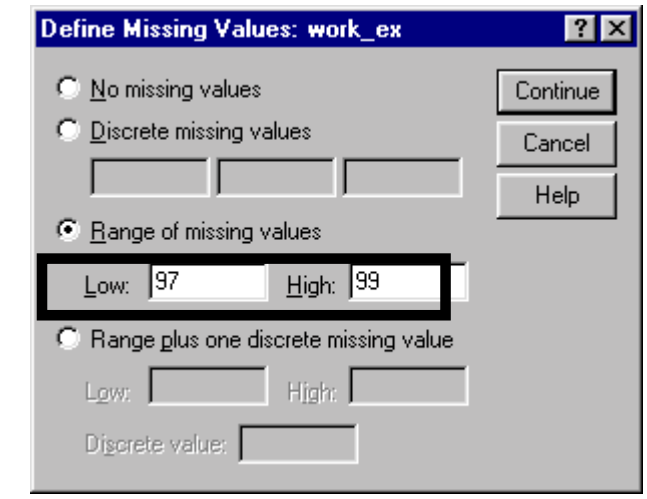
Enter the three numbers 97, 98, and 99 as shown.



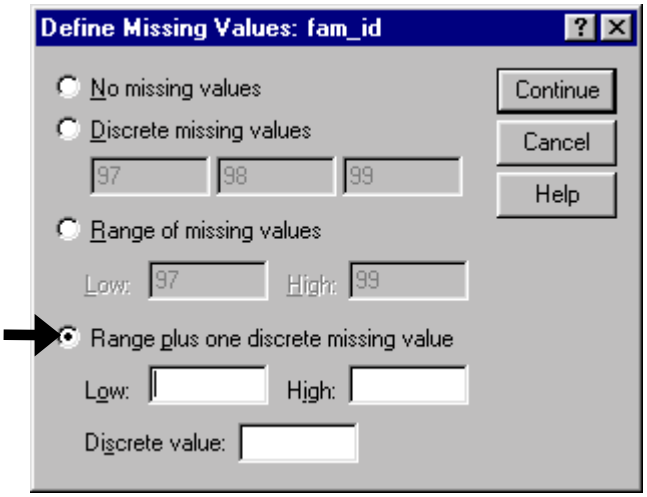
Another way to define the same missing values: choose the option "Range of Missing Values.”



Enter the range 97 (for "Low") and 99 (for "High") as shown. Now, any numbers between (and including) 97 and 99 will be considered as missing when SPSS uses the variable in a procedure.

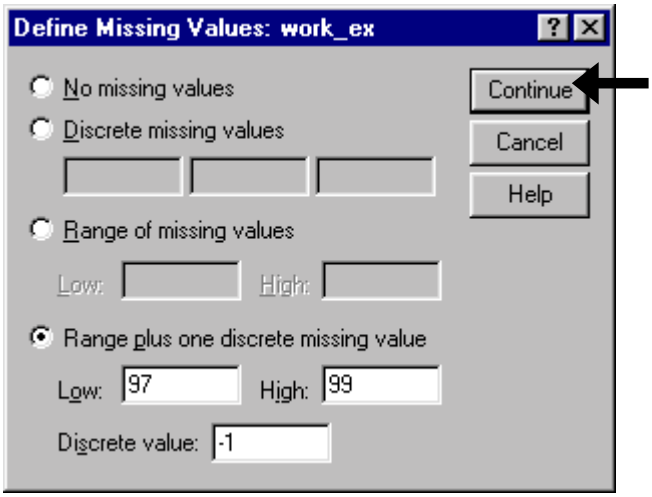


Yet another way of entering the same information: choose "Range plus one discrete missing value."



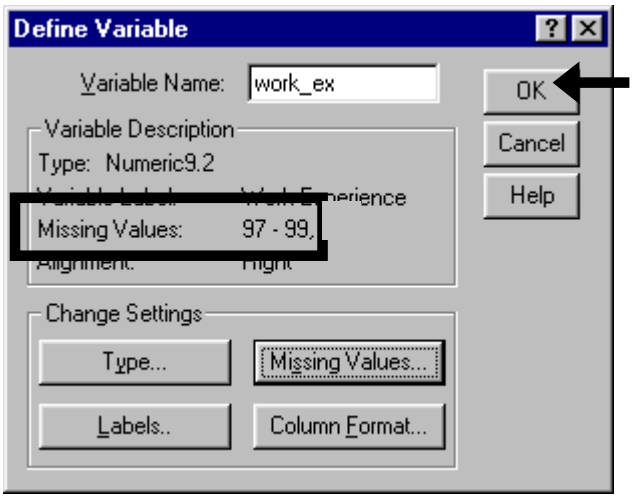
Enter the low to high range and the discrete value as shown.

After entering the values to be excluded using any of the three options above, click on the button "Continue."



Click on "OK."

In the area "Variable Description," you can see that the value range "97-99" is defined in the line "Missing Values."

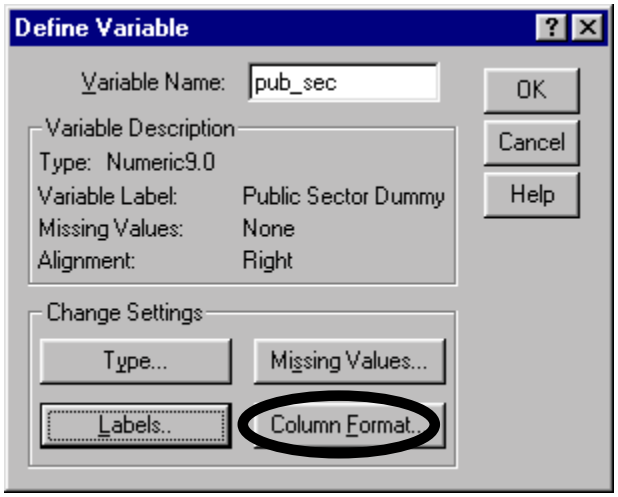


Ch 1. Section 2.c. Column Format

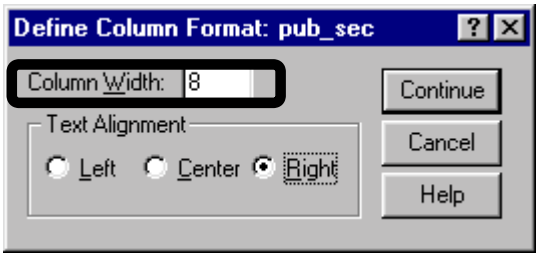
This option allows you to choose the width of the column as displayed on screen and to choose how the text is aligned in the column (left, center, or right aligned). For example, for the

dummy variables g*ender* and *pub_sec*, the column width can be much smaller than the default, which is usually 8.

Click on the data column *pub_sec*. Go to DATA/DEFINE VARIABLE. Click on the button "Column Format."

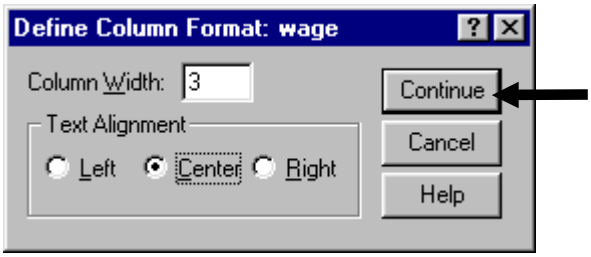


Click in the box "Column Width." Erase the number 8.



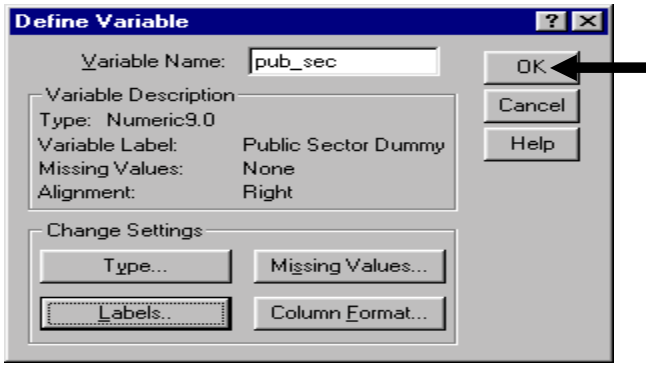
Type in the new column width "3."

Click on "Continue."



Click on "OK."

Remember: the change to column format has only a cosmetic effect. It has no effect on any calculations or procedures conducted that use the variable.



Ch 1. Section 2.d. Variable Labels

This feature allows you to type a description of the variable, other than the variable name, that will appear in the output. The usefulness of the label lies in the fact that it can be a long phrase, unlike the variable name, which can be only eight letters long. For example, for the variable

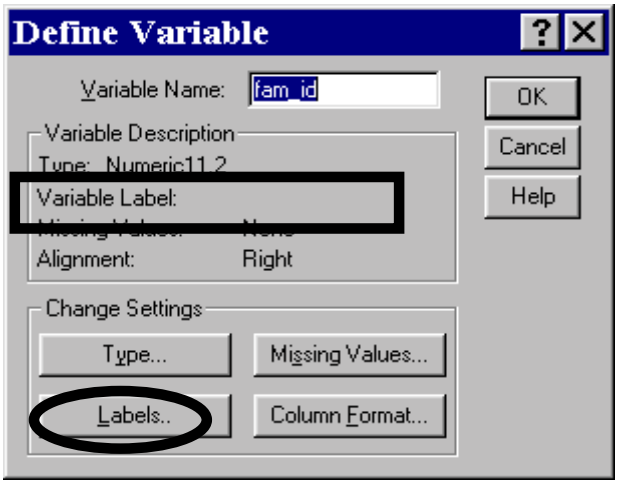
fam_id, you can define a label "Family Identification Number." SPSS displays the label (and not the variable name) in output charts and tables. Using a variable label will therefore improve the lucidity of the output.

Note: In order to make SPSS display the labels in output tables and charts, go to EDIT / OPTIONS. Click on the tab OUTPUT/NAVIGATOR LABELS. Choose the option "Label" for both "Variables" and "Values." This must be done only once for one computer. See also: Chapter 15.

Click on the variable *fam_id*. Go to DATA/DEFINE VARIABLE.

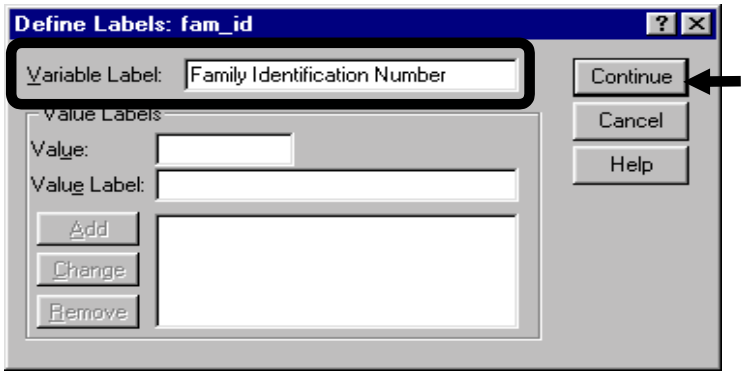
In the area "Variable Description," you can see that no label is defined in the line "Variable Label."

To define the label, click on the button "Labels."



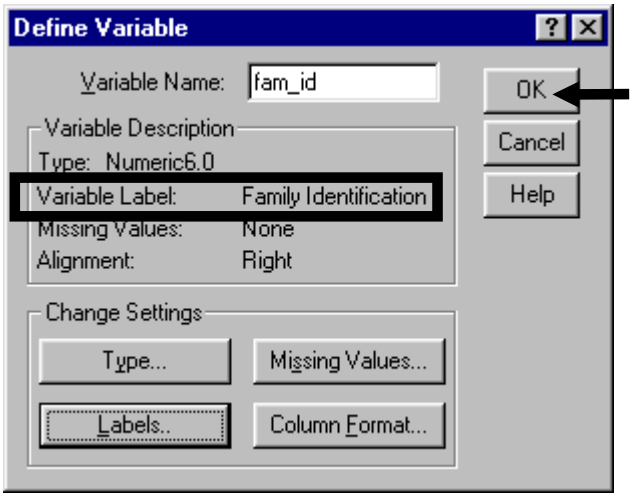
In the box "Variable Label," enter the label "Family Identification Number."

Click on "Continue."



Click on "OK."

In the area "Variable Description," you can see that the label "Family Identification Number" is defined in the line "Variable Label."



Note: You will find this simple procedure extremely useful when publishing and/or interpreting your output.

Ch 1. Section 2.e. Value Labels for Categorical and Dummy Variables

If the variable is a dummy (can have only one of two values) or categorical (can have only a few values, such as 0, 1, 2, 3, and 4) then you should define "value labels" for each of the possible values. You can make SPSS show the labels instead of the numeric values in output. For example, for the variable *pub_sec*, if you define the value 0 with the label "Private Sector Employee" and the value 1 with the label "Public Sector Employee," then reading the output will be easier. Seeing a frequency table with these intuitive text labels instead of the numeric values 0 or 1 makes it easier to interpret and looks more professional than a frequency table that merely displays the numeric values.

In order to make SPSS display the labels, go to EDIT / OPTIONS. Click on the tab OUTPUT/NAVIGATOR LABELS. Choose the option "Label" for both "Variables" and "Values." This must be done only once for one computer. See also: Chapter 15.

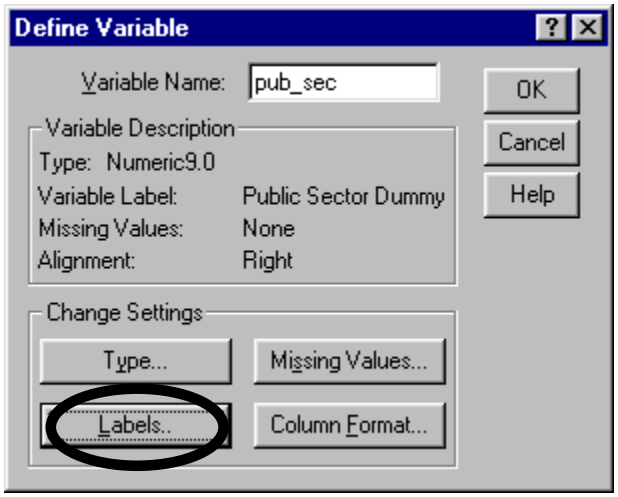
We show an example for one variable - *pub_sec*. The variable has two possible values: 0 (if the respondent is a private sector employee) or 1 (if the respondent is a public sector employee). We want to use text labels to replace the values 0 and 1 in any output tables featuring this variable.

Note: Defining value labels does not change the original data. The data sheet still contains the values 0 and 1.

Click on the data column *pub_sec*.

Go to DATA/DEFINE VARIABLE.

Click on the button “Labels.”

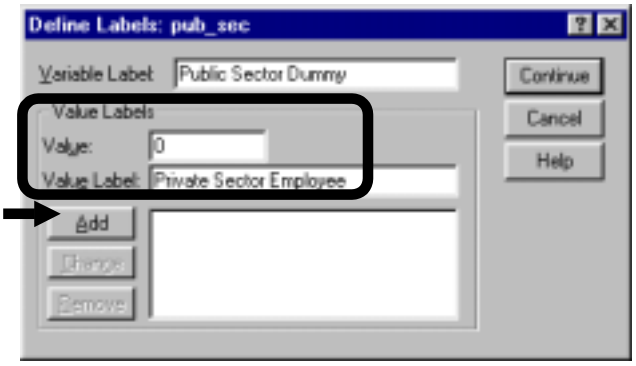


Now, you must enter the Value Labels.

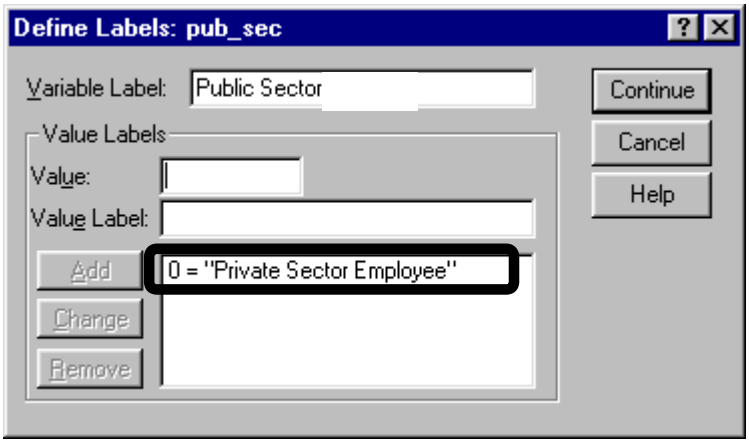
Go to the box “Value.”

Enter the number 0. Then enter its label “Private Sector Employee” into the box “Value Labels.”

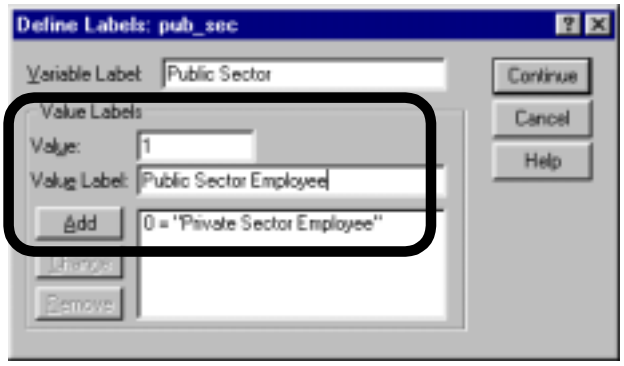
Click on the "Add" button.

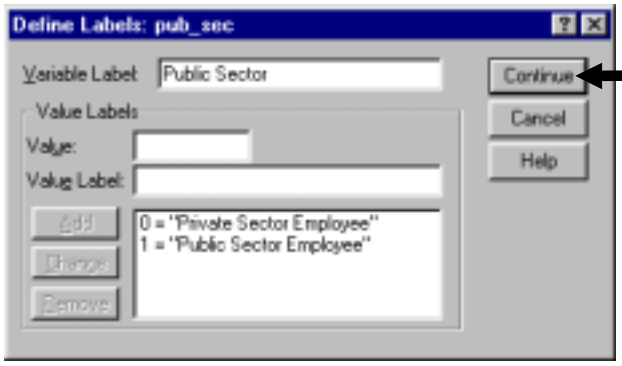


The boxes “Value” and “Value Label” will empty out and the label for the value 0 will be displayed in the large text box on the bottom.

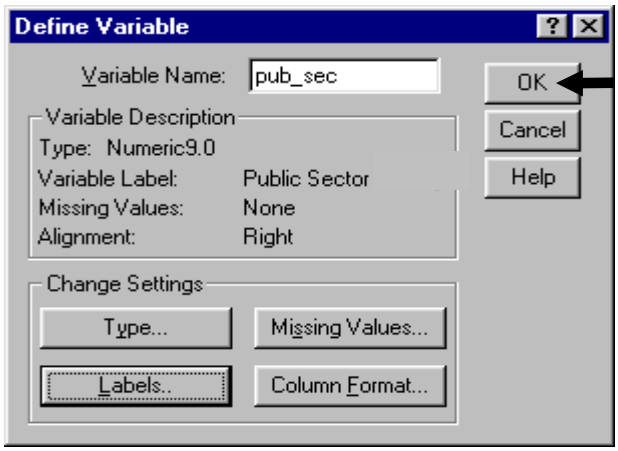


Repeat the above for the value 1, then click on the "Continue" button.





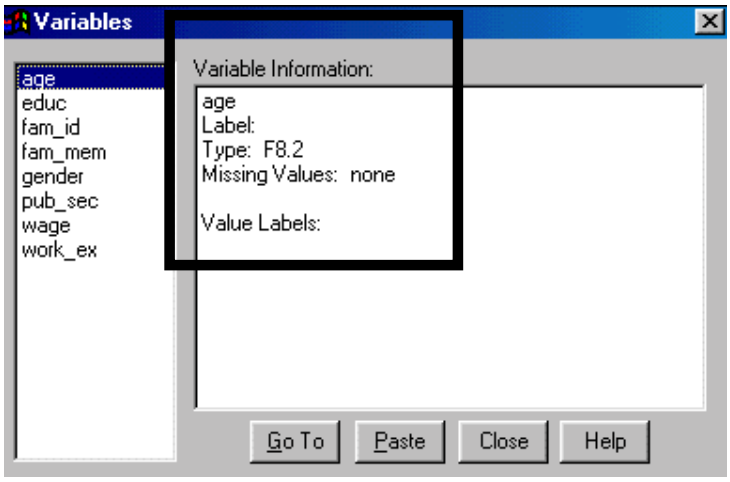
Click on "OK."



To see the labels on the screen, go to VIEW and click on the option "VALUE LABELS." Now, instead of 1s and 0s, you will see "Public Sector Employee" and "Private Sector Employee" in the cells of the column *pub_sec*. See also: chapter 15.

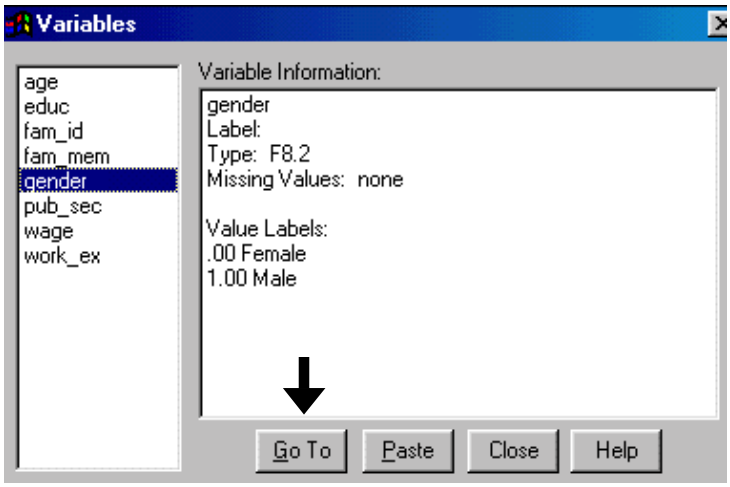
Ch 1. Section 2.f. Perusing the attributes of variables

Go to UTILITY/ VARIABLES. When you click on a variable's name in the left portion of the dialog box that comes up (shown below), the right box provides information on the attributes of that variable.



Locating a column (variable) in a data set with a large number of variables

Sometimes you may wish to access the column that holds the data for a series. However, because of the large number of columns, it takes a great deal of time to scroll to the correct variable. (Why would you want to access a column? Maybe to see the data visually or to define the attributes of the variable using procedures shown in section 1.2).



Luckily, there is an easier way to access a particular column. To do so, go to UTILITY / VARIABLES. When you click on a variable's name in the left portion of the dialog box that comes up (see picture on the right), and then press the button "Go To," you will be taken to that variable's column.

Ch 1. Section 2.g. The File information utility

In section 1.2, you learned how to define the attributes of a variable. You may want to print out a report that provides information on these attributes. To do so, go to UTILITY / FILE INFORMATION. The following information is provided.

```
WAGE      WAGE                                                          1
          Print Format: F9.2
          Write Format: F9.2

WORK_EX[11]   WORK EXPERIENCE[12]
2[13]
          Print Format[14]: F9
          Write Format: F9
          Missing Values[15]: 97 thru 99, -1

EDUC      EDUCATION                                                     3
          Print Format: F9
          Write Format: F9

FAM_ID    FAMILY IDENTIFICATION NUMBER (UNIQUE FOR EACH FAMILY)         4
          Print Format: F8
          Write Format: F8

FAM_MEM    FAMILY MEMBERSHIP NUMBER (IF MORE THAN ONE RESPONDENT FROM
           THE FAMILY)                                                  5
          Print Format: F8
          Write Format: F8

GENDER                                                                  6
          Print Format: F2
          Write Format: F2

          Value    Label[16]
            0     MALE
            1     FEMALE

PUB_SEC                                                                 7
          Print Format: F8
          Write Format: F8

          Value    Label
            0     PUBLIC SECTOR EMPLOYEE
            1     PRIVATE SECTOR EMPLOYEE
```

[11] This is the name of the variable.

[12] This is the "Variable Label."

[13] This is the column number of the variable.

[14] This is the "Data Type." A type "F9.2" means-- Numeric ("F" is for Numeric, "A" for String), width of 9, 2 decimal points.

[15] This is the "Missing Value" definition.

[16] This list gives the "Value Labels" for the variable *gender*.

```
AGE                                                                    8
          Print Format: F8
          Write Format: F8
```

Ch 1. Section 3 Weighing Cases

Statistical analysis is typically conducted on data obtained from “random” surveys. Sometimes, these surveys are not truly "random" in that they are not truly representative of the population. If you use the data as is, you will obtain biased (or less trustworthy) output from your analysis.

The agency that conducted the survey will usually provide a "Weighting Variable" that is designed to correct the bias in the sample. By using this variable, you can transform the variables in the data set into “Weighted Variables.” The transformation is presumed to have lowered the bias, thereby rendering the sample more "random."[17]

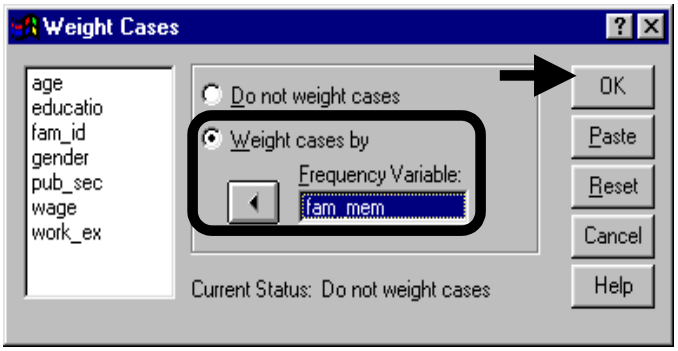
Let's assume that the variable *fam_mem* is to be used as the weighing variable.

Go to DATA/WEIGHT CASES.

Click on “Weight Cases By.”

Select the variable *fam_id* to use as the weight by moving it into the box “Frequency Variable.”

Click on “OK.”

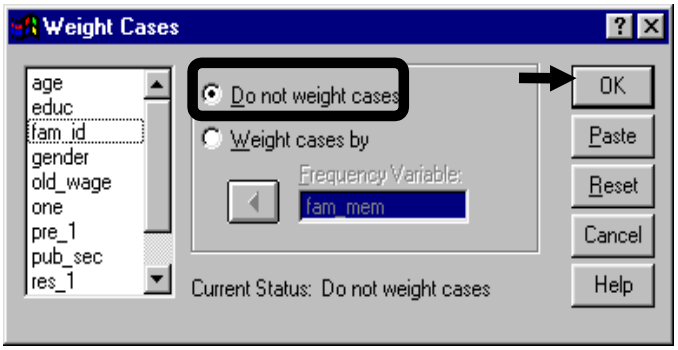


You can turn weighting off at any time, even after the file has been saved in weighted form.

To turn weighting off, go to DATA/WEIGHT CASES.

Click on “Do Not Weight Cases.”

Click on “OK.”



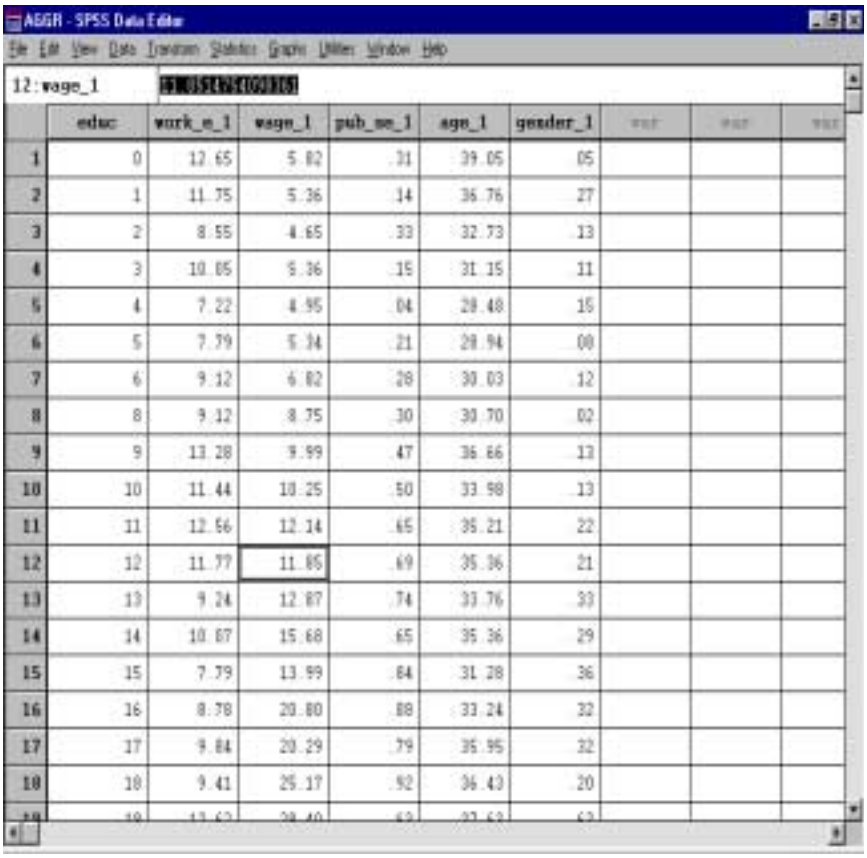
Ch 1. Section 4 Creating a smaller data set by aggregating over a variable

Aggregating is useful when you wish to do a more macro level analysis than your data set permits.

[17] Our explanation is simplistic, but we believe it captures the essence of the rationale for weighting.

Note: If this topic seems irrelevant, feel free to skip it. Most projects do not make use of this procedure.

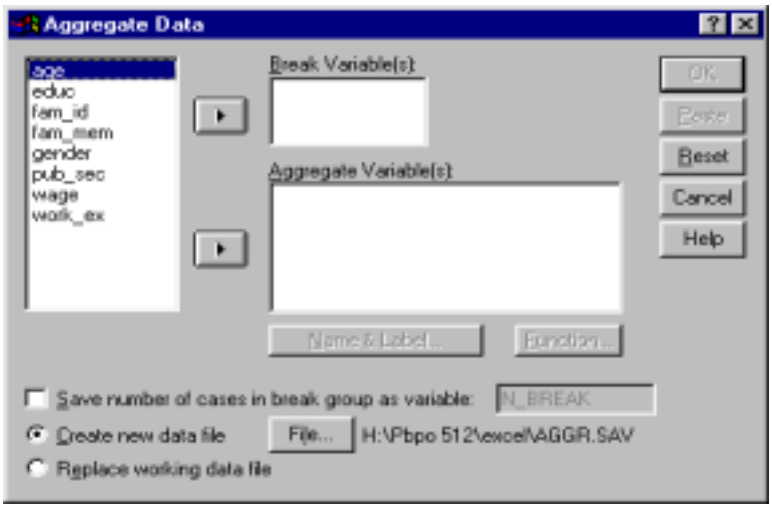
Let's assume that you are interested in doing analysis that compares across the mean characteristics of survey respondents with different *education* levels. You need each observation to be a unique *education* level. The household survey data set makes this cumbersome (as it has numerous observations on each education level).



A better way to do this may be to create a new data set, using DATA/ AGGREGATE, in which all the numeric variables are averaged for each *education* level. The new data set will have only 24 observations - one for each *education* level.

This new data set will look like the picture on the right. There are only 24 observations, one for each *education* level. *Education* levels from 0 to 23 constitute a variable. For the variable *age*, the mean for a respective *education* level is the corresponding entry.

To create this data set, go to DATA/ AGGREGATE.

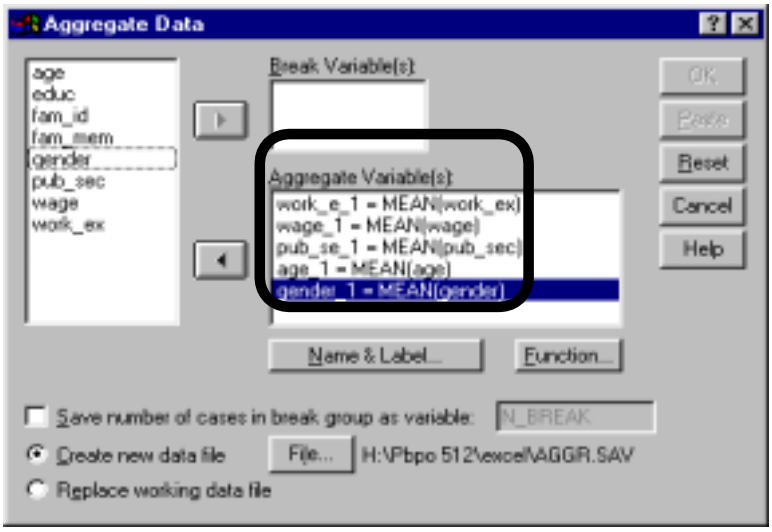


The white box on top ("Break Variable(s)" is where you place the variable that you wish to use as the criterion for aggregating the other variables over. The new data set will have unique values of this variable.

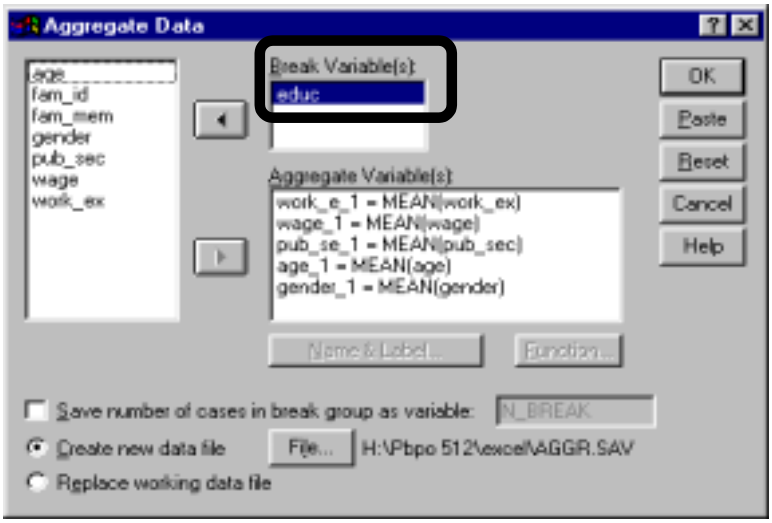
The box “Aggregate Variable(s)” is where you place the variables whose aggregates you want in the new data set.

Move the variables you want to aggregate into the box "Aggregate Variable(s)."

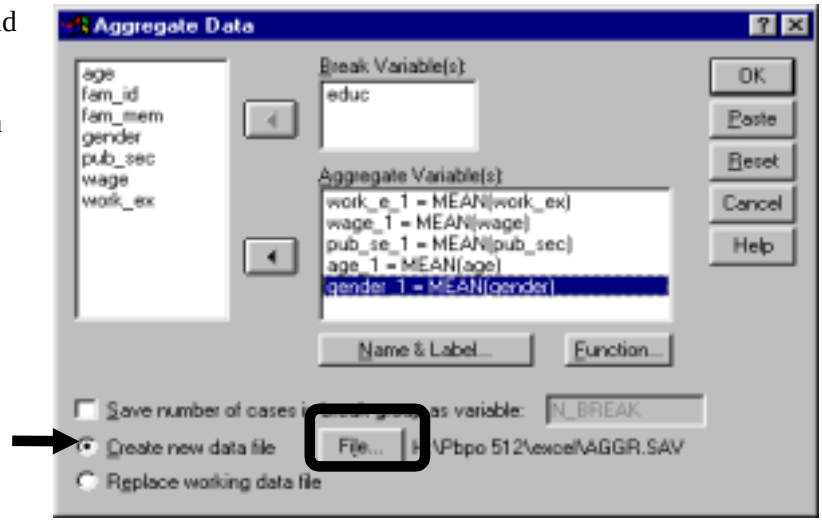
Note that the default function "MEAN" is chosen for each variable[18].



Move the variable whose values serve as the aggregation criterion (here it is *educ*) into the box "Break Variable(s)."



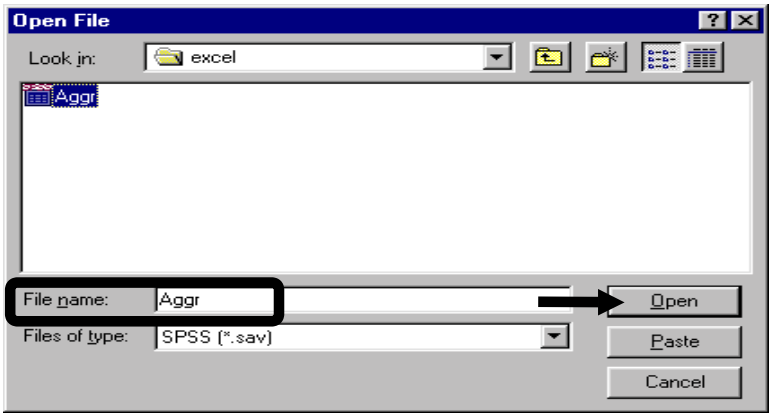
The aggregate data set should be saved under a new name. To do so, choose "Create New Data File" and click on the "File" button.



[18] In the next example, we show how to choose other statistics as the criterion for aggregation.

Select a location and name for the new file.

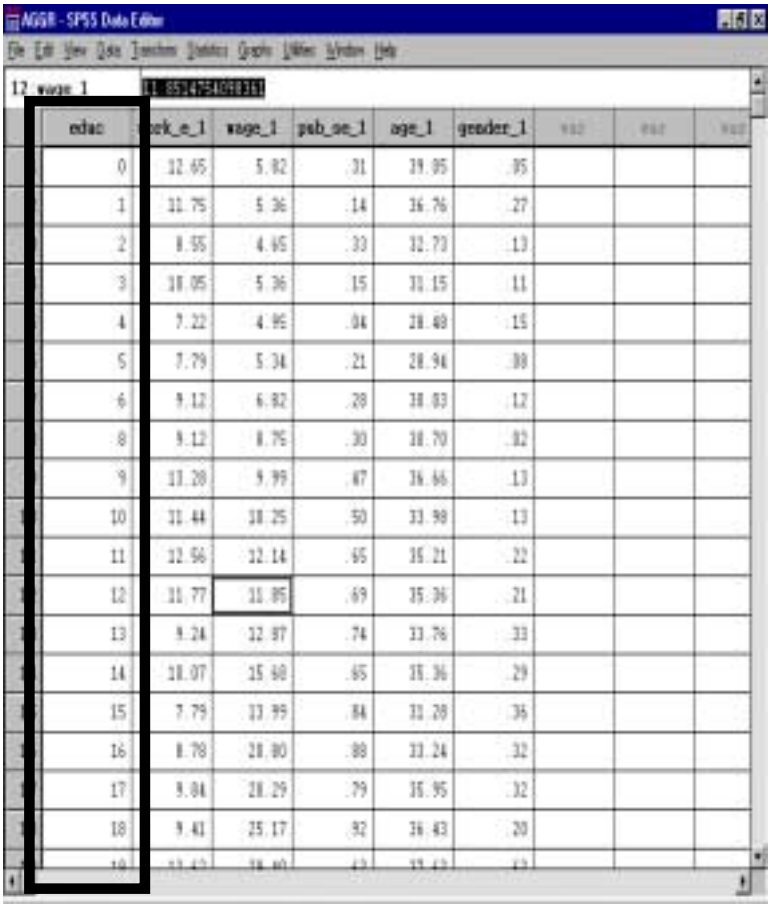
Click on "Open."



A new data file is created.

The variable *educ* takes on unique values. All the other variables are transformed values of the original variables.

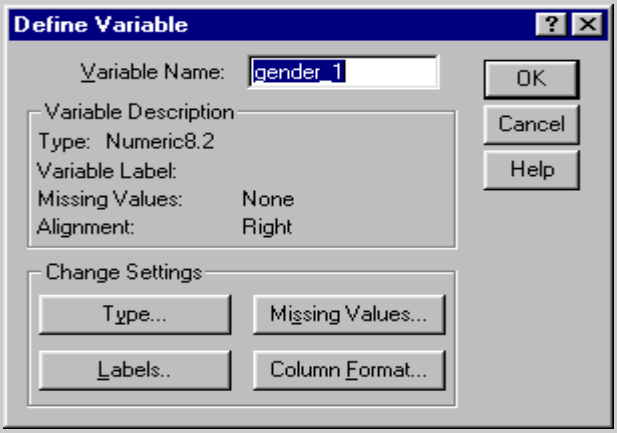
- *work_e_1* is the mean work experience (in the original data set) for each *education* level.
- *wage_1* is the mean wage (in the original data set) for each *education* level.
- *pub_se_1* is the proportion of respondents who are public sector employees (in the original data set) for each *education* level.
- *age_1* is the mean age (in the original data set) for each *education* level.
- *gender_1* is the proportion of respondents who are female (in the original data set) for each *education* level.



The variable gender_1 refers to the proportion of females at each *education* level[19]. We should define the attributes of this variable.

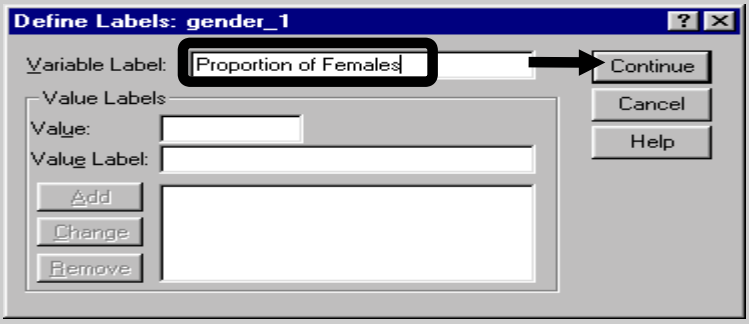
To do so, click on the variable *gender*_1 and go to DATA/ DEFINE VARIABLE.

Click on "Labels."



Enter an appropriate label.

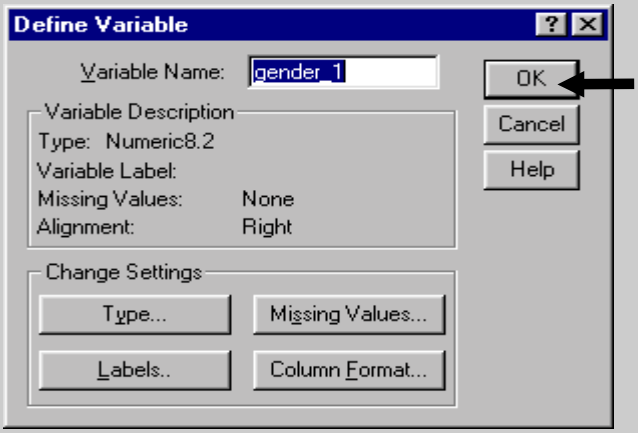
Click on "Continue."



Click on "OK."

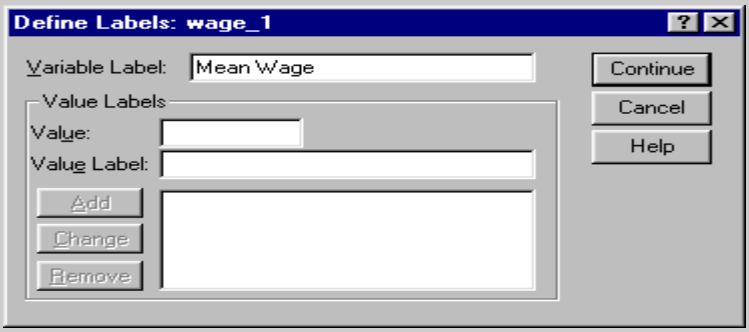
The new label for the variable reflects accurately the meaning of each value in the variable *gender*_1.

Do the same for the other "proportion" variable, *pub_se_1*.



The continuous variables are not the same as in the original data set.

You should redefine the labels of each continuous variable. For example, *wage*_1, *age*_1, and *work_e_1* should be labeled as "Means of Variable" (otherwise output will be difficult to interpret). Also,



[19] In our data set, females have a value of 1 in the variable *gender*.

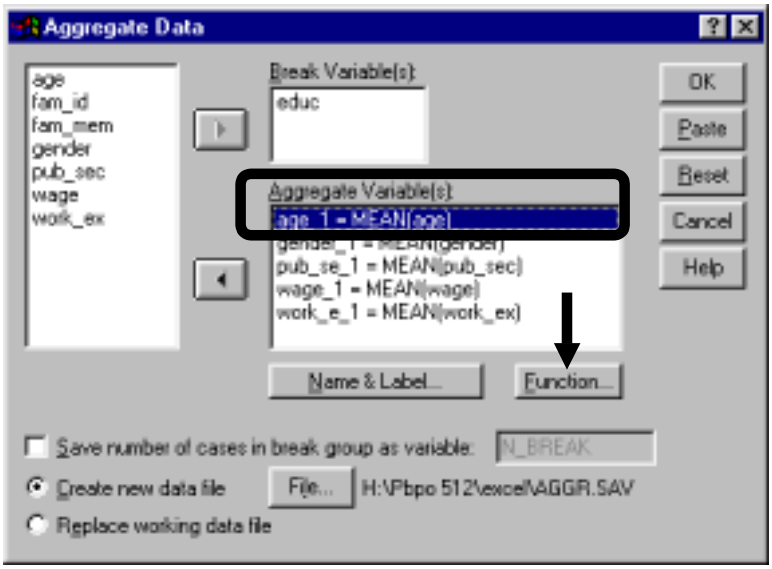
you may make the error of referring to the variable as "Wage" when, in reality, the new data set contains values of "Mean Wage."

Using other statistics (apart from mean)

To go back to the DATA / AGGREGATE procedure: you can use a function different from “mean.”

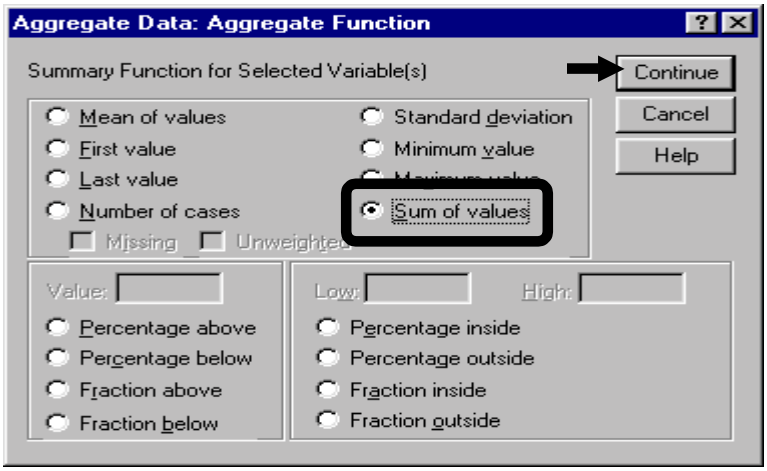
After entering the variables in the dialog box for DATA/ AGGREGATE, click on the variable whose summary function you wish to change.

Click on “Function.”



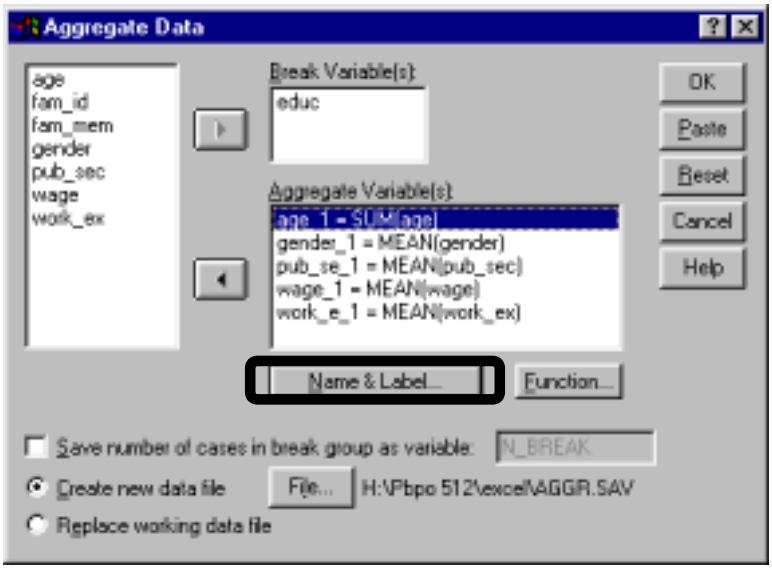
Choose the function you would like to use for aggregating *age* over each value of *education* level.

Click on “Continue.”



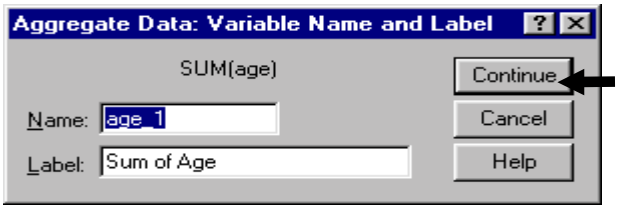
You can change the names and labels of the variables in the new data set.

Click on "Name & Label."



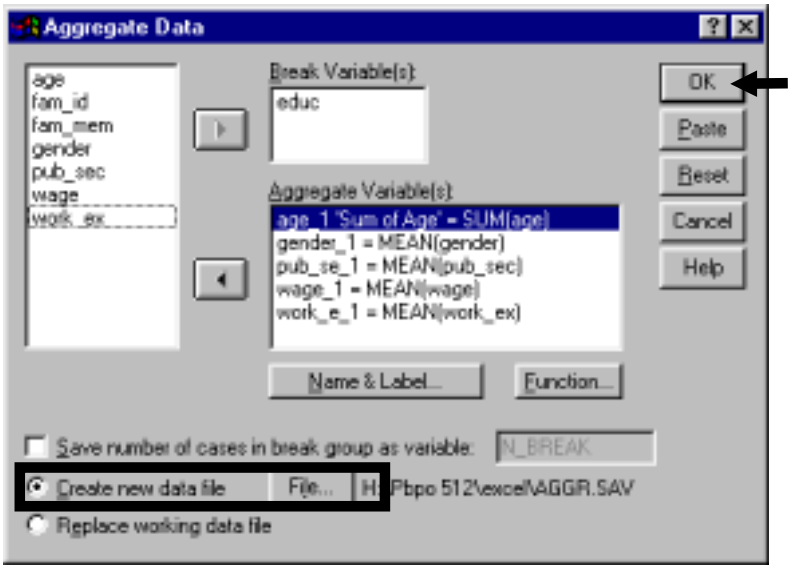
Change the variable name and enter (or change) the label.

Click on "Continue."



Save the file using a different name. To do so, click on "Create new data file" and the button "File." Select a path and name.

Click on "OK."



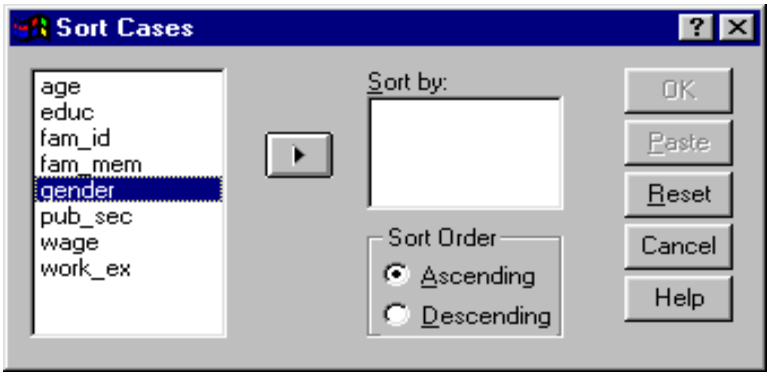
You can create several such aggregated files using different break variables (e.g. - *age* and *gender*, etc.). In the former there will be as many observations as there are *age* levels. In the latter the new data set will be aggregated to a further level (so that male and 12 years is one observation, female and 12 years is another).

Ch 1. Section 5 Sorting

Sorting defines the order in which data are arranged in the data file and displayed on your screen. When you sort by a variable, X, then you are arranging all observations in the file by the values of X, in either increasing or decreasing values of X. If X is text variable, then the order is alphabetical. If it is numerical, then the order is by magnitude of the value.

Sorting a data set is a prerequisite for several procedures, including split file, replacing missing values, etc.

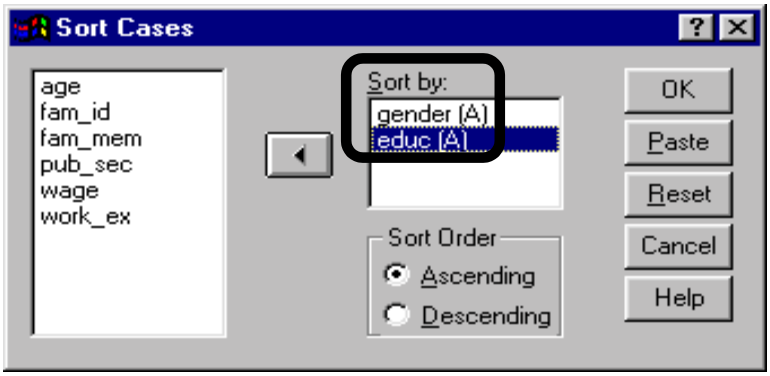
Go to DATA/ SORT.



Click on the variables by which you wish to sort.

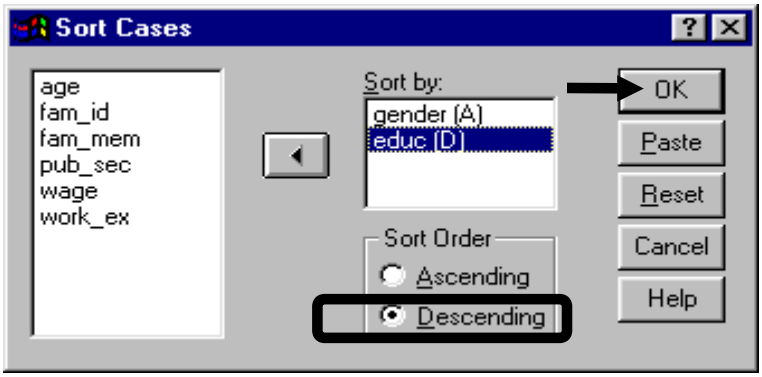
Move these variables into the box "Sort by."

The order of selection is important - first the data file will be organized by *gender*. Then within each *gender* group, the data will be sorted by *education*. So, all males (*gender*=0) will be before any female (*gender*=1). Then, within the group of males, sorting will be done by *education* level.



Let's assume you want to order *education* in reverse - highest to lowest. Click on *educ* in the box "Sort by" and then choose the option "Descending" in the area "Sort Order."

Click on "OK."



Example of how the sorted data will look (ascending in *gender*, then descending in *educ*)

gender	*educ*	*wage*	*age*	*work_ex*
0	21	34	24	2
0	15	20	23	2
0	8	21	25	5
0	0	6	35	20
1	12	17	45	25
1	8	14	43	27
1	6	11	46	25
1	3	7	22	2

Ch 1. Section 6 Reducing sample size

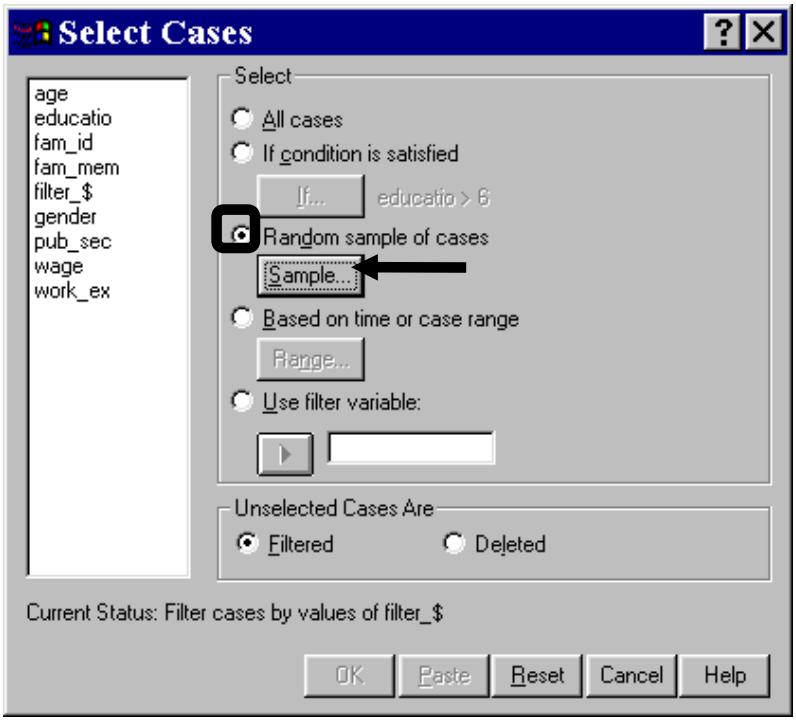
Ch 1. Section 6.a. Using random sampling

Let's assume you are dealing with 2 million observations. This creates a problem - whenever you run a procedure, it takes too much time, the computer crashes and/or runs out of disk space. To avoid this problem, you may want to pick only 100,000 observations, chosen randomly, from the data set.

Go to DATA/SELECT CASES.

Select the option "Random Sample of Cases" by clicking on the round button to the left of it.

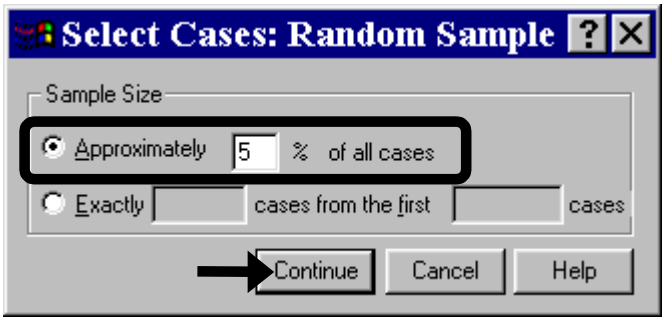
Click on the button "Sample."



Select the option "Approximately" by clicking on the round button to the left of it.

Type in the size of the new sample relative to the size of the entire data set. In this example the relative size is 5% of the entire data - SPSS will randomly select 100,000 cases from the original data set of 2 million.

Click on "Continue."

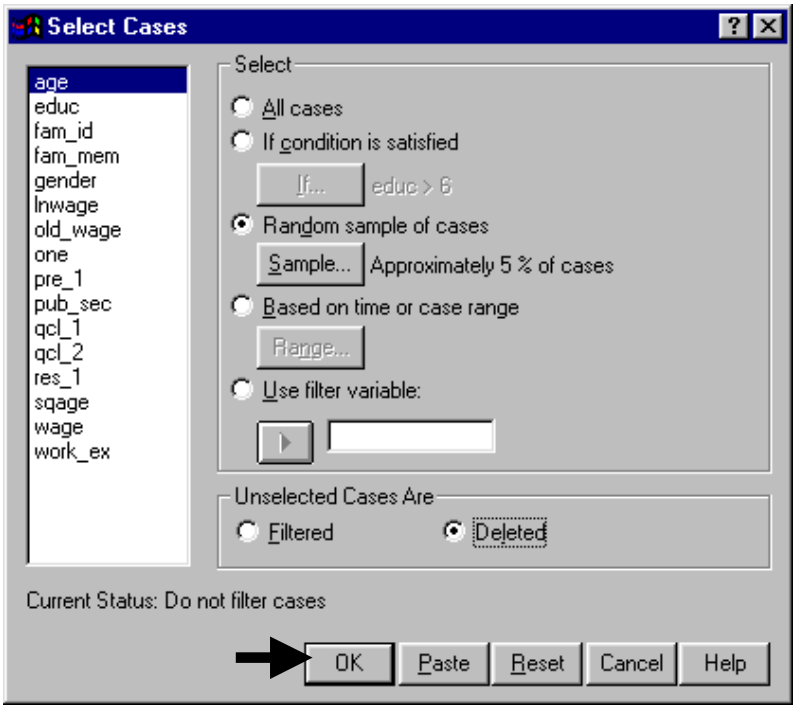


On the bottom, choose "Deleted."

Click on "OK"

Save this sample data set with a new file name.

Note: Be sure that you cannot use the original data set before opting to use this method. A larger data set produces more accurate results.



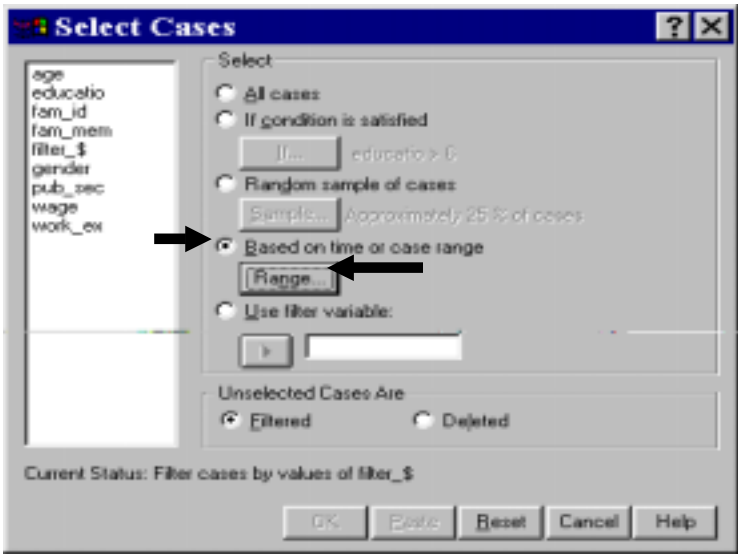
Ch 1. Section 6.b. Using a time/case range

You can also select cases based on time (if you have a variable that contains the data for time) or case range. For example, let's assume that you have time series data for 1970-1990 and wish to analyze cases occurring only after a particular policy change that occurred in 1982.

Go to DATA/SELECT CASES.

Select the option "Based on Time or Case Range" by clicking on the round button to the left of it.

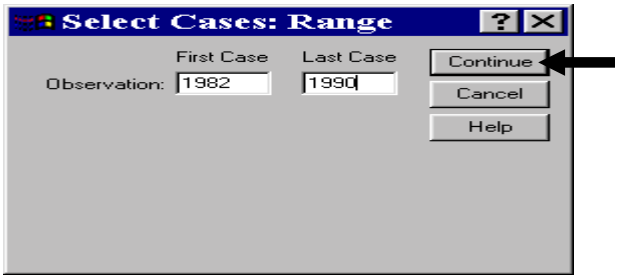
Click on the button "Range."



Enter the range of years to which you wish to restrict the analysis.

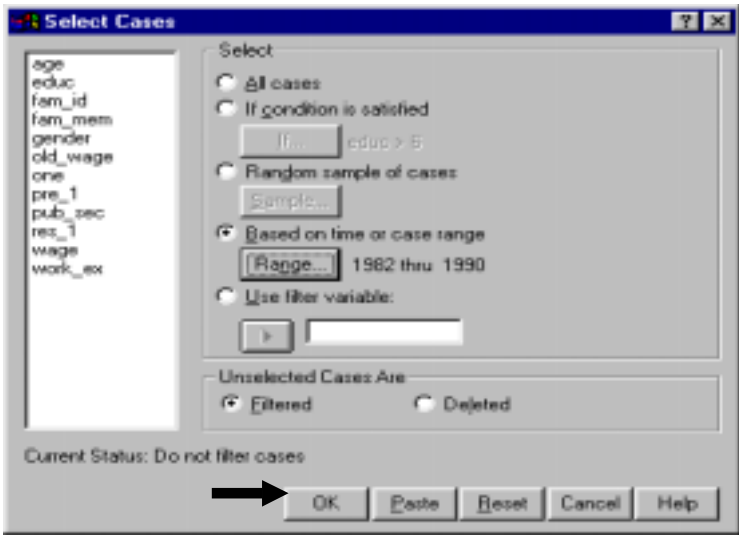
Note: The data set must have the variable "Year."

Click on “Continue.”



Select the option “Filtered” or “Deleted” in the bottom of the dialog box[20].

Click on “OK”



Ch 1. Section 7 Filtering data

It will often be the case that you will want to select a Sub-set of the data according to certain criteria. For example, let's assume you want to run procedures on only those cases in which *education* level is over 6. In effect, you want to temporarily “hide” cases in which *education* level is 6 or lower, run your analysis, then have those cases back in your data set. Such data manipulation allows a more pointed analysis in which sections of the sample (and thereby of the population they represent) can be studied while disregarding the other sections of the sample.

Similarly, you can study the statistical attributes of females only, adult females only, adult females with high school or greater *education* only, etc[21]. If your analysis, experience, research or knowledge indicates the need to study such sub-set separately, then use DATA/ SELECT CASE to create such sub-sets.

[20] If you choose “Filtered” then the cases that were not selected will not be used in any analysis, but will also not be deleted. Rather, they will be hidden. In the event that you wish to use those cases again, go to DATA/ SELECT CASES and choose the first option, “All Cases.”

[21] Apart from allowing you to concentrate on a Sub-set of the sample, SELECT CASE (or filtering, as it is often called) creates dummy variables on the basis of the subgroup you filter. Let's assume you have used DATA/ SELECT CASE to filter in adult females. SPSS will create a new variable that takes the value of 1 for the filtered in observations (i.e. - for adult females) and a value of 0 for all other observations. This dummy variable may be used in regression and other analysis and, even more importantly, in running a comparative analysis to compare the differences in statistical and graphical results for adult females versus the rest (see chapter 10). Ignore this footnote if it is too complex or seems irrelevant. We will get back to the use of filtering later in the book. Within the proper context, the usefulness will become apparent.

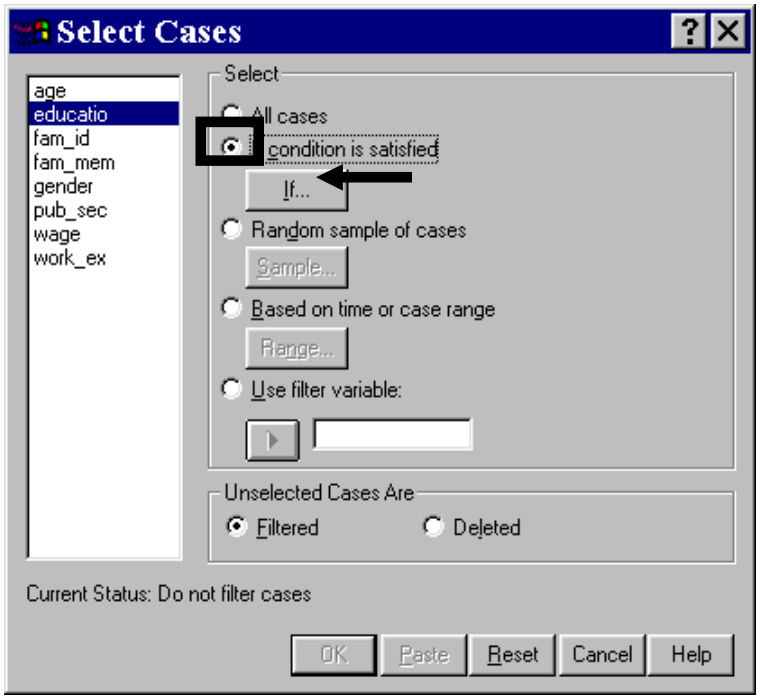
Ch 1. Section 7.a. A simple filter

Suppose you want to run an analysis on only those cases in which the respondents *education* level is greater than 6. To do this, you must filter out the rest of the data.

Go to DATA/ SELECT CASE

When the dialog box opens, click on "If condition is satisfied."

Click on the button "If."



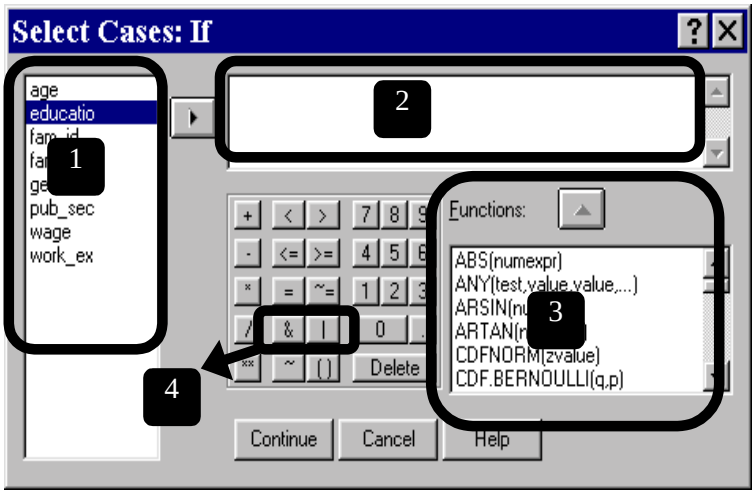
The white boxed area "2" in the upper right quadrant of the box is the space where you will enter the criterion for selecting a Sub-set.

Such a condition must have variable names. These can be moved from the box on the left (area "1").

Area "3" has some functions that can be used for creating complex conditions.

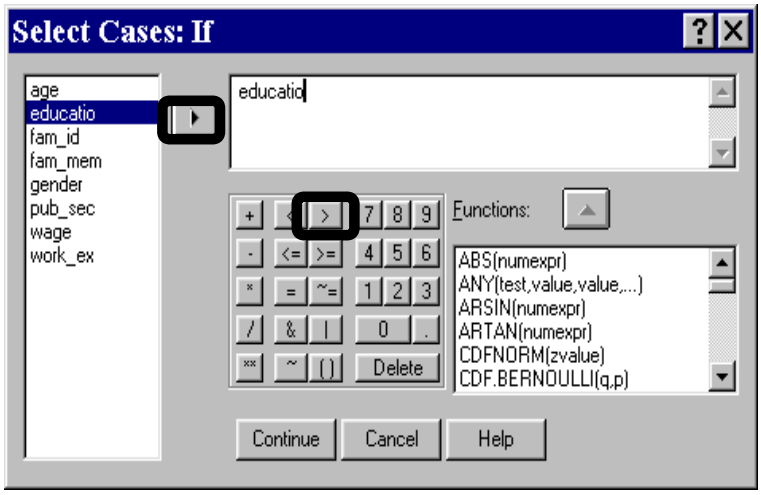
Area "4" has two buttons you will use often in filtering: "&" and "|" (for "or").

As you read this section, the purpose and role of each of these areas will become apparent.



Select the variable you wish to use in the filter expression (i.e. - the variable on the basis of whose values you would like to create a Sub-set). In this example, the variable is *educatio*.

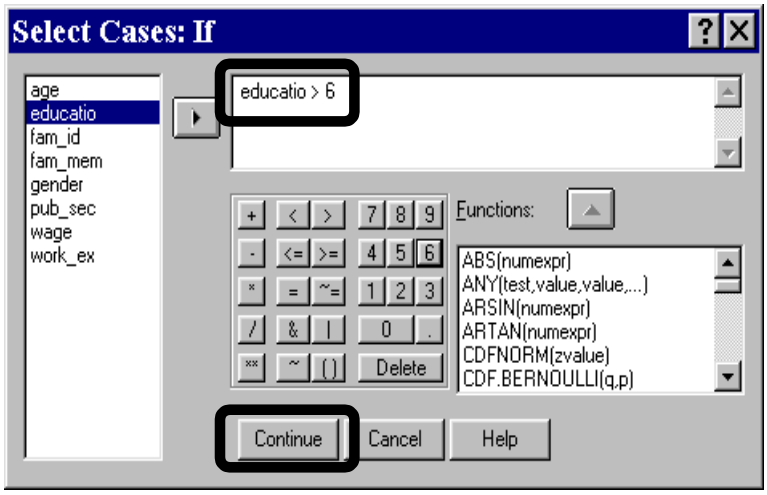
Click on the right arrow to move the variable over to the white area on the top of the box.



Using the mouse, click on the greater than symbol (">") and then the digit 6. (Or you can type in ">" and "6" using the keyboard.)

You will notice that SPSS automatically inserts blank spaces before and after each of these, so if you choose to type the condition you should do the same.

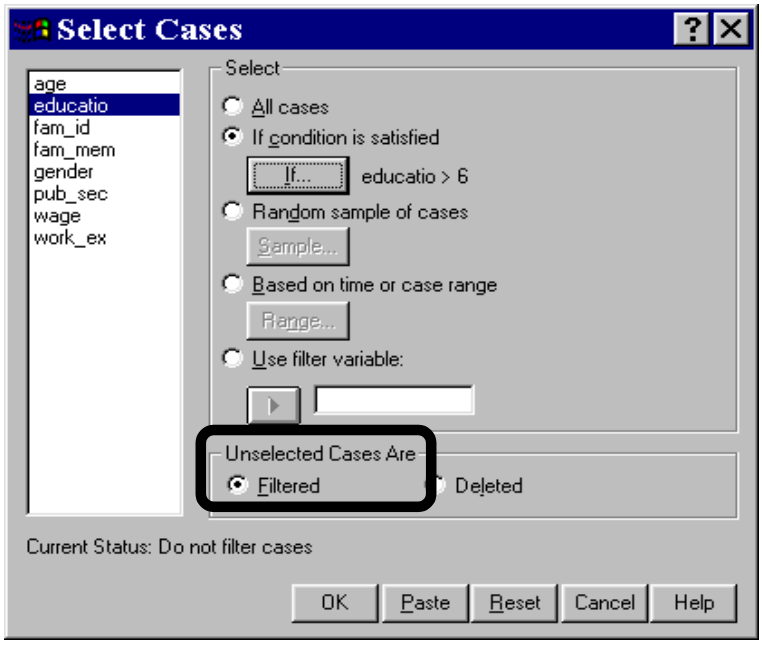
Click on "Continue."



You'll see that the condition you specified (If *educatio* > 6) is in this dialog box.

Move to the bottom of the box that says "Unselected Cases Are" and choose "Filtered." Click on "OK."[22]

Do not choose "Deleted" unless you intend to delete these cases permanently from your data set as, for example, when you want to reduce sample size. (We don't recommend that you delete cases. If you do want to eliminate some cases for purposes of analysis, just save the smaller data set with a different file name.)



[22] SPSS creates a "filter variable" each time you run a filter. This variable takes the value 1 if the case satisfies the filter criterion and 0 if it does not. This variable can be used as a dummy variable (see chapters 3-9). Also, if you

The filtered-out data have a diagonal line across the observation number. These observations are not used by SPSS in any analysis you conduct with the filter on.

25 : pub_noc

	fam_id	fam_mem	wage	pub_noc	work_ex	gender	age
14	170625	1	47.35	0	12.83	0	33
15	241224	1	30.87	0	18.00	0	52
16	110118	3	6.25	0	2.00	0	24
17	110122	3	6.82	0	1.33	0	35
18	110124	1	8.75	0	10.00	0	31
19	110132	4	4.26	0	3.00	0	27
20	110137	4	6.82	0	.33	0	28
21	110213	6	7.95	0	.92	0	27
22	110213	7	5.68	0	.08	0	21
23	110214	3	1.50	0	.33	0	18
24	110214	5	6.25	0	9.00	0	33
25	110230	1	8.52	0	.83	0	29
26	110316	10	6.19	0	.50	0	20
27	110335	4	6.82	0	8.00	0	23

Ch 1. Section 7.b. What to do after obtaining the sub-set

Now that the data set is filtered, you can run any analysis (see chapters 3-10 for discussions on different procedures). The analysis will use only the filtered cases (those not crossed out).

Ch 1. Section 7.c. What to do after the sub-set is no longer needed

After you have performed some procedures, use "All Cases" to return to the original data set.

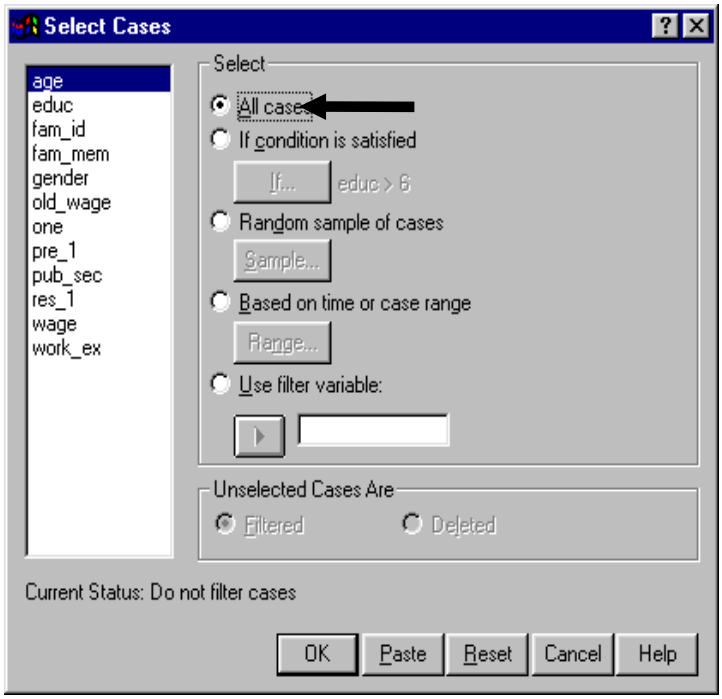
Do not forget this step. Reason: You may conduct other procedures (in the current or next SPSS session) forgetting that SPSS is using only a Sub-set of the full data set. If you do so, your interpretation of output would be incorrect.

want to compare more deeply between the filtered and unfiltered groups, use the filter variable as a criterion for comparative analysis (see chapter 10).

Go to DATA/ SELECT CASE

Select "All cases" at the top.

Click on "OK."



Ch 1. Section 7.d. Complex filter: choosing a Sub-set of data based on criterion from more than one variable

Often you will want or need a combination filter that bases the filtering criterion on more than one variable. This usually involves the use of "logical" operators. We first list these operators.

LOGICAL COMMAND	SYMBOL	DESCRIPTION
Blank	.	For choosing missing values.
Greater than	>	Greater than
Greater than or equal to	>=	Greater than or equal to
Equal to	=	Equal to
Not equal to	~=	Not equal to[23].
Less than	<	Less than
Less than or equal to	<=	Less than or equal to
Or	|	This means "satisfies EITHER criteria." Let's assume you want to run analysis on all females and all public *sector* employees. The condition would be (*gender*=1 | *pub_sec*=1).

[23] The symbol "~" can be obtained by pressing down on the shift button and clicking on the apostrophe button on the upper left portion of the keyboard.

LOGICAL COMMAND	SYMBOL	DESCRIPTION
And	**&**	This means "satisfies BOTH criteria." For example, if you want to isolate public sector (*pub_sec*=1) females (*gender*=1), your condition would be *pub_sec*=1 & *gender* =1.

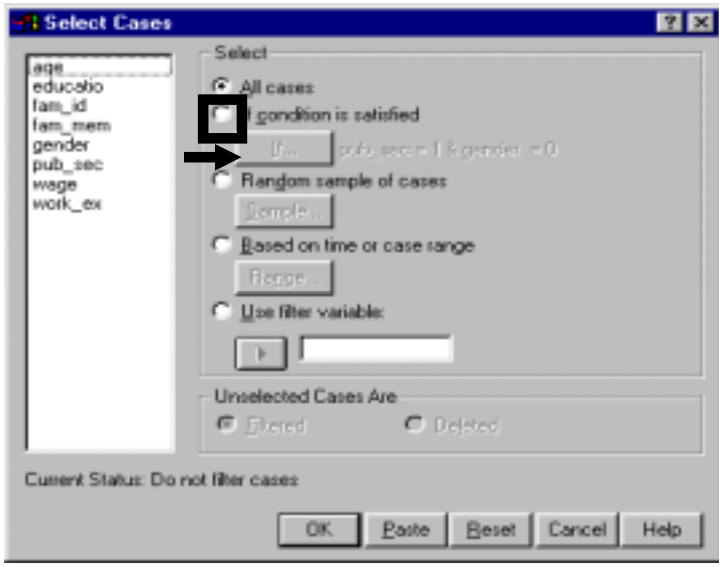
Example 1

Let's assume you want to select only those male employees (*gender*=0) who work in the public *sector* (*pub_sec* = 1).

Go to DATA/ SELECT CASE

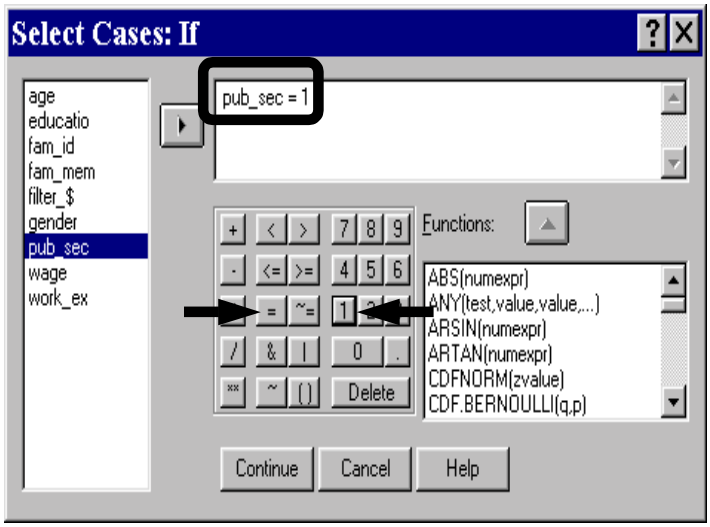
When the dialog box opens, click on "If condition is satisfied."

Click on the button "If."



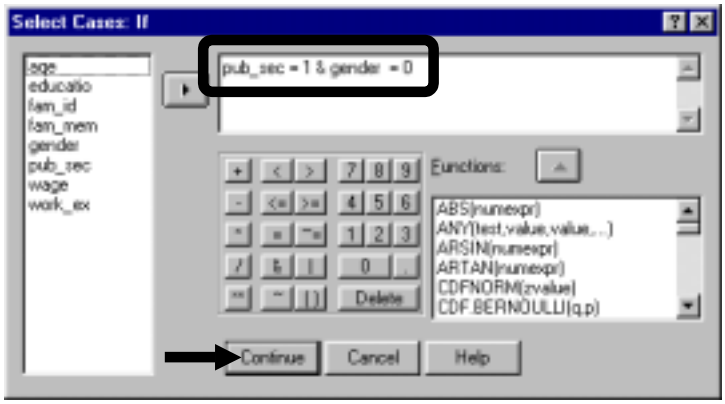
You want to look at cases in which the respondent is a public *sector* employee (*pub_sec* =1).

Select *pub_sec* from the list of variables, then choose or click on the buttons for "=" and then for "=1." (Or type in "=" and "1" from the keyboard.)



Use the "&" sign and then type the second condition to specify that both conditions must be met - in the filtered data, every respondent must be a male (*gender*=0) and must be working in the public sector.

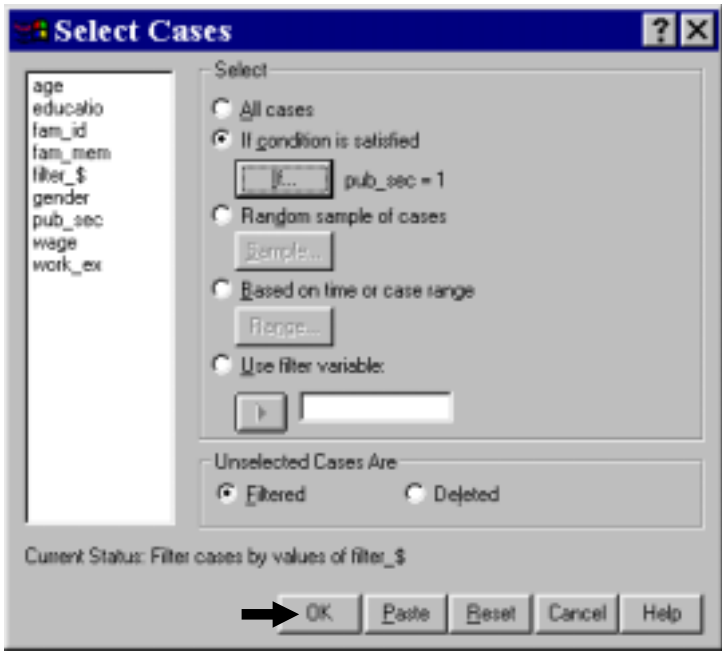
Click on "Continue."



Click on "OK."

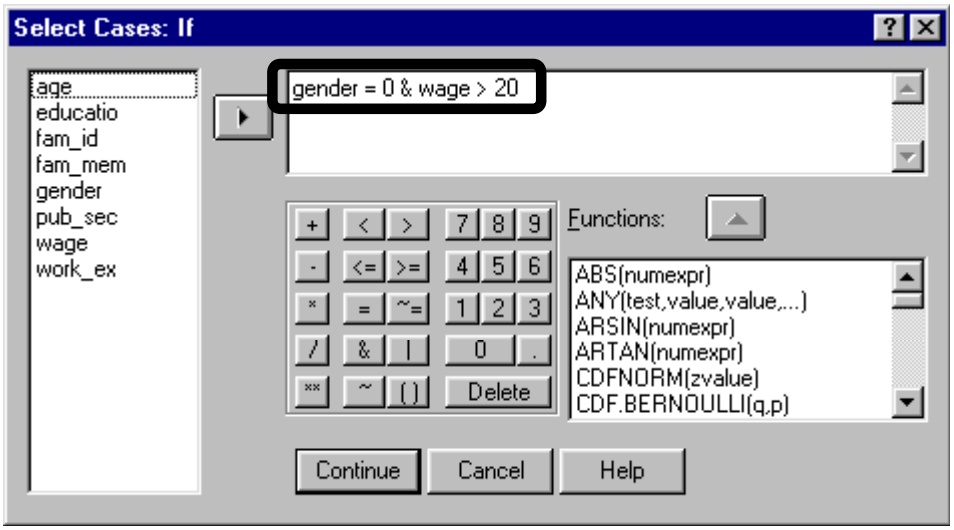
Now any analysis you conduct will use only those cases in which the respondents meet the filter criterion, i.e. - male public sector employees. (See section 1.7.b.)

Note: After completing the required analysis, remove the filter by going back to DATA/ SELECT CASE and choosing the topmost option "All Cases." (See section 1.7.c).



Example 2: Adult Females Only

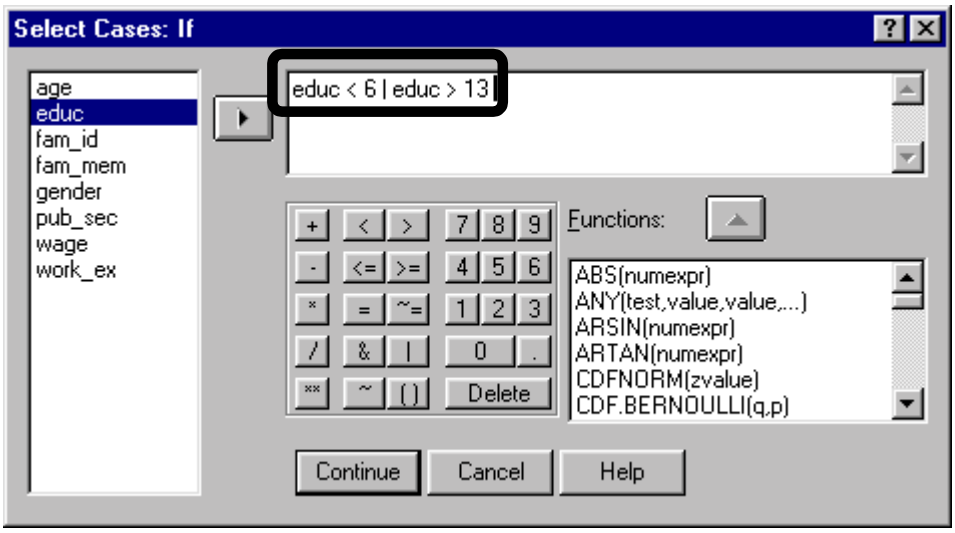
Now let us assume that you want to select cases where the respondent is a female (*gender*=1) and her *wage* is above twenty (*wage* > 20).
To do so, choose DATA / SELECT CASES, and "If Condition is Satisfied." (See section 1.7.a for details on the process involved.) In the large white window, you want to specify female (*gender* =1) and wages above twenty (*wage*>20). Select *gender* = 1 & *wage* > 20.



Now you can conduct analysis on "Adult Females only." (See sections 1.7.b and 1.7.c.)

Example 3: Lowest or Highest Levels of *Education*

Let's assume you want to choose the lowest or highest levels of *education* (*education* < 6 or *education* > 13). Under the DATA menu, choose SELECT CASES and "If Condition is Satisfied" (See section1.7.a for details on the process involved). In the large white window, you must specify your conditions. Remember that the operator for "or" is "|" which is the symbol that results from pressing the keyboard combination "SHIFT" and "\." Type in "*educ* < 6 | *educ* > 13" in the large white window.



Now you can conduct analysis on "Respondents with Low or High Education only." (See sections 1.7.b and 1.7.c.)

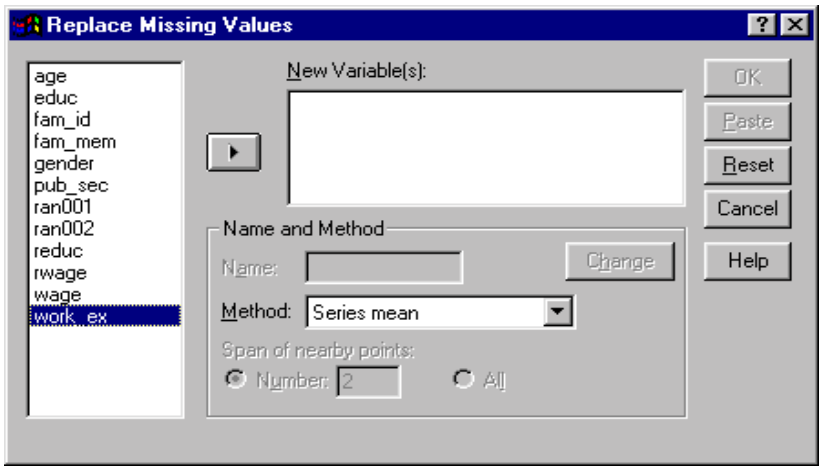
Ch 1. Section 8 Replacing missing values

In most data sets, missing values are a problem. For several procedures, if the value of even one of the variables is missing in an observation, then the procedure will skip that observation/case altogether!

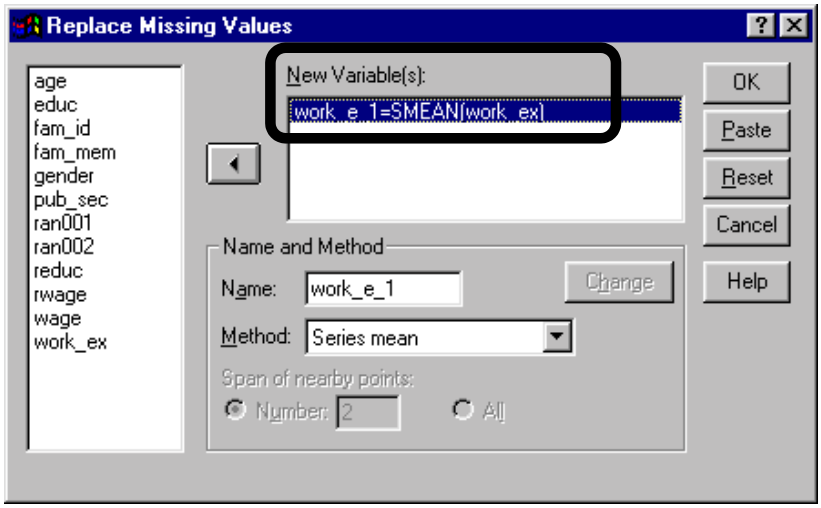
If you have some idea about the patterns and trends in your data, then you can replace missing values with extrapolations from the other non-missing values in the proximity of the missing value. Such extrapolations make more sense for, and are therefore used with, time series data. If you can arrange the data in an appropriate order (using DATA/ SORT) and have some sources to back your attempts to replace missing values, you can even replace missing values in cross-sectional data sets - but only if you are certain.

Let's assume *work_ex* has several missing values that you would like to fill in. The variable *age* has no missing values. Because *age* and *work_ex* can be expected to have similar trends (older people have more *work experience*), you can arrange the data file by *age* (using DATA /SORT and choosing *age* as the sorting variable - see section 1.5) and then replacing the missing values of *work_ex* by neighboring values of itself.

Go to TRANSFORM/ REPLACE MISSING VALUES.

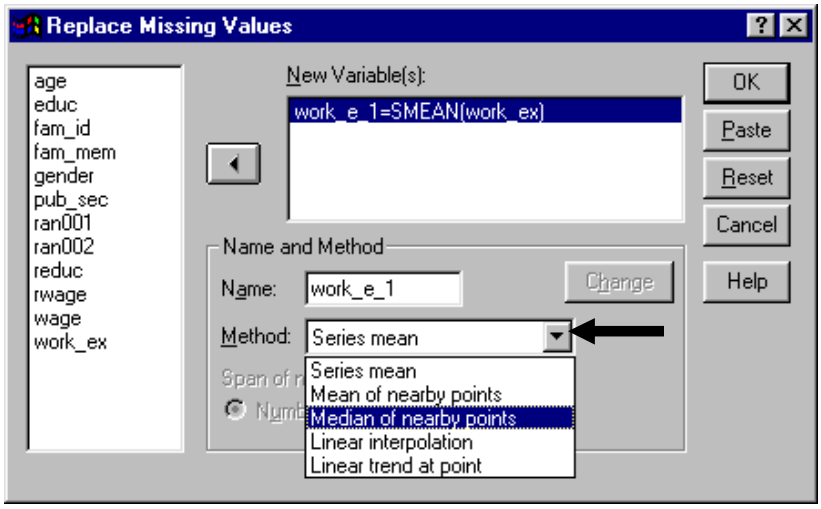


Select the variable *work_ex* and move it into the box "New Variable(s)."

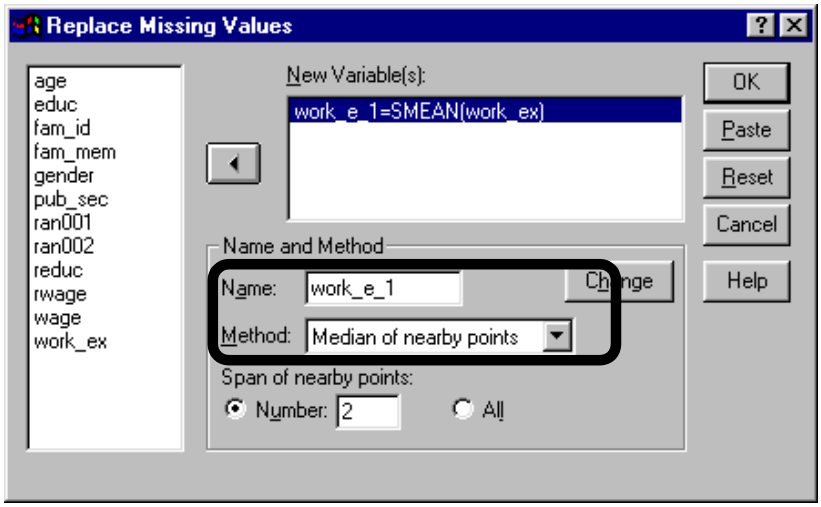


Click on the downward arrow next to the list box "Method."

Select the method for replacing missing values. We have chosen "Median of nearby points." (Another good method is "Linear interpolation," while "Series mean" is usually unacceptable).

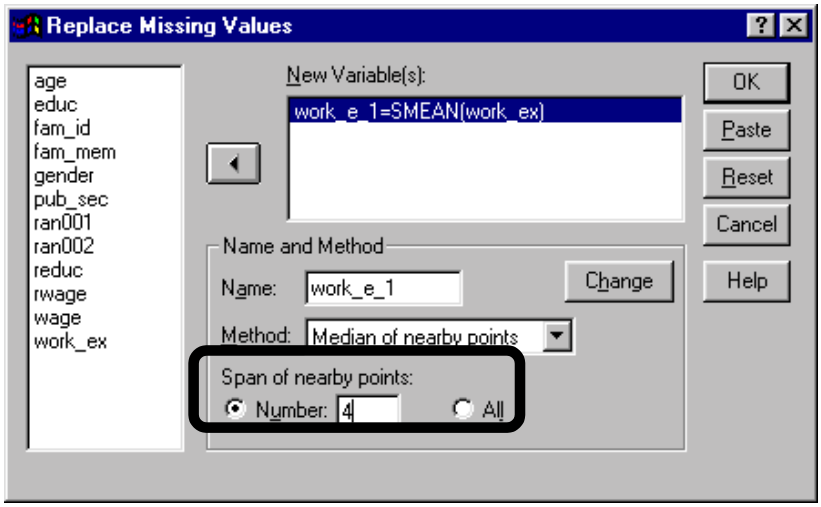


We do not want to change the original variable *work_ex*, so we will allow SPSS to provide a name for a new variable *work_e_1*[24].



The criterion for "nearby" must be given.

Go to the area in the bottom of the screen, "Span of nearby points," and choose a number (we have chosen 4). The median of the 4 nearest points will replace any missing value of *work_ex*.

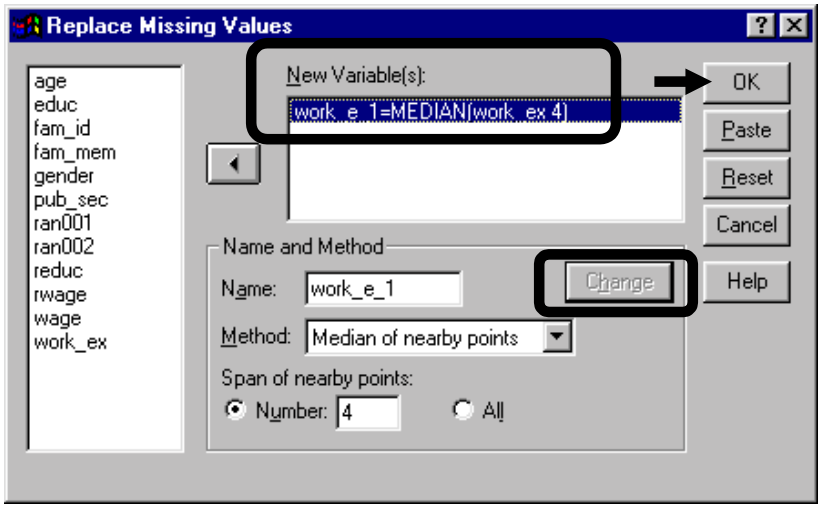


[24] This is a safe strategy - many statisticians are skeptical of analysis in which the analyst has let the computer fill in missing values. We therefore suggest that you not let the original variable change as you may need it again for future analysis.

Click on “Change.”

The box “New Variable(s) now contains the correct information.

Click on “OK.”

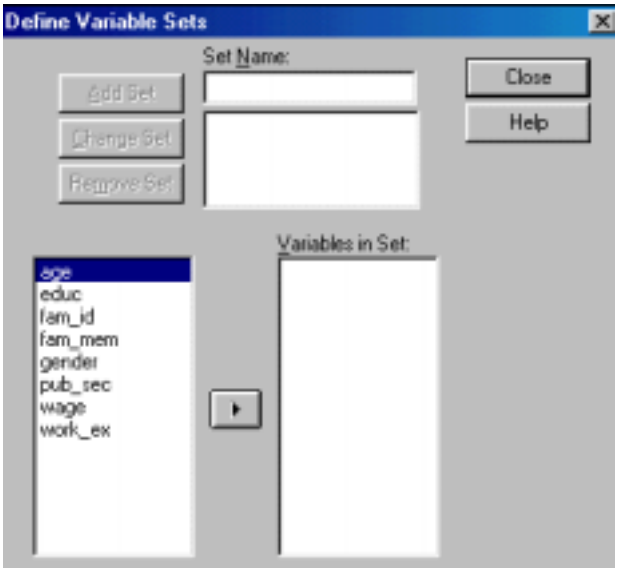


Ch 1. Section 9 Using Sub-sets of variables (and not of cases, as in section 1.7)

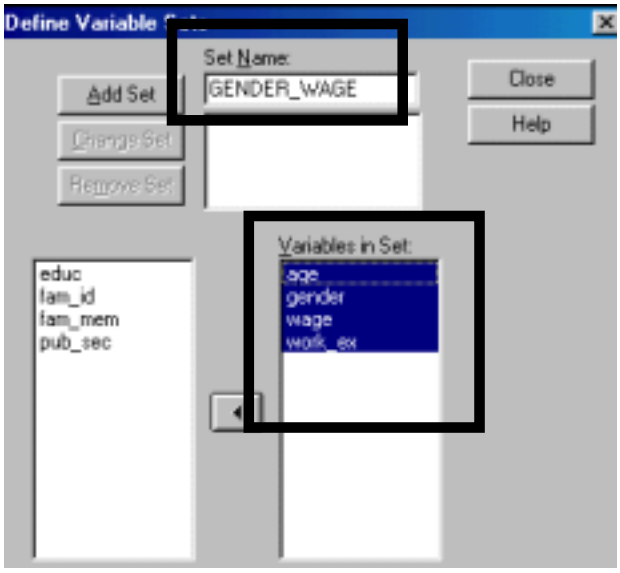
You may have a data set with a large number of variables. Each time you wish to run a procedure you will inevitably spend a great deal of time attempting to find the relevant variables. To assist you in this process SPSS includes a feature whereby restricts the variables shown in a procedure to those you wish to use This can be done by using options in the UTILITY menu.

Example: for a certain project (let's assume it is “Analysis of Gender Bias in Earnings”) you may need to use a certain Sub-set of variables. For a different project (let's assume it is “Sectoral Returns to Education”), you may need to use a different set of variables.

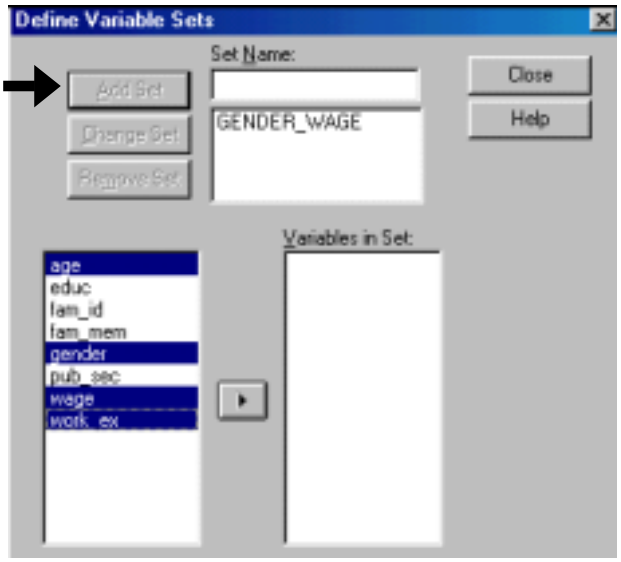
We first must define the two sets. Go to UTILITY / DEFINE SETS.



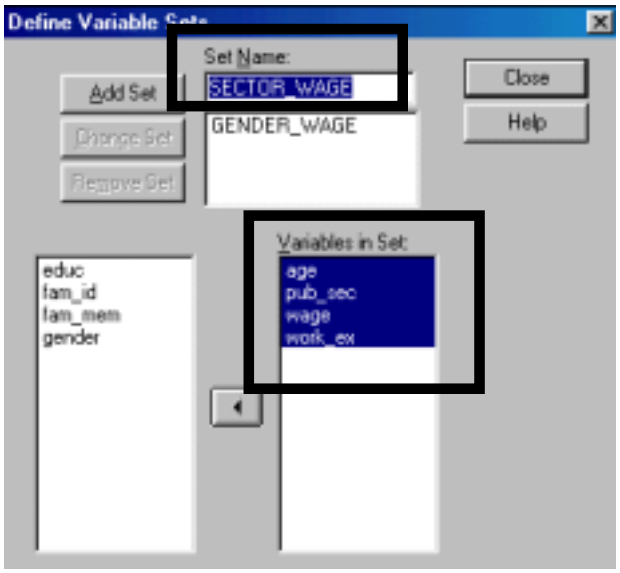
Move the variables you would like to be included in the set into the box "Variables in Set." Then name the set by typing in a name in the box "Set Name."



We still require one more set. To do this, first click on "Add Set."

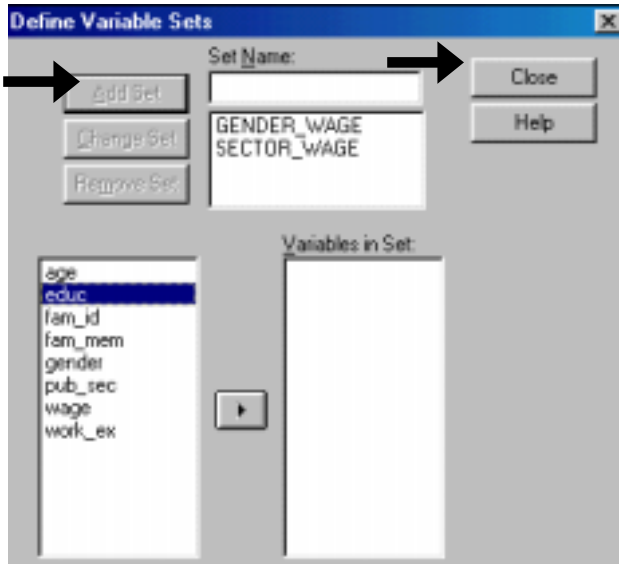


Move the variables you would like included in the set into the box "Variables in Set." Then name the set by typing in a name in the box "Set Name."



Click on “Add Set.”

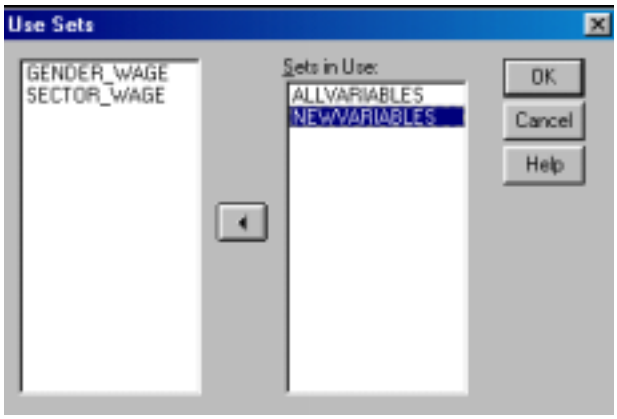
Click on “Close.”



Now, if you wish. you can restrict the variables shown in any dialog box for any procedure to those defined by you in the set. Go to UTILITY / USE SETS.

If you want to use the set “GENDER_WAGE,” then move it into the right box and move the default option “ALLVARIABLES” out. Click on "OK."

Now if you run a regression, the dialog box will only show the list of 4 variables that are defined in the set GENDER_WAGE.



To take quizzes on topics within each chapter, go to http://www.spss.org/wwwroot/spssquiz

Ch 2. CREATING NEW VARIABLES

Your project will probably require the creation of variables that are imputed/computed from the existing variables. Two examples illustrate this:

1. Let's assume you have data on the economic performance of the 50 United States. You want to compare the performance of the following regions: Mid-west, South, East Coast, West Coast, and other. The variable *state* has no indicator for "region." You will need to create the variable *region* using the existing variable *state* (and your knowledge of geography).

2. You want to run a regression in which you can obtain "the **%** effect on wages of a one year increase in education attainment." The variable *wage* does not lend itself to such an analysis. You therefore must create and use a new variable that is the natural log transformation of *wage*. In section 2.1, after explaining the concept of dummy and categorical variables, we describe how to create such variables using various procedures. We first describe recode, the most used procedure for creating such variables. Then we briefly describe other procedures that create dummy or categorical variables - automatic recode and filtering[25] (the variables are created as a by-product in filtering).

In section 2.2, we show how to create new variables by using numeric expressions that include existing variables, mathematical operators, and mathematical functions (like square root, logs, etc).

Section 2.3 explains the use of "Multiple Selection Sets." You may want to skip this section and come back to it after you have read chapters 3 and 6.

Section 2.4 describes the use of the count procedure. This procedure is used when one wishes to count the number of responses of a certain value across several variables. The most frequent use is to count the number of "yeses" or the "number of ratings equal to value X."

Let's assume that you wish to create a variable with the categories "High, mid, and low income groups" from a continuous variable *wage*. If you can define the exact criteria for deciding the range of values that define each income range, then you can create the new variable using the procedures shown in section 2.1. If you do not know these criteria, but instead want to ask SPSS to create the three "clusters" of values ("High," "Mid," and "Low") then you should use "Cluster Analysis" as shown in section 2.5.

You may want to use variables that are at a higher level of aggregation than in the data set you have. See section 1.4 to learn how to create a new "aggregated" data set from the existing file.

[25] See section 1.7 for a detailed discussion on filtering.

Ch 2. Section 1 Creating dummy, categorical, and semi-continuos variables using recode

TRANSFORM/ RECODE is an extremely important tool for social science statistical analysis. Social scientists are often interested in comparing the results of their analysis across qualitative sub-groups of the population, e.g. - male versus female, White-American compared to African-American, White-American compared to Asian-American, etc. A necessary requirement for such analysis is the presence in the data set of dummy or categorical variables that capture the qualitative categories of gender or race.

Once the dummy or categorical variables have been created, they can be used to enhance most procedures. In this book, any example that uses *gender* or *pub_sec* as a variable provides an illustration of such an enhancement. Such variables are used in many procedures:

- In regression analysis as independent variables (see chapters 7 and 8)
- In Logit as dependent and independent variables (see chapter 9)
- In bivariate and trivariate analysis as the criterion for comparison of means, medians, etc.[26]
- As the basis for "Comparative Analysis" (chapter 10). Using this, **all** procedures, including univariate methods like descriptives, frequencies, and simple graphs, can be used to compare across sub-groups defined by dummy or categorical variables.

Ch 2. Section 1.a. What are dummy and categorical variables?

A dummy variable can take only two values (usually 0 or 1)[27]. One of the values is the indicator for one category (e.g. - male) and the other for another category (e.g. - female).

Value	Category
0	Male
1	Female

Categorical variables can take several values, with each value indicating a specific category. For example, a categorical variable "Race" may have six values, with the values-to-category mapping being the following:

Value	Category
0	White-American
1	African-American
2	Asian-American
3	Hispanic-American

[26] Using graphs, boxplots, custom tables, etc. (see chapters 5 and 6).

[27] If you use the dummy variable in regression, Logit, and some other procedures, the coding must be 0 and 1. If the original coding is 1 and 2, you should change it to 0 and 1 using the procedure shown in section 2.1.c

Value	Category
4	Native-American
5	Other

Dummy and categorical variables can be computed on a more complex basis. For example:

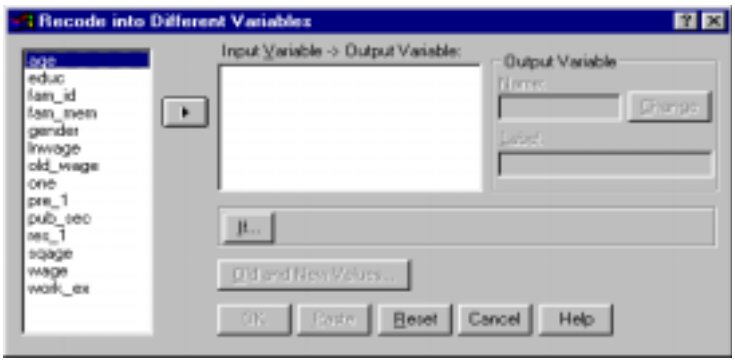
Value	Category
0	*wage* between 0 and 20
1	*wage* above 20

Ch 2. Section 1.b. Creating new variables using recode

Let's assume that we want to create a new dummy variable *basiced* (basic education) with the following mapping[28]:

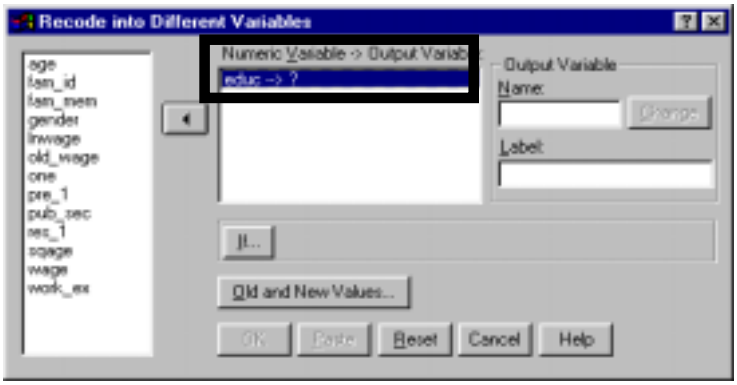
Old Variable- *educ*	New Variable - *basiced*
0-10	1
11 and above	0
Missing	Missing
All else	Missing

Go to TRANSFORM/ RECODE/ INTO NEW VARIABLES.

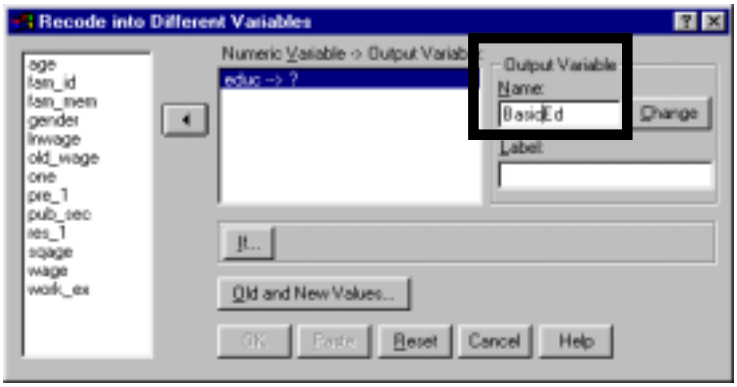


[28] It is a good practice to write down the mapping in tabular format before you start creating the new variable.

Select the variable you wish to use as the basis for creating the new variable and move it into the box “Numeric Variable.” In our example, the variable is *educ*.

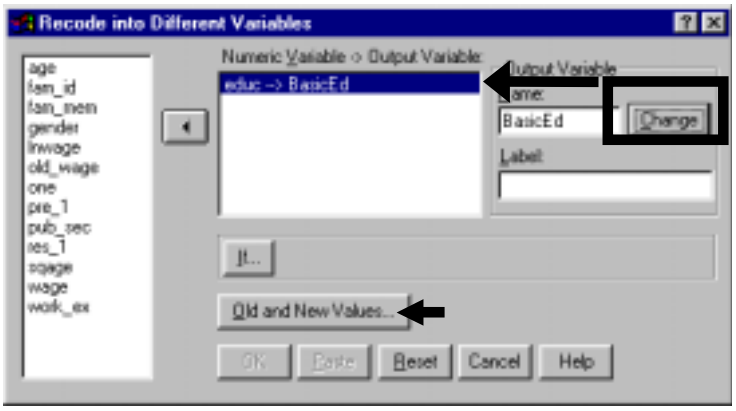


Enter the name for the new variable into the box “Output Variable.”



Click on the button “Change.” This will move the name *basiced* into the box "Numeric Variable→Output Variable."

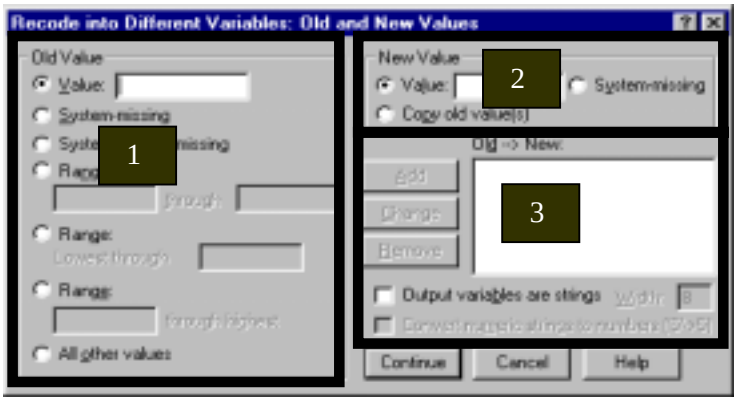
Click on the button “Old and New Values.”



This dialog box has three parts.

- Area 1 (“Old Value”) is the area in which you specify the values of the existing variable (that which is to be mapped from - in this example, that variable is *educ*).

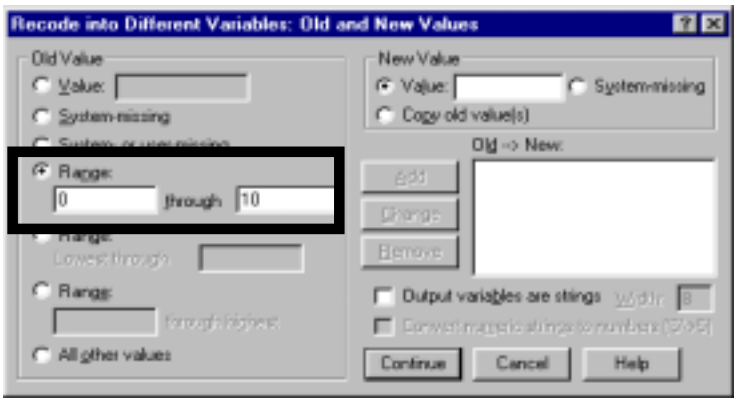
- Area 2 (“New Value”) is the area in which you specify the corresponding value in the new variable (that which is mapped into - in this example, that variable is



basiced).

- Area 3 ("Old→New") contains the mapping of the old variable to the new one.

Click on the button to the left of the label "Range." Enter the range 0 to 10.

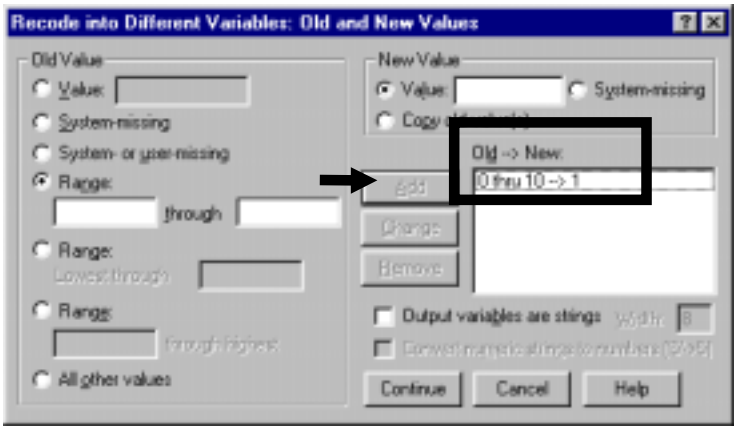


Now you must enter the new variable value that corresponds to 0 to 10 in the old variable.

In the "New Value" area, click on the button to the left of "Value" and enter the new value of 1 into the box to its right.

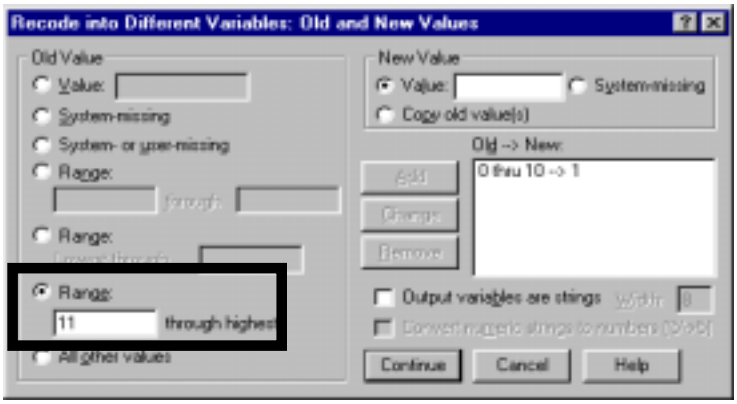
Click on the button "Add."

In the large white box "Old→New" you will see the mapping "0 thru 10→1."

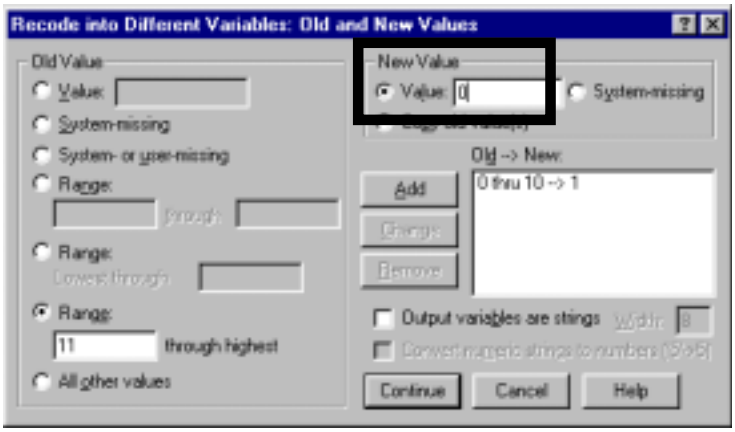


The second mapping we must complete is to make numbers 11 and higher in the old variable equal to 0 in the new variable.

Click on the button to the left of the label "Range ... through Highest" and enter the number 11.

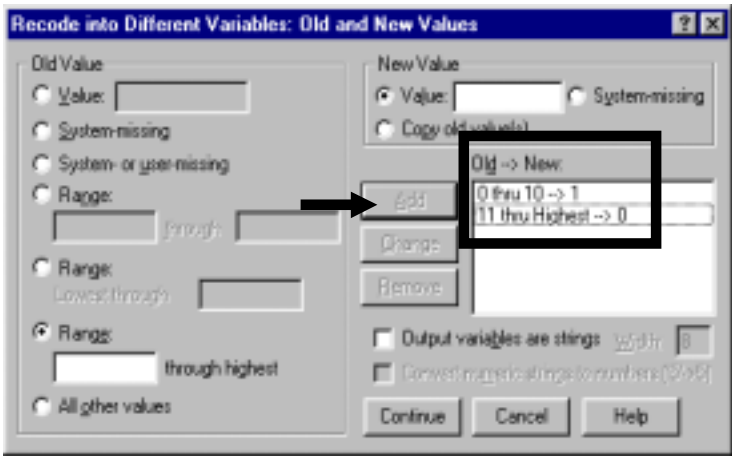


In the area "New Value" enter the number 0.



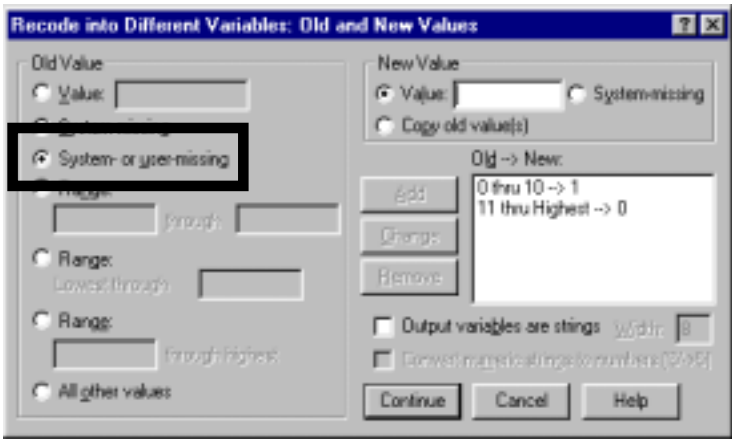
Click on "Add."

In the large white box "Old→New" you will see the mapping "11 thru Highest→0."

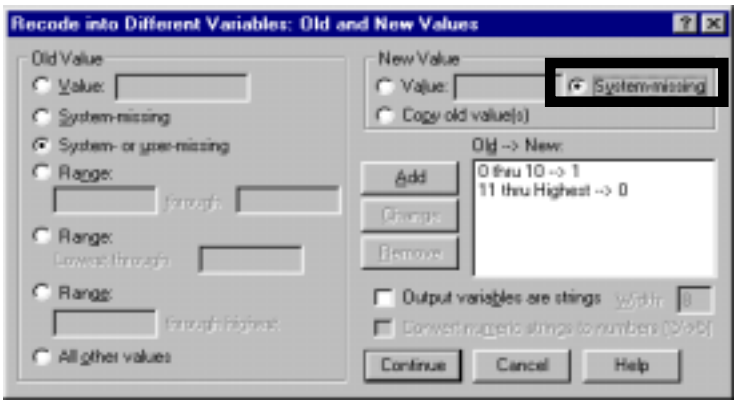


It is a good idea to specify what must be done to the missing values in the old variable.

To do this, click on the button to the left of the label "System or user-missing."[29]

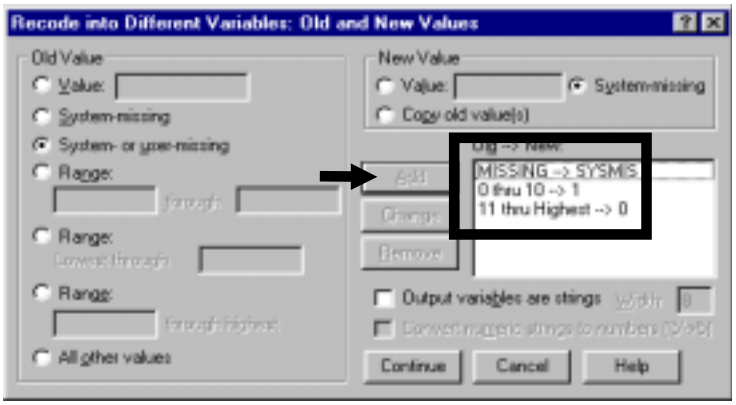


In the area "New Value," click on the button to the left of "System Missing."



Click on "Add."

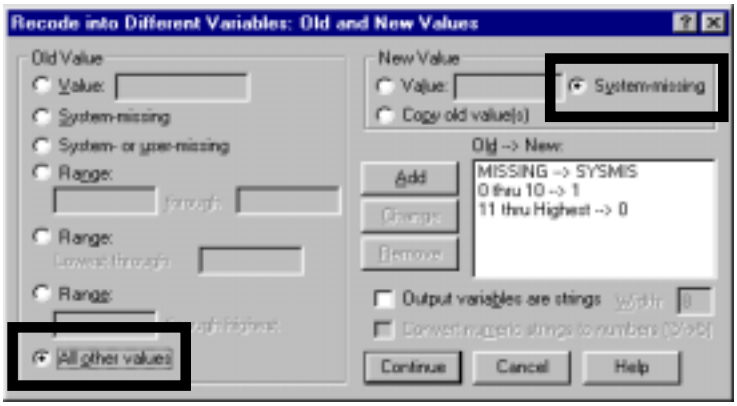
Compare this mapping to the required mapping (see the last table on page 2-2). It appears to be complete. It is complete, however, only if the original variable has no errors. But what if the original variable has values outside the logical range that was used for creating the original mapping? To forestall errors being generated from this possibility, we advise you to create one more mapping item.



[29] User-missing are the values defined by us (the "user") as missing in DATA/ DEFINE VARIABLE (see section 1.2.c). System-missing are the blank or empty data cells, defined by the system (i.e. - the data set) as missing. In the data sheet in SPSS, these cells have periods only. (In contrast, in Excel, the blank or empty cells have nothing in them.) Please be clear on the distinction between user-missing and system-missing. In the new variable the user-missing values from the original variable will be mapped into empty cells (with periods in them).

All other values (not between 0 and 10, not greater than 11, and not missing) are to be considered as missing[30].

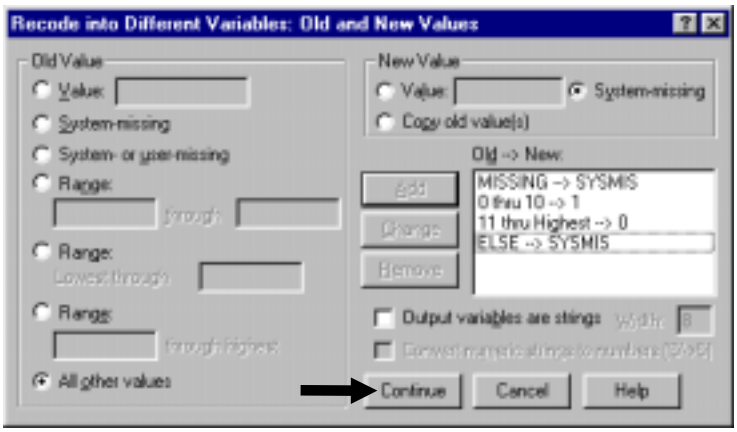
To do this, choose "All other values" in the area "Old Value" and choose the option "System missing" in the area "New Value."



Click on "Add."

The entire mapping can now be seen.

Click on "Continue."

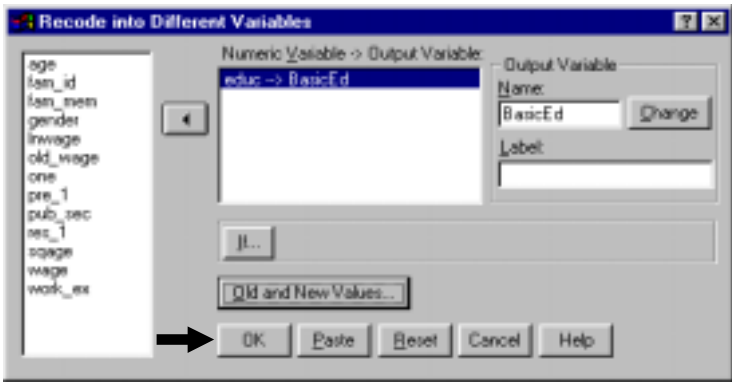


Click on "OK." A new variable *basiced* will be created.

The new variable will be in the last column of the data sheet.

Note: Go to DEFINE / VARIABLE and define the attributes of the new variable. See section 1.2 for examples of this process. In particular, you should create variable labels, value labels, and define the missing values.

The "If" option in the dialog box (see the button labeled "If") is beyond the scope of this book.



[30] This is a safety measure because we do not want any surprises from incorrect values in the new variable generated from incorrect data in the original variable. For example, the last mapping rule (see the entry ELSE→ SYSMIS) ensures that if any nonsensical or invalid values in the old variable (like, for example, an education level of "-1" in the variable *educ*) do not carry over into the new variable.

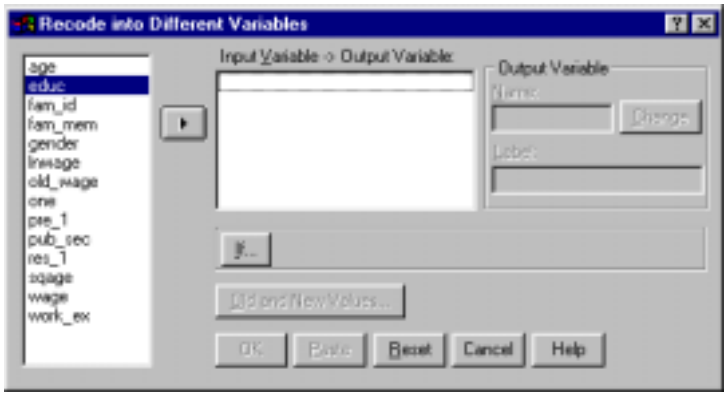
Example 2: continuous variable into a semi-continuous variable

Let's assume we want to create a new dummy variable, *educ2*, in which Master's or higher level education (17 or above) is recoded with one value, i.e. - 17. The other values remain as they are. The mapping is:

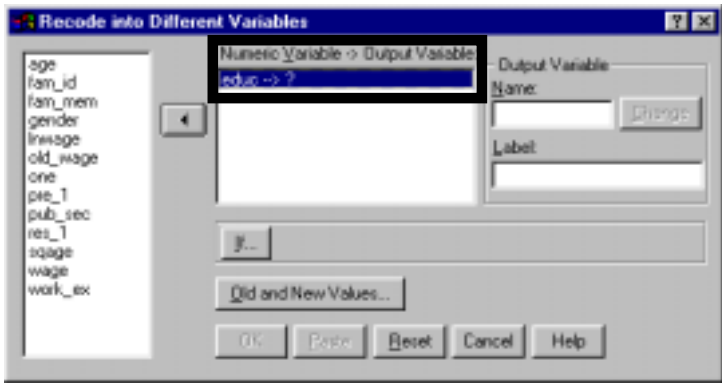
Old Variable- *educ*	**New Variable - *educ2***
17 and higher	17
0 to below 17	Same as old

Go to TRANSFORM/ RECODE/ INTO DIFFERENT VARIABLES.

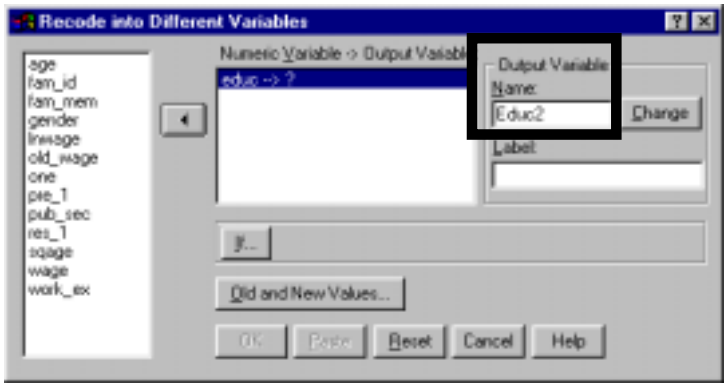
Note: We are repeating all the steps including those common to example 1. Please bear with us if you find the repetition unnecessary - our desire is to make this easier for those readers who find using SPSS dialog boxes difficult.



Select the variable you wish to use to create the new variable and move it into the box "Numeric Variable."

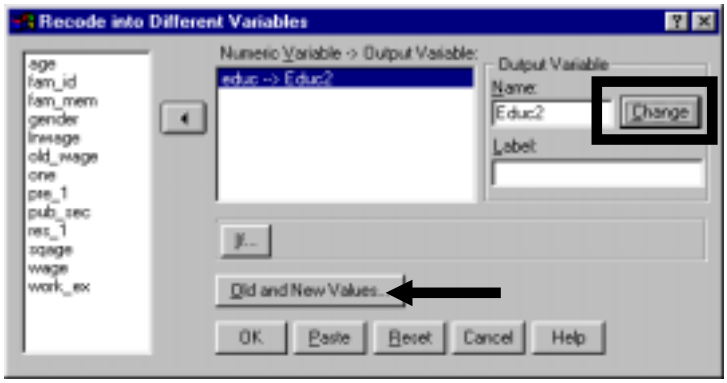


Enter the name for the new variable *educ2* into the box "Output Variable."



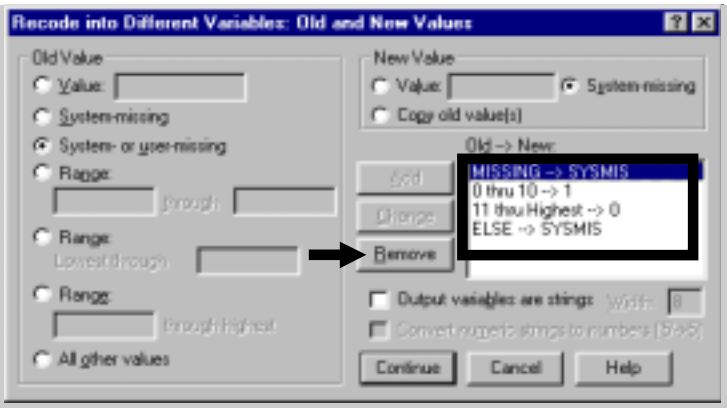
Click on the button "Change."

Click on the button "Old and New Values."

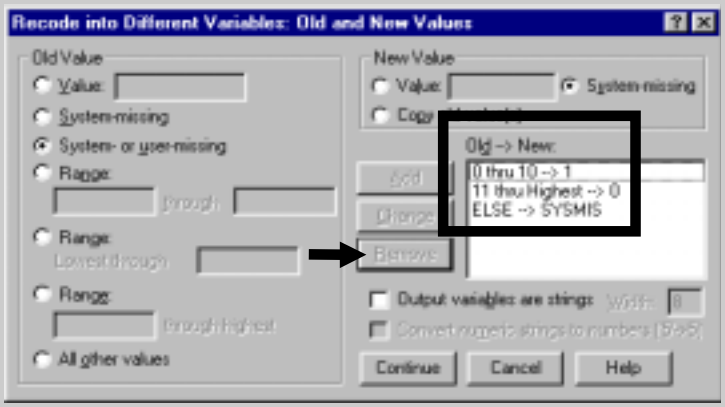


An aside: a simple way to save time in reusing a dialog box is presented on the right.
If you are working in a session in which you have previously created a recode, then you will have to remove the old mapping from the box "Old→New."

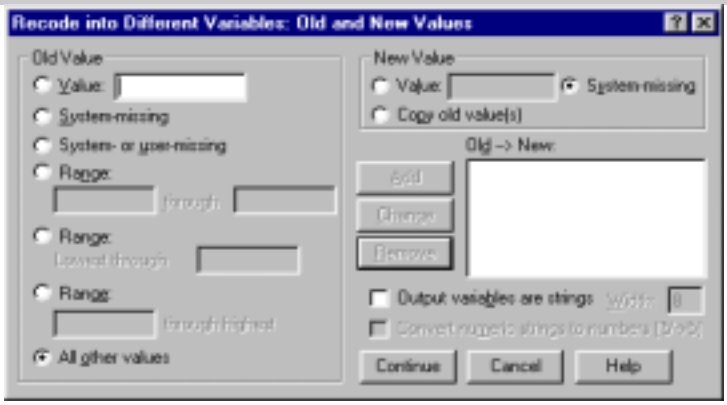
Click on the mapping entry "Missing→Sysmis" and click on "Remove."



Repeat the previous step (of pressing "Remove") for all old mapping entries.

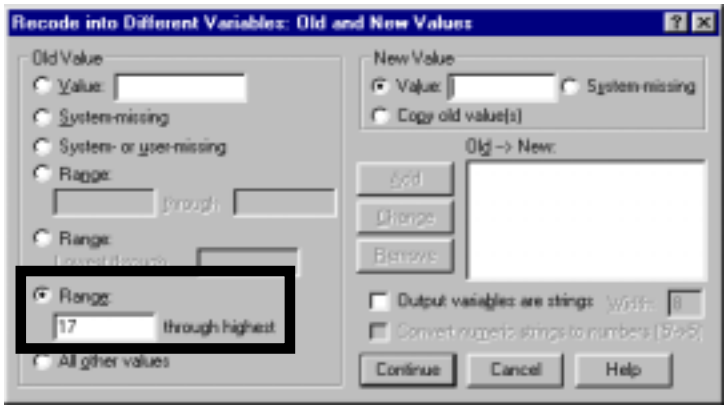


Now you are ready to type in the recodes.



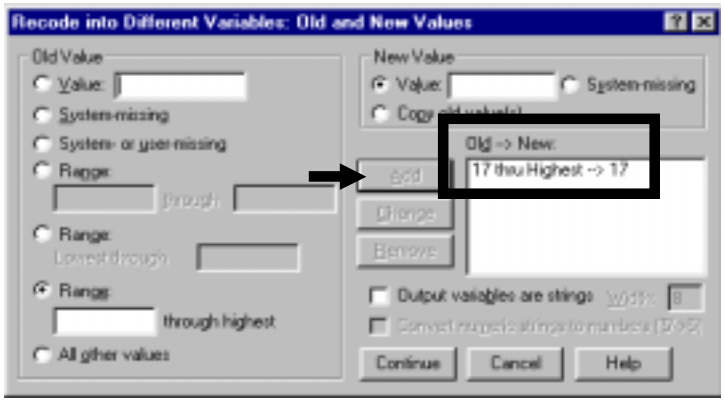
The first recoding item we wish to map is "17 and greater → 17."

Select “Range...thru Highest” and enter the number 17 so that the box reads “17 thru highest.”

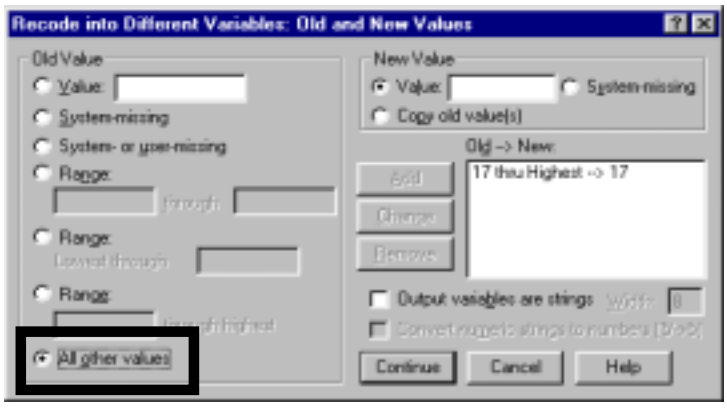


On the right side of the box, in the area “New Value,” choose “Value” and enter the number 17.

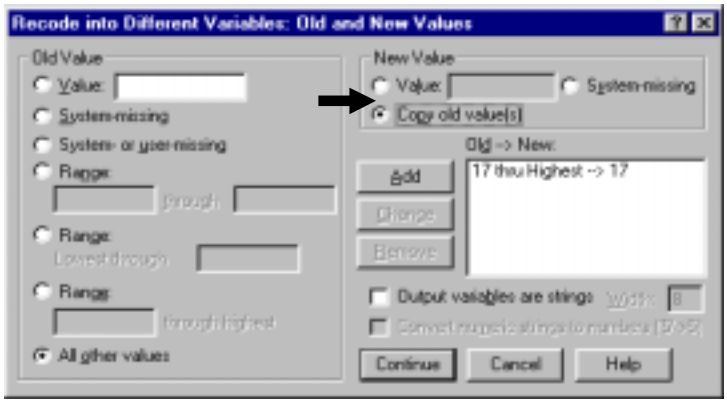
Click on “Add.” The mapping of “17,...,highest into 17” will be seen in the box “Old→New.”



In the area “Old Values,” choose “All other values.”



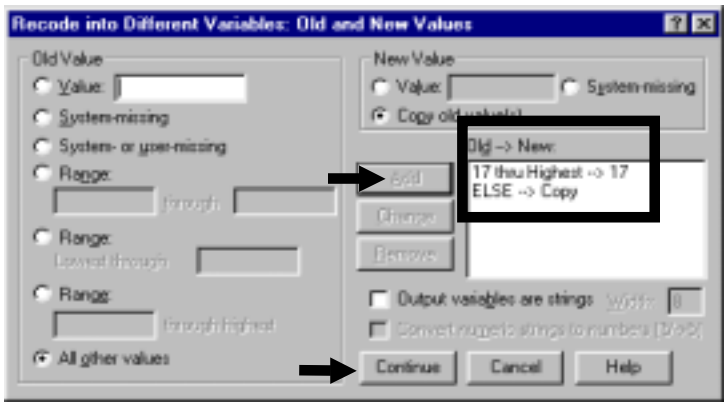
In the area "New Value," choose "Copy old value(s)."



Click on "Add."

The mapping is now complete.

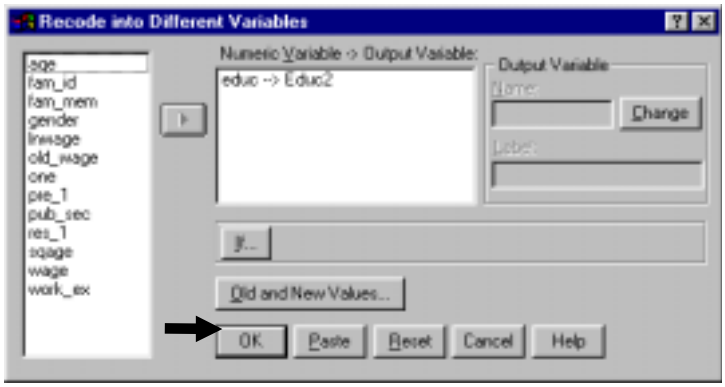
Click on "Continue."



Click on "OK."

A new variable, *educ2*, will be created.

Note: Go to DEFINE / VARIABLE and define the attributes of the new variable. See section 1.2 for examples of this process. In particular, you should create variable labels, value labels, and define the missing values.



Ch 2. Section 1.c. Replacing existing variables using recode

Sometimes you may prefer to change an existing variable with a categorical or dummy recoding of itself. This is the case when the coding of the old variable is misleading or inappropriate for the planned analysis[31]. Whenever you wish to replace an existing variable, you must be certain that the original version of the variable will not be needed for future analysis. (If you have any hesitation, then use the procedure described in sections 2.1.a and 2.1.b).

Let's assume you want to look at cases according to different age groups to test the hypothesis that workers in their forties are more likely to earn higher wages. To do this, you must recode

[31] A common use is recoding dummies from the codes 1 and 2 into the codes 0 and 1.

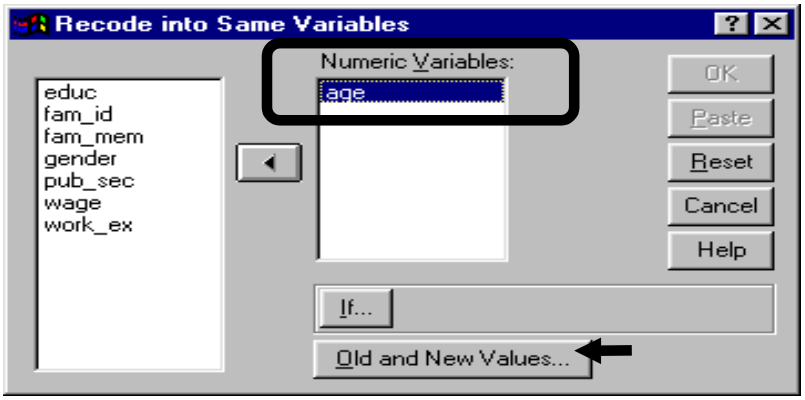
age into a variable with 5 categories: workers whose age is between 20-29 years, 30-39 years, 40-49 years, 50-59 years, and all other workers (i.e. - those who are 20 years old or younger and those who are 60 years old and over).

Values in Original Variable age	***Values in New Variable age***
20-29	1
30-39	2
40-49	0
50-59	3
0-20 and 60 through highest	4

Go to TRANSFORM/ RECODE/ INTO SAME VARIABLES[32].

Select *age* from the list of choices and click on the arrow to send it over to the box labeled "Numeric Variables."

Click on the button "Old and New Values."



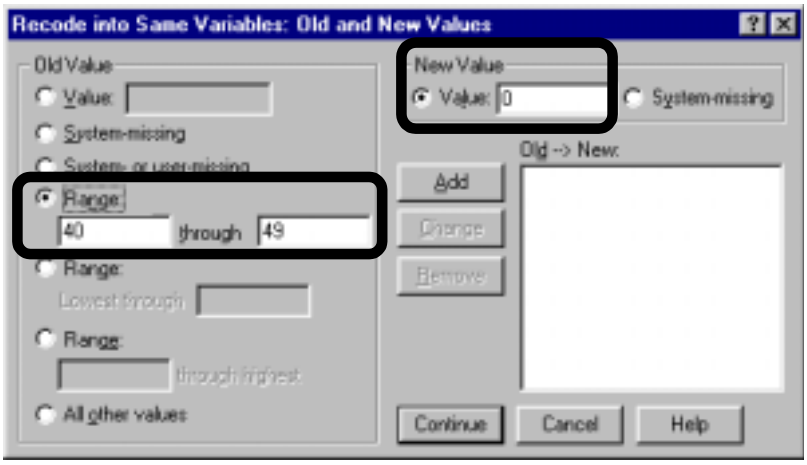
Select the option "Range," in which you can specify a minimum and maximum value.

[32] Note that this is a different sub-sub-menu compared to that used in the previous section (that menu option was TRANSFORM / RECODE / INTO NEW VARIABLES.

You must code workers with *age* 40-49 as your reference group. (i.e. - recode that range as zero.)

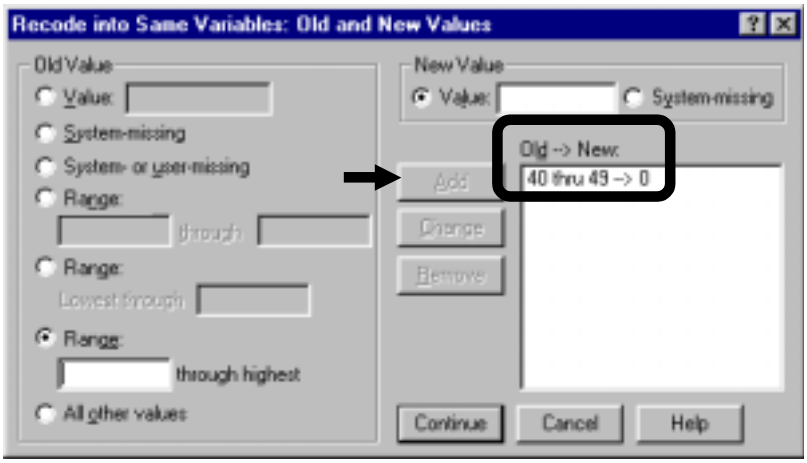
Under the "Range" option, type in the range of your reference group (40 through 49).

On the right side menu, select "Value" and type 0 into the box to the right of "Value."



Click on "Add" to move the condition to the "Old → New" box.

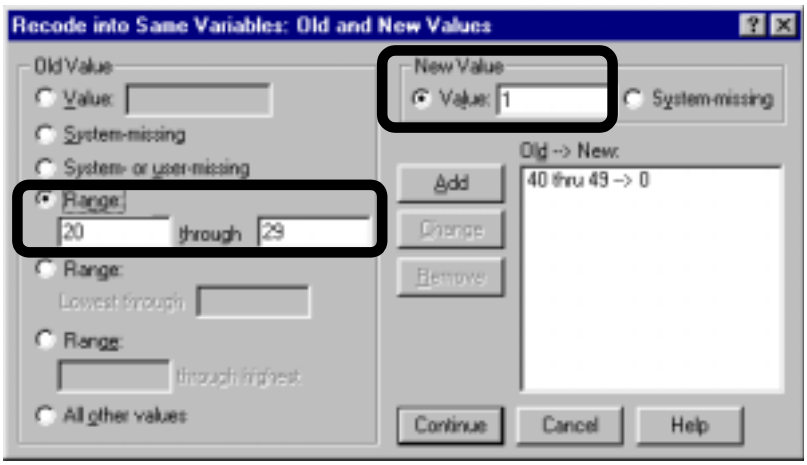
Now you should see the first mapping item: "40 thru 49→ 0."



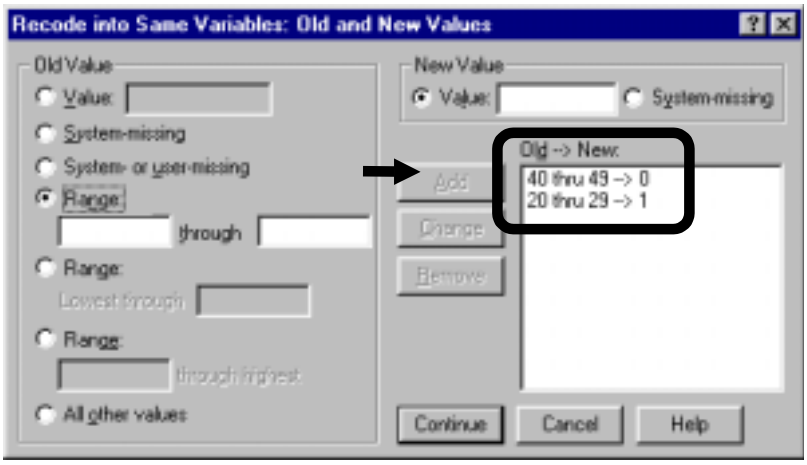
Continue specifying the other conditions. Specify all other age groups in the Range menu.

For example, select 20-29 as your range. This time, type in 1 as the new value.

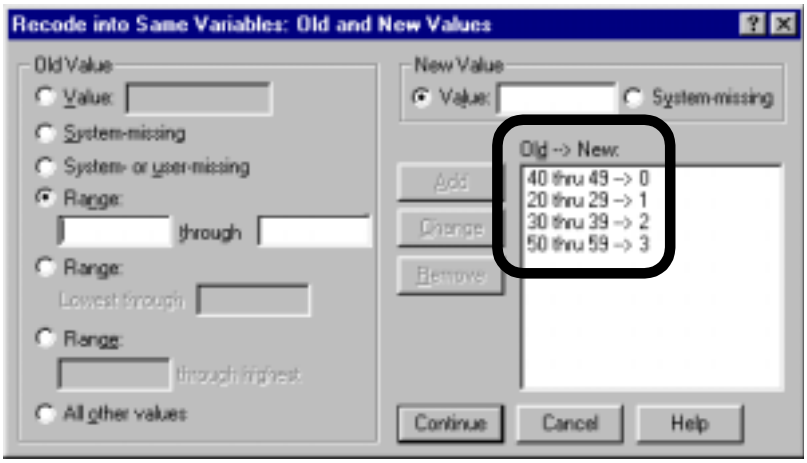
Reminder: Experiment with the different ways provided to define the "Old Value." Practice makes perfect!



Then click on "Add" to add it to the list.

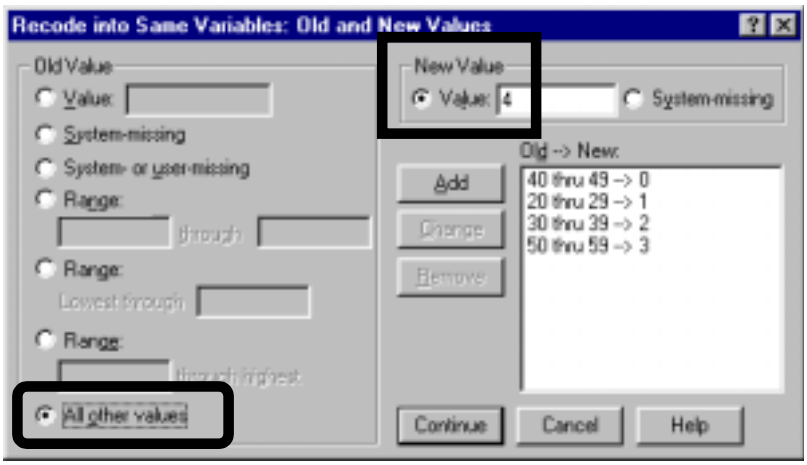


Continue the list of conditions: 20-29 = 1, 30-39 = 2, 50-59 = 3.



You also want to group the remaining values of *age* (below 20 and above 59) into another category.

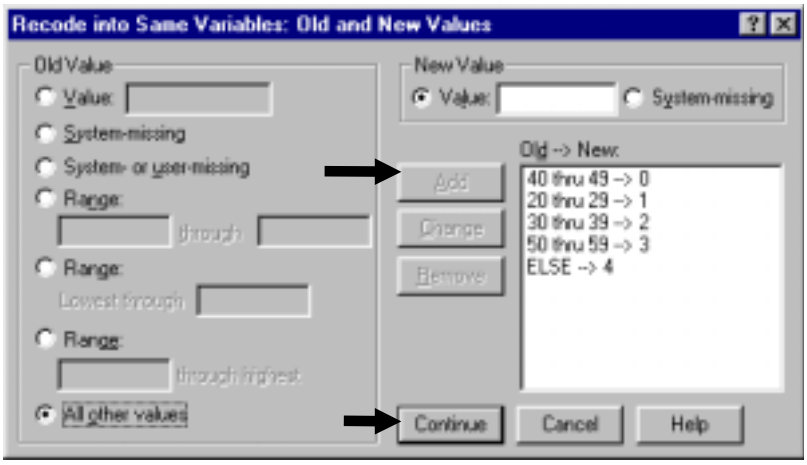
Select "All other values" at the bottom of the Old Value menu. Select 4 as the new value[33].



[33] You can choose a different way to achieve the same result. For example, you may prefer using the following three mapping items to replace the one item we use ("ELSE→4"): "lowest thru 19 → 4," "60 thru highest → 4," and "ELSE → System-Missing." The more finely you define the items, especially the values to be considered as missing, the lower the chances of creating a variable with incorrect values.

Click on “Add” to move it to the list of conditions.

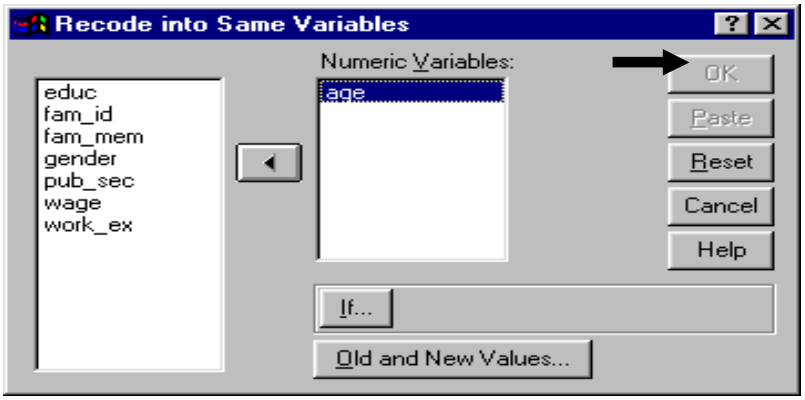
Click on “Continue.”



Click on “OK.”

The variable *age* will be replaced with the newly recoded variable *age*.

Note: Go to DEFINE / VARIABLE and define the attributes of the "new" variable. See section 1.2 for examples of this process. In particular, you should create value labels, e.g. - "1 → Young Adults," "2→ Adults," etc.



Ch 2. Section 1.d. Obtaining a dummy variable as a by-product of filtering

Recall that when you created Sub-sets of the data using DATA / SELECT CASE (see section 1.7), SPSS created a filter variable. Let's assume that the filter you created was “Filter in only those observations in which the respondent is an adult female” (i.e. - where *gender* =1 and *age* >20). The filter variable for that filter will contain two values mapped as:

Value	Category
0	Females of age 20 or under and all Males
1	Females of age 20 and above

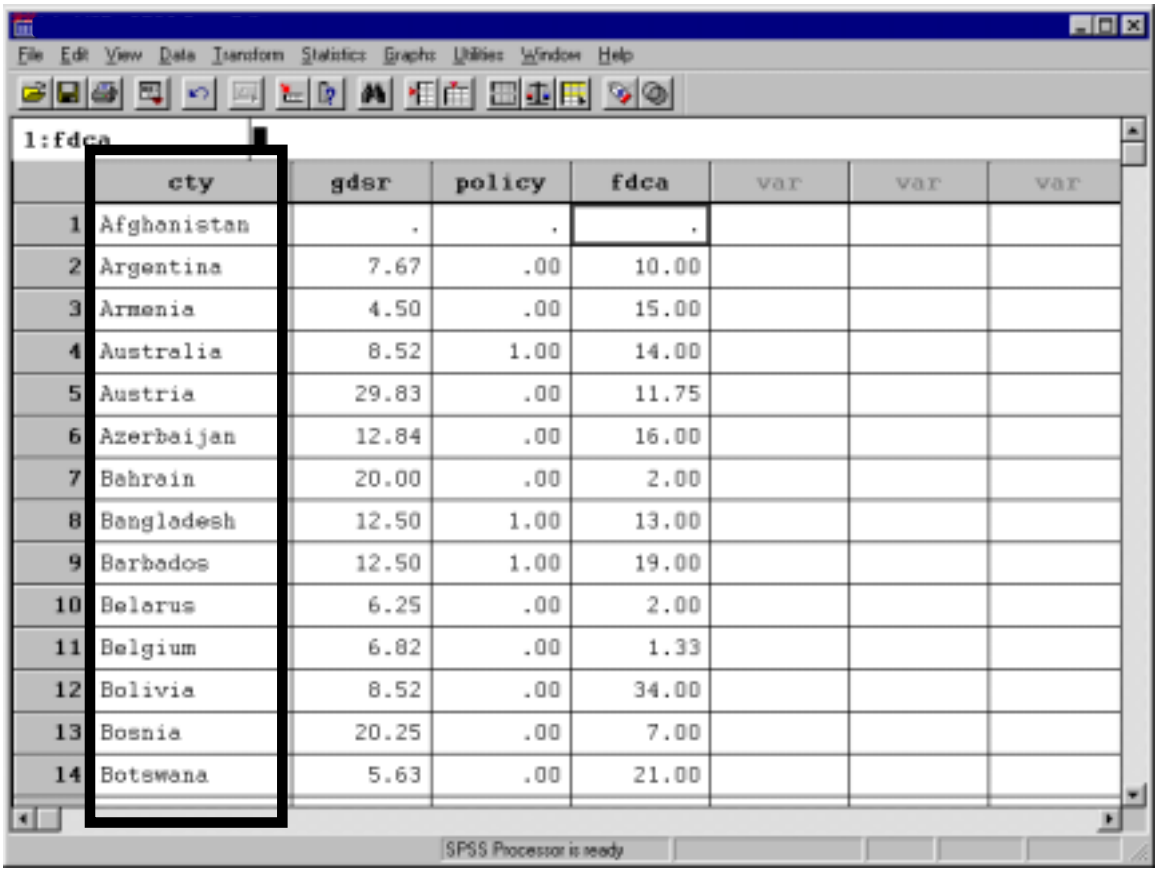
This dummy variable can be used as any other dummy variable. To use it, you must first turn the above filter off by going to DATA/ SELECT CASE and choosing “All cases” as shown in section 1.7.c.

Ch 2. Section 1.e. Changing a text variable into a numeric variable

You may want to create dummy or categorical variables using as criteria the values of a variable with text data, such as names of states, countries, etc. You must convert the variable with the text format into a numeric variable with numeric codes for the countries.

Tip: This procedure is not often used. If you think this topic is irrelevant for you, you may simply skip to the next section.

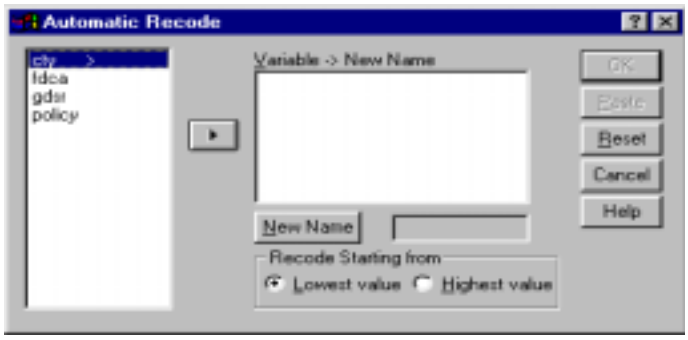
Let's assume that you have the names of countries as a variable *cty*. (See picture below.)



	cty	gdsr	policy	fdca	var	var	var
1	Afghanistan	.	.	.			
2	Argentina	7.67	.00	10.00			
3	Armenia	4.50	.00	15.00			
4	Australia	8.52	1.00	14.00			
5	Austria	29.83	.00	11.75			
6	Azerbaijan	12.84	.00	16.00			
7	Bahrain	20.00	.00	2.00			
8	Bangladesh	12.50	1.00	13.00			
9	Barbados	12.50	1.00	19.00			
10	Belarus	6.25	.00	2.00			
11	Belgium	6.82	.00	1.33			
12	Bolivia	8.52	.00	34.00			
13	Bosnia	20.25	.00	7.00			
14	Botswana	5.63	.00	21.00			

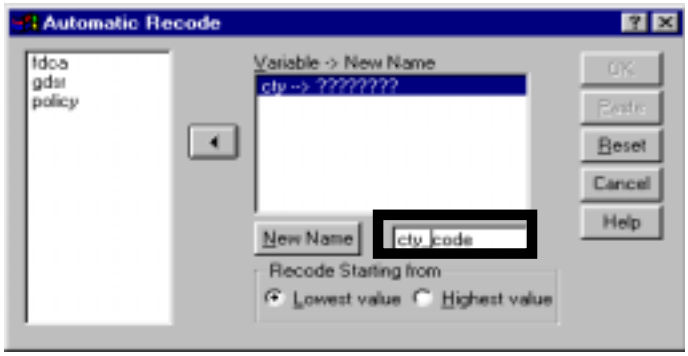
You want to create a new variable "c*ty_code"* where the countries listed in the variable "cty" are recoded numerically as 1,2,..... into a new variable, "*cty_code*." The recoding must be done in alphabetical order, with "Afghanistan" being recoded into 1, "Argentina" into 2, etc.

To do so, go to TRANSFORM/ AUTORECODE.



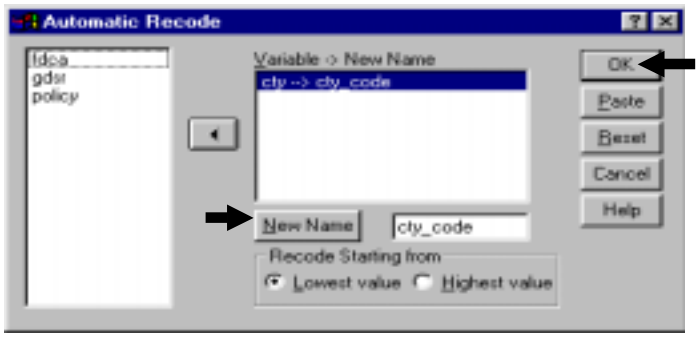
Select the text variable you wish to recode - move the variable *cty* into the white box "Variable→New Name."

Enter the new name *cty_code* for the variable into the small box on the right of the button "New Name."

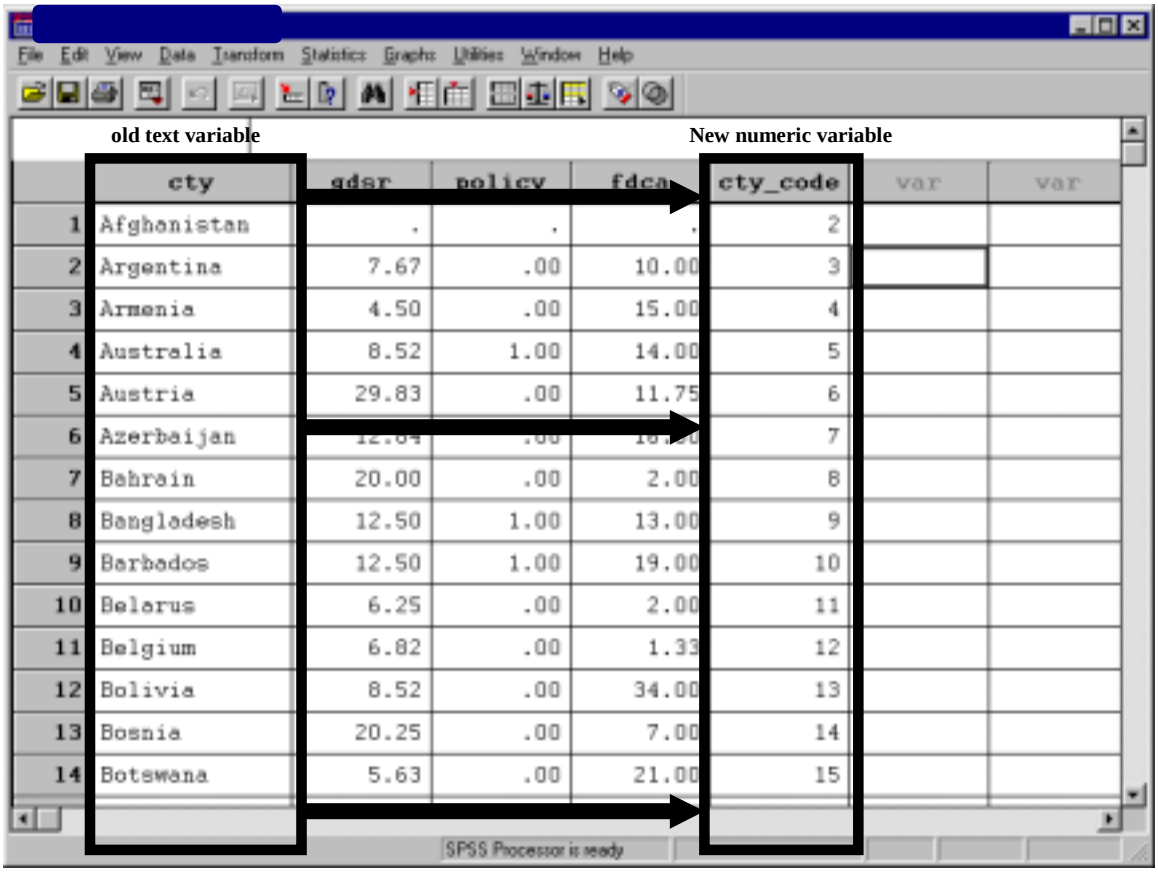


Click on the "New Name" Button.

Click on "OK."



The new variable has been created.



Now you can use the variable *cty_code* in other data manipulation, graphical procedures, and statistical procedures.

Ch 2. Section 2 Using mathematical computations to create new continuous variables: compute

New continuous variables must be computed for most analysis. The reasons may be:

1. Creation of a variable that is intuitively closer to the needs of your analysis.
2. Interactive dummies are used to enhance the information provided by a regression. Multiplying a dummy and a continuous variable creates interactive dummies.
3. Correct specification of regression (and other) models may require the creation of transformed variables of the original variables. Log transformations are a common tool used. See section 8.3 for an example.
4. Several tests and procedures (e.g. - the White's test for heteroskedasticity shown in section 7.5) require the use of specific forms of variables - squares, square roots, etc.

Don't worry if these terms/procedures are alien to you. You will learn about them in later chapters and/or in your class.)

Ch 2. Section 2.a. A simple computation

We show an example of computing the square of a variable. In mathematical terminology we are calculating the square of *age*:

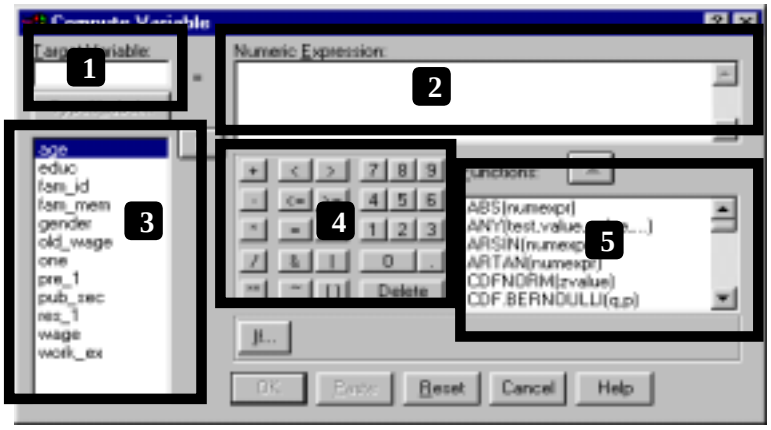
Sqage = $(age)^2$, or, equivalently, *Sqage* = (*age*)*(*age*)

Go to TRANSFORM/ COMPUTE.

In area 1, enter the name of the new variable.

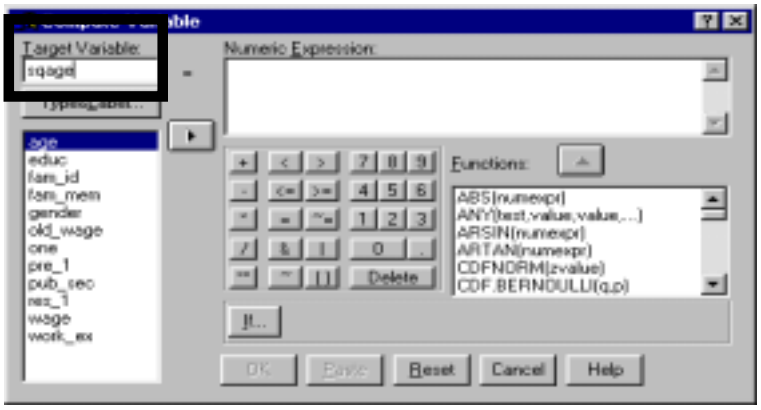
Area 2 is where the mathematical expression for the computing procedure is entered.

From area 3, you choose the existing variables that you want in the mathematical expression in area 2.



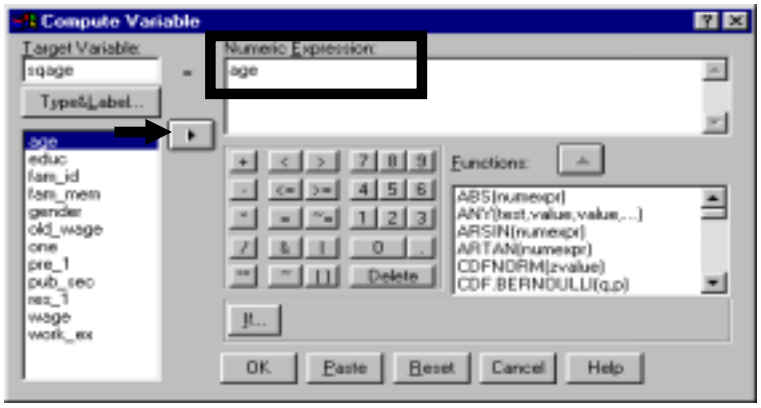
Area 4 has a keypad with numbers and operators. Area 5 has the "built-in" mathematical, statistical and other functions of SPSS.

In the box below the label "Target Variable," type in the name of the new variable you are creating (in this example, *sqage*).



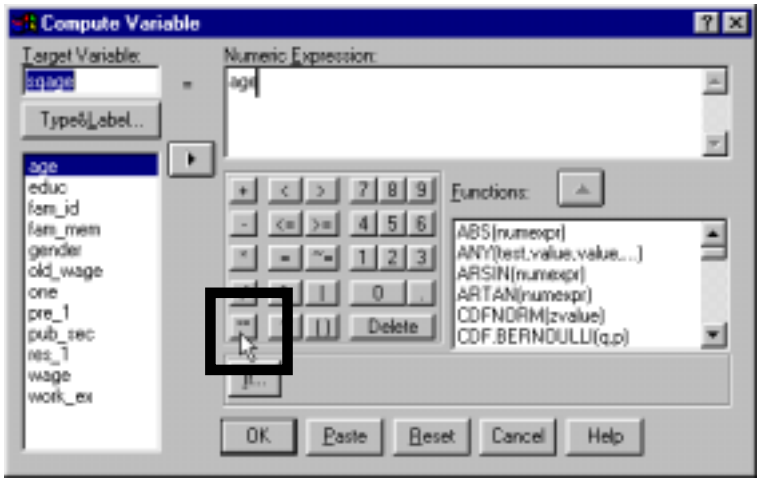
Now you must enter the expression/formula for the new variable.

First, click on the variable *age* and move it into the box below the label "Numeric Expression."

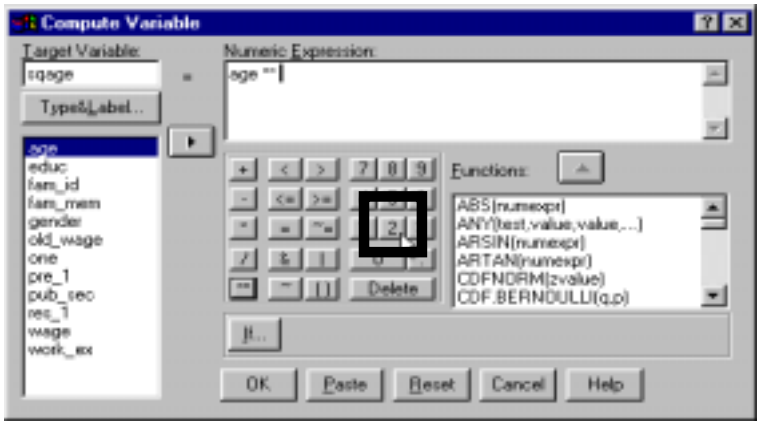


To square the variable *age*, you need the notation for the power function. Click on the button "** " (or type in " ^ ").

You may either type in the required number or operator or click on it in the keypad in the dialog box.



To square the variable *age*, it must be raised to the power of "2." Go to the button for two and click on it (or enter 2 from the keyboard).

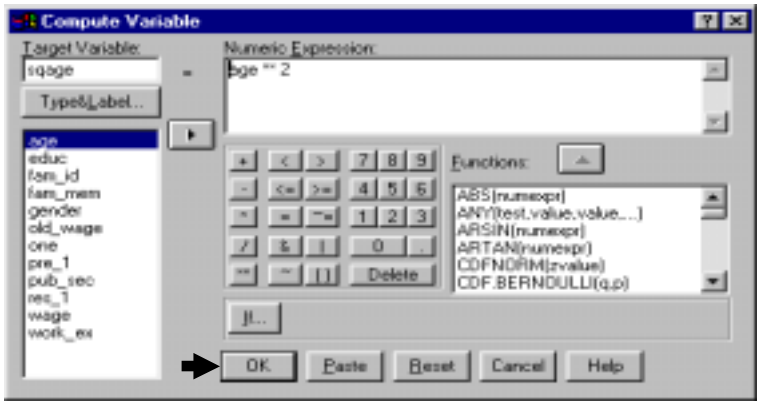


The expression is now complete.

Click on "OK."

A new variable has been created. Scroll to the right of the data window. The new variable will be the last variable.

Note: Go to DEFINE / VARIABLE and define the attributes of the new variable. See section 1.2 for examples of this process. In particular, you should create variable labels and define the missing values.



In the next table we provide a summary of basic mathematical operators and the corresponding keyboard digits.

Mathematical Operators

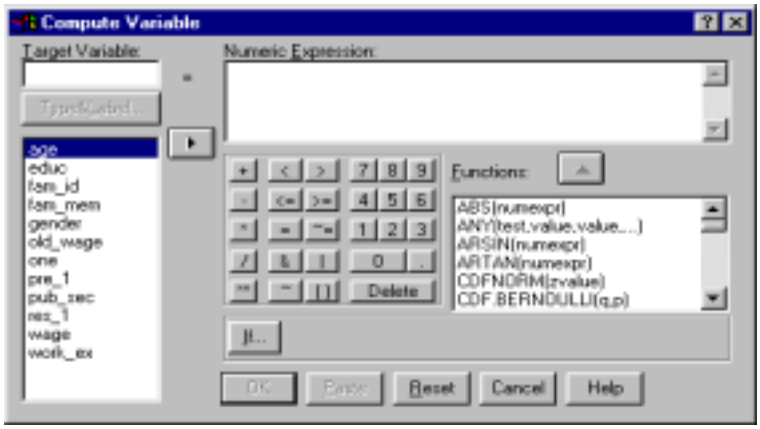
Operation	Symbol
Addition	+
Subtraction	-
Multiplication	*
Division	/
Power	** or ^

Ch 2. Section 2.b. Using built-in SPSS functions to create a variable

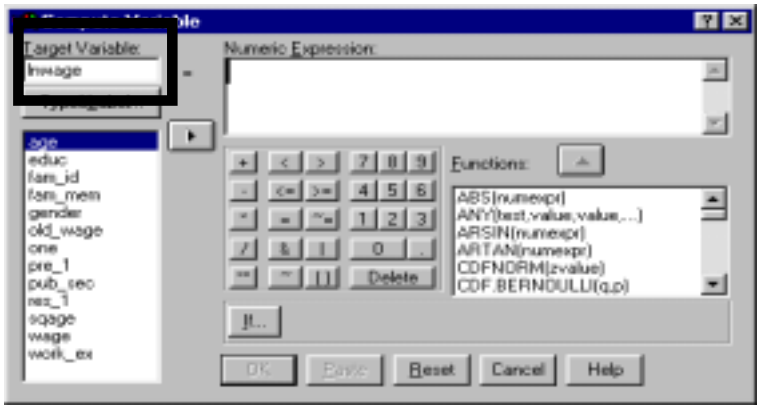
SPSS has several built-in functions. These include mathematical (e.g. - "Log Natural"), statistical, and logical functions. You will need to learn these functions only on an "as-needed" basis. In the examples that follow, we use the most useful functions.

Go to TRANSFORM/ COMPUTE.

Note: This dialog box is very complex. Please try a few examples on any sample data set.

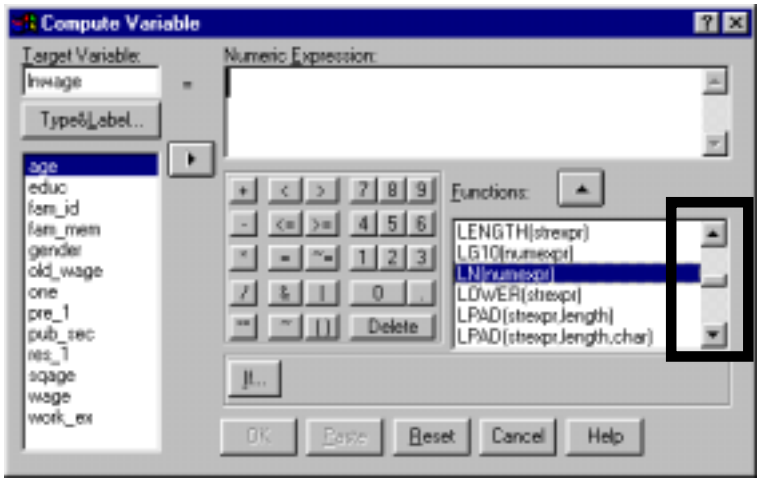


In the box below the label "Target Variable," type the name of the new variable you are creating (*lnwage*).



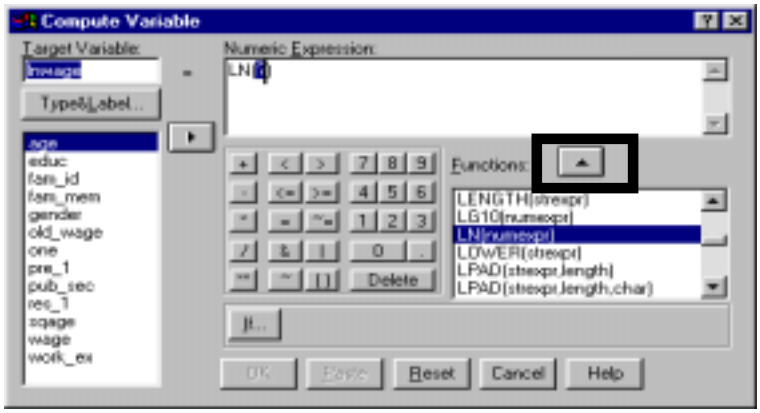
Now you must enter the formula for the new variable.

To do so, you first must find the log function. Go to the scroll bar next to the listed box "Functions" and scroll up or down until you find the function "LN (numexp)."



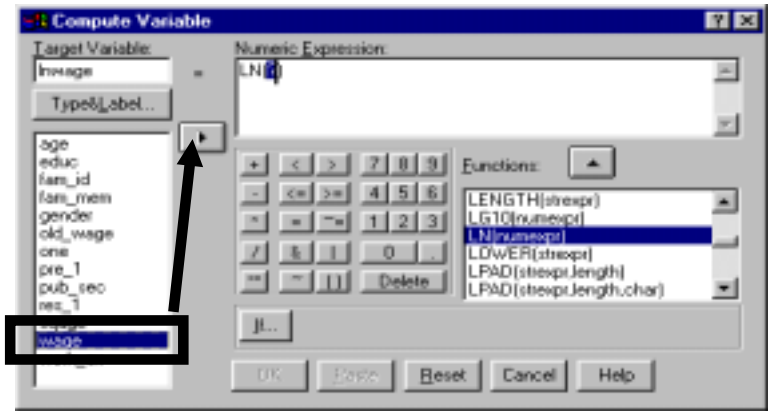
Click on the upward arrow. This moves the function LN into the expression box "Numeric Expression."

How does one figure out the correct function? Click on the help button for an explanation of each function or use the help section's find facility.



The question mark inside the formula is prompting you to enter the name of a variable.

Click on the variable *wage* and move it into the parenthesis after LN in the expression.



The expression is now complete.

Click on "OK."

A new variable, *lnwage*, is created. Scroll to the right of the data window. The new variable will be the last variable.

Note: Go to DEFINE / VARIABLE and define the attributes of the new variable. See section 1.2 for examples of this process. In particular, you should create variable labels and define the missing values.

The next table shows examples of the types of mathematical/statistical functions provided by SPSS.

Important/Representative Functions

Function	Explanation
LN(X)	Natural log of X
EXP(X)	Exponent of X
LG10(X)	Log of X to the base 10
MAX(X,Y,Z)	Maximum of variables X, Y and Z
MIN(X,Y,Z)	Minimum of variables X, Y and Z
SUM(X,Y,Z)	Sum of X, Y and Z (missing values assumed to be zero)
LAG(X)	1 time period lag of X
ABS(X)	Absolute value of X
CDF.BERNOULLI(X)	The cumulative density function of X, assuming X follows a Bernoulli distribution
PDF.BERNOULLI(X)	The probability density function of X, assuming X follows a Bernoulli distribution

Examples of other computed variables:

(1) Using **multiple variables**: the difference between *age* and *work experience*.

agework = age - work_ex

(2) Creating **interactive dummies**: you will often want to create an interactive term[34] in which a dummy variable is multiplied by a continuous variable. This enables the running of regressions in which differential slopes can be obtained for the categories of the

[34] Refer to your textbook for details on the usefulness of an interactive term.

dummy. For example, an interactive term of *gender* and *education* can be used in a *wage* regression. The coefficient on this term will indicate the difference between the rates of return to *education* for females compared to males.

*gen_educ = gender * educ*

(3) Using **multiple functions**: you may want to find the square root of the log of the interaction between *gender* and *education*. This can be done in one step. The following equation is combining three mathematical functions - multiplication of *gender* and *education*, calculating their natural log and, finally, obtaining the square root of the first two steps.

srlgened = SQRT (LN (gender * educ))

(4) Using **multi-variable mathematical functions**: you may want to find the maximum of three variables (the *wages* in three months) in an observation. The function MAX requires multi-variable input. (In the example below, *wage1*, *wage2*, and *wage3* are three separate variables.)

mage = MAX (*wage1*, *wage2*, *wage3*)

Ch 2. Section 3 Multiple response sets - using a "set" variable made up of several categorical variables

Nothing better illustrates the poor menu organization and impenetrable help menu of SPSS than the "Multiple Response Sets" options. They are placed, incorrectly, under STATISTICS. It would be preferable for them to be placed in DATA or TRANSFORM!

But despite its inadequacies, SPSS remains a useful tool...

In section 2.1, you learned how to use RECODE to create dummy and categorical variables. The RECODE procedure usually narrows down the values from a variable with (let's assume "M") more possible values into a new variable with fewer possible values, e.g. - the *education* to *basic education* recode mapped from the range 0-23 into the range 0-1.

What if you would like to do the opposite and take a few dummy variables and create one categorical variable from them? To some extent, Multiple Response Sets help you do that. If you have five dummy variables on *race* ("African-American or not," "Asian-American or not," etc.) but want to run frequency tabulations on race as a whole, then doing the frequencies on the five dummy variables will not be so informative. It would be better if you could capture all the categories (5 plus 1, the "or not" reference category) in one table. To do that, you must define the five dummy variables as one "Multiple Response Set."

Let us take a slightly more complex example. Continuing the data set example we follow in most of this book, assume that the respondents were asked seven more "yes/no" questions of the form -

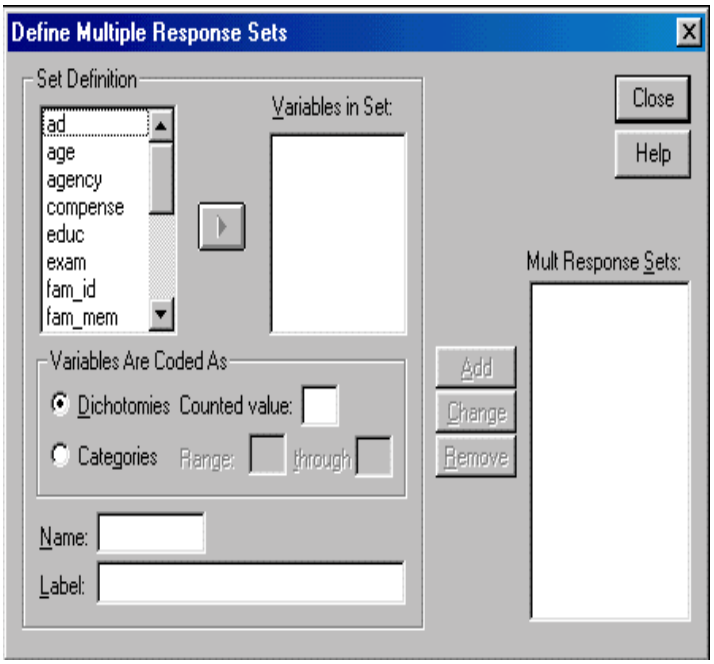
1. *Ad*: "Did the following resource help in obtaining current job - response to newspaper ad"
2. *Agency*: "Did the following resource help in obtaining current job - employment agency"
3. *Compense*: "Did the following resource help in obtaining current job - veteran or other compensation and benefits agency"

4. *Exam*: "Did the following resource help in obtaining current job - job entry examination"
5. *Family*: "Did the following resource help in obtaining current job - family members"
6. *Fed_gov*: "Did the following resource help in obtaining current job - federal government job search facility"
7. *Loc_gov*: "Did the following resource help in obtaining current job - local government job search facility"

All the variables are linked. Basically they are the "Multiple Responses" to the question "What resource helped in obtaining your current job?"

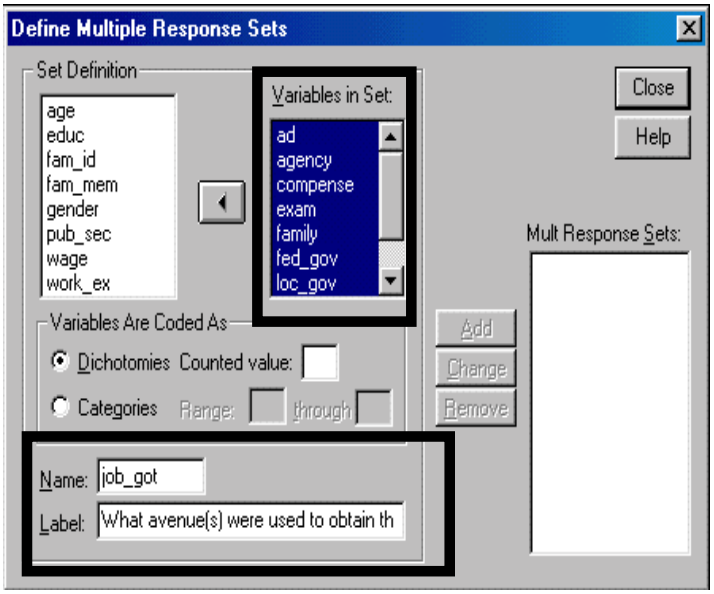
Let's assume you want to obtain a frequency table and conduct cross tabulations on this set of variables. Note that a respondent could have answered "yes" to more than one of the questions.

Go to STATISTICS / MULTIPLE RESPONSE/ DEFINE SETS.



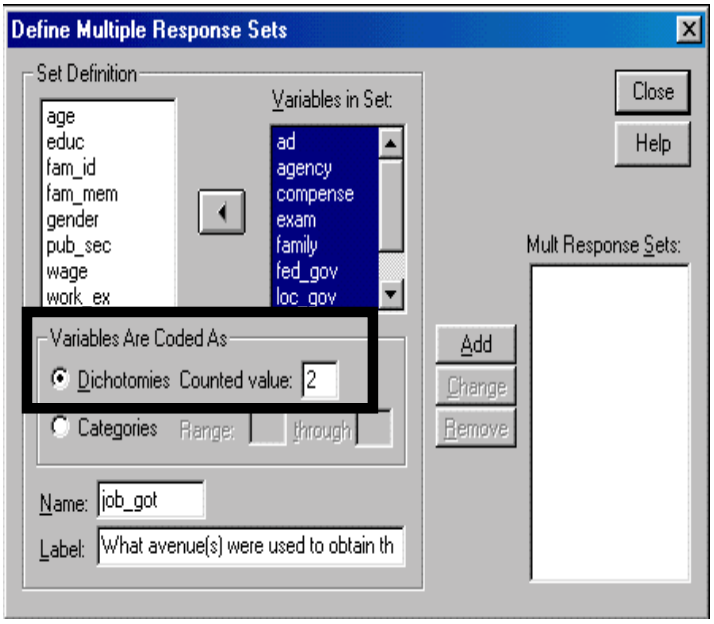
Enter a name and label for the set. (Note: no new variable will be created on the data sheet.)

Move the variables you want in the set into the box "Variables in Set."



Each of our seven variables is a "yes/no" variable. Thus, each is a "dichotomous" variable. So choose the option "Dichotomies" in the area "Variables are Coded As."

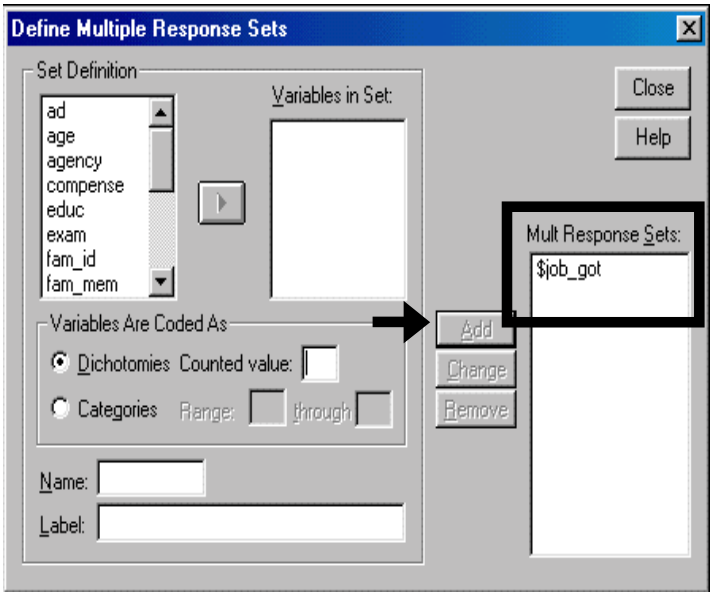
SPSS starts counting from the lowest value. So, by writing "2" in the box "Counted value," we are instructing SPSS to use the first two values as the categories of each variable constituting the set.



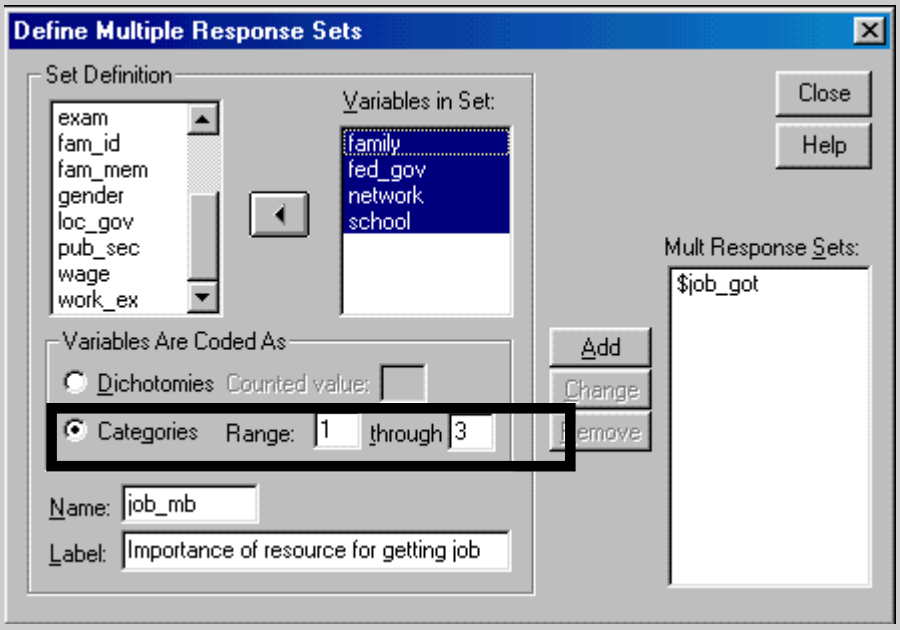
Click on "Add."

The new set is created and shown in the box "Multiple Response Sets."

Note: This feature can become very important if the data come from a survey with many of these "broken down" variables.

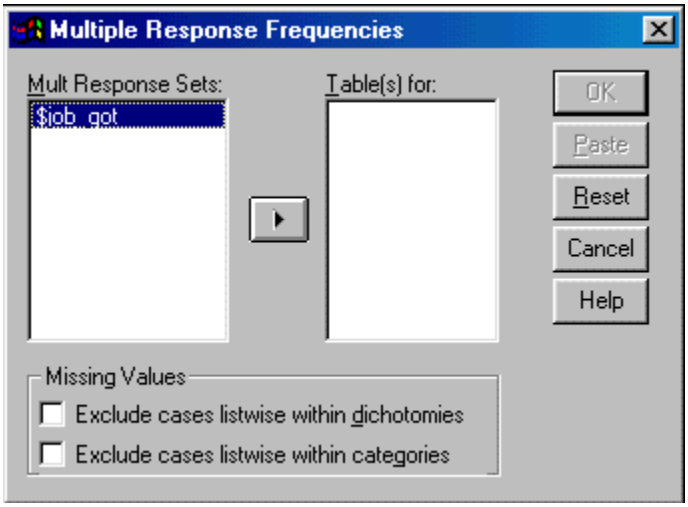


Note: you can also use category variables with more than two possible values in a multiple response set. Use the same steps as above with one exception: choose the option "Categories" in the area "Variables are Coded As" and enter the range of values of the categories.

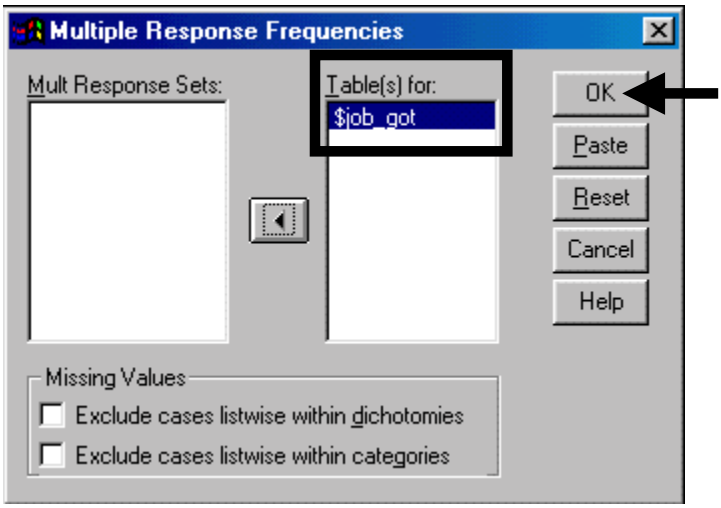


Now you can use the set. The set can only be used in two procedures: frequencies and cross tabulations.

To do frequencies, go to STATISTICS / MULTIPLE RESPONSE / MULTIPLE RESPONSE FREQUENCIES.

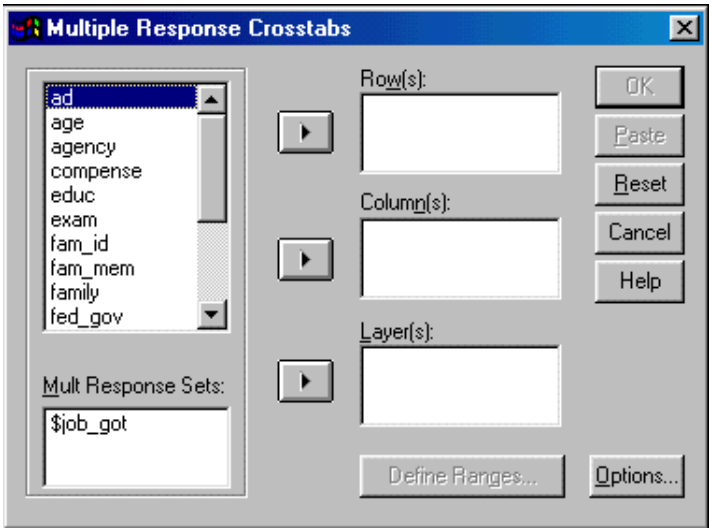


Choose the set for which you want frequencies and click on "OK."

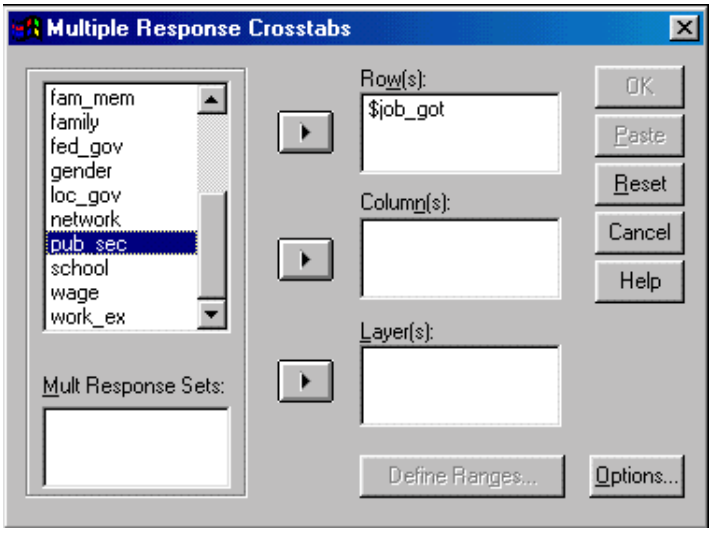


See section 3.2 for more on frequencies.

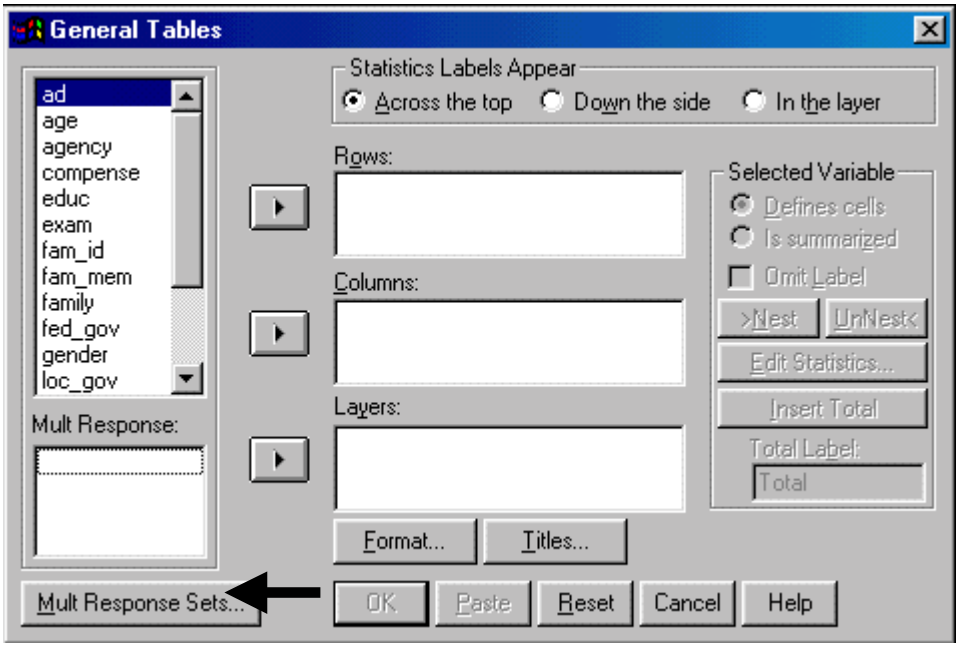
Similarly, to use the set in crosstabs, go to STATISTICS / MULTIPLE RESPONSE / MULTIPLE RESPONSE CROSSTABS.



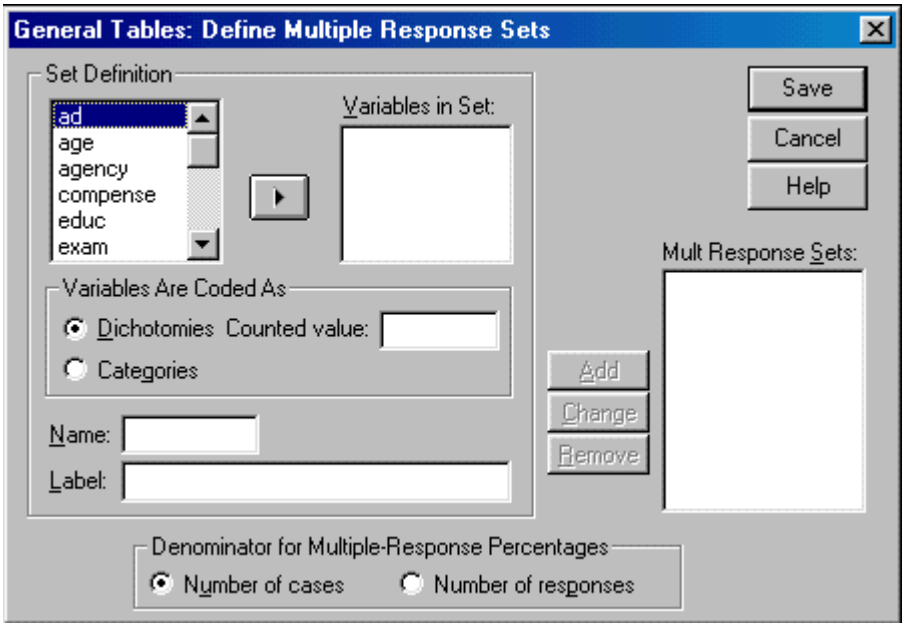
See the set as the criterion variable for a row, column, or layer variable.



To use multiple response sets in tables, go to STATISTICS / GENERAL TABLES. Click on "Mult Response Sets."



In the next dialog box, define the sets as you did above.



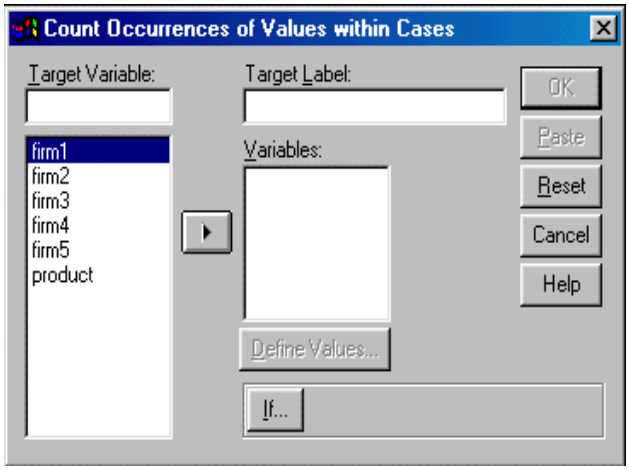
Ch 2. Section 4 Creating a "count" variable to add the number of occurrences of similar values across a group of variables

We will use a different data set to illustrate the creation of a “one-from-many-ratings-variables” variable using the COUNT procedure.

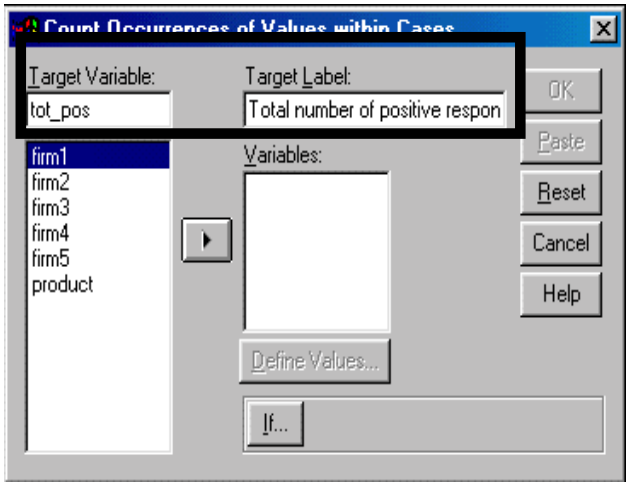
Let's assume a wholesaler has conducted a simple survey to determine the ratings given by five retailers (“firm 1,” “firm 2,” … , “firm 5”) to product quality on products supplied by this wholesaler to these retailers. The retailers were asked to rate the products on a scale from 0-10, with a higher rating implying a higher quality rating. The data was entered by product, with one variable for each retailer.

The wholesaler wants to determine the distribution of products that got a “positive” rating, defined by the wholesaler to be ratings in the range 7-10. To do this, a new variable must be created. This variable should “count” the number of firms that gave a “positive” rating (that is, a rating in the range 7-10) for a product.

To create this variable, go to TRANSFORM / COUNT.

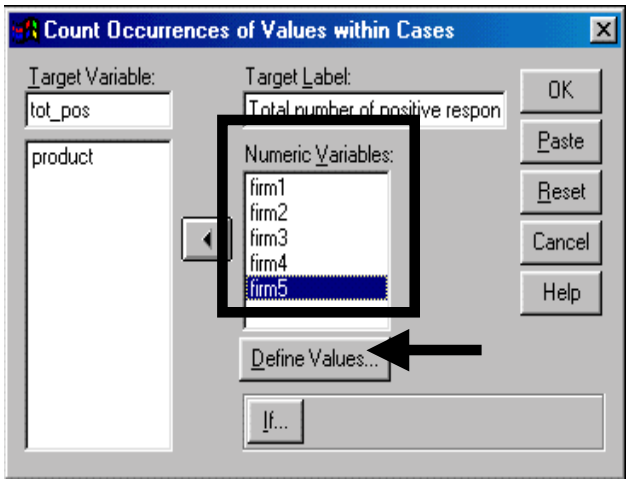


Enter the name and variable label for the new variable.

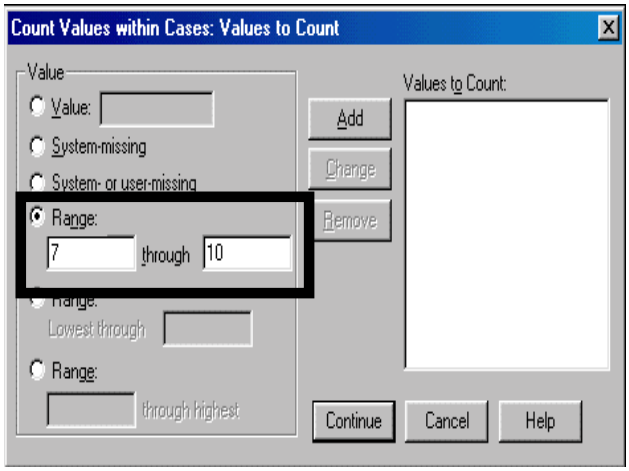


Move the variables whose values are going to be used as the criterion into the area “Numeric Variables”

Now the mapping must be defined, i.e. - we must define "what must be counted." To do this, click on “Define Values.”

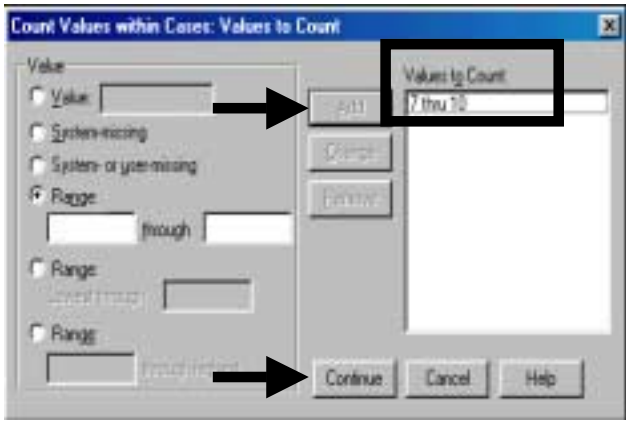


Enter the range you wish to define as the criterion. (See section 2.1 for more on how to define such range criterion.)

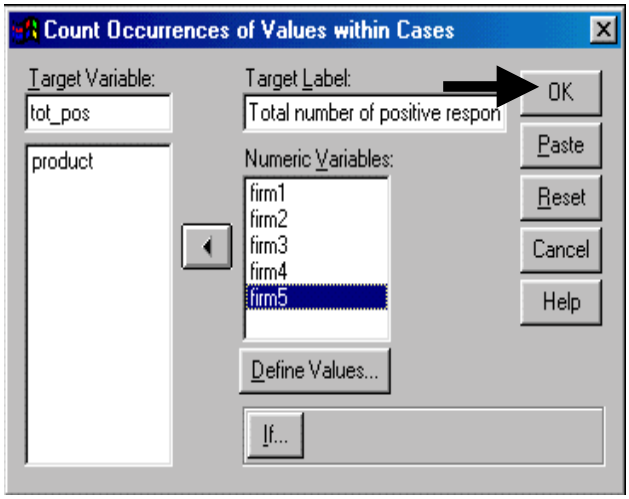


Click on "Add." The area "Values to Count" now contains the criterion.

If you wish to define more criteria, repeat the above two steps. Then click on "Continue."



Click on "OK."



Ch 2. Section 5 Continuous variable groupings created using cluster analysis

Using cluster analysis, a continuous variable can be grouped into qualitative categories based on the distribution of the values in that variable. For example, the variable *wage* can be used to

create a categorical variable with three values by making three groups of wage earnings - high income, mid income, and low income - with SPSS making the three groups.

The mapping is:

Value	Category
1	High income
2	Low income
3	Mid income

A very simplistic example of clustering is shown here.

Let's assume you want to use "income-group membership" as the variable for defining the groups in a comparative analysis. But let's also assume that your data have a continuous variable for income, but no categorical variable for "income-group membership." You therefore must use a method that can create the latter from the former. If you do not have pre-defined cut-off values for demarcating the three levels, then you will have to obtain them using methods like frequencies (e.g. - using the 33rd and 66th percentile to classify income into three groups), expert opinion, or by using the classification procedure. We show an example of the classification procedure in this section.

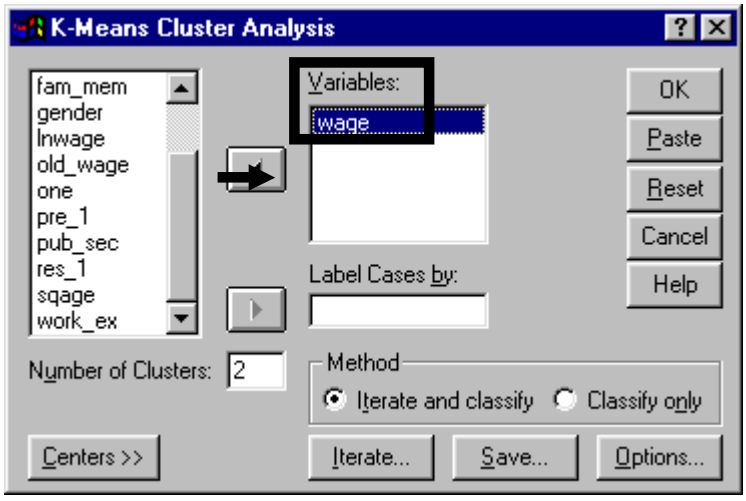
Note: The Classification procedure has many uses. We are using it in a form that is probably too simplistic to adequately represent an actual analysis, but is acceptable for the purposes of illustrating this point.

We show you how to make SPSS create groups from a continuous variable and then use those groups for comparative analysis.

Go to STATISTICS/ CLASSIFY/ K-MEANS CLUSTER.

Note: "K-Means Cluster" simply means that we want to create clusters around "K-number" of centers.

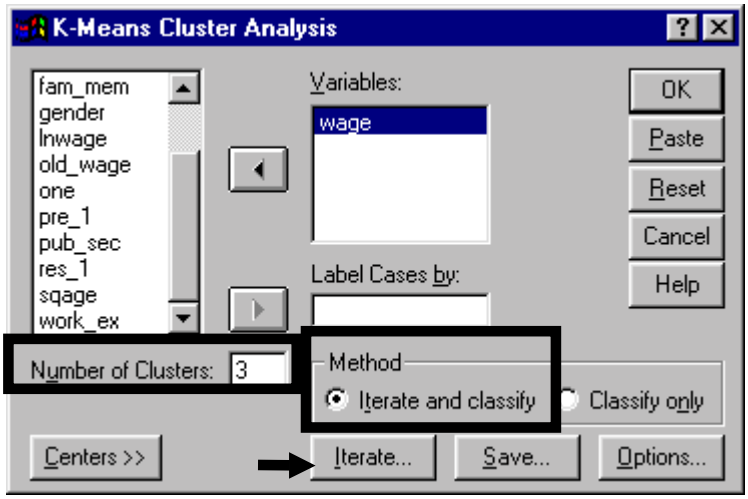
Select the variables on whose basis you wish to create groups. Move the variables into the box "Variables."



We want to divide the data into 3 income groups: low, mid, and high. Enter this number into the box "Number of Clusters."

Choose the method "Iterate and classify."

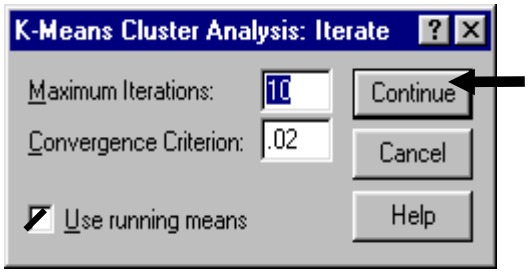
Click on the button "Iterate."



We recommend going with the defaults, though you may wish to decrease the convergence to produce a more fine-tuned classification.

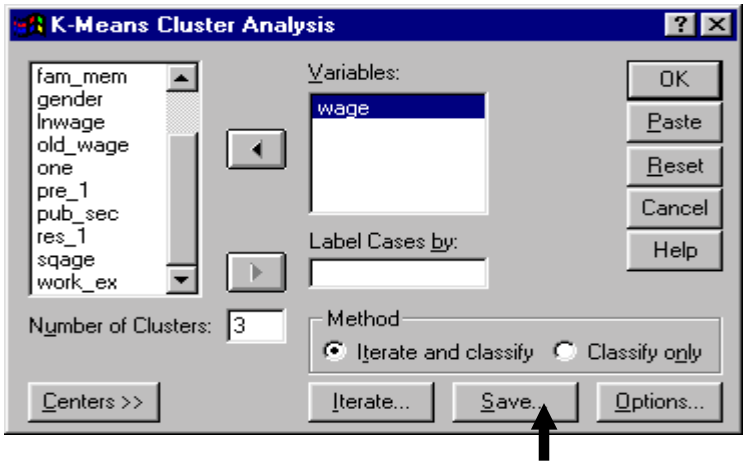
Choose the option "Use running means." This implies that each iteration (that is, each run of the cluster "algorithm") will use, as starting points, the 3 cluster "means" calculated in the previous iteration/run.

Click on "Continue."



Click on "Save."

Note: This step is crucial because we want to save the "index" variable that has been created and use it for our comparative analysis.



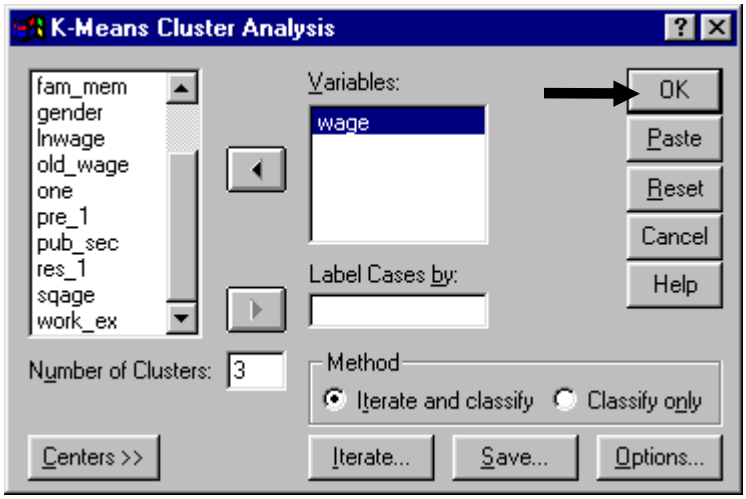
Choose to save "Cluster membership." This will create a new variable that will define three income groups.

Click on "Continue."

Click on "OK." A new variable with cluster membership will be created.

The variable will take three values: 1, 2, and 3, with each value signifying the income level of that observation.

The values in the index variable may not correspond in a monotonic fashion to the income categories low, mid, and high. For example, 1 may be low, 2 may be high, and 3 may be mid-income. See output below page - the above will become clear.



Results of Cluster Analysis

```
Convergence achieved due to no or small distance change.

 Final Cluster Centers.

Cluster       WAGE

        1       34.9612      ⟶   (high income)[35]

        2        4.6114      ⟶   (low income)

        3       14.8266      ⟶   (mid income)

Number of Cases in each Cluster.

  Cluster      unweighted cases    weighted cases

      1              66.0              66.0
      2            1417.0            1417.0
      3             510.0             510.0

      Variable with cluster membership created: qcl_2
```

Go to DATA/ DEFINE VARIABLE and define a variable label and value labels for the three values of the newly created variable *qcl_2* (see section 1.2 for instructions). On the data sheet, the new variable will be located in the last column. We use this variable to conduct an interesting analysis in section 10.1.a.

To take quizzes on topics within each chapter go to http://www.spss.org/wwwroot/spssquiz.asp

[35] SPSS does not label them as "Low", "Medium," or "High." To do so, go to DATA/ DEFINE VARIABLE and, following the steps shown in section 1.2, assign these labels to the values 1, 2, and 3.

Ch 3. UNIVARIATE ANALYSIS

A proper analysis of data must begin with an analysis of the statistical attributes of each variable in isolation - univariate analysis. From such an analysis we can learn:

- how the values of a variable are distributed - normal, binomial, etc.[36]
- the central tendency of the values of a variable (mean, median, and mode)
- dispersion of the values (standard deviation, variance, range, and quartiles)
- presence of outliers (extreme values)
- if a statistical attribute (e.g. - mean) of a variable equals a hypothesized value

The answer to these questions illuminates and motivates further, more complex, analysis. Moreover, failure to conduct univariate analysis may restrict the usefulness of further procedures (like correlation and regression). Reason: even if improper/incomplete univariate analysis may not directly hinder the conducting of more complex procedures, the interpretation of output from the latter will become difficult (because you will not have an adequate understanding of how each variable behaves).

This chapter explains different methods used for univariate analysis. Most of the methods shown are basic - obtaining descriptive statistics (mean, median, etc.) and making graphs. (Sections 3.2.e and 3.4.b use more complex statistical concepts of tests of significance.)

In section 3.1, you will learn how to use bar, line, and area graphs to depict attributes of a variable.

In section 3.2, we describe the most important univariate procedures - frequencies and distribution analysis. The results provide a graphical depiction of the distribution of a variable and provide statistics that measure the statistical attributes of the distribution. We also do the Q-Q and P-P tests and non-parametric testing to test the type of distribution that the variable exhibits. In particular, we test if the variable is normally distributed, an assumption underlying most hypothesis testing (the Z, T, and F tests).

Section 3.3 explains how to get the descriptive statistics and the boxplot (also called "Box and Whiskers plot" for each numeric variable. The boxplot assists in identifying outliers and extreme values.

Section 3.4 describes the method of determining whether the mean of a variable is statistically equal to a hypothesized or expected value. Usefulness: we can test to discover whether our sample is similar to other samples from the same population.

Also see chapter 14 for non-paramateric univariate methods like the Runs test to determine if a variable is randomly distributed.

[36] Check your textbook for descriptions of different types of distributions.

Ch 3. Section 1 Graphs (bar, line, area, and pie)

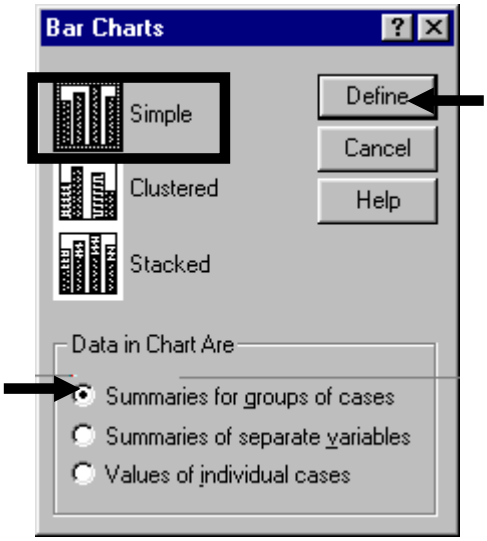
Ch 3. Section 1.a. Simple bar graphs

Bar graphs can be used to depict specific information like mean, median, cumulative frequency, cumulative percentage, cumulative number of cases, etc.

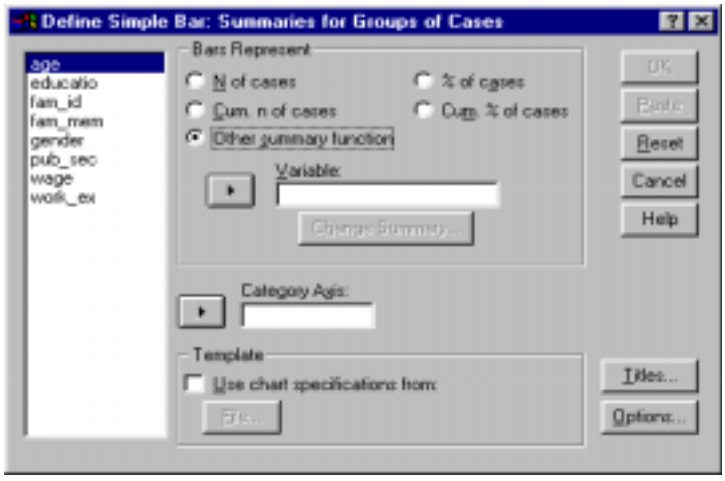
Select GRAPHS/BAR.

Select "Simple" and "Summaries of Groups of Cases."

Click on the button "Define."

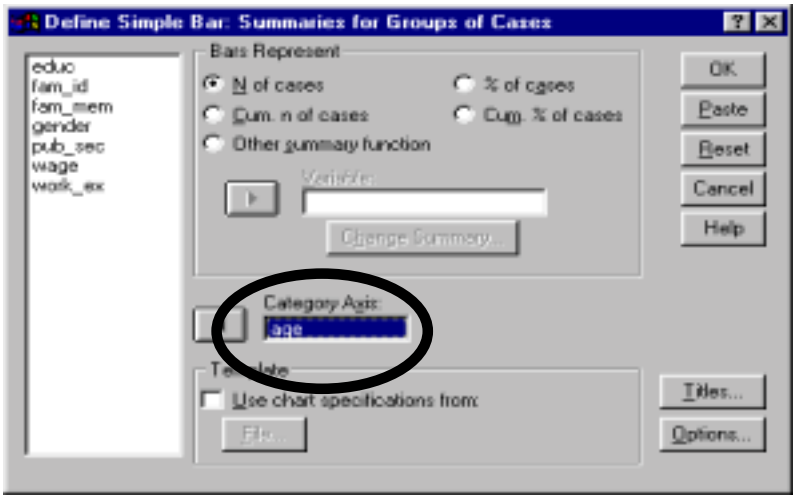


The following dialog box will open up.



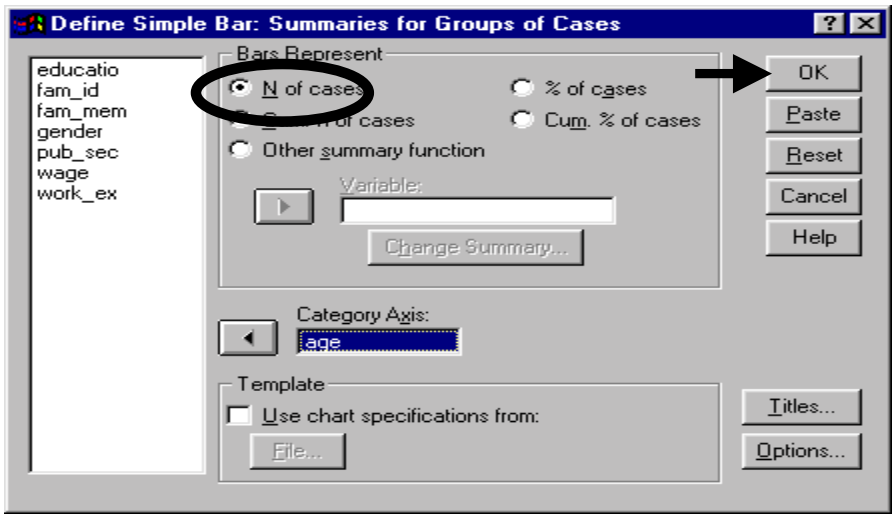
Note: You will see very similar dialog boxes if you choose to make a bar, line, area, or pie graph. Therefore, if you learn any one of these graph types properly you will have learned the other three. The choice of graph type should be based upon the ability and power of the graph to depict the feature you want to show.

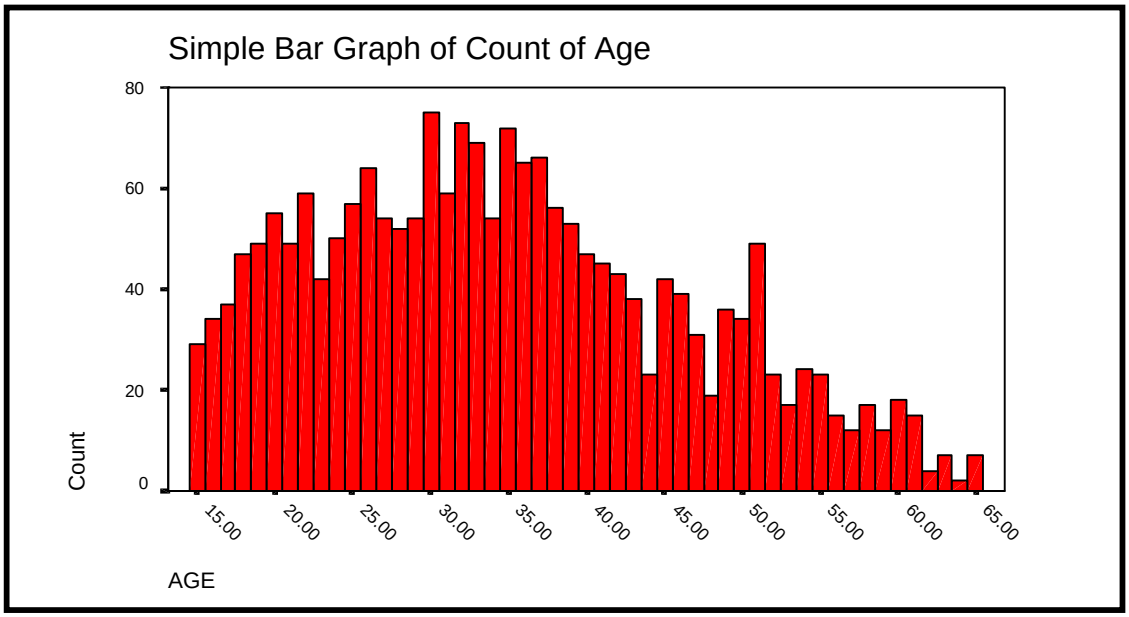
Select the variable *age*. Place it into the box "Category Axis." This defines the X-axis.



On the top of the dialog box you will see the options for the information on the variable *age* that can be shown in the bar graph. Select the option "N of Cases."

Click on "OK."





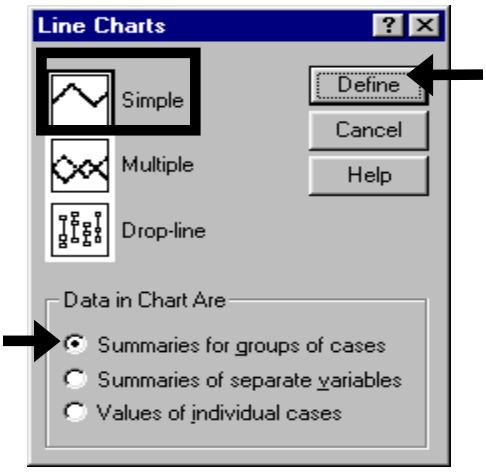
Ch 3. Section 1.b. Line graphs

If you prefer the presentation of a line (or area) graph, then the same univariate analysis can be done with line (or area) graphs as with bar charts.

Select GRAPHS/ LINE.

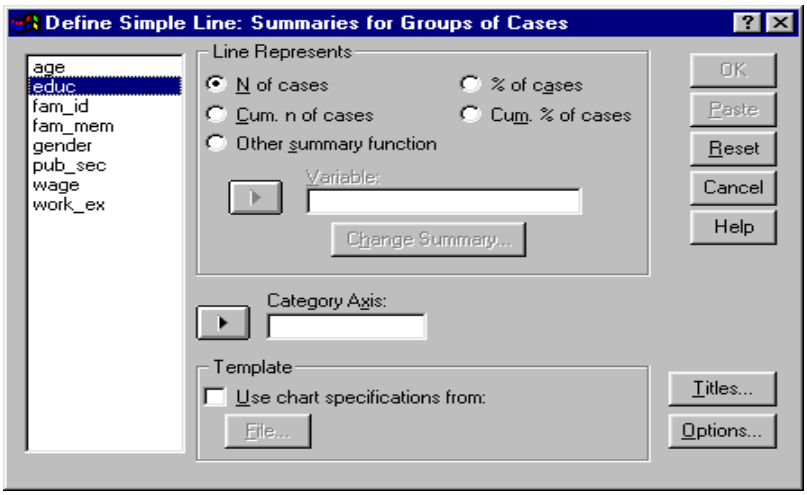
Select "Simple" and "Summaries of Groups of Cases."

Click on the button "Define."



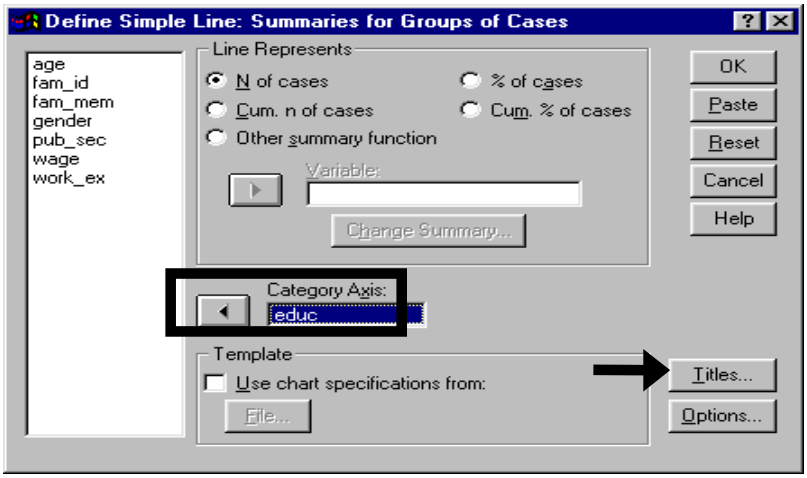
The following dialog box will open.

It looks the same as the box for bar graphs. The dialog boxes for bar, line, and area graphs contain the same options.



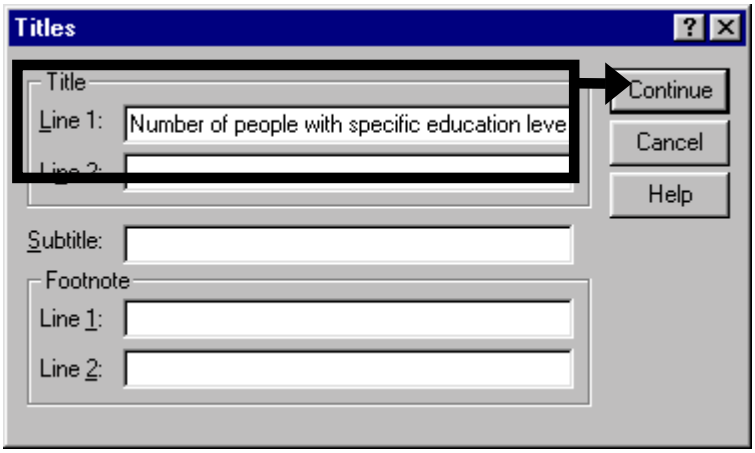
Place the variable *educ* into the box "Category Axis." This defines the X-axis.

Click on the button "Titles."



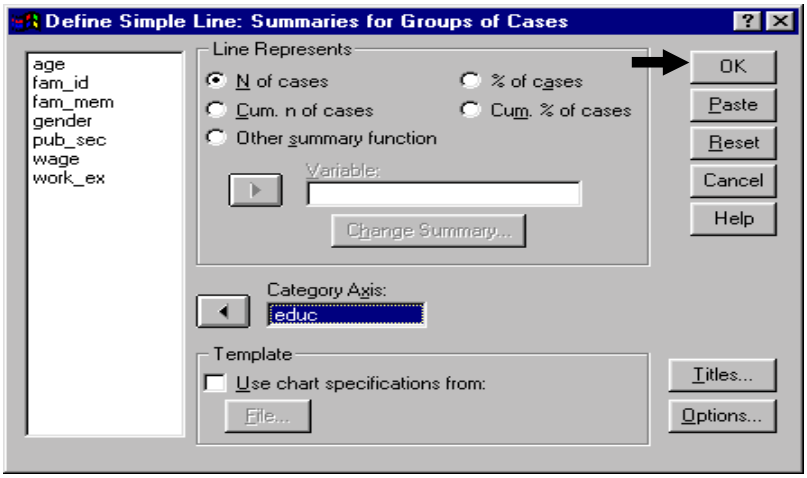
Enter text for the title and/or footnotes.

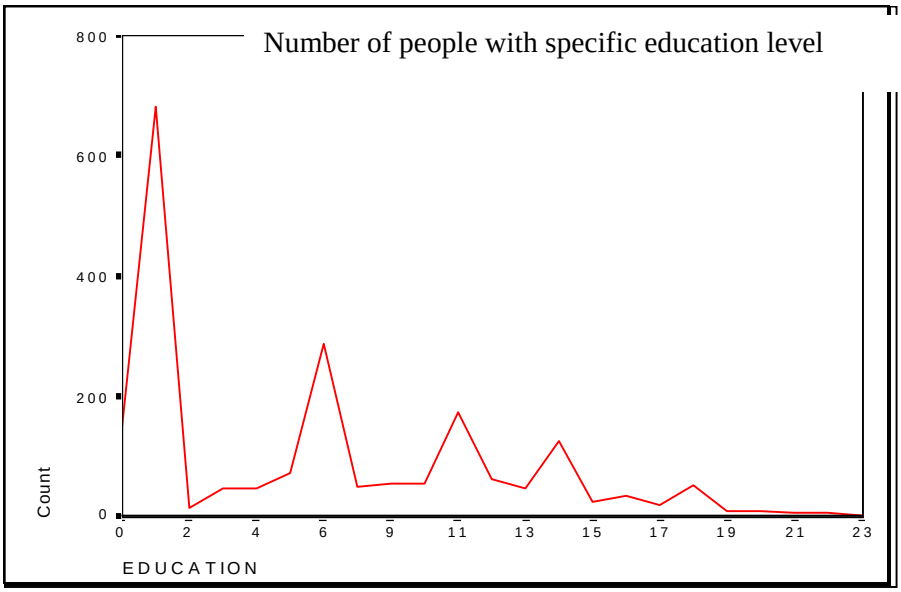
Click on "Continue."



Click on "OK."

Note: Either a bar or pie graph are typically better for depicting one variable, especially if the variable is categorical.





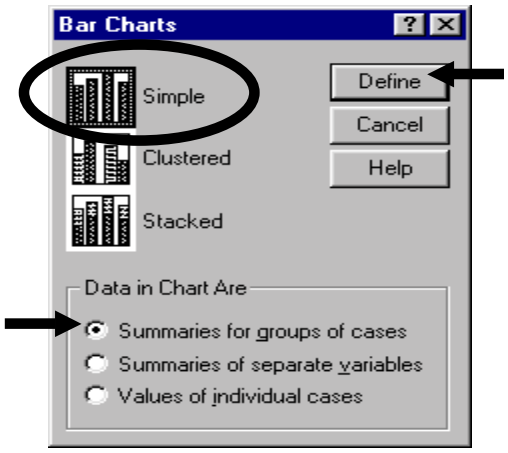
Ch 3. Section 1.c. Graphs for cumulative frequency

You may be interested in looking at the cumulative frequency or cumulative percentages associated with different values of a variable. For example, for the variable *age*, it would be interesting to see the rate at which the frequency of the variable changes as *age* increases. Is the increase at an increasing rate (a convex chart) or is at a decreasing rate (a concave chart)? At what levels is it steeper (i.e. - at what levels of *age* are there many sample observations)? Such questions can be answered by making cumulative bar, line, or area graphs.

Select GRAPHS/BAR[37].

Select "Simple" and "Summaries of Groups of Cases."

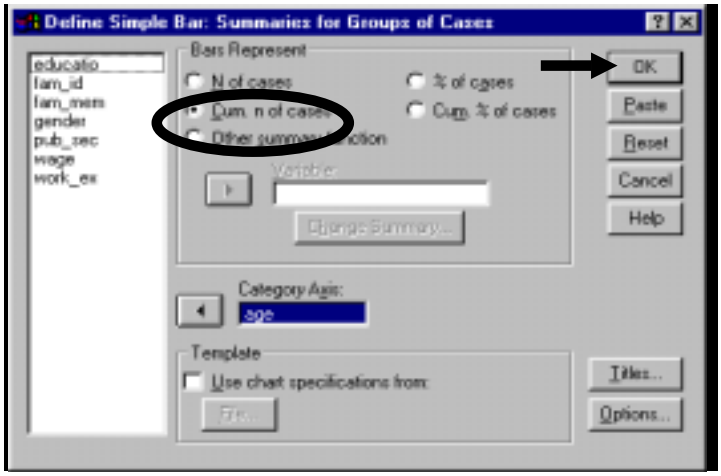
Click on the button "Define."



Select the variable *age*. Place it into the "Category Axis" box. This defines the X-axis.

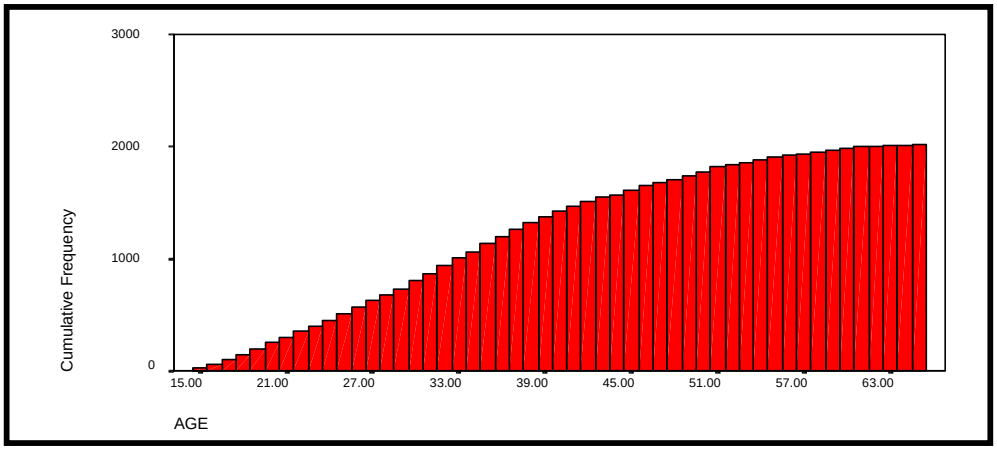
This time choose "Cum. n of cases" in the option box "Bars Represent." The result is the cumulative distribution of the variable *age*.

Note: if you choose the "Cum. % of cases," then the height of the bars will represent percentages. This may be better if you want to perform the procedure for several variables and compare the results.



Click on "OK."

[37] If you prefer to use line or area graphs, use similar steps.



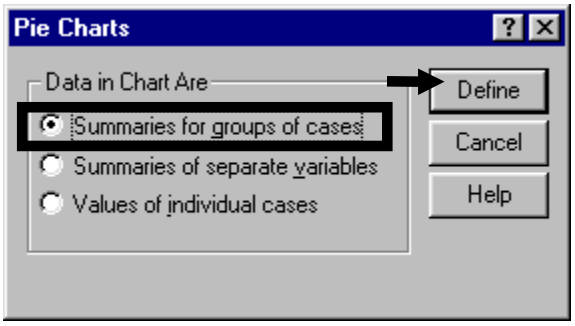
Ch 3. Section 1.d. Pie graph

These charts provide a lucid visual summary of the distribution of a dummy variable or a categorical variable with several categories. Pie charts are only used when the values a variable can take are limited[38]. In our data set, *gender* and *pub_sec* are two such variables.

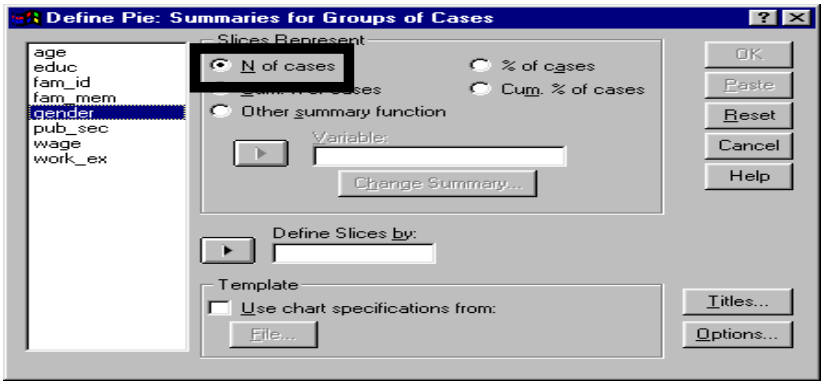
Go to GRAPHS/ PIE.

Select the option "Summaries for groups of cases."

Click on "Define."



Choose the option "N of Cases."



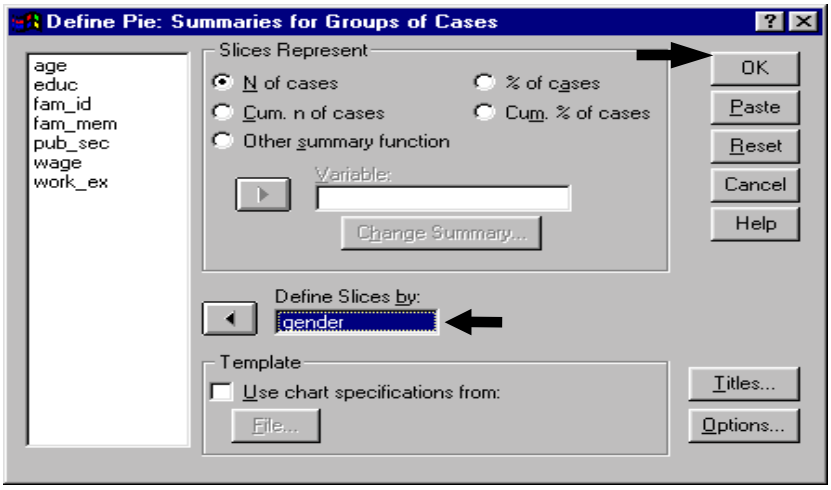
Note the similarity of the corresponding dialog boxes for bar, line, and area graphs. As we pointed out earlier, if you know how to make one of the graph types, you can easily render the other types of graphs.

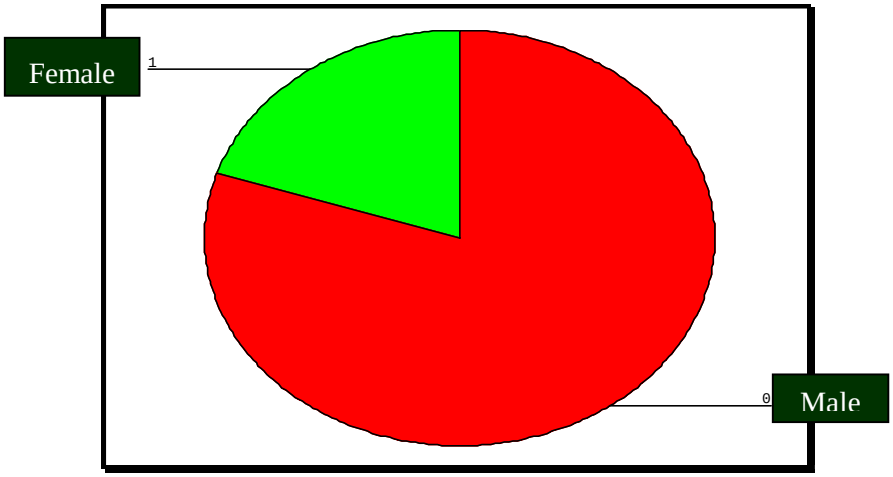
[38] A pie chart with too many "slices" of the pie is difficult to read.

Move the variable *gender* into the box "Define Slices by."

Note: Click on "Titles" and type in a title.

Click on "OK."





Ch 3. Section 2 Frequencies and distributions

This is the most important univariate procedure. Conducting it properly, and interpreting the output rigorously, will enable you to understand the major attributes of the frequency distribution of each variable[39].

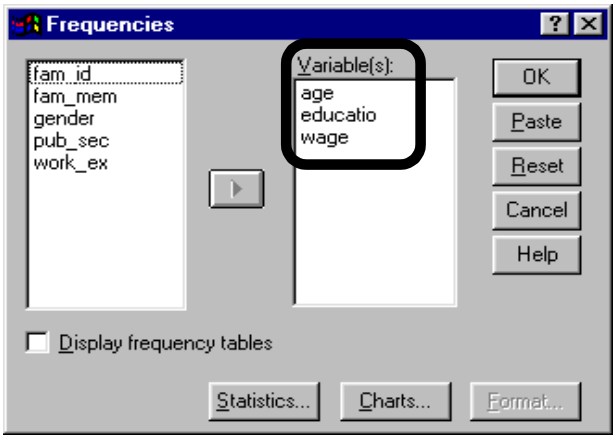
[39] Using histograms and frequency statistics, we can answer several questions about the distribution of individual variables. What is the nature of the distribution of a variable: normal, lognormal, exponential, uniform, etc? Is the variable distributed normally? Is it skewed, and if so, to the left or right? Is there a range in which many observations occur? Are there outliers, and if there are, where do they lie? What is the mode?

Note: check your statistics text for definitions/descriptions of the terms we use. We do not go into the details of statistical descriptions.

Ch 3. Section 2.a. The distribution of variables - histograms and frequency statistics

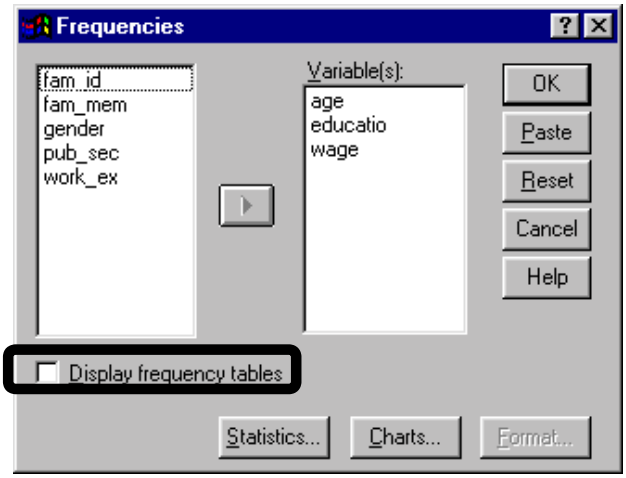
Go to STATISTICS/ SUMMARIZE/ FREQUENCIES.

Select the variables and move them into the box "Variable(s)."



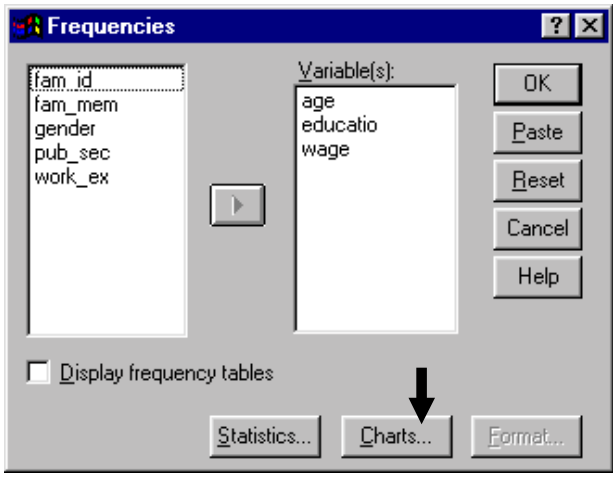
Creating Histograms of dummy (*gender* and *pub_sec*) or ID variables (*fam_id*) is not useful. The former have only two points on their histogram, the latter has too many points (as each ID is unique). We will therefore only make histograms of continuous or categorical variables.

Unless the variables you have chosen are categorical or dummy (i.e. - they have only a few discrete possible values), deselect the option "Display Frequency Tables." Otherwise, you will generate too many pages of output.

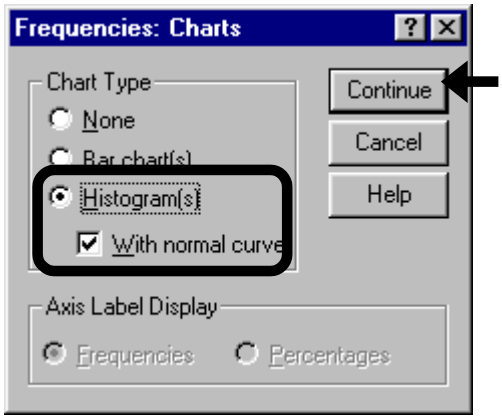


Note: Conduct the frequencies procedure twice - Once for continuous variables (deselecting the option "Display Frequency Tables") and once for categorical and dummy variables (this time choosing the option "Display Frequency Tables").

Now you must instruct SPSS to construct a histogram for each of the chosen variables. Click on the button "Charts."



Choose to draw a histogram with a normal curve - select the option "Histogram" and click on the box to the left of the title "With normal curve[40]."

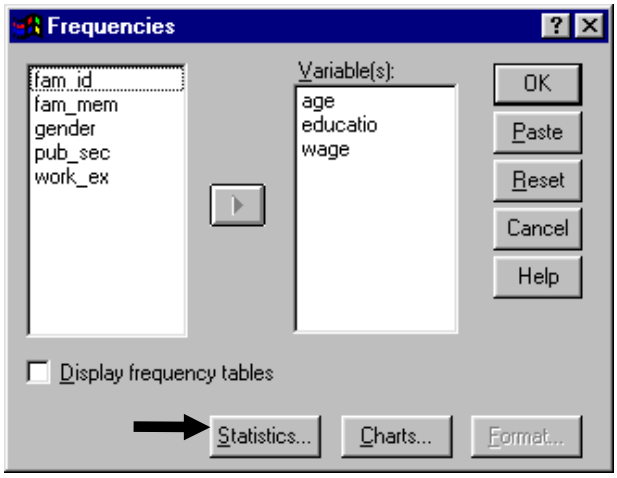


Throughout this chapter we stress methods of determining whether a variable is distributed normally. What is the normal distribution and why is it so important? A variable with a normal distribution has the same mode, mean, and median, i.e. - its most often occurring value equals the average of values and the mid-point of the values. Visually, a normal distribution is bell-shaped (see the "idealized normal curve" in the charts on page 3-12) - the left half is a mirror image of the right half. The importance stems from the assumption that "if a variable can be assumed to be distributed normally, then several inferences can be drawn easily and, more importantly, standardized tests (like the T and F tests shown in chapters 3-10) can be applied." In simpler terms: "normality permits the drawing of reliable conclusions from statistical estimates."

Note: We repeat - conduct the frequencies procedure twice. Once for continuous variables (deselecting the option "Display Frequency Tables" but choosing the option "With Normal Curve") and once for categorical and dummy variables (this time choosing the option "Display Frequency Tables" but deselecting "With Normal Curve").

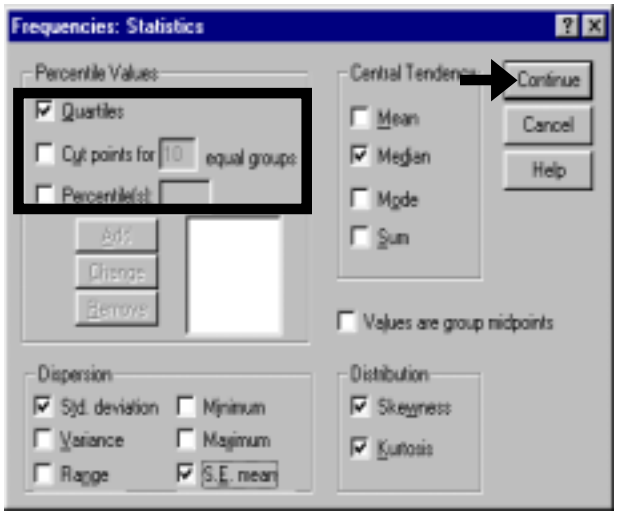
Click on "Continue."

We also want to obtain descriptives. Click on the button "Statistics."



[40] The latter will depict a normal distribution superimposed on the histogram. If the histogram and the normal curve are similar, then the variable is normally distributed. If they do not, then you must conduct Q-Q or P-P graphs to test for the type of distribution of each variable (see section 3.2.b).

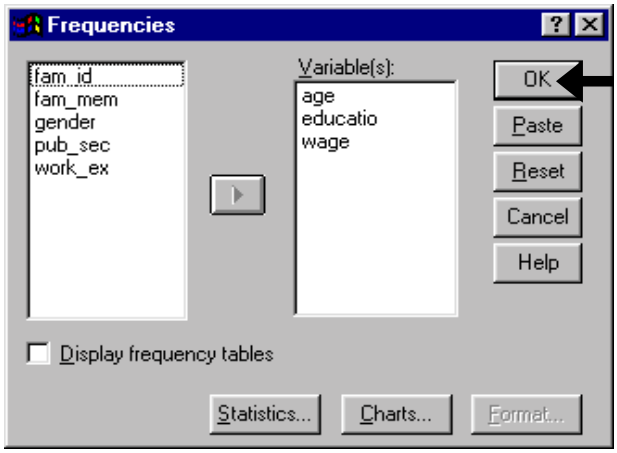
Select the options as shown. These statistics cover the list of "descriptive statistics.[41, 42]"



The options under "Percentile values" can assist in learning about the spread of the variable across its range. For a variable like *wage*, choosing the option "Quartiles" provides information on the *wage* ranges for the poorest 25%, the next 25%, the next 25%, and the richest 25%. If you wish to look at even more precise sub-groups of the sample, then you can choose the second option "Cut points for (let's say) 10 equal groups." The option percentile is even better - you can customize the exact percentiles you want - For instance: "poorest 10%, richest 10%, lower middle class (10-25%), middle class (25-75%), upper middle class (75-90%)," etc.

Click on "Continue."

Click on OK.



The output will have one frequency table for all the variables and statistics chosen and one histogram for each variable.

[41] The median and interquartile range (75^{th} - 25^{th} percentile or 3^{rd}- 1^{st} quartile) have a useful property - they are not affected by some outliers or extreme values. Another measure of dispersion is the Semi-Interquartile Range, defined as [(3^{rd} - 1^{st} quartile) divided by 2].

[42] Skewness measures the degree of symmetry in the distribution. A symmetrical distribution includes left and right halves that appear as mirror images. A positive skew occurs if skewness is greater than zero. A negative skew occurs if skewness is less than ten. A positive skewness indicates that the distribution is left heavy. You can consider values between 0 and 0.5 as indicating a symmetrical distribution. Kurtosis measures the degree to which the frequencies are distributed close to the mean or closer to the extremes. A bell-shaped distribution has a kurtosis estimate of around 3. A center-heavy (i.e. - close to the mean) distribution has an estimated kurtosis greater than 3. An extreme-heavy (or flat) distribution has a kurtosis estimate of greater than 3. (All in absolute terms)

Statistics

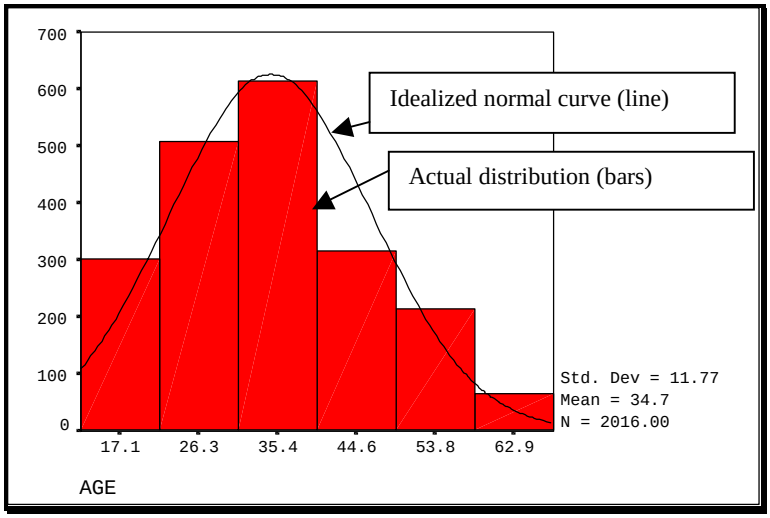
			AGE	education	WAGE/HR	WORK_EX
N	Valid	Statistic	2016	2016	2016	2016
	Missing	Statistic	0	0	0	0
Mean		Statistic	34.68	6.09	9.05	10.82
Median		Statistic	33.50	5.00	5.95	8.29
Std. Deviation		Statistic	11.77	5.60	11.23	9.17
Skewness		Statistic	.39	.69	6.29	1.10
		Std. Error	.055	.055	.055	.055
Kurtosis		Statistic	-.611	-.649	67.020	.850
		Std. Error	.109	.109	.109	.109
Percentiles	25.0000	Statistic	25.00	1.00	3.69	3.50
	50.0000	Statistic	33.50	5.00	5.95	8.29
	75.0000	Statistic	43.00	11.00	11.32	16.00

In the next three graphs, the heights of the bars give the relative frequencies of the values of variables. Compare the bars (as a group) with the normal curve (drawn as a bell-shaped line curve). All three variables seem to be left heavy relative to the relevant normal curves, i.e. - lower values are observed more often than higher values for each of the variables.

We advise you to adopt a broad approach to interpretation: consult the frequency statistics result (shown in the table above), the histograms (see next page), and your textbook.

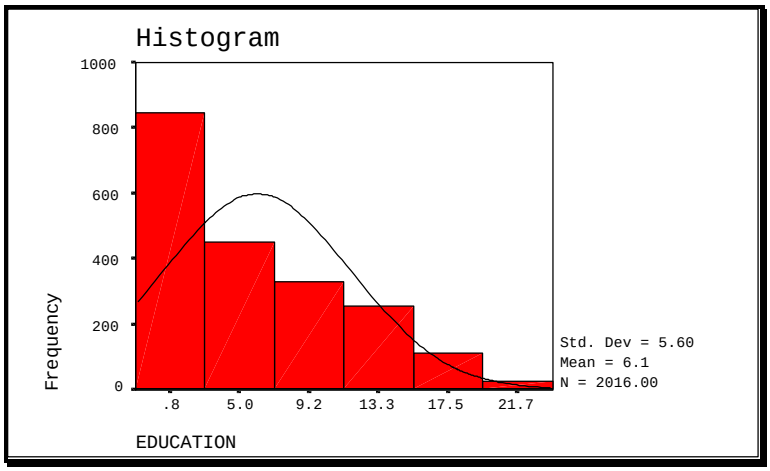
Age is distributed more or less normally[43] but with a slightly heavier distribution around the lower half.

On the lower-right corner, the chart provides the most important statistics - standard deviation, mean, and sample size. (The other statistics - like the median, mode, range, skewness, and kurtosis) are usually more visually identifiable from a histogram. The mode is the highest bar, the median has half the area (under the shaded bars) to its left, and the skewness and kurtosis are measures of attributes that are easily identifiable.



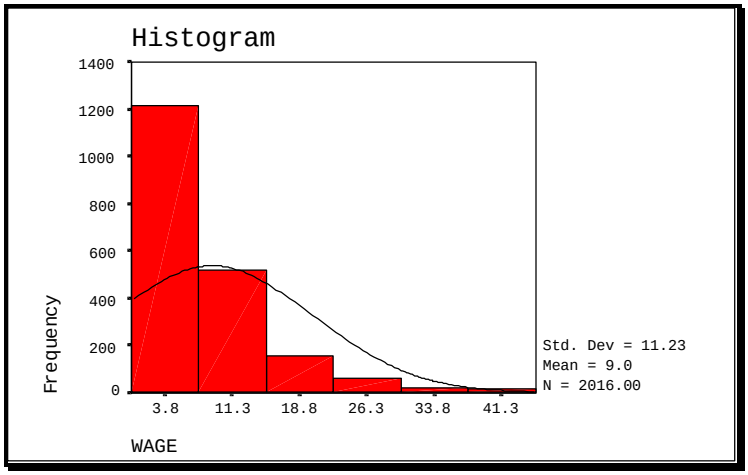
[43] See how the bars follow a similar pattern to the "idealised normal curve."

Education does not seem to have a normal distribution. It has a mode at its minimum (the mode is the value on the X-axis that corresponds to the highest bar).



Wage also does not look normally distributed. It is left-skewed.

The P-P or Q-Q tests and formal tests are used to make a more confident statement on the distribution of wage. These tests are shown in sections 3.2.b - 3.2.e.



Ch 3. Section 2.b. Checking the nature of the distribution of continuous variables

The next step is to determine the nature of the distribution of a variable.

The analysis in section 3.2.a showed that *education*, *age*, and *wage* might not be distributed normally. But the histograms provide only a rough visual idea regarding the distribution of a variable. Either the P-P or Q-Q procedure is necessary to provide more formal evidence[44]. The P-P tests whether the Percentiles (quartiles in the case of the Q-Q) of the variables' distribution match the percentiles (quartiles in the case of the Q-Q) that would indicate that the distribution is of the type being tested against.

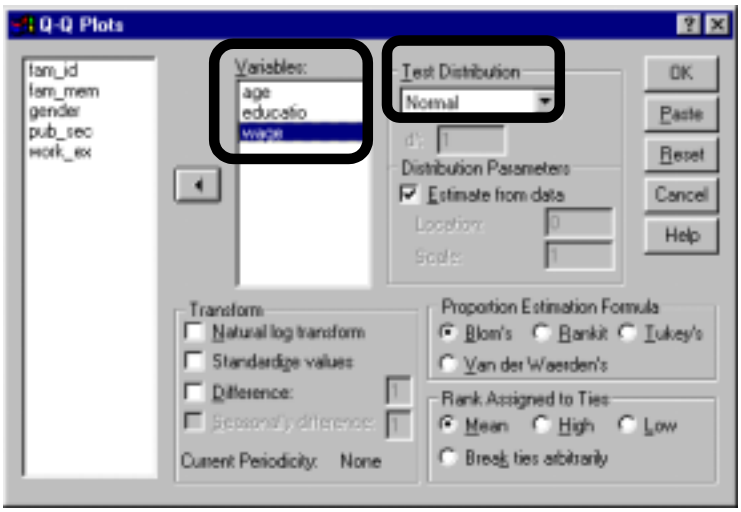
[44] The two methods are roughly equivalent, so we use only one of them here - the Q-Q test. Many statisticians consider these methods to be insufficient. Section 3.2.e shows the use of a more stringent/formal test.

Checking for normality of continuous variables

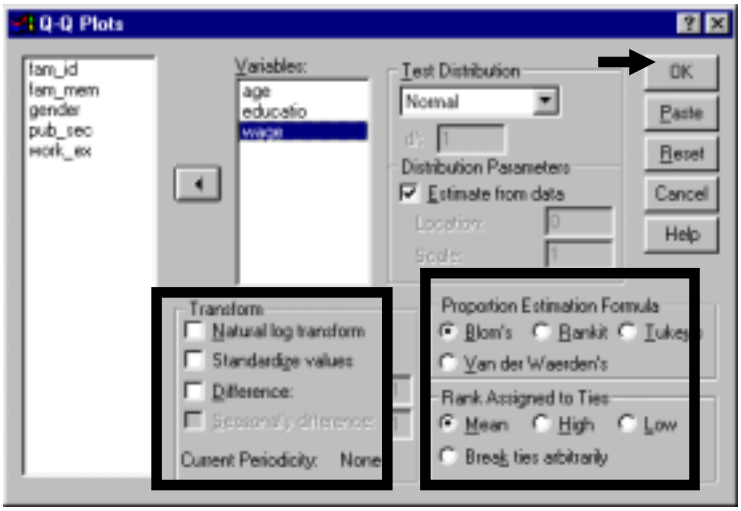
Go to GRAPHS/Q-Q.

Select the variables whose "normality" you wish to test.

On the upper-right side, choose the distribution "Normal" in the box "Test Distribution." This is indicating to SPSS to "test whether the variables *age*, *education*, and *wage* are normally distributed."



In the area "Transform," deselect all[45]. In the areas "Proportion Estimation Formula"[46] and "Rank Assigned to Ties,"[47] enter the options as shown.



A digression: The "Proportion Estimation Formula" uses formulae based on sample size and rank to calculate the "expected" normal distribution.

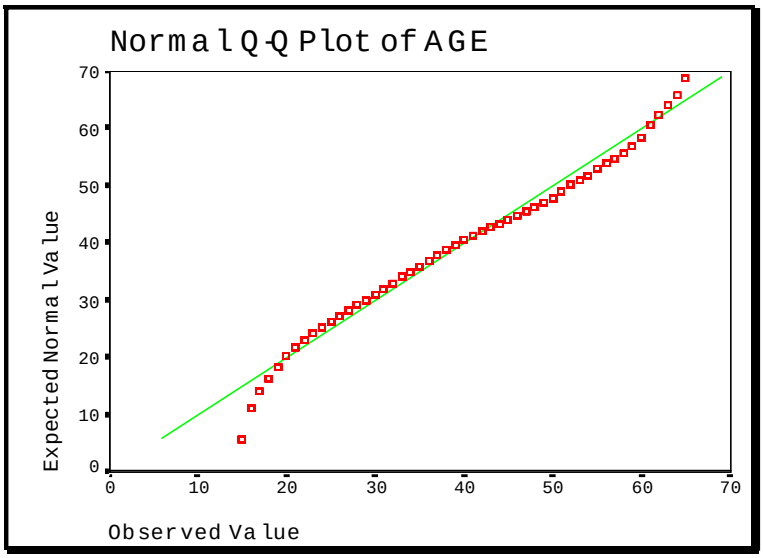
Click on "OK."

[45] The next section shows the use of the "Transformation" feature.

[46] A detailed explanation of this is beyond the scope of this book.

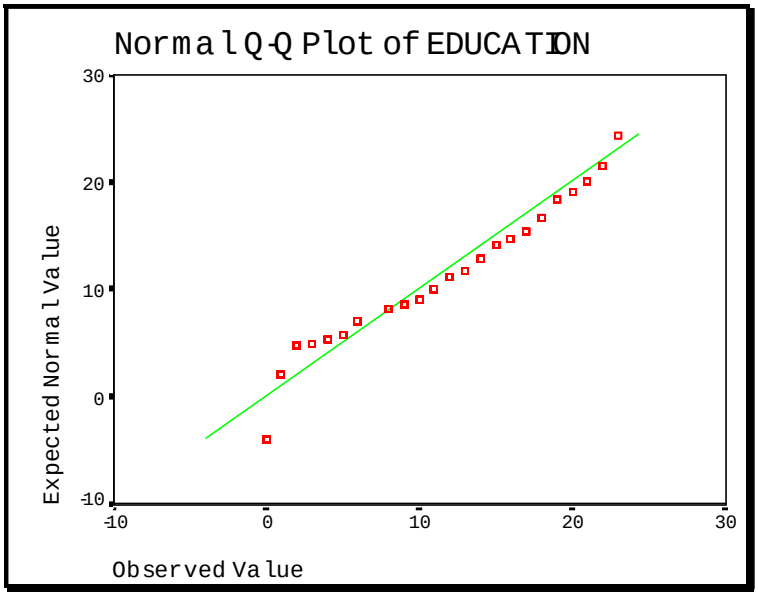
[47] If two values have the same rank (e.g. - if both are "18th largest") what rank should be given to them in the mathematical procedure underlying the P-P or Q-Q? The choice "Mean" implies that the mean rank would be used (continuing the example, this number would be 18.5).

In the following three graphs, observe the distance between the diagonal line and the dotted curve. The <u>smaller</u> the gap between the two, the higher the chance of the distribution of the variable being the same as the "Test Distribution," *which in this case is the normal distribution.*



The Q-Q of age suggests that it is normally distributed, as the Histogram indicated in section 3.2.a.

The Q-Q of education suggests that the variable is normally distributed, in contrast to what the histogram indicated in section 3.2.a.

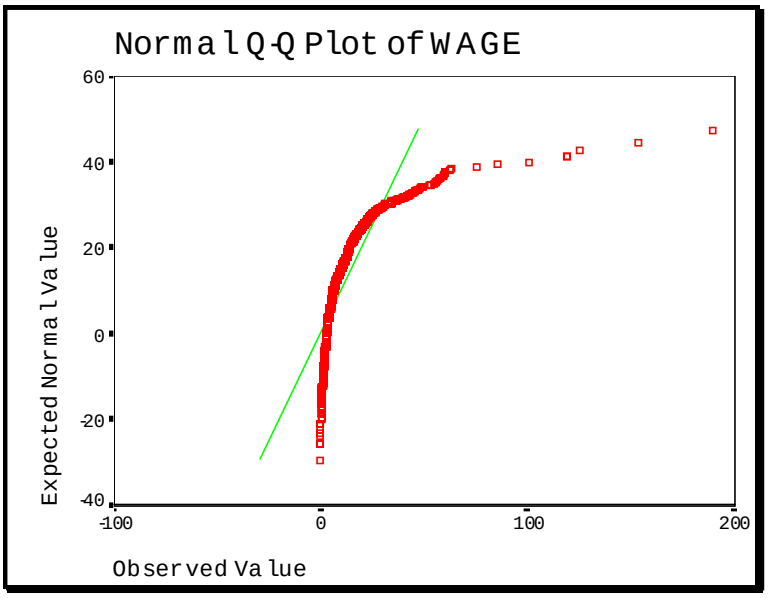


Note: The P-P and Q-Q are not formal tests and therefore cannot be used to render conclusive answers. For such answers, use the formal[48] testing method shown in section 3.2.e or other methods shown in your textbook.

[48] A "formal" testing method typically involves the use of a hypothesis test that, in turn, uses a test like the T, F, Z, etc. An "informal" testing method is typically a graphical depiction.

Wage is not normally distributed(the dotted curve definitely does not coincide with the straight line).

Although the histogram showed that all three variables might be non-normal, the Q-Q shows that only one variable (*wage*) is definitely not normally distributed.



Ch 3. Section 2.c. Transforming a variable to make it normally distributed

The variable *wage* is non-normal as shown in the chart above. The skew hints that the log of the variable may be distributed normally. As shown below, this is borne out by the Q-Q obtained when a log transformation[49] of *wage* is completed.

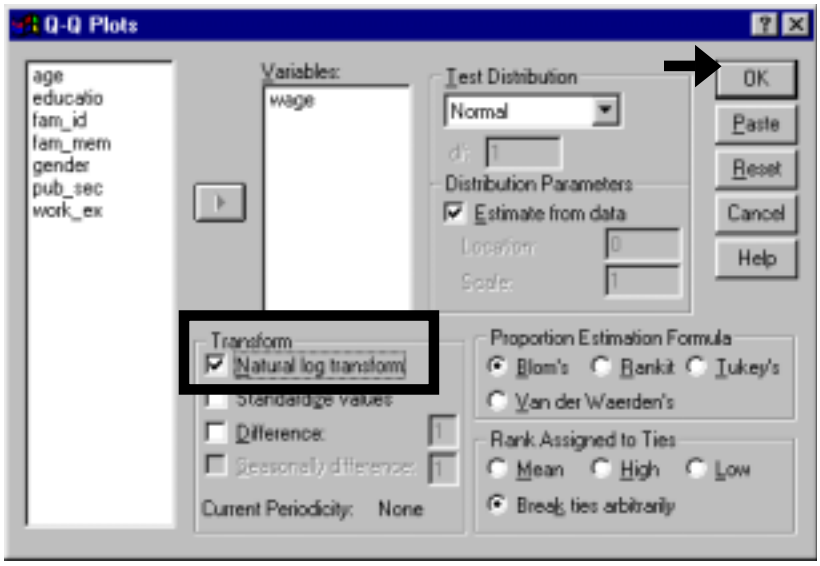
Go to GRAPHS/Q-Q.

Place the variable *wage* into the box "Variable."

On the right, choose "Normal" in the "Test Distribution" box.

In the "Transform" options area, choose "Natural Log Transform[50]."

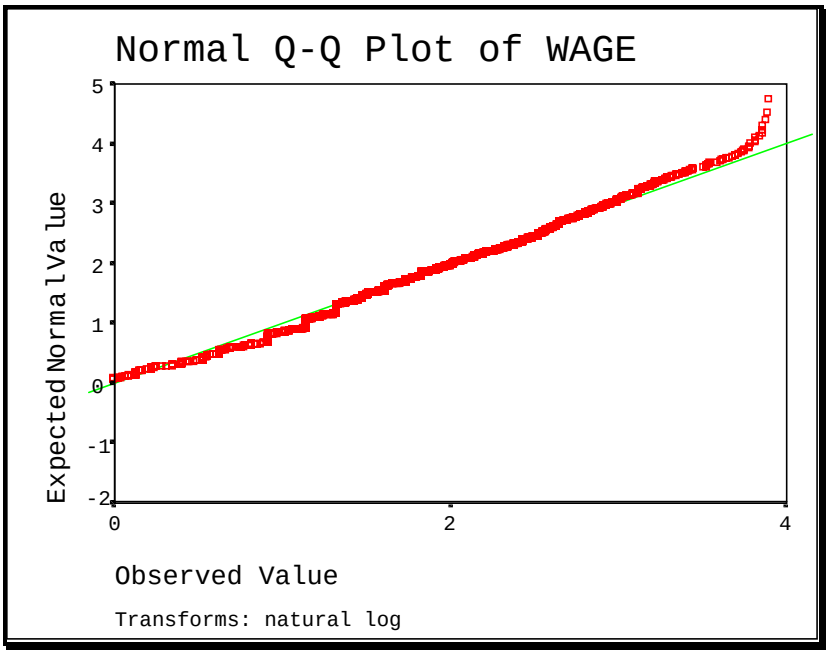
Click on "OK."



[49] If the term "log" or the concept of variable "transformation" are not familiar (and confusing), you can skip over to section 3.2.e.

[50] In effective, SPSS is not testing the variable *wage* for normality but, instead, the variable *log of wage*.

The log transformation of wage is normal as can be seen in the next chart (the dotted curve coincides with the straight line).



Ch 3. Section 2.d. Testing for other distributions

The Q-Q and P-P can be used to test for non-normal distributions. Following the intuition of the results above for the variable *wage*, we test the assumption that *wage* follows a lognormal distribution.

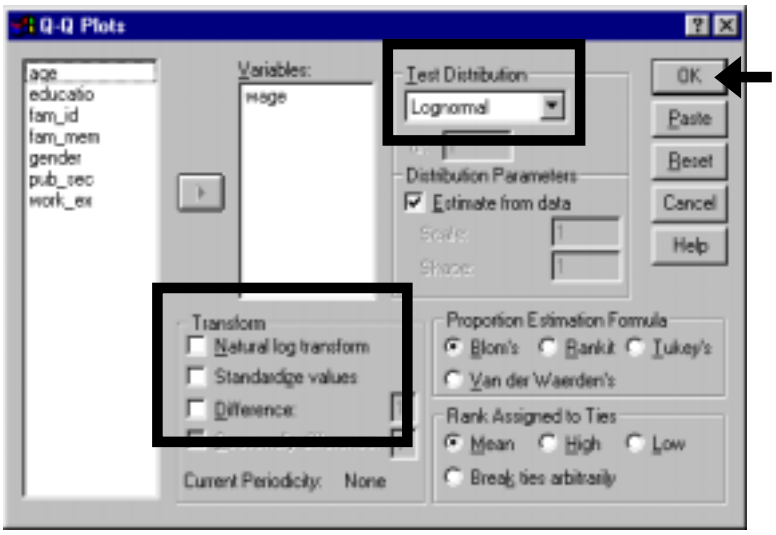
Note: Check your statistics book for descriptions of different distributions. For understanding this chapter all you need to know is that the lognormal is like a normal distribution but with a slight tilt toward the left side (lower values occur more frequently than in a normal distribution).

Place the variable *wage* into the box "Variable."

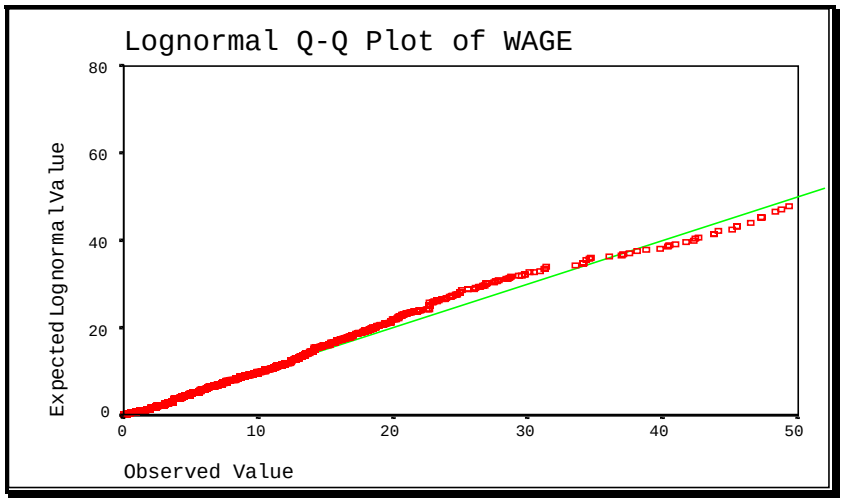
On the right, choose "Lognormal" in the box "Test Distribution."

In the "Transform" options area, deselect all.

Click on "OK."



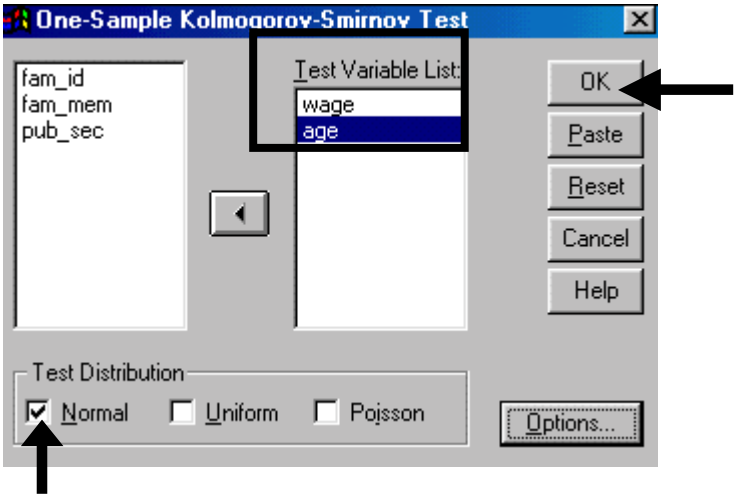
The Q-Q shows that wage is distributed lognomally (the dotted curve coincides with the straight line).



Note: If the terms (such as lognormal) are unfamiliar to you, do not worry. What you need to learn from this section is that the P-P and Q-Q test against several types of standard distributions (and not only the normal distribution).

Ch 3. Section 2.e. A Formal test to determine the distribution type of a variable

The P-P and Q-Q may not be sufficient for determining whether a variable is distributed normally. While they are excellent "visual" tests, they do not provide a mathematical hypothesis test that would enable us to say that the "hypothesis that the variable's distribution is normal can be accepted." For that we need a formal testing method. Your textbook may show several such methods (a common one is the Jacque-Berra). In SPSS, we found one such formal test - the "Kolmogorov-Smirnov" test. Using this test, we determine whether the variables are distributed normally.



Go to STATISTICS / NONPARAMETRIC TESTS / 1-SAMPLE K-S.

Move the variables whose normality you wish to test into the box "Test Variable List."

Choose the option "Normal." Click on "OK." The result is in the next table. The test statistic used is the Kolgoromov-Smirnov (or simply, K-S) Z. It is based upon the Z distribution.

One-Sample Kolmogorov-Smirnov Test

		Age in complete years	Hourly Net Income
N		1444	1446
Normal Parameters[a,b]	Mean	34.27	1.6137
	Std. Deviation	11.14	1.7733
Most Extreme Differences	Absolute	.111	.193
	Positive	.111	.171
	Negative	-.065	-.193
Kolmogorov-Smirnov Z		4.229	7.355
Asymp. Sig. (2-tailed)		.83	.21

a. Test distribution is Normal.

b. Calculated from data.

In class, you may have been taught to compare this estimated Z to the appropriate[51] value in the Z-distribution/test (look in the back of your book - the table will be there along with tables for the F, T , Chi-Square, and other distributions.) SPSS makes this process very simple! It implicitly conducts the step of "looking" at the appropriate table entry and calculates the "Significance" value. ALL YOU MUST DO IS LOOK AT THE VALUE OF THIS "SIGNIFICANCE" VALUE. The interpretation is then based upon where that value stands in the decision criterion provided after the next table.

If sig is less than 0.10, then the test is significant at 90% confidence (equivalently, the hypothesis that the distribution is normal can be rejected at the 90% level of confidence). This criterion is considered too "loose" by some statisticians.

If sig is less than 0.05, then the test is significant at 95% confidence (equivalently, the hypothesis that the distribution is normal can be rejected at the 95% level of confidence). This is the standard criterion used.

If sig is less than 0.01, then the test is significant at 99% confidence (equivalently, the hypothesis that the distribution is non-normal can be rejected at the 99% level of confidence). This is the strictest criterion used.

You should memorize these criteria, as nothing is more helpful in interpreting the output from hypothesis tests (including all the tests intrinsic to every regression and ANOVA analysis). You will encounter these concepts throughout sections or chapters 3.4, 4.3, 5, 7, 8, 9, and 10.

In the tests above, the sig value implies that the test indicated that both variables are normally distributed. (The null hypothesis that the distributions are normal cannot be rejected.)

[51] The aptness is based upon the degrees of freedom(s), level of significance, etc.

Ch 3. Section 3 Other basic univariate procedures (Descriptives and Boxplot)

The "Descriptives" are the list of summary statistics for many variables - the lists are arranged in a table.

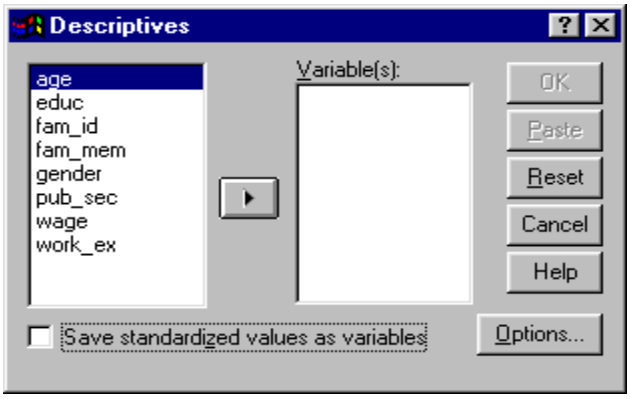
Boxplots are plots that depict the cut-off points for the four quartiles: 25th percentile, 50th percentile, 75th percentile, and the 99.99th percentile. Essentially, it allows us to immediately read off the values that correspond to each quarter of the population (if the variable used is *wage*, then "25% youngest," "50% youngest,"…and so on.) Section 3.3.b. has an example of boxplots and their interpretation.

Ch 3. Section 3.a. Descriptives

Section 3.2.a showed you how to obtain most of the descriptive statistics (and also histograms) using the "frequencies" procedure (so you may skip section 3.3.a).

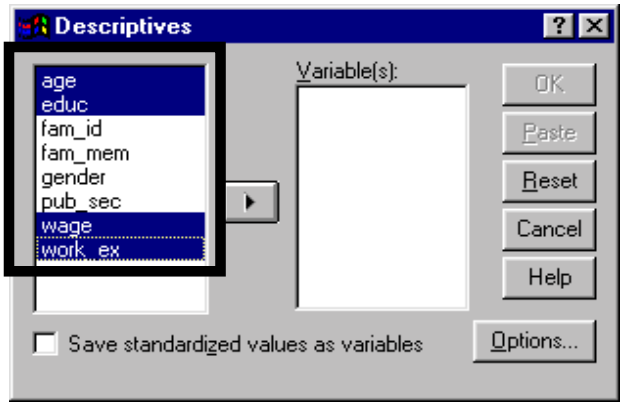
Another way to obtain the descriptives is described below.

Go to STATISTICS/SUMMARIZE/ DESCRIPTIVES. A very simple dialog box opens.



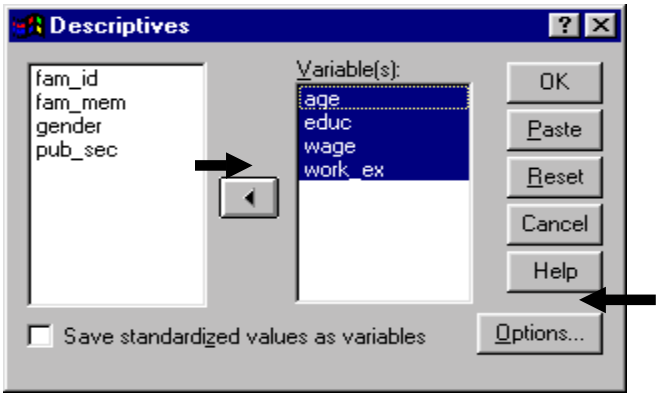
Select the variable whose descriptives you would like to find. Do not select dummy or categorical variables because they are qualitative (making any quantitative result like "mean=0.3" may be irrelevant).

To select multiple variables, click on the first one, press the CTRL key and, keeping the key pressed, choose the other variables.



Move them into the box "Variable(s)."

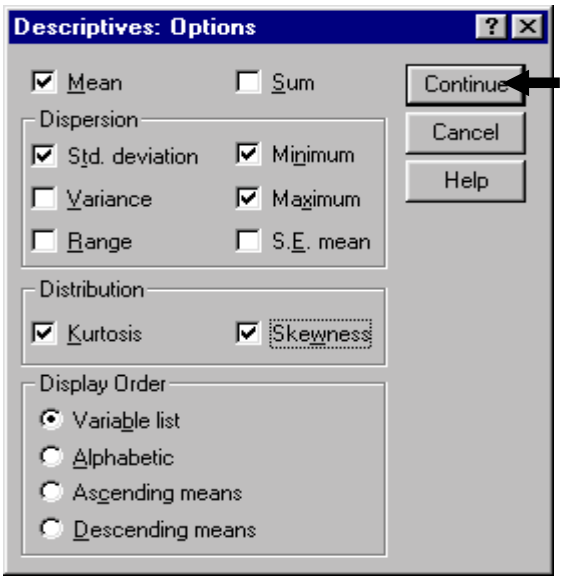
You must choose the statistics with which you want to work. Click on the button "Options."



Select the appropriate statistics[52].

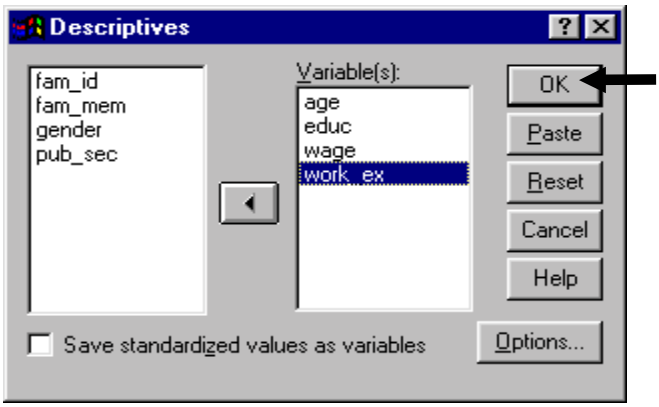
Note: Refer to your textbook for detailed explanations of each statistic.

Click on "Continue."



Click on "OK."

The output is shown in the next table. Interpretation is the same as in section 3.2.a. Note the poor formatting of the table. In section 11.1 you will learn how to improve the formatting of output tables such as this one.



[52] Note that almost all of these statistics can be obtained from STATISTICS/ SUMMARIZE/ FREQUENCIES (see section 3.1).

	N	Minimum	Maximum	Mean	Std. Error	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Age	2016	15.00	66.00	11.77	34.68	0.39	0.06	-0.61	0.11
Education	2016	0.00	23.00	5.60	6.09	0.69	0.06	-0.64	0.11
Wage	2016	0.00	189.93	11.23	9.04	6.28	0.06	67.00	0.11
Work Experience	2016	0.00	48.00	9.17	10.81	1.09	0.06	0.85	0.11
Valid N (listwise)	2016								

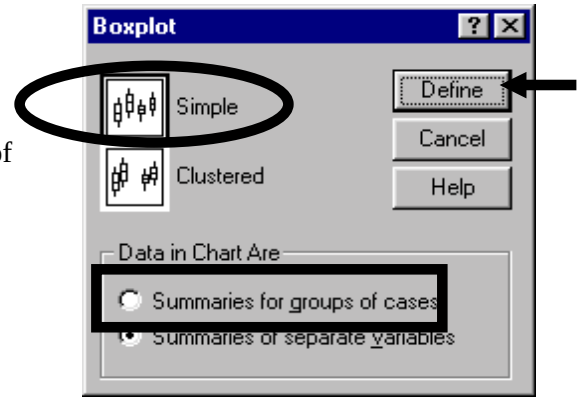
Ch 3. Section 3.b. Boxplots

The spread of the values can be depicted using boxplots. A boxplot chart provides the medians, quartiles, and ranges. It also provides information on outliers.

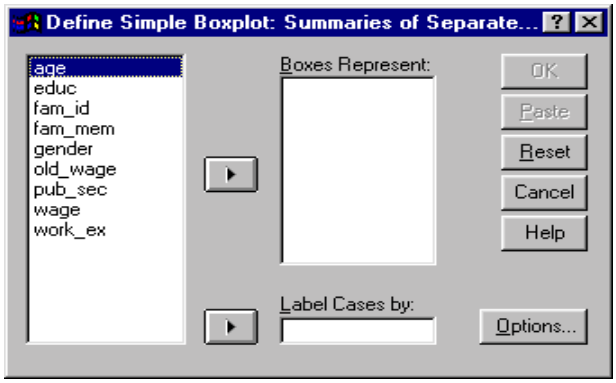
Go to GRAPHS/BOXPLOT.

Choose “Simple” and “Summaries for groups of cases.”

Click on “Define.”

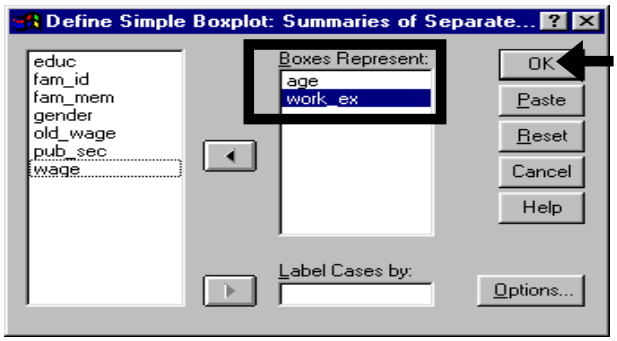


The following dialog box will open up.



Move the variables *age* and *work_ex* into the “Boxes Represent" box.

Click on "OK."



Interpretation:

a-b: lowermost quartile [0-25%]

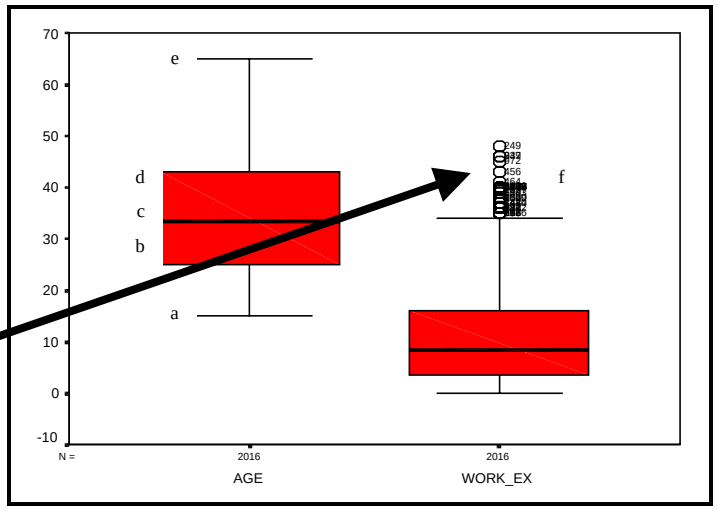
b-c: second lowest quartile [25-50%]

c: mean

c-d: second highest quartile [50-75%]

d-e: highest quartile [75-100%]

The individual cases above the highest quartile are the outliers.



Ch 3. Section 4 Testing if the mean is equal to a hypothesized number (the T-Test and error bar)

After you have obtained the descriptives, you may want to check whether the means you have are similar to the means obtained in:

- another sample on the same population
- a larger survey that covers a much greater proportion of the population

For example, say that mean *education* in a national survey of 100 million people was 6.2. In your sample, the mean is 6.09. Is this statistically similar to the mean from the national survey? If not, then your sample of *education* may not be an accurate representation of the actual distribution of *education* in the population.

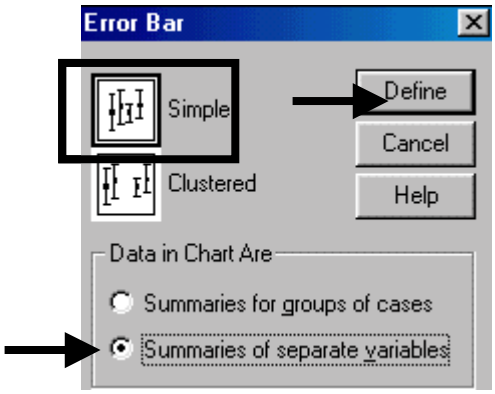
There are two methods in SPSS to find if our estimated mean is statistically indistinct from the hypothesized mean - the formal T-Test and the Error Bar. The number we are testing our mean against is called the hypothesized value. In this example that value is 6.2.

The Error Bar is a graph that shows 95% range within which the mean lies (statistically). If the hypothesized mean is within this range, then we have to conclude that "Our mean is statistically indistinct from the hypothesized number."

Ch 3. Section 4.a. Error Bar (graphically showing the confidence intervals of means)

The Error Bar graphically depicts the 95% confidence band of a variable's mean. Any number within that band may be the mean - we cannot say with 95% confidence that that number is not the mean.

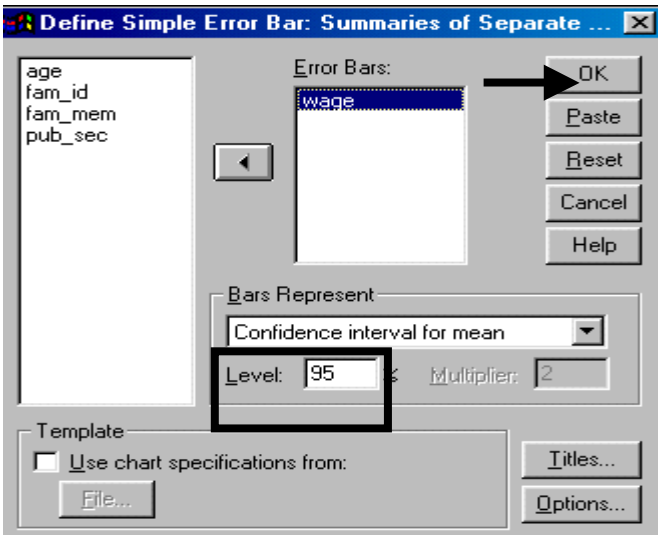
Go to GRAPHS / ERROR BAR. Choose "Simple" type. Select the option "Summaries of separate variables."
Click on "Define."

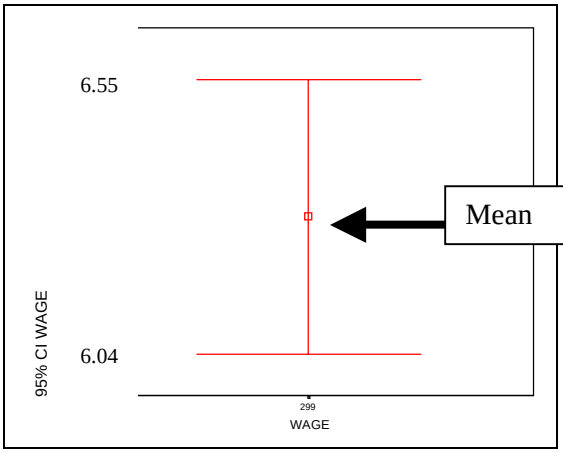


In the box "Error Bars," place the variables whose "Confidence interval for mean" you wish to determine (we are using the variable *wage*)

Choose the confidence level (the default is 95%. You can type in 99% or 90%).

Click on "OK."



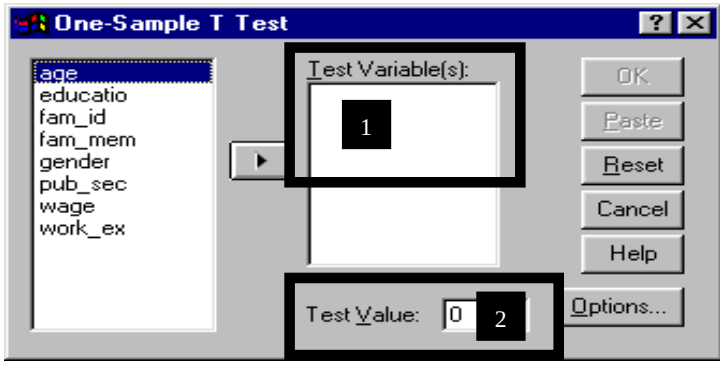


The Error Bar gives the 95% confidence interval for the mean[53]. After looking at the above graph you can conclude that we cannot say with 95% confidence that 6.4 is not the mean (because the number 6.4 lies within the 95% confidence interval).

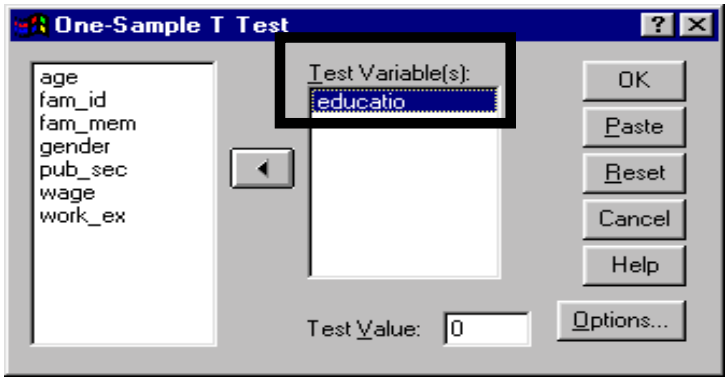
Ch 3. Section 4.a. A formal test: the T-Test

Go to STATISTICS/ MEANS/ ONE-SAMPLE T-TEST.

In area 1 you choose the variable(s) whose mean you wish to compare against the hypothesized mean (the value in area 2).

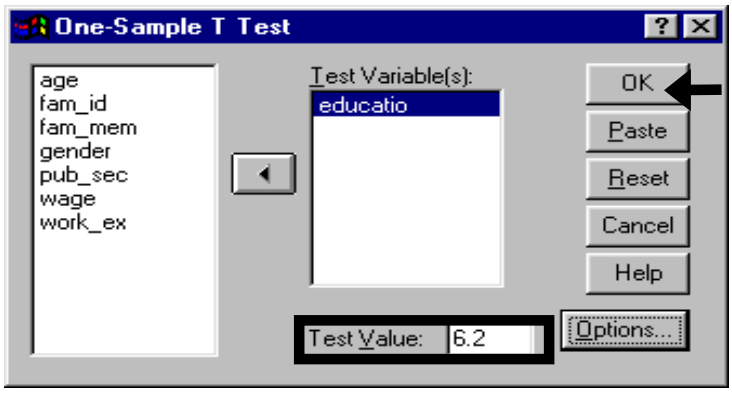


Select the variable *educatio* and put it in the box "Test Variable(s)."



[53] The Error Bar can also be used to depict the 95% confidence interval for the standard deviation (see section 4.3).

In the box "Test Value" enter the hypothesized value of the mean. In our example, the variable is *education* and its test value = 6.2.



SPSS checks whether 6.2 minus the sample mean is significantly different from zero (if so, the sample differs significantly from the hypothesized population distribution).

Click on "OK."

One-Sample Test

	Test Value = 6.2					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
EDUCATION	*-.875*	*2015*	*.382*	*-.11*	*-.35*	*.14*

The test for the difference in sample mean from the hypothesized mean is statistically insignificant (as it is greater than .1) even at the 90% level. We fail to reject the hypothesis that the sample mean does not differ significantly from the hypothesized number[54].

Note: If sig is less than 0.10, then the test is significant at 90% confidence (equivalently, the hypothesis that the means are equal can be rejected at the 90% level of confidence). This criterion is considered too "loose" by some.

If sig is less than 0.05, then the test is significant at 95% confidence (equivalently, the hypothesis that the means are equal can be rejected at the 95% level of confidence). This is the standard criterion used.

If sig is less than 0.01, then the test is significant at 99% confidence (equivalently, the hypothesis that the means are equal can be rejected at the 99% level of confidence). This is the strictest criterion used.

You should memorize these criteria, as nothing is more helpful in interpreting the output from hypothesis tests (including all the tests intrinsic to every regression, ANOVA and other analysis).

Your professors may like to see this stated differently. For example: "Failed to reject null hypothesis at an alpha level of .05." Use the terminology that the boss prefers!

Referring back to the output table above, the last two columns are saying that "with 95% confidence, we can say that the mean is different from the test value of 6.2 by -.35 to .14 - that is, the mean lies in the range '6.2-.35' to '6.2+.14' and we can say this with 95% confidence."

[54]The sample mean of *education* is statistically close to the hypothesized value.

To take quizzes on topics within each chapter go to http://www.spss.org/wwwroot/spssquiz.asp

Ch 4. COMPARING SIMILAR VARIABLES

Sometimes a data set may have variables that are similar in several respects - the variables measure similar entities, the units of measurement are the same, and the scale of the ranges is similar[55].

We debated the justification for a separate chapter on methods that are not used in a typical analysis. For the sake of completeness, and because the topic did not fit seamlessly into any other chapter, we decided to stick with this chapter. The chapter also reinforces some of the skills learned in chapter 3 and introduces some you will learn more about in chapter 5.

If you feel that your project/class does not require the skills taught in this section, you can simply skip to chapter 5.

In section 4.3, we describe how the means (or other statistical attributes) of user-chosen pairs of these variables are compared. For non-normal variables, a non-parametric method is shown.

In the remaining portion of the chapter we show how graphs are used to depict the differences between the attributes of variables. In section 4.2, we describe the use of boxplots in comparing several attributes of the variables - mean, interquartile ranges, and outliers.

Note: You could compare two variables by conducting, on each variable, any of the univariate procedures shown in chapter 3. Chapter four shows procedures that allow for more direct comparison.

Ch 4. Section 1 Graphs (bar, pie)

Let's assume you want to compare the present *wage* with the *old wage* (the wage before a defining event, such as a drop in oil prices). You naturally want to compare the medians of the two variables.

[55] Two examples:

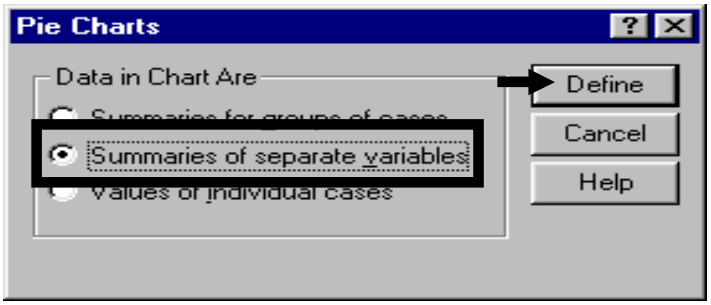
- Twelve variables, one for each month, that have spending in a particular month.
- Six variables, each of which captures the percentage increase in a stock index at a specific stock exchange in a different city (New York, London, Tokyo, Paris, Frankfurt, and Hong Kong).

An interesting analysis would be a comparison of these variables. In the first example, such an analysis can indicate the differences in spending across the months. In the second example, such an analysis can tell us about differences in average price movement in the major stock market indices. This chapter discusses such comparisons.

Go to GRAPHS/ PIE.

Note: You can use a bar graph instead.

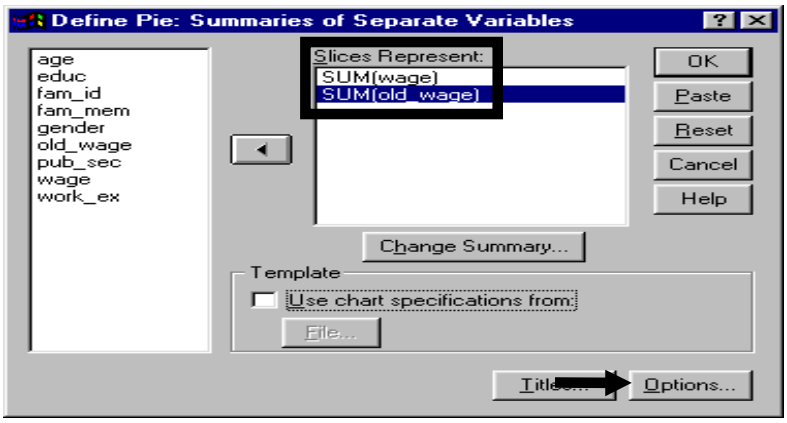
Select "Summaries of separate variables."



Click on "Define."

Move the two variables into the box "Slices Represent."

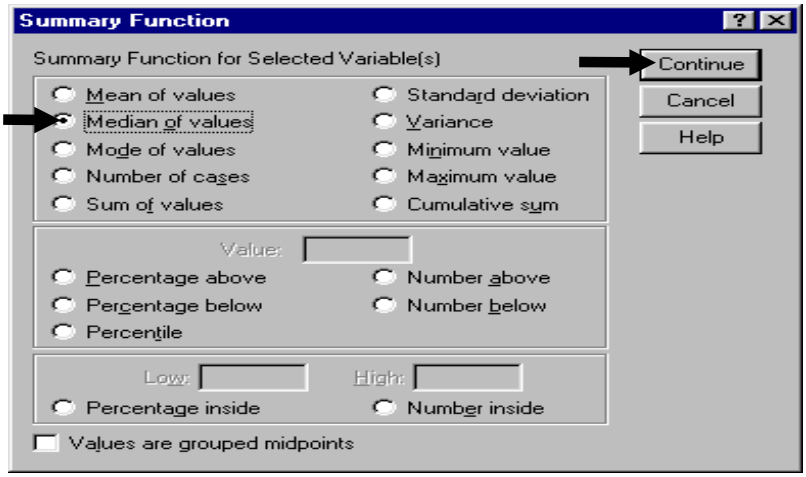
By default, the statistic used last time (in this case, "Sum") is assigned to them. Remember that you want to use the medians. To do so, click on the button "Options."



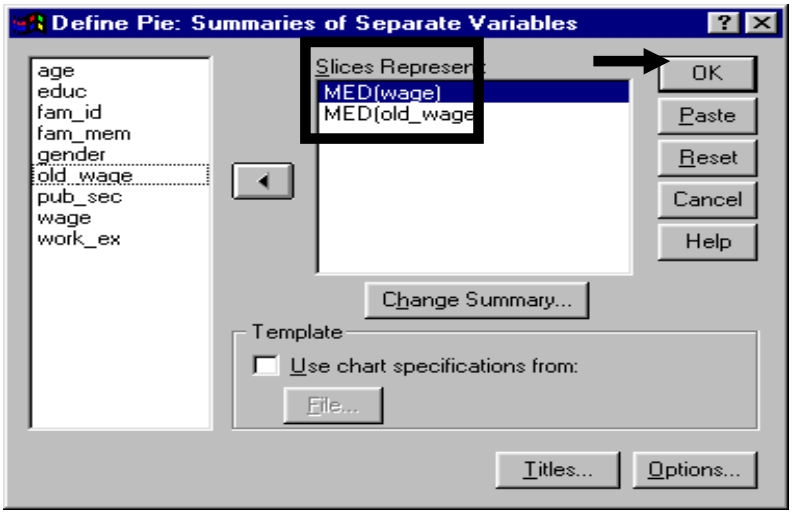
Select the option "Median of values."

Note: In all of the graphical procedures (bar, area, line, and pie), the option "Summary Function" provides the same list of functions.

Click on "Continue."

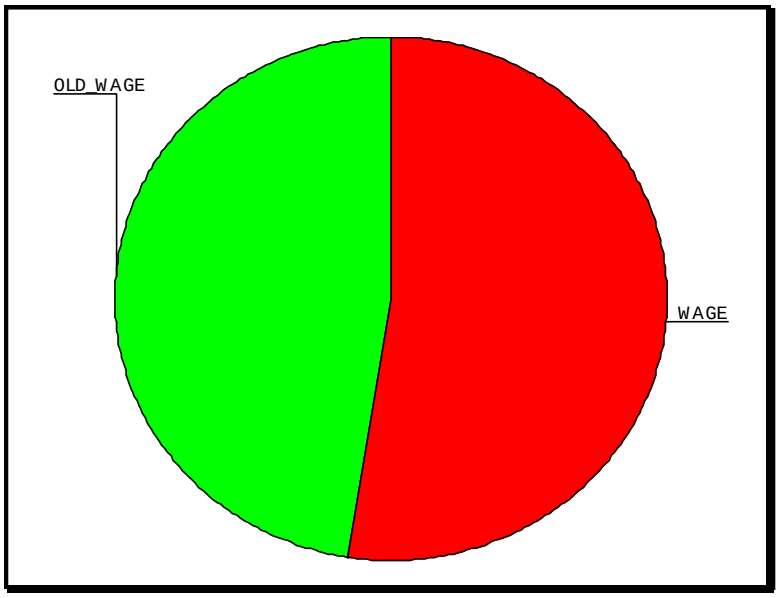


The functions change to median.



Click on "OK."

This method can compare several variables at the same time, with each "slice" of the pie representing one variable.



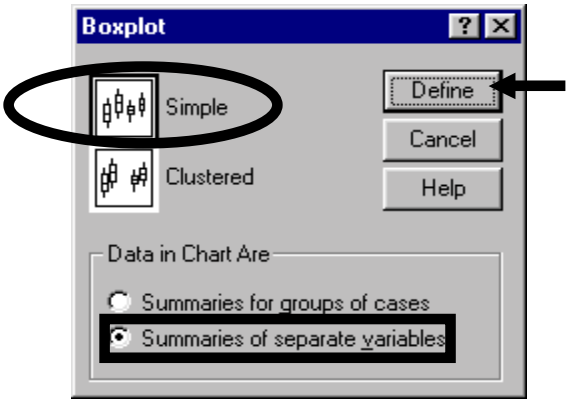
Interpretation: the median of *wage* is higher than that of *old_wage*.

Ch 4. Section 2 Boxplots

The spread of the values of two similar variables can be compared using boxplots. Let's assume that you want to compare *age* and *work experience*. A boxplot chart compares the medians, quartiles, and ranges of the two variables[56]. It also provides information on outliers.

[56] In separate boxplots on the same chart. As a reminder: the first quartile defines the value below which 25% of the variables' values lie, the second quartile (also called the median or mid-point) defines the value below which 50% of the variables' values lie, the third quartile defines the value below which 75% of the variables' values lie, and the interquartile range is defined by the values of the third and first quartiles.

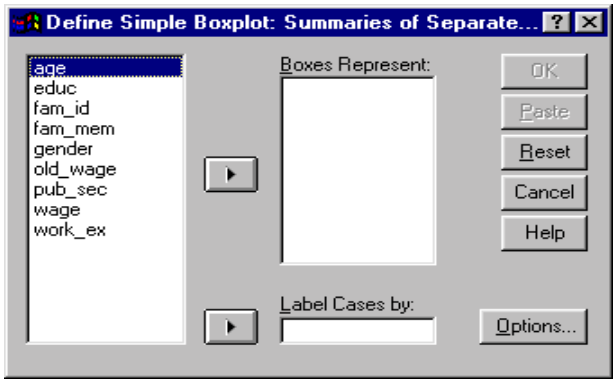
Go to GRAPHS/BOXPLOT.



Choose "Simple" and "Summaries of Separate Variables."

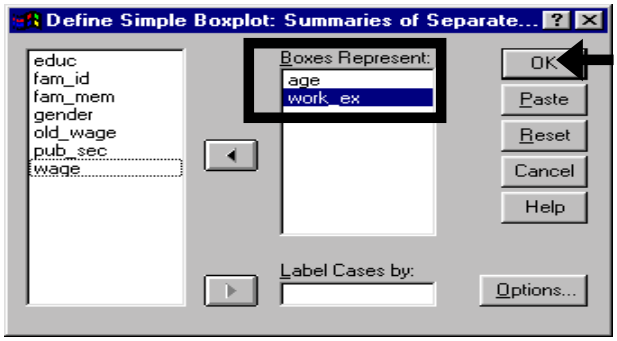
Click on "Define."

The following dialog box will open.



Move the variables *age* and *work_ex* into the box "Boxes Represent."

Click on "OK."



Interpretation:

a-b: lowermost quartile (0-25%)

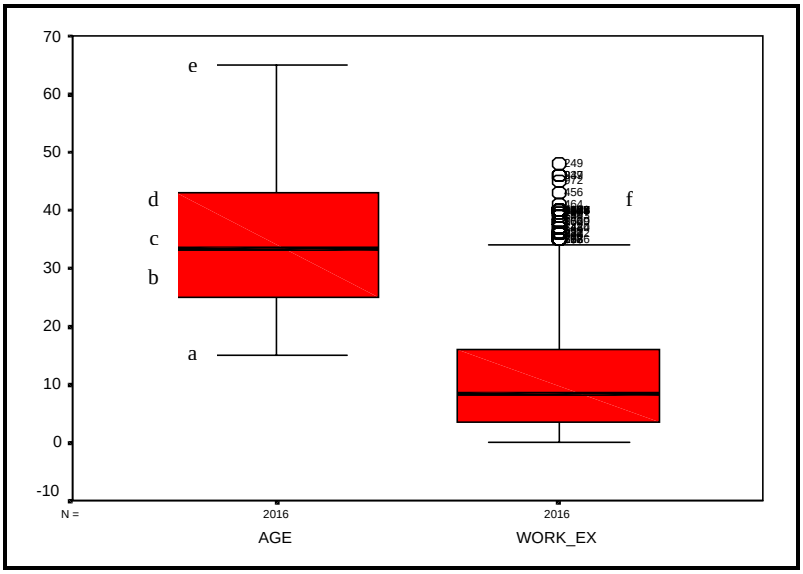
b-c: second lowest quartile (25-50%)

c: mean

c-d: second highest quartile (50-75%)

d-e: highest quartile (75-100%)

The individual cases above the highest quartile are the outliers.



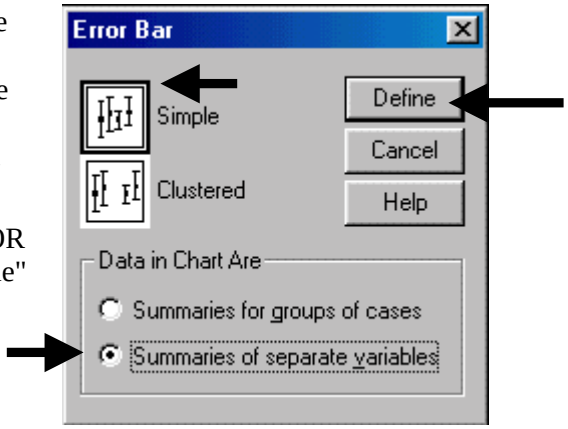
Ch 4. Section 3 Comparing means and distributions

Ch 4. Section 3.a. Error Bars

Error bars graphically depict differences in the confidence intervals of key statistics of the distributions of variables (note: use only if the variables are distributed normally).

Let's assume you want to compare aspects of the distribution of the current wage (*wage*) and the wage before (let us further assume) the company was bought (*old_wage*).

To do so, go to GRAPHS / ERROR BAR. Choose the options "Simple" and "Summaries of separate variables." Press "Define."

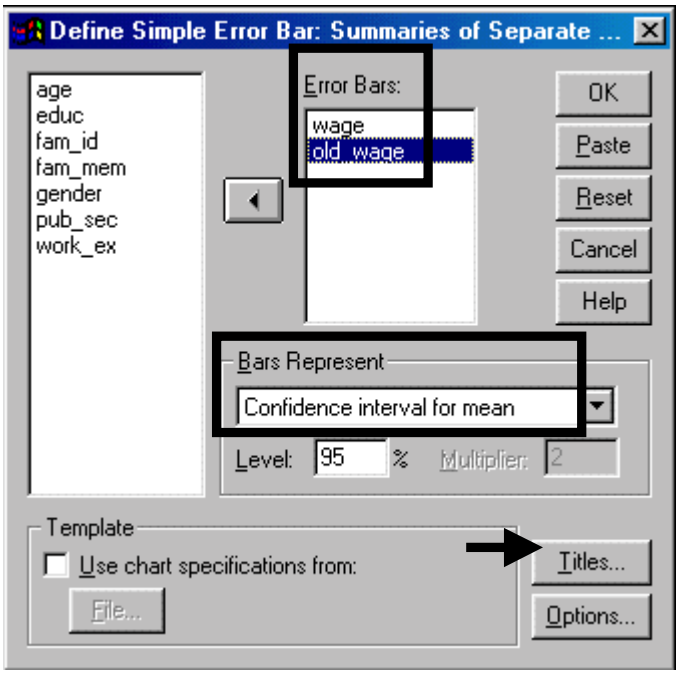


In the area "Error Bars," place the variables whose means you wish to compare.

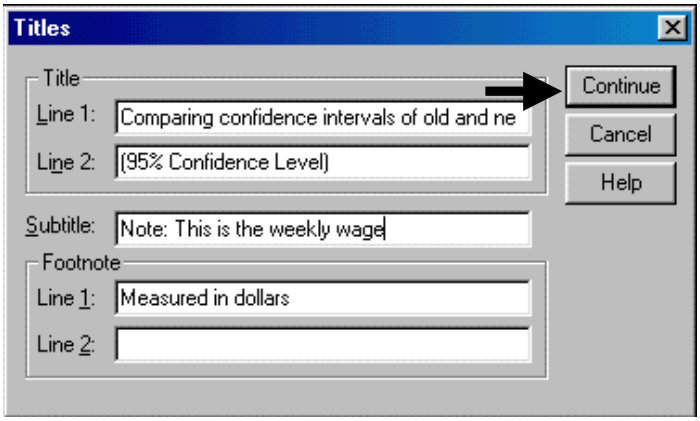
In the area "Bars Represent," choose the statistic whose confidence interval you want the error bars to depict. You will typically choose the "Confidence interval of mean" (below, we show examples of other statistics).

Choose the confidence level you want the Error Bars to depict. The default is 95%. We would advise choosing that level.

Click on "Titles."

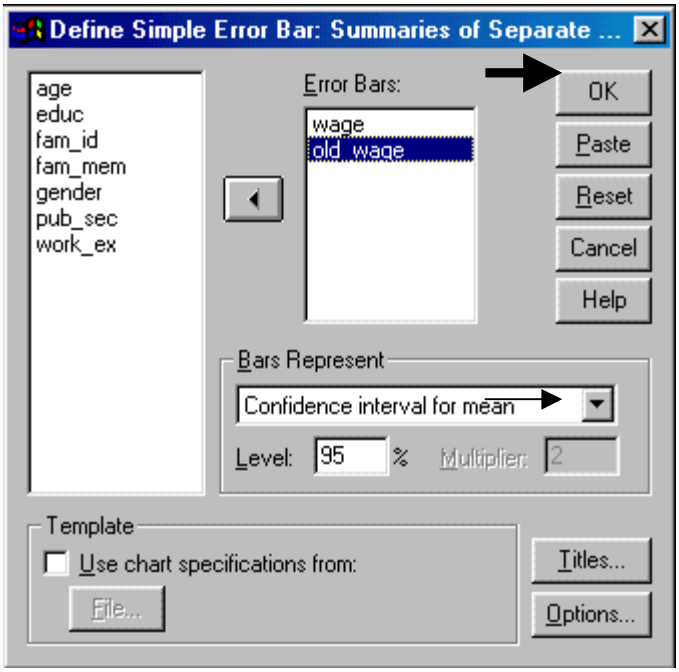


Enter a descriptive title, subtitle, and footnote.

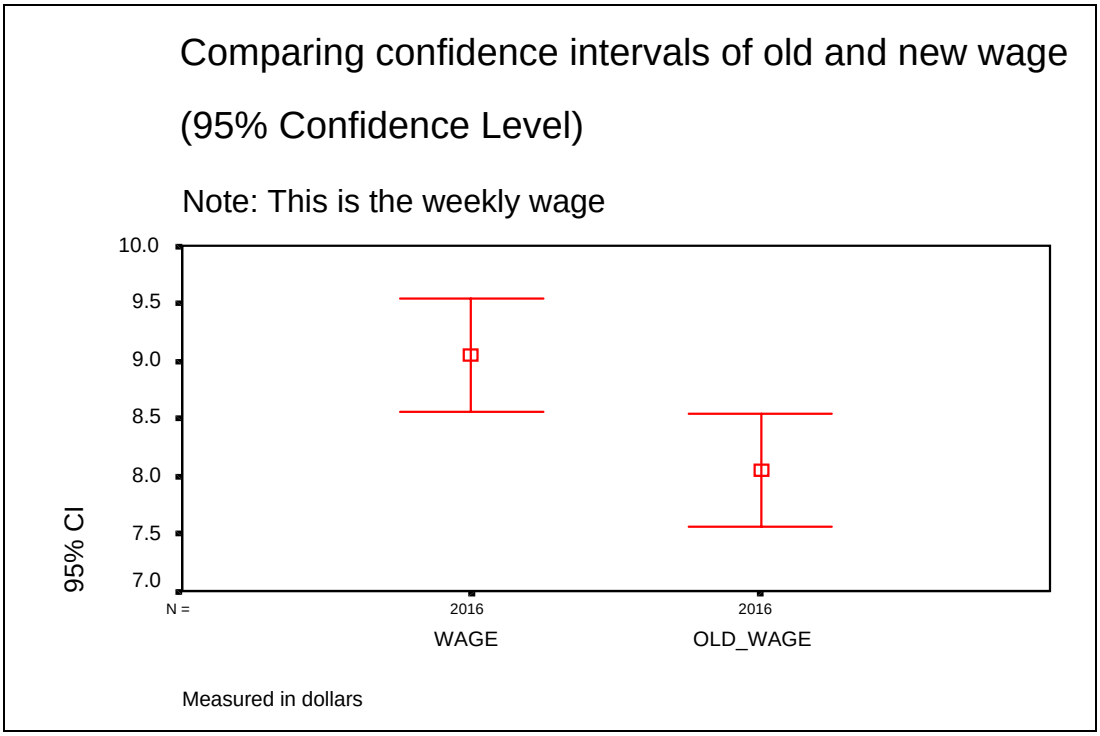


Note: Many SPSS procedures include the "Titles" option. To conserve space, we may skip this step for some of the procedures. We do, however, advise you to always use the "Titles" option.

Click on "OK."

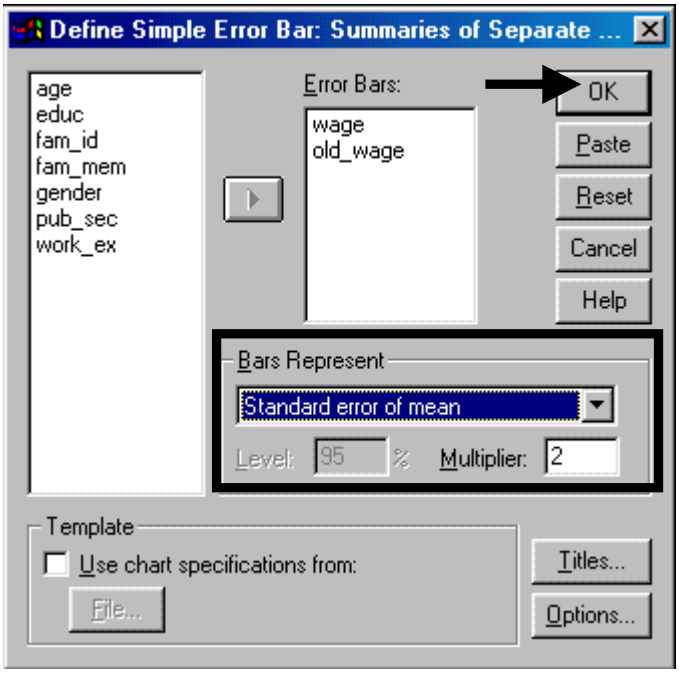


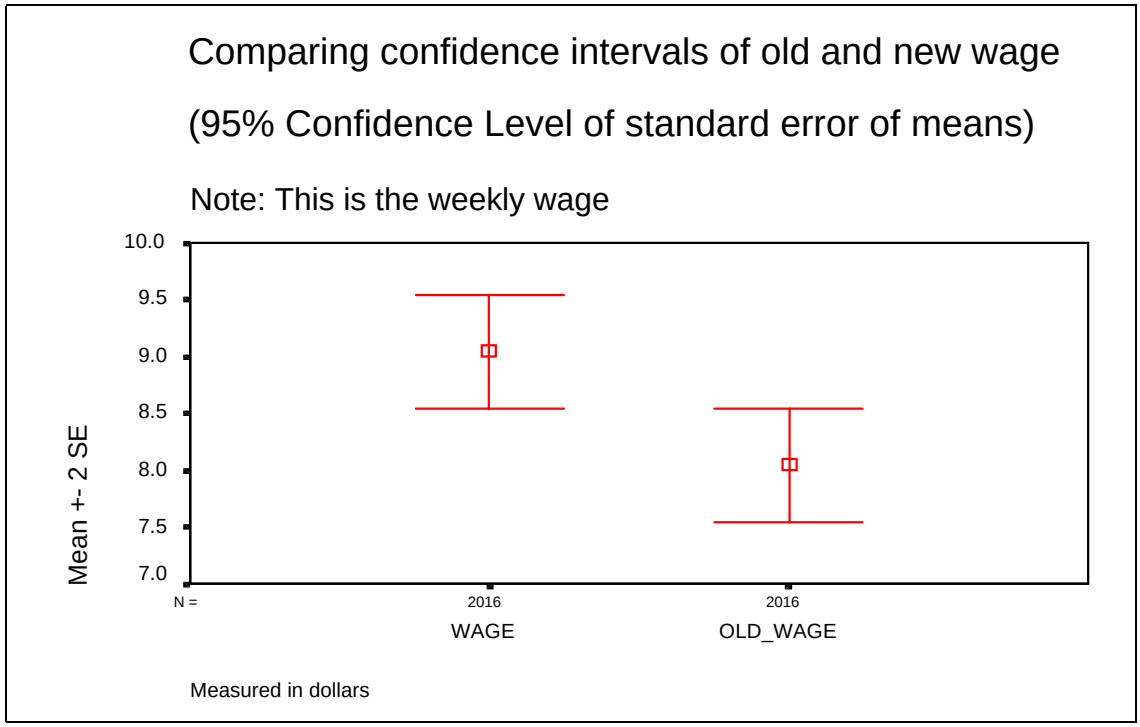
The output is shown below. Each "Error Bar" defines the range within which we can say, with 95% confidence, the mean lies. Another interpretation - we cannot reject the hypothesis that any number within the range may be the real mean. For example, though the estimated mean of wage is $9 (see the small box in the middle of the Error Bar), any value within 8.5 and 9.5 may be the mean. In essence, we are admitting to the fact that our estimate of the mean is subject to qualifications imposed by variability. The confidence interval incorporates both pieces of information - the estimate of the mean and its standard error.



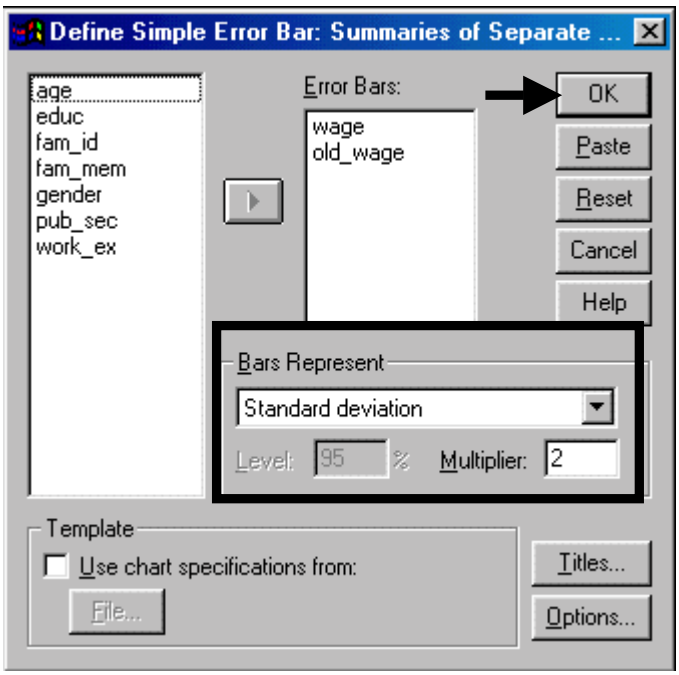
We now show an example of using Error Bars to depict the "Standard Error of mean." To do so, repeat all the steps from above except for choosing the option "Standard error of mean" (and typing in "2" to ask for "+/- 2 standard errors") in the area "Bars Represent."

The output is shown below. The graph looks the same as for the 95% confidence interval of the mean. Reason? The 95% interval is "mean + 2 standard errors," the 90% interval is "mean + 1 standard error," etc.



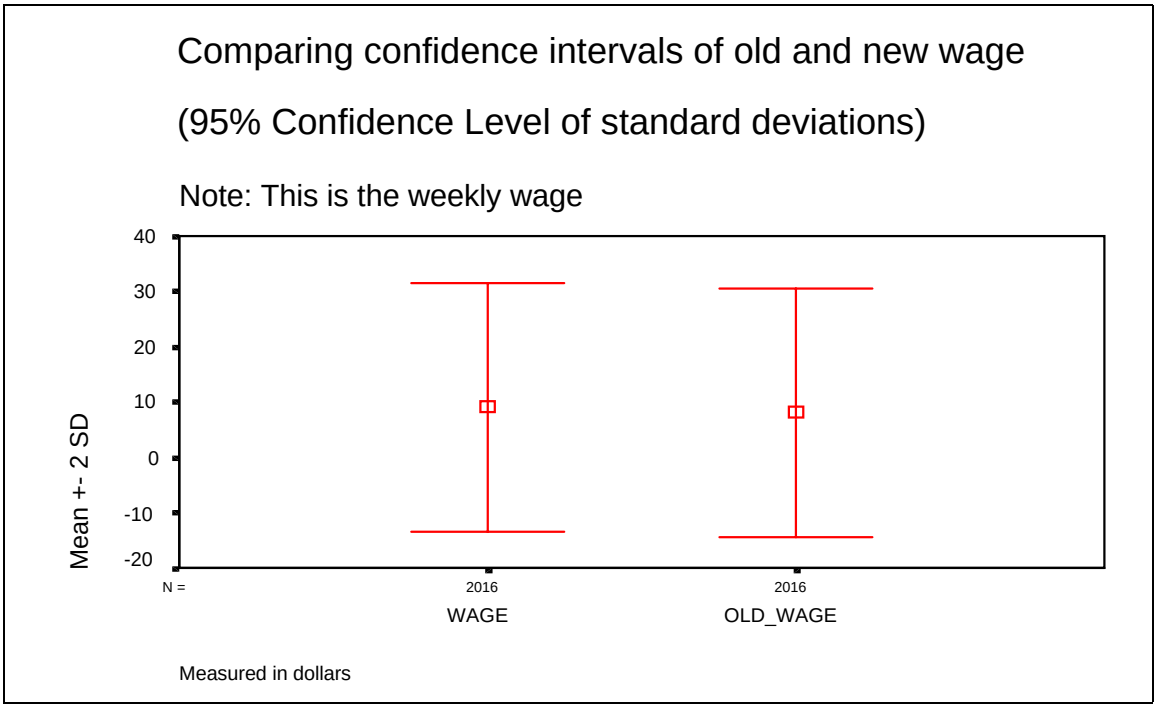


Another statistic the Error Bar can show is the "Standard deviation of the variable." To view this statistic, repeat all the steps from above except for choosing the option "Standard deviation" (and typing in "2" to ask for "+/- 2 standard errors") in the area "Bars Represent."



The output is shown below. Each "Error Bar" defines the range within which we can say, with 95% confidence, the standard deviation lies. Another interpretation: we cannot reject the hypothesis that any number within the range may be the real standard deviation. For example, though the estimated standard deviation of wage is $10 (see the small box in the middle of the Error Bar), any value within -8 and 32 may be the standard deviation. In essence, we are admitting to the fact that our estimate of the standard deviation is subject to qualifications imposed by variability. The confidence interval incorporates both pieces of information - the estimate of the standard deviation and its standard

error.



Ch 4. Section 3.b. The paired samples T-Test

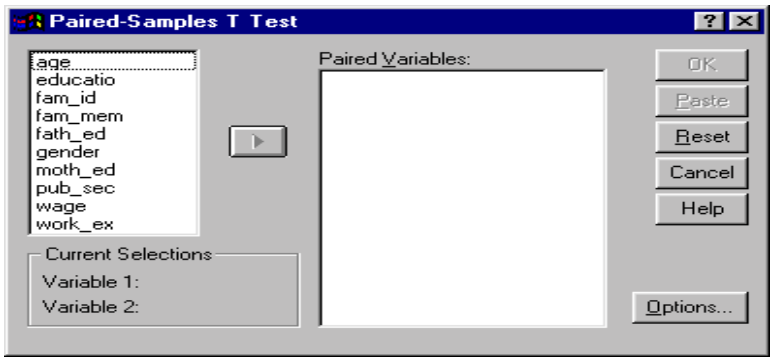
Let's assume you have three variables with which to work - education (respondent's education), *moth_ed* (mother's education), and *fath_ed* (father's education). You want to check if:

- The mean of the respondent's education is the same as that of the respondent's mother's
- The mean of the respondent's education is the same as that of the respondent's father's

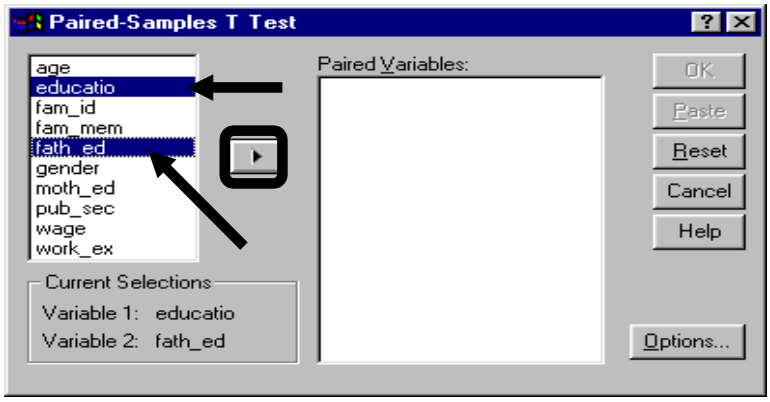
Using methods shown in sections 3.2.a and 3.3, you could obtain the means for all the above variables. A straightforward comparison could then be made. Or, can it? "Is it possible that our estimates are not really perfectly accurate?"

The answer is that our estimates are definitely not perfectly accurate. We must use methods for comparing means that incorporate the use of the mean's dispersion. The T-Test is such a method.

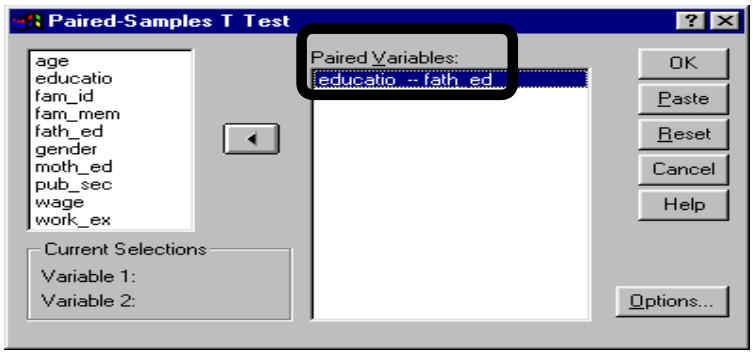
Go to STATISTICS/ MEANS/ PAIRED SAMPLES T-TEST.



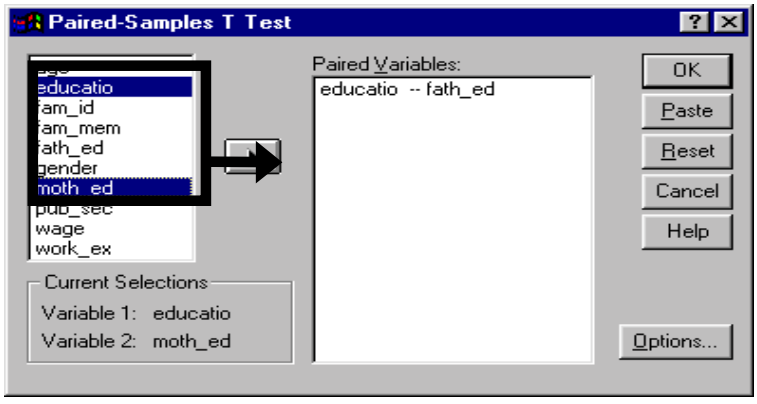
In the dialog box, choose the pair *educatio* and *fath_ed*. To do this, click on the variable *educatio* first. Then press the CTRL key on the keyboard and, keeping CTRL pressed, click on the variable *fath_ed*.



You have now successfully selected the first pair of variables[57].



To select the second pair, repeat the steps - click on the variable *educatio* first[58]. Then press the CTRL key on the keyboard and, keeping CTRL pressed, click on the variable *moth_ed*. Click on the selection button.

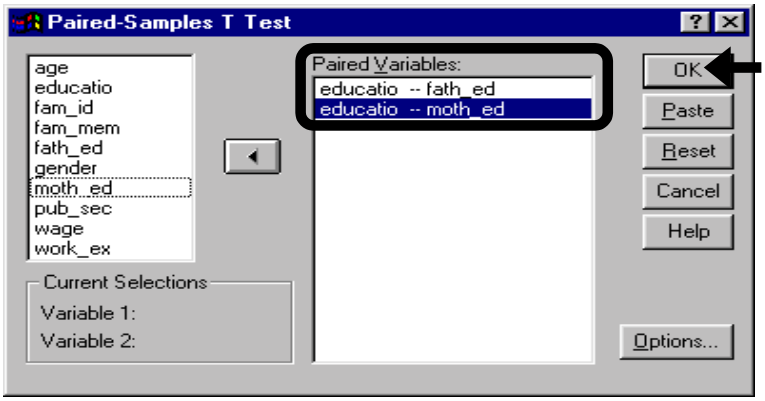


[57] Note that the first "Paired Variable" is defined as the difference between *educatio* and *fath_ed*, i.e. - "Respondents' education level MINUS Father's education level."

[58] We are choosing the variable *educatio* first so that the two pairs are symmetric.

You have now selected both pairs of variables[59].

Click on "OK."



The first output table shows the correlations within each pair of variables.

See section 5.3 for more on how to interpret correlation output.

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	EDUCATION & Father's Education Level	*2016*	*-.055*	*.014*
Pair 2	EDUCATION & Mother's Education Level	*2016*	*-.057*	*.010*

The next table gives the results of the tests that determine whether the difference between the means of the variables (in each pair) equals zero.

Paired Samples Test

			Pair 1: EDUCATION - Father's Education Level	Pair 2: EDUCATION - Mother's Education Level
Paired Differences	Mean		*-4.7260*	*3.5640*
	Std. Deviation		*11.0082*	*6.1380*
	Std. Error Mean		*.2452*	*.1367*
	95% Confidence Interval of the Difference	Lower	*-5.2068*	*3.2959*
		Upper	*-4.2452*	*3.8321*
t			*-19.276*	*26.071*
df			*2015*	*2015*
Sig. (2-tailed)			*.000*	*.000*

[59] Note that the two new variables are the differences between the variables in the pair. SPSS creates (only in its memory - no new variable is created on the data sheet) and uses two new variables:

- *Educatio* minus *fath_ed*
- *Educatio* minus *moth_ed*

The procedure is determining whether the means of these variables equal zero. If they do, then the paired variables have the same mean.

Both the pairs are significant (as the sig value is below 0.05)[60]. This is telling us:

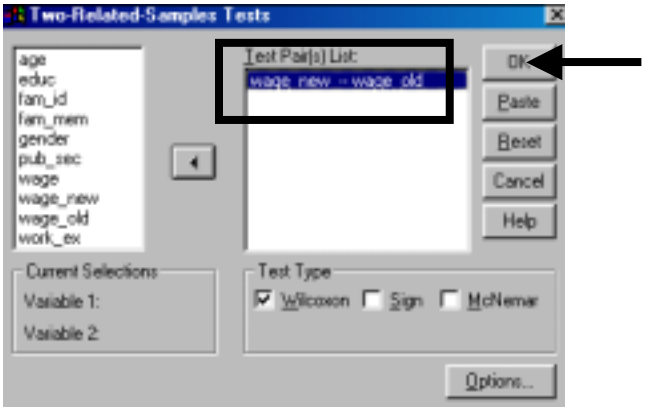
- The mean of the variable *father's education* is significantly different from that of the respondents. The negative Mean (-4.7) is signifying that the mean education of fathers is higher.
- The mean of the variable *mother's education* is significantly different from that of the respondents. The positive Mean (3.5) is signifying that the mean education of mothers is lower.

Ch 4. Section 3.c. Comparing distributions when normality cannot be assumed - 2 related samples non-parametric test

As we mentioned in section 3.2.e , the use of the T and F tests hinges on the assumption of normality of underlying distributions of the variables. Strictly speaking, one should not use those testing methods if a variable has been shown not to be normally distributed (see section 3.2). Instead, non-parametric methods should be used-- these methods do not make any assumptions about the underlying distribution types.

Let's assume you want to compare two variables: *old_wage* and *new_wage*. You want to know if the distribution of the *new_wage* differs appreciably from that of the old wage. You want to use the non-parametric method – "Two Related Samples Tests."

Go to "STATISTICS / NONPARAMETRIC / 2 RELATED SAMPLES TESTS." Choose the pair of variables whose distributions you wish to compare. To do this, click on the first variable name, press the CTRL key, and (keeping the CTRL key pressed) click on the second variable. Click on the middle arrow - this will move the pair over into the box "Test Pair(s) List" (note: You can also add other pairs).



Choose the "Wilcoxon" test in the area "Test

[60] The basic rules for interpreting the significance values should be firmly implanted in your memory. The rules, which are common for the interpretation of any significance test irrespective of test type (the most frequently used types are the T, F, Chi-Square, and Z tests) and context (as you will see in later chapters, the context may be regression, correlation, ANOVA, etc.), are:

- If the value in the significance table is less than .01, then the estimated coefficient can be believed with 99% confidence
- If the value in the significance table is less than .05, then the estimated coefficient can be believed with 95% confidence
- If the value in the significance table is less than .1, then the estimated coefficient can be believed with 90% confidence
- If the value in the significance table is greater than .1, then the estimated coefficient is not statistically significant, implying that the estimate should not be relied upon as reliable or accurate

Type." If the variables are dichotomous variables, then choose the McNemar test.

Click on "OK."

Ranks

		N	Mean Rank	Sum of Ranks
Wage_new - Wage_old	Negative Ranks	671[a]	788.45	529047.50
	Positive Ranks	1272[b]	1068.83	1359549
	Ties	73[c]		
	Total	2016		

a. Wage_new < Wage_old
b. Wage_new > Wage_old
c. Wage_new = Wage_old

The low Sig value indicates that the null hypothesis, that the two variables have similar distributions, can be rejected. Conclusion: the two variables have different distributions.

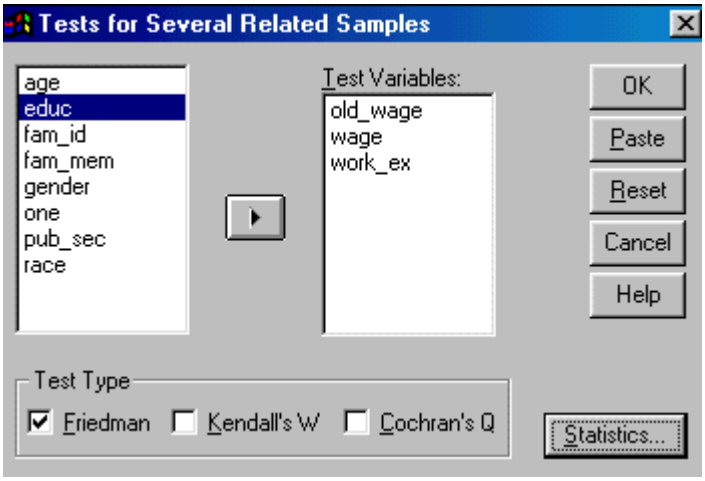
Test Statistics[b]

	Wage_new - Wage_old
Z	-16.792[a]
Asymp. Sig. (2-tailed)	.000

a. Based on negative ranks.
b. Wilcoxon Signed Ranks Test

If you want to compare more than two variables simultaneously, then use the option STATISTICS / NONPARAMETRIC / K RELATED SAMPLES TESTS. Follow the same procedures as shown above but with one exception:

- Choose the "Friedman" test in the area "Test Type." If all the variables being tested are dichotomous variables, then choose the "Cochran's Q" test.



We cannot make the more powerful statement that "the means are equal/unequal" (as we could with the T Test). You may see this as a trade-off: "The non-parametric test is more appropriate when the normality assumption does not hold, but the test does not produce output as rich as a parametric T test."

To take quizzes on topics within each chapter go to http://www.spss.org/wwwroot/spssquiz.asp

Ch 5. MULTIVARIATE STATISTICS

After performing univariate analysis (chapter 3) the next essential step is to understand the basic relationship between/across variables. For example, to "Find whether *education* levels are different for categories of the variable *gender* (i.e. - "male" and "female") and for levels of the categorical variable *age*."

Section 5.1 uses graphical procedures to analyze the statistical attributes of one variable categorized by the values/categories of another (or more than one) categorical or dummy variable. The power of these graphical procedures is the flexibility they offer: you can compare a wide variety of statistical attributes, some of which you can custom design. Section 5.1.c shows some examples of such graphs.

Section 5.2 demonstrates the construction and use of scatter plots.

In section 5.3, we explain the meaning of correlations and then describe how to conduct and interpret two types of correlation analysis: bivariate and partial. Correlations give one number (on a uniform and comparable scale of -1 to 1) that captures the relationship between two variables.

In section 5.3, you will be introduced to the term "coefficient." A very rough intuitive definition of this term is "an estimated parameter that captures the relationship between two variables." Most econometrics projects are ultimately concerned with obtaining the estimates of these coefficients. But please be careful not to become "coefficient-obsessed." The reasoning will become clear when you read chapters 7 and 8. Whatever estimates you obtain must be placed within the context of the reliability of the estimation process (captured by the "Sig" or "Significance" value of an appropriate "reliability-testing" distribution like the T or F[61]).

SPSS has an extremely powerful procedure (EXPLORE) that can perform most of the above procedures together, thereby saving time and effort. Section 5.4 describes how to use this procedure and illustrates the exhaustive output produced.

Section 5.5 teaches comparison of means/distributions using error bars, T-Tests, Analysis of Variance, and nonparametric testing.

[61] If you think of the hypothesis testing distributions as "reliability-testing," then you will obtain a very clear idea of the rationales behind the use of these distributions and their significance values.

Ch 5. Section 1 Graphs

Ch 5. Section 1.a. Graphing a statistic (e.g. - the mean) of variable "Y" by categories of X

One must often discover how the values of one variable are affected by the values of another variable. Does the mean of X increase as Y increases? And what happens to the standard deviation of X as Y increases? Bar graphs show this elegantly. Bar graphs are excellent and flexible tools for depicting the patterns in a variable across the categories of up to two other dummy or categorical variables[62].

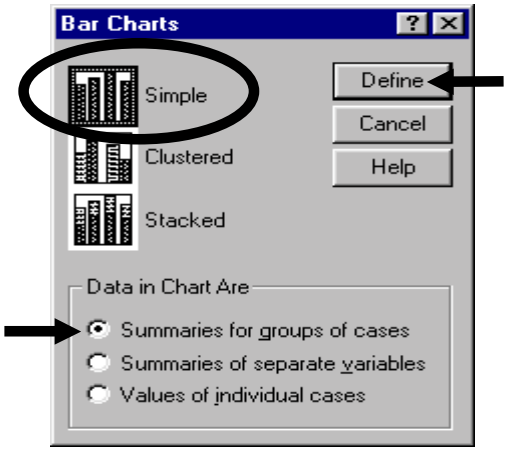
Note: Aside from the visual/graphical indicators used to plot the graph, the bar, line, area, and (for univariate graphs) pie graphs are very similar. The graph type you choose must be capable of showing the point for which you are using the graph (in your report/thesis). A bar graph typically is better when the X-axis variable takes on a few values only, whereas a line graph is better when the X-axis variable can take on one of several values and/or the graph has a third dimension (that is, multiple lines). An area graph is used instead of a line graph when the value on the Y-axis is of an aggregate nature (or if you feel that area graphs look better than line graphs), and a pie graph is preferable when the number of "slices" of the pie is small. **The dialog boxes for these graph types (especially bar, line, and area) are very similar.** Any example we show with one graph type can also be applied using any of the other graph types.

Example 1: Bar graph for means

Select GRAPHS/BAR.

Select "Simple" and "Summaries of Groups of Cases."

Click on the button "Define."



[62] If you have a category variable with numerous categories and/or if you want to compare the cases of two or more variables, then a line or area graph is better. This section includes examples of area and line graphs.

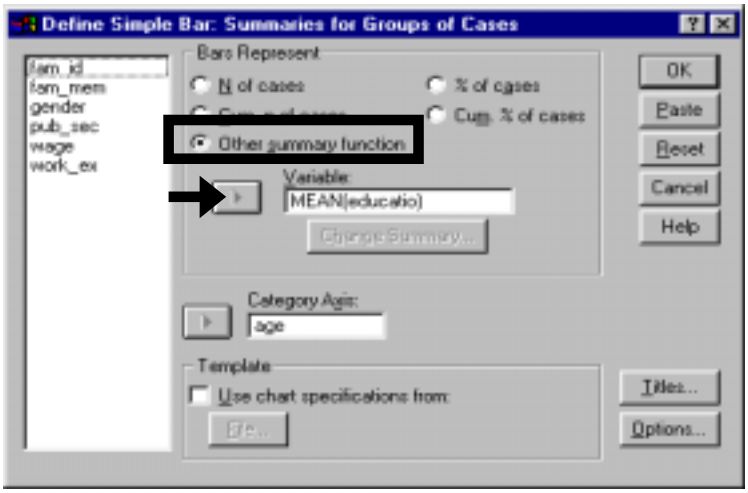
Select the variable *age*. Place it into the "Category Axis" box.

This defines the X-axis.

Select the variable *education* and move it over into the "Variable" box by clicking on the uppermost rightward-pointing arrow.

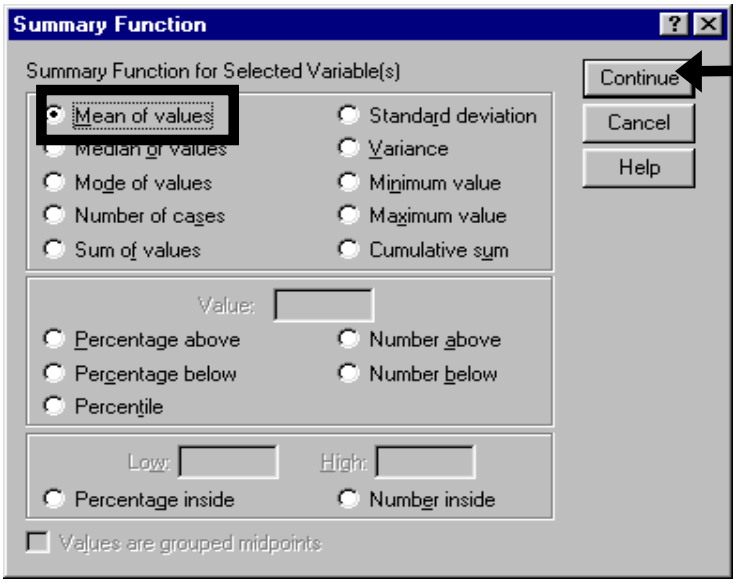
Select the option "Other Summary Function."

Press the button "Change Summary."



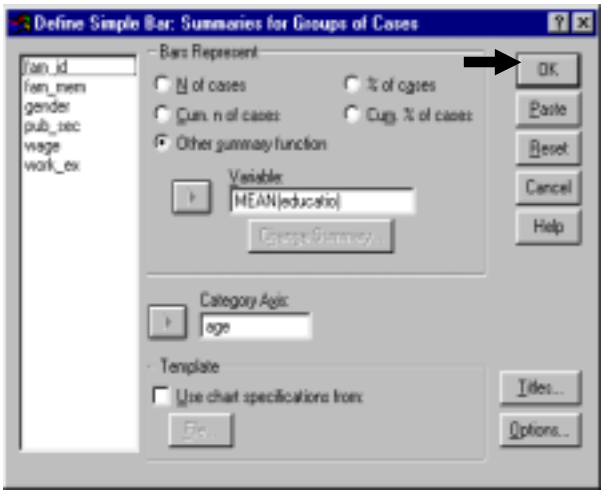
Select the summary statistic you want (in this case "Mean of Values"), and then press "Continue."

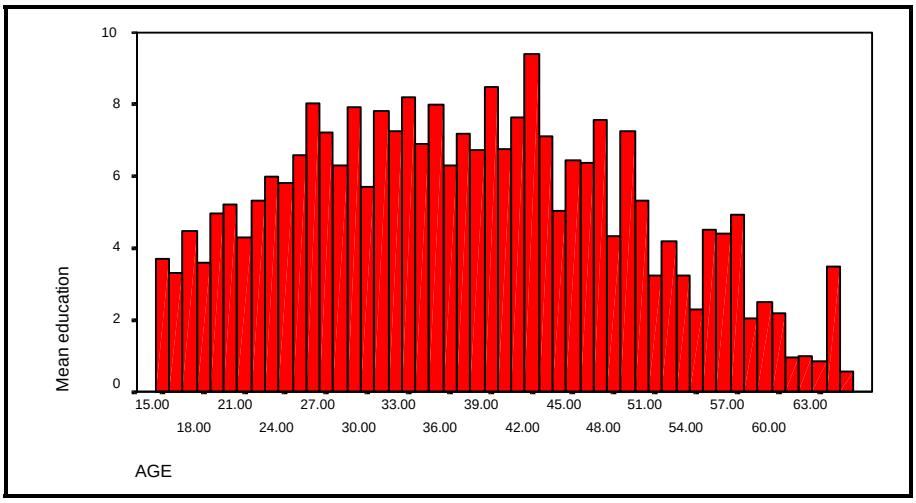
The "Summary Statistic" defines the attributes depicted by the bars in the bar graph (or the line, area, or slice in the respective graph type) and, consequently, the scale and units of the Y-axis.



Press "OK"

The graph produced is shown below. The X-axis contains the categories or levels of age. The Y-axis shows the mean education level for each age category/level.





In the above bar graph, each bar gives the mean of the *education* level for each *age* (from 15 to 65). The mean *education* is highest in the *age* group 25-50.

Example 2: Bar graph for standard deviations

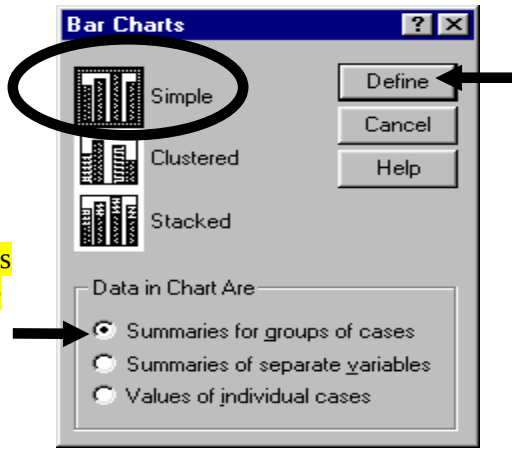
Let's assume that you want to know whether the deviations of the *education* levels around the mean are different across *age* levels? Do the lower educational levels for 15- and 64-year-olds imply a similar dispersion of individual *education* levels for people of those *age* groups? To answer this, we must see a graph of the standard deviations of the *education* variable, separated by *age*.

Select GRAPHS/BAR.

Select "Simple" and "Summaries of Groups of Cases."

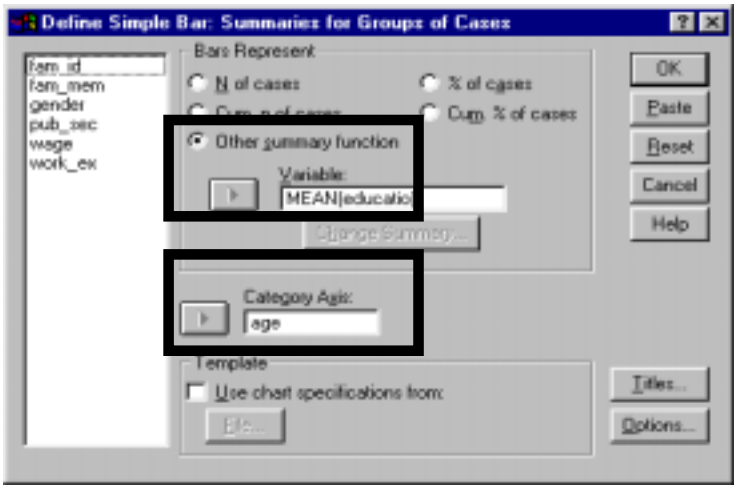
Click on the button "Define."

Note: In this example, we repeat some of the steps that were explained in the previous example. We apologize for the repetition, but we feel that such repetition is necessary to ensure that the reader becomes familiar with the dialog boxes.

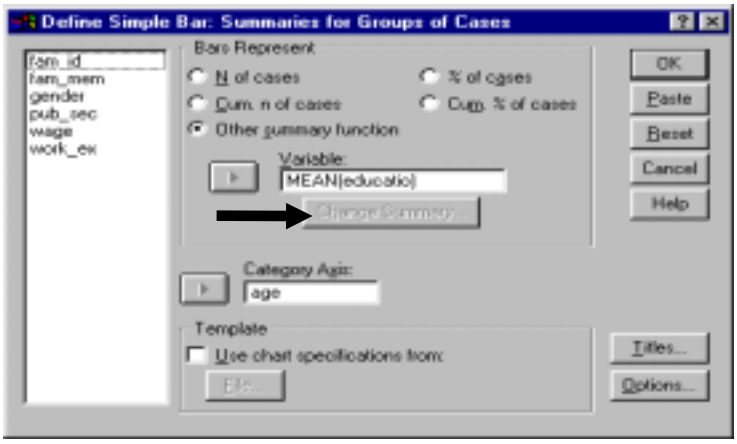


Select the variable *age*. Place it into the "Category Axis" box.

Select the variable *education* and move it over into the "Variable" box by clicking on the uppermost rightward-pointing arrow.

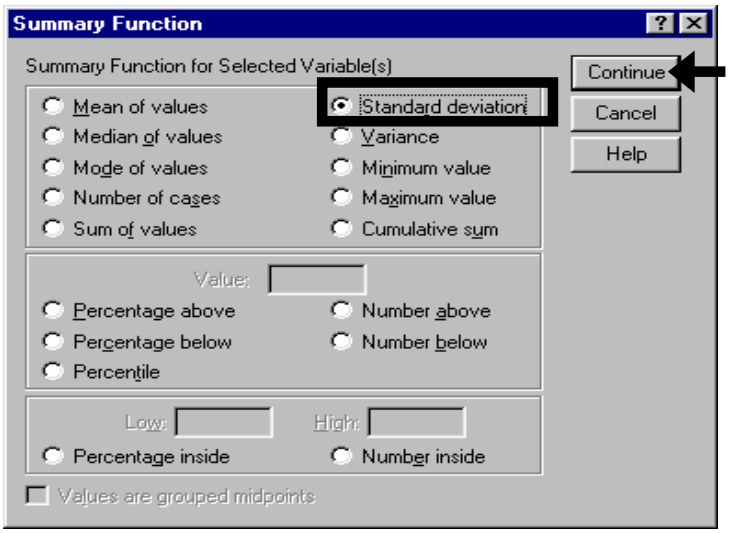


This example requires the statistic "Standard Deviation." The dialog box still shows the statistic "Mean." To change the statistic, press the button "Change Summary."



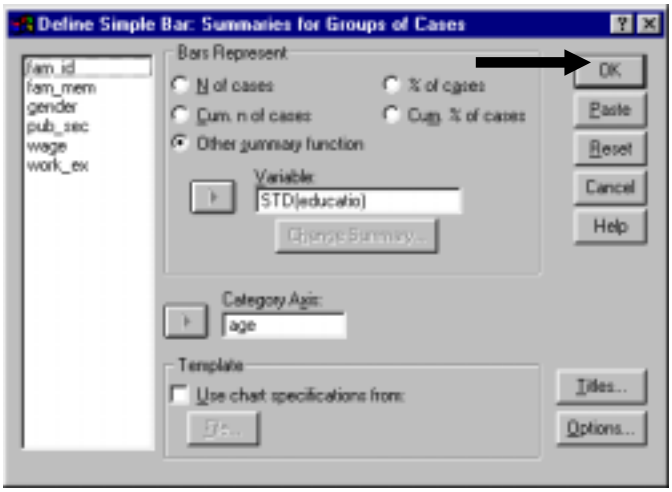
Select the summary statistic you want (in this case "Standard Deviation").

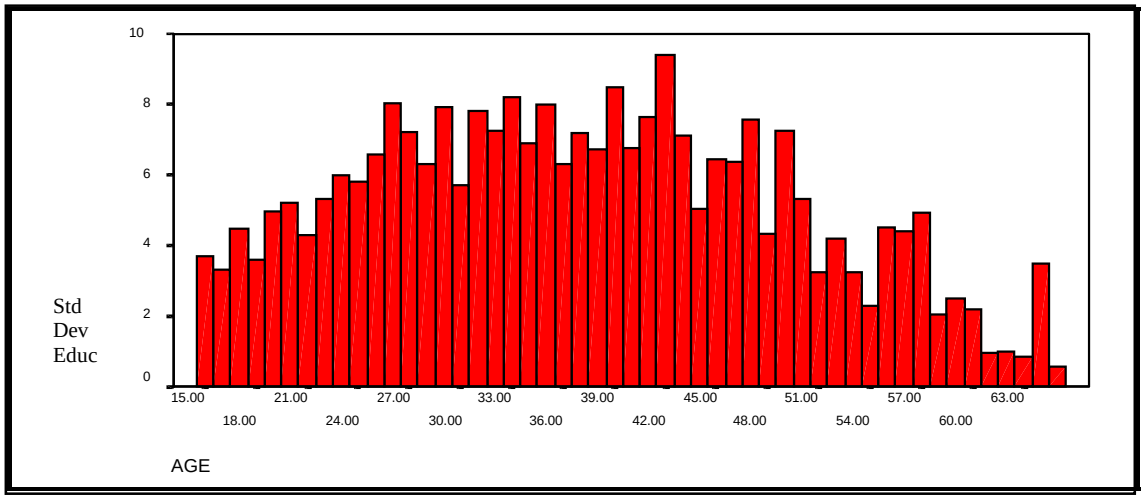
Click on "Continue."



Click on "OK."

Note: the resulting graph is similar to the previous graph, but there is one crucial difference - in this graph, the Y-axis (and therefore the bar heights) represents the standard deviation of education for each value of age.





Ch 5. Section 1.b. Graphing a statistic (e.g. - the mean) of variable "Y" by categories of "X" and "Z"

We can refer to these graphs as 3-dimensional, where dimension 1 is the X-axis, dimension 2 is each line, and dimension 3 is the Y-axis. A line or area graph chart is more appropriate than a bar graph if the category variable has several options.

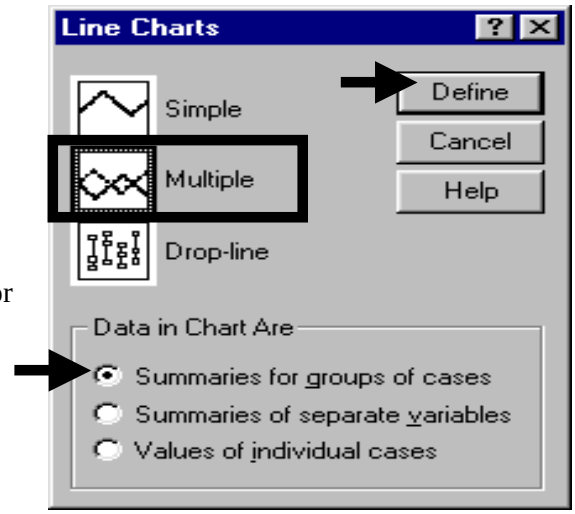
Note: Aside from the visual/graphical indicators used to plot each graph, the bar, line, area, and (for univariate graphs) pie graphs are very similar. The graph type you choose must be capable of showing the point you are using the graph for (in your report/thesis). A bar graph is typically better when the X-axis variable takes on only a few values, whereas a line graph is better when the X-axis variable can take on one of several values and/or the graph has a third dimension (that is, multiple lines). An area graph is used instead of a line graph when the value on the Y-axis is of an aggregate nature (or you feel that area graphs look better than line graphs), and a pie graph is preferable when the number of "slices" of the pie is small. The dialog boxes for these graph types (especially bar, line, and area) are very similar. Any example we show with one graph type can also be applied using any of the other graph types.

Example 1: Line graph for comparing median

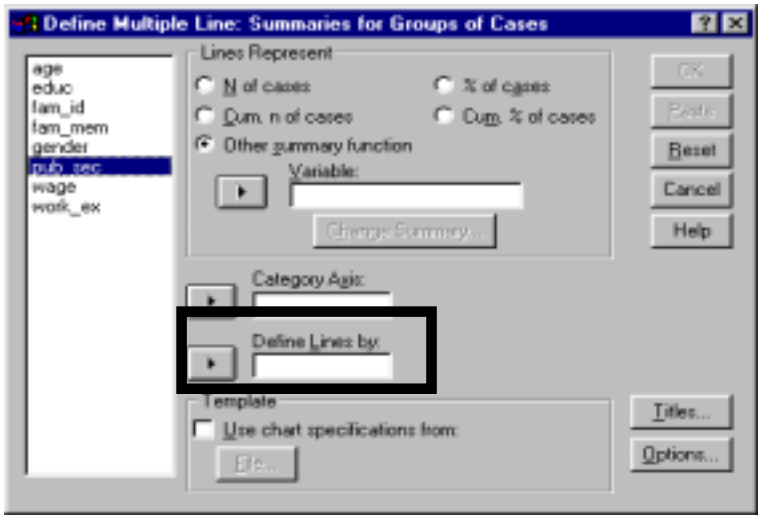
Let's assume that we want to compare the median education levels of males and females of different ages. We must make a multiple line graph.

To make a multiple line graph, go to GRAPHS/ LINE.

Select "Multiple," and "Summaries for groups of cases." Click on "Define."



The following dialog box opens:



Compare it to the dialog boxes in section 5.1. The "Define lines by" area is the only difference. Essentially, this allows you to use three variables -

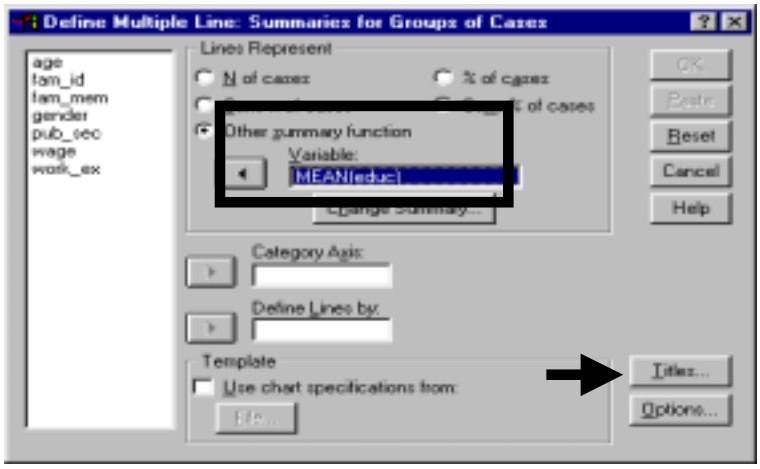
- in the box "Variable" you may place a continuous variable (as you did in section 5.1.b),
- in the box "Category Axis[63]" you may place a category variable (as you did in section 5.1.b) and
- in the box "Define lines by[64]" you may place a dummy variable or a categorical variable with few categories. This is the 3rd dimension we mentioned in the introduction to this section.

[63] The variable in "Category Axis" defines the X-axis.

[64] Each line in the graph depicts the line for one category of the variable placed in "Define Lines by."

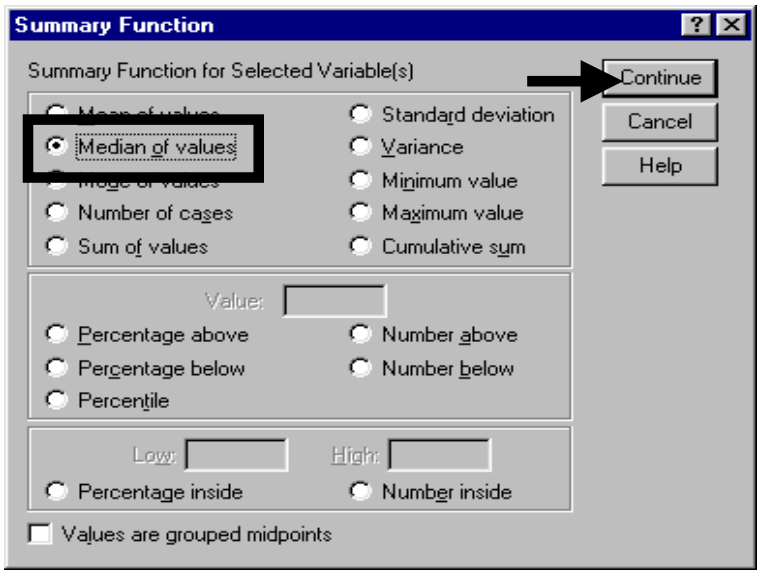
Place the continuous variable *educ* into the box "Variable."

Click on "Options."

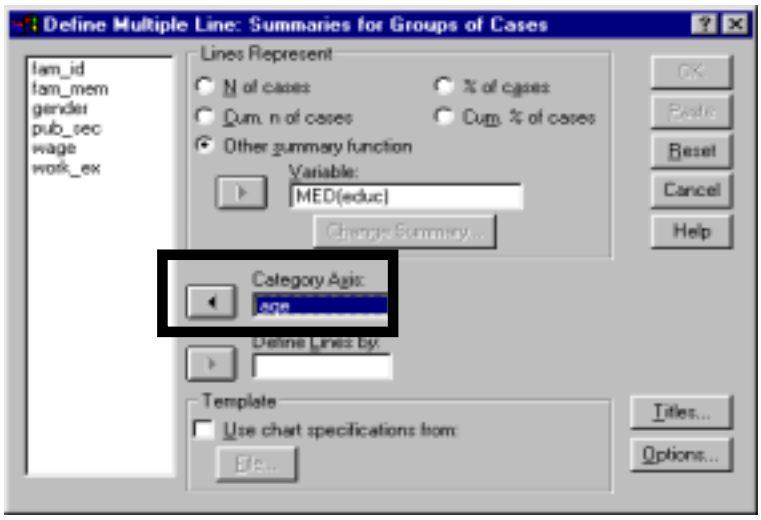


Select the summary statistic you desire. We have chosen "Median of values."

Click on "Continue."

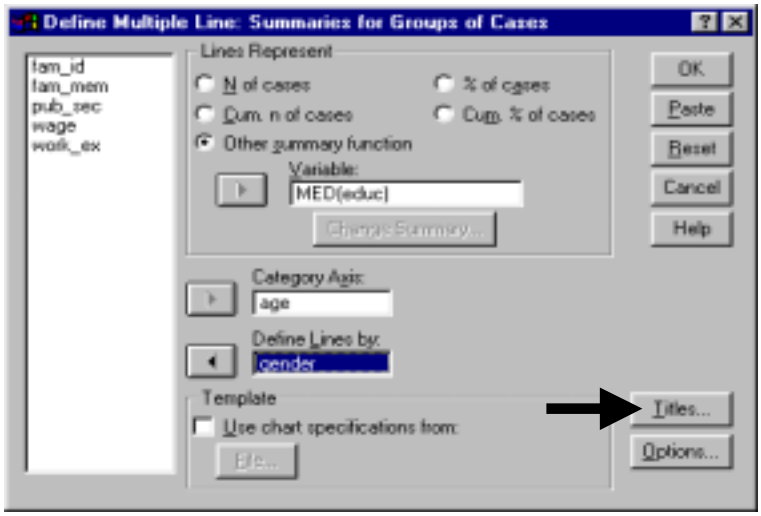


We want to have *age* levels on the X-axis. To do this, move the variable *age* into the box "Category axis."

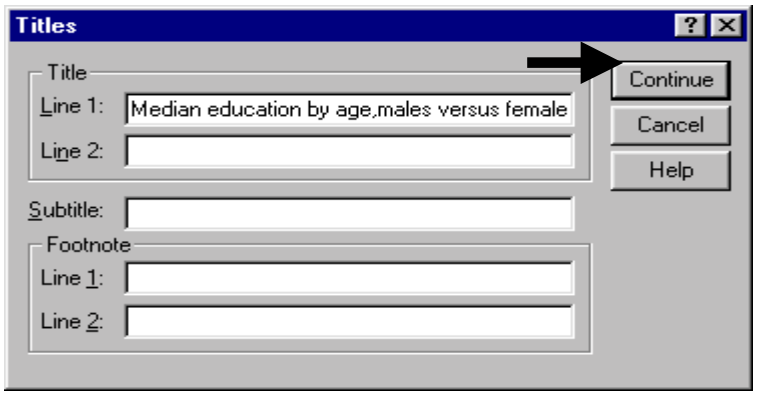


Further, we wish to separate lines for males and females. To achieve this, move the variable *gender* into the box "Define Lines by."

Click on "Titles."



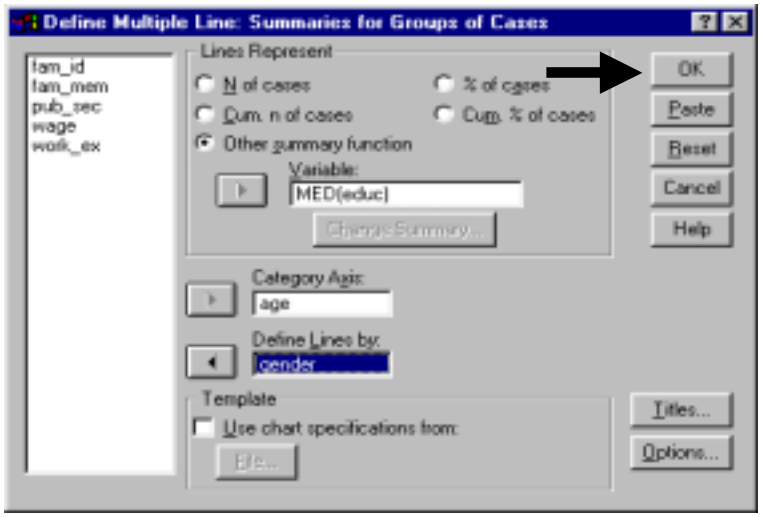
Enter an appropriate title. Click on "Continue."

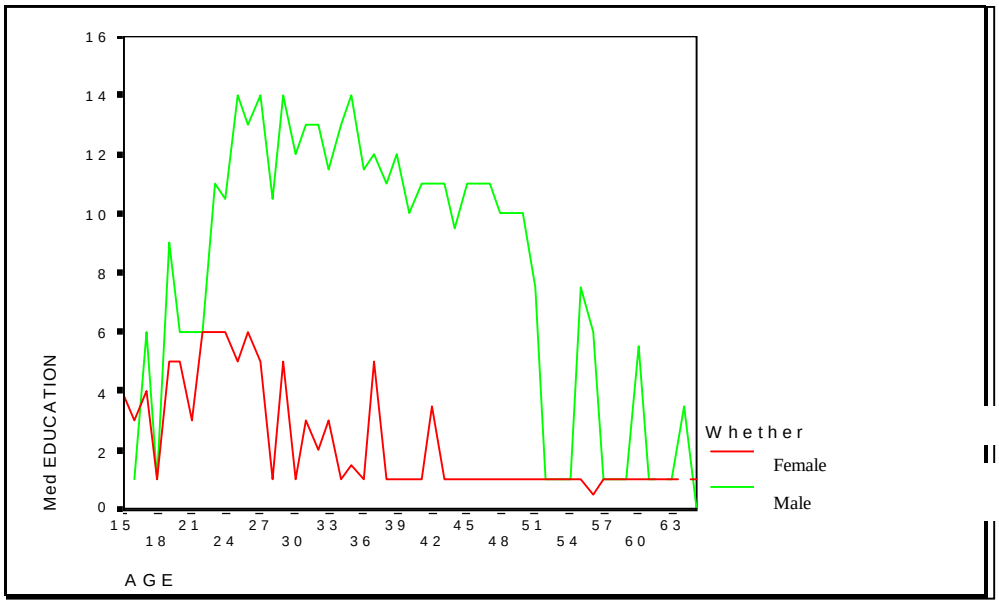


Click on "OK."

The next graph shows the results. Notice: Dimension 1 (the X-axis) is *age*, dimension 2 (each line) is *gender*, and dimension 3 is the median *education* (the Y-axis).

Would it not have been better to make a bar graph? Experiment and see which graph type best depicts the results.



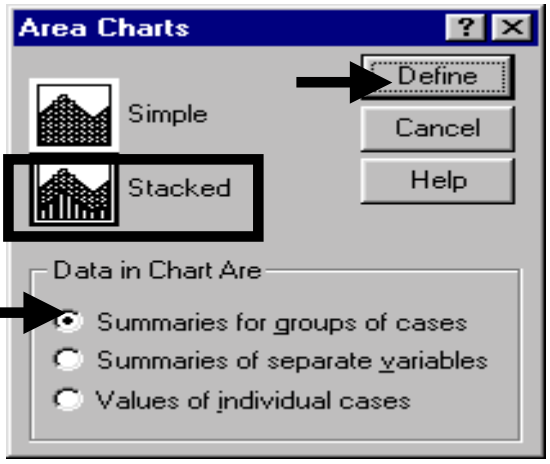


Example 2: Area graph for comparing aggregate statistics

Go to GRAPHS/ AREA.

Select "Stacked" and "Summaries for groups of cases."

Click on "Define."

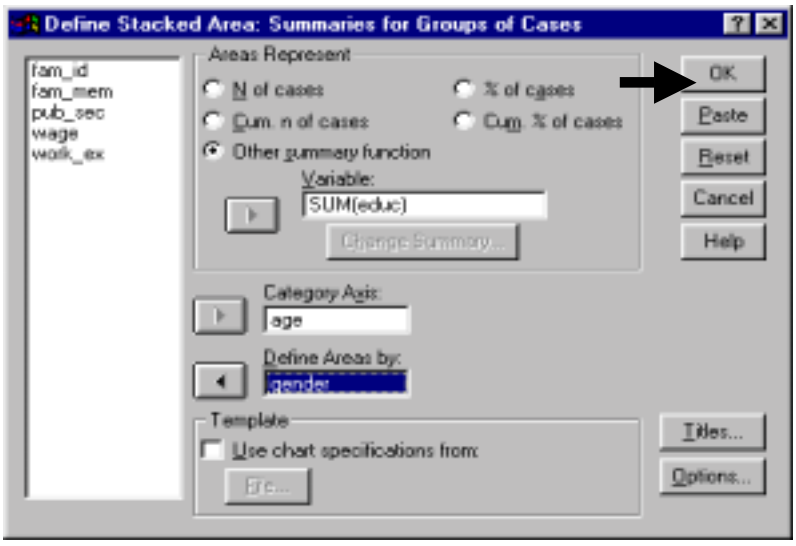


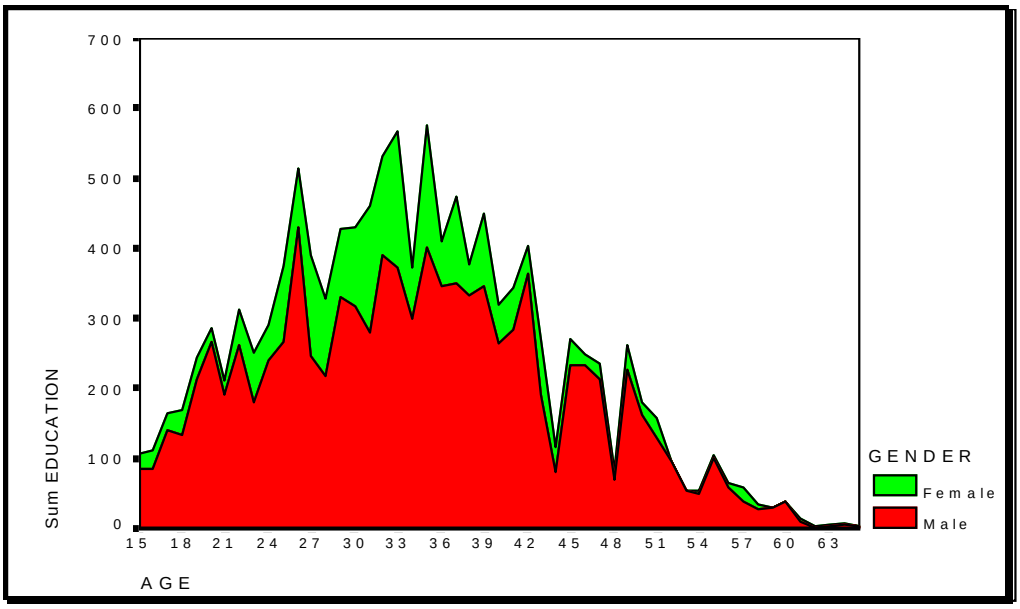
This time we will skip some of the steps.

Enter information as shown on the right (see example 1 of this section for details on how to do so).

Click on "OK."

The resulting graph is shown below. Dimension 1 (the X-axis) is *age*, dimension 2 (each area) is *gender* and, dimension 3 (the statistic shown in the graph and therefore the Y-axis label and unit) is the sum of *education*. Note that each point on both area curves are measured from the X-axis (and not, as in some Excel graphs, from the other curve).





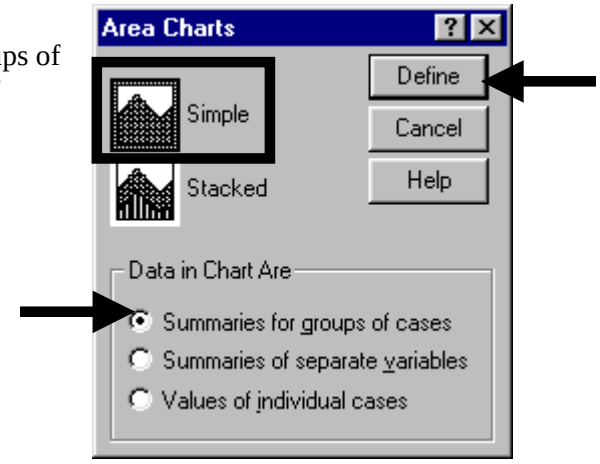
All the examples above used a standard statistic like the mean, median, sum, or standard deviation. In section 5.1.c we explore the capacity of SPSS to use customized statistics (like "Percent Below 81").

Ch 5. Section 1.c. Using graphs to capture user-designed criteria

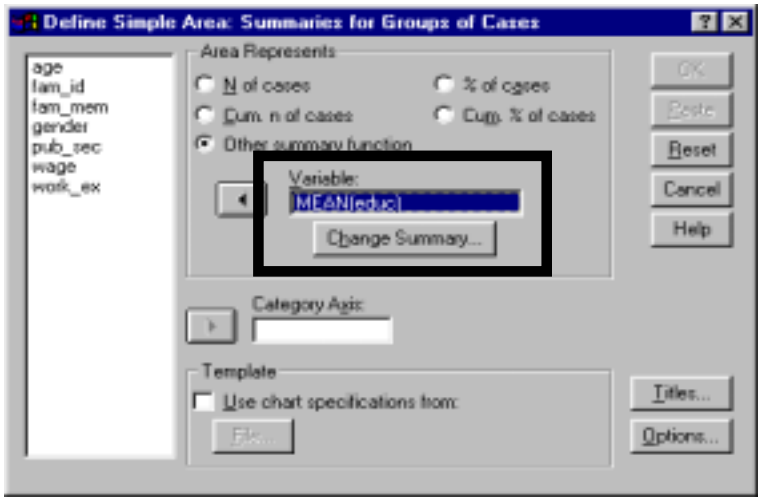
Apart from summary measures like mean, median, and standard deviation, SPSS permits some customization of the function/information that can be depicted in a chart.

Example 1: Depicting "Respondents with at least primary education"

Go to GRAPHS/ AREA. Select "Simple" and "Summaries of groups of cases" and then click on "Define."



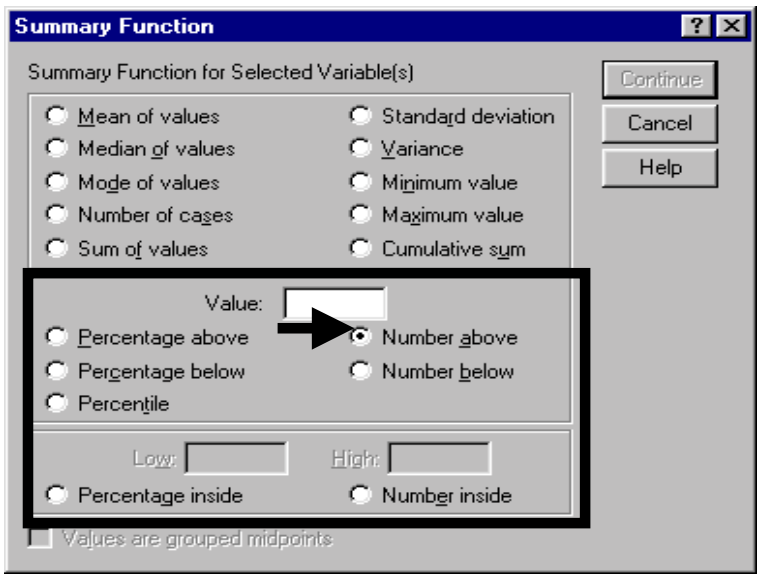
After moving *educ*, click on the button "Change Summary."



We want to use the statistic "Respondents with at least primary education." In more formal notation, we want "Number > 6" (assuming primary education is completed at grade 6).

Click on "Number above."

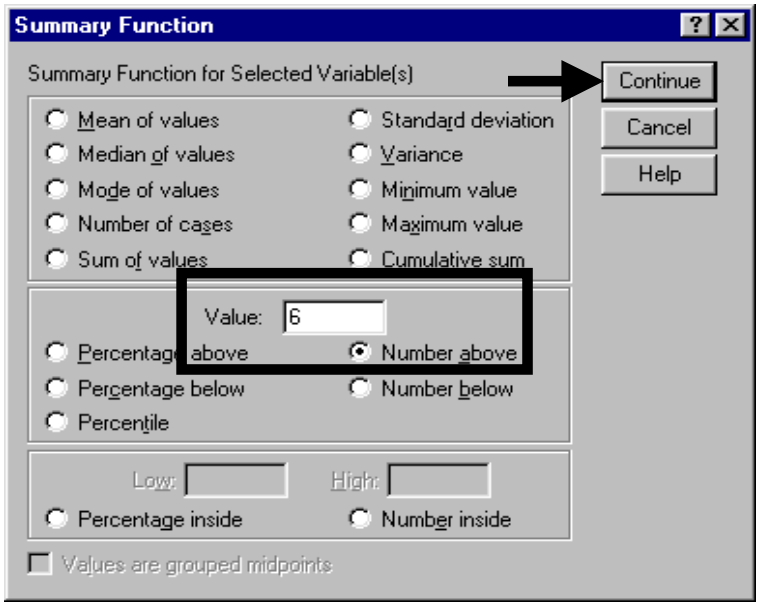
Note: The area inside the dark-bordered rectangular frame shows the options for customization.



Enter the relevant number. This number is 6 (again assuming that primary schooling is equivalent to 6 years).

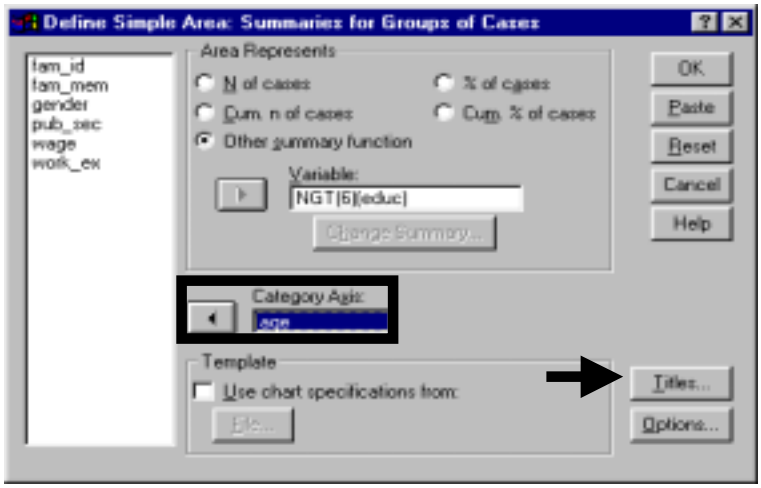
Note: You may prefer using "Percentage above 6." Experiment with the options until you become familiar with them.

Click on "Continue."



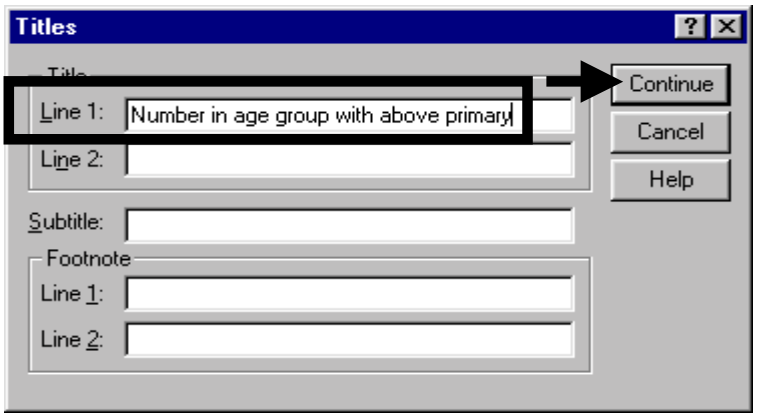
Enter “Age” into the box “Category Axis.”

Click on “Titles.”

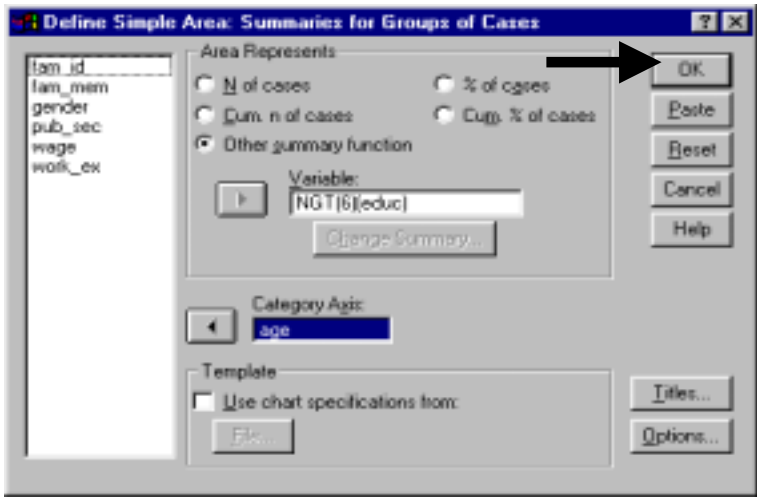


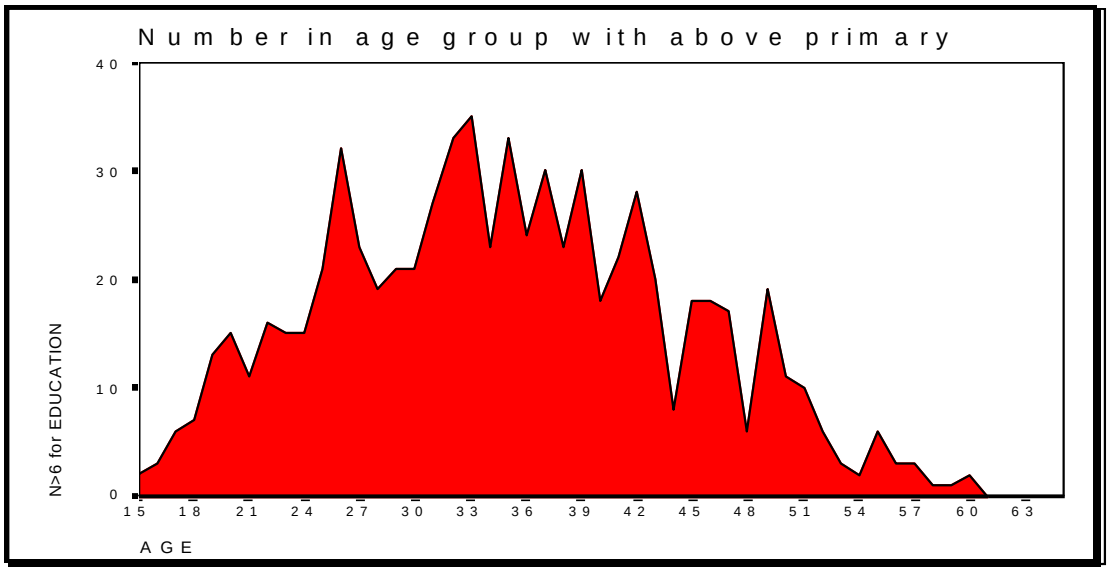
Enter an appropriate title.

Click on “Continue.”



Click on “OK.”

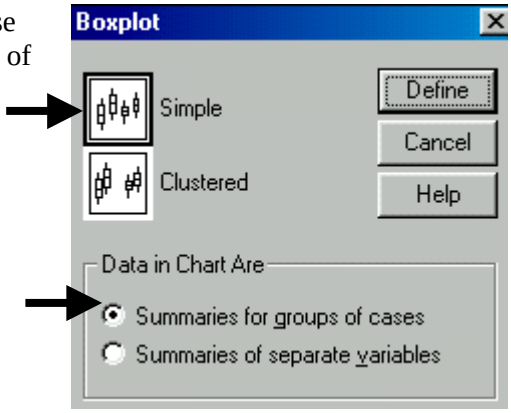




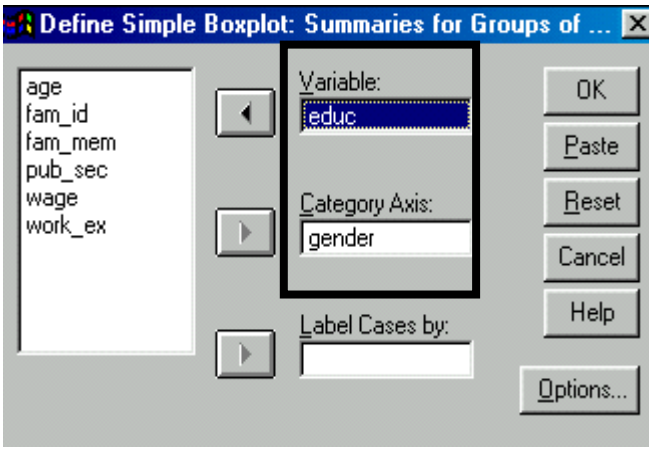
Ch 5. Section 1.d. Boxplots

Boxplots provide information on the differences in the quartile distributions of sub-groups of one variable, with the sub-groups being defined by categories of another variable. Let's assume that we want to compare the quartile positions for *education* by the categories of *gender*.

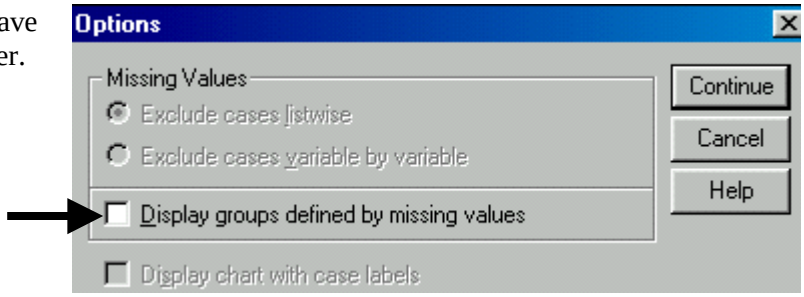
Go to GRAPHS / BOXPLOT. Choose "Simple" and "Summaries of Groups of Cases."



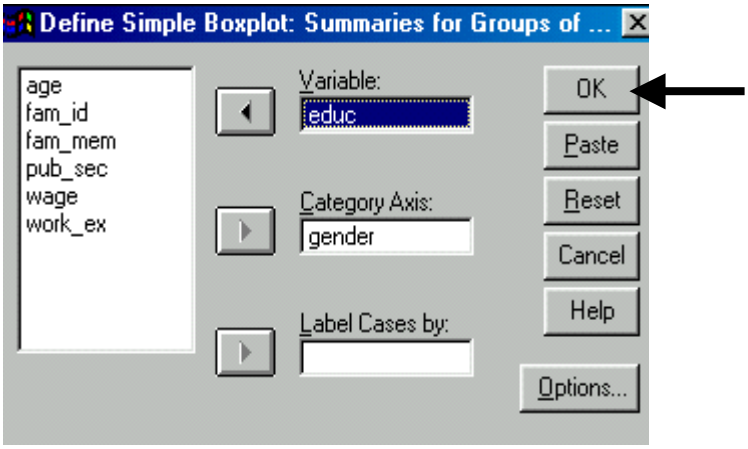
Place the variable whose boxplot you wish to analyze into the box "Variable." Place the categorical variable, which defines each boxplot, into the box "Category Axis."



Click on options and choose not to have a boxplot for missing values of gender. Click on "Continue."

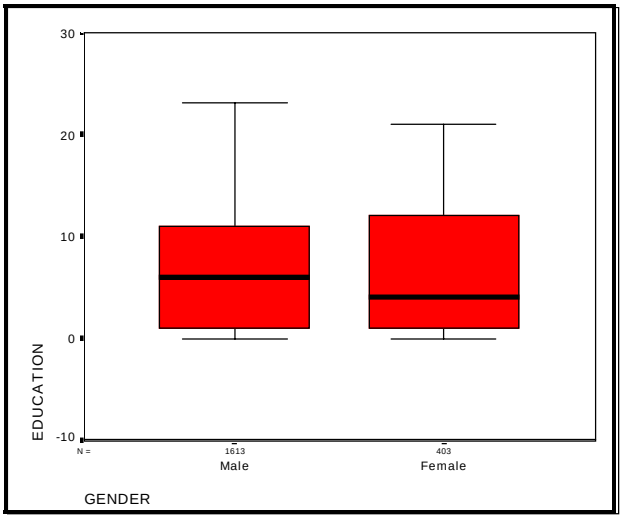


Click on "OK."



The lowest quartile is very similar for males and females, whereas the second quartile lies in a narrower range for females. The median (the dark horizontal area within the shaded area) is lower for females and, finally, the third quartile is wider for females.

Note: See 4.2 for a more detailed interpretation of boxplots.



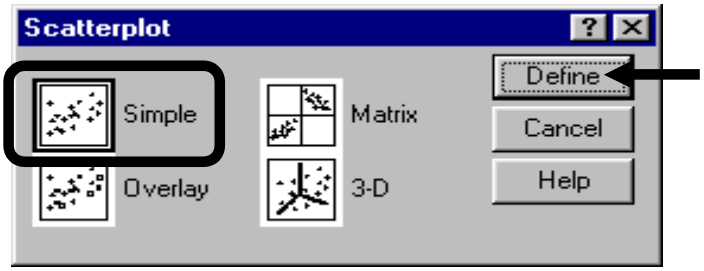
Ch 5. Section 2 Scatters

Ch 5. Section 2.a. A simple scatter

Scatters are needed to view patterns between two variables.

Go to GRAPHS/SCATTER.

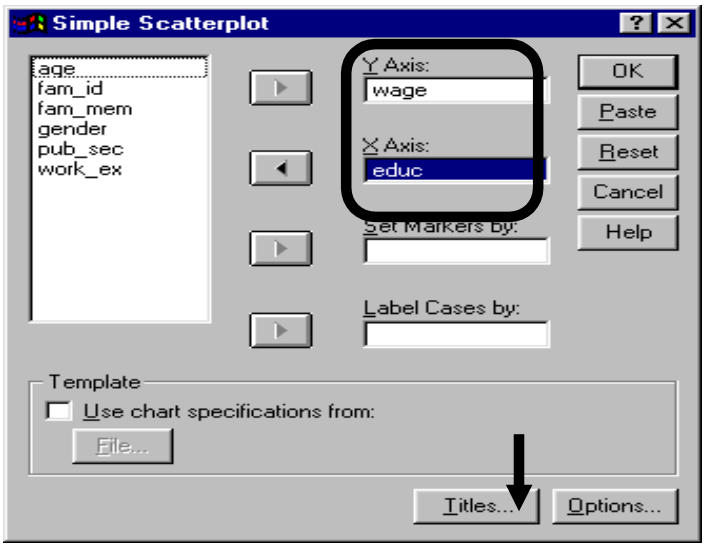
Select the option "Simple" and click on "Define."



Select the variable *wage*. Place it in the box "Y Axis."

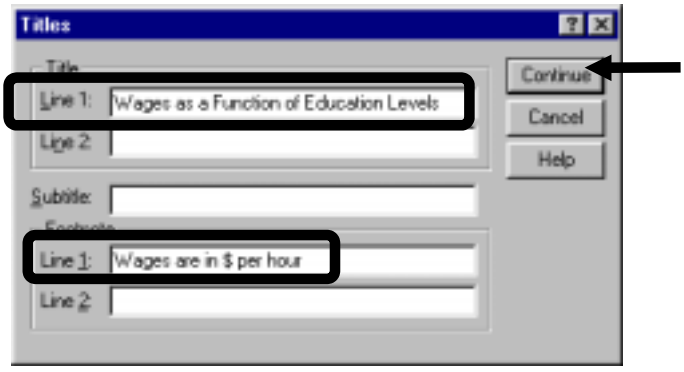
Select the variable *educ*. Place it in the box "X-Axis."

Click on "Titles."

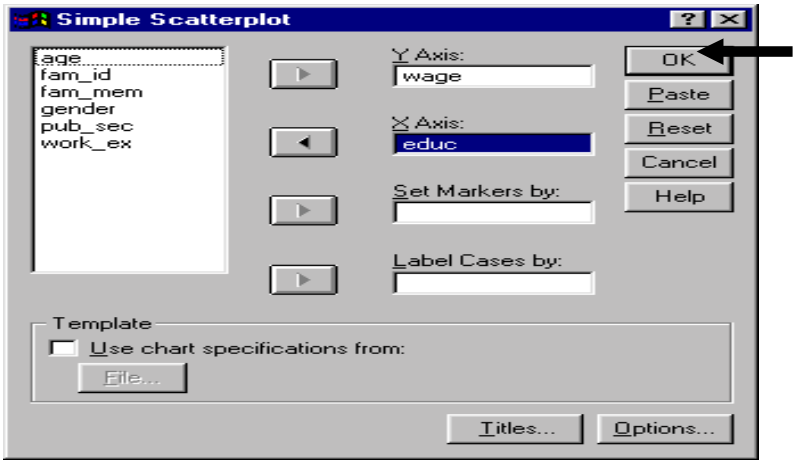


Type in a title and footnote.

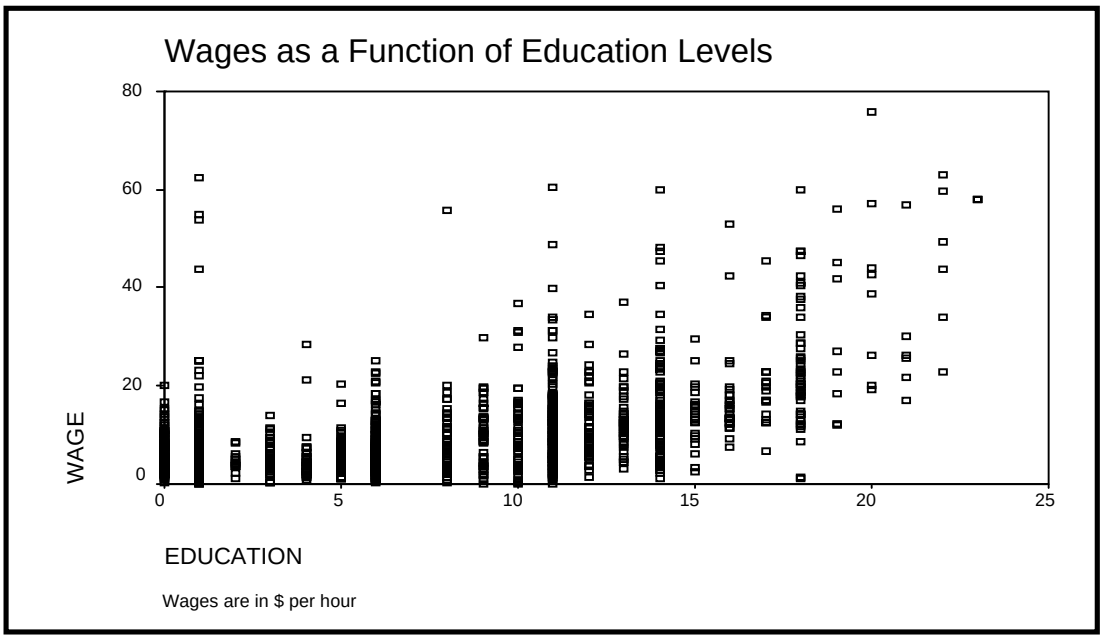
Click on "Continue."



Click on "OK."



Scatter plots often look different from those you may see in textbooks. The relation between the variables is difficult to determine conclusively (sometimes changing the scales of the X and/or Y axis may help - see section 11.2 for more on that). We use methods like correlation (see section 5.3) to obtain more precise answers on the relation between the variables.

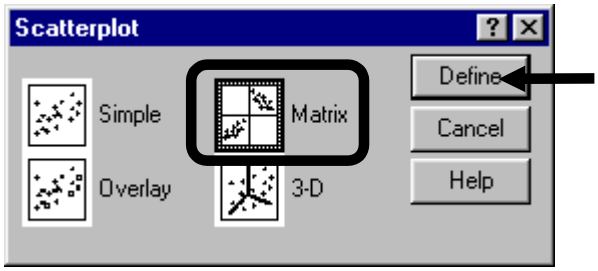


Ch 5. Section 2.b. Plotting scatters of several variables against one other

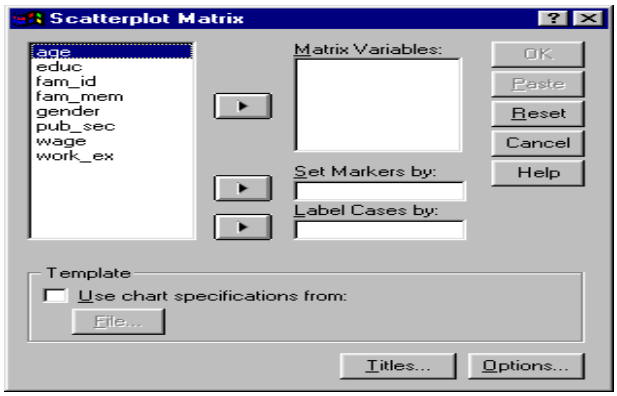
If you want to create scatters that include multiple variables, you can use SPSS to create several simple scatter plots (rather than executing "simple scatters" four times, you can use "matrix scatter" feature). This feature is useful for saving time.

Go to GRAPHS/SCATTER.

Select the option "Matrix" and click on "Define."

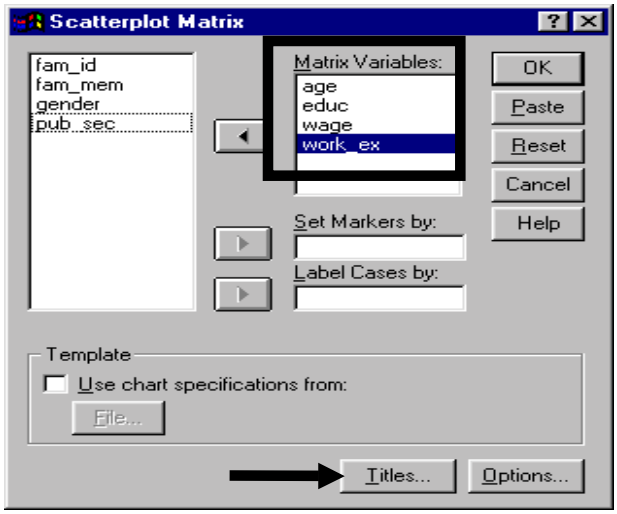


The following dialog box will open.



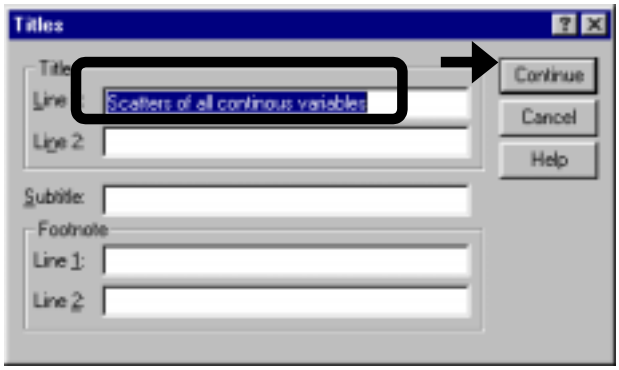
Select the variables whose scatters you wish to view. A scatter of each combination will be produced.

Click on "Titles."



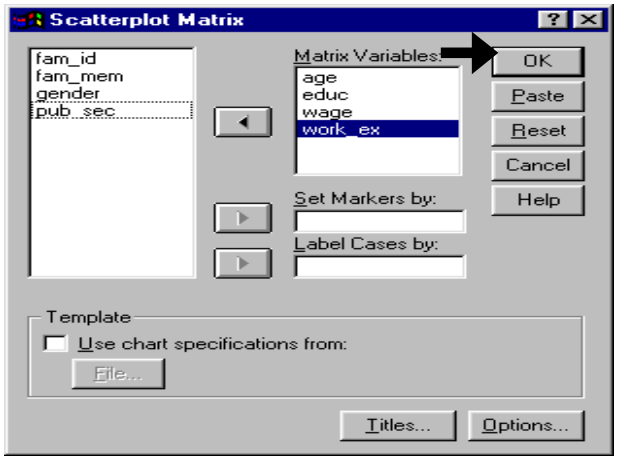
Enter a title.

Click on "Continue."



Click on "OK."

Scatters of all possible pairs of the four variables will be created. They will be shown in one block.



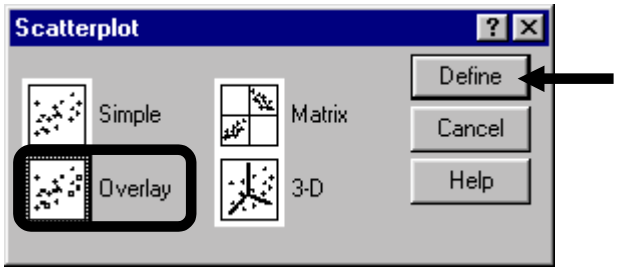
Ch 5. Section 2.c. Plotting two X-variables against one Y

If two independent variables are measured on the same scale and have similar values, an overlay chart can be used to plot scatters of both these variables against the dependent variable on one chart. The goal is to compare the differences in the scatter points.

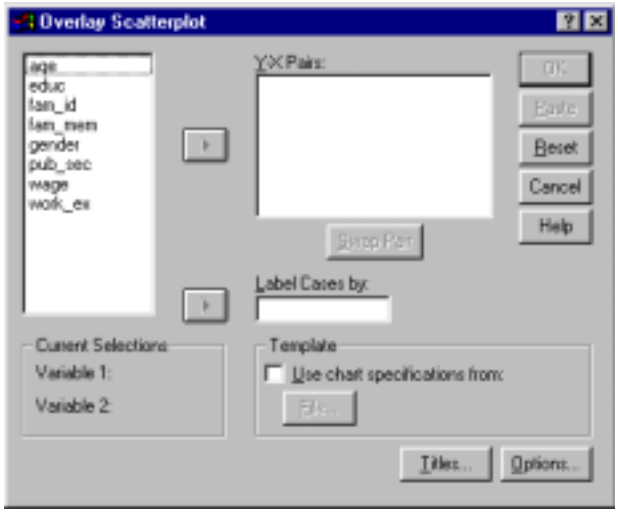
Let's assume you want to compare the relationship between *age* and *wage* with the relationship between *work experience* and *wage*.

Go to GRAPHS/SCATTER.

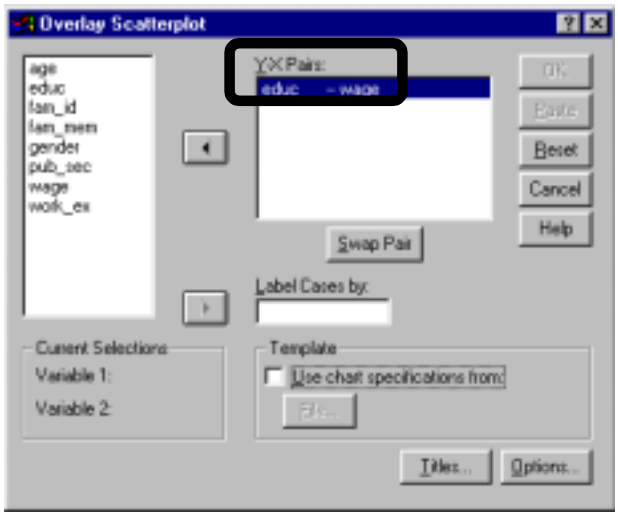
Select the option "Overlay" and click on "Define."



The following dialog box will open.

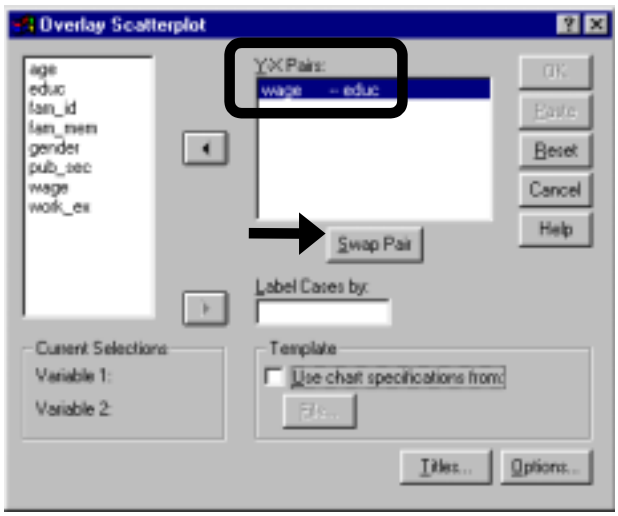


Click on *educ*. Press CTRL and click on *wage*. Click on the right-arrow button and place the chosen pair into the box "Y-X Pairs."



The first variable in the pair should be the Y-variable - in our example *wage*. But we currently have this reversed (we have *educ-wage* instead of *wage-educ*). To rectify the error, click on the button "Swap pair."

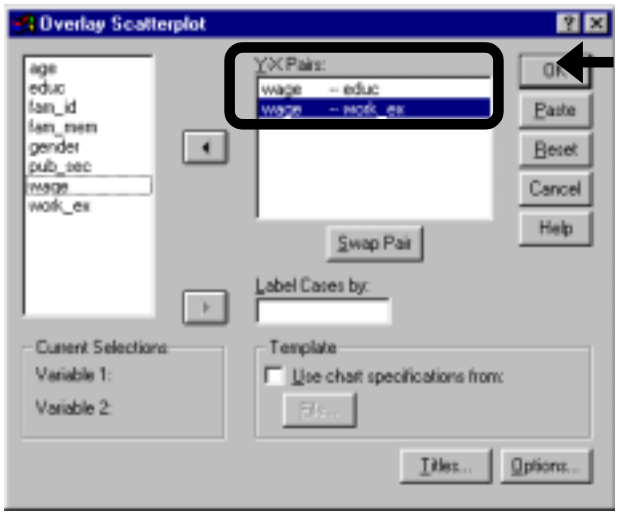
Note: Click on "Title" and include an appropriate title.

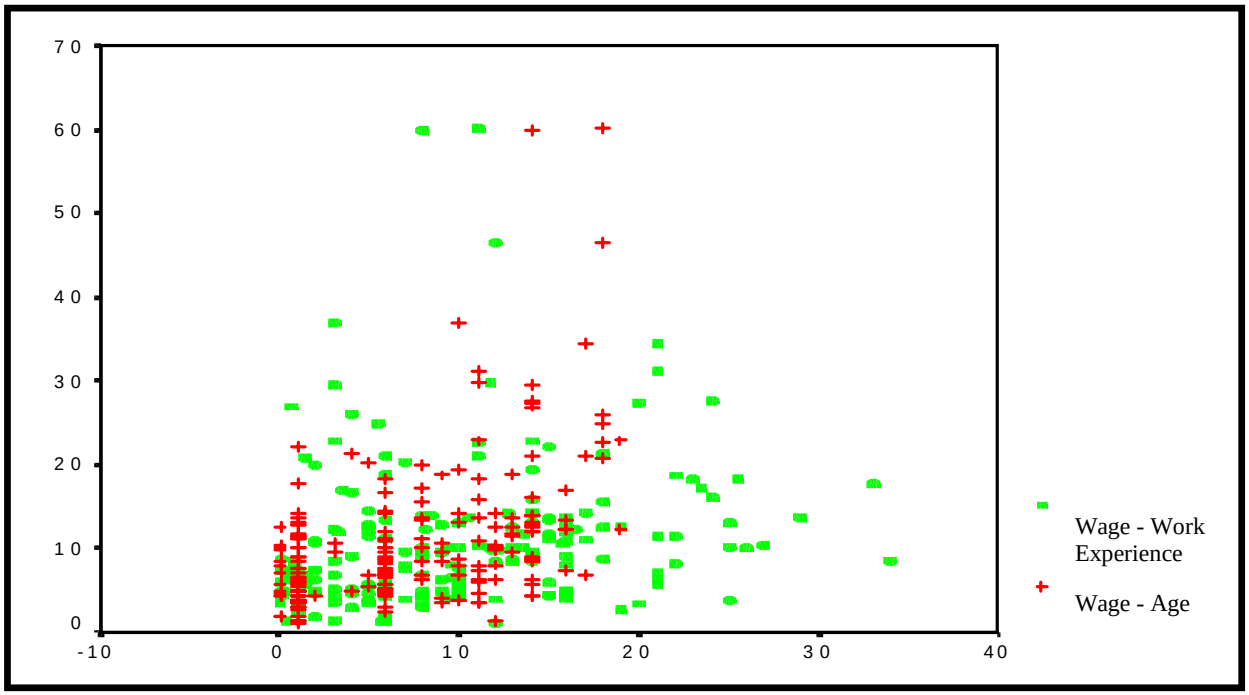


Repeat the previous two steps for the pair *wage* and *work_ex*.

Click on "OK."

An overlay scatter plot will be made. The next figure shows the plot




70
60
50
40
30
20
10
0
-10
0
10
20
30
40
Wage - Work Experience
Wage - Age

Ch 5. Section 3 Correlations

The correlation coefficient depicts the basic relationship across two variables[65]: "Do two variables have a tendency to increase together or to change in opposite directions and, if so, by how much[66]?

Bivariate correlations estimate the correlation coefficients between two variables at a time, ignoring the effect of all other variables. Sections 5.3.a and 5.3.b describe this procedure.

Section 5.3.a shows the use of the Pearson correlation coefficient. The Pearson method should be used only when each variable is quantitative in nature. Do not use it for ordinal or unranked qualitative[67] variables. For ordinal variables (ranked variables), use the Spearman correlation coefficient. An example is shown in section 5.3.b.

The base SPSS system does not include any of the methods used to estimate the correlation coefficient if one of the variables involved is unranked qualitative.

There is another type of correlation analysis referred to as "Partial Correlations." It controls for the effect of selected variables while determining the correlation between two variables[68]. Section 5.3.c shows an example of obtaining and interpreting partial correlations.

Note: See section 5.2 to learn how to make scatter plots in SPSS. These plots provide a good visual image of the correlation between the variables. The correlation coefficients measure the linear correlation, so look for such linear patterns in the scatter plot. These will provide a rough idea about the expected correlation and will show this correlation visually.

[65] Do not confuse correlation with regression. While the former does not presume any causal link between X and Y, the latter does.

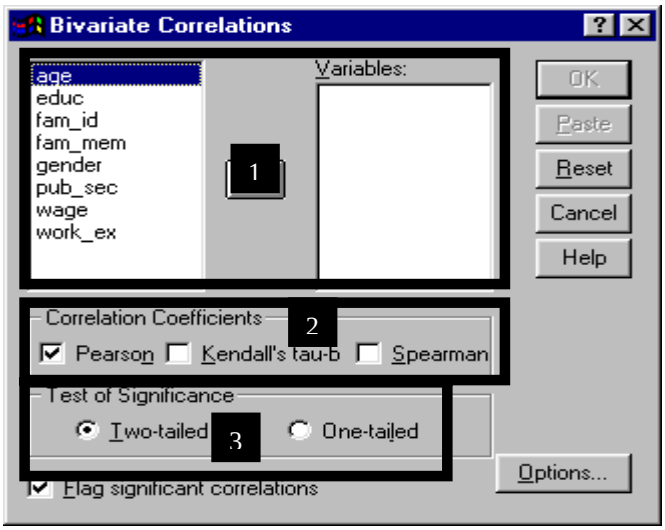
[66] The term "correlation" means "Co (together)" + "Relation." If variable X is higher (lower) when variable Z is higher (higher), then the two variables have a positive (negative) correlation. A correlation captures the linear co-relation, if any, shown in a scatter between the graphs (see section 5.2.)

[67] Example of ranked variables: *GPA* (taking on the ranked category values A+, A,…, C-). Example of an unranked qualitative variable (sometimes referred to as a nominal variable): *Gender* (there is no ranking between the categories male and female).

[68] In our data set, a bivariate correlation of *wage* and *work experience* will ignore all other variables. But is that realistic and/or intuitive? We know (from previous research and after doing other analysis) that *gender* and *education* play a major role in *wage* determination and *age* plays a major role in the determination of *work experience*. So, ideally, a "pure" correlation between *wage* and *work experience* should account for the effect of the variables *gender* and *education*. Partial correlation does this.

Ch 5. Section 3.a. Bivariate correlations

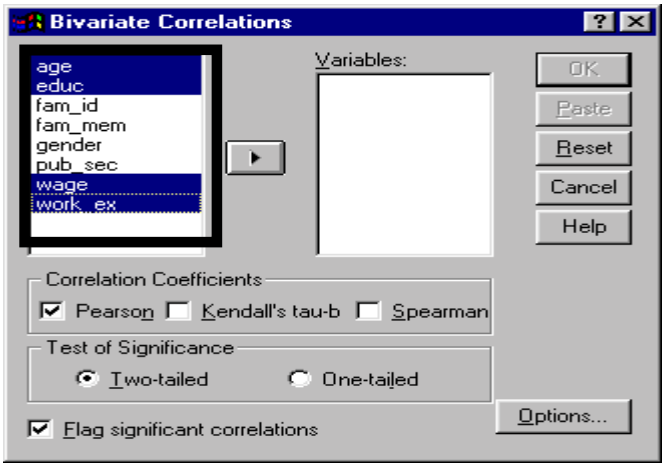
Go to STATISTICS/ CORRELATE/ BIVARIATE.



Area 1 allows you to choose the variables whose correlations you would like to determine. Correlations are produced in pairs between all the variables chosen in the box "Variables."

Area 2 is where you can choose the method for calculating the correlation coefficients.

In area 3 you can choose the direction of the significance test. Two-tailed is the typically selected option. However, if you are looking specifically for the significance in one direction, use the one-tailed test[69].



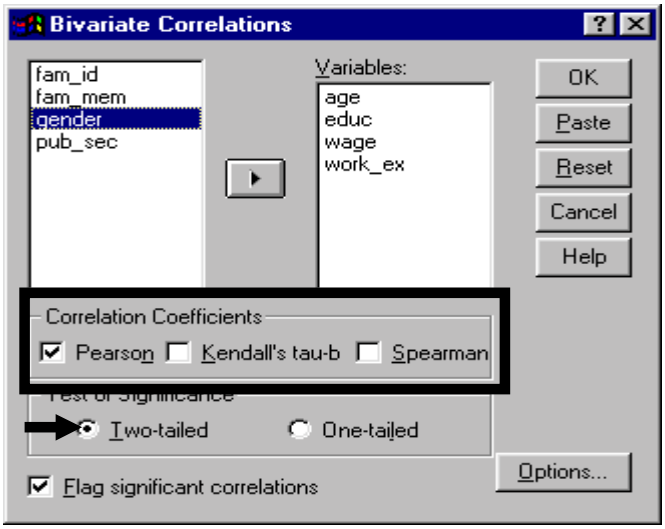
Choose the pairs of variables between which you wish to find bivariate correlation coefficients. To do so, click on the first variable name, then press the CTRL button and click on the other variable names. Then press the arrow button.

[69] Check your statistics book for a description of "one-tailed" and "two-tailed."

Select the method(s) of finding the correlations. The default is "Pearson."[70]

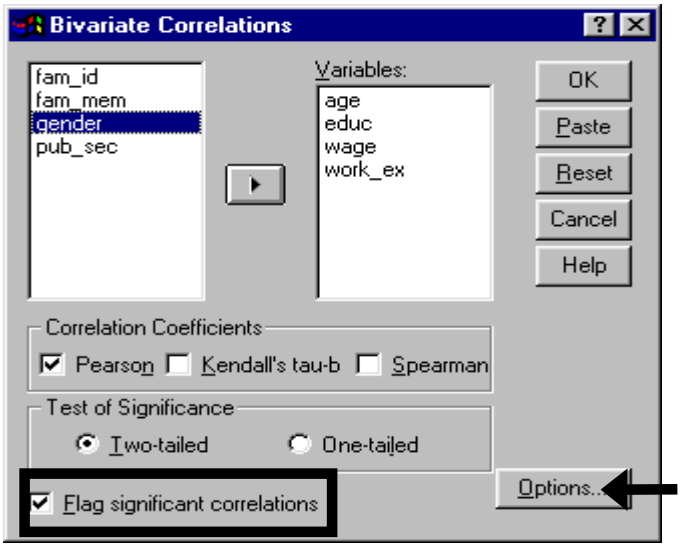
Select "Two-Tailed" in the area "Test of Significance."

The two-tailed test is checking whether the estimated coefficient can reliably be said to be above 0 (tail 1) or below 0 (the second tail). A one-tailed test checks whether the same can be said for only one side of 0 (e.g. - set up to check if the coefficient can be reliably said to be below 0).



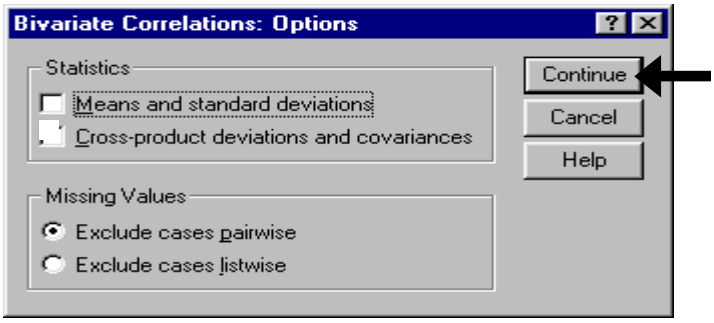
On the bottom, there is an option "Flag Significant Coefficients." If you choose this option, the significant coefficients will be indicated by * signs.

Click on "Options."



In "Options," choose not to obtain Mean and Standard deviations[71].

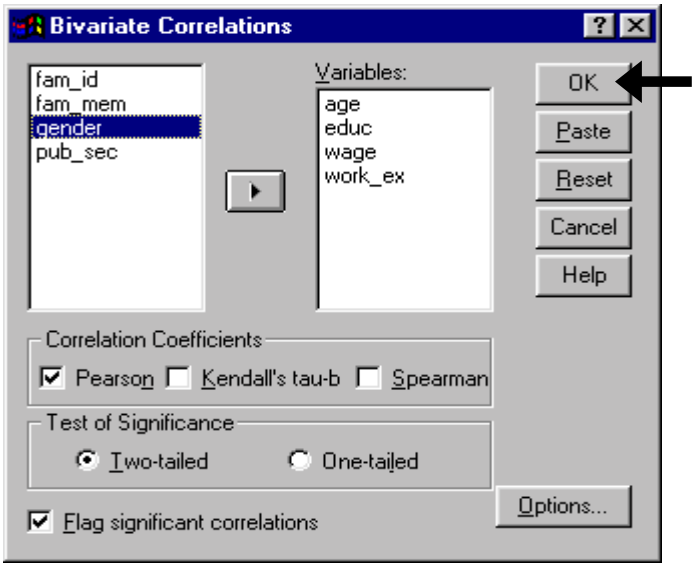
Click on "Continue."



[70] If even one of the variables is ordinal (dummy or categorical) or not normally distributed, you cannot use the method "Pearson." Our approach is simplistic. Your textbook may explain several different types of correlations and the guidelines regarding the use of each. Nevertheless, we follow an approach that is professionally accepted practice. Note that Pearson's method requires all variables to be distributed normally. Most researchers don't even bother to check if this is true! If the sample size is large enough (above 30) then most distributions start behaving like the normal distribution - the oft-quoted Central Limit Theorem!

[71] The mean and standard deviation are usually obtained in descriptives (see sections 3.2 and 3.3).

Click on "OK."



Note:
A high level of correlation is implied by a correlation coefficient that is greater than 0.5 in absolute terms (i.e - greater than 0.5 or less than –0.5).

A mid level of correlation is implied if the absolute value of the coefficient is greater than 0.2 but less that 0.5.

A low level of correlation is implied if the absolute value of the coefficient is less than 0.2.

The output gives the value of the correlation (between -1 and 1) and its level of significance, indicating significant correlations with one or two * signs. First, check whether the correlation is significant (look for the asterisk). You will then want to read its value to determine the magnitude of the correlation.

Make this a habit. Be it correlation, regression (chapter 7 and 8), Logit (chapter 9), comparison of means (sections 4.4 and 5.5), or the White's test (section 7.5), you should always follow this simple rule - **first look at the significance**. If, and only if, the coefficient is significant, then rely on the estimated coefficient and interpret its value.

This row contains the correlation coefficients between all the variables.

Correlations

		AGE	EDUCATION	WAGE	WORK_EX
Pearson Correlation	AGE	1.000	-.051*	.274**	.674**
	EDUCATION	-.051*	1.000	.616**	-.055*
	WAGE	.274**	.616**	1.000	.254**
	WORK_EX	.674**	-.055*	.254**	1.000
Sig. (2-tailed)	AGE	.	.021	.000	.000
	EDUCATION	.021	.	.000	.014
	WAGE	.000	.000	.	.000
	WORK_EX	.000	.014	.000	.
N	AGE	2016	2016	1993	2016
	EDUCATION	2016	2016	1993	2016
	WAGE	1993	1993	1993	1993
	WORK_EX	2016	2016	1993	2016

*. Correlation is significant at the 0.05 level (2-tailed).

**. Correlation is significant at the 0.01 level (2-tailed).

Correlation coefficient is > 0. This implies that the variables *age* and *work experience* change in the same direction. If one is higher, then so is the other. This result is expected. The two asterisks indicate that the estimate of 0.674 is statistically significant at the 0.01 level - a 99% degree of confidence.

This result is interesting. Age and education change in different directions, though the magnitude of the relation is very small (-.05) . This is an unexpected result. Serendipity? Perhaps. But that is what a detailed bivariate analysis will yield - unanticipated insight!

The coefficient of determination can be roughly interpreted as the proportion of variance in a variable that can be explained by the values of the other variable. The coefficient is calculated by squaring the correlation coefficient. So, in the example above, the coefficient of determination between age and work experience is the square of the correlation coefficient.

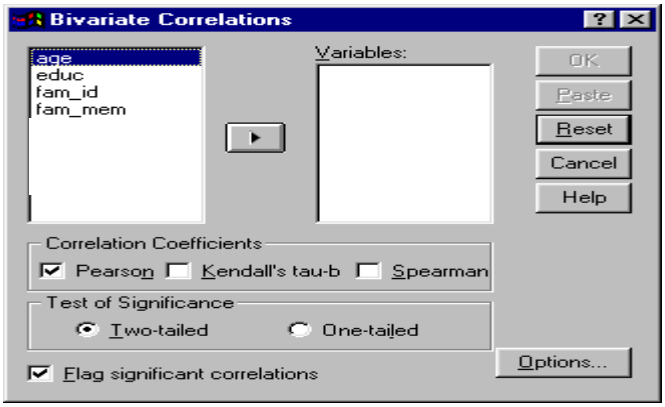
Coefficient of determination (age, work experience) = [correlation(age, work experience)]^2 = [0.674] ^{2} = 0.454

[or, 45.4% of the variance of one variable can be explained by the other one]

Ch 5. Section 3.b. Bivariate correlation if one of the variables is ordinal (ranked categorical) or not normally distributed

If even one of the variables is ordinal (ranked categorical) or non-normal, you cannot use the method "Pearson."[72] You must use a "non-parametric" method (see chapter 14 for a definition of non-parametric methods). Age and education may be considered to be such variables (though strictly speaking, the Spearman's is better used when each variable has a few levels or ordered categories). These facts justify the use of Spearman's correlation coefficient.

Go to STATISTICS/ CORRELATE/ BIVARIATE.

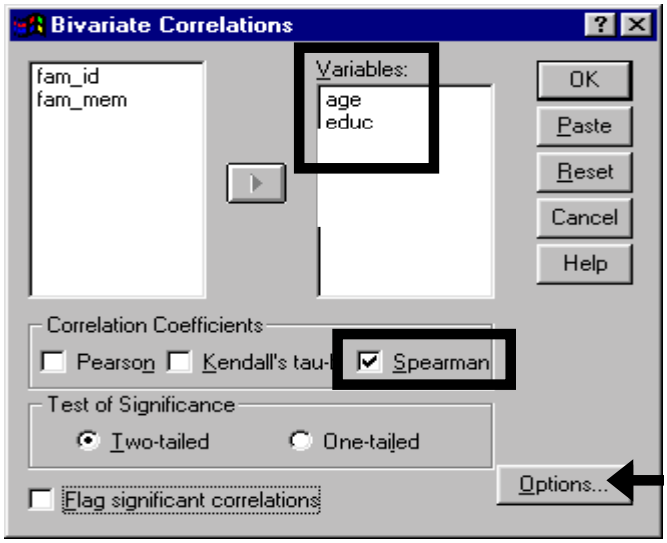


[72] In practice, many researchers ignore this rule! They therefore ignore all we have in chapters 3-6, going straight from descriptives (section 3.3) to regression (chapters 7 and 8). Completing all the steps (chapters 3-11) will engender a thorough command over the project and an understanding of the implications of each result.

Select the variables for the analysis.

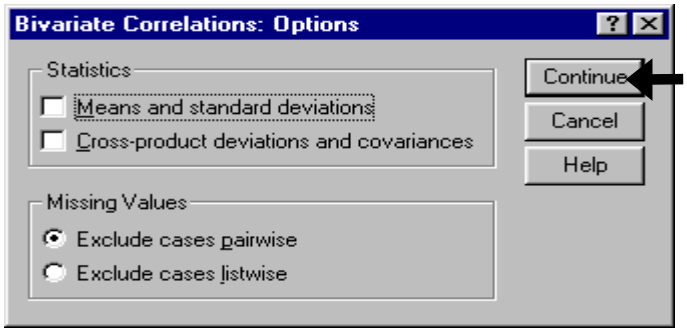
Click on "Spearman" in the area "Correlation Coefficients" after deselecting "Pearson."

Click on "Options."

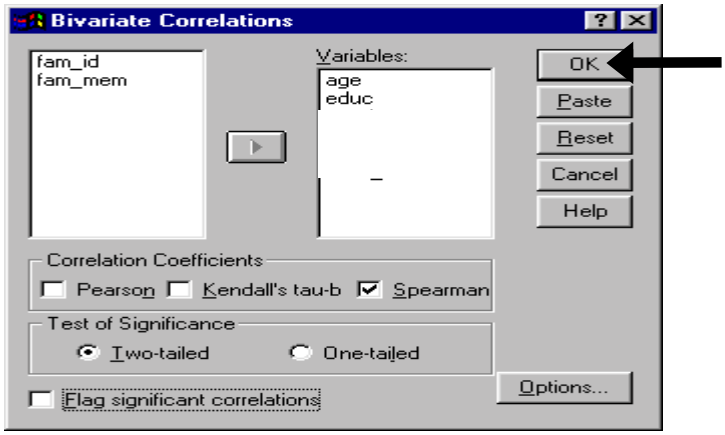


Deselect all.

Click on "Continue."



Click on "OK."



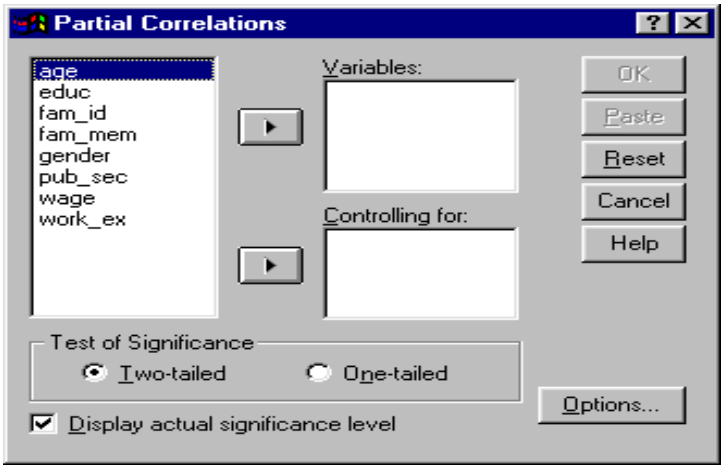
The output table looks similar to that using Pearson's in section 5.3.a. The difference is that a different algorithm is used to calculate the correlation coefficient. We do not go into the interpretations here - check your textbook for more detailed information.

Ch 5. Section 3.c. Partial correlations

With partial correlations, the correlation coefficient is measured, controlling for the effect of other variables on both of them. For example, we can find the correlation between *age* and *wage* controlling for the impact of *gender*, *sector*, and *education* levels.

Note: Partial Correlation is an extremely powerful procedure that, unfortunately, is not taught in most schools. In a sense, as you shall see on the next few pages, it provides a truer picture of the correlation than the "bivariate" correlation discussed in section 5.3.a.

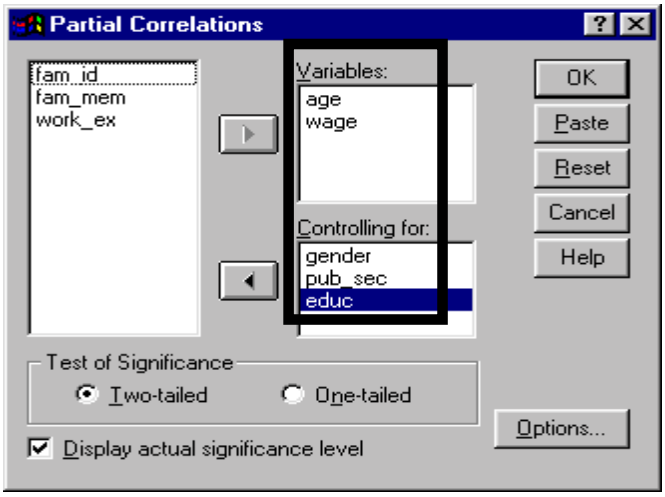
Go to STATISTICS/ CORRELATE/ PARTIAL CORRELATIONS.



Move the variables whose correlations you wish to determine into the box "Variables."

Move the variables whose impact you want to control for into the box "Controlling for."

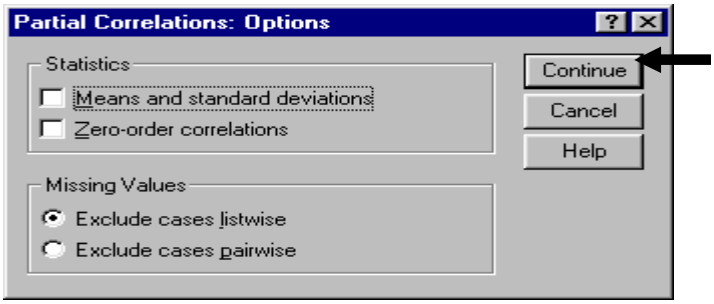
Select the option "Two-tailed" in the area "Test of Significance." This sets up a test to determine whether the estimated correlation coefficient can reliably be said to be either lower (tail 1) or higher (tail 2) than 0.



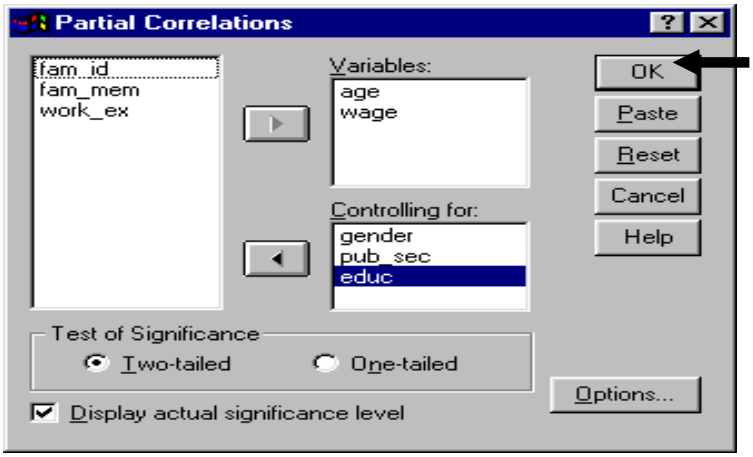
Click on "Options."

Deselect all options.

Click on "Continue."



Click on "OK."



Reminder: In section 5.3 we did not control for the effect of other variables while calculating the correlation between a pair of variables.

```
- - -  P A R T I A L   C O R R E L A T I O N   C O E F F I C I E N T S
- - -

Controlling for GENDER, PUB_SEC, EDUC

                    AGE          WAGE

AGE              1.0000         .3404
                 P= .           P= .000

WAGE              .3404         1.0000
                 P= .000        P= .

(Coefficient / (D.F.) / 2-tailed Significance)

" . " is printed if a coefficient cannot be computed
```

The correlation is significant at the 0.01 % level (as P< .01).

The interesting fact is that the partial correlation is higher than the bivariate (see section 5.3.a), implying that once one has removed the impact of *gender*, *sector*, and *education*, then *age* and *wage* have an even stronger relationship.

Ch 5. Section 4 Conducting several bivariate explorations simultaneously

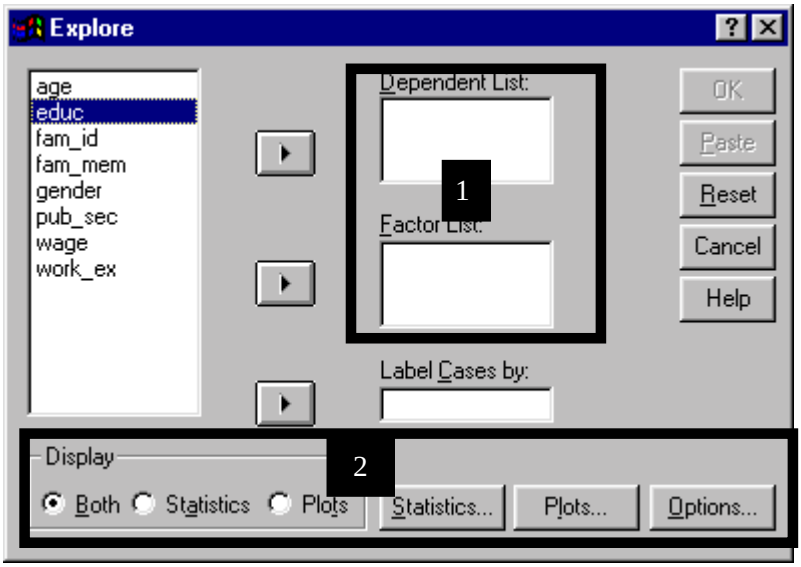
Comparing the attributes of a variable by another variable can be done in many ways, including boxplots, error bars, bar graphs, area graphs, line graphs, etc. Several of these variables can be done together using STATISTICS/ SUMMARIZE/ EXPLORE? This saves both time and effort.

Let's assume we want to find the differences in the attributes of the variables *education* and *wage* across the categories of *gender* and *sector*.

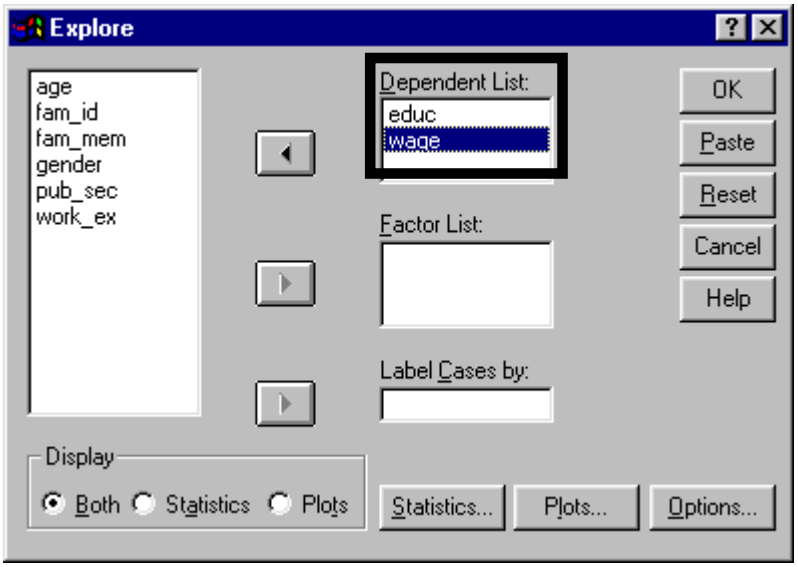
Go to STATISTICS/ SUMMARIZE/ EXPLORE.

In area 1 you will choose the list of variables whose categories you want to use as criteria for comparison and the variables whose attributes you want to compare.

In area 2, you choose the statistics (tables) and plots with which you would like to work.

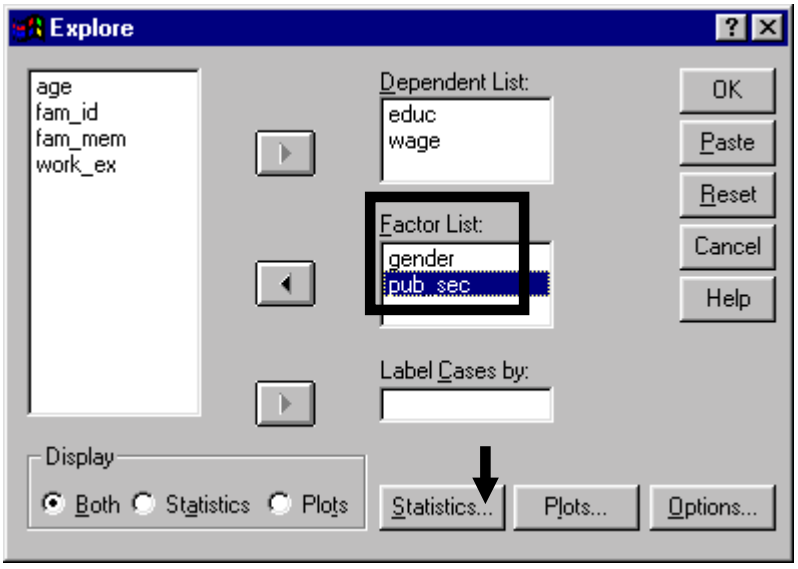


Move the variables *educ* and *wage* into the box "Dependants."



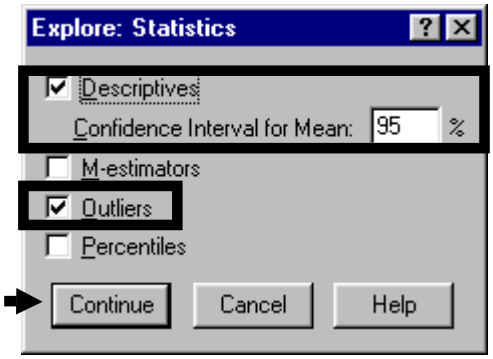
Move the dummy or categorical variables by whose categories you want to compare *educ* and *wage* into the "Factor List."

Click on the button "Statistics."



Select the statistics you want compared across the categories of *gender* and *sector*.

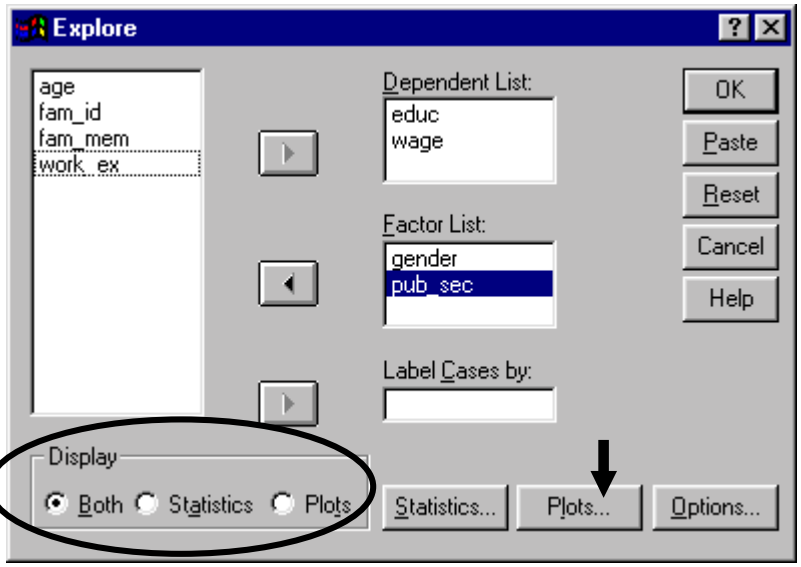
Here, the option "Descriptives" contains a wide range of statistics, including the confidence interval for the mean. "Outliers" gives the outliers by each sub-group (only male, only female, etc.).



"M-estimators" is beyond the scope of this book.

Click on "Continue."

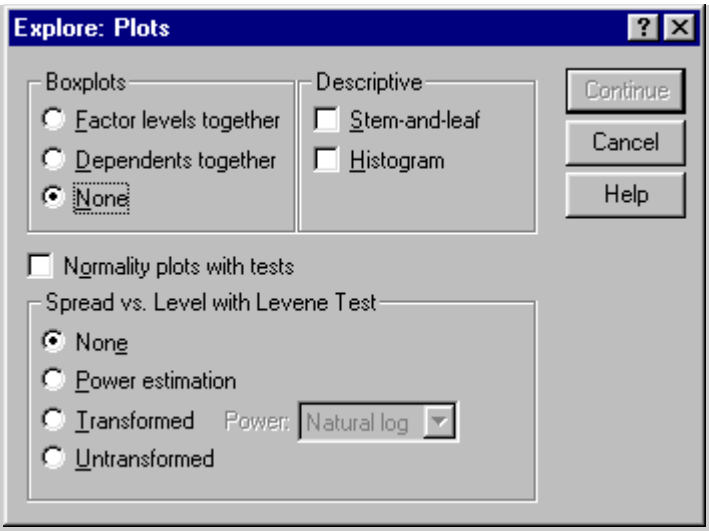
Click on the button "Plots."



Several plots can be generated.

Here we have deselected all options to show that, if you like, you can dictate the manner of the output, e.g. - that only the tables are displayed, only the plots are displayed, or both - EXPLORE is flexible.

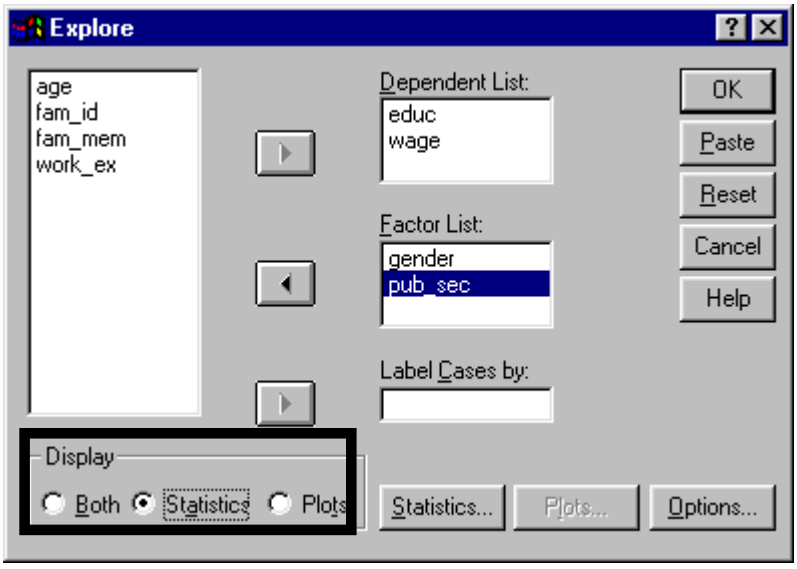
Click on "Cancel" (because you have deselected all options, the continue button may not become highlighted).



As mentioned above, you want to view only the tables with statistics. To do this, go to the area "Display" and choose the option "Statistics."

Click on "OK."

Several tables will be produced - basic tables on the cases in each variable and sub-group, tables with descriptives, and a table with information on the outliers.



Summary of *education* and *wage* by gender

Case Processing Summary

		Cases					
		Valid		Missing		Total	
	GENDER	N	Percent	N	Percent	N	Percent
EDUCATION	Male	*1613*	*100.0%*	*0*	*.0%*	*1613*	*100.0%*
	Female	*403*	*100.0%*	*0*	*.0%*	*403*	*100.0%*
WAGE	Male	*1613*	*100.0%*	*0*	*.0%*	*1613*	*100.0%*
	Female	*403*	*100.0%*	*0*	*.0%*	*403*	*100.0%*

Summary of *education* and *wage* by sector

Case Processing Summary

		Cases					
		Valid		Missing		Total	
	SECTOR	N	Percent	N	Percent	N	Percent
EDUCATION	Private Sector	*1292*	*100.0%*	*0*	*.0%*	*1292*	*100.0%*
	Public Sector	*724*	*100.0%*	*0*	*.0%*	*724*	*100.0%*
WAGE	Private Sector	*1292*	*100.0%*	*0*	*.0%*	*1292*	*100.0%*
	Public Sector	*724*	*100.0%*	*0*	*.0%*	*724*	*100.0%*

On the next few pages you will see a great deal of output. We apologize if it breaks from the narrative, but by showing you the exhaustive output produced by EXPLORE, we hope to impress upon you the great power of this procedure. The descriptives tables are excellent in that they provide the confidence intervals for the mean, the range, interquartile range (75^{th} - 25^{th} percentile), etc. The tables located two pages ahead show the extreme values.

Tip: Some of the tables in this book are poorly formatted. Think it looks unprofessional and sloppy? Read chapter 11 to learn how not to make the same mistakes that we did!

Descriptives of *education* and *wage* by gender

Descriptives

GENDE				Statistic	Std. Error
EDUCATIO	Male	Mean		6.00	.14
		95% Confidence Interval for Mean	Lower Boun	5.73	
			Upper Boun	6.27	
		5% Trimmed Mean		5.63	
		Median		6.00	
		Variance		30.083	
		Std. Deviation		5.48	
		Minimum		0	
		Maximum		23	
		Range		23	
		Interquartile Range		10.00	
		Skewness		.737	.061
		Kurtosis		-.452	.122
	Female	Mean		6.45	.30
		95% Confidence Interval for Mean	Lower Boun	5.86	
			Upper Boun	7.04	
		5% Trimmed Mean		6.12	
		Median		4.00	
		Variance		36.602	
		Std. Deviation		6.05	
		Minimum		0	
		Maximum		21	
		Range		21	
		Interquartile Range		11.00	
		Skewness		.530	.122
		Kurtosis		-1.244	.243
WAGE	Male	Mean		9.4243	.2721
		95% Confidence Interval for Mean	Lower Boun	8.8906	
			Upper Boun	9.9580	
		5% Trimmed Mean		7.9114	
		Median		6.2500	
		Variance		119.438	
		Std. Deviation		10.9288	
		Minimum		.00	
		Maximum		153.88	
		Range		153.88	
		Interquartile Range		7.6100	
		Skewness		5.166	.061
		Kurtosis		44.173	.122
	Female	Mean		7.5437	.6108
		95% Confidence Interval for Mean	Lower Boun	6.3429	
			Upper Boun	8.7445	
		5% Trimmed Mean		6.1636	
		Median		4.3800	
		Variance		150.363	
		Std. Deviation		12.2623	
		Minimum		.23	
		Maximum		189.39	
		Range		189.16	
		Interquartile Range		7.9600	
		Skewness		9.657	.122
		Kurtosis		129.010	.243

Descriptives of *education* and *wage* by sector

Descriptives

SECTOR				Statistic	Std. Error
EDUCATIC	Private Sector	Mean		4.06	.12
		95% Confidence Interval for Mean	Lower Bou	3.82	
			Upper Bou	4.29	
		5% Trimmed Mean		3.67	
		Median		1.00	
		Variance		18.220	
		Std. Deviation		4.27	
		Minimum		0	
		Maximum		20	
		Range		20	
		Interquartile Range		5.00	
		Skewness		1.228	.068
		Kurtosis		.742	.136
	Public Sector	Mean		9.72	.22
		95% Confidence Interval for Mean	Lower Bou	9.29	
			Upper Bou	10.15	
		5% Trimmed Mean		9.69	
		Median		11.00	
		Variance		34.404	
		Std. Deviation		5.87	
		Minimum		0	
		Maximum		23	
		Range		23	
		Interquartile Range		8.00	
		Skewness		-.196	.091
		Kurtosis		-.910	.181
WAGE	Private Sector	Mean		6.3359	.2609
		95% Confidence Interval for Mean	Lower Bou	5.8241	
			Upper Bou	6.8477	
		5% Trimmed Mean		5.0529	
		Median		4.3800	
		Variance		87.931	
		Std. Deviation		9.3771	
		Minimum		.00	
		Maximum		153.88	
		Range		153.88	
		Interquartile Range		3.3950	
		Skewness		8.411	.068
		Kurtosis		97.830	.136
	Public Sector	Mean		13.8889	.4669
		95% Confidence Interval for Mean	Lower Bou	12.9721	
			Upper Bou	14.8056	
		5% Trimmed Mean		12.4272	
		Median		11.8050	
		Variance		157.859	
		Std. Deviation		12.5642	
		Minimum		.01	
		Maximum		189.39	
		Range		189.38	
		Interquartile Range		8.6350	
		Skewness		5.729	.091
		Kurtosis		61.492	.181

Extreme values (outliers included) of *education* and wage across categories of *sector* and *gender*

Extreme Values

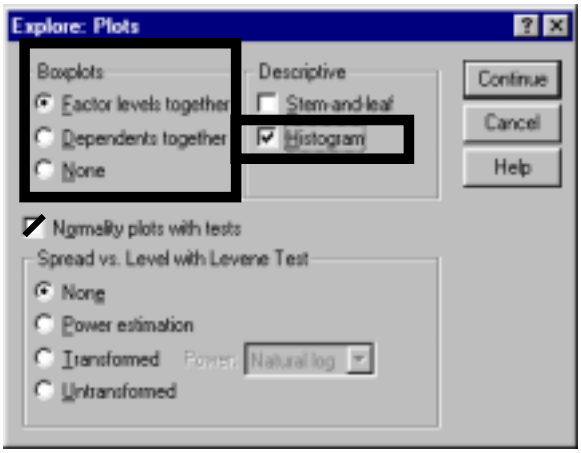
	Whether Public			Case Number	Value
EDUCATION	Private Sector	Highest	1	4	20
			2	993	20
			3	1641	20
			4	1614	19
			5	1629	.[a]
		Lowest	1	222	0
			2	688	0
			3	551	0
			4	709	0
			5	75	.[b]
	Public Sector	Highest	1	1033	23
			2	1034	23
			3	1037	22
			4	1503	22
			5	1035	.[c]
		Lowest	1	1260	0
			2	1582	0
			3	1268	0
			4	1273	0
			5	1278	.[b]
WAGE	Private Sector	Highest	1	5	153.88
			2	3	125.00
			3	2	119.32
			4	1616	101.13
			5	4	75.76
		Lowest	1	731	.00
			2	776	.01
			3	87	.13
			4	1759	.23
			5	237	.[b]
	Public Sector	Highest	1	1876	189.39
			2	1037	119.32
			3	1036	85.23
			4	1039	63.13
			5	1038	60.42
		Lowest	1	1230	.01
			2	1119	.11
			3	2016	.25
			4	2015	.28
			5	1987	.33

a. Only a partial list of cases with the value 19 are shown in the table of upper extremes.

b. Only a partial list of cases with the value 0 are shown in the table of lower extremes.

c. Only a partial list of cases with the value 22 are shown in the table of upper extremes.

In the previous example, we chose no plots. Let us go back and choose some plots.



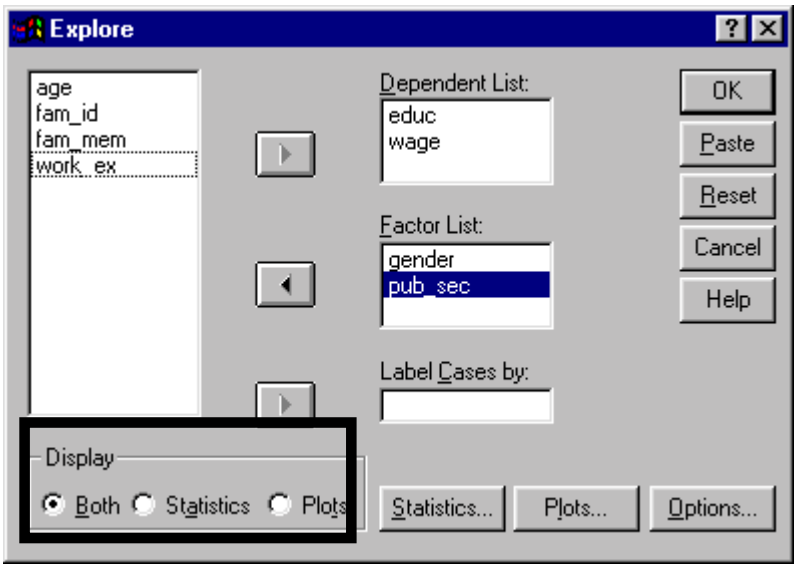
Select "Factor levels together"[73] in the area "Boxplots." Select "Histograms" in the area "Descriptives."

You should check "Normality plots with tests." This will give the Q-Q plots and the K-S test for normality. In the interest of saving space, the output below does not reproduce these charts - see section 3.2 for learning how to interpret the Q-Q and the K-S test for normality.

"Spread vs. Level" is beyond the scope of this book.

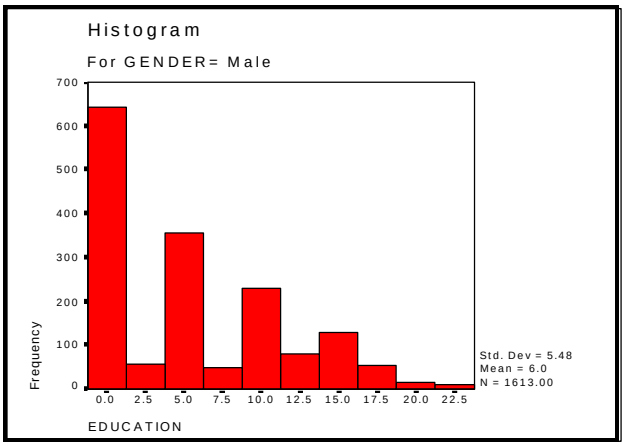
Click on "Continue."

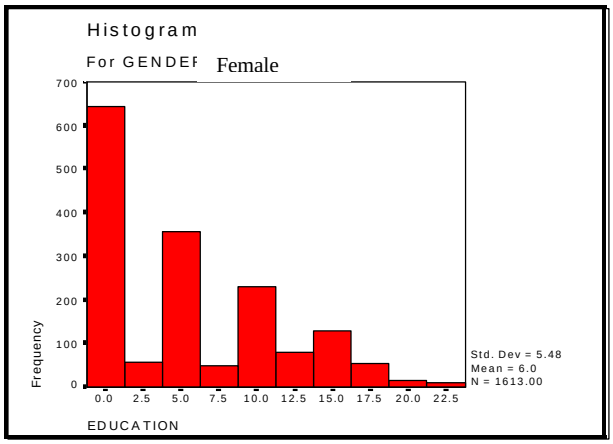
You must choose "Plots" or "Both" in the option area "Display." Click on "OK."

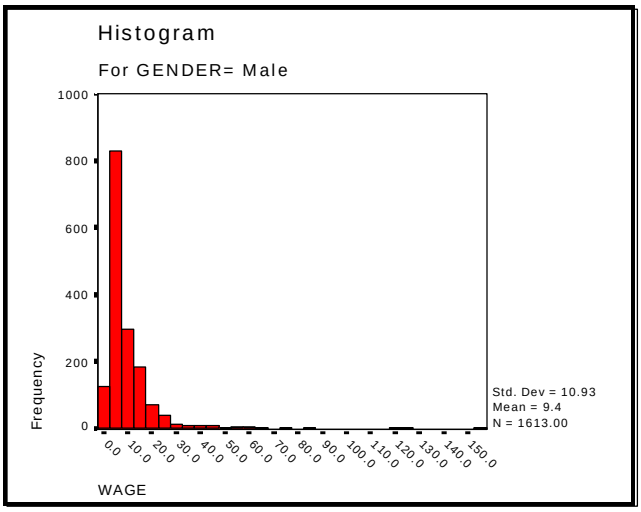


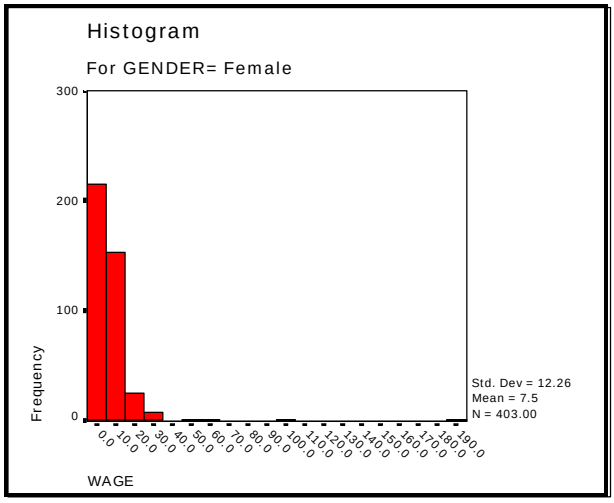
Several charts are drawn, including eight histograms and four boxplots.

[73] Choosing "Dependants together" will change the order in which the plots are displayed in the output.

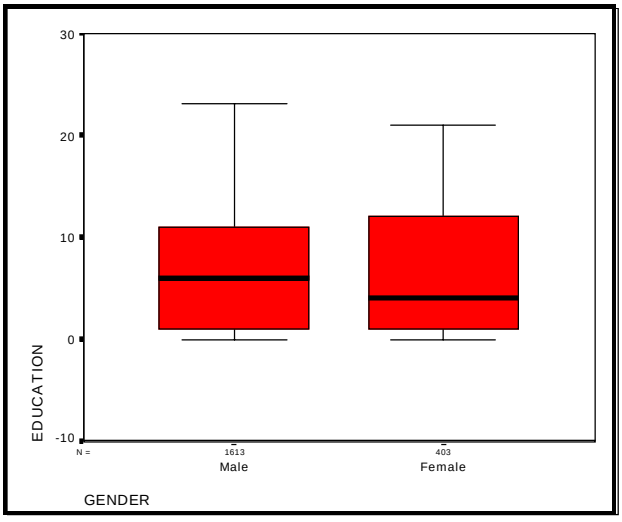


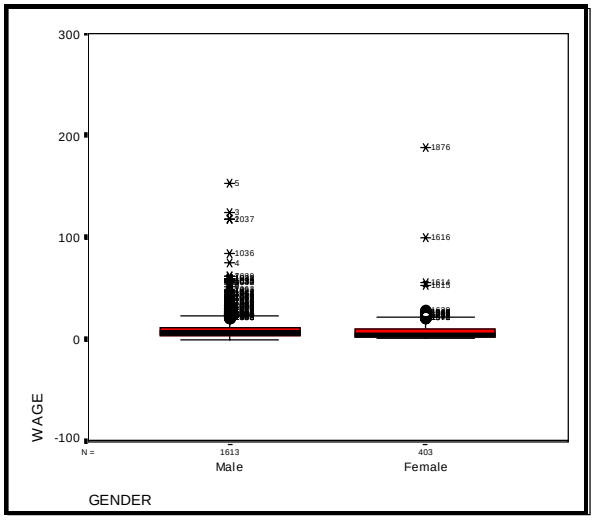






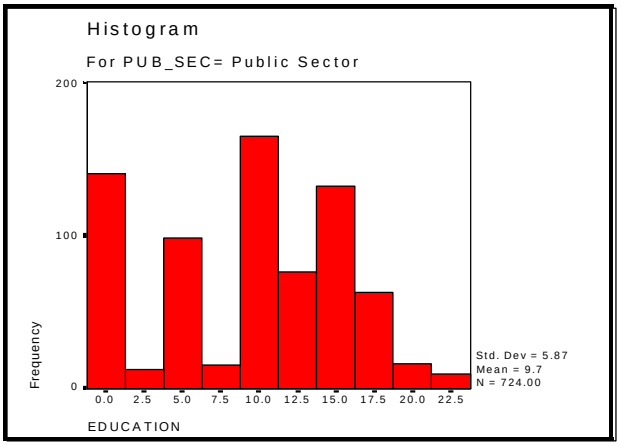
You should re-scale the axis (using procedures shown in section 11.2) to ensure that the bulk of the plot is dominated by the large bars.

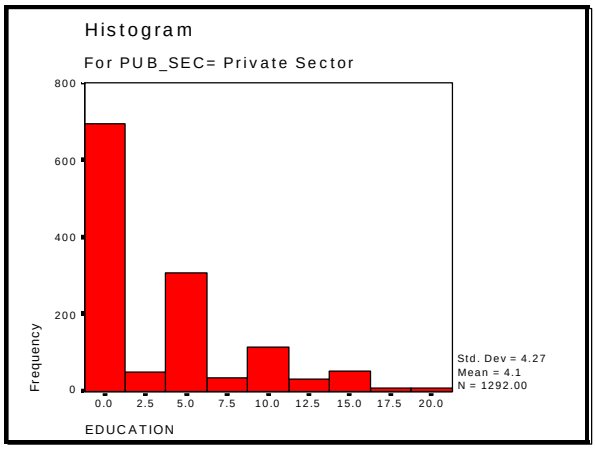


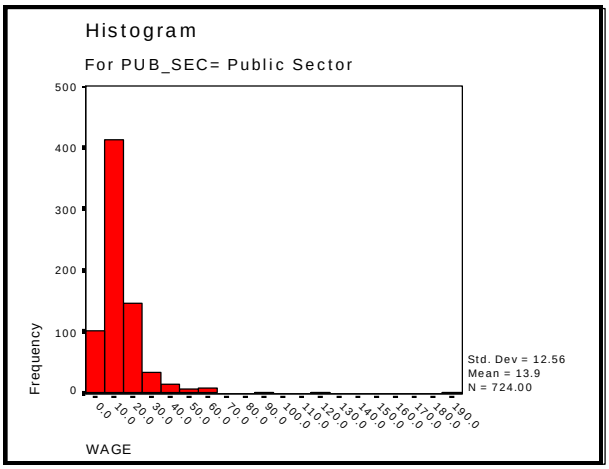


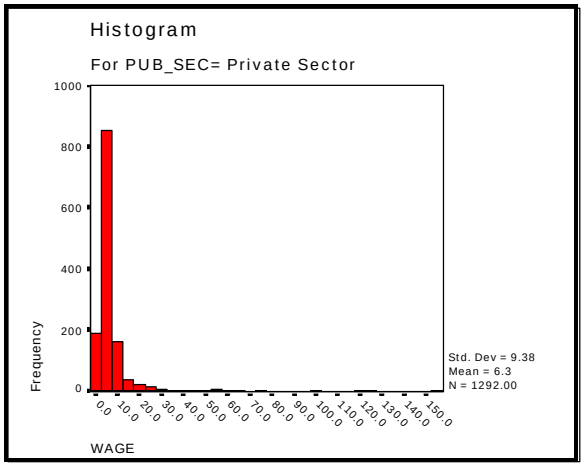
You may want to re-scale the axis so that the boxplot can be seen more clearly. See section

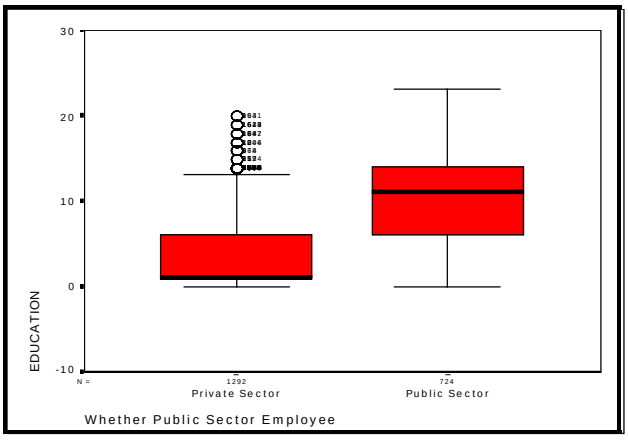
11.2 on the method of rescaling.

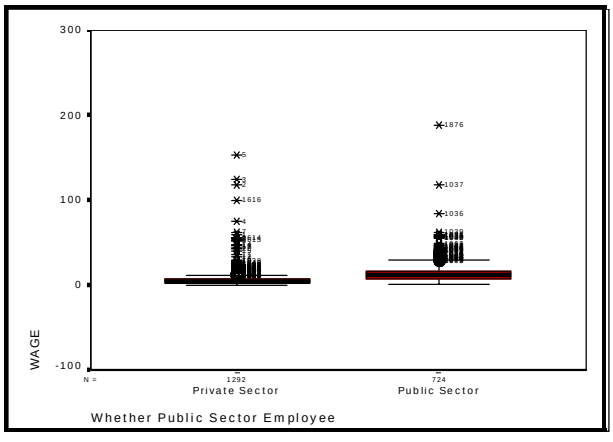












Reminder: The median is the thick horizontal line in the middle of the shaded area The shaded area defines the 75th to 25th percentile range and the outliers are the points above (or below) the "box and whiskers."

Note: in the boxplot on the upper right and the histogram above it, the depictive power of the graph can be increased significantly by restricting the range of X-values (for the histogram) and Y-values for the boxplot. See section 11.2 to learn how to change the formatting.

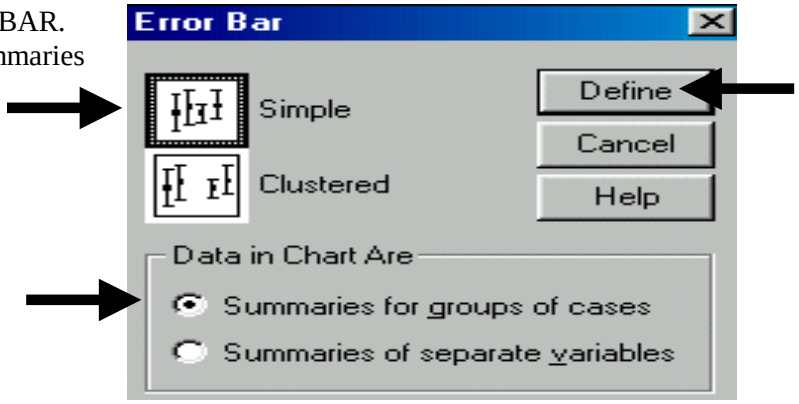
Ch 5. Section 5 Comparing the means and distributions of sub-groups of a variable -- Error Bar, T-Test, ANOVA and Non-Parametric Tests

Ch 5. Section 5.a. Error bars

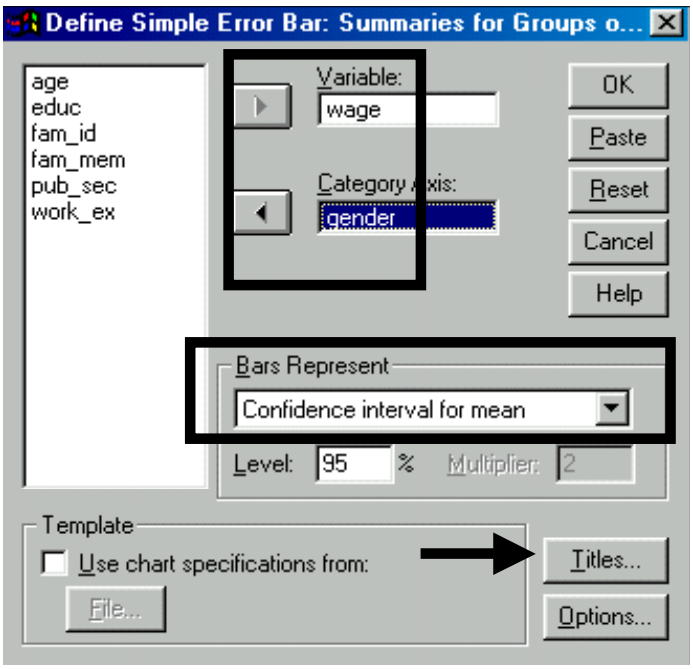
Though a bar graph can be used to determine whether the estimated wage for males is higher than that of females, that approach can be problematic. For example, what if the wage for males is higher but the standard error of the mean is also much higher for males? In that case, saying that "males have a higher wage" may be misleading. It is better to compare and contrast the range within which we can say with 95% confidence that the mean may lie (confidence intervals incorporate both pieces of information - the mean and its standard error). Error bars depict the confidence intervals very well.

Go to GRAPHS / ERROR BAR. Choose "Simple" and "Summaries of Groups of cases."

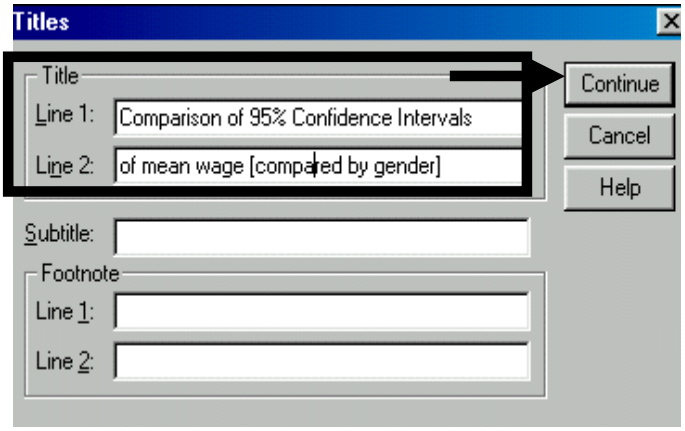
Click on "Define."



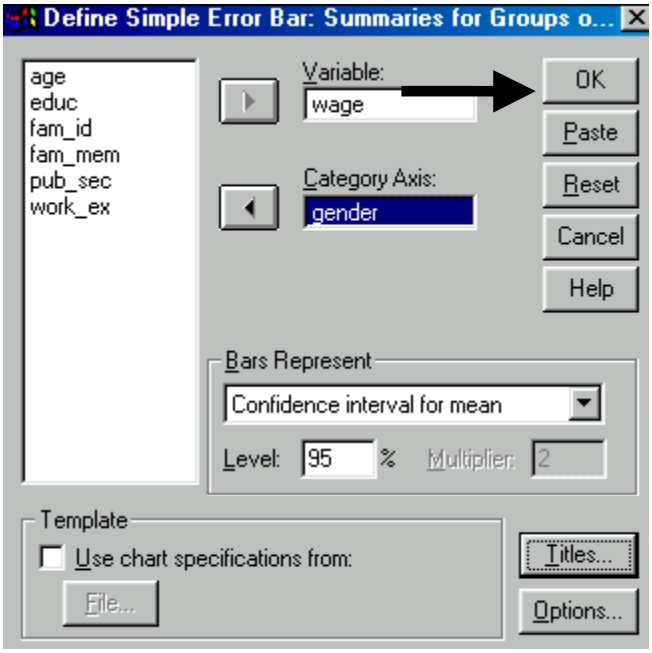
Place the appropriate variable (that whose mean's confidence intervals you wish to determine) into the box "Variable." Place the variable whose categories define the X-axis into the box "Category Axis." Type in the appropriate level of confidence (we recommend 95%).



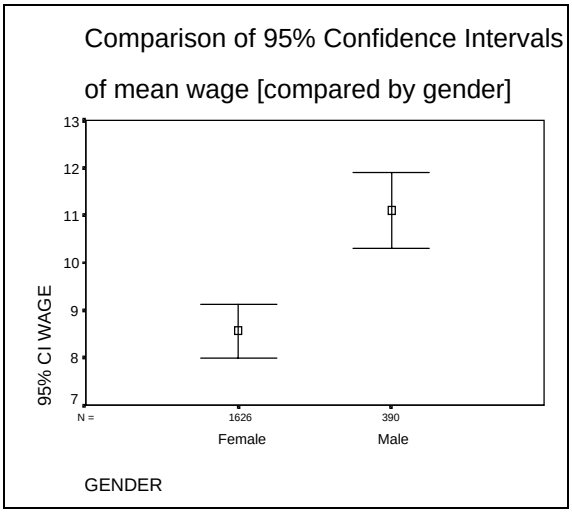
Click on "Titles" and enter an appropriate title. Click on "Continue."



Click on "OK."



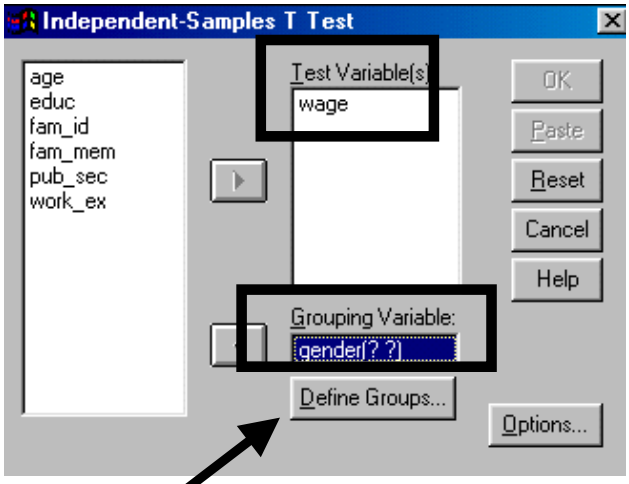
In addition to the mean (the small box in the middle of each error bar) being higher for males, the entire 95% confidence interval is higher for males. This adds great support to any statement on differentials in wages.



Ch 5. Section 5.b. The T-Test for comparing means

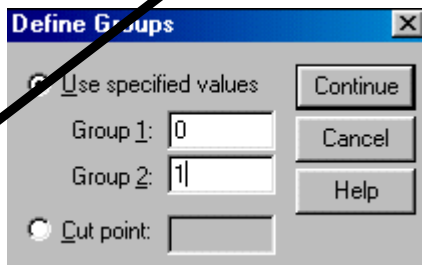
We want to test the hypothesis that the mean *wage* for males is the same as that for females. The simplest test is the "Independent-Samples T Test."

Go to STATISTICS / COMPARE MEANS / INDEPENDENT-SAMPLES T TEST." In the box "Test Variable(s)," move the variable whose subgroups you wish to compare (in our example, the *wage*.) You can choose more than one quantitative variable.



The variable that defines the groups is *gender*. Move it into the box "Grouping Variable."

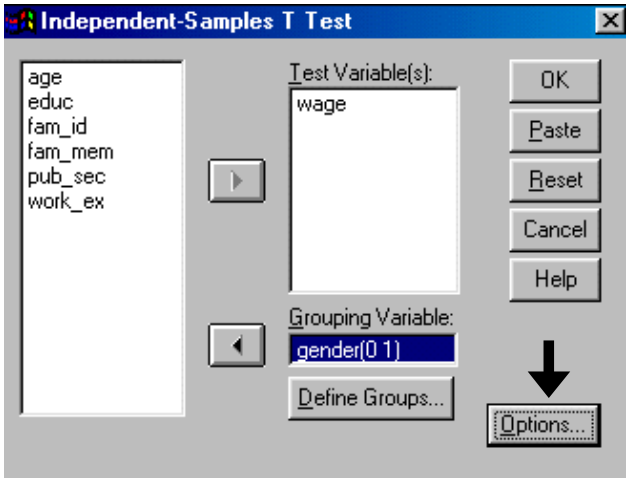
Observe the question marks in the box "Grouping Variable." SPSS is requesting the two values of gender that are to be used as the defining characteristics for each group. Click on "Define Groups."



Enter the values (remember that these numbers, 0 and 1, must correspond to categories of the variable *gender*, i.e. - male and female.)

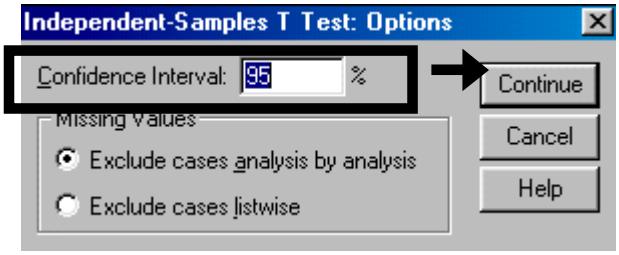
See the option "Cut Point." Let's assume you wanted to compare two groups, one defined by education levels above 8 and the other by education levels below 8. One way to do this would be to create a new dummy variable that captures this situation (using methods shown in sections 2.1 and 1.7). An easier way would be to simply define 8 as the cut point. To do this, click on the button to the left of "Cut Point" and enter the number 8 into the text box provided.

Click on "Options."

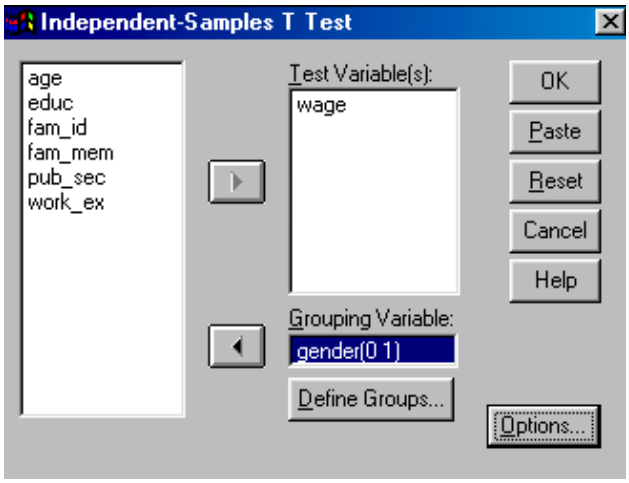


You can choose the confidence interval that the output tables will use as criteria for hypothesis testing.

Click on "Continue."



Click on "OK."



Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Mean	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
WAGE	Equal variances assumed	.717	.397	-4.041	2014	.000	-2.5488	.6308	-3.7858	-1.3117
	Equal variances not assumed			-5.122	856.401	.000	-2.5488	.4976	-3.5254	-1.5721

The interpretation of the output table (above) is completed in five steps:

1. The first three columns test the hypothesis that "the two groups of *wage* observations have the same (homogenous) variances." Because the Sig value for the F is greater than 0.1, we fail to reject the hypothesis (at the 90% confidence level) that the variances are equal.

2. The F showed us that we should use the row "Equal variances assumed." Therefore, when looking at values in the 4th to last columns (the T, Sig, etc.), use the values in the 1st row (i.e. - the row that has a T of –4.04. In the next table we have blanked out the other row).

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Mean	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
WAGE	Equal variances assumed	.717	.397	-4.041	2014	.000	-2.5488	.6308	-3.7858	-1.3117
	Equal variances not assumed			-5.122	856.401	.000	-2.5488	.4976	-3.5254	-1.5721

3. Find whether the T is significant. Because the "Sig (2-tailed)" value is below .05, the coefficient is significant at 95% confidence.

4. The "coefficient" in this procedure is the difference in mean wage across the two groups. Or stated differently, Mean (*wage* for gender=1 or female) – Mean(*wage* for gender=0 or male). The mean difference of –2.54 implies that we can say, with 95% confidence, that "the mean *wage* for males is –2.54 higher than that for females.

5. The last two columns provide the 95% confidence interval for this difference in mean. The interval is (-3.78, -1.31).

Let's assume you have a variable with three values - 0, 1, and 2 (representing the concepts "conservative," "moderate," and "liberal"). Can you use this variable as the grouping variable, i.e. - first compare across "conservative" and "moderate" by using the values 0 and 1 in the "Define Groups" dialog box, then compare "conservative" to "liberal" by using the values 0 and 2 in the same dialog box? The answer is no, one cannot break up a categorical variable into pairs of groups and then use the "Independent Samples T Test." Certain biases are introduced into the procedure if such an approach is employed. We will not get into the details of these biases, for they are beyond the scope of this book. However, the question remains - If the "Independent Samples T Test" cannot be used, what should be used? The answer is the ANOVA. In the next section we show an example of a simple "One-Way ANOVA."

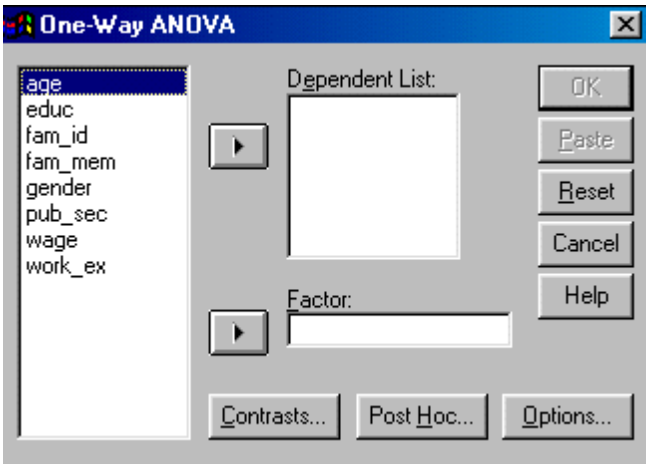
One can argue, correctly, that the T or F tests cannot be used for testing a hypothesis about the variable *wage* because the variable is not distributed normally - see section 3.2. Instead, non-parametric methods should be used - see section 5.5.d. Researchers typically ignore this fact and proceed with the T-Test. If you would like to hold your analysis to a higher standard, use the relevant non-parametric test shown in section 5.5.d.

Ch 5. Section 5.c. ANOVA

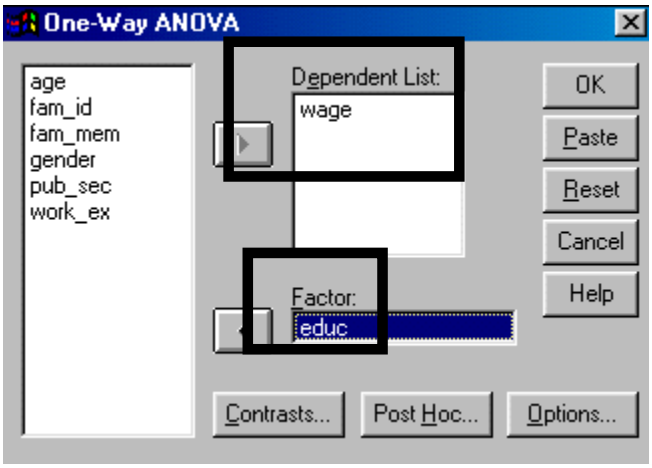
Let's assume you want to compare the mean *wage* across *education* levels and determine whether it differs across these various levels. The variable *education* has more than two values, so you therefore cannot use a simple T-Test. An advanced method called ANOVA (Analysis of Variance) must be used. We will conduct a very simple case study in ANOVA.

ANOVA is a major topic in itself, so we will show you only how to conduct and interpret a basic ANOVA analysis.

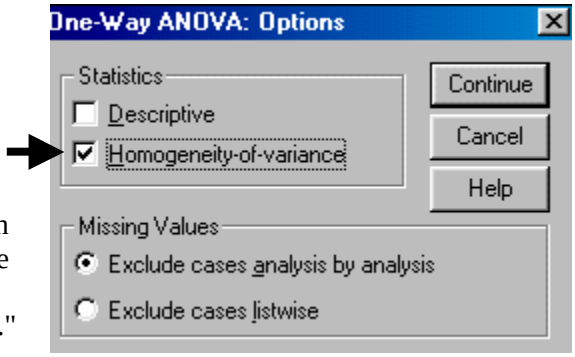
Go to STATISTICS / MEANS / 1-WAY ANOVA



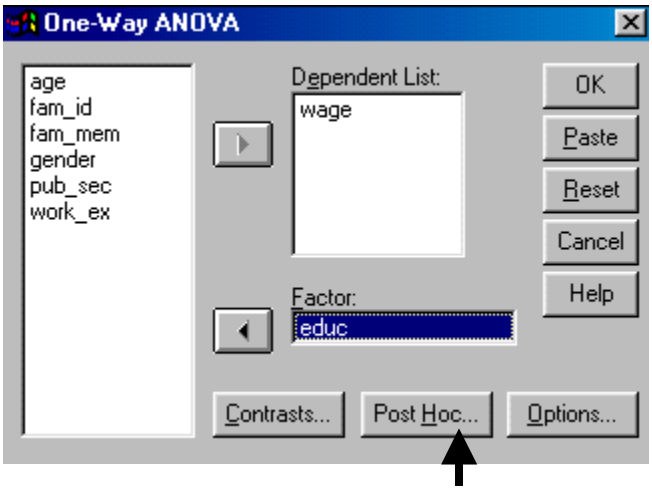
We want to see if the mean *wage* differs across *education* levels. Place the variable wage into the box "Dependent List" and education into "Factor" (note: you can choose more than one dependent variable).



ANOVA runs different tests for comparisons of means depending on whether the variances across sub-groups of wage (defined by categories of education) differ or are similar. Therefore, we first must determine, via testing, which path to use. To do so, click on the button "Options" and choose the option "Homogeneity of variance." Click on "Continue."



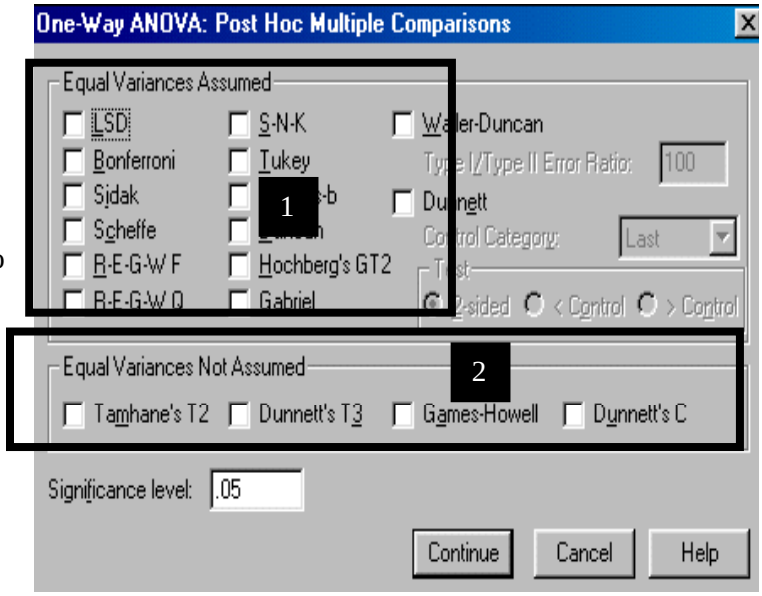
Now we must choose the method for testing in the event that the means are different. Click on "Post Hoc" (we repeat - our approach is basic).



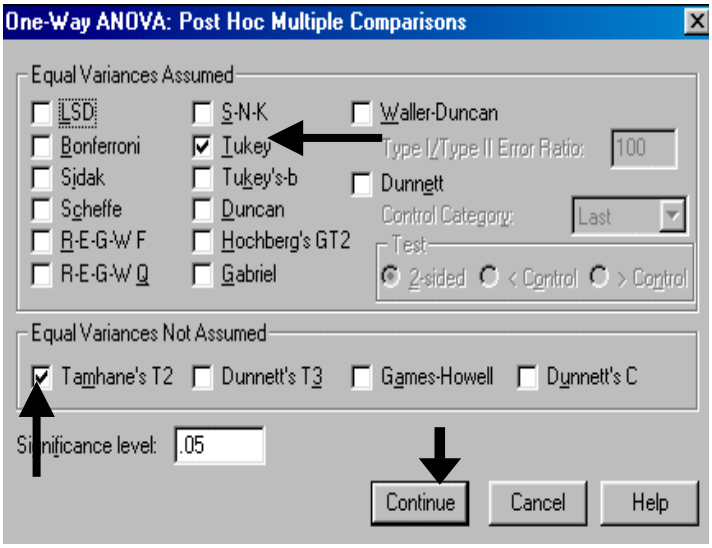
Note: In your textbook, you may encounter two broad approaches for choosing the method - *a priori* and a *posteriori*. *A priori* in this context can be defined as "testing a hypothesis that was proposed before any of the computational work." In contrast, *a posteriori* can be defined as "testing a hypothesis that was proposed after the computational work." For reasons beyond the scope of this book, *a posteriori* is regarded as the approach that is closer to real research methods. "Post Hoc" is synonymous with *a posteriori*.

Area 1 allows the user choices of tests to use in the event that the variances between sub-groups of wage (as defined by categories of education) are found to be equal (note: this is very rarely the case). There are many options. Consult your textbook for the best option to use.

Area 2 asks for choices of tests to use if the variances between sub-groups of wage (as defined by categories of education) are not found to be equal.

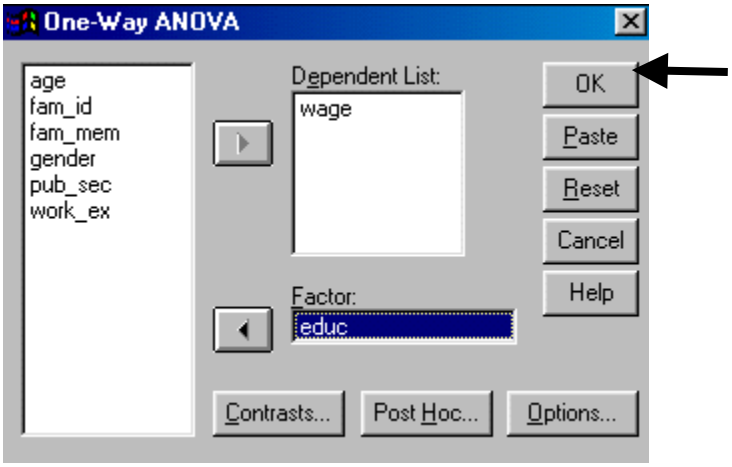


We chose to use "Tukey" and "Tamhane's T2" because they are the most used test statistics by statisticians. SPSS will produce two tables with results of mean comparisons. One table will be based on "Tukey" and the other on "Tamhane's T2." How does one decide which to use? In the output, look for the "Test of Homogenity of Variances." If the Sig value is significant (less than .1 for 90% confidence level), then the variances of the subgroups are not homogenous. Consequently, one should use the numbers estimated using "Tamhane's T2."



Click on "Continue."

Click on "OK."



Note that we had asked for testing if the variances were homogenous across sub-groups of *wage* defined by categories of *education*. The Sig value below shows that the hypothesis of homogeneity can not be accepted. Heterogeneity is assumed as correct. "Tamhane's" method for comparing means should therefore be used

Test of Homogeneity of Variances

	Levene Statistic	df1	df2	Sig.
WAGE	8.677	22	1993	.000

The ANOVA table below tests whether the difference between groups (i.e. - the deviations in wages explained by differences in *education* level)[74] is significantly higher than the deviations within each education group. The Sig value indicates that the "Between Groups" variation can explain a relatively large portion of the variation in wages. As such, it makes sense to go further and compare the difference in mean *wage* across education levels (this point is more clear when the opposite scenario is encountered). If the "Between Groups" deviations' relative importance is not so large, i.e. - the F is not significant, then we can conclude that differences in *education* levels do not play a major role in explaining deviations in *wages*.

[74] The between groups sum of squares is, computationally, the sum of squares obtained if each group were seen as one "observation," with this "observation" taking on the value of the mean of the group.

Note: The "analysis" of variance is a key concept in multivariate statistics and in econometrics. A brief explanation: the sum of squares is the sum of all the squared deviations from the mean. So for the variable *wage*, the sum of squares is obtained by:
[a] obtaining the mean for each group.
[b] re-basing every value in a group by subtracting the mean from this value. This difference is the "deviation."
[c] Squaring each deviation calculated in "b" above.
[d] Summing all the squared values from "c" above. By using the "squares" instead of the "deviations," we permit two important aspects. When summing, the negative and positive deviations do not cancel each other out (as the squared values are all positive) and more importance is given to larger deviations than would be if the non-squared deviations were used (e.g. - let's assume you have two deviation values 4 and 6. The second one is 1.5 times greater than the first. Now square them. 4 and 6 become 16 and 36. The second one is 2.25 times greater than the first).

ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
WAGE	Between Groups	78239.146	22	3556.325	40.299	.000
	Within Groups	175880.9	1993	88.249		
	Total	254120.0	2015			

This shows that the sub-groups of wage (each sub-group is defined by an education level) have unequal (i.e. - heterogeneous) variances and, thus, we should only interpret the means-comparison table that uses a method (here "Tamhane's T2") that assumes the same about the variances.

SPSS will produce tables that compares the means. One table uses "Tukeys" method; the other will use "Tamhane's" method. We do not reproduce the table here because of size constraints.

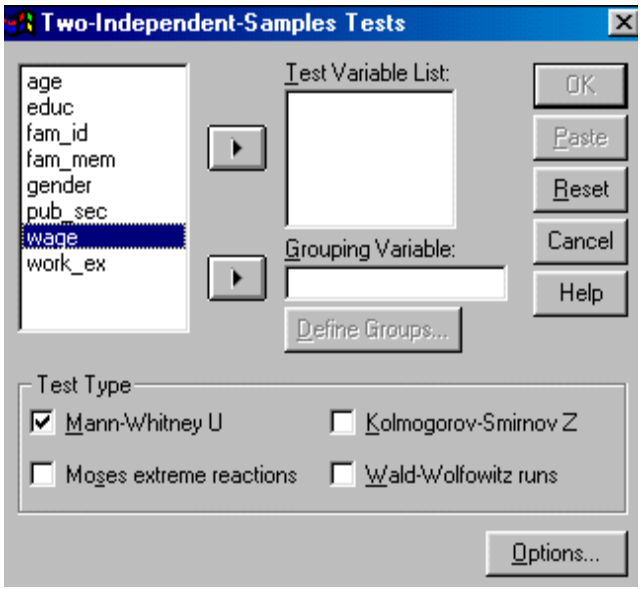
Rarely will you have to use a method that assumes homogenous variances. In our experience, real world data typically have heterogeneous variances across sub-groups.

Ch 5. Section 5.d. Nonparametric testing methods

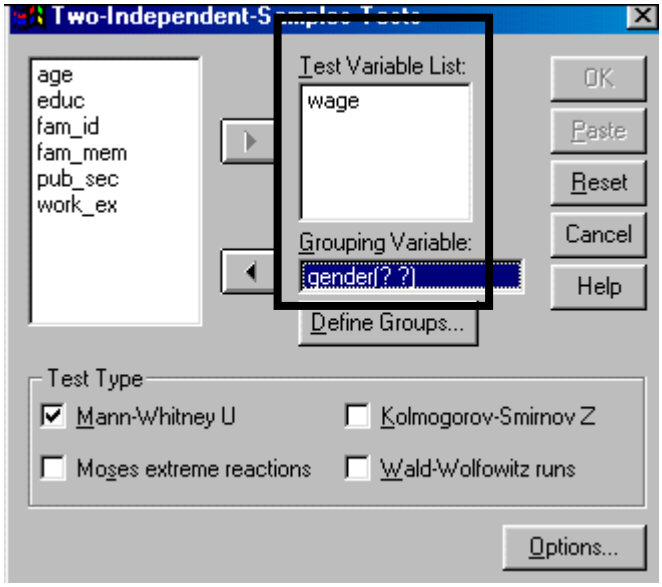
Let's assume the histogram and P-P showed that the variable wage is not distributed normally. Can we still use the method shown in section 5.5.b? Strictly speaking, the answer is "No." In recent years, some new "Non-parametric" testing methods have been developed that do not assume underlying normality (a test/method that must assume specific attributes of the underlying distribution is, in contrast, a "Parametric" method). We used one such method in section 5.3.c. We show its use for comparing distributions.

Go to STATISTICS / NONPARAMETRIC TESTS / TWO-INDEPENDENT SAMPLES TESTS.

Basically it tests whether the samples defined by the each category of the grouping variable have different distribution attributes. If so, then the "Test Variable" is not independent of the "Grouping Variable." The test does not provide a comparison of means.

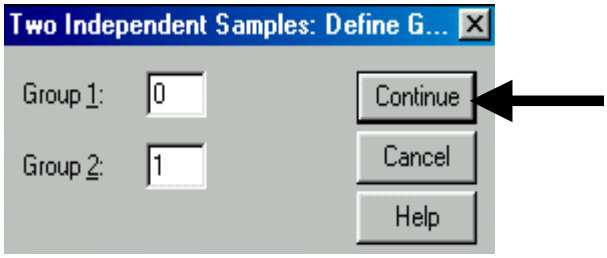


Place the variables into the appropriate boxes.



Click on "Define Groups." In our data, *gender* can take on two values - 0 if male and 1 if female. Inform SPSS of these two values.

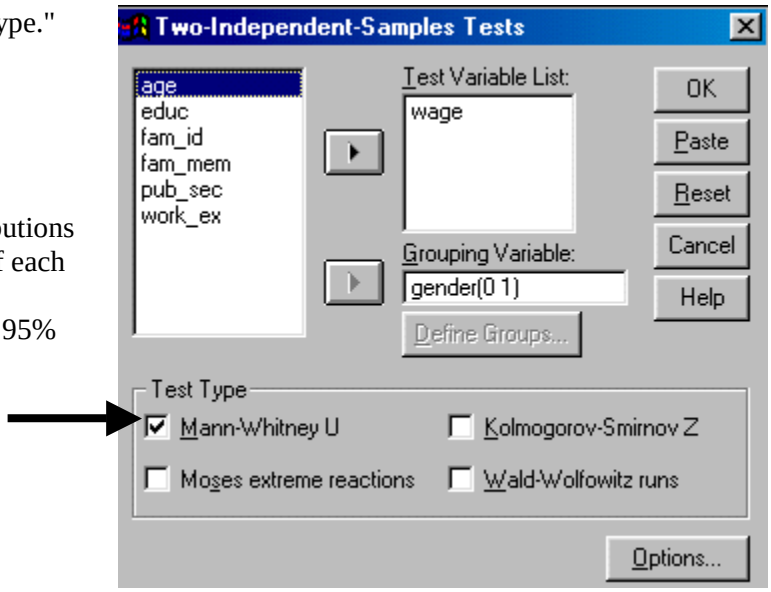
Click on "Continue."



Choose the appropriate "Test Type." The most used type is "Mann-Whitney[75]."

Click on "OK."

The results show that the distributions can be said to be independent of each other and different (because the "Asymp. Sig" is less than .05, a 95% confidence level).



Test Statistics[a]

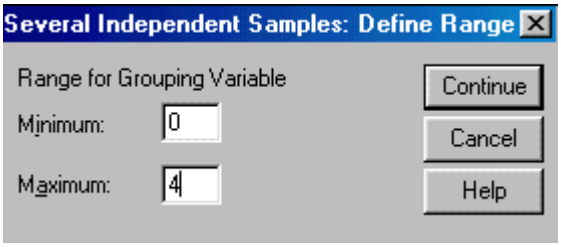
	WAGE
Mann-Whitney U	211656.5
Wilcoxon W	1534408
Z	-10.213
Asymp. Sig. (2-tailed)	.000

a. Grouping Variable: GENDER

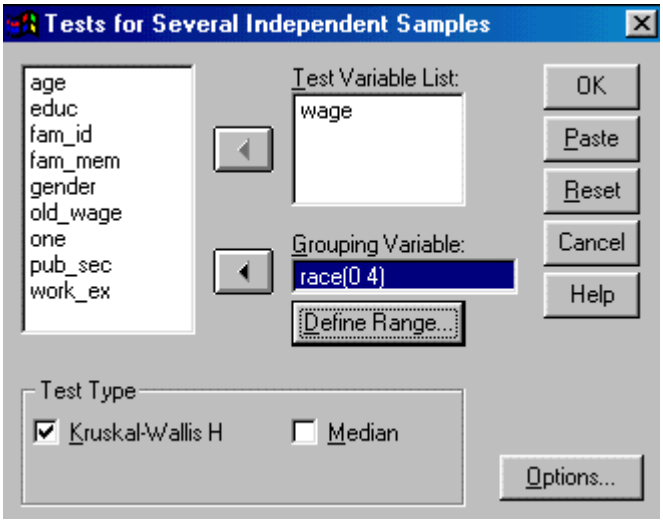
Note: if you have several groups, then use STATISTICS / NONPARAMETRIC TESTS / K [SEVERAL]INDEPENDENT SAMPLES TESTS. **In effect, you are conducting the non-parametric equivalent of the ANOVA**. Conduct the analysis in a similar fashion here, but with two exceptions:

1. Enter the range of values that define the group into the box that is analogous to that on the right. For example:

[75] An explanation of the differences between these test types is beyond the scope of this book.



2. Choose the "Kruskal-Wallis H test" as the "Test type" unless the categories in the grouping variable are ordered (i.e. - category 4 is better/higher than category 1, which is better/higher than category 0).



To take quizzes on topics within each chapter go to http://www.spss.org/wwwroot/spssquiz.asp

Ch 6. TABLES

In this chapter, you will learn how to extend your analysis to a disaggregated level by making tables (called "Custom Tables"). SPSS can make excellent, well-formatted tables with ease.

Tables go one step further than charts[76]: they enable the production of numeric output at levels of detail chosen by the user. Section 6.1 describes how to use custom tables to examine the patterns and values of statistics (i.e. - mean, median, standard deviation, etc.) of a variable across categories/values of other variables.

Section 6.2 describes how to examine the frequencies of the data at a disaggregated level. Such an analysis complements and completes analysis done in section 6.1.

For understanding Multiple Response Sets and using them in tables, refer to section 2.3 after reading this chapter.

Note: the SPSS system on your computer may not include the Custom Tables procedures.

Ch 6. Section 1 Tables for statistical attributes

Tables are useful for examining the "Mean/Median/other" statistical attribute of one or more variables Y across the categories of one or more "Row" variables X and one or more "Column" variables Z.

If you are using Excel to make tables, you will find the speed and convenience of SPSS to be a comfort. If you are using SAS or STATA to make tables, the formatting of the output will be welcome.

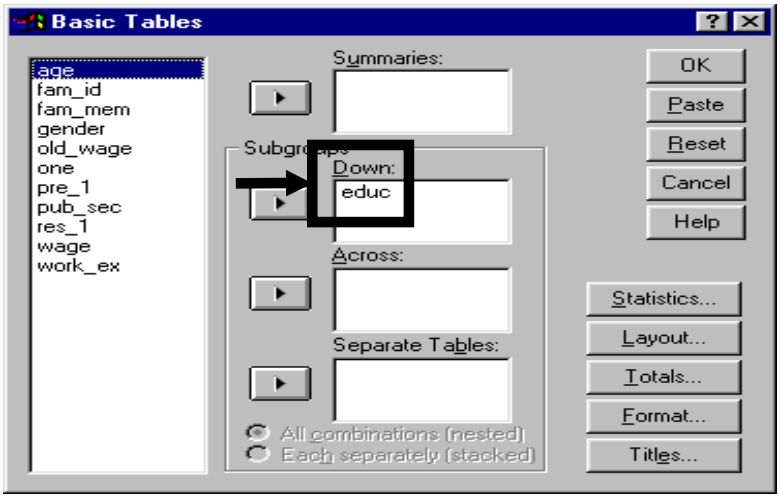
Ch 6. Section 1.a. Summary measure of a variable

Example: making a table to understand the relations/patterns between the variables *wage*, *gender*, and *education* - what are the attributes of *wage* at different levels of *education* and how do these attributes differ across *gender*[77]?

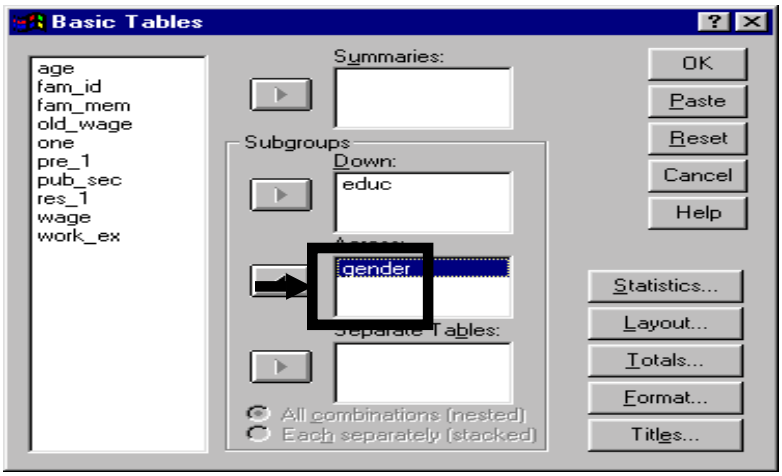
[76] The power of graphs is that the patterns are easily viewed. However, once you want to delve deeper into the data, you want numeric information in addition to simple visual depictions.

[77] Are the patterns the same for higher education levels? Does the pattern reverse itself for certain gender-age combinations? Questions like these can be answered using custom tables. Interpretation of these tables also strengthens one's understanding of the forces driving all the results in the analysis.

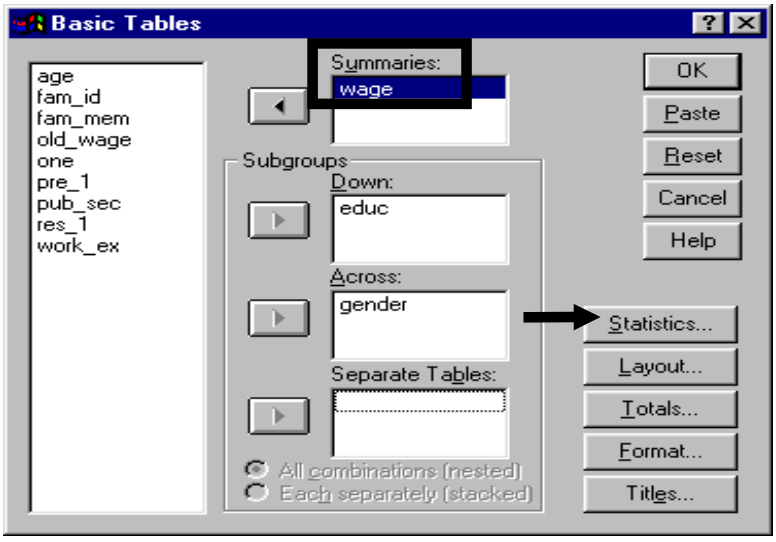
Go to STATISTICS/CUSTOM TABLES[78].



Place *education* into the box "Down." The rows of the table will be levels of *education*.



Place *gender* into the box "Across." The columns of the table will be based on the values of gender.

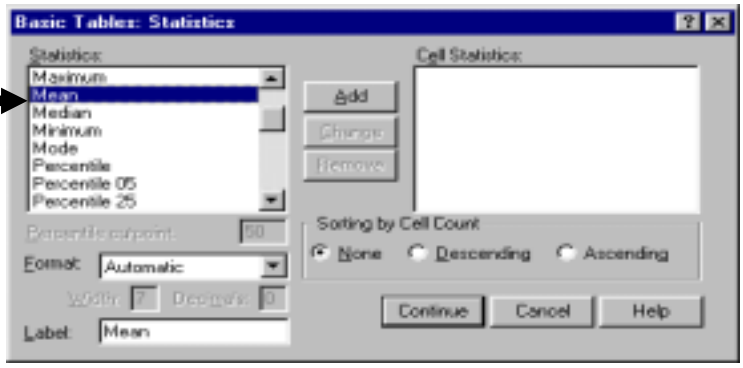


Place *wage* into the box "Summaries." This implies that the data in the cells of the table will be one or more statistic of *wage*.

The next step is to choose the statistic(s) to be displayed in the table. To do so, click on "Statistics."

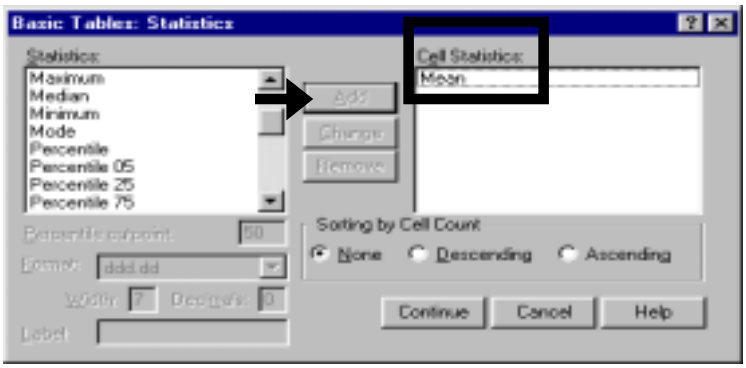
[78] Note: the base SPSS installed in your computer system may not include the Custom Tables procedures.

In the list of the left half of the box, click on the statistic you wish to use. In this example, the statistic is the mean.



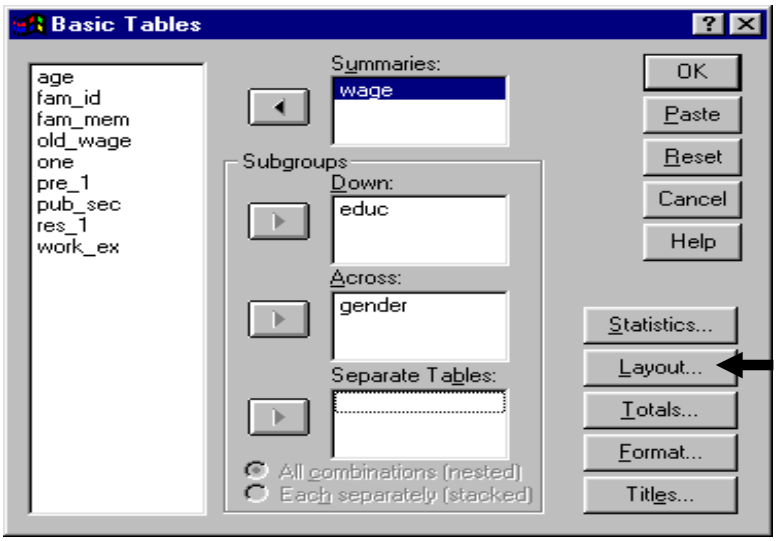
Click on "Add" to choose the statistic.

Click on "Continue."



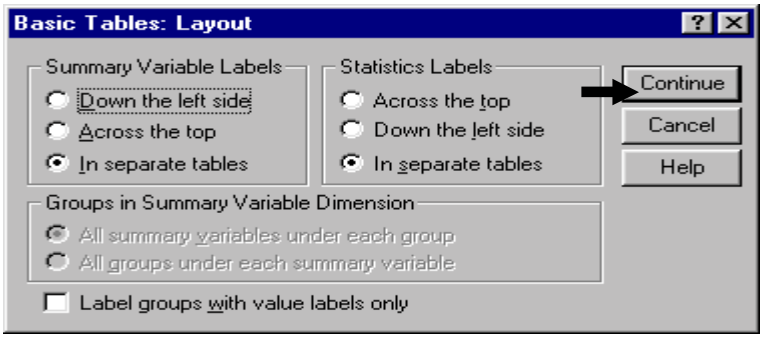
Click on the button "Layout."

Layouts help by improving the layout of the labels of the rows and columns of the custom table.

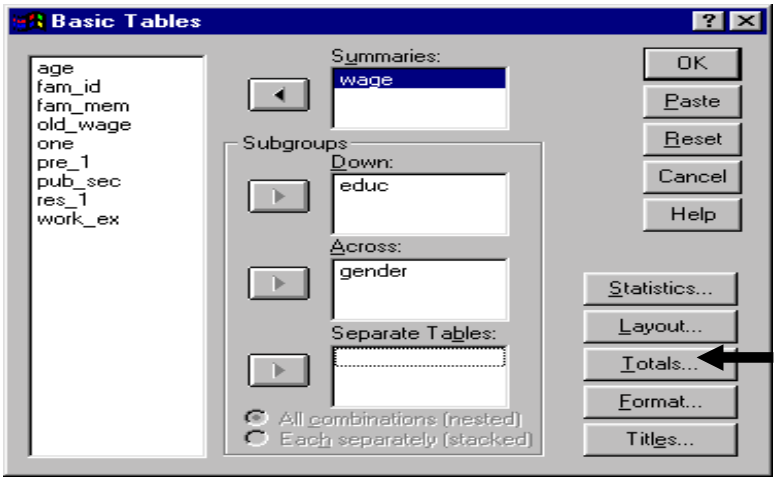


Select the options as shown. We have chosen "In Separate Tables" to obtain lucid output. Otherwise, too many labels will be produced in the output table.

Click on "Continue."



Click on the button "Totals."



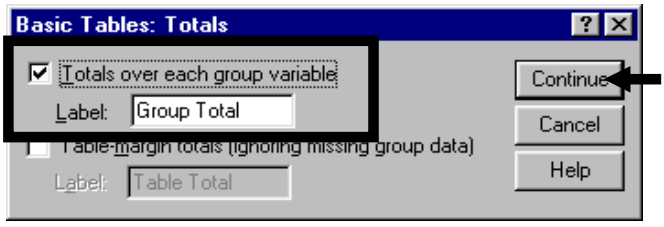
"Totals" enable you to obtain a macro-level depiction of the data.

Data are effectively displayed at three levels of aggregation:

- at the lowest level, where each value is for a specific education-gender combination (these constitute the inner cells in the table on page 6-6),
- at an intermediate level, where each value is at the level of either of the two variables[79] (these constitute the last row and column in the table on page 6-6), and
- at the aggregate level, where one value summarizes all the data (in the last, or bottom right, cell in the table on page 6-6)[80].

You should request totals for each group[81].

Click on "Continue."

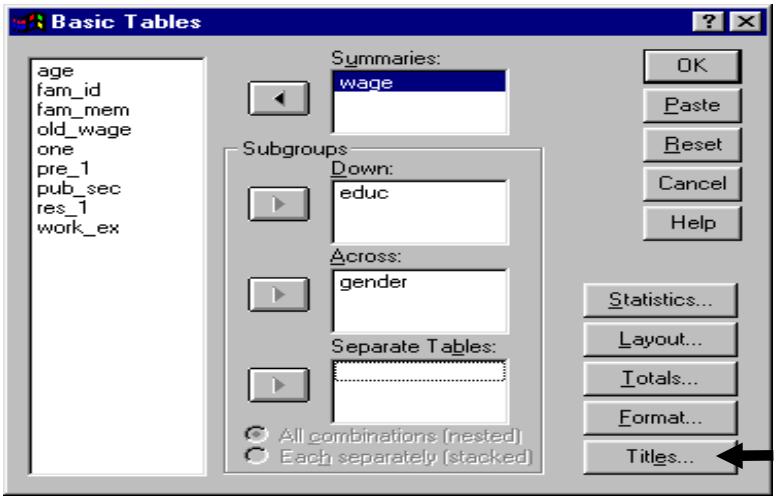


[79] For example, "For all males," "For everyone with education level X," etc.

[80] The three levels of aggregation will become apparent when you look at the output table.

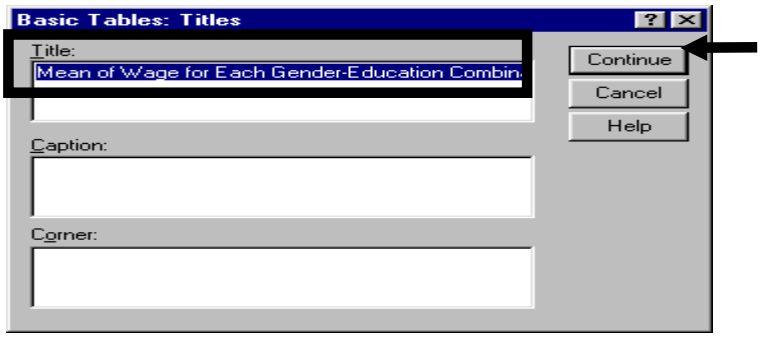
[81] Mean for all females and for all males irrespective of their education levels - in the last row, and means for each education level irrespective of gender - in the last column.

Click on the button "Titles."



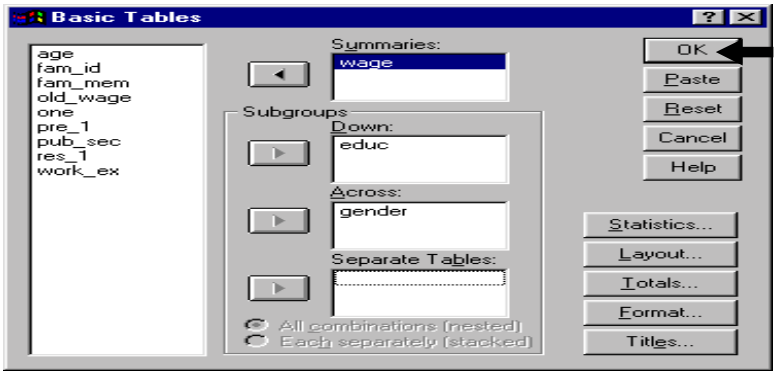
Enter a title for the table.

Click on "Continue."



Click on OK.

The table is shown on the next page. From this table, you can read the numbers of interest. If you are interested in the wages of females who went to college, then look in rows "*education*=13-16" and column "*gender*=1."



Mean of Wage for Each Gender-Education Combination

WAGE Mean

		Male	Female	Group Total
EDUCATION	0	5.97	3.04	5.82
	1	5.32	3.03	4.72
	2	4.32	6.82	4.65
	3	5.49	4.23	5.36
	4	5.25	3.25	4.95
	5	5.50	3.51	5.34
	6	6.78	3.71	6.40
	8	7.85	4.97	7.79
	9	10.05	9.59	9.99
	10	10.78	6.68	10.25
	11	12.47	9.63	11.85
	12	12.09	10.96	11.85
	13	13.30	11.99	12.87
	14	15.77	11.36	14.47
	15	13.81	14.31	13.99
	16	17.09	12.87	15.64
	17	22.09	16.40	20.29
	18	24.72	17.42	23.23
	19	37.97	16.33	25.61
	20	33.00	26.33	31.89
	21	23.63	26.14	24.13
	22	37.50	.	37.50
	23	.	.	.
Group Total		8.63	6.62	8.23

The "Total" for all the data

You can also compare across cells to make statements like "males with only a high school education earn more, on average, than females who completed two years of college[82]."

Another interesting fact emerges when one looks at females with low levels of education: "females with 2 years of education earn more than females with 3-8 years of education and more than men with up to 5 years of education." Why should this be? Is it because the number of females with 2 years of education is very small and an outlier is affecting the mean? To understand this, you may want to obtain two other kinds of information - the medians (see section 6.1.b) and the frequency distribution within the cells in the table (see section 6.2).

Ch 6. Section 1.b. Obtaining more than one summary statistic

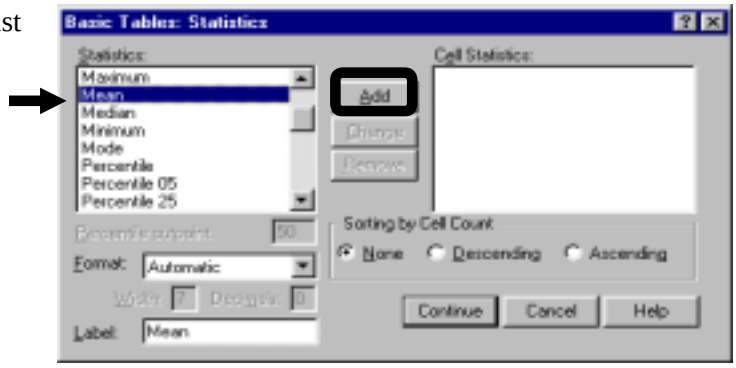
We will repeat the example in section 6.1.a with one exception: we will choose mean and median as the desired statistics. Follow the steps in section 6.1.a except, while choosing the statistics, do the following:

Click on "Statistics"

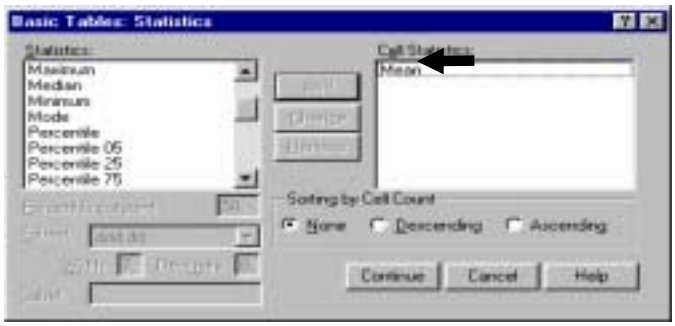
[82] Compare the value in row "*education* = 12" and column "*gender* = 0 (male)" to the value in cell "*education* = 14" and column "*gender*=1 (female)." The double-tailed arrow points to these values.

Click on the first statistic you want in the list in the left half of the box. In this example, the first statistic is the mean.

Click on "Add" to choose this statistic.

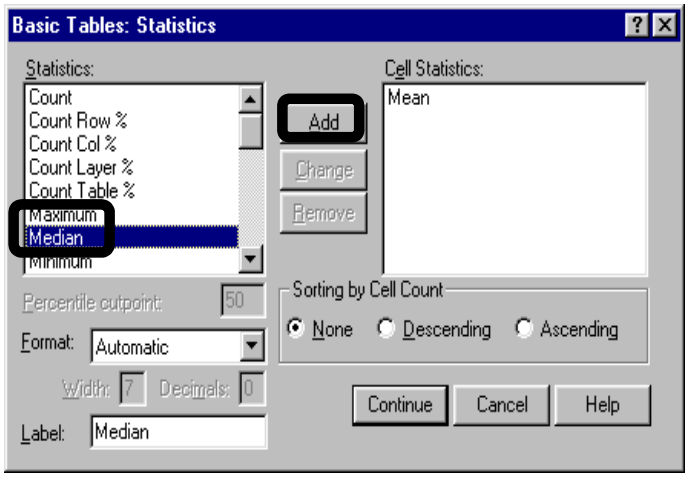


The statistic chosen is displayed in the window "Cell Statistics."

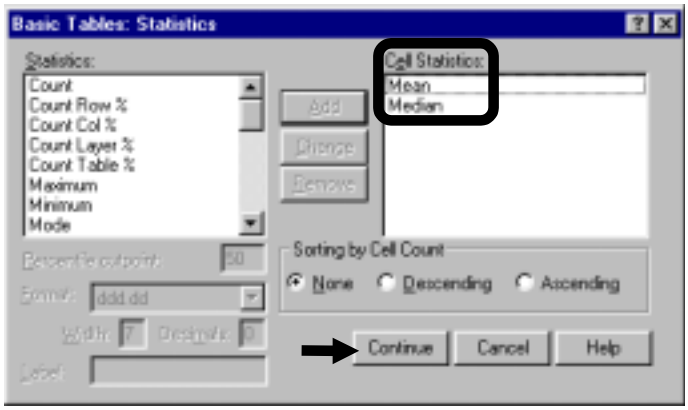


Click on the second statistic you want in the list in the left half of the box. In this example, the second statistic is the median.

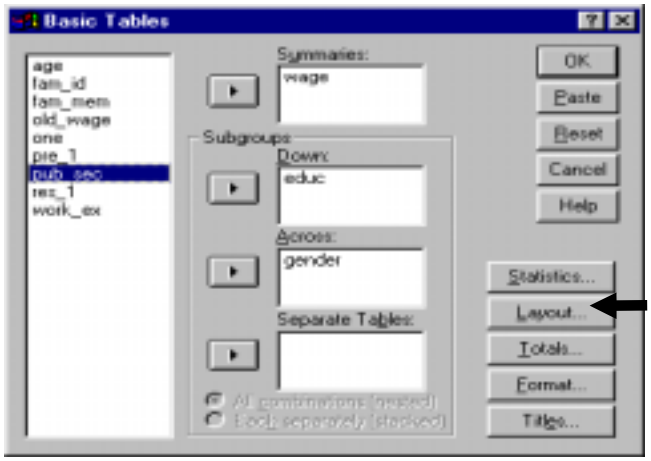
Click on the button "Add."



Click on "Continue."

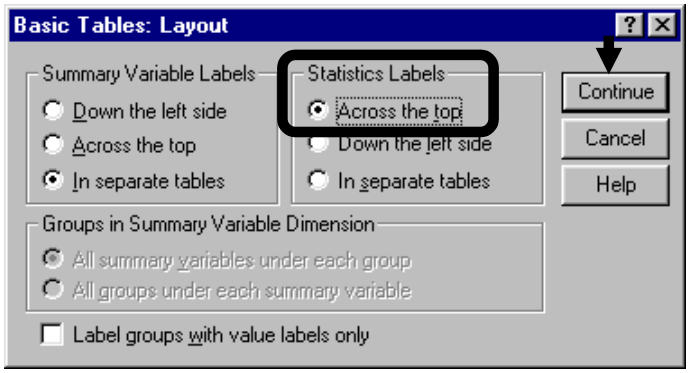


You will need an indicator to distinguish between mean and median in the table. For that, click on "Layout."

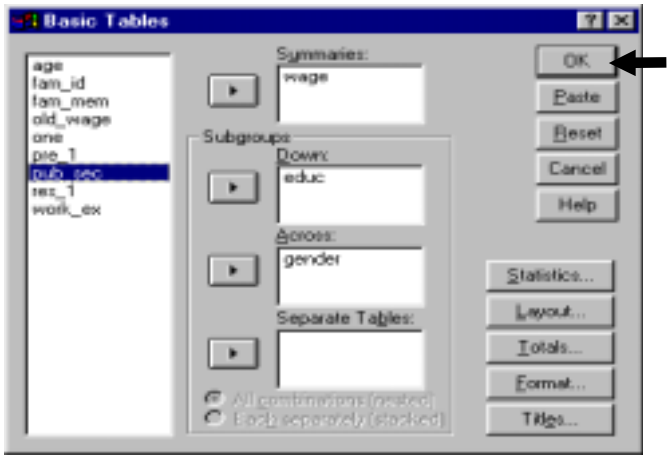


Select "Across the Top" in the options area "Statistics Labels." This will label the mean and median columns. Try different layouts until you find that which produces output to your liking.

Click on "Continue."



Click on "OK."



Mean and Median Wage for Each Gender-Education Combination

		GENDER				Group Total	
		Male		Female			
		Mean	**Median**	**Mean**	**Median**	**Mean**	**Median**
EDUCATION	0	*5.97*	*5.00*	*3.04*	*3.41*	*5.82*	*5.00*
	1	*5.32*	*4.38*	*3.03*	*2.84*	*4.72*	*3.76*
	2	*4.32*	*4.03*	*6.82*	*6.82*	*4.65*	*4.26*
	3	*5.49*	*4.75*	*4.23*	*3.69*	*5.36*	*4.75*
	4	*5.25*	*4.00*	*3.25*	*3.13*	*4.95*	*3.87*
	5	*5.50*	*4.50*	*3.51*	*3.14*	*5.34*	*4.38*
	6	*6.78*	*5.76*	*3.71*	*3.13*	*6.40*	*5.68*
	8	*7.85*	*6.54*	*4.97*	*4.97*	*7.79*	*6.25*
	9	*10.05*	*10.18*	*9.59*	*8.88*	*9.99*	*10.13*
	10	*10.78*	*8.52*	*6.68*	*7.95*	*10.25*	*8.24*
	11	*12.47*	*11.86*	*9.63*	*9.09*	*11.85*	*10.80*
	12	*12.09*	*11.55*	*10.96*	*9.75*	*11.85*	*10.26*
	13	*13.30*	*12.05*	*11.99*	*12.14*	*12.87*	*12.10*
	14	*15.77*	*13.18*	*11.36*	*11.91*	*14.47*	*12.50*
	15	*13.81*	*13.64*	*14.31*	*14.20*	*13.99*	*14.20*
	16	*17.09*	*15.45*	*12.87*	*13.07*	*15.64*	*13.53*
	17	*22.09*	*20.74*	*16.40*	*16.85*	*20.29*	*19.03*
	18	*24.72*	*21.97*	*17.42*	*18.61*	*23.23*	*20.83*
	19	*37.97*	*41.76*	*16.33*	*15.33*	*25.61*	*22.73*
	20	*33.00*	*38.83*	*26.33*	*26.33*	*31.89*	*32.58*
	21	*23.63*	*23.68*	*26.14*	*26.14*	*24.13*	*25.57*
	22	*37.50*	*38.92*	.	.	*37.50*	*38.92*
	23	.	.	.	.	.	.
Group Total		*8.63*	*6.25*	*6.62*	*4.37*	*8.23*	*5.82*

Total for each gender category ("column total").

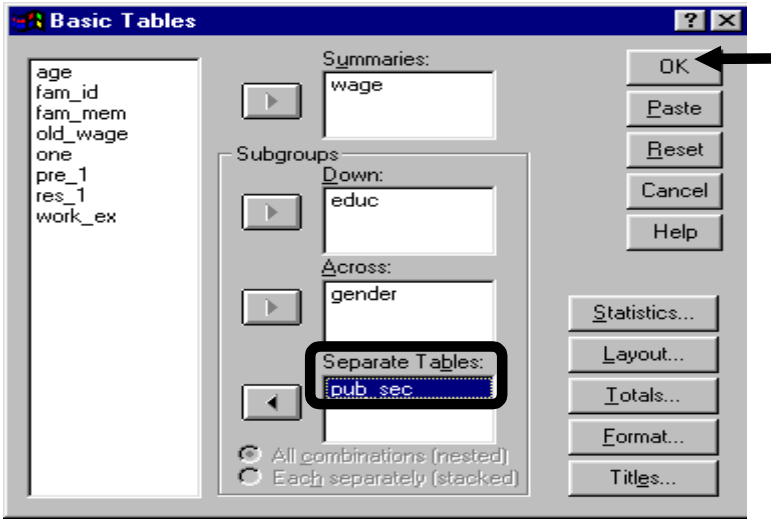
Total for each education category ("row total").

Inspect the table carefully. Look at the patterns in means and medians and compare the two. For almost all the *education-gender* combinations, the medians are lower than the means, implying that a few high earners are pushing the mean up in each unique *education-gender* entry.

Ch 6. Section 1.c. Summary of a variable's values categorized by three other variables

Let's assume we want to find the mean of *wage* for each *education* level, each *gender*, and each *sector* of employment.

Repeat all the steps of the example in section 6.1.a and add one more step - move the variable *pub_sec* into the box "Separate Tables." Now two tables will be produced: one for public sector employees and one for private sector employees.



Note: A better method of doing a 4 (or more) dimensional table construction exercise is to combine (1) a 3-dimensional Custom Table procedure with (2) A single or multidimensional comparative analysis using DATA/ SPLIT FILE. See chapter 10 for more.

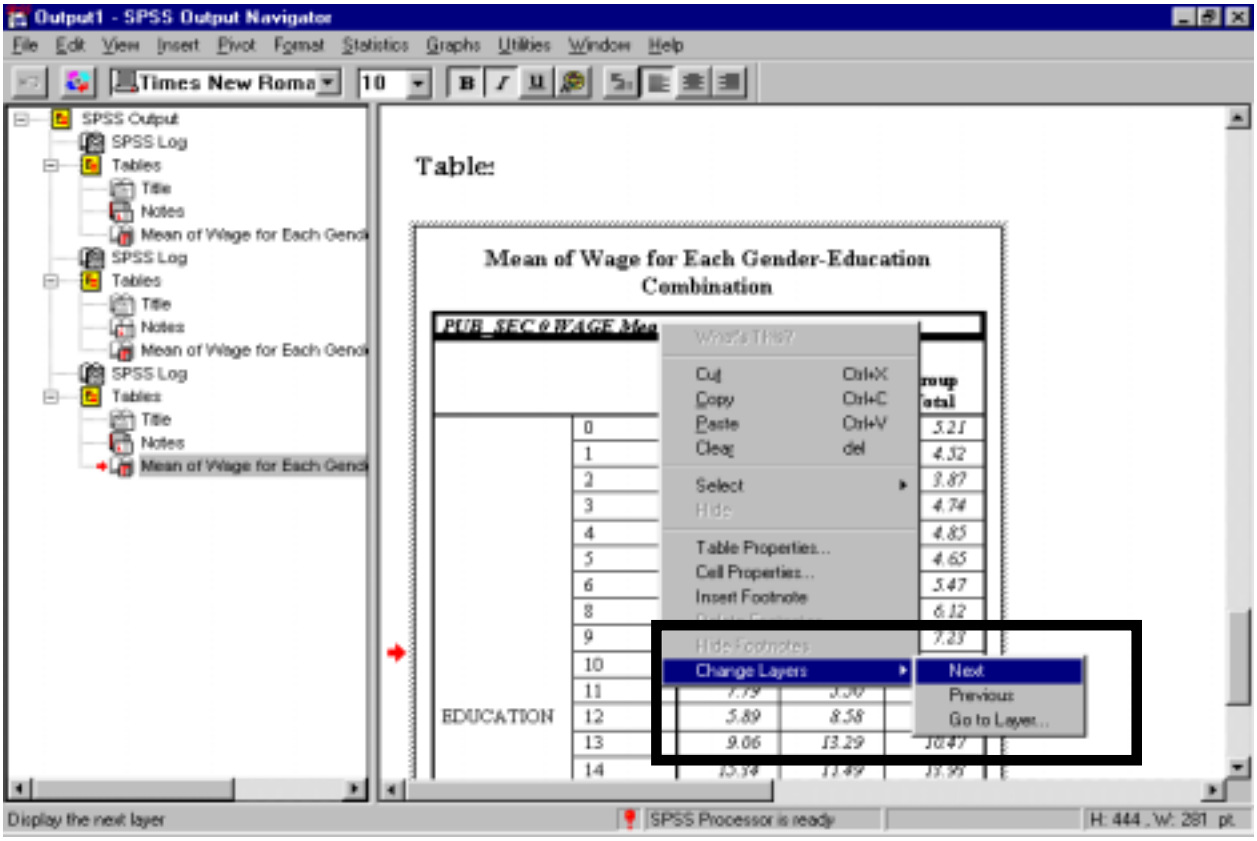
The first table will be for private sector employees (*pub_sec*=0) and will be displayed in the output window. The second table, for public sector employees, will not be displayed.

Mean of Wage for Each Gender-Education Combination

Private Sector

		GENDER		
		Male	Female	**Group Total**
EDUCATION	0	*5.26*	*3.80*	*5.21*
	1	*5.11*	*2.98*	*4.52*
	2	*3.73*	*5.11*	*3.87*
	3	*4.88*	*3.50*	*4.74*
	4	*5.20*	*2.61*	*4.85*
	5	*4.80*	*3.13*	*4.65*
	6	*5.80*	*3.60*	*5.47*
	8	*6.15*	*4.97*	*6.12*
	9	*7.12*	*8.07*	*7.23*
	10	*8.33*	*3.60*	*7.80*
	11	*7.79*	*5.50*	*7.41*
	12	*5.89*	*8.58*	*6.32*
	13	*9.06*	*13.29*	*10.47*
	14	*15.34*	*11.49*	*13.93*
	15	*7.37*	*9.66*	*7.94*
	16	*16.86*	.	*16.86*
	17	*20.65*	*17.05*	*19.75*
	18	*35.04*	*22.59*	*28.81*
	19	.	*17.33*	*17.33*
	20	*19.32*	*26.33*	*22.83*
	21	.	.	.
	22	.	.	.
	23	.	.	.
Group Total		*5.95*	*4.40*	*5.64*

You need to view and print the second table (for *pub_sec*=1). To view it, first double click on the table above in the output window. Click on the right mouse. You will see several options.



Select "Change Layers." Select the option "Next." The custom table for *pub_sec*=1 will be shown.

Mean of Wage for Each Gender-Education Combination

Public Sector

		GENDER		
		Male	Female	
EDUCATION	0	*7.62*	*2.29*	*7.17*
	1	*6.51*	*3.53*	*6.00*
	2	*5.66*	*8.52*	*6.23*
	3	*9.17*	*7.14*	*8.88*
	4	*7.39*	*7.12*	*7.26*
	5	*8.08*	*5.40*	*7.90*
	6	*8.97*	*4.56*	*8.76*
	8	*11.58*	.	*11.58*
	9	*13.53*	*10.72*	*13.08*
	10	*13.33*	*9.00*	*12.69*
	11	*15.25*	*11.16*	*14.26*
	12	*15.19*	*11.67*	*14.35*
	13	*14.77*	*11.51*	*13.72*
	14	*15.96*	*11.26*	*14.74*
	15	*15.29*	*14.89*	*15.14*
	16	*17.11*	*12.87*	*15.55*
	17	*22.53*	*16.27*	*20.44*
	18	*24.16*	*16.13*	*22.73*
	19	*37.97*	*15.33*	*28.92*
	20	*36.42*	.	*36.42*
	21	*23.63*	*26.14*	*24.13*
	22	*37.50*	.	*37.50*
	23	.	.	.
Group Total		*13.42*	*10.71*	*12.89*

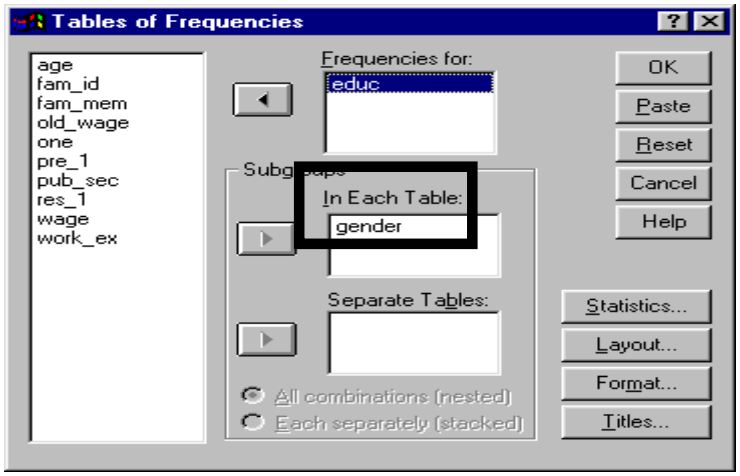
Ch 6. Section 2 Tables of frequencies

A table of frequencies examines the distribution of observations of a variable across the values of other category variable(s). The options in this procedure are a Sub-set of the options in section 6.1.

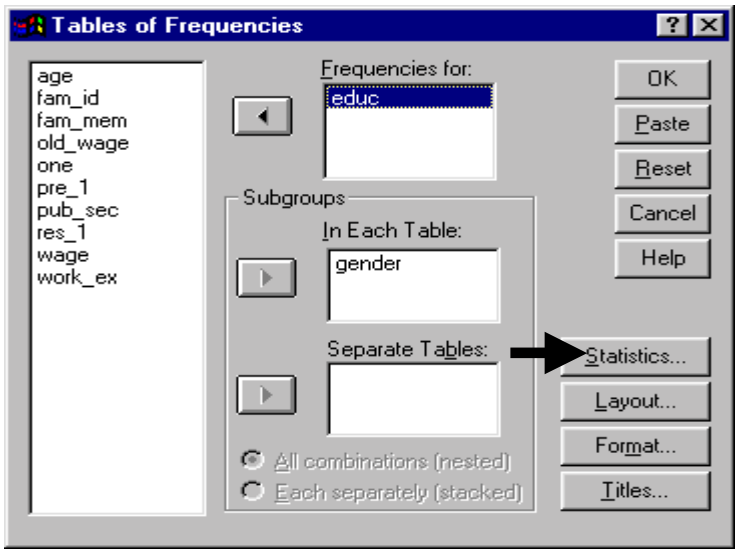
Go to STATISTICS/ CUSTOM TABLES/ TABLES OF FREQUENCIES.

Move *educ* to the box "Frequencies for."

Move *gender* into the box "In Each Table."

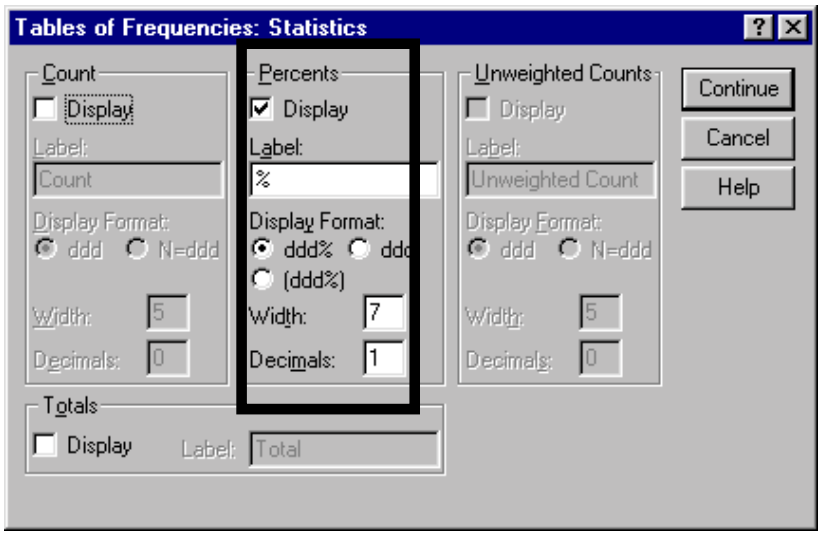


Click on the button "Statistics."

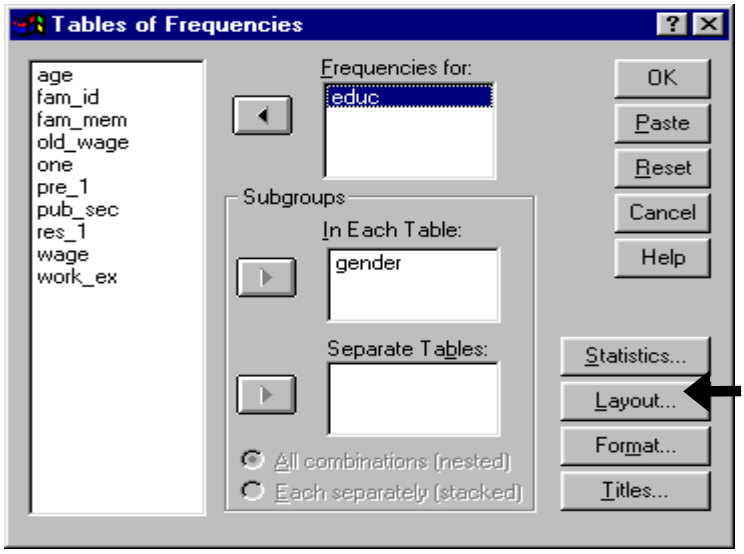


There are two types of statistics displayed: "Count" and "Percent." The latter is preferred.

Select the options in this dialog box and press "Continue."

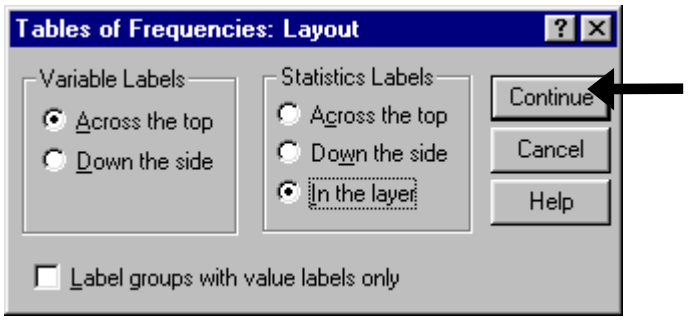


Click on the button "Layout."

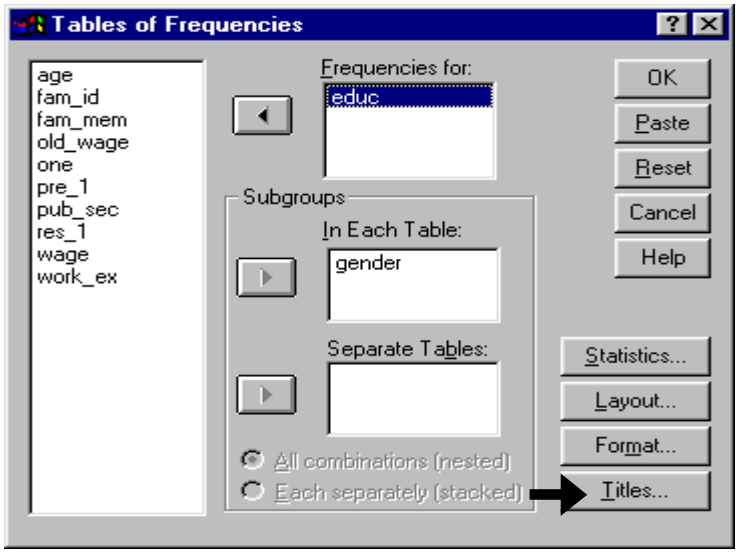


Select options as shown.

Click on "Continue."



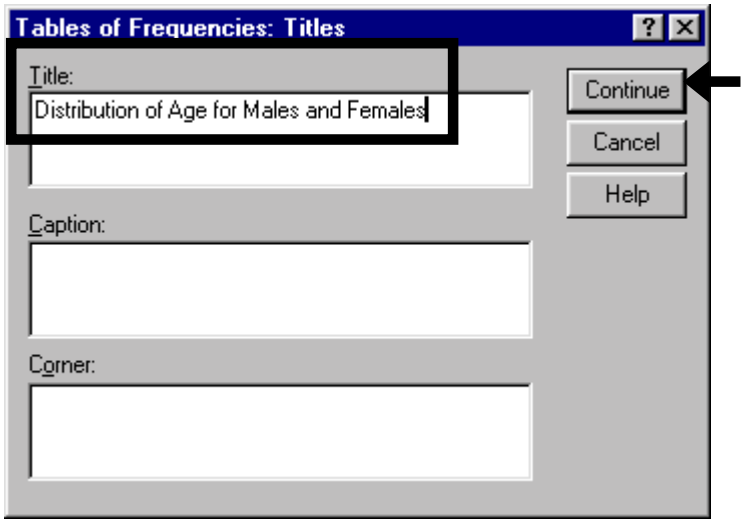
Click on "Titles."



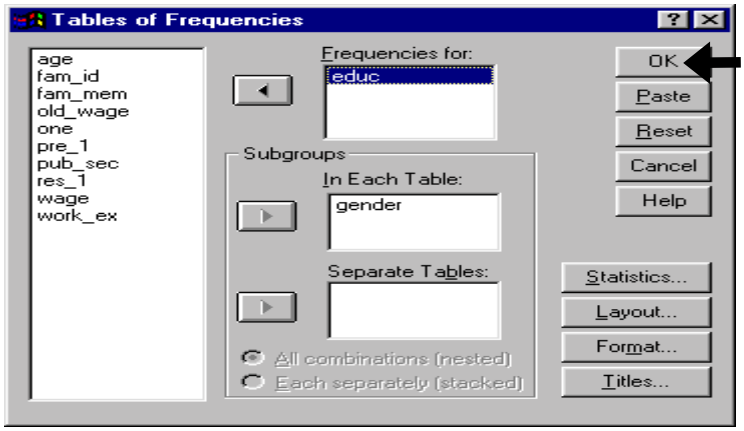
Write a title for your table.

Click on "Continue."

Note: In some of the sections we skip this step. We advise you to always use the title option so that output is easy to identify and publish.



Click on "OK."



Distribution of Age for Males and Females

%

	GENDER	
	0	1
	AGE	AGE
15	*1.3%*	*2.5%*
16	*1.5%*	*2.5%*
17	*1.5%*	*3.0%*
18	*2.3%*	*2.5%*
19	*2.4%*	*2.5%*
20	*2.9%*	*2.0%*
21	*2.7%*	*1.5%*
22	*2.9%*	*3.0%*
23	*1.9%*	*2.7%*
24	*2.4%*	*2.7%*
25	*2.7%*	*3.2%*
26	*3.3%*	*2.7%*
27	*2.5%*	*3.5%*
28	*2.5%*	*2.7%*
29	*2.7%*	*2.5%*
30	*3.7%*	*4.0%*
31	*2.7%*	*4.0%*
32	*3.3%*	*4.7%*
33	*3.2%*	*4.2%*
34	*2.7%*	*2.7%*
35	*3.6%*	*3.5%*
36	*3.3%*	*2.7%*
37	*3.0%*	*4.2%*
38	*2.9%*	*2.2%*
39	*2.6%*	*2.7%*
40	*2.5%*	*1.5%*
41	*2.4%*	*1.5%*
42	*2.3%*	*1.5%*
43	*1.8%*	*2.2%*
44	*1.1%*	*1.5%*
45	*2.4%*	*1.0%*
46	*2.2%*	*1.0%*
47	*1.7%*	*1.0%*
48	*1.1%*	*.5%*
49	*1.8%*	*1.7%*
50	*1.7%*	*1.5%*
51	*2.5%*	*2.0%*
52	*1.4%*	
53	*.9%*	*.5%*
54	*1.2%*	*1.2%*
55	*1.2%*	*.7%*
56	*.6%*	*1.2%*
57	*.6%*	*.5%*
58	*.8%*	*1.0%*
59	*.7%*	*.2%*
60	*1.1%*	
61	*.6%*	*1.2%*
62	*.1%*	*.7%*
63	*.3%*	*.5%*
64	*.1%*	*.2%*
65	*.3%*	*.5%*

The observations are pretty well spread out with some clumping in the range 25-40, as expected. You can read interesting pieces of information from the table: “The number of young females (< 19) is greater than males,” “females seem to have a younger age profile, with many of observations in the 30-38 age range," etc.

Compare these facts with known facts about the distribution of the population. Do the cells in this table conform to reality?

- Also note that at this stage you have been able to look at a very micro-level aggregation.

To take quizzes on topics within each chapter go to http://www.spss.org/wwwroot/spssquiz.asp

ADVERTISEMENT

Ch 7. LINEAR REGRESSION

Regression procedures are used to obtain statistically established causal relationships between variables. Regression analysis is a multi-step technique. The process of conducting "Ordinary Least Squares" estimation is shown in section 7.1.

Several options must be carefully selected while running a regression, because the all-important process of interpretation and diagnostics depends on the output (tables and charts produced from the regression procedure) of the regression and this output, in turn, depends upon the options you choose.

Interpretation of regression output is discussed in section 7.2[83]. Our approach might conflict with practices you have employed in the past, such as always looking at the R-square first. As a result of our vast experience in using and teaching econometrics, we are firm believers in our approach. You will find the presentation to be quite simple - everything is in one place and displayed in an orderly manner.

The acceptance (as being reliable/true) of regression results hinges on diagnostic checking for the breakdown of classical assumptions[84]. If there is a breakdown, then the estimation is unreliable, and thus the interpretation from section 7.2 is unreliable. Section 7.3 lists the various possible breakdowns and their implications for the reliability of the regression results[85].

Why is the result not acceptable unless the assumptions are met? The reason is that the strong statements inferred from a regression (i.e. - "an increase in one unit of the value of variable X causes an increase in the value of variable Y by 0.21 units") depend on the presumption that the variables used in a regression, and the residuals from the regression, satisfy certain statistical properties. These are expressed in the properties of the distribution of the residuals *(that explains why so many of the diagnostic tests shown in sections 7.4-7.5 and the corrective methods shown chapter 8 are based on the use of the residuals).* If these properties are satisfied, then we can be confident in our interpretation of the results.

The above statements are based on complex formal mathematical proofs. Please check your textbook if you are curious about the formal foundations of the statements.

Section 7.4 provides a schema for checking for the breakdown of classical assumptions. The testing usually involves informal (graphical) and formal (distribution-based hypothesis tests like

[83] Even though interpretation precedes checking for the breakdown of classical assumptions, it is good practice to first check for the breakdown of classical assumptions (sections 7.4-7.5), then to correct for the breakdowns (chapter 8), and then, finally, to interpret the results of a regression analysis.

[84] We will use the phrase "Classical Assumptions" often. Check your textbook for details about these assumptions. In simple terms, regression is a statistical method. The fact that this generic method can be used for so many different types of models and in so many different fields of study hinges on one area of commonality - the model rests on the bedrock of the solid foundations of well-established and proven statistical properties/theorems. If the specific regression model is in concordance with the certain assumptions required for the use of these properties/theorems, then the generic regression results can be inferred. The classical assumptions constitute these requirements.

[85] If you find any breakdown(s) of the classical assumptions, then you must correct for it by taking appropriate measures. Chapter 8 looks into these measures. After running the "corrected" model, you again must perform the full range of diagnostic checks for the breakdown of classical assumptions. This process will continue until you no longer have a serious breakdown problem, or the limitations of data compel you to stop.

the F and T) testing, with the latter involving the running of other regressions and computing of variables.

Section 7.5 explores in detail the many steps required to run one such formal test: White's test for heteroskedasticity**.**

Similarly, formal tests are typically required for other breakdowns. Refer to a standard econometrics textbook to review the necessary steps.

Ch 7. Section 1 OLS Regression

Assume you want to run a regression of *wage* on *age*, *work experience*, *education*, *gender*, and a dummy for *sector of employment* (whether employed in the public sector).

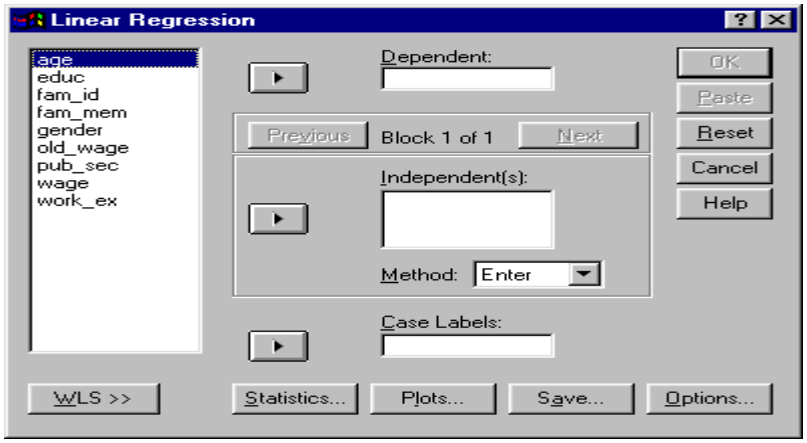
wage = function(*age*, *work experience*, *education*, *gender*, *sector*)

or, as your textbook will have it,

$wage = \beta_1 + \beta_2 * age + \beta_3 * work\ experience + \beta_4 * education + \beta_5 * gender + \beta_6 * sector$

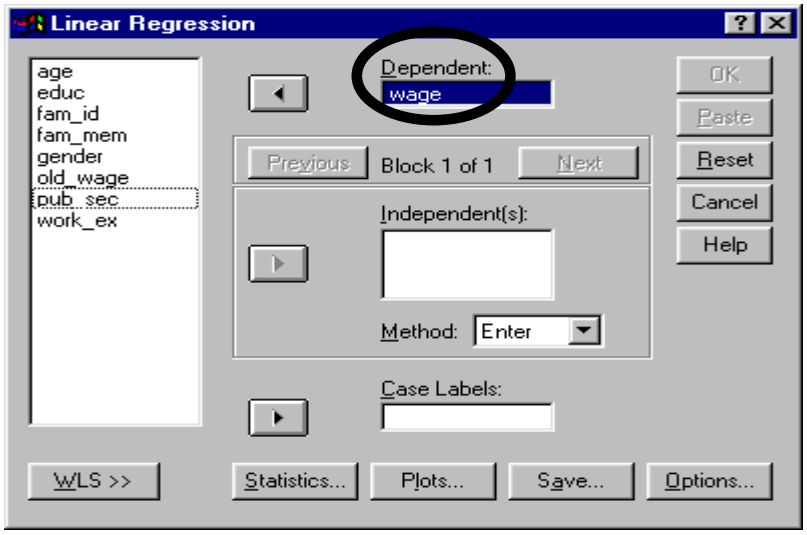
Go to STATISTICS/REGRESSION/ LINEAR

Note: Linear Regression is also called OLS (Ordinary Least Squares). If the term "Regression" is used without any qualifying adjective, the implied method is Linear Regression.



Click on the variable *wage*. Place it in the box "Dependent" by clicking on the arrow on the top of the dialog box.

Note: The dependent variable is that whose values we are trying to predict (or whose dependence on the independent variables is being studied). It is also referred to as the "Explained" or "Endogenous" variable, or as the "Regressand."

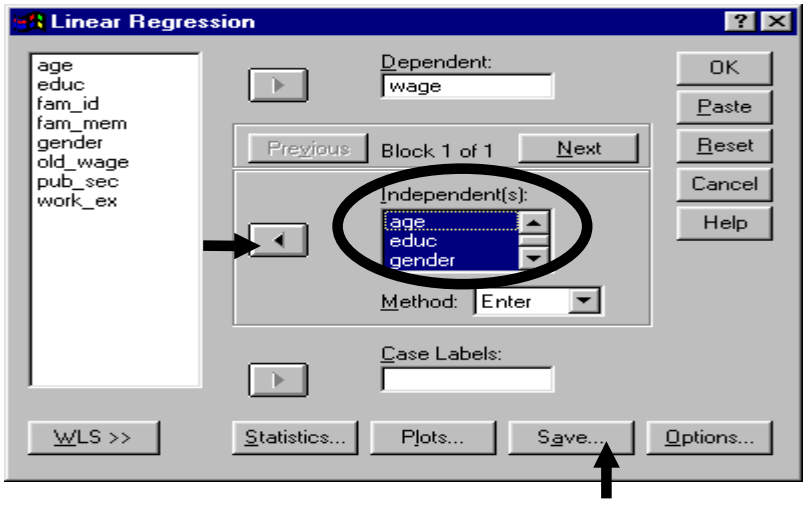


Select the independent variables.

Note: The independent variables are used to explain the values of the dependent variable. The values of the independent variables are not being explained/determined by the model - thus, they are "independent" of the model. The independent variables are also called "Explanatory" or "Exogenous" variables. They are also referred to as "Regressors."

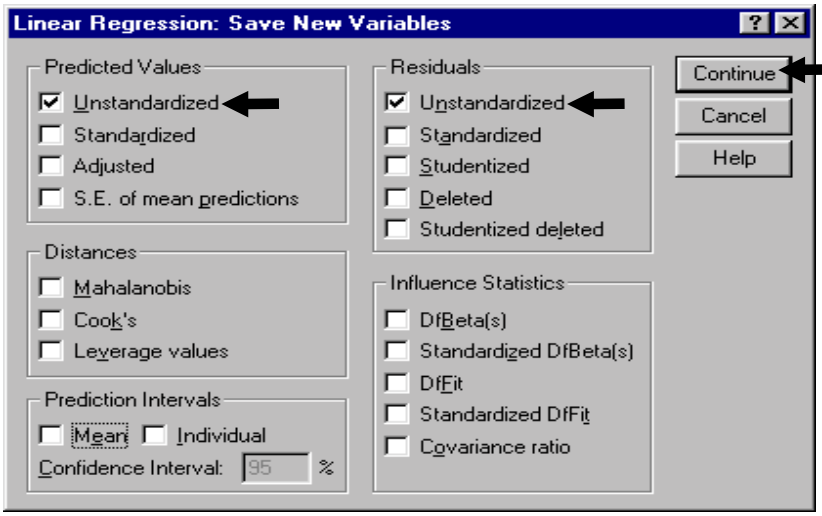
Move the independent variables by clicking on the arrow in the middle.

For a basic regression, the above may be the only steps required. In fact, your professor may only inform you of those steps. However, because comprehensive diagnostics and interpretation of the results are important (as will become apparent in the rest of this chapter and in chapter 8), we advise that you follow all the steps in this section.



Click on the button “Save."

Select to save the unstandardized predicted values and residuals by clicking on the boxes shown.



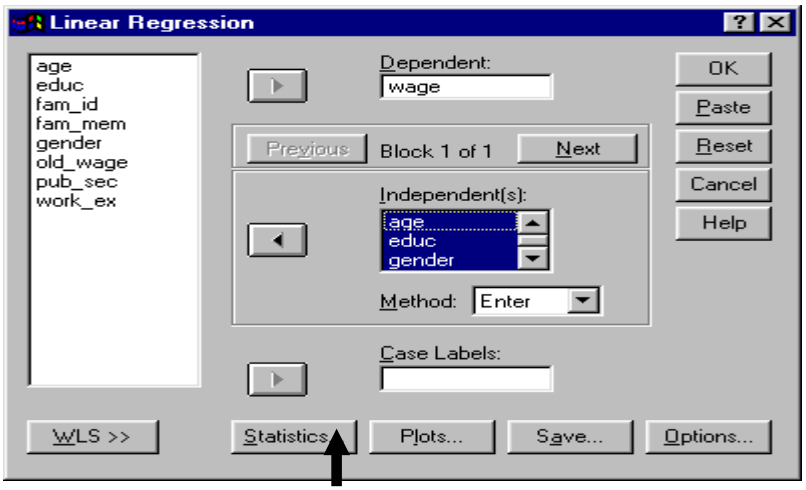
Choosing these variables is not an essential option. We would, however, suggest that you choose these options because the saved variables may be necessary for checking for the breakdown of classical assumptions[86].

For example, you will need the residuals for the White's test for heteroskedasticity (see section 7.5), and the residuals and the predicted values for the RESET test, etc.

The use of statistics shown in the areas "Distances"[87] and "Influence Statistics" are beyond the scope of this book. If you choose the box "Individual" in the area "Prediction Intervals," you will get two new variables, one with predictions of the lower bound of the 95% confidence interval.

Click on "Continue."

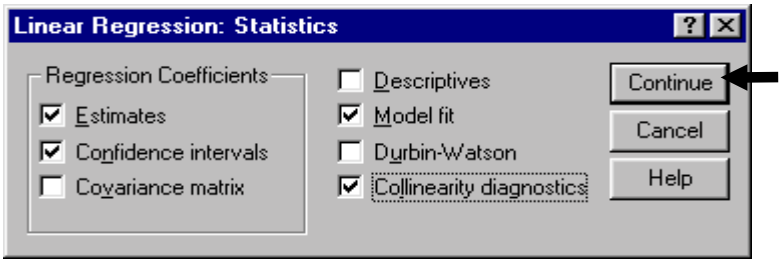
Now we will choose the output tables produced by SPSS. To do so, click on the button "Statistics."



The statistics chosen here provide what are called "regression results."

Select "Estimates" & "Confidence Intervals[88]."

"Model Fit" tells if the model fitted the data properly[89].



[86] For example, the residuals are used in the White's test while the predicted dependent variable is used in the RESET test. (See section 7.5.)

[87] "Distance Measurement" (and use) will be dealt with in a follow-up book and/or the next edition of this book in January, 2000. The concept is useful for many procedures apart from Regressions.

[88] These provide the estimates for the coefficients on the independent variables, their standard errors & T-statistics and the range of values within which we can say, with 95% confidence, that the coefficient lies.

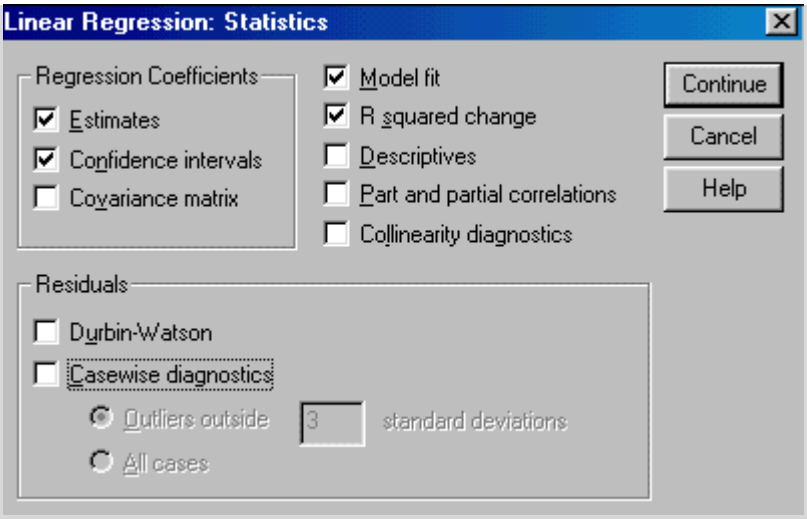
fitted the data properly[89].

Note: We ignore Durbin-Watson because we are not using a time series data set.

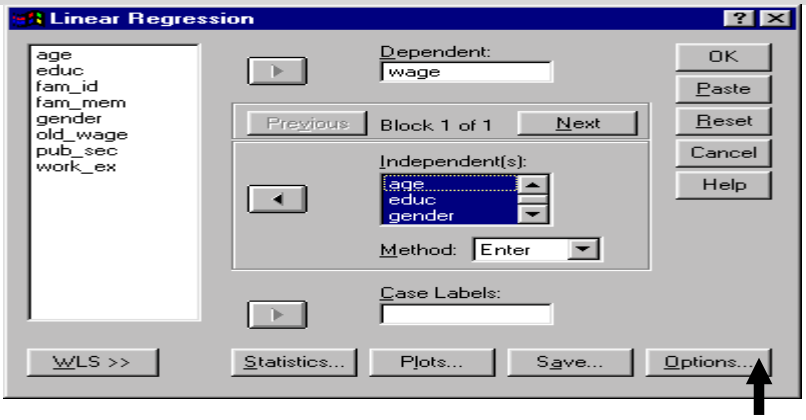
If you suspect a problem with collinearity (and want to use a more advanced test then the simple rule-of-thumb of "a correlation coefficient higher than 0.8 implies collinearity between the two variables"), choose "Collinearity Diagnostics." See section 7.4.

Click on "Continue."

In later versions of SPSS (7.5 and above), some new options are added. Usually, you can ignore these new options. Sometimes, you should include a new option. For example, in the Linear Regression options, choose the statistic "R squared change."

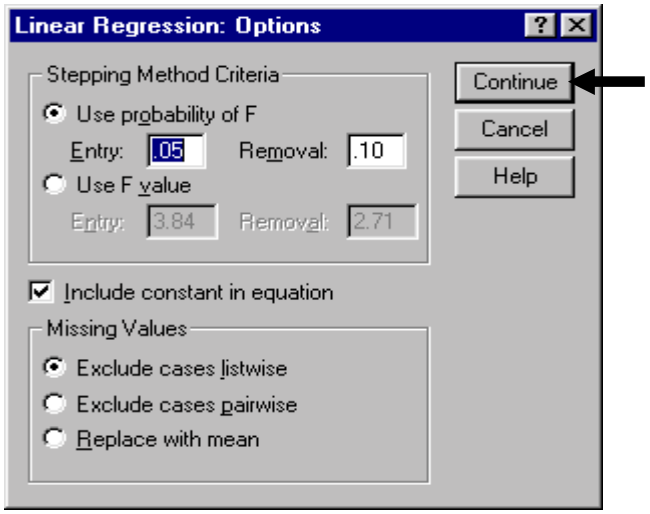


Click on the button "Options."



[89] If the model fit indicates an unacceptable F-statistic, then analyzing the remaining output is redundant - if a model does not fit, then none of the results can be trusted. Surprisingly, we have heard a professor working for Springer-Verlag dispute this basic tenet. We suggest that you ascertain your professor's view on this issue.

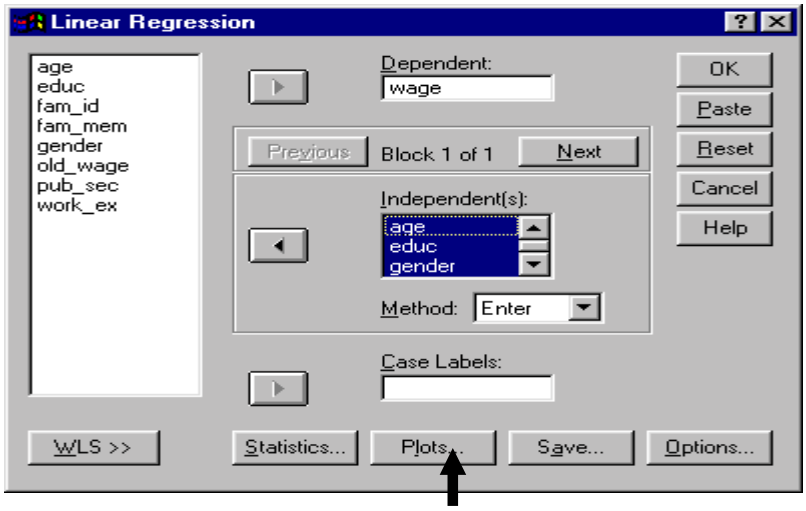
It is typically unnecessary to change any option here.



Note: Deselect the option "Include Constant in Equation" if you do not want to specify any intercept in your model.

Click on "Continue."

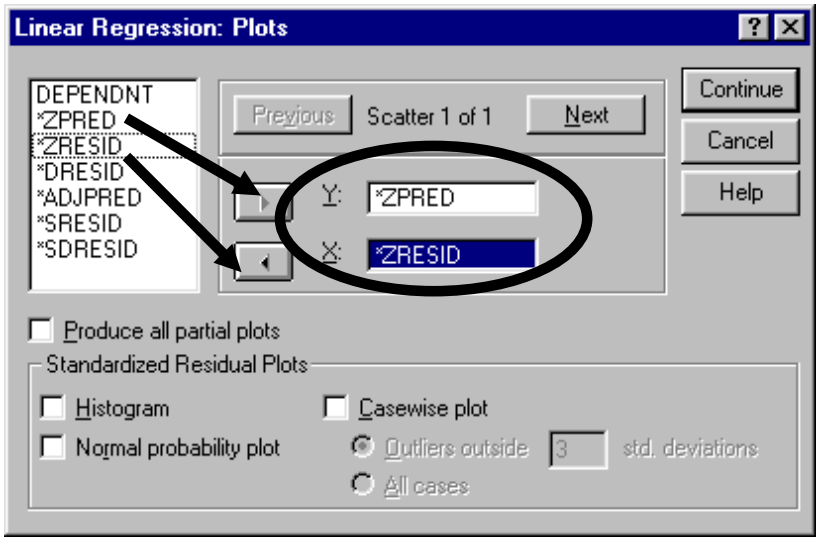
Click on "Plots."



We think that the plotting option is the most important feature to understand for two reasons: (1) Despite the fact that their class notes and econometric books stress the importance of the visual diagnosis of residuals and plots made with the residuals on an axis, most professors ignore them. (2) SPSS help does not provide an adequate explanation of their usefulness. The biggest weakness of SPSS, with respect to basic econometric analysis, is that it does not allow for easy diagnostic checking for problems like mis-specification and heteroskedasticity (see section 7.5 for an understanding of the tedious nature of this diagnostic process in SPSS). In order to circumvent this lacuna, always use the options in plot to obtain some visual indicators of the presence of these problems.

We repeat: the options found here are essential - they allow the production of plots which provide summary diagnostics for violations of the classical regression assumptions.

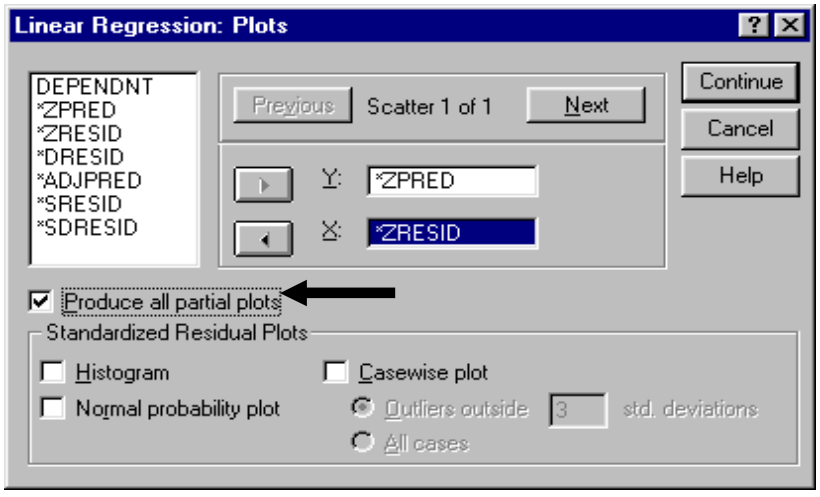
Select the option "ZPRED" (standard normal of predicted variable) and move it into the box "Y." Select the option "ZRESID" (standard normal of the regression residual) and move it into the box "X."



Any pattern in that plot will indicate the presence of heteroskedasticity and/or mis-specification due to measurement errors, incorrect functional form, or omitted variable(s). See section 7.4 and check your textbook for more details.

Select to produce plots by clicking on the box next to "Produce all partial plots."

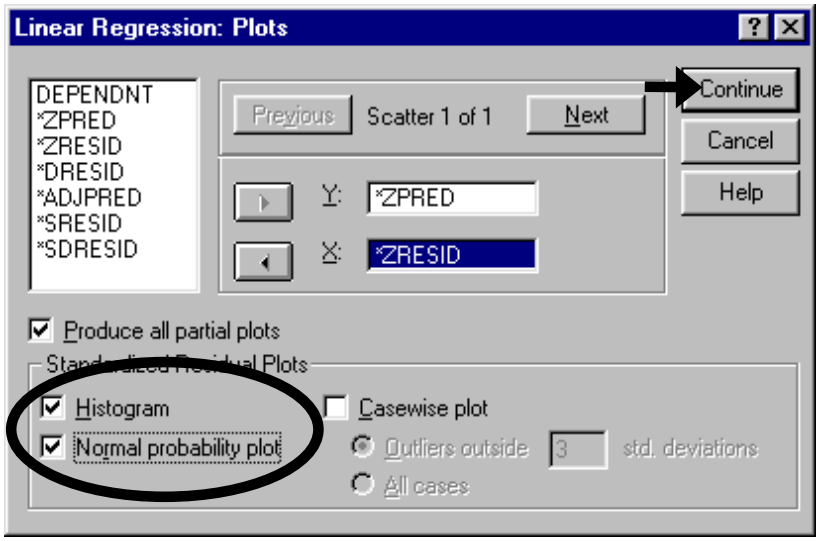
Patterns in these plots indicate the presence of heteroskedasticity.



You may want to include plots on the residuals.

If the plots indicate that the residuals are not distributed normally, then mis-specification, collinearity, or other problems are indicated (section 7.4 explains these issues. Check your textbook for more details on each problem).

Note: Inquire whether your professor agrees with the above concept. If not, then interpret as per his/her opinion.

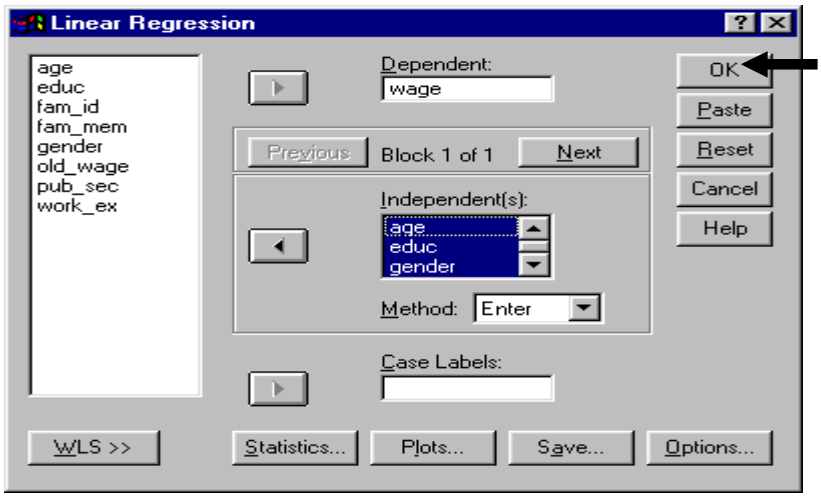


Click on "Continue."

Click on "OK."

The regression will be run and several output tables and plots will be produced (see section 7.2).

Note: In the dialog box on the right, select the option "Enter" in the box "Method." The other methods available can be used to make SPSS build up a model (from one explanatory/independent variable to all) or build "down" a model until it finds the best model. Avoid using those options - many statisticians consider their use to be a dishonest practice that produces inaccurate results.

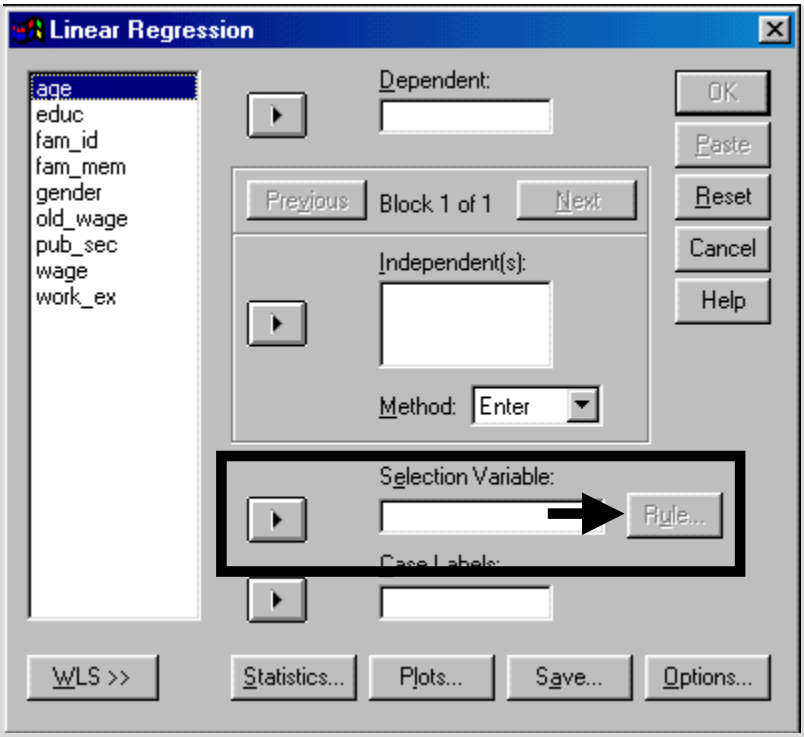


A digression:

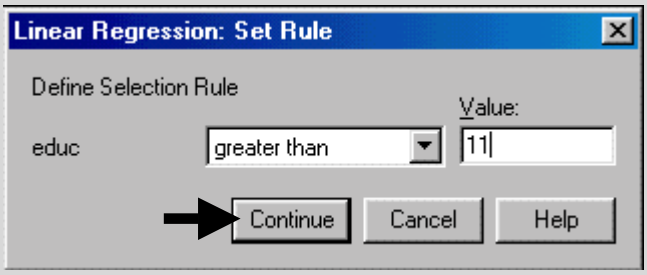
In newer versions of SPSS you will see a slightly different dialog box.

The most notable difference is the additional option, "Selection Variable." Using this option, you can restrict the analysis to a Sub-set of the data.

Assume you want to restrict the analysis to those respondents whose *education* level was more than 11 years of schooling. First, move the variable *education* into the area "Selection Variable." Then click on "Rule."



Enter the rule. In this case, it is "educ>11." Press "Continue" and do the regression with all the other options shown earlier.



Ch 7. Section 2 Interpretation of regression results

Always look at the model fit ("ANOVA") first. **Do not make the mistake of looking at the R-square before checking the goodness of fit**. The last column shows the goodness of fit of the model. The lower this number, the better the fit. Typically, if "Sig" is greater than 0.05, we conclude that our model could not fit the data[90].

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	54514.39	5	10902.88	414.262	.000[b]
	Residual	52295.48	1987	26.319		
	Total	106809.9	1992			

a. Dependent Variable: WAGE

b. Independent Variables: (Constant), WORK_EX, EDUCATION, GENDER, PUB_SEC, AGE

[90] If Sig < .01, then the model is significant at 99%, if Sig < .05, then the model is significant at 95%, and if Sig <.1, the model is significant at 90%. Significance implies that we can accept the model. If Sig>.,1 then the model was not significant (a relationship could not be found) or "R-square is not significantly different from zero."

In your textbook you will encounter the terms TSS, ESS, and RSS (Total, Explained, and Residual Sum of Squares, respectively). The TSS is the total deviations in the dependent variable. The ESS is the amount of this total that could be explained by the model. The R-square, shown in the next table, is the ratio ESS/TSS. It captures the percent of deviation from the mean in the dependent variable that could be explained by the model. The RSS is the amount that could not be explained (TSS minus ESS). In the previous table, the column "Sum of Squares" holds the values for TSS, ESS, and RSS. The row "Total" is TSS (106809.9 in the example), the row "Regression" is ESS (54514.39 in the example), and the row "Residual" contains the RSS (52295.48 in the example).

The "Model Summary" tells us:

- which of the variables were used as independent variables[91],
- the proportion of the **variance** in the dependent variable (*wage*) that was explained by variations in the independent variables[92],
- the proportion of the **variation** in the dependent variable (*wage*) that was explained by variations in the independent variables[93]
- and the dispersion of the dependent variables estimate around its mean (the "Std. Error of the Estimate" is 5.13[94]).

Model Summary [a,b]

Model	Variables Entered	Variables Removed	R Square	Adjusted R Square	Std. Error of the Estimate
1	*WORK_EX, EDUCATION, GENDER, PUB_SEC, AGE* [c,d]	.	*.510*	*.509*	*5.1302*

a. Dependent Variable: WAGE

b. Method: Enter

c. Independent Variables: (Constant), WORK_EX, EDUCATION, GENDER, PUB_SEC, AGE

d. All requested variables entered.

[91] Look in the column "Variables/Entered."

[92] The "**Adjusted** R-Square" shows that 50.9% of the variance was explained.

[93] The "R-Square"' tells us that 51% of the variation was explained.

[94] Compare this to the mean of the variable you asked SPSS to create - "Unstandardized Predicted." If the Std. Error is more than 10% of the mean, it is high.

The table "Coefficients" provides information on:

- the effect of individual variables (the "Estimated Coefficients"--see column "B") on the dependent variable and
- the confidence with which we can support the estimate for each such estimate (see the column "Sig.").

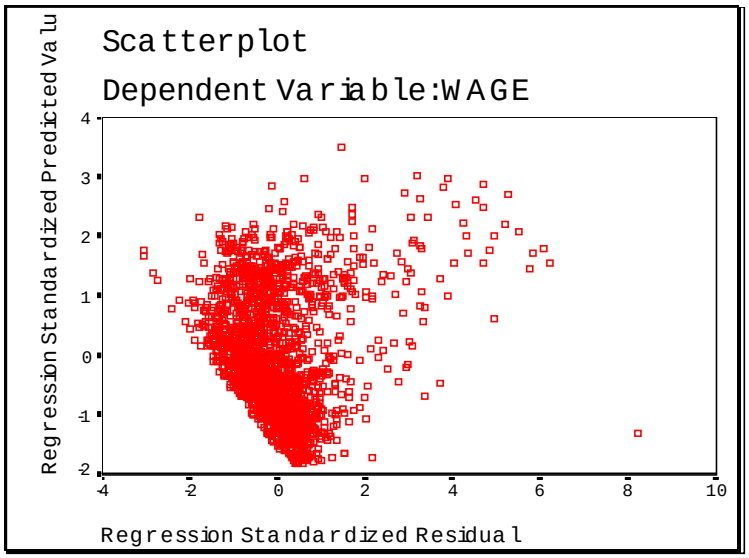
If the value in "Sig." is less than 0.05, then we can assume that the estimate in column "B" can be asserted as true with a 95% level of confidence[95]. Always interpret the "Sig" value first. **If this value is more than .1 then the coefficient estimate is not reliable because it has "too" much dispersion/variance.**

Coefficients[a]

Model		Unstandardized Coefficients		t	Sig.	95% Confidence Interval for B	
		B	Std. Error			Lower Bound	Upper Bound
1	(Constant)	*-1.820*	*.420*	*-4.339*	*.000*	*-2.643*	*-.997*
	AGE	*.118*	*.014*	*8.635*	*.000*	*.091*	*.145*
	EDUCATION	*.777*	*.025*	*31.622*	*.000*	*.729*	*.825*
	GENDER	*-2.030*	*.289*	*-7.023*	*.000*	*-2.597*	*-1.463*
	PUB_SEC	*1.741*	*.292*	*5.957*	*.000*	*1.168*	*2.314*
	WORK_EX	*.100*	*.017*	*5.854*	*.000*	*.067*	*.134*

a. Dependent Variable: WAGE

This is the plot for "ZPRED versus ZRESID." The pattern in this plot indicates the presence of mis-specification[96] and/or heteroskedasticity.



A formal test such as the RESET Test is required to conclusively prove the existence of mis-specification. This test requires the running of a new regression using the variables you saved in this regression - both the predicted and residuals. You will be required to create other transformations of these variables (see section 2.2 to

[95] If the value is greater than 0.05 but less than 0.1, we can only assert the veracity of the value in "B" with a 90% level of confidence. If "Sig" is above 0.1, then the estimate in "B" is unreliable and is said to not be statistically significant. The confidence intervals provide a range of values within which we can assert with a 95% level of confidence that the estimated coefficient in "B" lies. For example, "The coefficient for *age* lies in the range .091 and .145 with a 95% level of confidence, while the coefficient for *gender* lies in the range -2.597 and -1.463 at a 95% level of confidence."

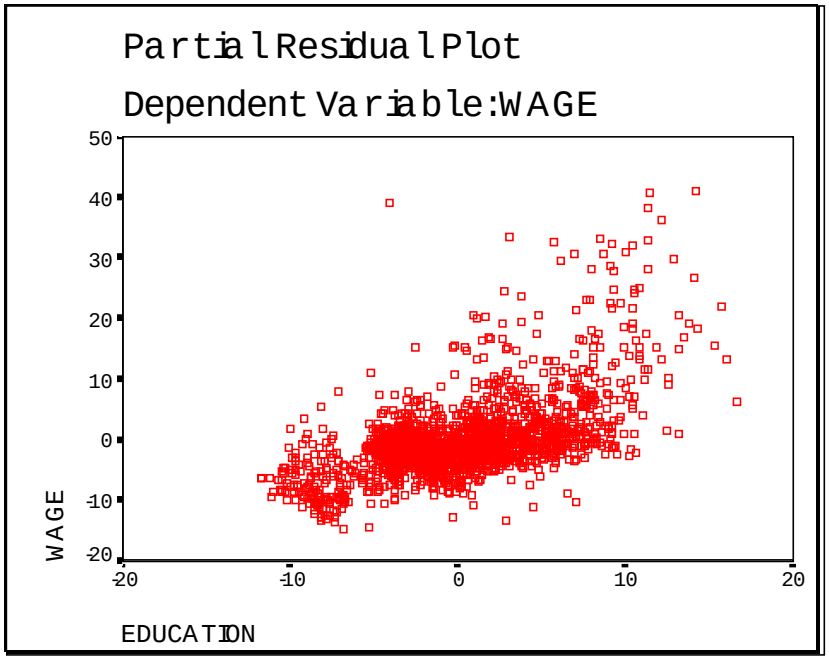
[96] Incorrect functional form, omitted variable, or a mis-measured independent variable.

learn how). Review your textbook for the step-by-step description of the RESET test.

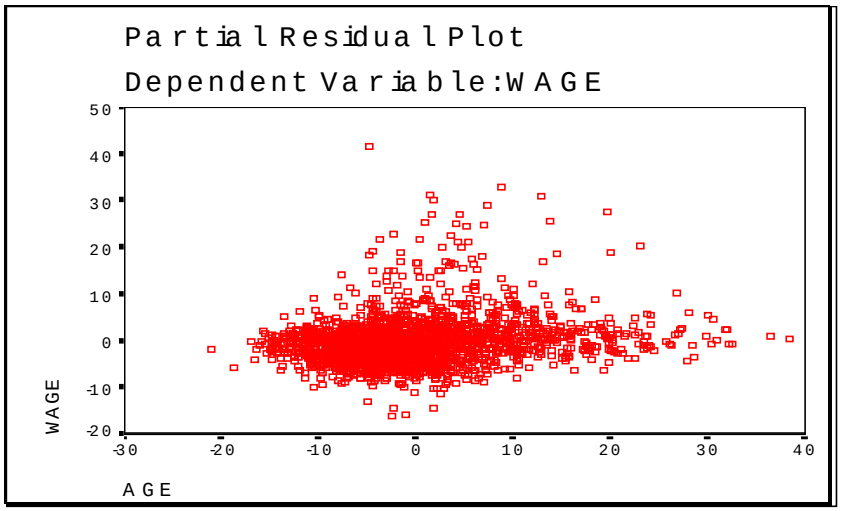
A formal test like the White's Test is necessary to conclusively prove the existence of heteroskedasticity. We will run the test in section 7.5.

This is the partial plot of residuals versus the variable *education*. The definite positive pattern indicates the presence of heteroskedasticity caused, at least in part, by the variable education.

A formal test like the White's Test is required to conclusively prove the existence and structure of heteroskedasticity (see section 7.5).



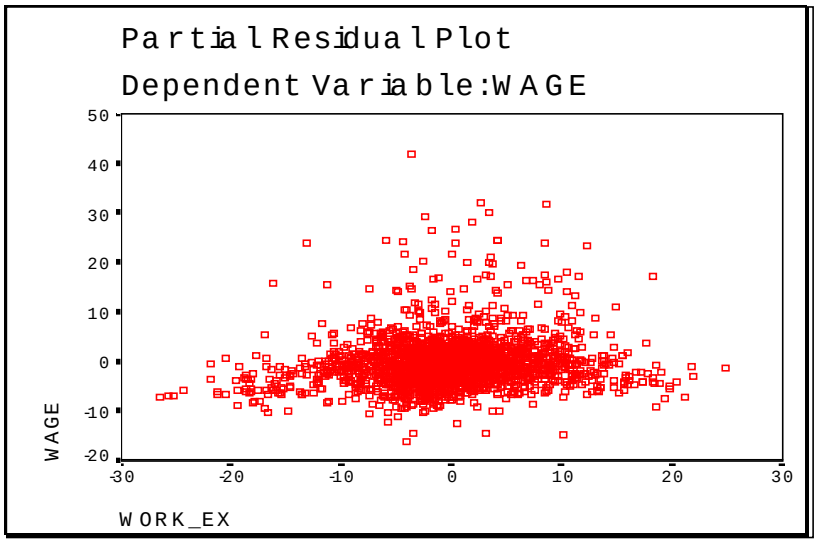
The partial plots of the variables *age* and *work experience* have no pattern, which implies that no heteroskedasticity is caused by these variables.



Note: Sometimes these plots may not show a pattern. The reason may be the presence of extreme values that widen the scale of one or both of the axes, thereby "smoothing out" any patterns. If you suspect this has happened, as would be the case if most of the graph area were empty save for a few dots at the extreme ends of the graph, then rescale the axes using the methods shown in section 11.2. This is true for all graphs produced, including the ZPRED-ZRESID shown on the previous page.

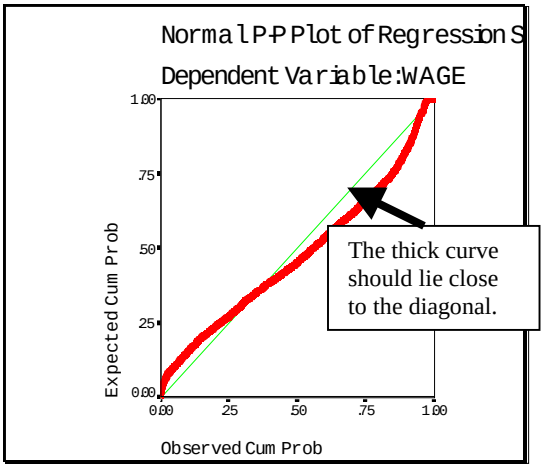
Note also that the strict interpretation of the partial plots may differ from the way we use

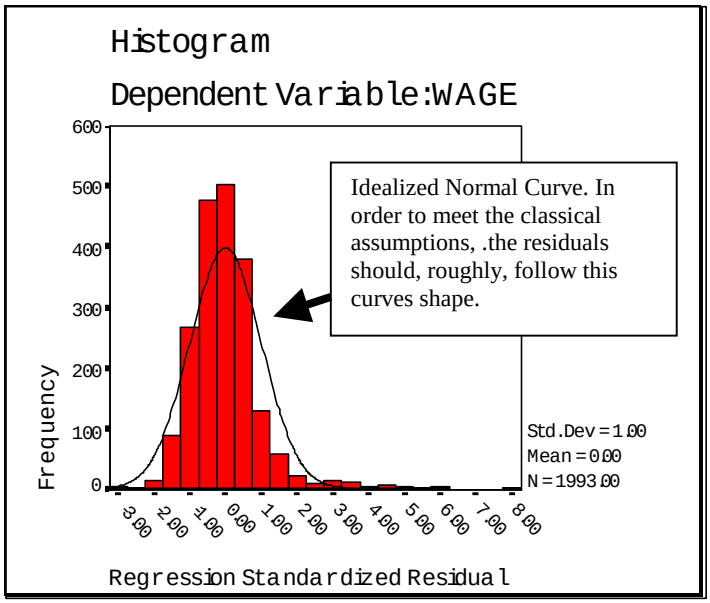
the partial plots here. Without going into the details of a strict interpretation, we can assert that the best use of the partial plots vis-à-vis the interpretation of a regression result remains as we have discussed it.



The histogram and the P-P plot of the residual suggest that the residual is probably normally distributed[97].

You may want to use the Runs test (see chapter 14) to determine whether the residuals can be assumed to be randomly distributed.





[97] See chapter 3 for interpretation of the P-P. The residuals should be distributed normally. If not, then some classical assumption has been violated.

Regression output interpretation guidelines

Name Of Statistic/ Chart	**What Does It Measure Or Indicate?**	**Critical Values**	**Comment**
Sig.-F *(in the ANOVA table)*	Whether the model as a whole is significant. It tests whether R-square is significantly different from zero	- below .01 for 99% confidence in the ability of the model to explain the dependent variable - below .05 for 95% confidence in the ability of the model to explain the dependent variable - below 0.1 for 90% confidence in the ability of the model to explain the dependent variable	**The first statistic to look for in SPSS output**. If Sig.-F is insignificant, then the regression as a whole has failed. No more interpretation is necessary (although some statisticians disagree on this point). You must conclude that the "Dependent variable cannot be explained by the independent/explanatory variables." The next steps could be rebuilding the model, using more data points, etc.
RSS, ESS & TSS *(in the ANOVA table)*	The main function of these values lies in calculating test statistics like the F-test, etc.	The ESS should be high compared to the TSS (the ratio equals the R-square). Note for interpreting the SPSS table, column "Sum of Squares": "Total" =TSS, "Regression" = ESS, and "Residual" = RSS	If the R-squares of two models are very similar or rounded off to zero or one, then you might prefer to use the F-test formula that uses RSS and ESS.
SE of Regression *(in the Model Summary table)*	The standard error of the estimate predicted dependent variable	There is no critical value. Just compare the std. error to the mean of the predicted dependent variable. The former should be small (<10%) compared to the latter.	You may wish to comment on the SE, especially if it is too large or small relative to the mean of the predicted/estimated values of the dependent variable.
R-Square *(in the Model Summary table)*	Proportion of variation in the dependent variable that can be explained by the independent variables	Between 0 and 1. A higher value is better.	This often mis-used value should serve only as a summary measure of Goodness of Fit. Do not use it blindly as a criterion for model selection.

Name Of Statistic/ Chart	**What Does It Measure Or Indicate?**	**Critical Values**	**Comment**
Adjusted R-square *(in the Model Summary table)*	Proportion of variance in the dependent variable that can be explained by the independent variables or R-square adjusted for # of independent variables	Below 1. A higher value is better	Another summary measure of Goodness of Fit. Superior to R-square because it is sensitive to the addition of irrelevant variables.
T-Ratios *(in the Coefficients table)*	The reliability of our estimate of the individual beta	Look at the p-value (in the column "Sig.") it must be low: - below .01 for 99% confidence in the value of the estimated coefficient - below .05 for 95% confidence in the value of the estimated coefficient - below .1 for 90% confidence in the value of the estimated coefficient	For a one-tailed test (at 95% confidence level), the critical value is (approximately) 1.65 for testing if the coefficient is greater than zero and (approximately) -1.65 for testing if it is below zero.
Confidence Interval for beta *(in the Coefficients table)*	The 95% confidence band for each beta estimate	The upper and lower values give the 95% confidence limits for the coefficient	Any value within the confidence interval cannot be rejected (as the true value) at 95% degree of confidence
Charts: Scatter of predicted dependent variable and residual *(ZPRED & ZRESID)*	Mis-specification and/or heteroskedasticity	There should be no discernible pattern. If there is a discernible pattern, then do the RESET and/or DW test for mis-specification or the White's test for heteroskedasticity	Extremely useful for checking for breakdowns of the classical assumptions, i.e. - for problems like mis-specification and/or heteroskedasticity. At the top of this table, we mentioned that the F-statistic is the first output to interpret. Some may argue that the ZPRED-ZRESID plot is more important (their rationale will become apparent as you read through the rest of this chapter and chapter 8).

Name Of Statistic/ Chart	What Does It Measure Or Indicate?	Critical Values	Comment
Charts: Partial plots	Heteroskedasticity	There should be no discernible pattern. If there is a discernible pattern, then perform White's test to formally check.	Common in cross-sectional data. If a partial plot has a pattern, then that variable is a likely candidate for the cause of heteroskedasticity.
Charts: Histograms of residuals	Provides an idea about the distribution of the residuals	The distribution should look like a normal distribution	A good way to observe the actual behavior of our residuals and to observe any severe problem in the residuals (which would indicate a breakdown of the classical assumptions)

Ch 7. Section 3 Problems caused by breakdown of classical assumptions

The fact that we can make bold statements on causality from a regression hinges on the classical linear model. If its assumptions are violated, then we must re-specify our analysis and begin the regression anew. It is very unsettling to realize that a large number of institutions, journals, and faculties allow this fact to be overlooked.

When using the table below, remember the ordering of the severity of an impact.

- The worst impact is a bias in the F (then the model cant be trusted)
- A second disastrous impact is a bias in the betas (the coefficient estimates are unreliable)
- Compared to the above, biases in the standard errors and T are not so harmful (these biases only affect the reliability of our confidence about the variability of an estimate, not the reliability about the value of the estimate itself)

Violation ↓ / Impact →	β	Std err (of estimate)	Std err (of β)	T	F	R^2
Measurement error in dependent variable	👍	👍	X ↑	X ↓	👍	👍
Measurement error in independent variable	X	X	X	X	X	X
Irrelevant variable	👍	👍	X ↑	X ↓	👍	👍
Omitted variable	X	X	X	X	X	X

Violation ↓ / Impact →	β	Std err (of estimate)	Std err (of β)	T	F	R^2
Incorrect functional form	X	X	X	X	X	X
Heteroskedasticity	👍	X	X	X	X	👍
Collinearity	👍	👍	X ↑	X ↓	👍	👍
Simultaneity Bias	X	X	X	X	X	X

👍 The statistic is still reliable and unbiased.

X The statistic is biased, and thus cannot be relied upon.

↑ Upward bias.

↓ Downward bias.

Ch 7. Section 4 Diagnostics

This section lists some methods of detecting for breakdowns of the classical assumptions.

With experience, you should develop the habit of doing the diagnostics before interpreting the model's significance, explanatory power, and the significance and estimates of the regression coefficients. If the diagnostics show the presence of a problem, you must first correct the problem (using methods such as those shown in chapter 8) and then interpret the model. Remember that the power of a regression analysis (after all, it is extremely powerful to be able to say that "data shows that X causes Y by this slope factor") is based upon the fulfillment of certain conditions that are specified in what have been dubbed the "classical" assumptions.

Refer to your textbook for a comprehensive listing of methods and their detailed descriptions.

Ch 7. Section 4.a. Collinearity[98]

Collinearity between variables is always present. A problem occurs if the degree of collinearity is high enough to bias the estimates.

Note: Collinearity means that two or more of the independent/explanatory variables in a regression have a linear relationship. This causes a problem in the interpretation of the

[98] Also called Multicollinearity.

regression results. If the variables have a close linear relationship, then the estimated regression coefficients and T-statistics may not be able to properly isolate the unique effect/role of each variable and the confidence with which we can presume these effects to be true. The close relationship of the variables makes this isolation difficult. Our explanation may not satisfy a statistician, but we hope it conveys the fundamental principle of collinearity.

Summary measures for testing and detecting collinearity include:

- Running bivariate and partial correlations (see section 5.3). A bivariate or partial correlation coefficient greater than 0.8 (in absolute terms) between two variables indicates the presence of significant collinearity between them.
- Collinearity is indicated if the R-square is high (greater than 0.75[99]) and only a few T-values are significant.
- In section 7.1, we asked SPSS for "Collinearity diagnostics" under the regression option "statistics." Here we analyze the table that is produced. Significant collinearity is present if the condition index is >10. If the condition index is greater than 30, then severe collinearity is indicated (see next table). Check your textbook for more on collinearity diagnostics.

Collinearity Diagnostics [a]

Dimension	Eigenvalue	Condition Index	Variance Proportions					
			(Constant)	AGE	EDUCATION	GENDER	PUB_SEC	WORK_EX
1	4.035	1.000	.00	.00	.01	.01	.02	.01
2	.819	2.220	.00	.00	.00	.85	.03	.01
3	.614	2.564	.01	.01	.14	.01	.25	.09
4	.331	3.493	.03	.00	.34	.09	.49	.08
5	.170	4.875	.11	.03	.43	.04	.15	.48
6	3.194E-02	11.239	.85	.96	.08	.00	.06	.32

a. Dependent Variable: WAGE

Ch 7. Section 4.b. Mis-specification

Mis-specification of the regression model is the most severe problem that can befall an econometric analysis. Unfortunately, it is also the most difficult to detect and correct.

Note: Mis-specification covers a list of problems discussed in sections 8.3 to 8.5. These problems can cause moderate or severe damage to the regression analysis. Of graver importance is the fact that most of these problems are caused not by the nature of the data/issue, but by the modeling work done by the researcher. It is of the utmost importance that every researcher realise that the responsibility of correctly specifying an econometric model lies solely on them. A proper specification includes determining curvature (linear or not), functional form (whether to use logs, exponentials, or squared variables), and the accuracy of measurement of each variable, etc.

Mis-specification can be of several types: incorrect functional form, omission of a relevant independent variable, and/or measurement error in the variables. Sections 7.4.c to 7.4.f list a few summary methods for detecting mis-specification. Refer to your textbook for a comprehensive listing of methods and their detailed descriptions.

[99] Some books advise using 0.8.

Ch 7. Section 4.c. Incorrect functional form

If the correct relation between the variables is non-linear but you use a linear model and do not transform the variables[100], then the results will be biased. Listed below are methods of detecting incorrect functional forms:

- Perform a preliminary visual test. To do this, we asked SPSS for the plot ZPRED and Y-PRED while running the regression (see section 7.1). Any pattern in this plot implies mis-specification (and/or heteroskedasticity) due to the use of an incorrect functional form or due to omission of a relevant variable.
- If the visual test indicates a problem, perform a formal diagnostic test like the RESET test[101] or the DW test[102].
- Check the mathematical derivation (if any) of the model.
- Determine whether any of the scatter plots have a non-linear pattern. If so, is the pattern log, square, etc?
- The nature of the distribution of a variable may provide some indication of the transformation that should be applied to it. For example, section 3.2 showed that *wage* is non-normal but that its log is normal. This suggests re-specifying the model by using the log of *wage* instead of *wage*.
- Check your textbook for more methods.

Ch 7. Section 4.d. Omitted variable

Not including a variable that actually plays a role in explaining the dependent variable can bias the regression results. Methods of detection [103] include:

- Perform a preliminary visual test. To do this, we asked SPSS for the plot ZPRED and Y-PRED while running the regression (see section 7.1). Any pattern in this plot implies mis-specification (and/or heteroskedasticity) due to the use of an incorrect functional form or due to the omission of a relevant variable.
- If the visual test indicates a problem, perform a formal diagnostic test such as the RESET test.
- Apply your intuition, previous research, hints from preliminary bivariate analysis, etc. For example, in the model we ran, we believe that there may be an omitted variable bias because of the absence of two crucial variables for wage determination - whether the labor is unionized and the professional sector of work (medicine, finance, retail, etc.).
- Check your textbook for more methods.

[100] In section 8.3, you will learn how to use square and log transformations to remove mis-specification.

[101] The test requires the variables “predicted Y” and “predicted residual.” We obtained these when we asked SPSS to save the "unstandardized" predicted dependent variable and the unstandardized residuals, respectively (see section 7.1).

[102] Check your textbook for other formal tests.

[103] The first three tests are similar to those for Incorrect Functional form.

Ch 7. Section 4.e. Inclusion of an irrelevant variable

This mis-specification occurs when a variable that is not actually relevant to the model is included[104]. To detect the presence of irrelevant variables:

- Examine the significance of the T-statistics. If the T-statistic is not significant at the 10% level (usually if T< 1.64 in absolute terms), then the variable may be irrelevant to the model.

Ch 7. Section 4.f. Measurement error

This is not a very severe problem if it only afflicts the dependent variable, but it may bias the T-statistics. Methods of detecting this problem include:

- Knowledge about problems/mistakes in data collection
- There may be a measurement error if the variable you are using is a proxy for the actual variable you intended to use. In our example, the wage variable includes the monetized values of the benefits received by the respondent. But this is a subjective monetization of respondents and is probably undervalued. As such, we can guess that there is probably some measurement error.
- Check your textbook for more methods

Ch 7. Section 4.g. Heteroskedasticity

Note: Heteroskedasticity implies that the variances (i.e. - the dispersion around the expected mean of zero) of the residuals are not constant, but that they are different for different observations. This causes a problem: if the variances are unequal, then the relative reliability of each observation (used in the regression analysis) is unequal. The larger the variance, the lower should be the importance (or weight) attached to that observation. As you will see in section 8.2, the correction for this problem involves the downgrading in relative importance of those observations with higher variance. The problem is more apparent when the value of the variance has some relation to one or more of the independent variables. *Intuitively, this is a problem because the distribution of the residuals should have no relation with any of the variables (a basic assumption of the classical model).*

Detection involves two steps:

- Looking for patterns in the plot of the predicted dependent variable and the residual (the partial plots discussed in section 7.2)
- If the graphical inspection hints at heteroskedasticity, you must conduct a formal test like the White's test. Section 7.5 teaches you how to conduct a White's test[105]. Similar multi-step methods are used for formally checking for other breakdowns.

[104] By dropping it, we improve the reliability of the T-statistics of the other variables (which are relevant to the model). But, we may be causing a far more serious problem - an omitted variable! An insignificant T is not necessarily a bad thing - it is the result of a "true" model. Trying to remove variables to obtain only significant T-statistics is bad practice.

[105] Other tests: Park, Glejser, Goldfelt-Quandt. Refer to your text book for a comprehensive listing of methods and their detailed descriptions.

Ch 7. Section 5 Checking formally for heteroskedasticity: White's test

The least squares regression we ran indicated the presence of heteroskedasticity because of the patterns in the partial plots of the residual with regards to the variables *education* and *work_ex*. We must run a formal test to confirm our suspicions and obtain some indication of the nature of the heteroskedasticity.

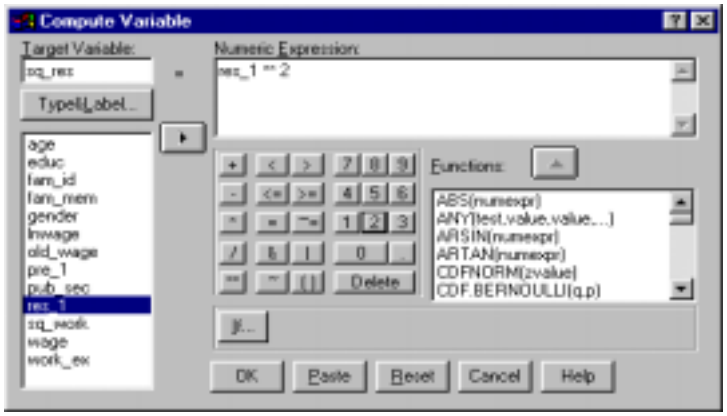
The White's test is usually used as a test for heteroskedasticity. In this test, a regression of the squares of the residuals[106] is run on the variables suspected of causing the heteroskedasticity, their squares, and cross products.

$$(\text{residuals})^2 = b_0 + b_1\, educ + b_2\, work\_ex + b_3\,(educ)^2 + b_4\,(work\_ex)^2 + b_5\,(educ*work\_ex)$$

To run this regression, several new variables must be created. This is a limitation of SPSS - many tests that are done with the click of a button in E-Views and with simple code in SAS must be done from scratch in SPSS. This applies to the tests for mis-specification (RESET and DW tests) and other tests for heteroskedasticity.

Go to TRANSFORM/ COMPUTE[107].

Create the new variable *sqres* (square of residual).

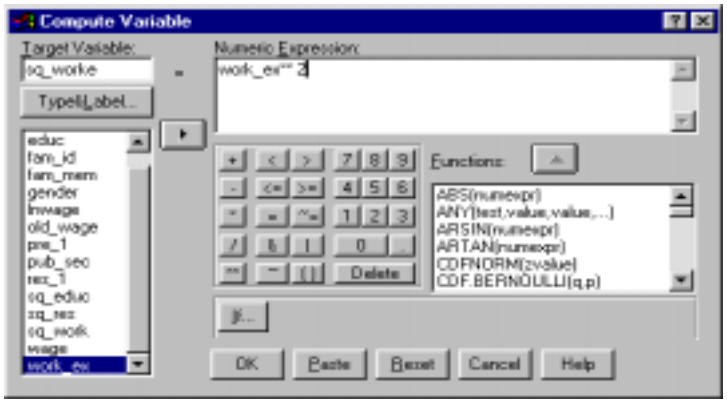


[106] The test requires the variables "predicted residual." We obtained this when we asked SPSS to save the unstandardized residuals (see section 7.1).

[107] If you are unfamiliar with this procedure, please refer to section 2.2.

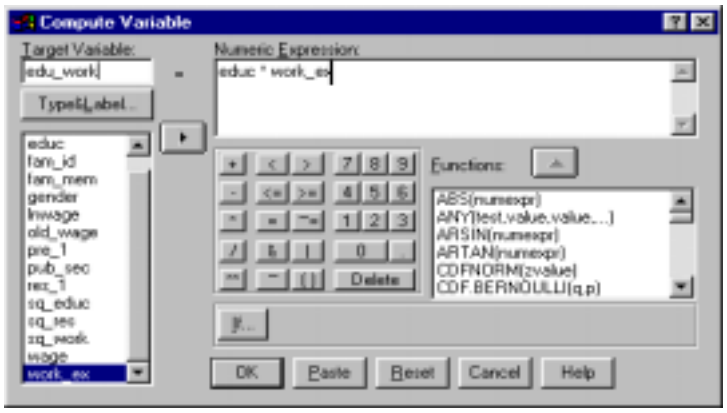
Create *sq_worke* (square of *work experience*).

Similarly, create *sq_educ* (square of *educ*).



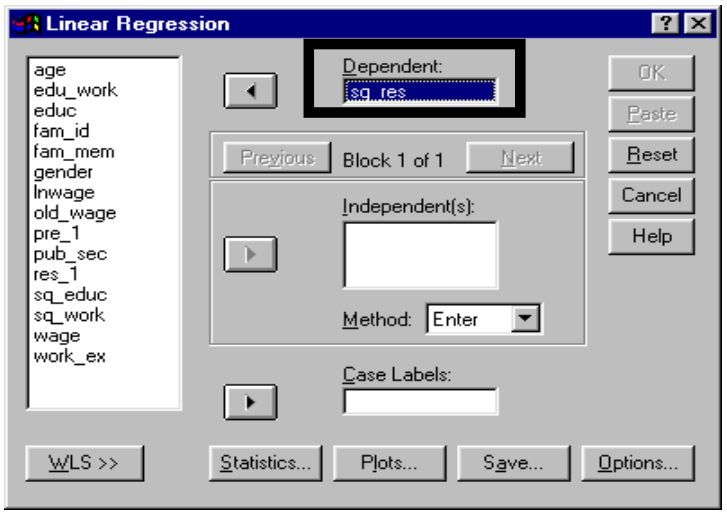
Create the cross product of *educ* and *work_ex*.

Now you are ready to do the White's test - you have the dependent variable square of the residuals, the squares of the independent variables, and their cross products.

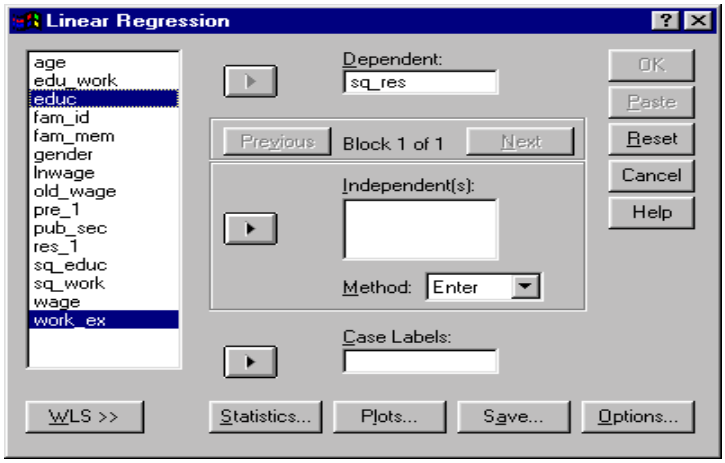


Go to STATISTICS/ REGRESSION/ LINEAR.

Place the variable *sq_res* into the box "Dependent."

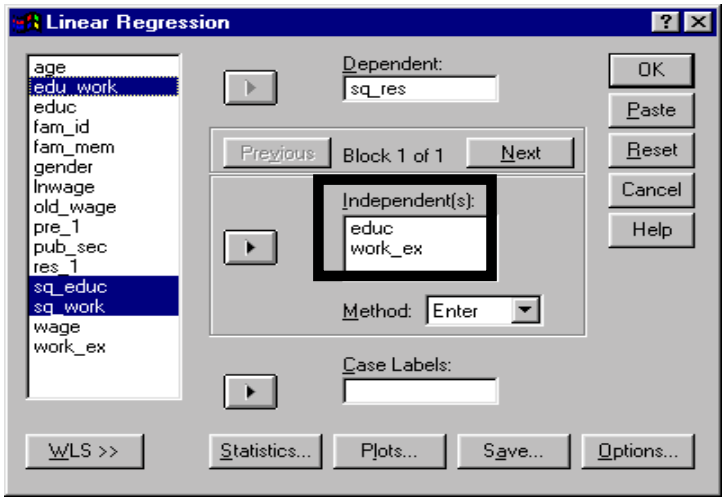


Select the variables *educ* and *work_ex* and move them into the box “Independent(s)."



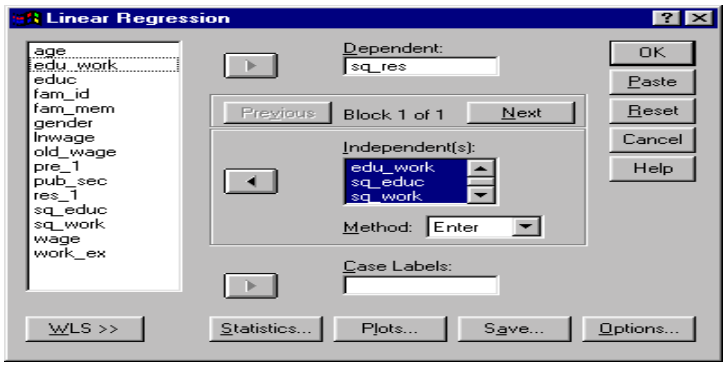
Place the variables *sq_educ*, *sq_work* and *edu_work* into the box “Independents."

Note: On an intuitive level, what are we doing here? We are trying to determine whether the absolute value of the residuals ("absolute" because we use the squared residuals) can be explained by the independent variable(s) in the original case. This should not be the case because the residuals are supposedly random and non-predictable.



Click on “OK."

Note: We do not report the F-statistic and the table ANOVA as we did in section 7.2 (it is significant). If the F was not significant here, should one still proceed with the White's test? We think you can argue both ways, though we would lean towards not continuing with the test and concluding that "there is no heteroskedasticity."



Model Summary [a]

Variables Entered	R Square	Adjusted R Square	Std. Error of the Estimate
SQ_WORK, SQ_EDUC, EDU_WORK, Work Experience, EDUCATION	*.037*	*.035*	*.2102*

a. Dependent Variable: SQ_RES

White's Test

- Calculate $n*R^2$ → $R^2 = 0.037$, n=2016 → Thus, $n*R^2 = .037*2016 = 74.6$.
- Compare this value with $\chi^2(n)$, i.e. with $\chi^2(2016)$ (χ^2 is the symbol for the Chi-Square distribution)

$\chi^2(2016) = 124$ obtained from χ^2 table. (For 955 confidence) As $n*R^2 < \chi^2$, heteroskedasticity can not be confirmed.

Note: Please refer to your textbook for further information regarding the interpretation of the White's test. If you have not encountered the Chi-Square distribution/test before, there is no need to panic! The same rules apply for testing using any distribution - the T, F, Z, or Chi-Square. First, calculate the required value from your results. Here the required value is the sample size ("n") multiplied by the R-square. You must determine whether this value is higher than that in the standard table for the relevant distribution (here the Chi-Square) at the recommended level of confidence (usually 95%) for the appropriate degrees of freedom (for the White's test, this equals the sample size "n") in the table for the distribution (which you will find in the back of most econometrics/statistics textbooks). If the former is higher, then the hypothesis is rejected. Usually the rejection implies that the test could not find a problem[108].

To take quizzes on topics within each chapter, go to http://www.spss.org/wwwroot/spssquiz.asp

[108] We use the phraseology "Confidence Level of "95%." Many professors may frown upon this, instead preferring to use "Significance Level of 5%." Also, our explanation is simplistic. Do not use it in an exam! Instead, refer to the chapter on "Hypothesis Testing" or "Confidence Intervals" in your textbook. A clear understanding of these concepts is essential.

Ch 8. CORRECTING FOR BREAKDOWN OF CLASSICAL ASSUMPTIONS

A regression result is not acceptable unless the estimation satisfies the assumptions of the Classical Linear regression model. In sections 7.4 through 7.5, you learned how to diagnose the viability of the model by conducting tests to determine whether these assumptions are satisfied.

In the introduction to this chapter, we place some notes containing intuitive explanations of the reasons why the breakdowns cause a problem. (These notes have light shading.) Our explanations are too informal for use in an exam. Our explanation may not satisfy a statistician, but we hope it gets the intuitive picture across. We include them here to help you understand the problems more clearly.

Why is the result not acceptable unless the assumptions are met? The reason is simple - the strong statements inferred from a regression (e.g. - "an increase in one unit of the value of variable X causes an increase of the value of variable Y by 0.21 units") depend on the presumption that the variables used in a regression, and the residuals from that regression, satisfy certain statistical properties. These are expressed in the properties of the distribution of the residuals. *That explains why so many of the diagnostic tests shown in sections 7.4-7.5 and their relevant corrective methods, shown in this chapter, are based on the use of the residuals.* If these properties are satisfied, then we can be confident in our interpretation of the results. The above statements are based on complex, formal mathematical proofs. Please refer to your textbook if you are curious about the formal foundations of the statements.

If a formal[109] diagnostic test confirms the breakdown of an assumption, then you must attempt to correct for it. This correction usually involves running another regression on a transformed version of the original model, with the exact nature of the transformation being a function of the classical regression assumption that has been violated[110].

In section 8.1, you will learn how to correct for collinearity (also called multicollinearity)[111].

Note: Collinearity means that two or more of the independent/explanatory variables in a regression have a linear relationship. This causes a problem in the interpretation of the regression results. If the variables have a close linear relationship, then the estimated regression coefficients and T-statistics may not be able to properly isolate the unique impact/role of each variable and the confidence with which we can presume these impacts to be true. The close relationship of the variables makes this isolation difficult.

[109] Usually, a "formal" test uses a hypothesis testing approach. This involves the use of testing against distributions like the T, F, or Chi-Square. An "informal' test typically refers to a graphical test.

[110] Don't worry if this line confuses you at present - its meaning and relevance will become apparent as you read through this chapter.

[111] We have chosen this order of correcting for breakdowns because this is the order in which the breakdowns are usually taught in schools. Ideally, the order you should follow should be based upon the degree of harm a particular breakdown causes. First, correct for mis-specification due to incorrect functional form and simultaneity bias. Second, correct for mis-specification due to an omitted variable and measurement error in an independent variable. Third, correct for collinearity. Fourth, correct for heteroskedasticity and measurement error in the dependent variable. Fifth, correct for the inclusion of irrelevant variables. Your professor may have a different opinion.

In section 8.2 you will learn how to correct for heteroskedasticity.

Note: Heteroskedasticity implies that the variances (i.e. - the dispersion around the expected mean of zero) of the residuals are not constant - that they are different for different observations. This causes a problem. If the variances are unequal, then the relative reliability of each observation (used in the regression analysis) is unequal. The larger the variance, the lower should be the importance (or weight) attached to that observation. As you will see in section 8.2, the correction for this problem involves the downgrading in relative importance of those observations with higher variance. The problem is more apparent when the value of the variance has some relation to one or more of the independent variables. *Intuitively, this is a problem because the distribution of the residuals should have no relation with any of the variables (a basic assumption of the classical model).*

In section 8.3 you will learn how to correct for mis-specification due to incorrect functional form.

Mis-specification covers a list of problems discussed in sections 8.3 to 8.5. These problems can cause moderate or severe damage to the regression analysis. Of graver importance is the fact that most of these problems are caused not by the nature of the data/issue, but by the modeling work done by the researcher. It is of the utmost importance that every researcher realise that the responsibility of correctly specifying an econometric model lies solely on them. A proper specification includes determining curvature (linear or not), functional form (whether to use logs, exponentials, or squared variables), and the measurement accuracy of each variable, etc.

Note: Why should an incorrect functional form lead to severe problems? Regression is based on finding coefficients that minimize the "sum of squared residuals." Each residual is the difference between the predicted value (the regression line) of the dependent variable versus the realized value in the data. If the functional form is incorrect, then each point on the regression "line" is incorrect because the line is based on an incorrect functional form. A simple example: assume Y has a log relation with X (a log curve represents their scatter plot) but a linear relation with "Log X." If we regress Y on X (and not on "Log X"), then the estimated regression line will have a systemic tendency for a bias because we are fitting a straight line on what should be a curve. The residuals will be calculated from the incorrect "straight" line and will be wrong. If they are wrong, then the entire analysis will be biased because everything hinges on the use of the residuals.

Section 8.4 teaches 2SLS, a procedure that corrects for simultaneity bias.

Note: Simultaneity bias may be seen as a type of mis-specification. This bias occurs if one or more of the independent variables is actually dependent on other variables in the equation. For example, we are using a model that claims that income can be explained by investment and education. However, we might believe that investment, in turn, is explained by income. If we were to use a simple model in which income (the dependent variable) is regressed on investment and education (the independent variables), then the specification would be incorrect because investment would not really be "independent" to the model - it is affected by income. Intuitively, this is a problem because the simultaneity implies that the residual will have some relation with the variable that has been incorrectly specified as "independent" - the residual is capturing (more in a metaphysical than formal mathematical sense) some of the unmodeled reverse relation between the "dependent" and "independent" variables.

Section 8.5 discusses how to correct for other specification problems: measurement errors, omitted variable bias, and irrelevant variable bias.

Note: Measurement errors causing problems can be easily understood. Omitted variable bias is a bit more complex. Think of it this way - the deviations in the dependent variable are in reality

explained by the variable that has been omitted. Because the variable has been omitted, the algorithm will, mistakenly, apportion what should have been explained by that variable to the other variables, thus creating the error(s). Remember: our explanations are too informal and probably incorrect by strict mathematical proof for use in an exam. We include them here to help you understand the problems a bit better.

Our approach to all these breakdowns may be a bit too simplistic or crude for purists. We have striven to be lucid and succinct in this book. As such, we may have used the most common methods for correcting for the breakdowns. Please refer to your textbook for more methods and for details on the methods we use.

Because we are following the sequence used by most professors and econometrics textbooks, we first correct for collinearity and heteroskedasticity. Then we correct for mis-specification. It is, however, considered standard practice to correct for mis-specification first. It may be helpful to use the table in section 7.3 as your guide.

Also, you may sense that the separate sections in this chapter do not incorporate the corrective procedures in the other sections. For example, the section on misspecification (section 8.3) does not use the WLS for correcting for heteroskedasticity (section 8.2). The reason we have done this is to make each corrective procedure easier to understand by treating it in isolation. In practice, you should always incorporate the features of corrective measures.

Ch 8. Section 1 Correcting for collinearity

Collinearity can be a serious problem because it biases the T-statistics and may also bias the coefficient estimates.

The variables *age* and *work experience* are correlated (see section 7.3). There are several[112] ways to correct for this. We show an example of one such method: "Dropping all but one of the collinear variables from the analysis[113]."

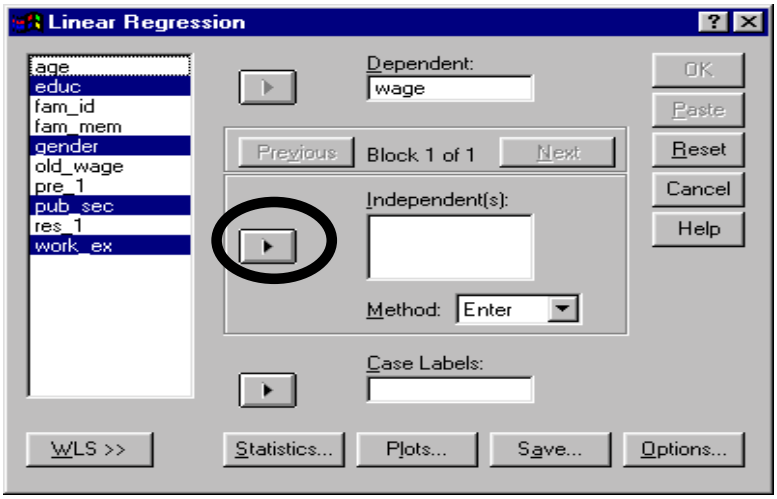
[112] Sometimes adding new data (increasing sample size) and/or combining cross-sectional and time series data can also help reduce collinearity. Check your textbook for more details on the methods mentioned here.

[113] Warning--many researchers, finding that two variables are correlated, drop one of them from the analysis. However, the solution is not that simple because this may cause mis-specification due to the omission of a relevant variable (that which was dropped), which is more harmful than collinearity.

Ch 8. Section 1.a. Dropping all but one of the collinear variables from the model

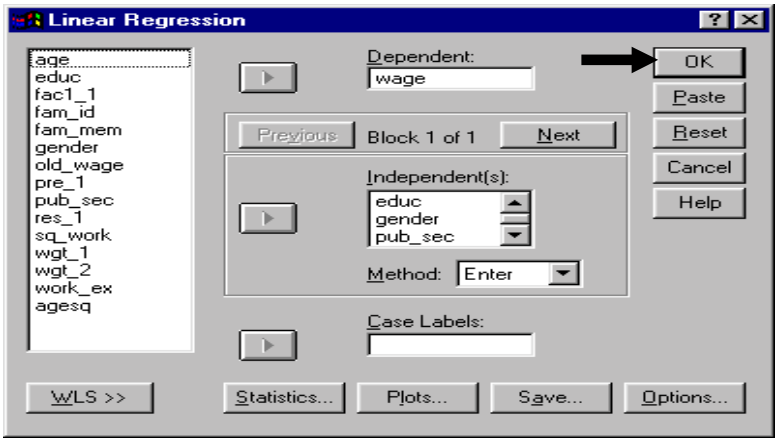
Go to STATISTICS/REGRESSION/ LINEAR.

Choose the variables for the analysis. First click on *educ*. Then, press CTRL, and while keeping it pressed, click on *gender*, p*ub_sec*, and *work_ex*. Do not choose the variable *age* (we are dropping it because it is collinear with *work experience*). Click on the arrow to choose the variables.



Repeat all the other steps from section 7.1.

Click on "OK."



We know the model is significant because the "Sig." of the F-statistic is below .05.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	*52552.19*	*4*	*13138.05*	*481.378*	*.000*
	Residual	*54257.68*	*1988*	*27.293*		
	Total	*106809.9*	*1992*			

a. Dependent Variable: WAGE

b. Independent Variables: (Constant), WORK_EX, EDUCATION, GENDER, PUB_SEC

Although the adjusted R-square has dropped, this is a better model than the original model (see sections 7.1 and 7.2) because the problem of collinear independent variables does not bias the results here.

Reminder: it is preferable to keep the collinear variables in the model if the option is Omitted Variable bias because the latter has worse implications, as shown in section 7.3.

Model Summary[a,b]

Model	Variables Entered	Variables Removed	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	*WORK_EX, EDUCATION, GENDER, PUB_SEC*[c]	.	.701	.492	.491	5.2242

a. Dependent Variable: WAGE

b. Method: Enter

c. Independent Variables: (Constant), WORK_EX, EDUCATION, GENDER, PUB_SEC

The coefficients have changed slightly from the original model (see sections 7.1 and 7.2). A comparison is worthless, because the coefficients and/or their T-statistics were unreliable in the model in chapter 7 because of the presence of collinearity.

Note: we have suppressed other output and its interpretation. Refer back to sections 7.1 and 7.2 for a recap on those topics.

Coefficients[a]

Model		Unstandardized Coefficients B	Std. Error	t	Sig.	95% Confidence Interval for B Lower Bound	Upper Bound
1	(Constant)	1.196	.237	5.055	.000	.732	1.660
	EDUCATIO	.746	.025	30.123	.000	.697	.794
	GENDER	-1.955	.294	-6.644	.000	-2.532	-1.378
	PUB_SEC	2.331	.289	8.055	.000	1.763	2.898
	WORK_EX	.196	.013	14.717	.000	.169	.222

a. Dependent Variable: WAGE

Ch 8. Section 2 Correcting for heteroskedasticity

In our model, the variable *education* is causing heteroskedasticity. The partial plot in section 7.2 showed that "as education increases, the residuals also increase," but the exact pattern of the plot was not clear.

Because we are following the sequence used by most professors and econometrics textbooks, we have first corrected for collinearity and heteroskedasticity. We will later correct for mis-specification. It is, however, considered standard practice to correct for mis-specification first as it has the most severe implications for interpretation of regression results. It may be helpful use the table in section 7.3 as your guide.

Ch 8. Section 2.a. WLS when the exact nature of heteroskedasticity is not known

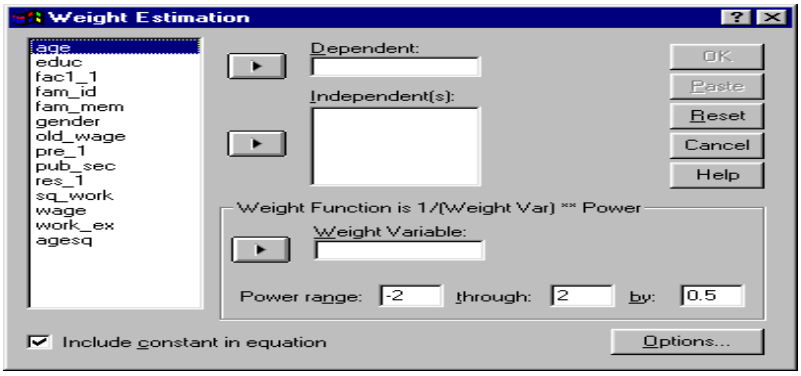
We believe that education is causing heteroskedasticity, but we do not know the pattern. As the weighting variable, what transformation of education should we use? Some options include:

- *Education*
- $Education^{0.5}$

- *Education*[1.5]

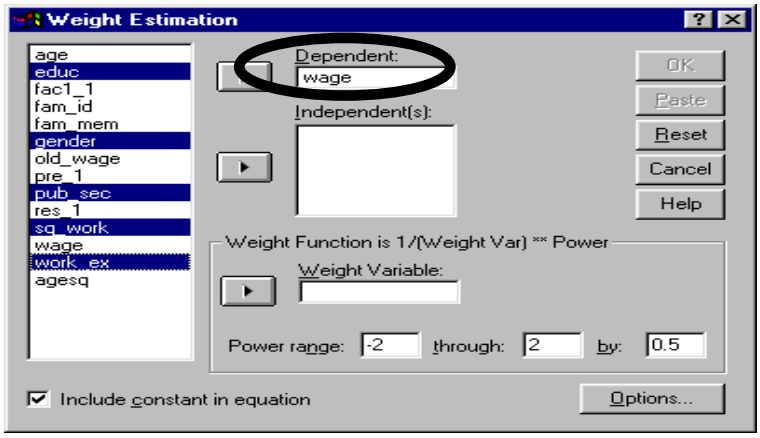
We firmly believe that *education* should be used[114], and we further feel that one of the above three transformations of *education* would be best. We can let SPSS take over from here[115]. It will find the best transformation of the three above, and then run a WLS regression with no threat of heteroskedasticity.

Go to
STATISTICS/REGRESSION/
WEIGHT ESTIMATION

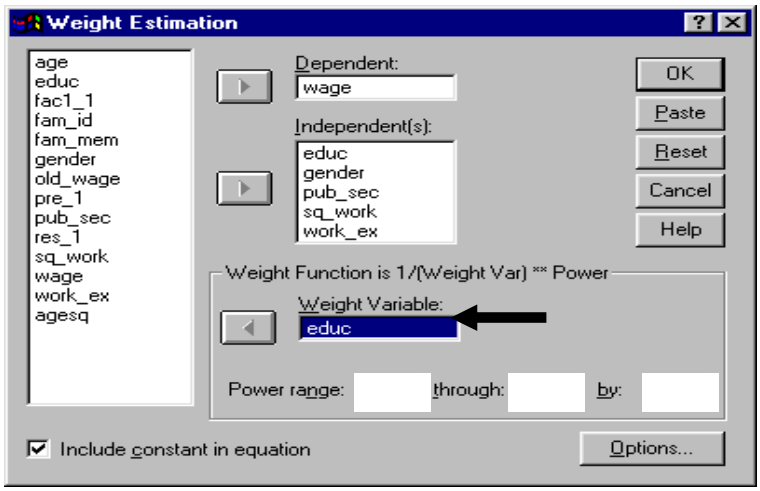


Select the variable *wage* and place it in the box for the "Dependent" variable.

Select the independent variables and place them into the box "Independents."



Move the variable *educ* into the box "Weight Variable."



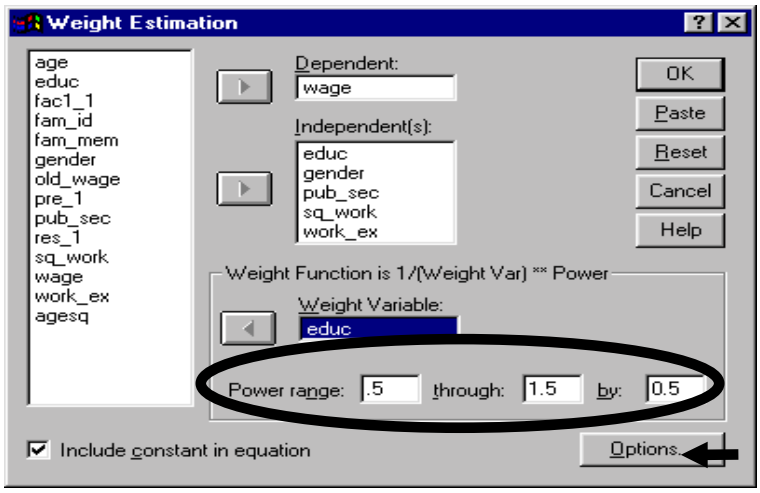
[114] See sections 7.2 and 7.5 for justification of our approach.

[115] There exists another approach to solving for heteroskedasticity: *White's Heteroskedasticity Consistent Standard Errors*. Using this procedure, no transformations are necessary. The regression uses a formula for standard errors that automatically corrects for heteroskedasticity. *Unfortunately, SPSS does not offer this method/procedure.*

In our example, the pattern in the plot of residual versus education hints at a power between .5 and 1.5 (See section 7.2).

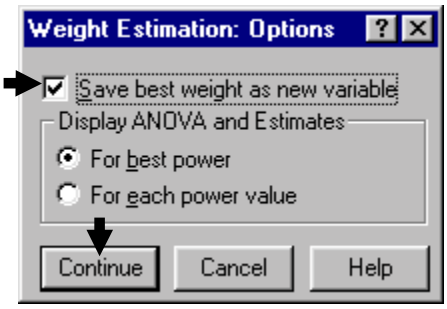
To provide SPSS with the range within which to pick the best transformation, enter "Power Range .5 through 1.5 by .5." This will make SPSS look for the best weight in the range from powers of .5 to 1.5 and will increment the search by .5 each time[116].

Click on "Options."



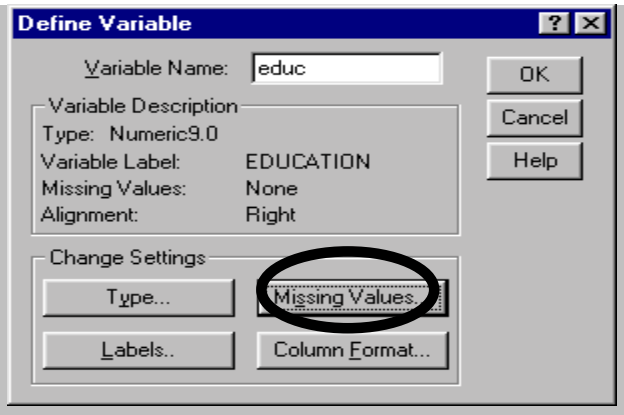
Select "Save best weight as new variable." This weight can be used to run a WLS using STATISTICS / REGRESSION / LINEAR or any other appropriate procedure.

Click on "Continue."



A problem will arise if we use the above weights: if education takes the value of zero, the transformed value will be undefined. To avoid this, we remove all zero values of education from the analysis. This may bias our results, but if we want to use only education or its transformed power value as a weight, then we must assume that risk.

To redefine the variable education, choose the column with the data on education and then go to DATA/ DEFINE VARIABLE (See section 1.2.). Click on "Missing Values."

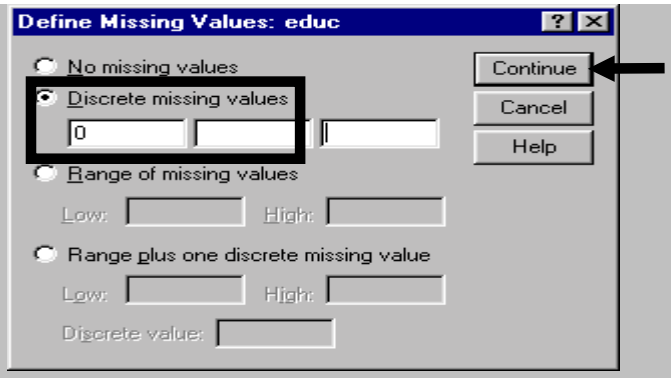


[116] SPSS will search through

- .5+0=.5
- .5+.5 = 1 and
- .5+.5+.5 = 1.5

Enter zero as a missing value. Now, until you come back and redefine the variable, all zeroes in education will be taken to be missing values.

Note: this digressive step is not always necessary. We put it in to show what must be done if SPSS starts producing messages such as "Weighted Estimation cannot be done. Cannot divide by zero."

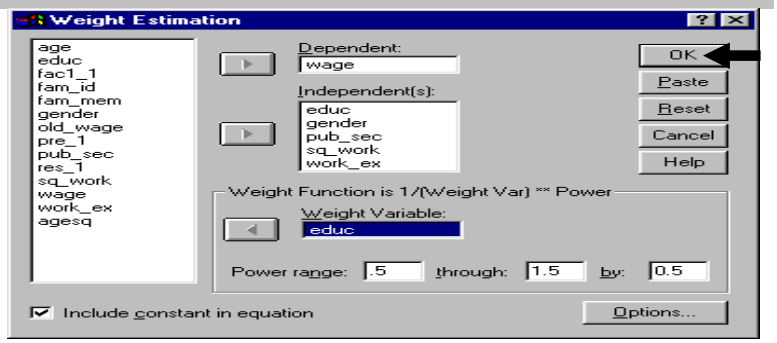


Now go back to STATISTICS/REGRESSION/ WEIGHT ESTIMATION

Re-enter the choices you made before moving on to re-define the variable.

Click on "OK."

Note: Maximum Likelihood Estimation (MLE) is used. This differs from the Linear Regression methodology used elsewhere in chapters 7 and 8. You will learn a bit more about MLE in chapter 9.




```
Source variable. EDUC                 Dependent variable. WAGE

Log-likelihood Function =-5481        POWER value =   .5
Log-likelihood Function =-5573        POWER value =  1
Log-likelihood Function =-5935        POWER value =  1.5

The Value of POWER Maximizing Log-likelihood Function =   .5

Source variable.    EDUC                  POWER value =   .5

Dependent variable. WAGE

R Square              .451
Adjusted R Square     .449
Standard Error       3.379

            Analysis of Variance:

                 DF   Sum of Squares        Mean Square

Regression        5        17245            3449.04
Residuals      1836        20964            11.41

F =     302      Signif F =  .0000

------------------ Variables in the Equation ------------------

Variable         B        SE B       Beta         T         Sig. T
```

The best weight is education to the power .5.

The model is significant

```
EDUC            .687      .025          .523      26.62       .0000
GENDER        -1.564      .247        - .110      -6.36       .0000
PUB_SEC        2.078      .273          .151       7.61       .0000
SQ_WORK         -.004     .0008       - .280      -5.54       .0000
WORK_EX         .293      .031          .469       9.20       .0000
(Constant)     1.491      .242                     6.14       .0000

      Log-likelihood Function =-5481

The following new variables are being created:
  Name        Label

  WGT_2       Weight for WAGE from WLS, MOD_2  EDUC**  -.500
```

All the variables are significant

Each coefficient can be interpreted directly (compare this to the indirect method shown at the end of section 8.2.b.). The results do not suffer from heteroskedasticity. Unfortunately, the output is not so rich (there are no plots or output tables produced) as that obtained when using STATISTICS/REGRESSION/LINEAR (as in the earlier sections of this chapter, chapter 7, and section 8.2.b).

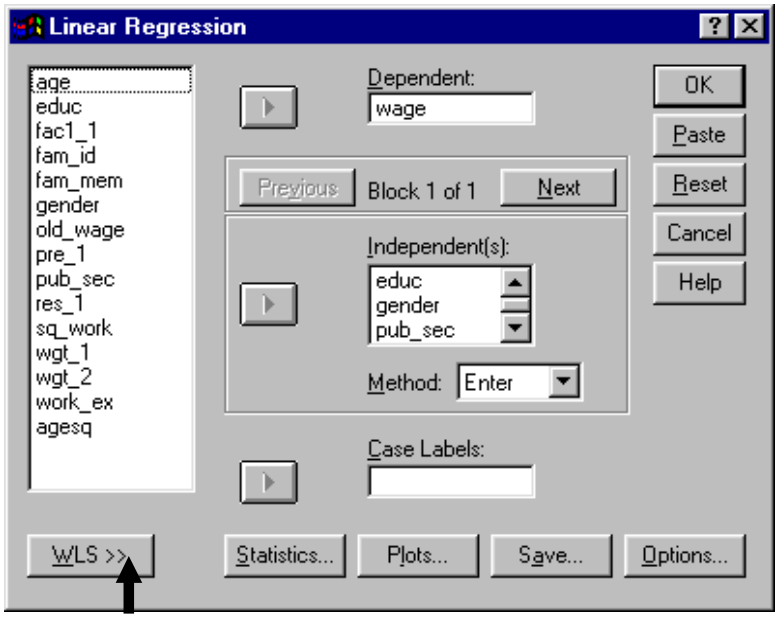
A new variable *wgt_2* is created. This represents the best heteroskedasticity-correcting power of *education*.

Ch 8. Section 2.b. Weight estimation when the weight is known

If the weight were known for correcting heteroskedasticity, then WLS can be performed directly using the standard linear regression dialog box.

Go to STATISTICS/REGRESSION/ LINEAR.

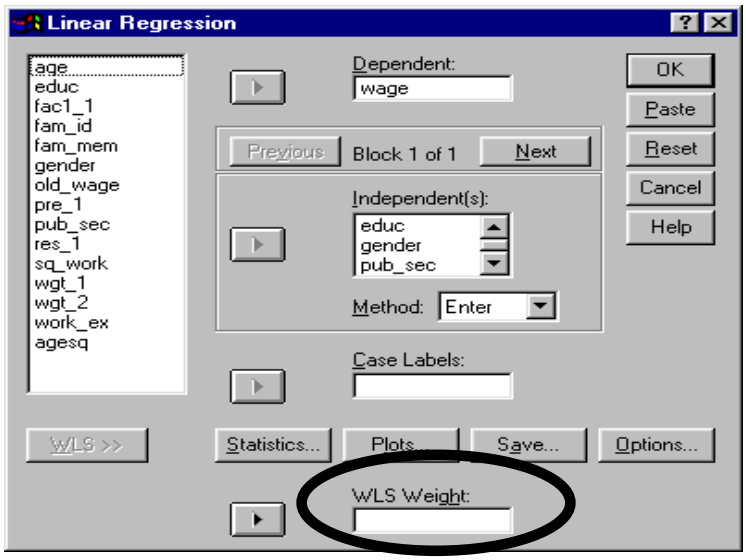
Click on the button "WLS."



[117] The weight is = $(1/(education)^{.5} = education^{-.5}$

A box labeled "WLS Weight" will open up at the bottom of the dialog box.

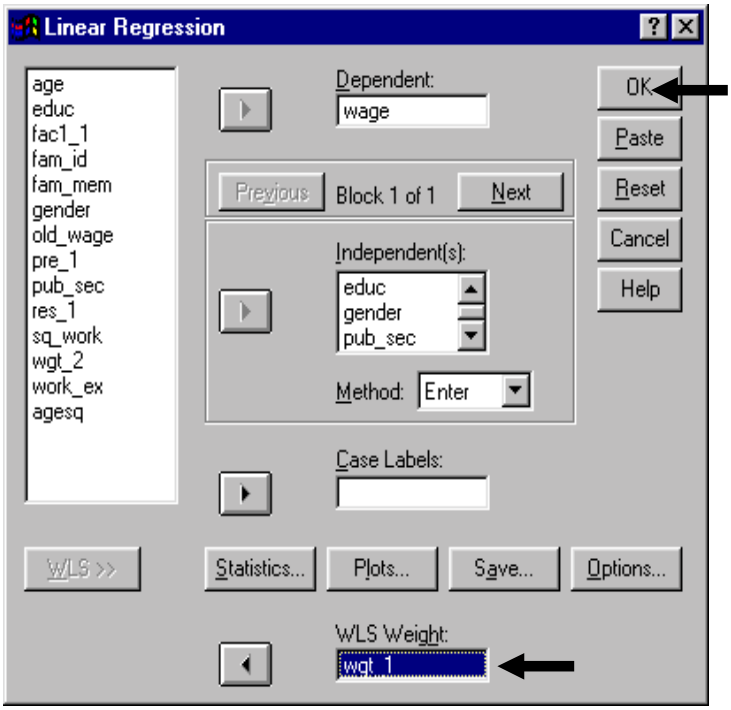
The weight variable is to be placed here.



Place the weight variable in the box "WLS Weight."

Repeat all other steps from section 7.1.

Press "OK."



The variables have been transformed in WLS. Do not make a direct comparison with the OLS results in the previous chapter.

To make a comparison, you must map the new coefficients on the "real" coefficients on the original (unweighted) variables. This is in contrast to the direct interpretation of coefficients in section 8.2.a. Refer to your econometrics textbook to learn how to do this.

Coefficients[a,b]

Model		Unstandardized Coefficients		t	Sig.	95% Confidence Interval for B	
		B	Std. Error			Lower Bound	Upper Bound
1	(Constant)	*-3.571*	*.849*	*-4.207*	*.000*	*-5.235*	*-1.906*
	EDUCATION	*.694*	*.026*	*26.251*	*.000*	*.642*	*.746*
	GENDER	*-1.791*	*.245*	*-7.299*	*.000*	*-2.272*	*-1.310*
	PUB_SEC	*1.724*	*.279*	*6.176*	*.000*	*1.177*	*2.272*
	AGESQ	*-3.0E-03*	*.001*	*-4.631*	*.000*	*-.004*	*-.002*
	AGE	*.328*	*.049*	*6.717*	*.000*	*.232*	*.423*

a. Dependent Variable: WAGE

b. Weighted Least Squares Regression - Weighted by Weight for WAGE from WLS, MOD_1 EDUC** -.500

Note: other output suppressed and not interpreted. Refer to section 7.2 for detailed interpretation guidelines.

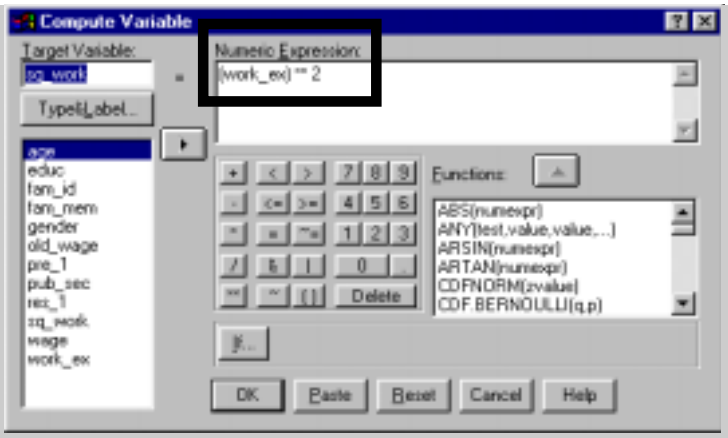
Ch 8. Section 3 Correcting for incorrect functional form

Because we are following the sequence used by most professors and econometrics textbooks, we have first corrected for collinearity and heteroskedasticity. We will now correct for mis-specification. It is, however, considered standard practice to correct for mis-specification first. It may be helpful use the table in section 7.3 as your guide. You may sense that the separate sections in this chapter do not incorporate the corrective procedures in the other sections. For example, this section does not use WLS for correcting for heteroskedasticity. The reason we have done this is to make each corrective procedure easier to understand by treating it in isolation from the other procedures. In practice, you should always incorporate the features of all corrective measures.

We begin by creating and including a new variable, *square of work experience*[118]. The logic is that the incremental effect on *wages* of a one-year increase in *experience* should reduce as the experience level increases.

[118] Why choose this transformation? Possible reasons for choosing this transformation: a hunch, the scatter plot may have shown a slight concave curvature, or previous research may have established that such a specification of age is appropriate for a wage determination model.

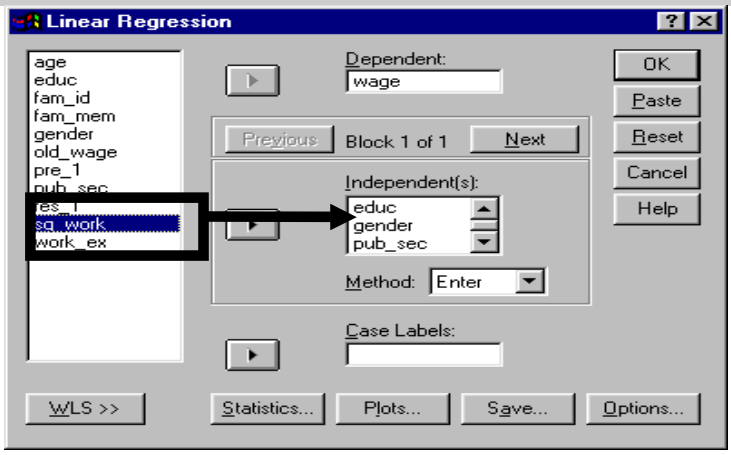
First, we must create the new variable "square of work experience." To do so, go to TRANSFORM/COMPUTE. Enter the label *sq_work* in to the box "Target variable" and the formula for it in the box "Numeric Expression." See section 2.2 for more on computing variables.



Now we must go back and run a regression that includes this new variable.

Go to STATISTICS/REGRESSION/LINEAR. Move the variable you created (*sq_work)* into the box of independent variables. Repeat all other steps from section 7.1.

Click on "OK."



We cannot compare the results of this model with those of the mis-specified model (see sections 7.1 and 7.2) because the latter was biased.

Although the addition of the new variable may not increase adjusted R-square, and may even lower it, this model is superior to the one in earlier sections (7.1 and 8.1).

Model Summary [a,b]

Model	Variables Entered	Variables Removed	R Square	Adjusted R Square	Std. Error of the Estimate
1	*SQ_WORK, EDUCATION, GENDER, PUB_SEC, WORK_EX* [c,d]	.	*.503*	*.501*	*5.1709*

a. Dependent Variable: WAGE

b. Method: Enter

c. Independent Variables: (Constant), SQ_WORK, EDUCATION, GENDER, PUB_SEC, WORK_EX

d. All requested variables entered.

The coefficient on *sq_work* is negative and significant, suggesting that the increase in wages resulting from an increase in *work_ex* decreases as *work_ex* increases.

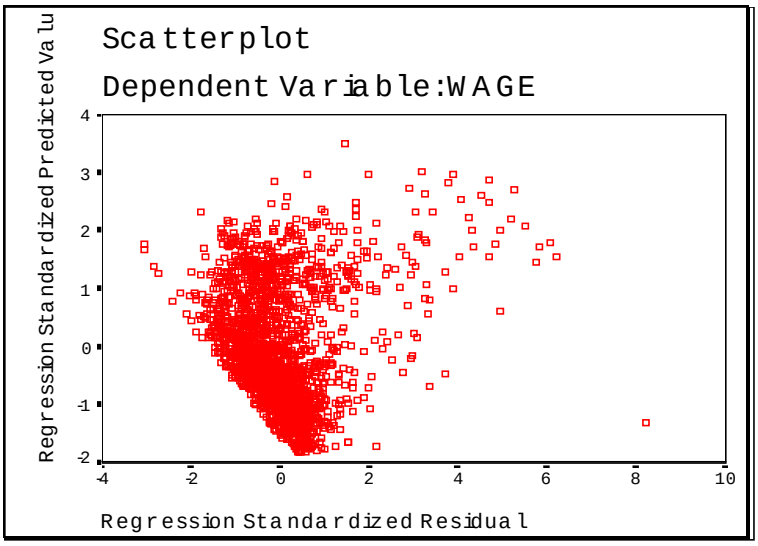
Coefficients[a]

Model		Unstandardized Coefficients		t	Sig.	95% Confidence Interval for B	
		B	Std. Error			Lower Bound	Upper Bound
1	(Constant)	.220	.278	.791	.429	-.326	.766
	EDUCATION	.749	.025	30.555	.000	.701	.797
	GENDER	-1.881	.291	-6.451	.000	-2.452	-1.309
	PUB_SEC	2.078	.289	7.188	.000	1.511	2.645
	WORK_EX	.422	.037	11.321	.000	.349	.495
	SQ_WORK	-7.1E-03	.001	-6.496	.000	-.009	-.005

a. Dependent Variable: WAGE

The ZPRED-ZRESID still has a distinct pattern, indicating the presence of mis-specification.

We used the square of the variable *work experience* to correct for mis-specification. This did not solve the problem[119].



What else may be causing mis-specification? Omitted variable bias may be a cause. Our theory and intuition tells us that the nature of the wage-setting environment (whether unionized or not) and area of work (law, administration, engineering, economics, etc.) should be relevant variables, but we do not have data on them.

Another cause may be the functional form of the model equation. Should any of the variables (apart from *age*) enter the model in a non-linear way? To answer this, one must look at:

- The models used in previous research on the same topic, possibly with data on the same region/era, etc.
- Intuition based on one's understanding of the relationship between the variables and the manner in which each variable behaves
- Inferences from pre-regression analyses such as scatter-plots

[119] We only did a graphical test. For formal tests like the RESET test, see a standard econometrics textbook like Gujarati. The test will require several steps, just as the White's Test did in section 7.5.

In our case, all three aspects listed below provide support for using a log transformation of *wages* as the dependent variable.

- Previous research on earnings functions has successfully used such a transformation and thus justified its use.
- Intuition suggests that the absolute change in *wages* will be different at different levels of wages. As such, comparing percentage changes is better than comparing absolute changes. This is exactly what the use of logs will allow us to do.
- The scatters showed that the relations between wage and education and between wage and work experience are probably non-linear. Further, the scatters indicate that using a log dependent variable may be justified. We also saw that *wage* is not distributed normally but its log is. So, in conformity with the classical assumptions, it is better to use the log of *wages*.

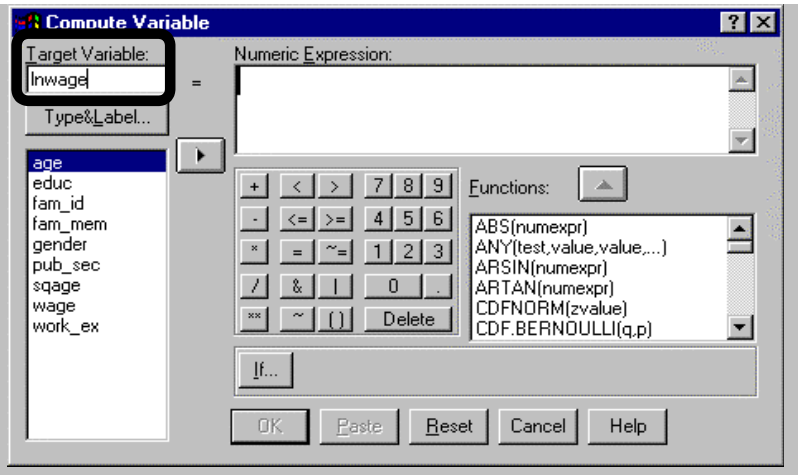
Arguably, mis-specification is the most debilitating problem an analysis can incur. As shown in section 7.3, it can bias all the results. **Moreover, unlike measurement errors, the use of an incorrect functional form is a mistake for which the analyst is to blame.**

To run the re-specified model, we first must create the log transformation of *wage*.

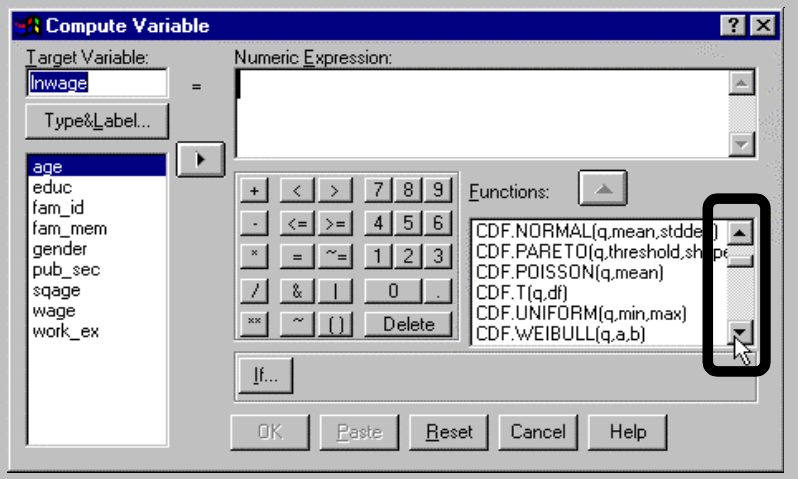
Note: The creation of new variables was shown in section 2.2. We are repeating it here to reiterate the importance of knowing this procedure.

Go to TRANSFORM/COMPUTE.

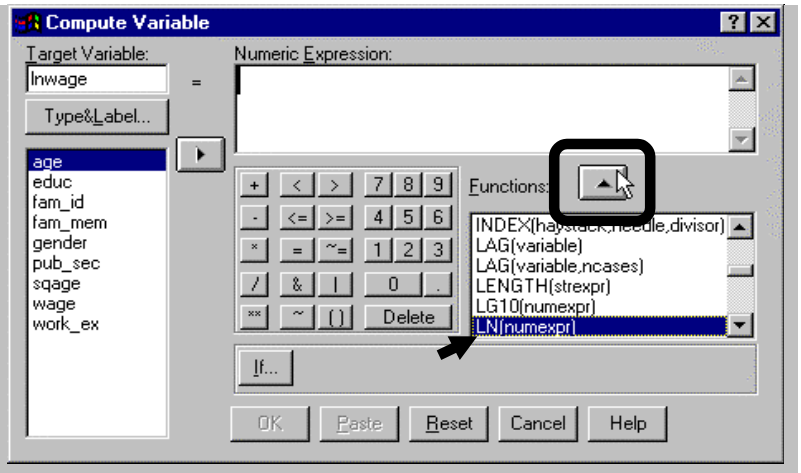
Enter the name of the new variable you wish to create.



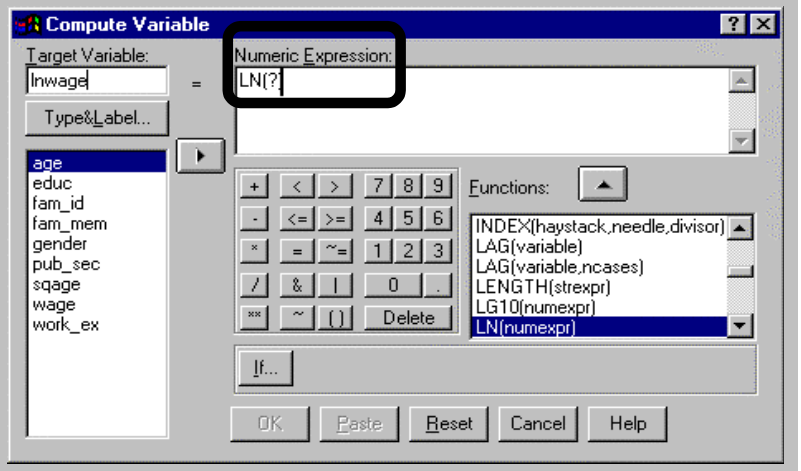
In the box "Numeric Expression," you must enter the function for logs. To find it, scroll in the box "Functions."



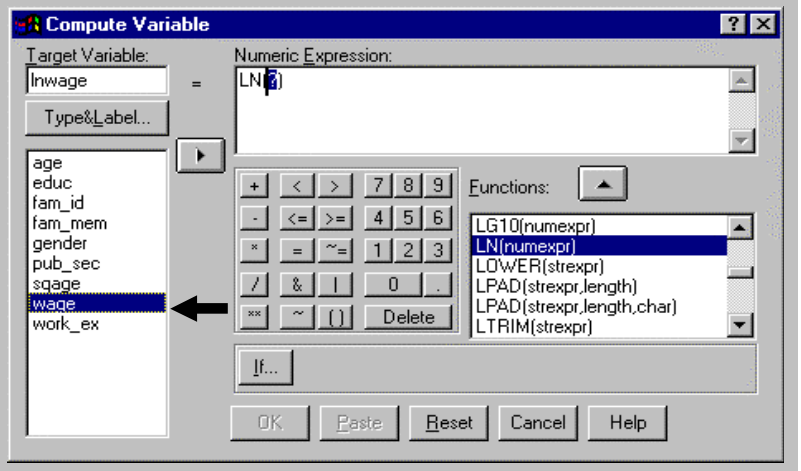
Select the function "LN" and click on the upward arrow.



The log function is displayed in the box "Numeric Expression."



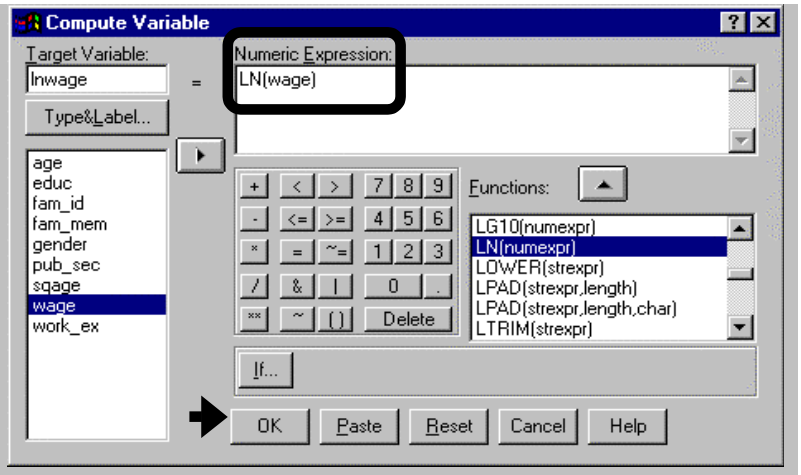
Click on the variable *wage*.



Click on the arrow pointing to the right.

The expression is complete.

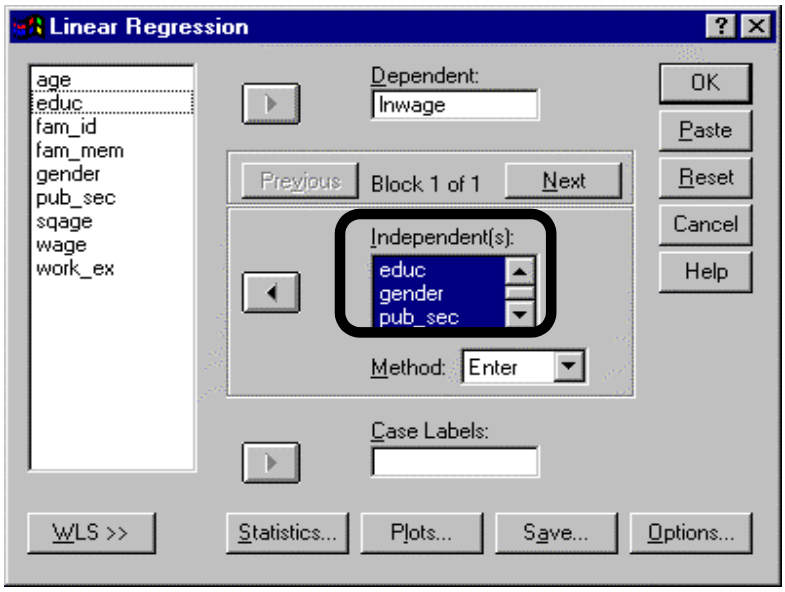
Click on "OK."



Now that the variable *lnwage* has been created, we must run the re-specified model.

Go to STATISTICS/ LINEAR/REGRESSION. Move the newly created variable *lnwage* into the box "Dependent."

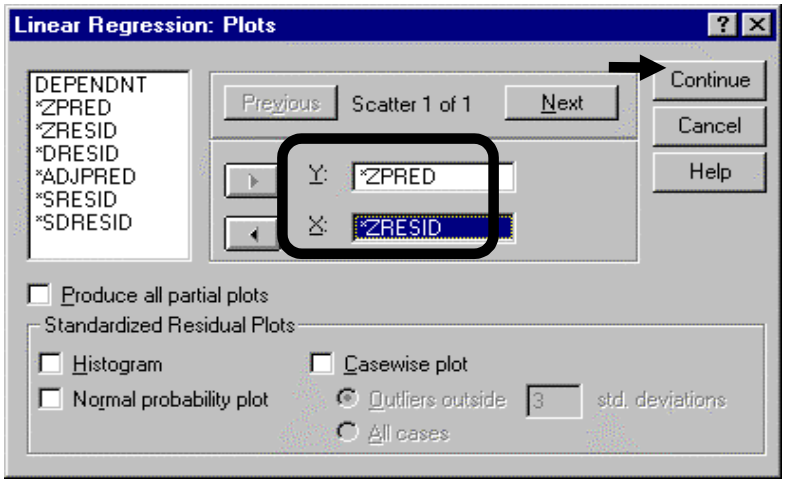
Select the independent variables. They are the same as before. Choose other options as shown in section 7.1.



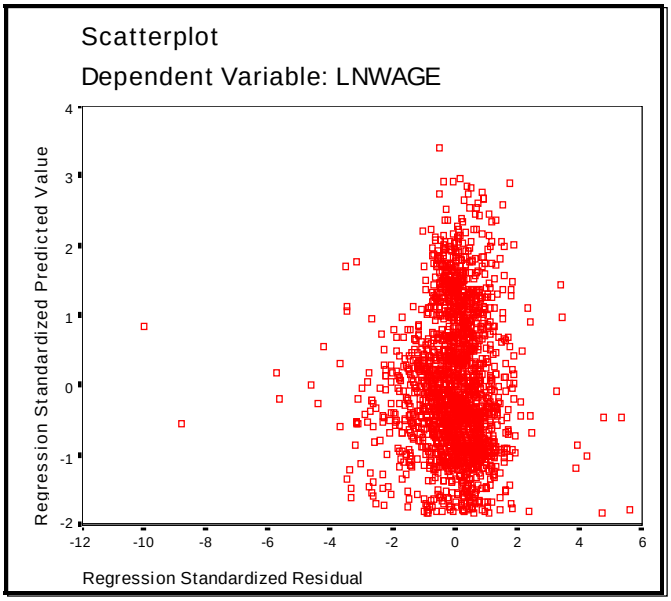
In particular, choose to plot the standardized predicted (ZPRED) against the standardized residual (ZRESID). This plot will indicate whether the mis-specification problem has been removed.

Click on "Continue."

Click on "OK."



The plot of predicted versus residual shows that the problem of mis-specification is gone!



Now the results can be trusted. They have no bias due to any major breakdown of the classical assumptions.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	*732.265*	*5*	*146.453*	*306.336*	*.000[b]*
	Residual	*960.463*	*2009*	*.478*		
	Total	*1692.729*	*2014*			

a. Dependent Variable: LNWAGE

b. Independent Variables: (Constant), Work Experience, EDUCATION, GENDER, Whether Public Sector Employee, SQAGE

Model Summary [a]

	Variables Entered	R Square	Adjusted R Square	Std. Error of the Estimate
	Work Experience, EDUCATION, GENDER, Whether Public Sector Employee, SQAGE	*.433*	*.431*	*.6914*

a. Dependent Variable: LNWAGE

Coefficients [a]

Model		Unstandardized Coefficients		t	Sig.	95% Confidence Interval for B	
		B	Std. Error			Lower Bound	Upper Bound
1	(Constant)	.938	.035	26.473	.000	.868	1.007
	EDUCATION	8.309E-02	.003	25.589	.000	.077	.089
	GENDER	-.368	.039	-9.478	.000	-.444	-.291
	Whether Public Sector Employee	.283	.039	7.285	.000	.207	.359
	SQAGE	1.445E-04	.000	6.015	.000	.000	.000
	Work Experience	1.197E-02	.002	5.269	.000	.008	.016

a. Dependent Variable: LNWAGE

None of the regression results before this (in chapter 7 and sections 8.1-8.2) can be compared to this, as they were all biased due to mis-specification. **This is the most important issue in regression analysis. Focus your attention on diagnosing and correcting for the breakdowns in the classical assumptions (and not on the R-square).**

Ch 8. Section 4 Correcting for simultaneity bias: 2SLS

2-stage least squares is used when the residual is correlated with one or more of the independent variables because of simultaneity bias. Simultaneity bias occurs when one of the independent variables is not truly independent and, instead, is a function of one or more other independent variables.

Let's assume you want to estimate the model:

wage = function(*education, work experience*)

But what if the "independent" variable *education* is actually "dependent" on the variable *gender*? Using the equation above would then be incorrect because one of the right-hand-side variables (*education*) is not truly independent. If you just ran the equation above, simultaneity bias would result, severely compromising the reliability of your results.

Instead, using 2SLS, you can run the real model that consists of two equations, one to explain *wage* and another to explain *education*:

wage = function(*education, work experience*)

education = function(*gender*)

The above model is run using 2SLS:

1. In 2SLS, SPSS first runs a regression of *education* on all of the independent variables[120] (first stage regression), and saves the predicted *education*.

Education = function(*gender, work experience*) ➔ pred(*education*)

[120] In our example, *gender* and *work experience*.

2. Then, in the second stage regression, it will run the regression of interest to us - *wage* on *work experience* and the predicted *education* from the first regression.

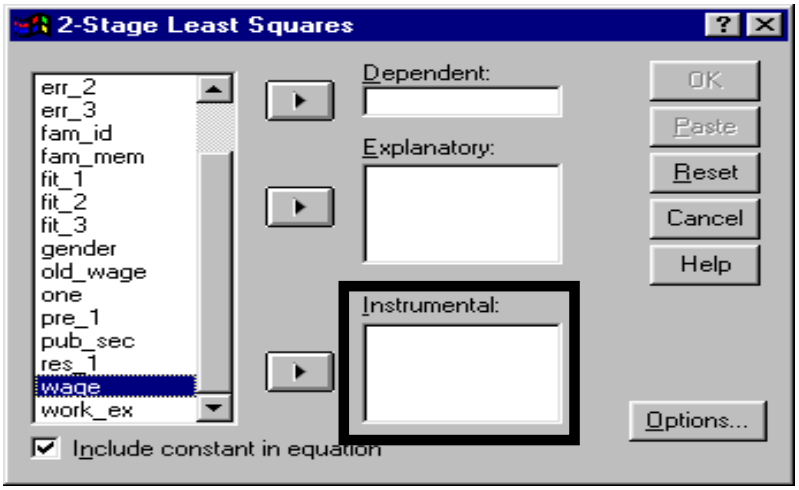
   Wage = function(*work experience*, pred(*education*))

   The output will only report one result:

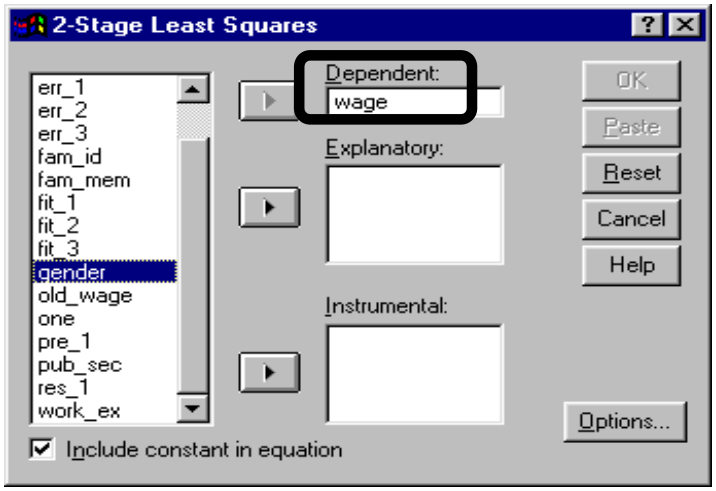
Wage = function(*work experience*, *education*)

Go to STATISTICS/ REGRESSION/ 2SLS.

Note the box "Instrumental." You do not see this label in any other procedure. This box is where you put the "proxies" for the variables in the main equation.

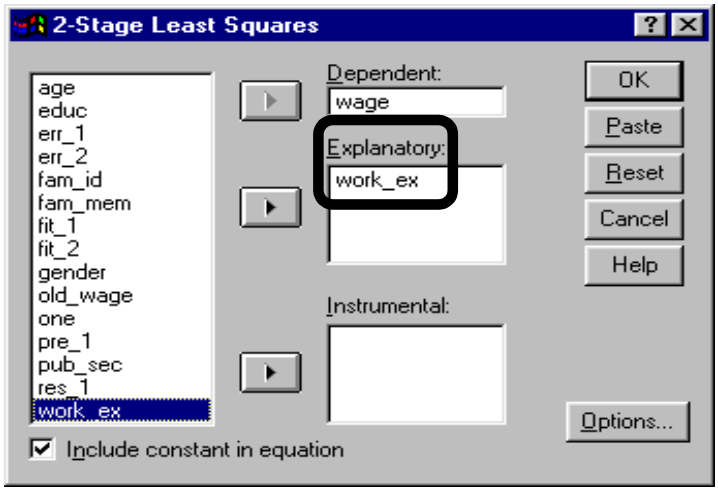


Move the variable *wage* into the box "Dependent."



Move the first explanatory variable (*work_ex*) into the box "Explanatory."

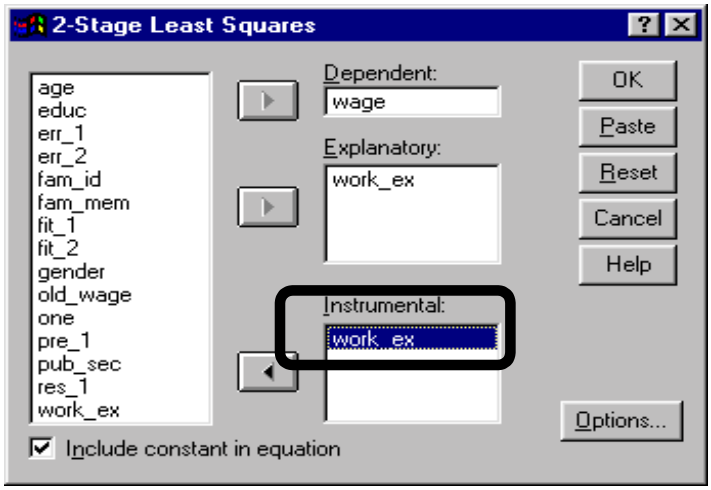
As of now, you have the model:
wage = function(*work_ex*).



Move the same variable into the box "Instrumental."

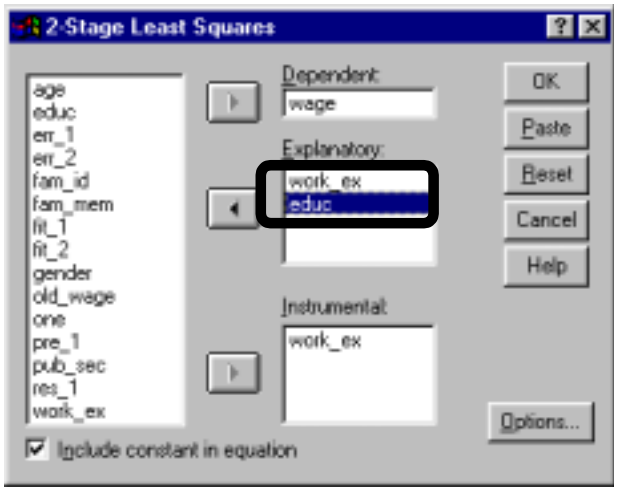
The fact that *work_ex* is its own instrument implies that it is truly independent in our model and is not a cause of the correlation between the residual and independent variables.

As of now, you still have the same model:
wage = function(*work_ex*).

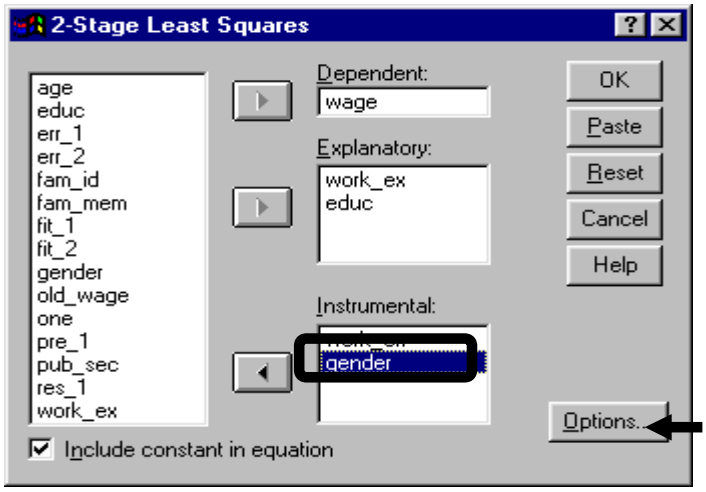


Move the variable *educ* into the box "Explanatory."

As of now, you have the model:
wage = function (*work_ex*, *education*)



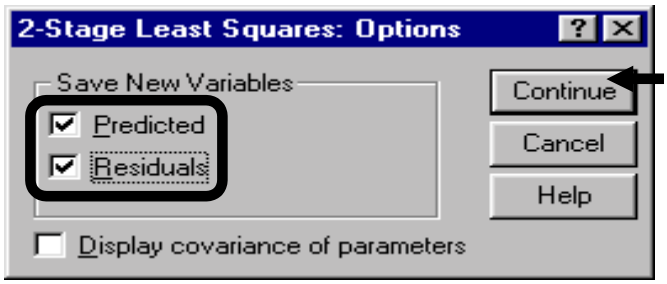
Educ is presumed to be a function of gender. As a result, we place *gender* as the instrument for the variable *educ*.



Effectively, the model is now run as "*wage* as a function of work experience and *education*, which itself has been transformed to correct for the influence of *gender*." **By not placing *educ* in the box "Instrumental" (but only in "Explanatory"), we have implicitly told SPSS that it is an endogenous variable. Consequently, the first-stage regression will be: *educ* on <u>all</u> the variables in the area "Instrumental."**

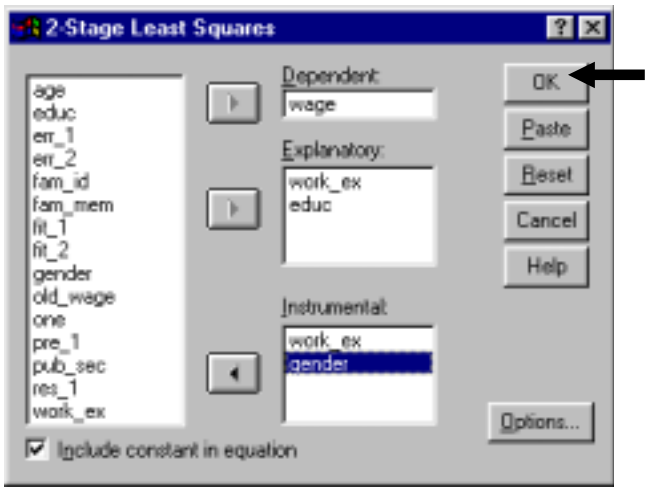
Click on the button "Options."
Request the predicted *wage* and *residuals* by clicking on the boxes to the left of the labels "Predicted" and "Residuals."

Click on "Continue."



Click on "OK."

If our model is correct, then the results are startling: once the influence of *gender* has been accounted for, *education* levels and *work experience* do not make a significant contribution to *wage* determination.




```
Dependent variable. WAGE

R Square              .0084
Adjusted R Square     .0074
Standard Error      20.7887

            Analysis of Variance:

                 DF   Sum of Squares      Mean Square

Regression        2          7210.5       3605.2
Residuals      1965        849218.1        432.1
```

```
F =       8.34220       Signif F =  .0002
```

The model is significant

```
------------------ Variables in the Equation ------------------

Variable             B        SE B       Beta         T       Sig. T

EDUC                -2.79     2.161    -2.103     -1.29        .196
WORK_EX               .093     .095      .116       .97        .329
(Constant)          24.090   13.91                1.73        .083

The following new variables are being created:

  Name        Label
  FIT_3       Fit for WAGE from 2SLS, MOD_5 Equation 1
```

Do not worry if the R-square is too "low." The R-square is a function of the model, the data, sample size, etc. It is better to have a properly specified model (one conforming to the classical assumptions) with a low R-square compared to an improperly specified model with a high R-square. Honesty is a good policy - trying to inflate the R-square is a bad practice that an incredible number of economists have employed (including so-called experts at Universities and major research institutes).

Ch 8. Section 5 Correcting for other breakdowns

We provide some tips for correcting other breakdowns. The list is not exhaustive, nor are the explanations descriptive. Please refer to your textbook for more information.

Ch 8. Section 5.a. Omitted Variable

As we saw in section 7.3, the absence of a relevant variable is harmful for a linear regression. This is a difficult problem to overcome if the data on any omitted "relevant" variable is difficult to obtain. Our analysis may have some omitted variables such as unionisation, family background, etc. Using proxy variables may be one way to avoid this problem.

Be careful not to cause this problem inadvertently while correcting for the problems of collinearity[121] or the inclusion of an irrelevant variable.

Please refer to your textbook for more information.

Ch 8. Section 5.a. Irrelevant Variable

As we saw in section 7.3, the presence of an irrelevant variable is not very harmful for a linear regression. As a result, it may be reasonable to do nothing.

The other option would be to remove the "irrelevant" variables, a distressingly common practice. Be careful - this approach has two problems:

[121] When you use the correction method of dropping "all but one" of the collinear variables from the model.

- If, by error, one removes a "relevant" variable, then we may be introducing an omitted variable bias, a far worse breakdown in comparison to the presence of an irrelevant variable.

- A tendency to remove all variables that have an insignificant T-statistic may result in a choice to ignore theory and instead use statistics to construct regression models, an incorrect approach. The aim of regression analysis is to prove/support certain theoretical and intuitive beliefs. All models should be based upon these beliefs.

The fact that the T is insignificant is itself a result. It shows that that variable does not have a significant effect. Or, it can be interpreted as "the impact of the variable as measured by the beta coefficient is not reliable because the estimated probability distribution of this beta has a standard error that is much too high."

Please refer to your textbook for more information.

Ch 8. Section 5.b. Measurement error in dependent variable

This is not a major problem (see section 7.3). It can be ignored, or a proxy variable can be used. For example, it may be better to use accurate *GNP* compared to mis-measured *GDP*. However, this may be a limitation one has to live with.

Please refer to your textbook for more information.

Ch 8. Section 5.c. Measurement error in independent variable(s)

This is a serious problem (see section 7.3) that is often ignored by researchers. One manner of getting around this problem would be to use Instrumental Variables. These are proxies for the mis-measured variables and must have two properties:

- high correlation with the mis-measured variable

- low correlation with the residuals

Just replace the independent variable with its proxy.

Please refer to your textbook for more information.

Your professor may scoff at the simplicity of some of our approaches. In cases of conflict, always listen to the person who is grading your work.

To take quizzes on topics within each chapter, go to http://www.spss.org/wwwroot/spssquiz.asp

Ch 9. MLE: Logit and Non-Linear Regression

Linear regression cannot be used for estimating relationships when:

1. The dependent[122] variable is not continuous and quantitative. Rather, it is qualitative (dichotomous or categorical)[123]. In such situations, the Logit model/method is used. This is discussed in section 9.1.
2. When the functional form that captures the relation between the variables cannot be modeled in a linear equation. That is, in intuitive and simple terms, when the regression "line" is not a straight line but is, instead, a curve. The use of this method is shown in section 9.2.

All of these methods use an estimation technique called Maximum Likelihood Estimation (MLE)[124], an advanced algorithm that calculates the coefficients that would maximize the likelihood of viewing the data distributions as seen in the data set. MLE is a more powerful method than linear regression. More importantly, it is not subject to the same degree to the classical assumptions (mentioned *ad nauseam* in chapters 7 and 8) that must be met for a reliable Linear Regression.

The output from MLE differs from that of Linear Regression. In addition, since these models are not based on properties of the residuals, as is the case with OLS, there are different goodness-of-fit tests. We will not delve into the details of MLE and related diagnostics. Those topics are beyond the scope of this book.

Ch 9. Section 1 Logit

Logit (also called logistic) estimates models in which the dependent variable is a dichotomous dummy variable - the variable can take only two values, 1 and 0. These models are typically

[122] When an independent variable is a dummy, it can be used in a linear regression without a problem as long as it is coded properly (as 0 and 1). What is the problem if the dependent variable is a dummy? If we run such a regression, the predicted values will lie within and in the vicinity of the two values of the original dependent variable, namely the values 0 and 1. What is the best interpretation of the predicted value? Answer: "The probability that the dependent variable takes on the quality captured by the value 1." In a linear regression, such predicted probabilities may be estimated at values less than 0 or greater than 1, both nonsensical. Also, for reasons we will not delve into here, the R-square cannot be used, normality of the residuals is compromised, and a severe case of heteroskedasticity is always present. For all these reasons, the linear regression should not be used. A stronger and simpler argument is that imposing a linear regression on what is a non-linear model (as will become apparent with the Logit example later) constitutes serious mis-specification (incorrect functional form).

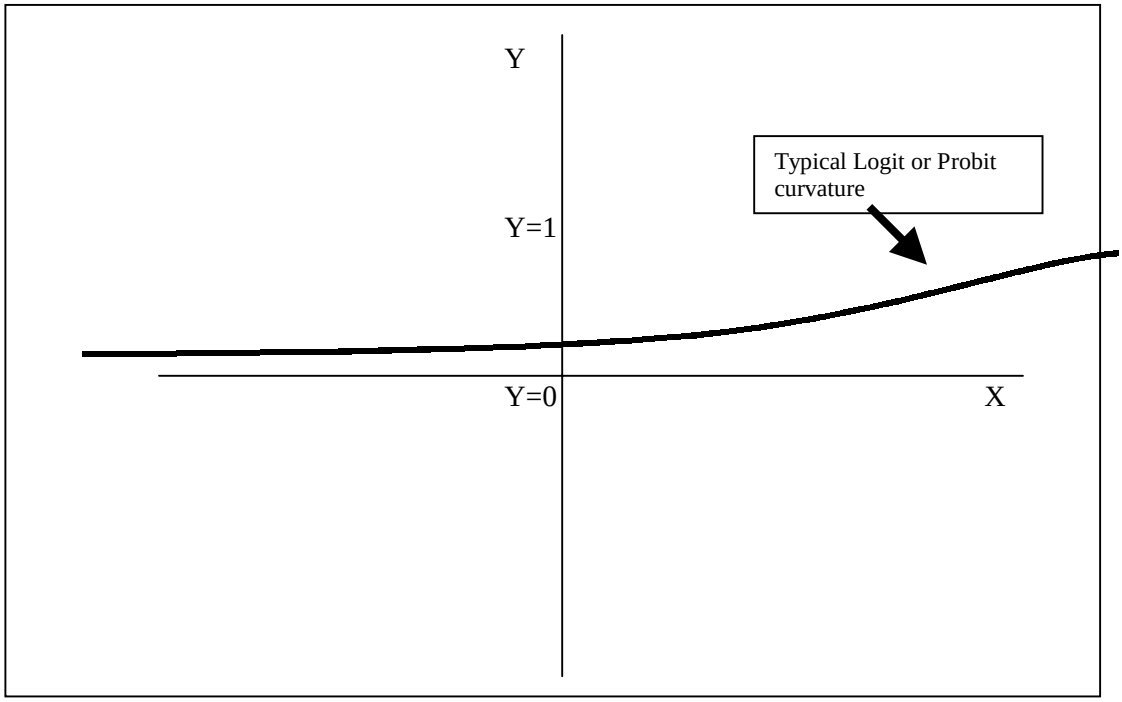
[123] Dummy and categorical variables are also called "Qualitative" variables because the values of the variable describe a quality and not a quantity. For example, the dummy variable gender can take on two values - 0 and 1. The former if the respondent is male and the latter if the respondent is a female, both of which are qualities.

[124] You can estimate a linear model using the procedure "Non-Linear Regression." This may be useful if you want to show that “the results are robust in the sense that the estimates from Least Squares Estimation (linear regression) and MLE are the same” or if violations of the classical assumptions are proving particularly difficult to overcome.

used to predict whether or not some event will occur, such as whether a person will vote "yes" or "no" on a particular referendum, or whether a person will graduate this year (or not) from high school, etc.

The other model used for estimating models with dichotomous models is the Probit. The Logit and Probit techniques are similar and both use Maximum Likelihood Estimation methods. The Logit is used more frequently because it is easier to interpret. That is why we only show the Logit in this book.

In any case, the Probit procedure in SPSS (STATISTICS/REGRESSION/PROBIT) is for analysis with grouped data where the dependent variable is a calculated "proportion of cases" and not a dichotomy. We will include an example of such a Probit model in the next edition.



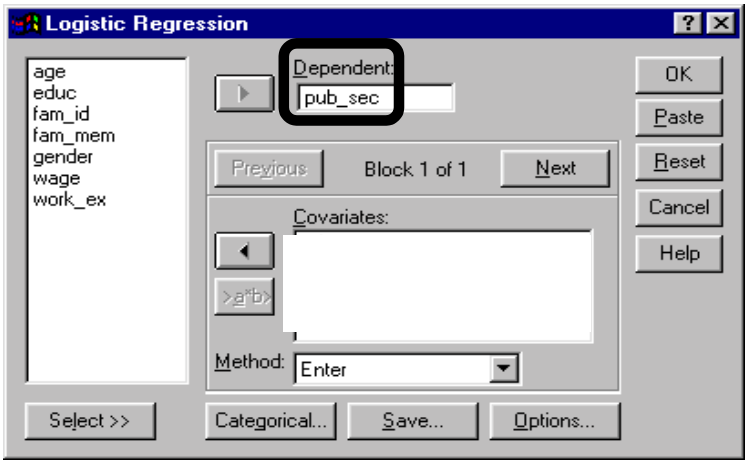
If you look at a graph of the Logit or Probit (see graph above), you will notice a few striking features: as the value on the X-axis increases, the value on the Y-axis gradually tends towards 1 but never reaches it. Conversely, as the value on the X-axis tends towards negative infinity, the Y-value never drops below zero. The fact that the Y-value remains inside the bounds of 0 and 1 provides the intuitive rationale for using the Logit or Probit. The X-axis represents the independent variable(s) and the Y represents the probability of the dependent variable taking the value of 1. Because of the nature of the curve, the probability always remains within the range of 0 and 1, regardless of the values of the independent variables. This is a requirement for estimating the predicted value of a dummy variable because the predicted value is interpreted as a probability. A probability must lie between 0 and 1.

Pub_sec is a dichotomous dummy variable, coded as 0 if the respondent is an employee of the private sector and 1 if the respondent is an employee of the public sector. Let's assume you want to estimate the impact of *gender* and *education* on the probability of working in the public sector (relative to working in the private sector). You can perform this estimation by running a Logit in which the dependent variable is *Pub_sec* and the independent variables are *gender* and *education*.

Go to STATISTICS/REGRESSION/ LOGISTIC.

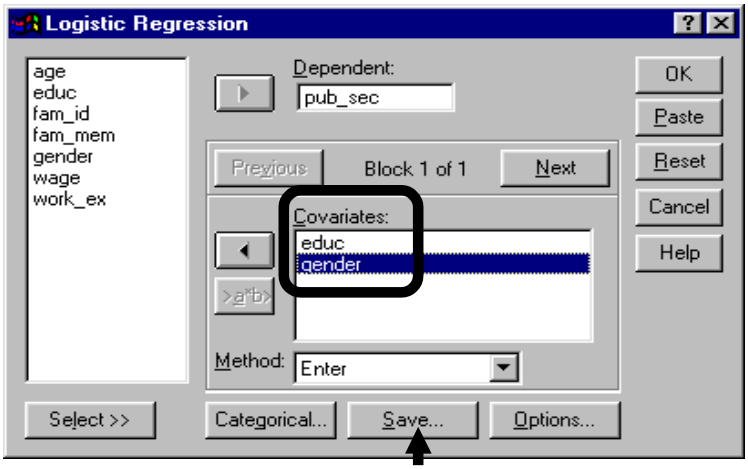
The following dialog box will open.

Pub_sec is the dependent variable. Move it into the box "Dependent.[125]"



Select *gender* and *educ* as your independent variables by moving them into the "Covariates" window.

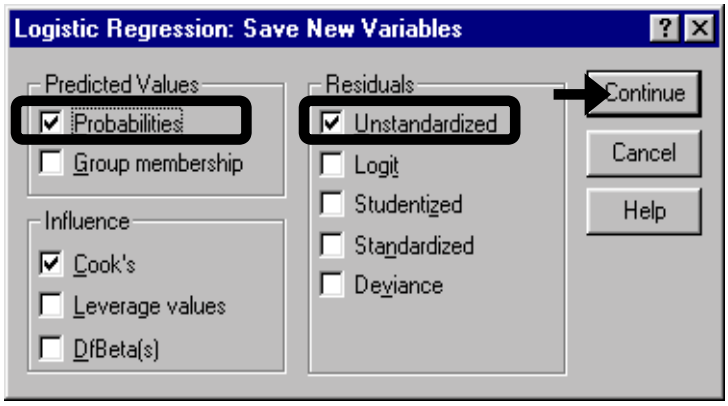
Click on the button "Save."



[125] The dependent variable can take only two values, 0 or 1. If you have any other values, then SPSS will generate an error. If that happens, go back to the data editor and remove the other values by using the appropriate procedure(s) from:

- DATA/DEFINE VARIABLE/MISSING (see section 1.2)
- TRANSFORM/RECODE/INTO SAME VARIABLES (see section 2.1)

Choose to save "Probabilities" and "Unstandardized residuals[126]."

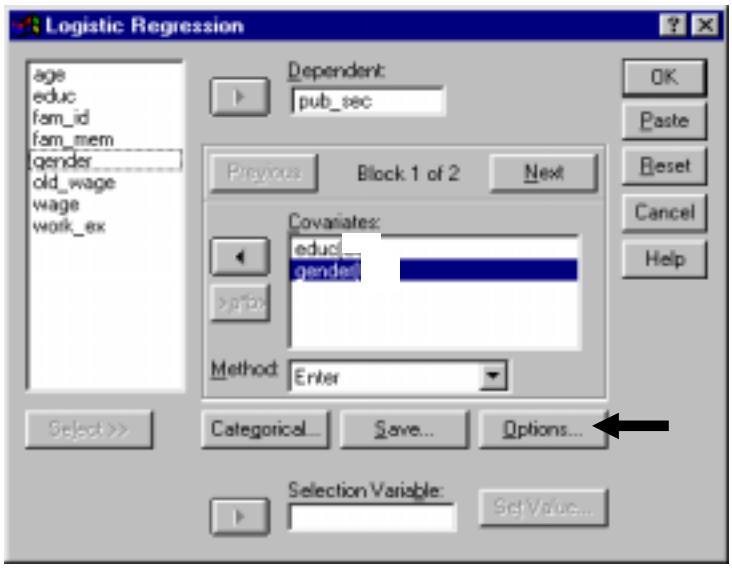


The probability variable contains "the predicted probability that the dependent variable equals one[127]."

Click on "Continue."

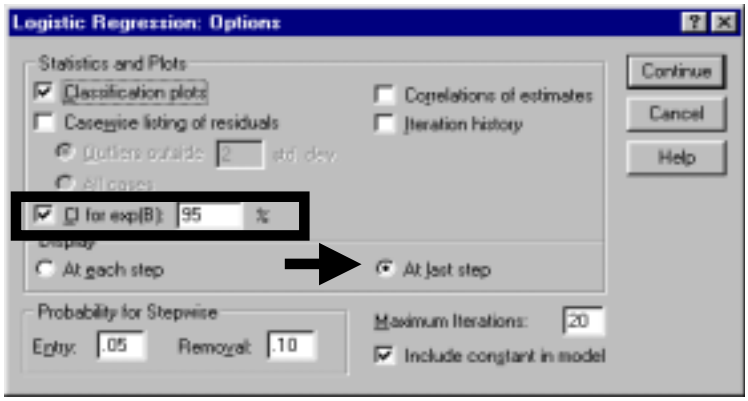
Note: "Influence" statistics are beyond the scope of this book.

You must customize the output you obtain. To do that, click on "Options."



It is useful to obtain a confidence interval for each of the coefficient estimates. To do this, click on the box to the left of the label "CI for exp(B)."

Choose "At last step" in the area "Display." By doing so, you are telling SPSS to give only the solution that emerges after the "final" iteration or run of the algorithm.

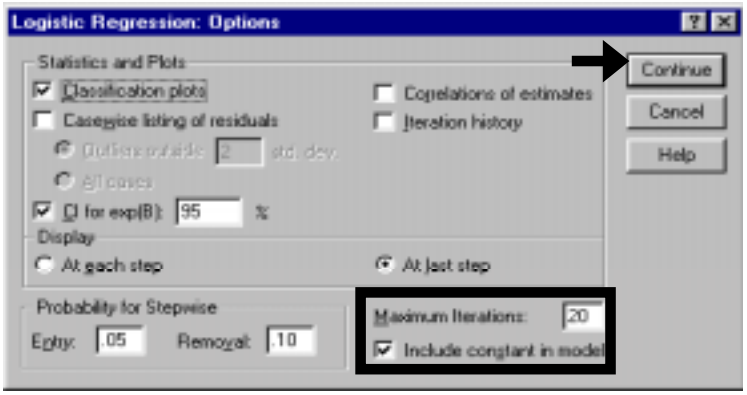


[126] The option "Logit" in residuals gives the residuals on a scale defined by the logistic distribution. Mathematically, the "Logit" values will be = (residual)/(p*(1-p)) where p is the (predicted) probability.

[127] In our example, the probability that the respondent works in the public *sector*

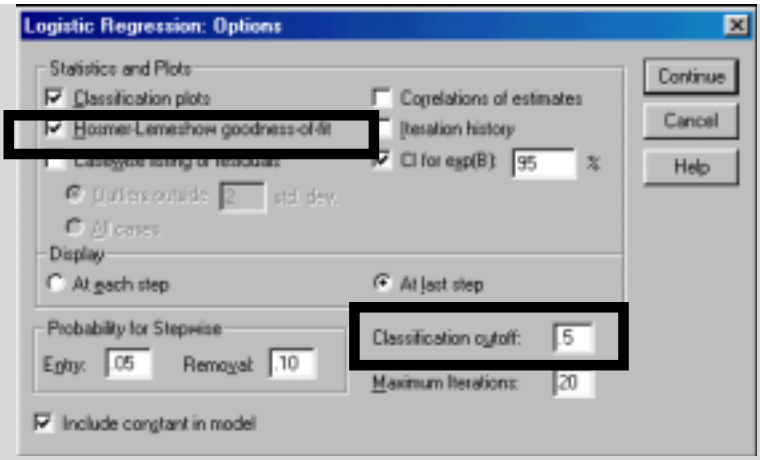
On the bottom of the dialog box, you can choose the probability level and number of iterations SPSS should use when it runs the model[128]. It is usually best to leave the probability level unchanged.

Click on "Continue."



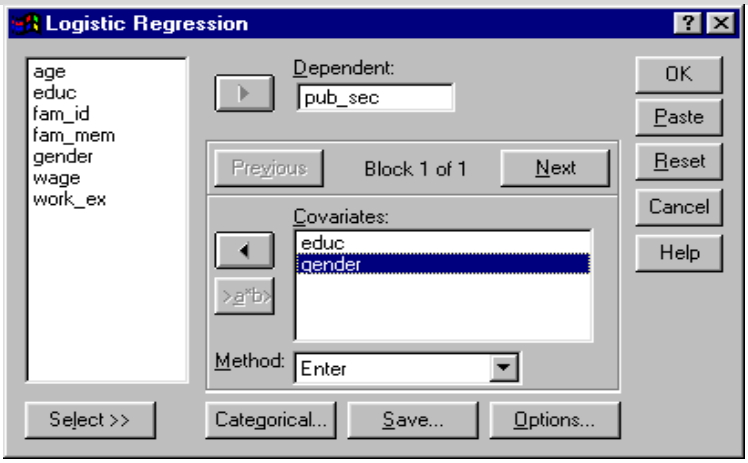
In newer versions of SPSS, some additional features will be shown in the "Options" dialog box. We suggest you choose the "H-L goodness-of-fit" test (it can be interpreted in the same way as a F test for linear regression).

You may want to change the value in the box "Classification cutoff." The use of the value .5 implies that: "If the value of the predicted dependent variable is less than (greater than) .5 then classify it as a prediction of 0 (1)." Unless you have solid rationale for changing the cutoff value, leave it at .5.



Click on "OK."

Note: The newer SPSS versions also allow you to restrict the analysis to a Sub-set of the data using the option "Selection Variable." We advise you to use the SELECT CASE procedure (see section 1.7) to restrict analysis as this feature allows for great flexibility in restricting the analysis.



The MLE algorithm can be crudely described as "Maximizing the log of the likelihood that what is observed in the data will occur." The endogenous or choice variables for these algorithms are the coefficient estimates. The likelihood function is the joint probability of the distribution of the data as captured by a logistic function (for the joint distribution). For those who are curious as to why the "log" is used in MLE, we offer this explanation: When you take

[128] If you run a Logit and the output informs you that the model did not converge or a solution could not be found, then come back to this dialog box and increase the number of iterations specified in the box "Maximum Iterations." MLE runs an algorithm repetitively until the improvement in the result is less than the "Convergence Criterion," which in this case equals .01, or until the maximum number of iterations (20) is reached.

the logs of a joint distribution (which, in essence, is the multiplication of the distribution of each observation), the algorithm is converted into an additive function. It is much simpler to work with a function with 20,000 additive components (if your sample size is 20,000) than to work with 20,000 multiplicative components.

Logistic Regression

This tells you that a solution was found. If not, then go back to the options box and increase the number of iterations.

```
Dependent Variable. PUB_SEC

-2 Log Likelihood   2632.5502  [after first iteration]

Estimation terminated at iteration number 3 because Log Likelihood
decreased by less than .01 percent.

 -2 Log Likelihood     2137.962  [after last iteration]
 Goodness of Fit       2058.148

                     chi-square    df Significance

 Model chi-square       494.588     2         .0000
 Improvement            494.588     2         .0000
```

If this is below .1, then the model was significant, equivalent to the "Sig -F" in a linear regression.

Classification Table for *PUB_SEC*

Observed	*Predicted* 0	1	*% Correct*
0	1112	180	**86.07%**
1	294	430	**59.39%**
		Overall	**76.49%**

This says that, if the model were to predict the Y-values as 0 or 1, the model would be correct 76.5% times, a high number

For 294 cases, the model predicted the value 0, but the actual value (observed value) was 1.

Variables in the Equation

Variable	B	S.E.	Wald[129]	df	Sig.	R	Exp(B)
EDUC	.202	.0103	388.21	1	.0000	.3830	1.2249
GENDER	-.208	.1366	2.31	1	.1278	.0110	.8121
Constant	-1.888	.0911	429.53	1	.0000		

Look at the "Sig." If it is below 0.1, then the variable is significant at the 90% level. In this example, *education* is significant, but *gender* is not.

Let's interpret the coefficient on *education*. Look in the column "Exp (B)." The value is 1.2249[130]. First subtract 1 from this: 1.2249-1 = .2249.

Then multiply the answer by 100:

100*(.2249) = 22.49 %.

This implies that for a 1 unit increase in *education* (i.e. - one more year of *education*), the odds[131] of joining the public *sector* increase by 22.49%.

The "odds" interpretation may be less intuitive than an interpretation in terms of probability. To do that, you will have to go back to column "B" and perform some complex calculations. Note that the slope is not constant, so a one unit change in an independent variable will have a different impact on the dependent variable, depending on the starting value of the independent variables.

Note: Consult your textbook and class notes for further information on interpretation.

To compare between models when Maximum Likelihood Estimation is used (as it is throughout this chapter), the relevant statistic is the "-2 Log Likelihood." In this example, the number is 2137.962. Consult your textbook for more details on the testing process.

Ch 9. Section 1 Non-linear regression

In linear regressions, each coefficient is a slope (or first derivative) and its interpretation is straightforward. In non-linear estimation, the interpretation of coefficients is more complex because the slopes are **not** constant (as was the case in Logit).

The main advantage of non-linear estimation is that it allows for the use of highly flexible functional forms.
We first use curve estimation to implement a simple 2 variable, 1 function, non-linear model (See section 9.2.a). In section 9.2.b., we describe a more flexible method that can be used to estimate complex multi-variable, multi-function models.

[129] The Wald is equivalent to the T-test in Linear regression.

[130] Remember that Logit is non-linear. As such, the impact of *education* on probability will depend upon the level of *education*. An increase in *education* from 11 to 12 years may have a different effect on the problem of joining the public sector than an increase in *education* from 17 to 18. Do not interpret it in the same manner as a linear regression.

[131] The odds of "yes" = Probability("yes")/Probability("no")

You may find that this entire section is beyond the scope of what you must know for your thesis/exam/project.

Ch 9. Section 1.a. Curve estimation

Curve estimation is a Sub-set of non-linear estimation. The latter (shown in section 9.2.b) is far more flexible and powerful. Still, we decided to first show an example of curve estimation so that you do not get thrown headlong into non-linear estimation.

We estimate the function[132] *wage* = B_0 + LOG B_1* (*work_ex*).

A zero value of the independent variable can disrupt our analysis. To avoid that, go to DATA/DEFINE VARIABLE and define zero values of the independent variable (*work_ex*) as missing values that are not to be included in the analysis (See section 1.2). You are now ready for curve estimation

Go to STATISTICS/REGRESSION/ CURVE ESTIMATION.

Note: The underlying estimation method used is Maximum Likelihood Estimation (MLE) and it remains the method used in all sections of this chapter.

Curve estimation is useful for quickly running several bivariate regressions. Typically, you use one independent variable and several dependent variables. For each dependent variable you specify, SPSS will run one curve estimation.

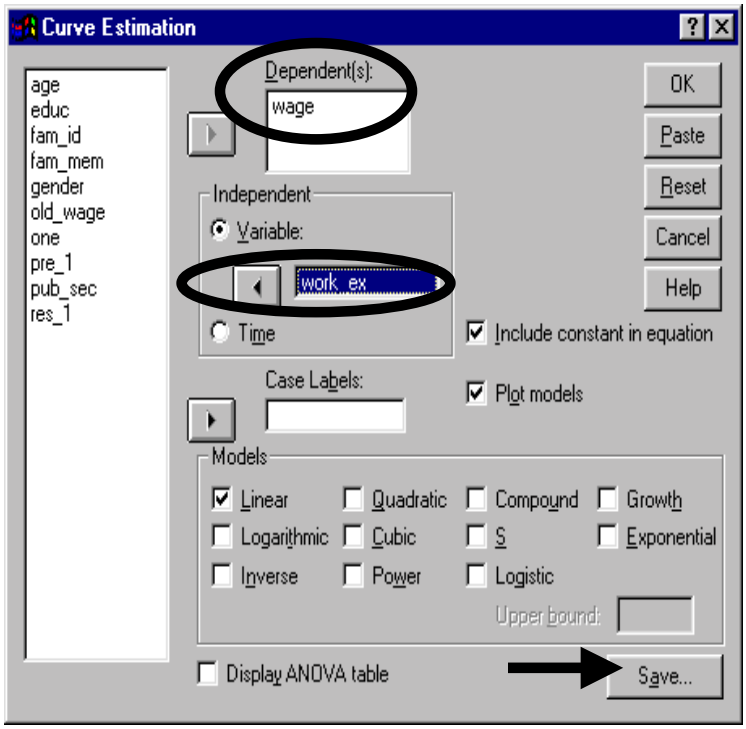
[132] Curve estimation can only estimate 2 variable, 1 function models.

Click on and place the dependent variable *wage* into the box "Dependent."

Click on and place the variable *work_ex* into the box "Independent."

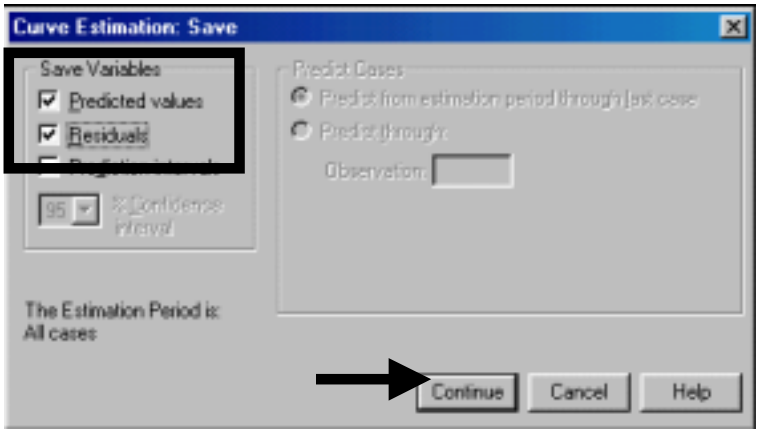
Select to include the constant (intercept).

Click on "Save."



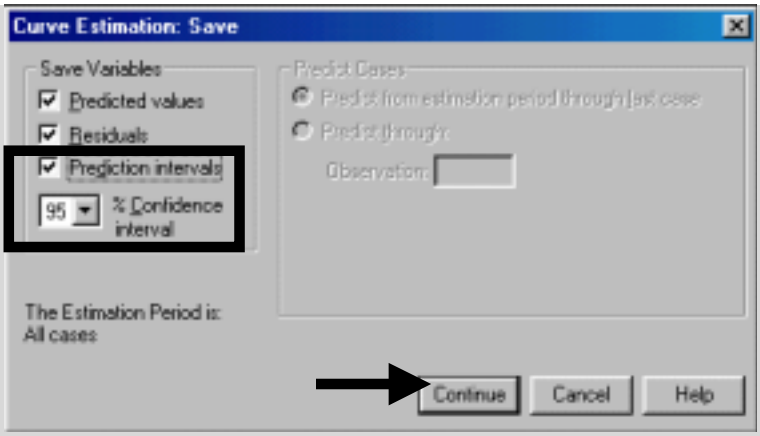
You should save the residuals and the predicted values (for the dependent variable). These will be added as new variables at the end of your data sheet. You may want to use them in tests, diagnostics, etc.

Click on "Continue."



A digression: You can also ask for the lower and upper bounds of the 95% confidence interval for the predicted dependent variable for each observation. SPSS will produce two variables - one for the upper bound of the confidence interval and one for the lower bound.

Click on "Continue."

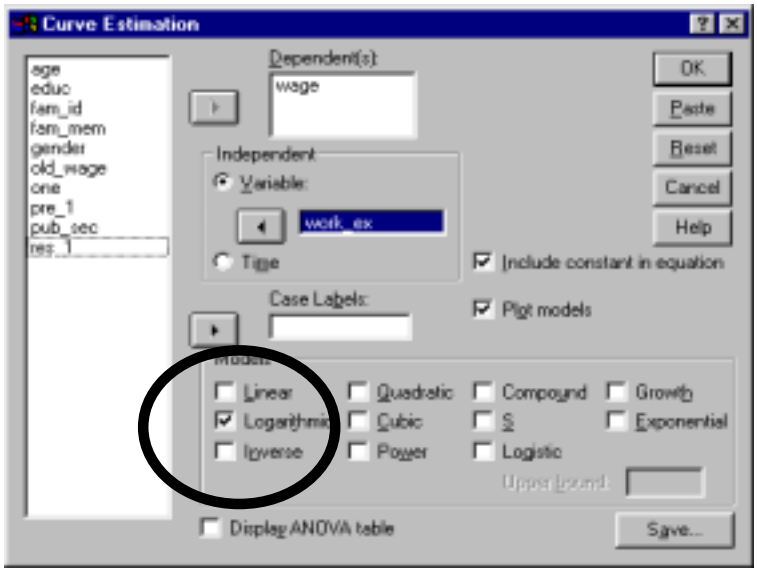


Select the model "Logarithmic" in the area "Models." Deselect everything else.

Note: SPSS provides several functions from which you can choose (other than Logarithmic). Your theory, the opinions of experts in the field, and mathematical modeling should suggest the appropriate function to use.

Click on OK.

Note: You can click on "Save" and choose to save the predicted values as a new variable.



As presented in the shaded text below, the model is significant at 95% (as Sig. F < .05), the intercept equals 5.41 (bo = 5.41), and the coefficient is 1.52. The R-square is a psuedo-rsquare, but can be roughly interpreted as the R-square in Linear Regression. Note that the coefficient estimate is not the slope. The slope at any point is dependent on the coefficient and the value of the independent variable at the point at which the slope is being measured.

```
Independent:  WORK_EX

Dependent Mth(function) R-square    F       Sig. f        b0        b1
WAGE        LOG          .066       139.1   .000          5.41      1.52
```

Ch 9. Section 1.b. General non-linear estimation (and constrained estimation)

This is a more flexible method than curve estimation. It places almost **no limitation** on the structure of the function. In the same model, you can use logs, exponentials, cubic powers, or combinations thereof.

In addition, you can place constraints on the possible values coefficients can take. For example, $\beta_3 \geq 1.5$

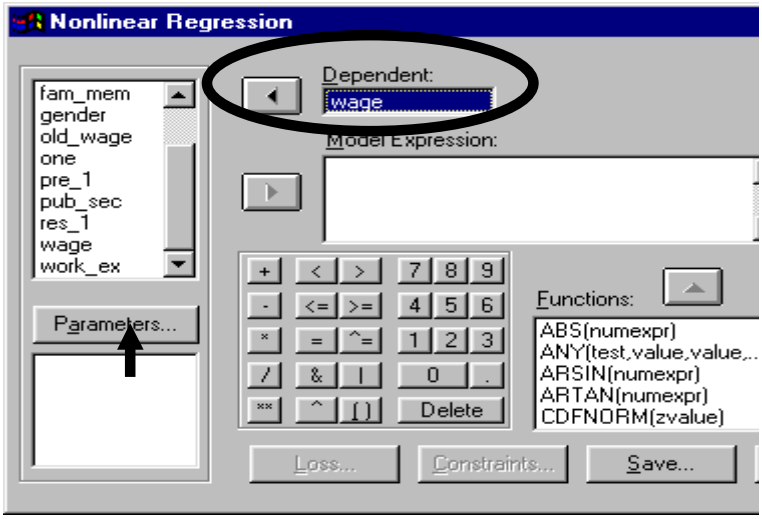
In Linear regression, the "sum of squared residuals" was minimized. The Non Linear Estimation procedure allows you to minimize other functions (e.g. - deviations in predicted values) also.

We estimate the model:

$$wage = B_1 * educ + B_2 * gender + B_3 * pub\_sec + \text{LN } B_4 * work\_ex$$

Go to STATISTICS/REGRESSION/ NON-LINEAR. Place the variable *wage* into the box "Dependent."

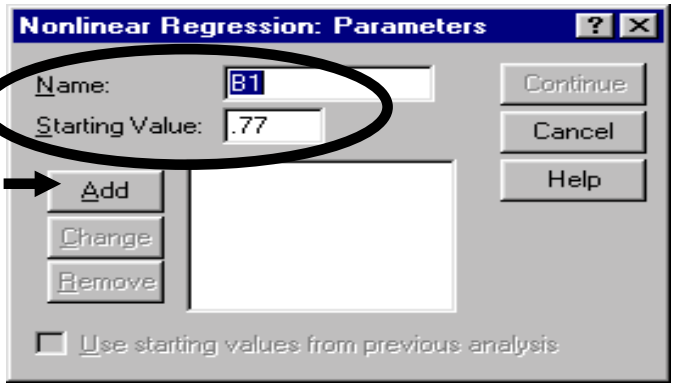
Click on the button "Parameters."



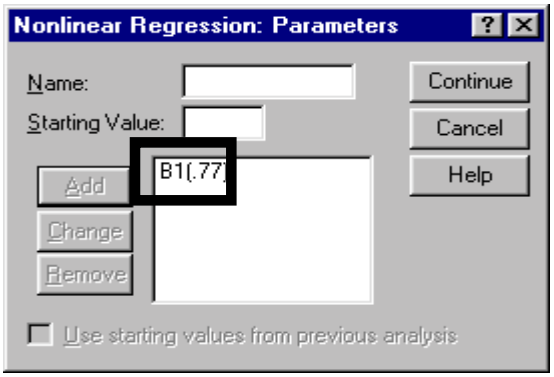
We define the parameters for the model in this dialog box. For initial values, we are using the coefficients obtained in the initial mis-specified linear regression shown in section 7.2. as a first approximation.

In the box "Name," place a name for the parameter. This corresponds to the coefficient for the variable *education*. Enter the initial value into the box "Starting Value."

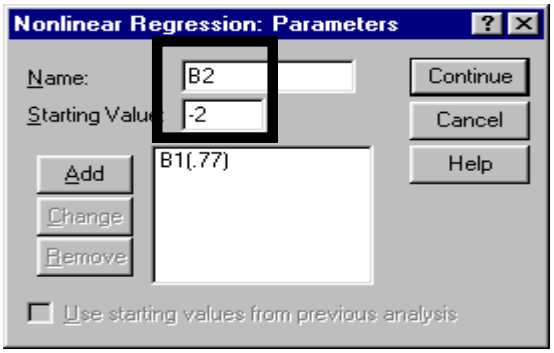
Click on the button "Add."

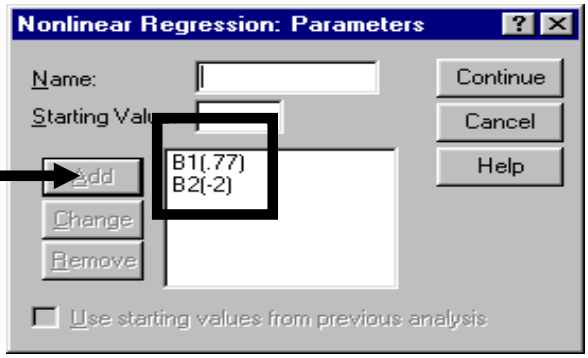


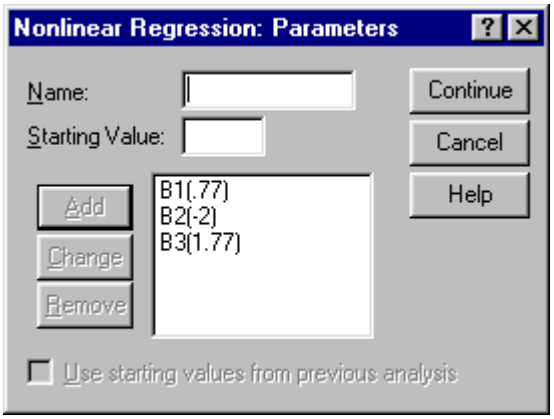
The starting values for parameter B1 are registered in the box.

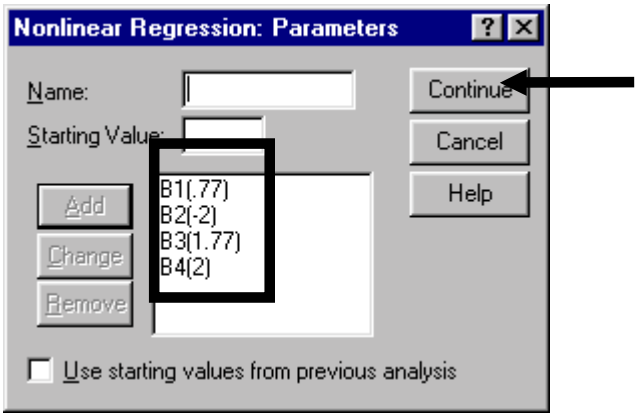


Do the same for the second to fourth parameters/coefficients.



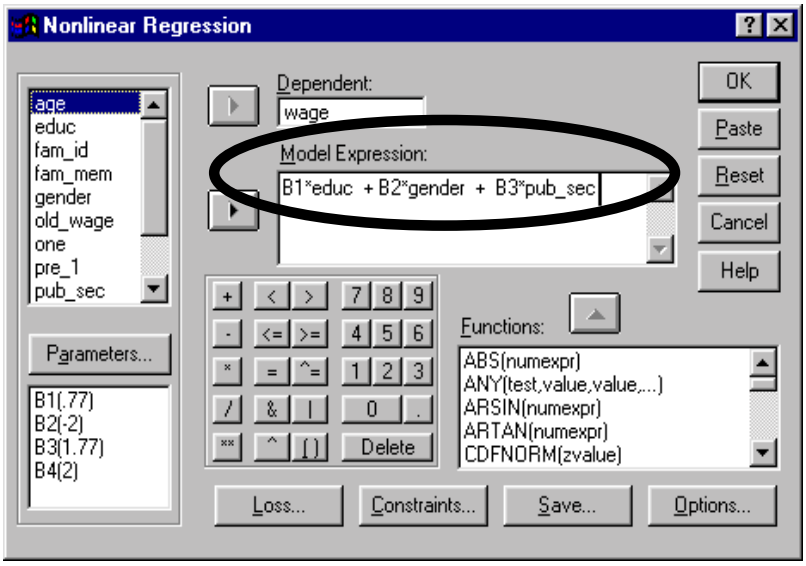




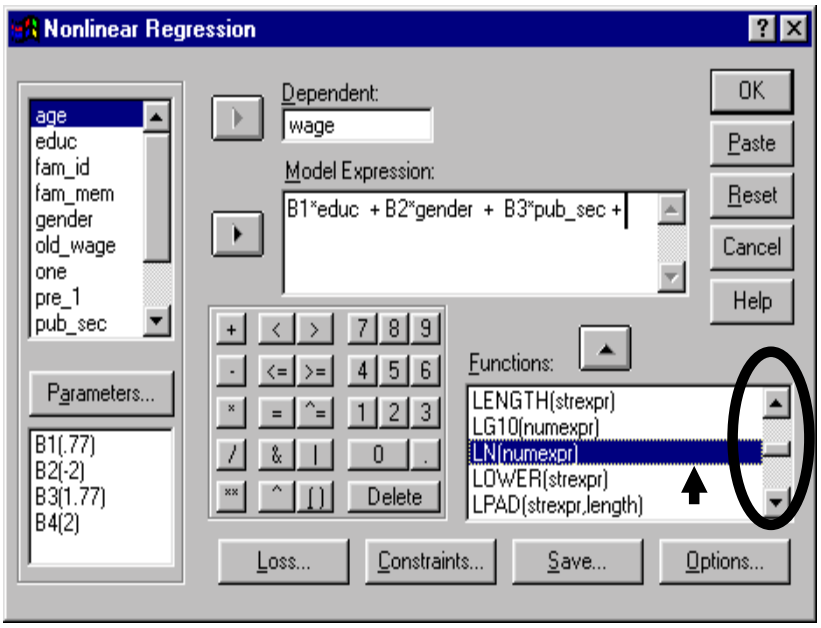


Click on "Continue."

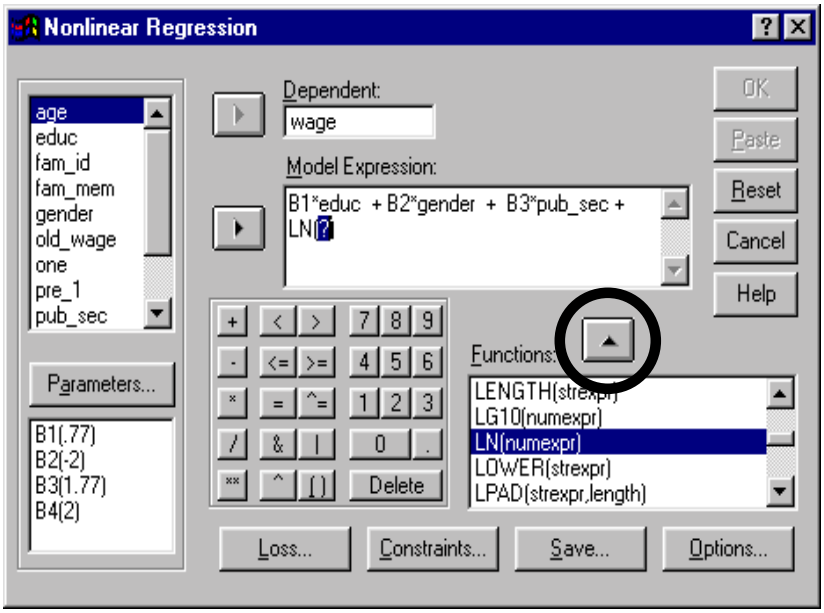
Place the cursor inside the box/area "Model Expression" and click. Then type in the equation using the parameter names you defined in the box "Parameters."



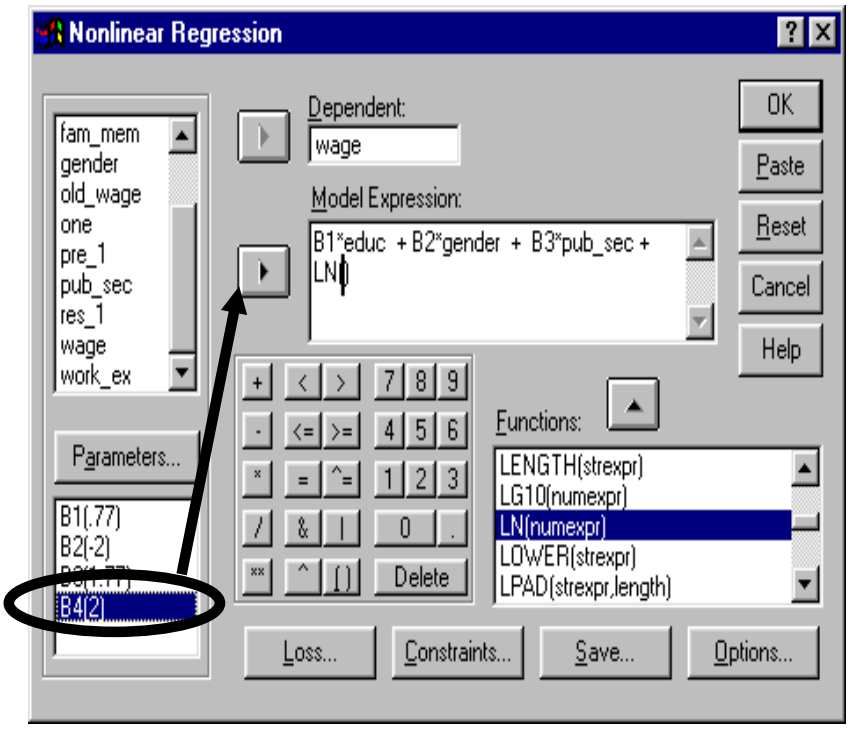
We need to add a functional form for the parameter associated with work experience (B4). To do so, click on the scroll bar in the box "Functions" and find the function "LN."



Click on the upward arrow to place the LN function into the equation.

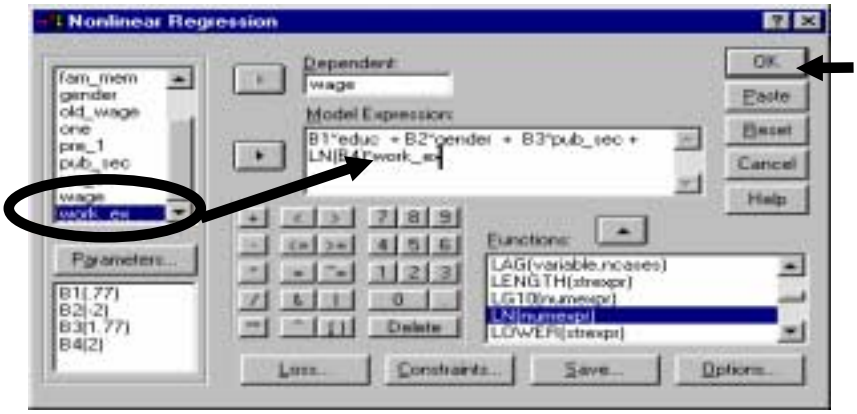


Click on B4 (2) in the area "Parameters." Then click on the arrow to move B4 (2) into the expression box.



Place the variable *work_ex* into the equation.

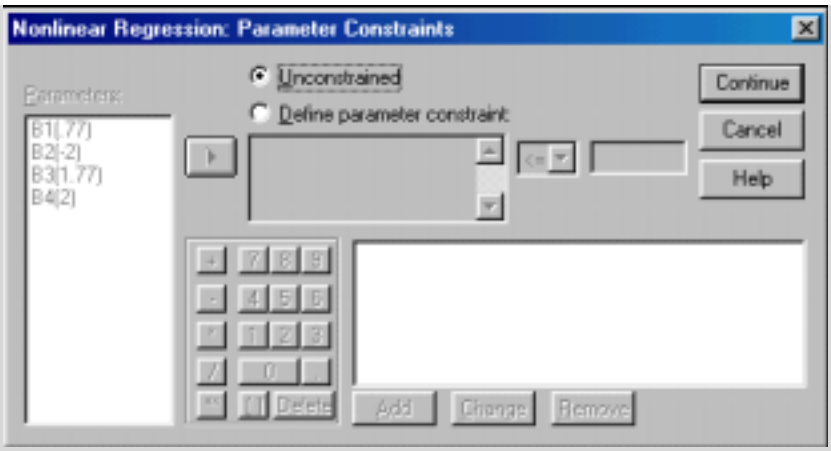
Click on OK.



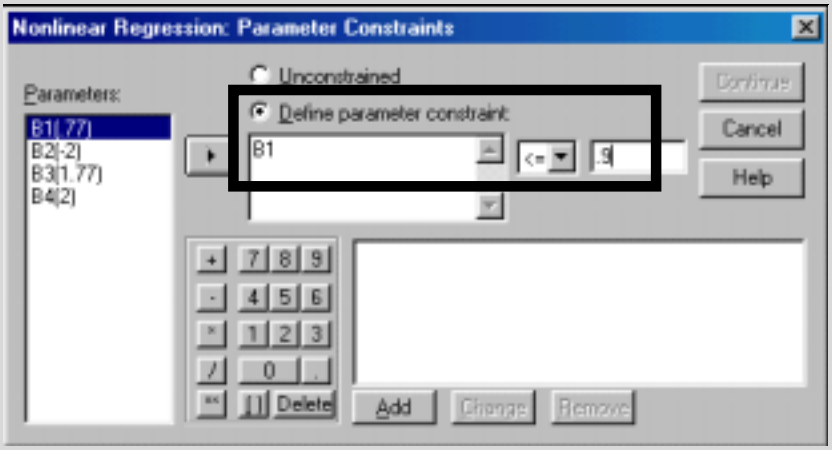
Addendum

You can ask SPSS to do a "constrained optimization."

Click on the button "Constraints" in the main dialog box. The following dialog box will open.

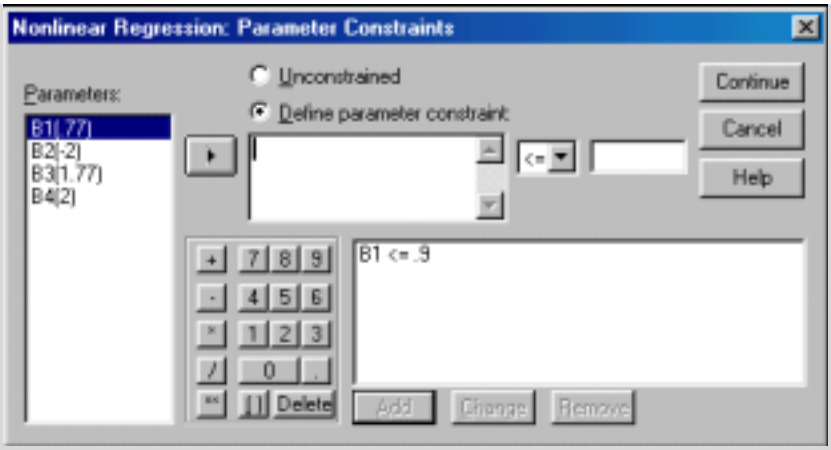


Click on "Define parameter constraint" and type in a constraint as shown.



Click on the "Add" button.

You can have several constraints.



The model had a good fit as shown in the output reproduced in the shaded text below.

```
Run stopped after 9 model evaluations and 5 derivative evaluations.
Iterations have been stopped because the magnitude of the largest
correlation between the residuals and any derivative column is at
most RCON = 1.000E-08
```

Implies that a solution was found. If it was not found, then go back and increase the number of iterations by pressing the button "Options."

```
Non-linear Regression Summary Statistics     Dependent Variable WAGE

  Source                 DF   Sum of Squares     Mean Square
  Regression              4         186720           46680
  Residual             1964          54677           27.84
  Uncorrected Total    1968         241398

  (Corrected Total)    1967   106029

R squared = 1 - Residual SS / Corrected SS =     .48

                          Asymptotic      Asymptotic 95 %
                                          Confidence Interval
  Parameter   Estimate    Std. Error      Lower          Upper
  B1           .8125          .021          .770          .8550
  B2         -1.564           .288        -2.129         -.999
  B3          2.252           .292         1.677         2.826
  B4          1.268           .013         1.242         1.294
```

Note: No T-Statistic is produced. The confidence intervals give the range within which we can say (with 95% confidence) that the coefficient lies. To obtain a "rough" T-estimate, divide the "Asymptotic Std. Error" by the estimate. If the absolute value of the result is greater than 1.96 (1.64), then the coefficient is significant at the 95% (90%) level.

The estimated model:

wage = 0.81**educ* – 1.56**gender* + 2.25**pub_sec* + LN 1.26**work_ex*

The coefficients on the first three variables can be interpreted as they would be in a linear regression. The coefficient on the last variable cannot be interpreted as a slope. The slope depends on the value of the estimated coefficient and the value of the variable at that point.

Note: The R-square is a psuedo-square, but can be roughly interpreted in the same way as in Linear regression.

To take quizzes on topics within each chapter, go to http://www.spss.org/wwwroot/spssquiz.asp

Ch 10. COMPARATIVE ANALYSIS

"Comparative analysis" is used to compare the results of any and all statistical and graphical analysis across sub-groups of the data defined by the categories of dummy or categorical variable(s). Comparisons always provide insight for social science statistical analysis because many of the variables that define the units of interest to the analyst (e.g. - gender bias, racial discrimination, location, and milieu, etc.) are represented by categorical or dummy variables.

For example, you can compare the regression coefficients on *age*, *education*, and *sector* of a regression of wages for male and female respondents. You can also compare descriptives like mean *wage*, median *wage*, etc. for the same subgroups.

This is an extremely powerful procedure for extending all your previous analysis (see chapters 3-9). You can see it as an additional stage in your project - procedures to be completed after completing the desired statistical and econometric analysis (discussed in chapters 3-9), but before writing the final report on the results of the analysis[133].

In section 10.1, we first show the use of one categorical variable (*gender*) as the criterion for comparative analysis. We then show how to conduct comparisons across groups formed by the interaction of three variables.

[133] The note below may help or may confuse you, depending on your level of understanding of statistics, the process of estimation, and this book. Nevertheless, we feel that the note may be of real use for some of you, so read on.

After you have conducted the procedures in chapters 3-9, you should interpret and link the different results. Look at the regression results. What interesting or surprising result have they shown? Does that give rise to new questions? Link these results to those obtained from chapters 3-6. Is a convincing, comprehensive, and logically consistent story emerging? To test this, see if you can verbally, without using numbers, describe the results from the initial statistical analysis (chapters 3-6), link them to the expected regression results (including the breakdown of classical assumptions), and then trace them back to the initial statistics.

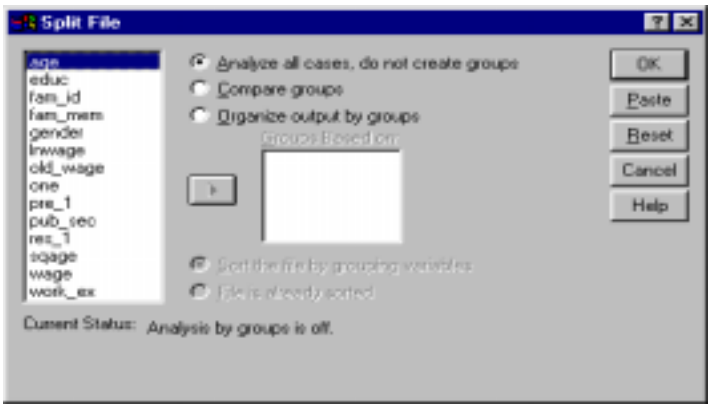
After the econometric procedures, you can do several things:

- Manipulate the data so that you can filter in (using SELECT CASE as shown in section 1.7) only those cases that are required for the analysis inspired by the above insights and queries.
- Create variables from continuous variables. Then use these variables in a re-specified regression model, and in other graphical and statistical procedures.
- Make more detailed custom tables (chapter 6) and graphs (chapters 3-5) to better understand the data.

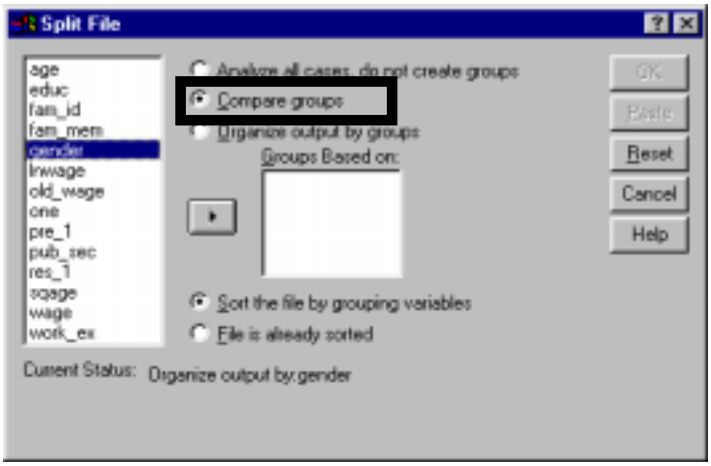
Ch 10. Section 1 Using Split File to compare results

We plan to compare the descriptive statistics (means, medians, standard deviations, etc.) and run a regression analysis for the variables *education* and *wage* across the categories of the variable *gender*.

Go to DATA/SPLIT FILE.



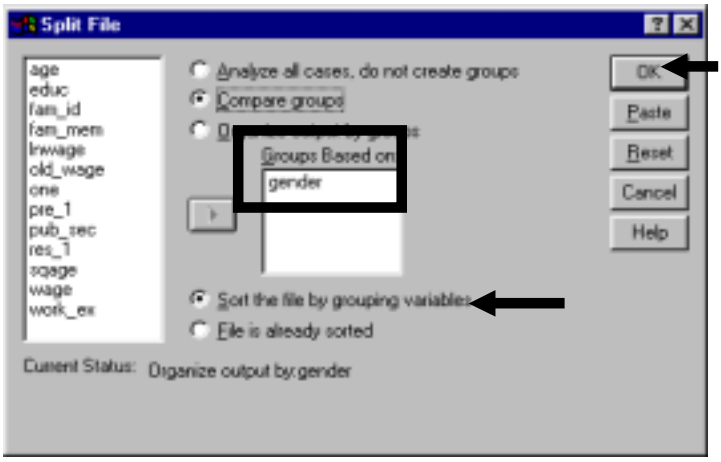
Click on the button to the left of the label "Compare Groups."



In this chapter, we never use the third option, "Organize output by groups." You should experiment with it. It is similar to the second option with one notable difference - the output produced is arranged differently.

Move the variable *gender* into the box "Groups Based on." Each "group" will be defined by the values of the variable chosen. The two groups will be "Male" and "Female."

Click on the button to the left of "Sort the file by grouping variables."



If you do not choose this option, and the data are not pre-sorted by *gender* (see chapter 1 for learning how to sort), then too many groups will be confused. For example, SPSS will start with Males. Each time it finds a different entry ("Females"), it will create a new

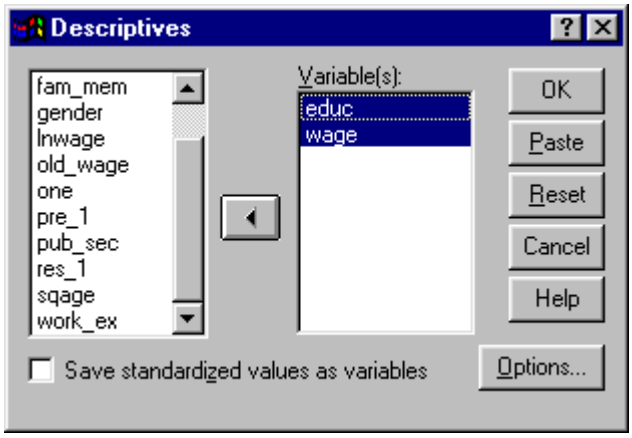
group. If it finds the entry ("Male") again, it will create a third group. Sorting avoids such problems.

Click on "OK."

Now, any procedure you conduct will be split in two - one for males and one for females.

For example, assume you want to compare descriptive statistics for males and females.

Go to STATISTICS/ DESCRIPTIVES and fill in the dialog box as shown.



Note: the variable *gender* cannot be a variable used in the analysis because we are doing the comparison by *gender* categories.

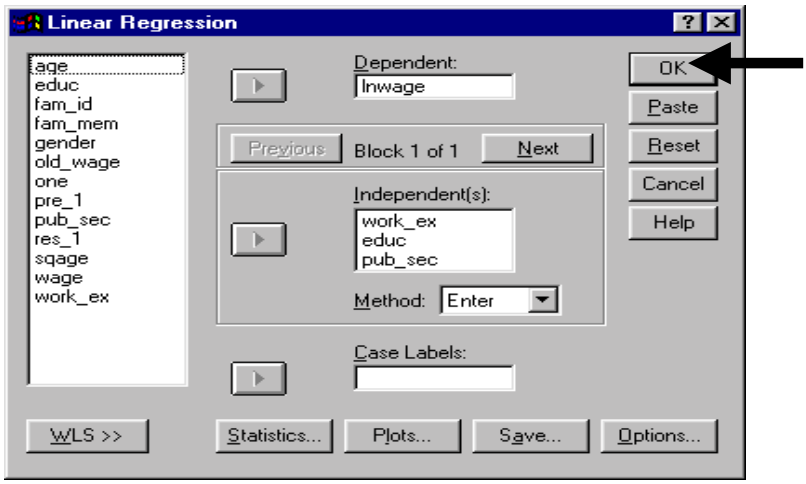
The statistics are shown separately for males and females. From this table you can compare males and females easily. "The mean *wage* of males is 0.2 higher than that for females, even though the mean *education* of males is lower by 0.45 years."

Descriptive Statistics

GENDER		N	Minimum	Maximum	Mean	Std. Deviation
Male	EDUCATION	1613	0	23	6.00	5.48
	WAGE	1594	.00	49.43	8.6346	7.5878
	Valid N (listwise)	1594				
Female	EDUCATION	403	0	21	6.45	6.05
	WAGE	399	.23	29.36	6.6161	5.8868
	Valid N (listwise)	399				

You can also compare regression results.

Go to STATISTICS/ REGRESSION/ LINEAR and fill in the dialog box as shown. (Refer to chapters 7 and 8 for more on regression analysis).



Note: *gender* cannot be a variable used in the analysis because we are doing the comparison by *gender* categories.

Click on "OK."

The fit of the model is better for females - the adjusted R-square is higher for them. We checked that the F was significant for all models but we did not reproduce the table here in order to save space.

Model Summary[a]

GENDER	Model	Variables Entered	R Square	Adjusted R Square	Std. Error of the Estimate
Male	1	*Whether Public Sector Employee, Work Experience, EDUCATION*	*.394*	*.392*	*.6464*
Female	1	*Whether Public Sector Employee, Work Experience, EDUCATION*	*.514*	*.511*	*.7037*

a. Dependent Variable: LNWAGE

Note: In chapters 7 and 8 we stressed the importance of mis-specification and omitted variable bias. Is the fact that we are splitting the data (and thus the regression) into 2 categories leaving us vulnerable to omitted variable bias? Not really.

Reason: Now the population being sampled is "only males" or "only females," so gender is not a valid explanatory agent in any regression using these samples as the parent population and thereby the valid model has changed. Of course, some statisticians may frown at such a simplistic rationale. Usually such comparative analysis is acceptable and often valued in the workplace.

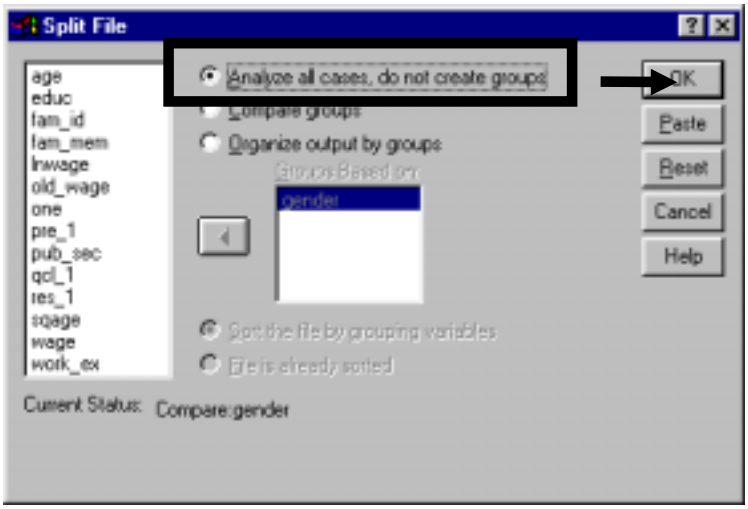
All the coefficients are significant for all the models. The rate of return to an additional year of schooling is higher for females by 4% (compare the coefficient values of .07 [7%] and .11 [11%] in column B) while the impact of *work experience* is the same across *gender*.

Coefficients[a]

GENDER	Model		Unstandardized Coefficients B	Std. Error	t	Sig.
Male	1	(Constant)	*1.07*	*.03*	*33.54*	*.00*
		Work Experience	*.02*	*.00*	*11.51*	*.00*
		EDUCATION	*.07*	*.00*	*19.44*	*.00*
		Whether Publi Sector Employee	*.39*	*.04*	*9.86*	*.00*
Female	1	(Constant)	*.53*	*.06*	*8.22*	*.00*
		Work Experience	*.02*	*.00*	*4.85*	*.00*
		EDUCATION	*.11*	*.01*	*14.76*	*.00*
		Whether Publi Sector	*.17*	*.09*	*1.86*	*.06*

Once the comparative analysis is completed and you do not want to do any more in the current session, go to DATA/ SPLIT FILE and remove the split by clicking on the button to the left of "Analyze all cases."

Click on "OK."



Ch 10. Section 1.a. Example of a detailed comparative analysis

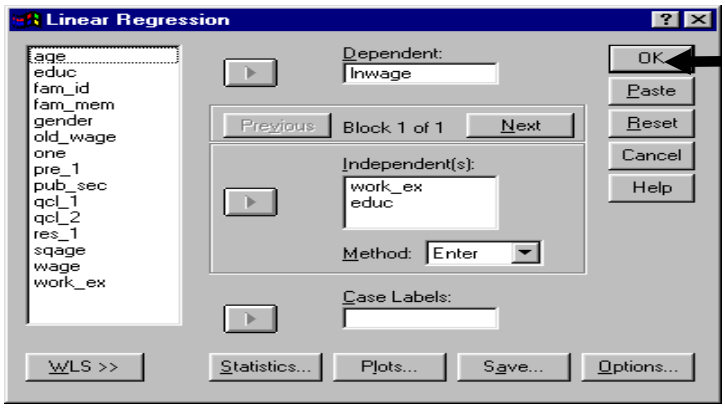
In the previous two examples, we used only one variable as the basis for comparative analysis. You can use more than one variable and this will often provide a greater insight into the differences in attributes of subgroups of the data.

Go to DATA/ SPLIT FILE. Select "Compare Groups" and move three variables into the box "Groups Based on." Use *sector*, income group (*qcl_2*), and g*ender*.

Click on "OK."

Run a regression as shown.

The output has twelve groups. We can compare groups like "Public sector high-income females" and "Public sector low-income males." The next table shows these comparative results for the coefficient estimates and the T-statistics (significance tests).

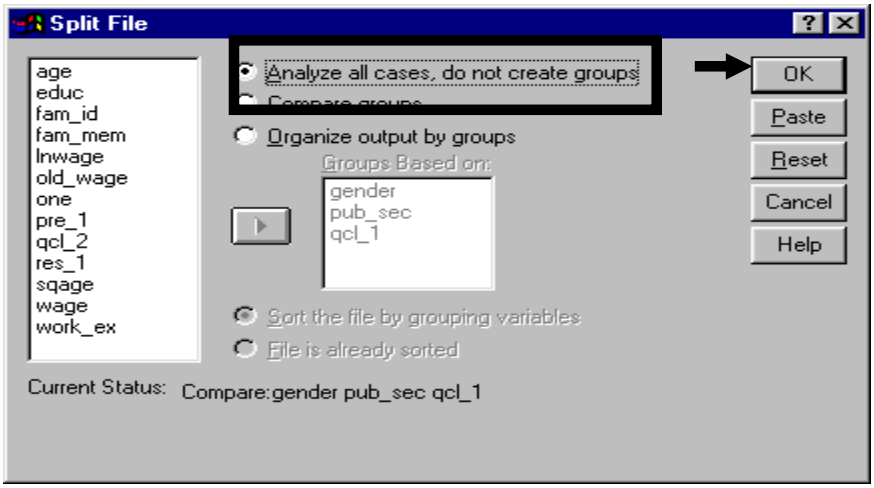


Coefficients[b]

GENDER	Whether Public Sector Employee	QCL_1			Unstandardized Coefficients		t	Sig.
					B	Std. Error		
Male	Private Sector	High Income		(Constant)	3.53	.20	17.43	.00
				Work_Ex	.00	.01	-.34	.74
				EDUCATION	.01	.01	.65	.53
		Low Income		(Constant)	1.33	.04	33.58	.00
				Work_Ex	.01	.00	2.43	.02
				EDUCATION	.01	.01	1.52	.13
		Mid Income		(Constant)	2.45	.07	33.87	.00
				Work_Ex	.01	.00	2.21	.03
				EDUCATION	.02	.01	3.37	.00
	Public Sector	High Income		(Constant)	3.16	.16	20.17	.00
				Work_Ex	.00	.00	.92	.36
				EDUCATION	.02	.01	2.75	.01
		Low Income		(Constant)	1.33	.10	13.27	.00
				Work_Ex	.01	.01	2.57	.01
				EDUCATION	.04	.01	4.03	.00
		Mid Income		(Constant)	2.42	.05	46.93	.00
				Work_Ex	4.915E-03	.002	3.133	.002
				EDUCATION	1.944E-02	.003	6.433	.000
Female	Private Sector	High Income		(Constant)	2.050	.068	30.301	.000
				Work_Ex	2.816E-03	.004	.629	.537
				EDUCATION	2.587E-02	.005	4.786	.000
		Low Income		(Constant)	2.931	.459	6.392	.000
				Work_Ex	1.007E-02	.009	1.073	.319
				EDUCATION	8.557E-03	.026	.326	.754
		Mid Income		(Constant)	.704	.080	8.820	.000
				Work_Ex	.01	.01	1.29	.20
				EDUCATION	.05	.01	3.54	.00
	Public Sector	High Income		(Constant)	1.86	.12	15.10	.00
				Work_Ex	.01	.00	3.09	.00
				EDUCATION	.04	.01	4.75	.00
		Low Income		(Constant)	1.88	.50	3.74	.00
				Work_Ex	.02	.01	2.19	.05
				EDUCATION	.06	.02	2.30	.04
		Mid Income		(Constant)	.45	.28	1.57	.13
				Work_Ex	.03	.03	1.18	.24
				EDUCATION	.05	.03	1.90	.07

b. Dependent Variable: LNWAGE

To remove the comparative analysis, go to DATA/SPLIT FILE and choose the option "Analyze all cases." Click on the button "OK."



Note: Your boss/professor may never realize the ease and power of using SPLIT FILE. Here is your chance to really impress them with "your" efficiency and speed.

To take quizzes on topics within each chapter, go to http://www.spss.org/wwwroot/spssquiz.asp

Ch 11. FORMATTING AND EDITING OUTPUT

The professional world demands well-presented tables and crisp, accurate charts. Apart from aesthetic appeal, good formatting has one more important function - it ensures that the output shows the relevant results clearly, with no superfluous text, data, or markers.

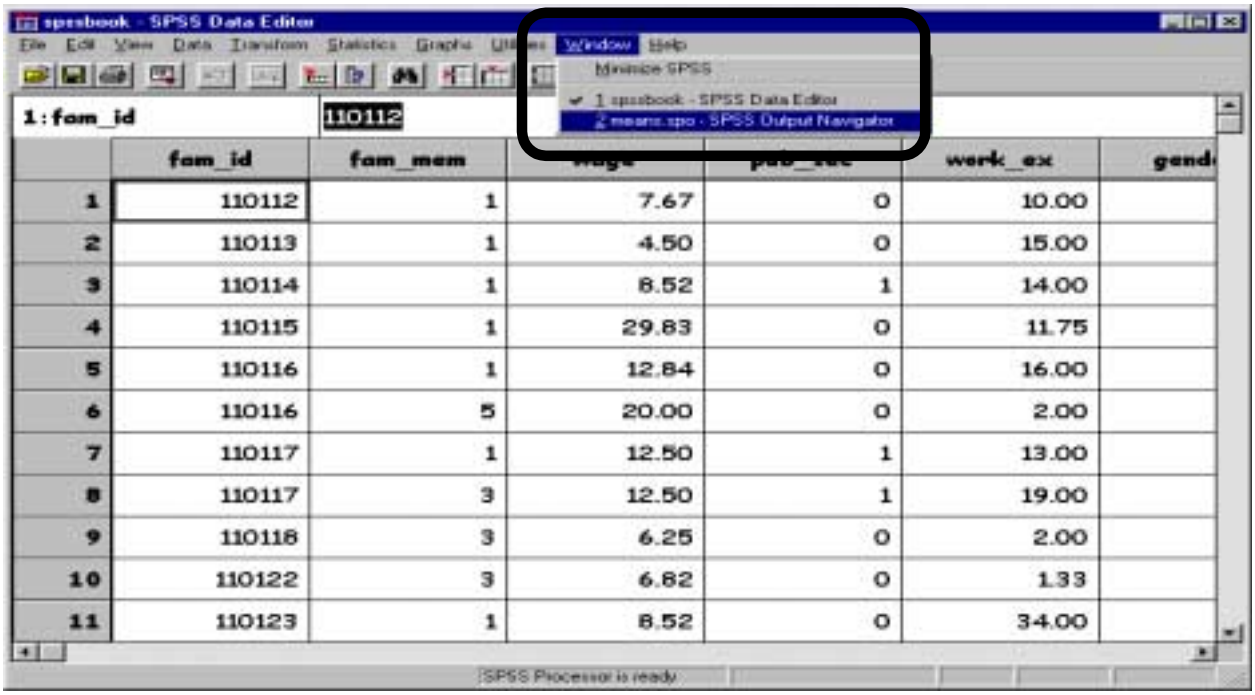
Section 11.1 shows how to format output tables.

Section 11.2 shows how to format charts.

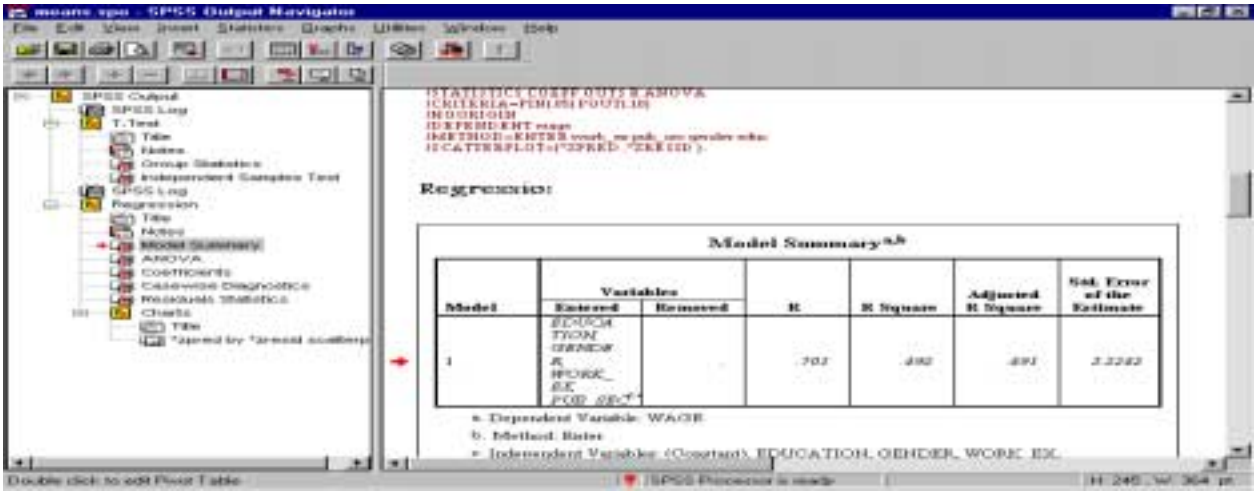
Ch 11. Section 1 Formatting and editing tables

Ch 11. Section 1.a. Accessing the window for formatting / editing tables

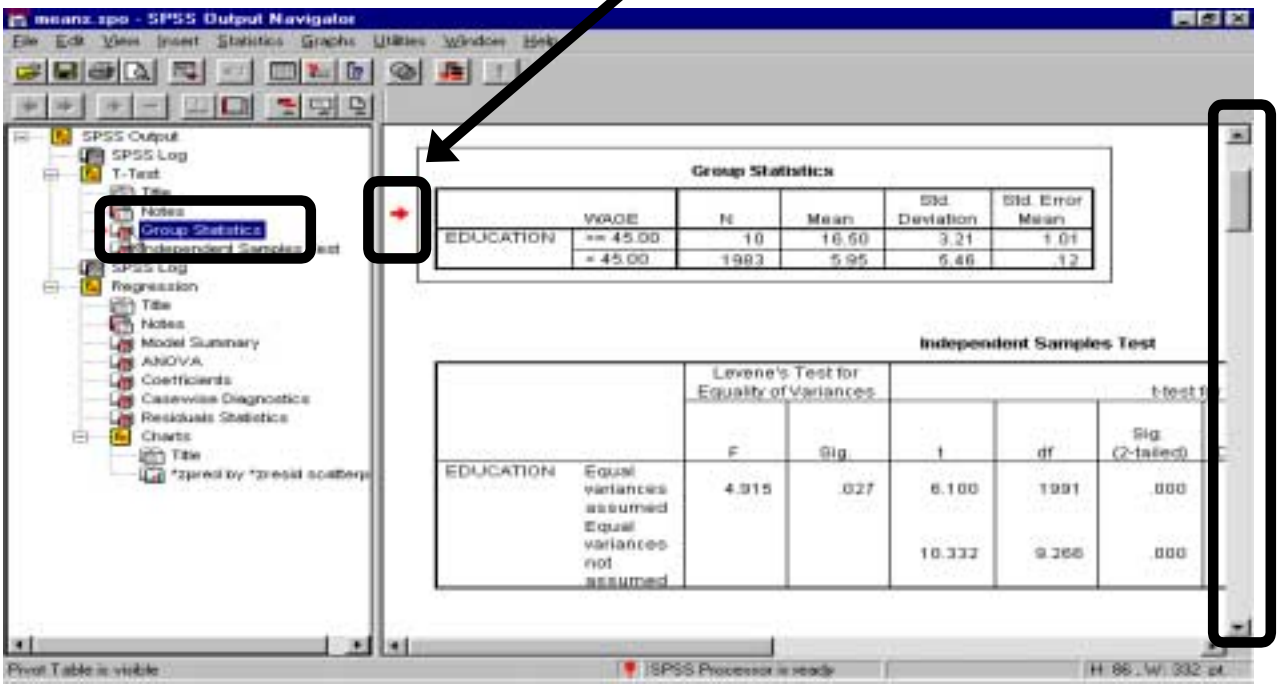
To format tables, you must first enter the formatting/editing window. To do so, go to WINDOWS/SPSS OUTPUT NAVIGATOR.



The navigator window shows all the output tables and charts (see next picture).

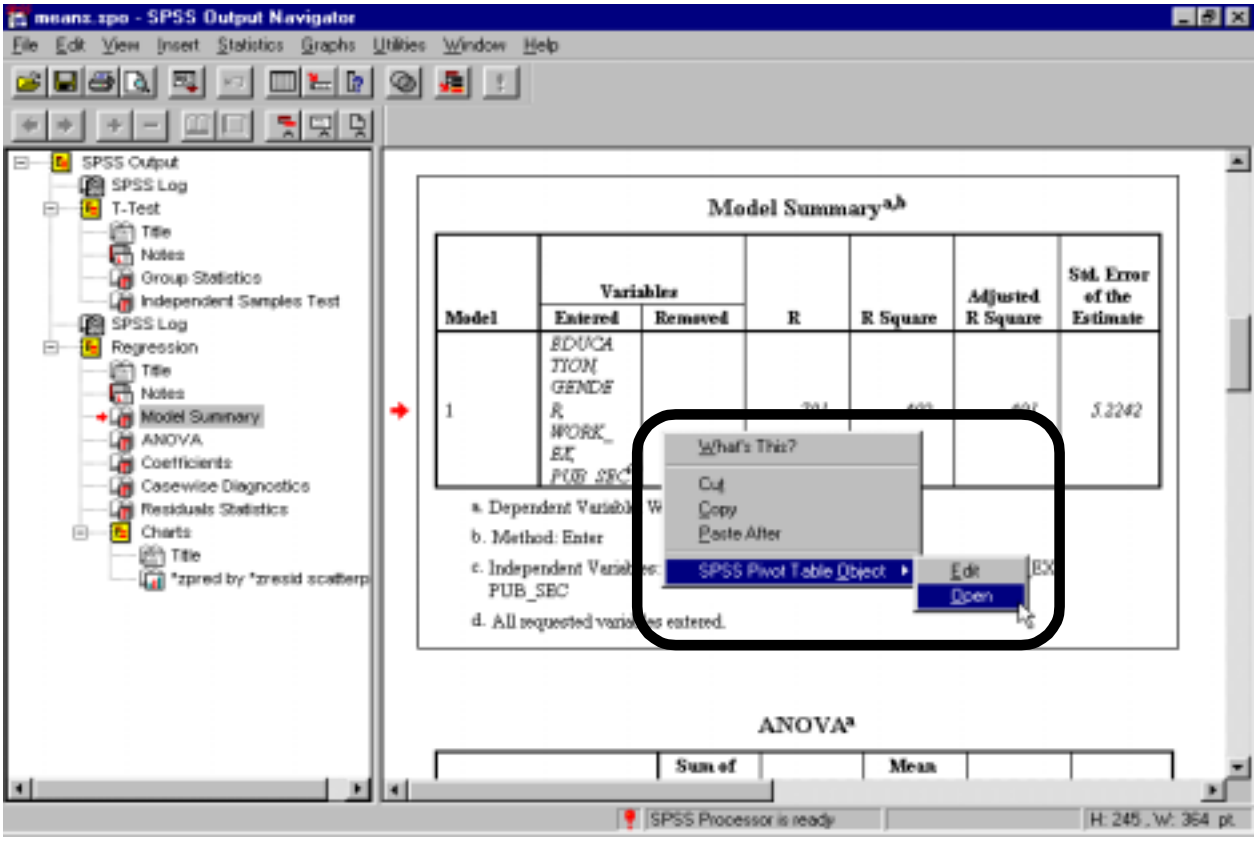


Using the scroll bar on the extreme right, scroll to the table you want to format or click on the table's name in the left half of the window. If you see a red arrow next to the table, then you have been successful in choosing the table.



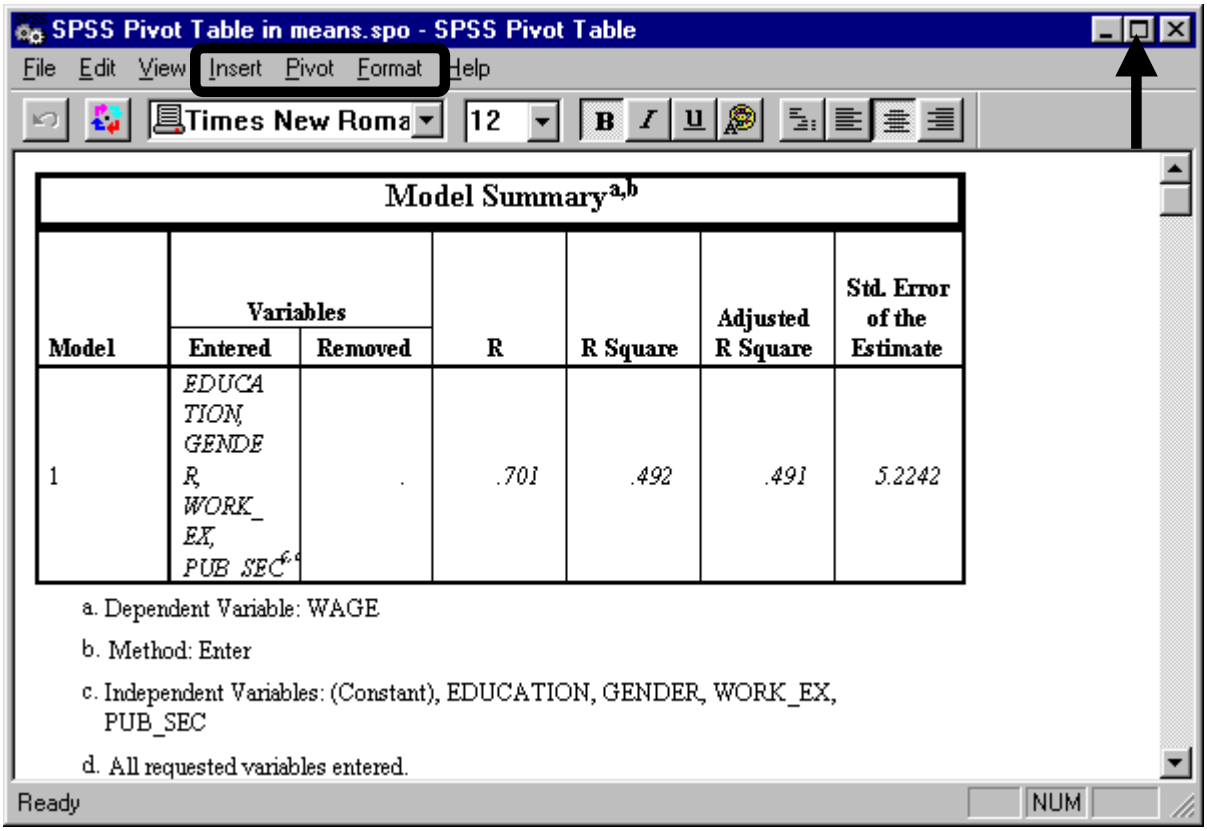
Click on the right mouse button.

Several options will open. Select the last one - "SPSS Pivot Table Object." Within that, choose the option "open" (see picture on next page).



A new window will open (see picture below). This window has only one item - the table you chose. Now you can edit/format this table.

Click on the "maximize" button to fill the screen with this window. See the arrowhead at the top right of the next picture to locate the "maximize" button.



This window has three menus you did not see in the earlier chapters: INSERT, PIVOT, and FORMAT. These menus are used for formatting tables.

1. INSERT is used to insert footnotes, text, etc.
2. PIVOT is used to change the layout of the table.
3. FORMAT has several features to format the table and individual cells within the table. The formatting features include shading, font type, font size and attributes, and borders.

Ch 11. Section 1.b. Changing the width of columns

Often, the width of a column is not large enough to display the text/data in cells in the column. In the table below, the column "Entered" is too small - variable names like *education* do not fit into one line.

Model Summary

Model	Variables Entered	Variables Removed	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	*EDUCATION, GENDER, WORK_EX, PUB_SEC*		.701	.492	.491	5.2242

This can be easily corrected. Using the mouse, go to the column dividing line. Now you can manually change the width of the column by dragging with the left mouse button. The next table shows the effect of doing this - the column "Entered" has been widened.

Model Summary

Model	Variables Entered	Variables Removed	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	*EDUCATION, GENDER, WORK_EX, PUB_SEC*	.	*.701*	*.492*	*.491*	*5.2242*

Ch 11. Section 1.c. Deleting columns

You may want to delete certain columns if they are not necessary for your particular project. For example, the column "R" may not be needed in a table with output from a regression.

Original table (with the column "R")

Model Summary

Model	Variables Entered	Variables Removed	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	*EDUCATION, GENDER, WORK_EX, PUB_SEC*	.	*.701*	*.492*	*.491*	*5.2242*

Inside the Table/Chart Editing Window, click on the cell “R." Then choose EDIT/SELECT. Select DATA CELLS and LABEL. Press the keyboard key “delete.”

This will yield the table on the right. Note that the column "R" has been deleted.

Model Summary

Model	Variables Entered	Variables Removed	R Square	Adjusted R Square	Std. Error of the Estimate
1	*EDUCATION, GENDER, WORK_EX, PUB_SEC*	.	*.492*	*.491*	*5.2242*

Ch 11. Section 1.d. Transposing

You may want to flip the rows and columns, so that the above table looks like:

Model Summary

		Model
		1
Variables	Entered	*EDUCATION, GENDER, WORK_EX, PUB_SEC*
	Removed	.
R Square		*.492*
Adjusted R Square		*.491*
Std. Error of the Estimate		*5.2242*

To do this, choose PIVOT/TURN ROWS INTO COLUMNS. Compare the table above to the one before it. Notice that the rows in the previous table have become columns in this example and the columns in the previous table have become rows in this example.

Ch 11. Section 1.e. Finding appropriate width and height

You may want to let SPSS resize the rows and columns so that the labels and data fit exactly. To do this, choose FORMAT/AUTOFIT. The result is shown in the next table. Compare it with the original (the table above).

Model Summary

		Model
		1
Variables	Entered	*EDUCATION, GENDER, WORK_EX, PUB_SEC*
	Removed	.
R Square		*.492*
Adjusted R Square		*.491*
Std. Error of the Estimate		*5.2242*

Autofit is a quick method to ensure that the row heights and column widths are adequate to display the text or data in each individual cell.

Ch 11. Section 1.f. Deleting specific cells

You may want to delete specific entries in a cell. In the example above, the entries “Model” and “1” are superfluous. To delete them, choose the cell with “Model," then press "delete" on the keyboard. Then choose the cell with “1," and press "delete" to get:

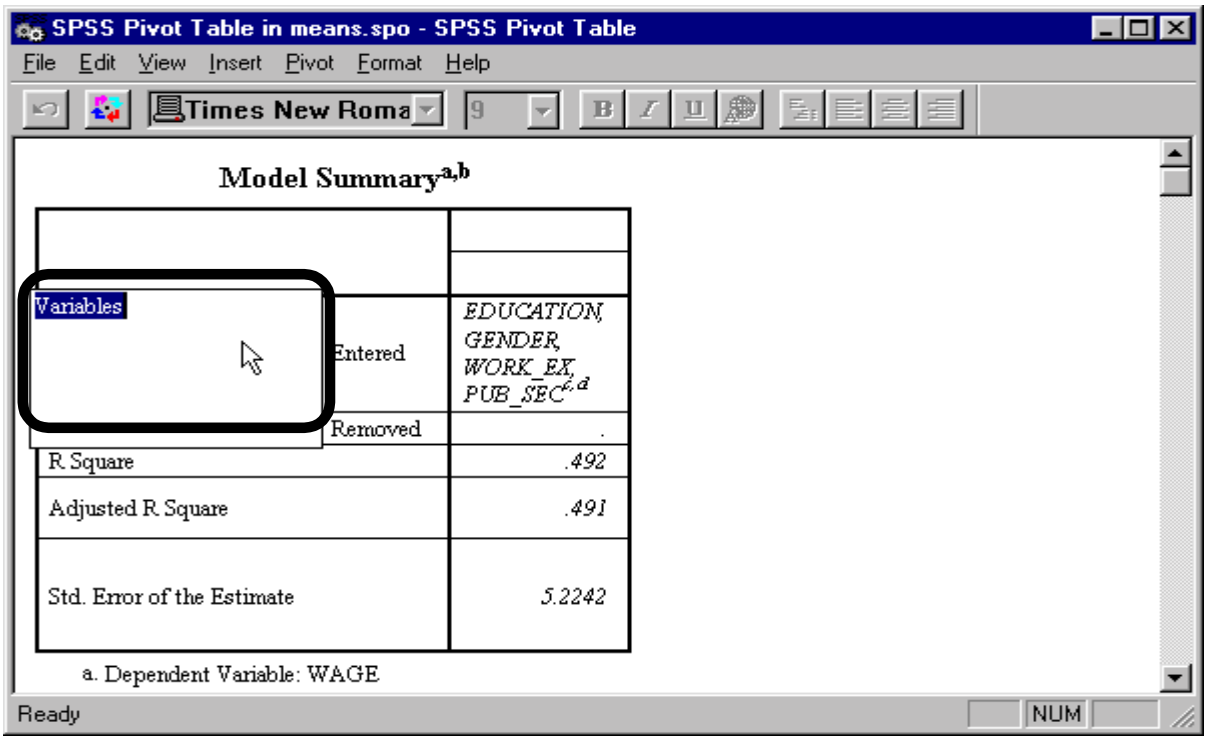
Model Summary

Variables	Entered	*EDUCATION, GENDER, WORK_EX, PUB_SEC*
	Removed	.
R Square		*.492*
Adjusted R Square		*.491*
Std. Error of the Estimate		*5.2242*

Note: If you make a mistake, go to EDIT/UNDO.

Ch 11. Section 1.g. Editing (the data or text in) specific cells

You may want to edit specific cells. To edit a cell, choose the cell by clicking on it with the left mouse then double click on the left mouse. The result of these mouse moves is, as you see in the next picture, that the cell contents are highlighted, implying that you are in edit mode. You are not restricted to the editing of cells - you can use a similar method to edit the title, footnotes, etc.



Now, whatever you type in will replace the old text or number in the cell. Type the new text "Explanatory Variables" on top of the highlighted text “Variables." The table now looks like:

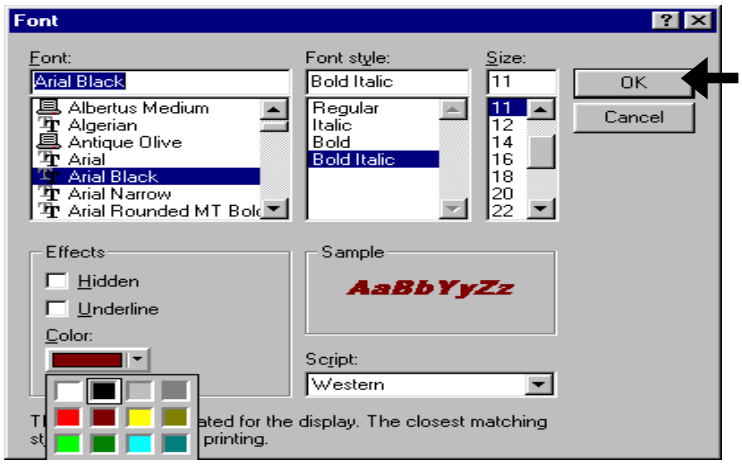
Model Summary

Explanatory Variables	Entered	*EDUCATION, GENDER, WORK_EX, PUB_SEC*
	Removed	.
R Square		*.492*
Adjusted R Square		*.491*
Std. Error of the Estimate		*5.2242*

Ch 11. Section 1.h. Changing the font

Choose the cells by clicking on the first cell and lightly dragging the mouse over the others until they are all highlighted. Go to FORMAT/FONT. Select the font you want. Click on "OK."

The next table is the same as the previous example except that the cell with the variable names is displayed using a different font.



Model Summary

Explanatory Variables	Entered	***EDUCATION, GENDER, WORK_EX, PUB_SEC***
	Removed	.
R Square		*.492*
Adjusted R Square		*.491*
Std. Error of the Estimate		*5.2242*

The font has changed.

Advice: Return to sections 11.1.b - 11.1.h and observe how the table has changed in appearance. Try the methods on any SPSS output table you have.

Ch 11. Section 1.i. Inserting footnotes

You may want to insert a footnote to clarify or qualify a certain statistic or value. For example, in the above table, you may want to place a footnote reference to the cell “Model Summary." To do so, you must first choose the cell that has the text "Model Summary." Then, go to INSERT/FOOTNOTE. With the footnote highlighted, double click with the left mouse. Type in the desired text

Model Summary [a]

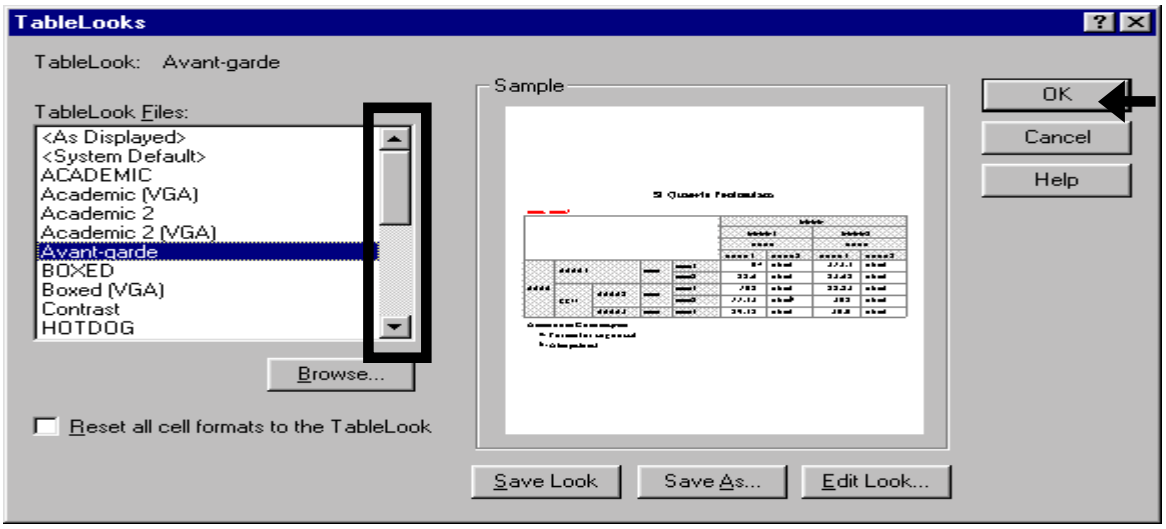
Explanatory Variables	Entered	***EDUCATION, GENDER, WORK_EX, PUB_SEC***
	Removed	.
R Square		*.492*
Adjusted R Square		*.491*
Std. Error of the Estimate		*5.2242*

a. Dependent variable is wage

Footnote indicator (the superscript "a")

Ch 11. Section 1.j. Picking from pre-set table formatting styles

To quickly re-set the table style (implying a combination of style features: text orientation, font type and size, shading, coloring, gridlines, etc.), choose FORMAT/TABLELOOK[134].



Scroll through the "TableLooks." A sample of the TableLook you have highlighted will be shown in the area "Sample." Select a look that you prefer. We have chosen the look "AVANT-GARDE." The table will be displayed using that style:

[134] To learn how to change the default look, see section 12.1.

Model Summary [a]

Explanatory Variables	**Entered**	***EDUCATION, GENDER, WORK_EX, PUB_SEC***
	Removed	.
R Square		***.492***
Adjusted R Square		***.491***
Std. Error of the Estimate		***5.2242***

a. Dependent variable is wage

Notice how so many of the formatting features have changed with this one procedure. Compare this table with that in the previous example - the fonts, borders, shading, etc. have changed.

If you want to set a default look for all tables produced by SPSS, go to EDIT/OPTIONS within the main data windows interface and click on the tab "Pivot Table." You will see the same list of TableLooks as above. Choose those you desire. Click on "Apply" and then on "OK." See also: chapter 15.

Ch 11. Section 1.k. Changing specific style properties

You can make more specific changes to formatting attributes of the entire table (or chosen cells) using FORMAT/TABLE PROPERTIES (or FORMAT/CELL PROPERTIES). The attributes include:

- Shading
- Data format
- Alignment
- Footnotes
- Borders
- Gridlines

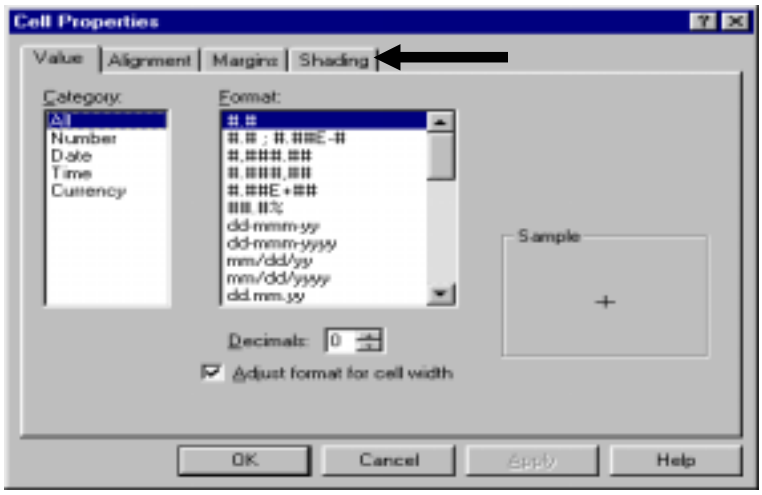
The next few sub-sections show how to change these features.

Ch 11. Section 1.l. Changing the shading of cells

Select those cells whose shading you wish to change. Click on the first cell and then lightly move the cursor over the cells.

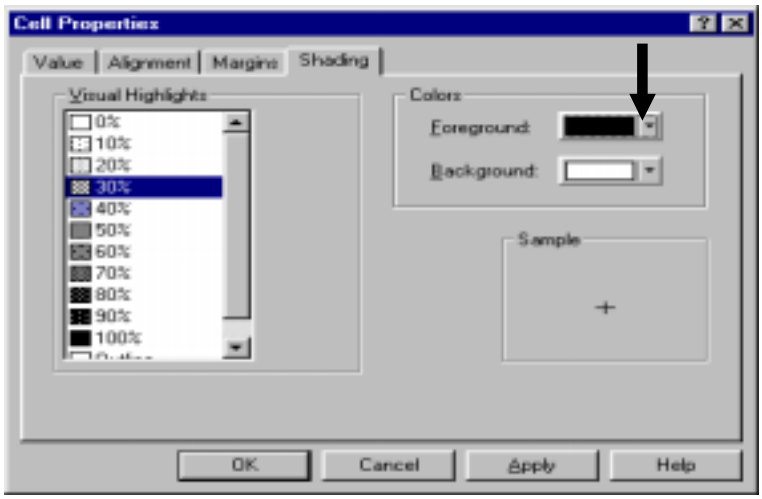
Go to FORMAT/CELL PROPERTIES.

Click on the tab “Shading.”



Click on the down arrow next to the black rectangle to the right of the label “Foreground.”

You will have a choice of shadings.



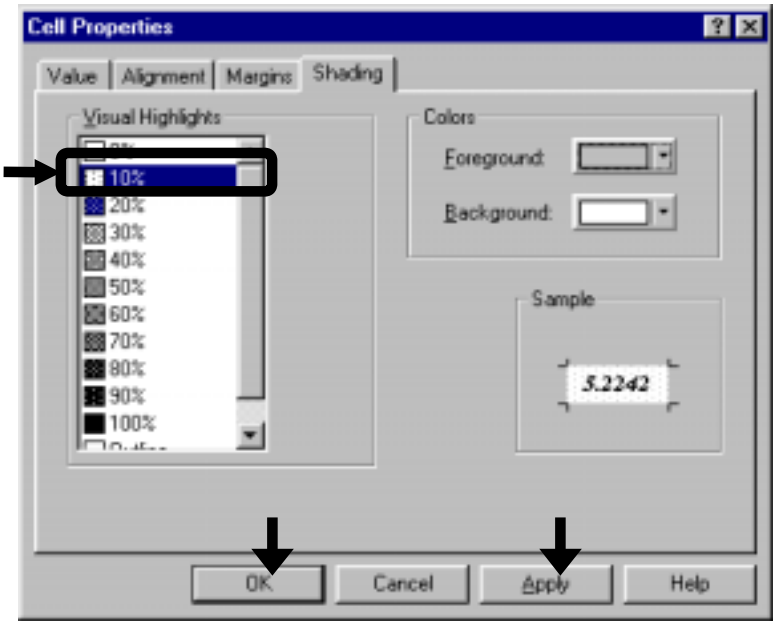
Click on the shade and color you desire.

You may want to change the patterns. To do so, click in the area "Visual Highlights" and make a choice.

Click on "Apply."

Click on "OK."

Note: Result not shown.



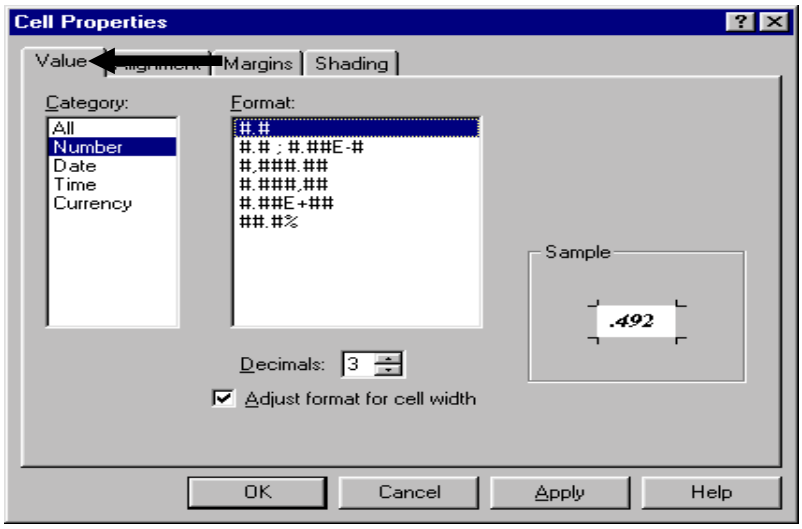
Ch 11. Section 1.m. Changing the data format of cells

Let's assume you want to reduce the number of decimals in the data cells of the table.

To do so, you must first select the cells. Then, click on the first cell and lightly move the cursor over the cells.

Go to FORMAT/CELL PROPERTIES.

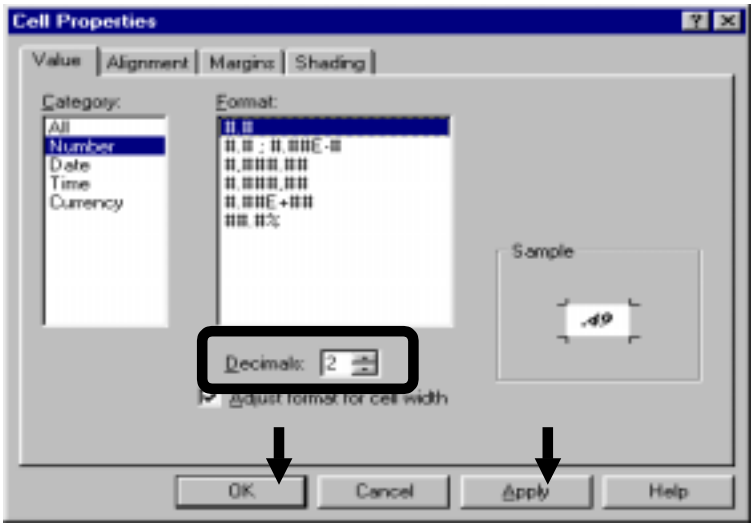
Click on the tab "Value."



Reduce the number of decimal places to 2 by clicking on the down arrow next to the label "Decimals."

Click on "Apply."

Click on "OK."



Note that the numbers are displayed with only two decimal places (Compare to the tables in sections 11.1.b to 11.1.j, all of which had 3 decimal places).

Model Summary [a]

Explanatory Variables	**Entered**	***EDUCATION, GENDER, WORK_EX, PUB_SEC***
	Removed	.
R Square		***.49***
Adjusted R Square		***.49***
Std. Error of the Estimate		***5.22***

a. Dependent variable is wage

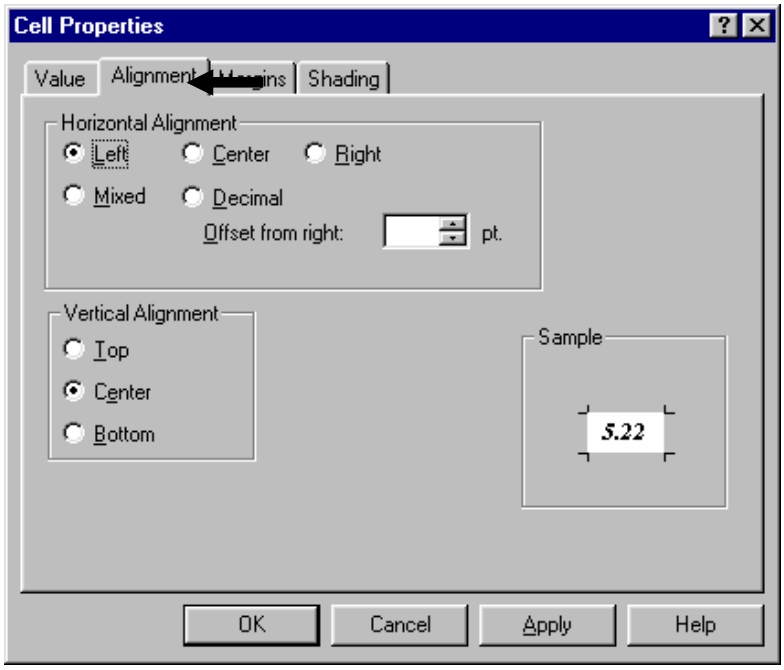
Ch 11. Section 1.n. Changing the alignment of the text or data in cells

Let's assume you want to center the numbers in the data cells.

To do so, you must first select the cells. Then, click on the first cell and lightly move the cursor over the cells.

Go to FORMAT/CELL PROPERTIES.

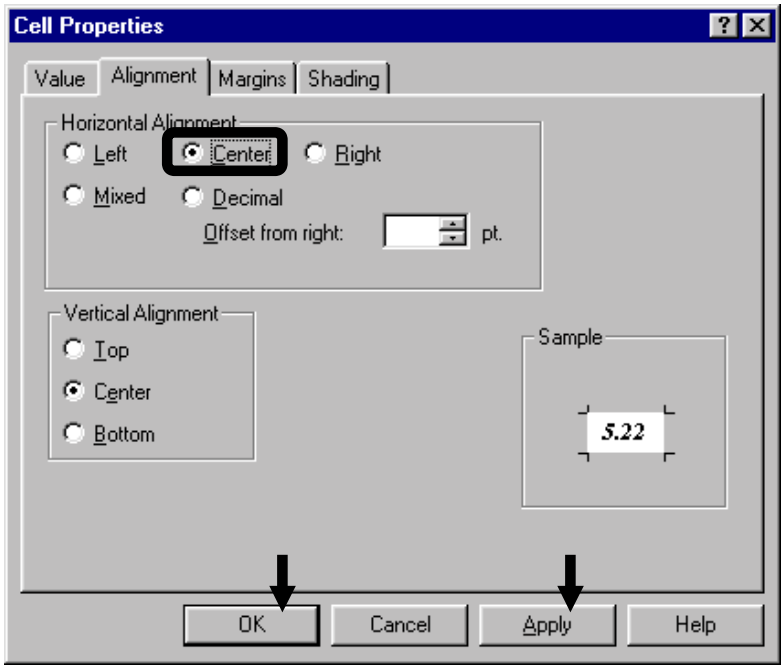
Click on the tab "Alignment."



Click on the round button to the left of "Center" in the area "Horizontal Alignment."

Click on "Apply."

Click on "OK."



Now the cells that include data are aligned in the center. Compare these cells to the corresponding cells in the table produced in section 11.1.m.

Model Summary[a]

Explanatory Variables	**Entered**	***EDUCATION, GENDER, WORK_EX, PUB_SEC***
	Removed	.
R Square		***.49***
Adjusted R Square		***.49***
Std. Error of the Estimate		***5.22***

a. Dependent variable is wage

Ch 11. Section 1.o. Formatting footnotes

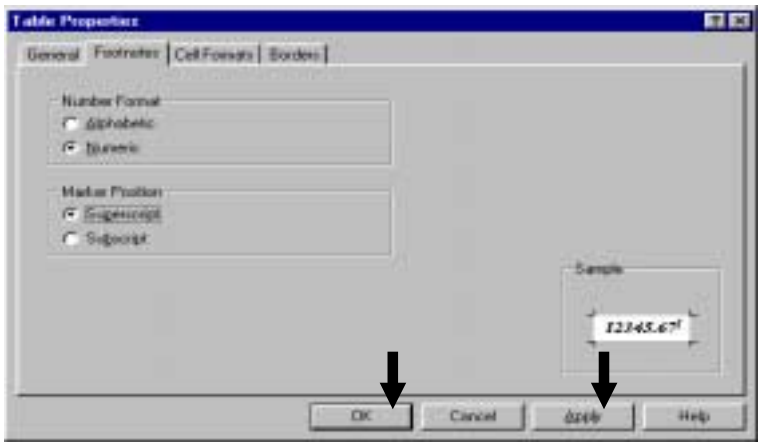
The table above uses numbers to mark the footnote reference. Let's assume you want to use alphabets instead of numbers.

To do so, go to FORMAT/TABLE PROPERTIES.

Select the formats for the footnotes.

Click on "Apply."

Click on "OK."



The footnote is referenced as "1" now, rather than "a" as it was in the table above

Note: To change the text in a footnote, double click on the footnote and follow the steps in section 11.1.g. To change the font, follow section 11.1.h. To change the shading, follow section 11.1l.

Model Summary[1]

Explanatory Variables	**Entered**	**EDUCATION, GENDER, WORK_EX, PUB_SEC**
	Removed	.
R Square		**.49**
Adjusted R Square		**.49**
Std. Error of the Estimate		**5.22**

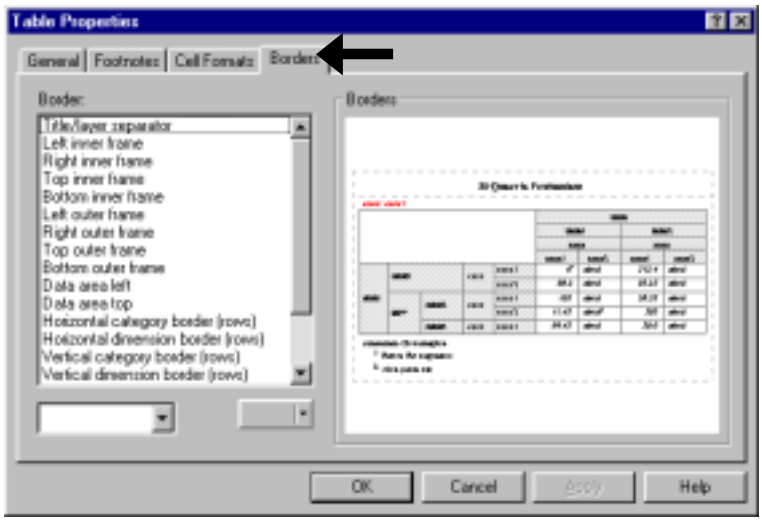
1. Dependent variable is wage

Ch 11. Section 1.p. Changing borders and gridlines

Go to FORMAT/TABLE PROPERTIES.

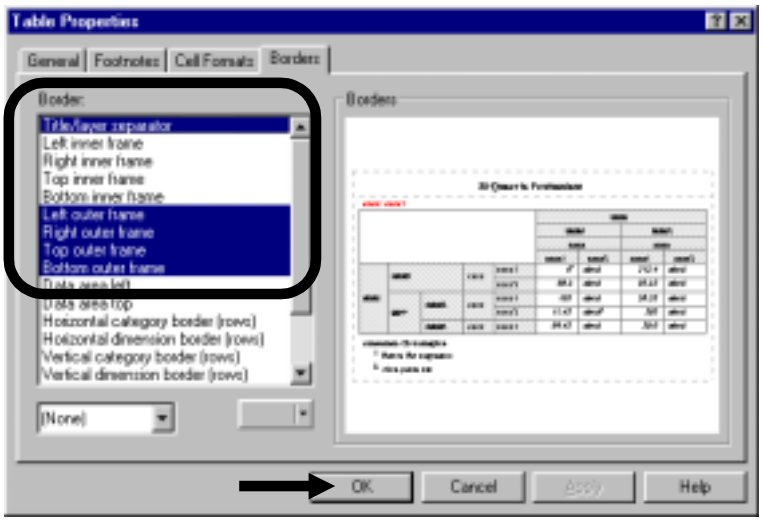
Click on the tab “Borders.”

You can see a number of items on the left side of the dialog box. These define the specific components of an output table.



Select the borders you desire by clicking on the options in the left half of the dialog box, under the title “Border.”

Click on “OK.”



The requested borders are displayed.

Note: This feature does not always work.

Model Summary[1]

Explanatory Variables	**Entered**	***EDUCATION, GENDER, WORK_EX, PUB_SEC***
	Removed	**.**
R Square		***.49***
Adjusted R Square		***.49***
Std. Error of the Estimate		***5.22***

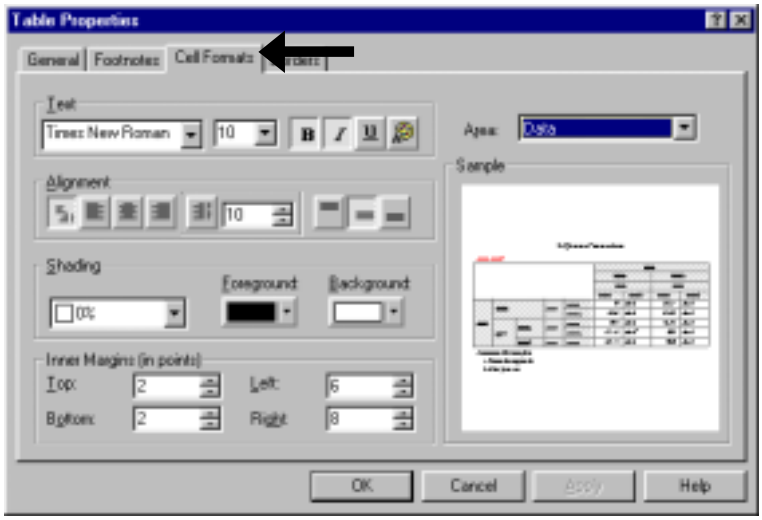
1. Dependent variable is wage

Ch 11. Section 1.q. Changing the font of specific components (data, row headers, etc.)

Let's assume you want to change the font of all the "row headers" in a table (but not the font of any other item). The method shown in section 11.1.h can be used, but it can be tedious. An easier method is shown here.

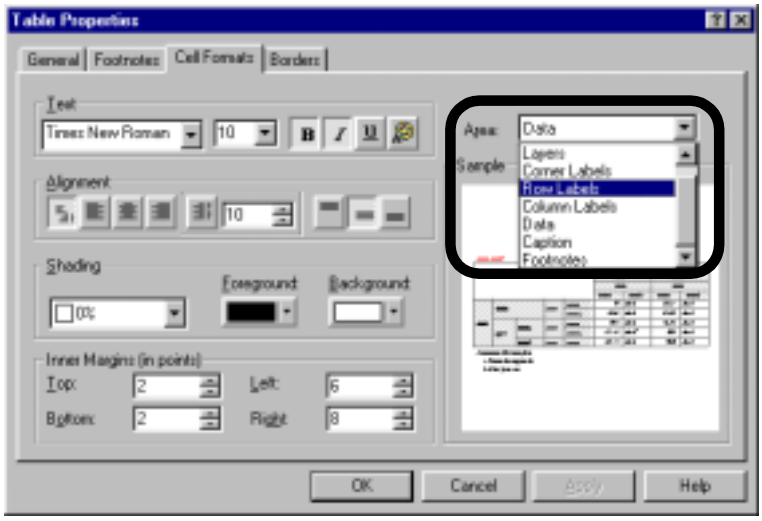
Go to FORMAT/TABLE PROPERTIES.

Click on the tab “Cell Formats.”



Select the component of the table whose font you wish to change. As an example, we have chosen “Row Labels.”

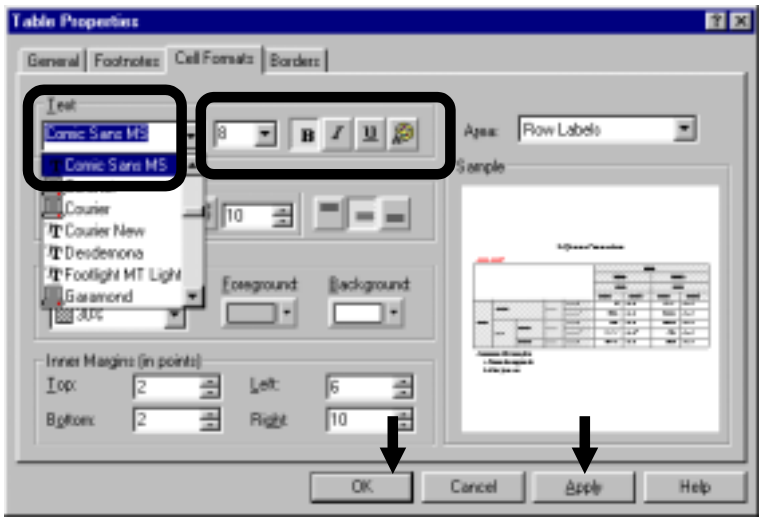
Note: Look at the list of components. Can you understand them? The "Row Labels" are the titles for each row. The "Column Labels" are the titles for each column. The "Data" are the cells that contain the numeric results. The "Caption" is the table title.



Make the choices. We have made the choice of font type “Comic Sans MS, size 8.” We have also removed the italics by clicking on the button “I.”

Click on “Apply.”

Click on “OK.”



The font style will be changed for all the row headers, but nothing else changes.

Once you are satisfied with the formatting, go to EDIT/COPY TABLE and open Word (or any other software you are using) and choose EDIT/PASTE. To save on storage space, you may want to choose EDIT/PASTE SPECIAL/PICTURE.

Model Summary[1]

Explanatory Variables	Entered	*EDUCATION, GENDER, WORK_EX, PUB_SEC*
	Removed	.
R Square		*.49*
Adjusted R Square		*.49*
Std. Error of the Estimate		*5.22*

1. Dependent variable is wage

Ch 11. Section 2 Formatting and editing charts

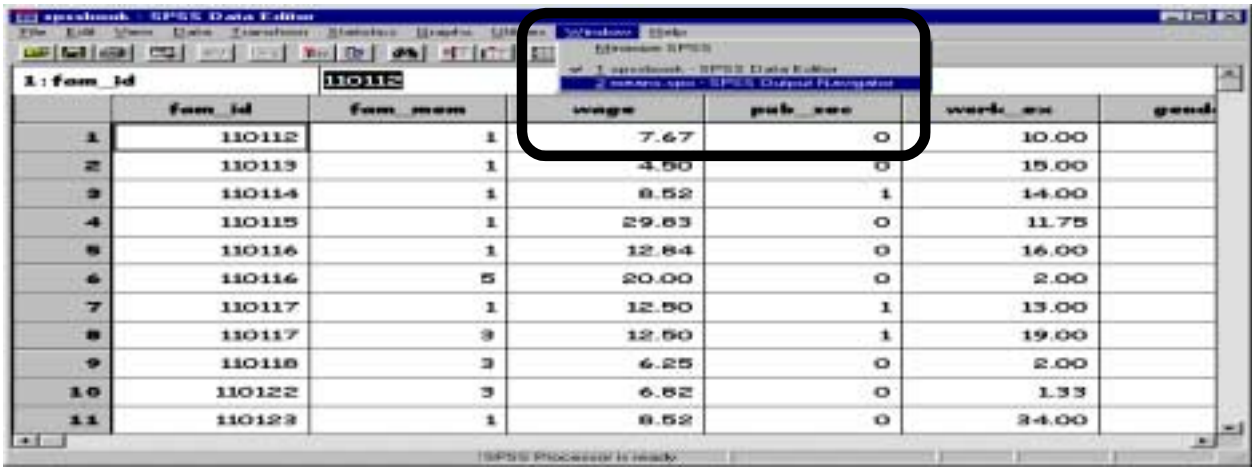
Well-presented tables are important, but properly edited and formatted charts are absolutely essential. There are two reasons for this:

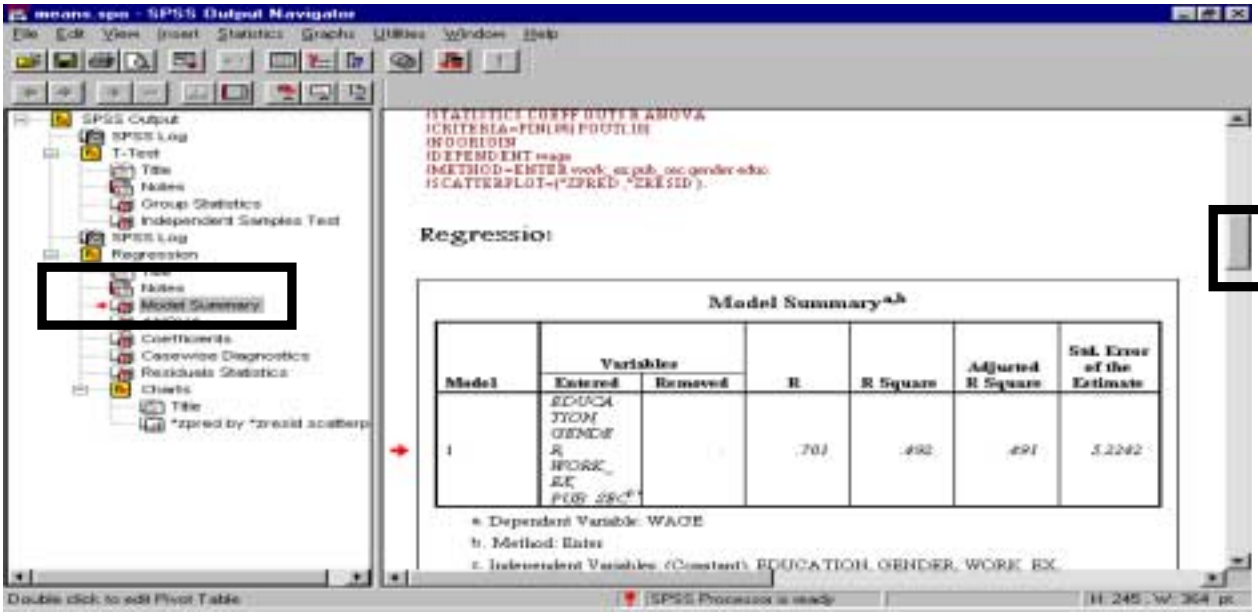
1. The main purpose of charts is to depict trends and to enable visual comparisons across categories/variables. As such, it is imperative that the chart shows exactly what you want. Otherwise, the powerful depictive capability of a chart may highlight inessential features or hide important ones[135].
2. SPSS charts may not have the "proper" formatting more often than tables. Again, this arises because of the nature of charts – a few extreme values can change the axis scales dramatically, thereby hiding the visual depiction of trends.

Ch 11. Section 2.a. Accessing the window for formatting / editing charts

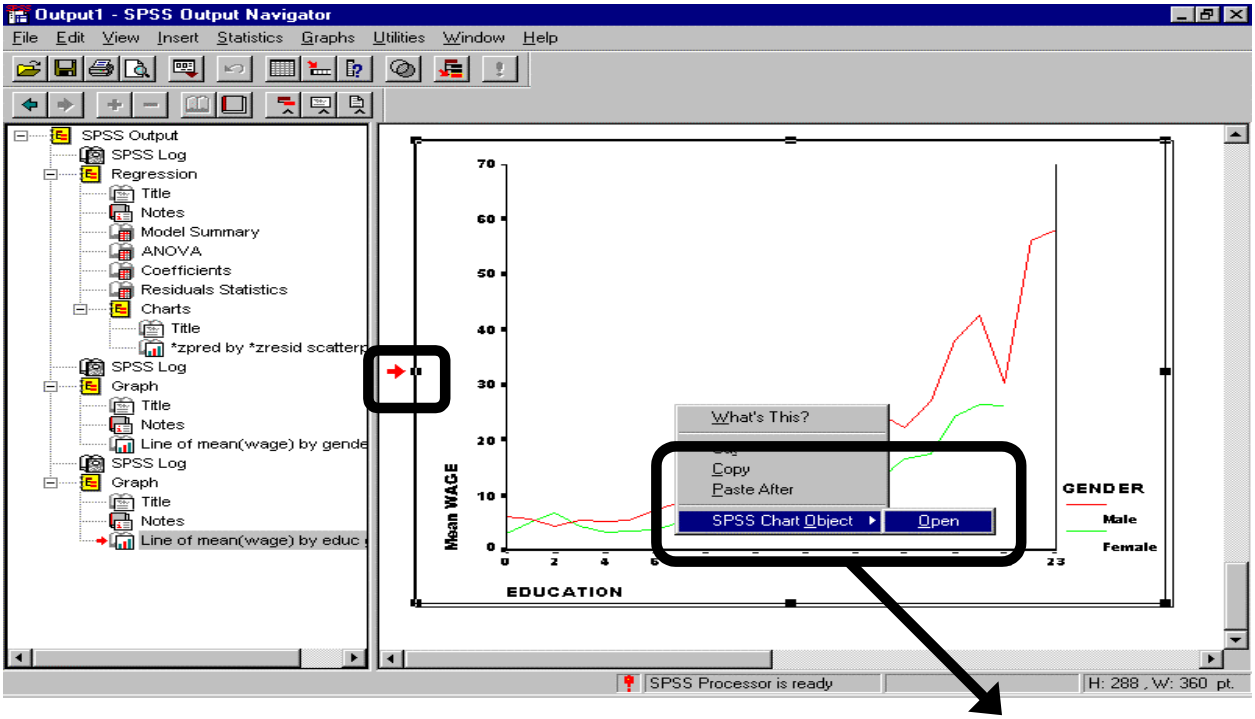
To format charts, you must enter the formatting/editing window. To do so, go to WINDOWS/SPSS OUTPUT NAVIGATOR.

[135] Formatting reduces the ever-present possibility of confusion. A common problem is the presence of outliers. These tend to expand massively the scale of an axis, thereby flattening out any trends.



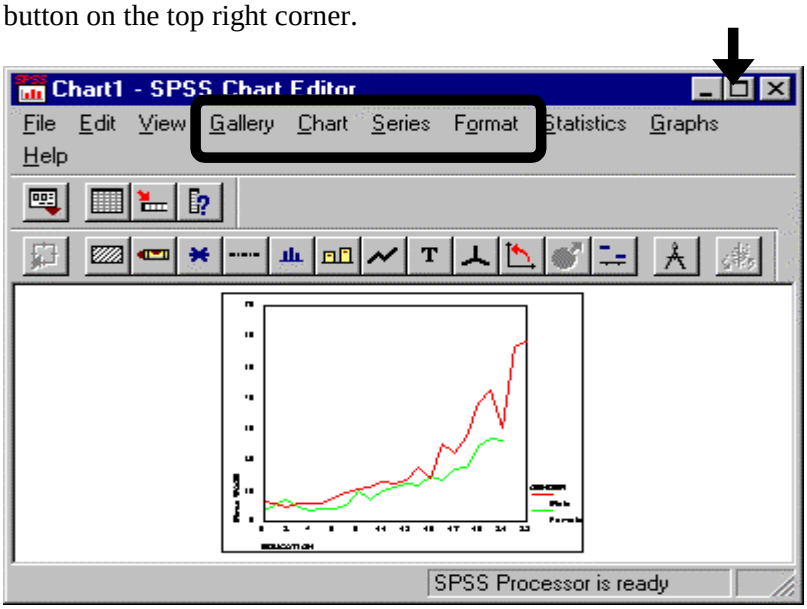


Using the scroll bar on the extreme right, scroll to the chart you want to format or click on the chart's name in the left half of the window. If you see a red arrow next to the chart, then you have been successful in choosing the chart.



To edit/format the chart, click on the right mouse and choose the option "SPSS Chart Object/Open" or double click on the chart with the left mouse.

A new window called the "Chart Editor" will open. Maximize it by clicking on the maximize button on the top right corner.



Notice that there are four new menus. The menus are:

1. GALLERY: this allows you to change the chart type. You can make a bar chart into a line, area, or pie chart. You can change some of the data into lines and some into bars, etc. So, if you made a bar chart and feel that a line chart would have been better, you can make the change right here. If you have too many variables in the chart, then you might want to mix the chart types (bar, line, and area). On the next few pages, we illustrate the use of this menu.

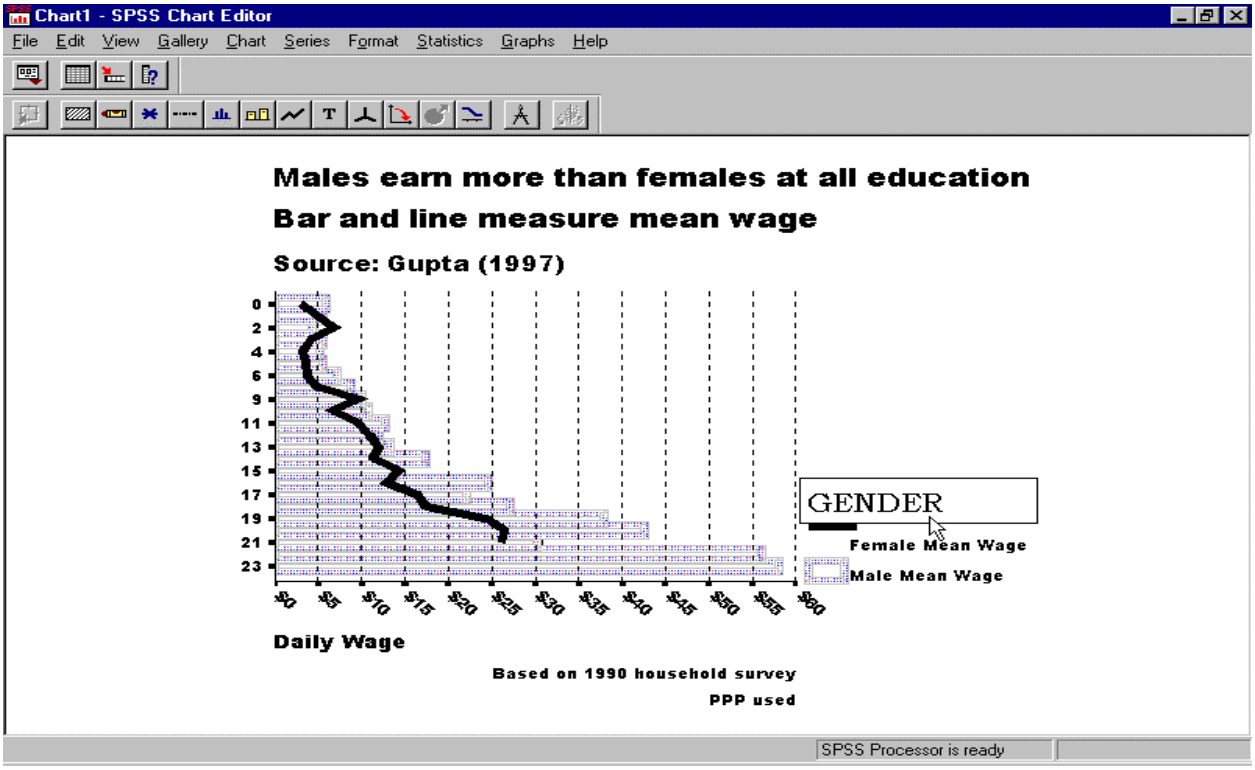
2. CHART: using this, you can change the broad features that define a chart. These include the frames around and in the chart and titles, sub-titles, footnotes, legends, etc.

3. SERIES: this allows you to remove certain series (variables) from a chart.

4. FORMAT: using this, you can format the fonts of text in labels, titles, or footnotes, format an axis, rescale an axis, swap the X and Y axes, and change the colors, patterns, and markers on data lines/bars/areas/pie slices, etc.

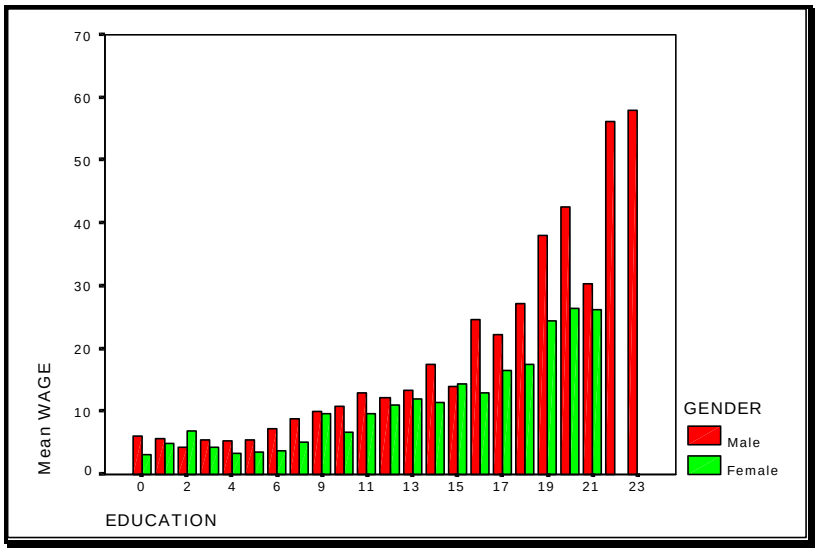
Ch 11. Section 2.b. Using the mouse to edit text

Click on the text you wish to edit. A box-like image will be shown around the text (look at the label "GENDER" below). To edit this, double click inside the box and make the desired changes.



Ch 11. Section 2.c. Changing a chart from bar type to area/line type (or vice versa)

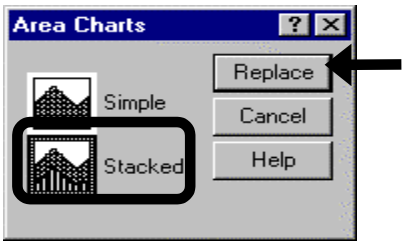
We will use this chart as the original chart. It was made using the GRAPHS/BAR function (see chapter 5 to learn how to make this chart).



You can use the GALLERY menu to convert the above chart into a different type of chart.

Go to GALLERY/AREA[136].

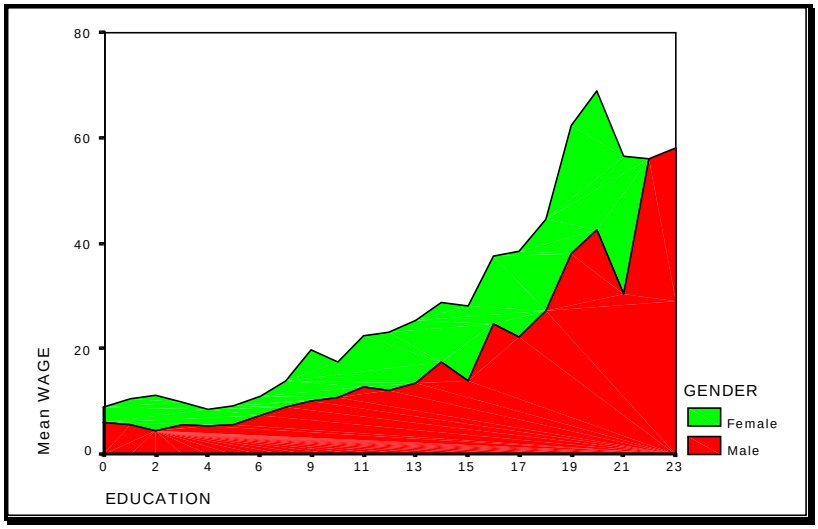
Select "Stacked" and click on "Replace."



The original bar chart is replaced with the following area chart.

No other formatting feature (title, fonts, axis, etc.) has changed.

Note: Different manners of indicating values (bar, line, area, pie, etc.) may be appropriate given the purpose of the graph. Line graphs are usually good when trends must be compared. Bar graphs are good when the X-axis has only a few categories. Area graphs are used mostly for depicting aggregate functions.



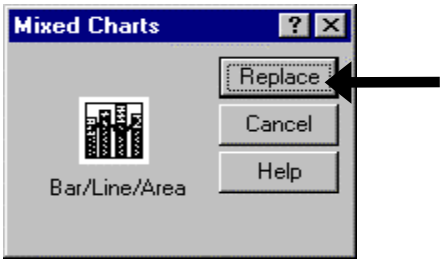
[136] If you chose GALLERY/LINE, a line graph would be the result of the conversion from a bar graph.

Ch 11. Section 2.d. Making a mixed bar/line/area chart

A very powerful use of GALLERY is to make a kind of chart that cannot be made using the regular SPSS GRAPH menu options (in all SPSS versions till 8.0). This is the "Mixed" chart, so named because it mixes different chart styles on one plot. Using such a chart, different series can be depicted using different types of charting tools. For an example, see the graph on the next page.

Go to GALLERY/MIXED CHARTS.

Click on "Replace."

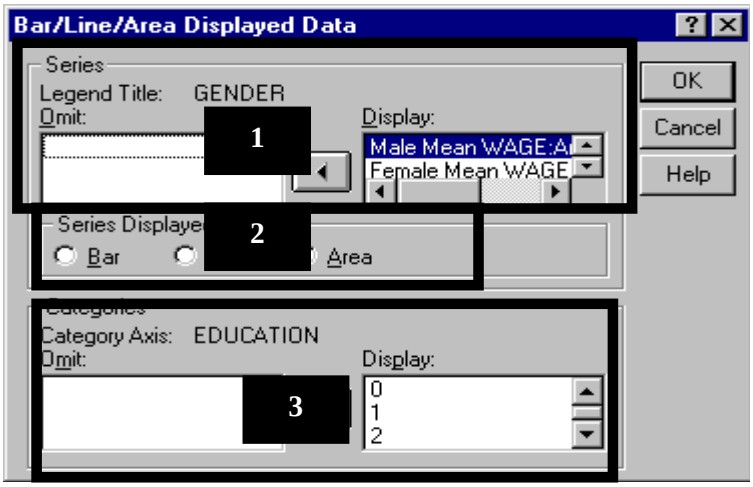


The following dialog box opens.

Area 1 allows you to choose which series to omit/display.

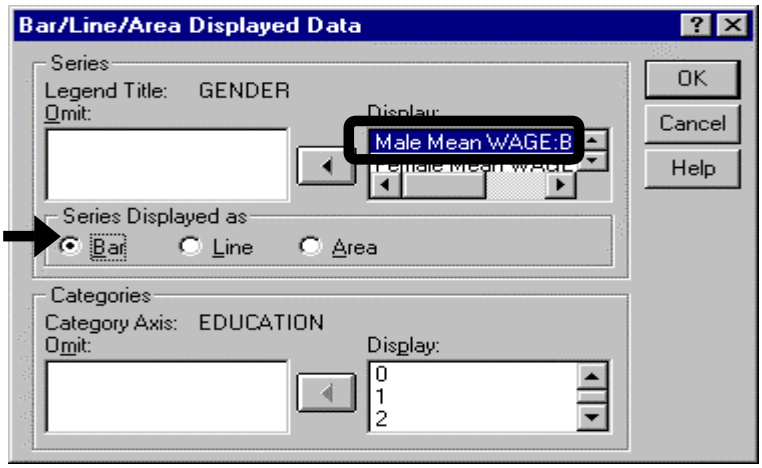
In area 2 you can select the style used to depict a series chosen for display in area 1.

In area 3 you can choose to omit a range of values from the X (or category) axis[137].



Let's assume you want to use bars to display the series *Male Mean Wage* (or any other series already used in the chart you are editing).

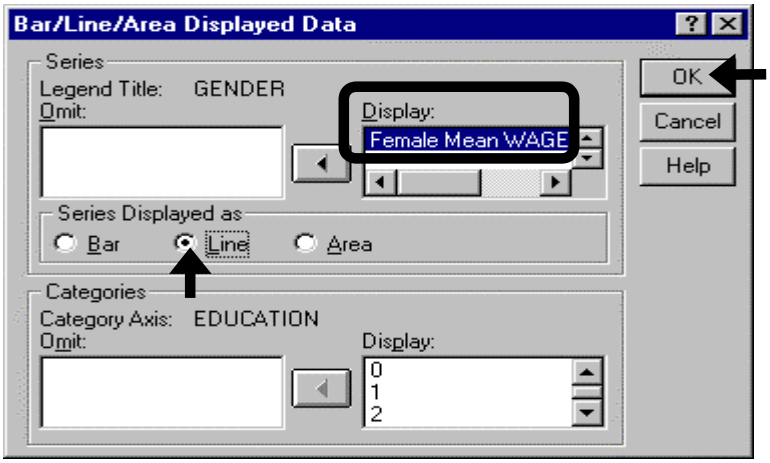
To do so, click on the series *Male Mean Wage* in the box "Display." Then go to the area "Series Displayed As" and choose the option "Bar." Now, *Male Mean Wage* will be displayed using bars.



[137] This feature is rarely used.

Assume you want to use a line to display the series *Female Mean Wage*. Click on the series name *Female Mean Wage* in the area "Display." Select the option "Line" from the options in "Series Displayed As."

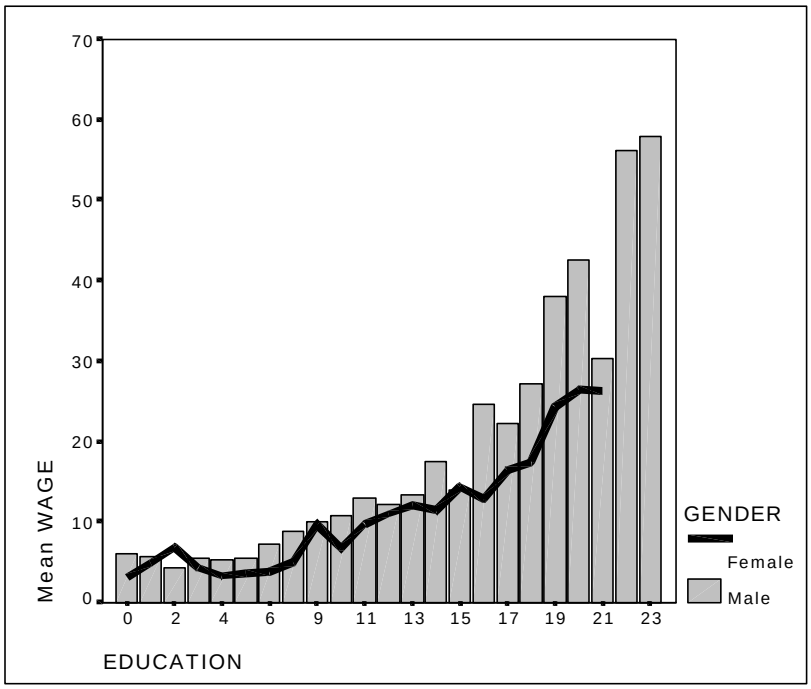
Click on "OK."



The chart has been transformed. Now, the mean wage of females is displayed using a line.

This type of chart is called "Mixed" because, as is plainly shown in the plot on the right, it mixes different styles of displaying series - bars, lines, etc.

Note: In this chart, one can compare the mean wage across gender more easily than in a bar chart because the line cuts through the bars. In a pure bar chart, the indicators for males and females would be adjacent to each other, making visual comparison more difficult.



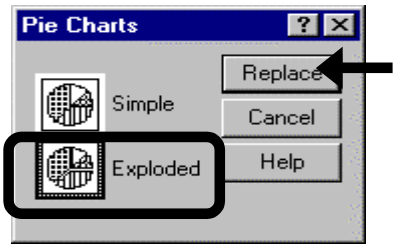
Ch 11. Section 2.e. Converting into a pie chart

The data for only **one** series can be converted to a pie chart[138].

Go to GALLERY/PIE.

Select the option "Simple" or "Exploded" (the difference is purely cosmetic).

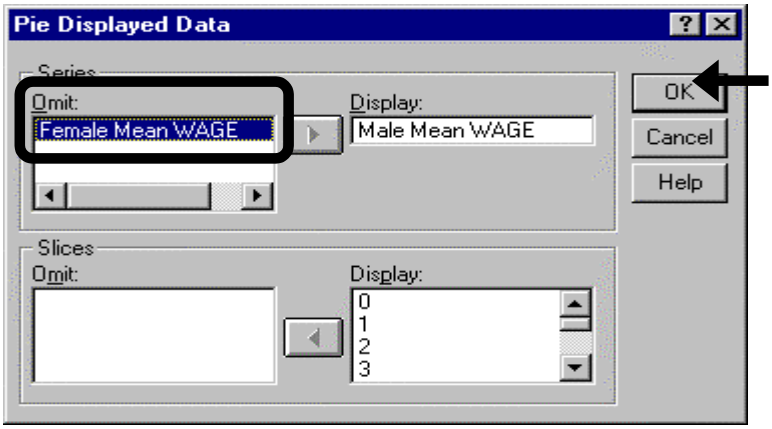
Click on "Replace."



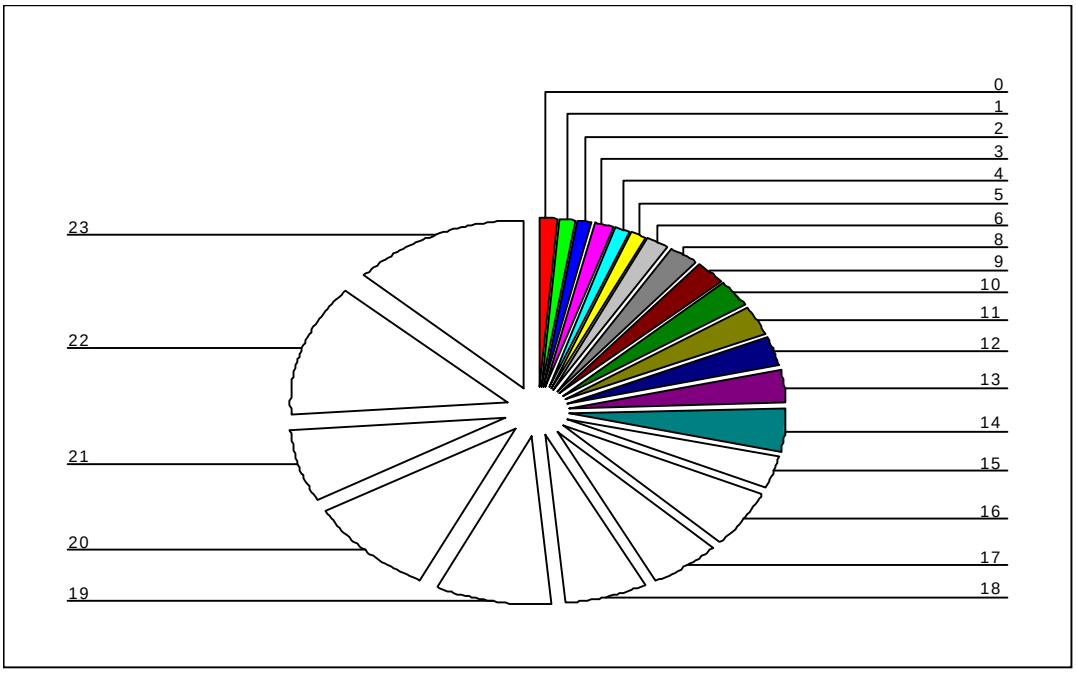
[138] The other series must therefore be hidden.

SPSS will place all the series into the box “Omit." Then you can choose the one series that you want to display by moving it into the box “Display" (a pie chart can display only one series).

Click on “OK."

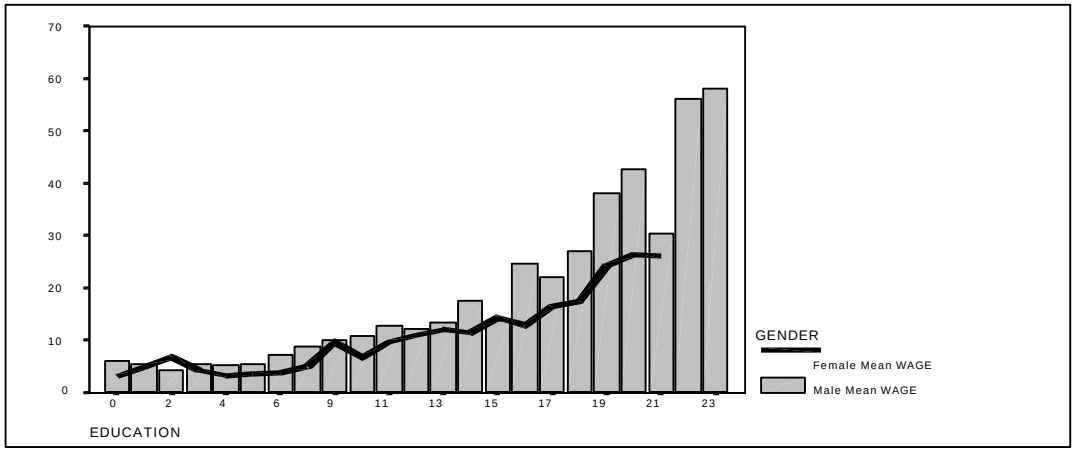


Each slice represents the mean wage for males with a specific level of educational attainment. For example, the slice “6” shows the mean wage for males who had an educational attainment of six years (primary schooling).



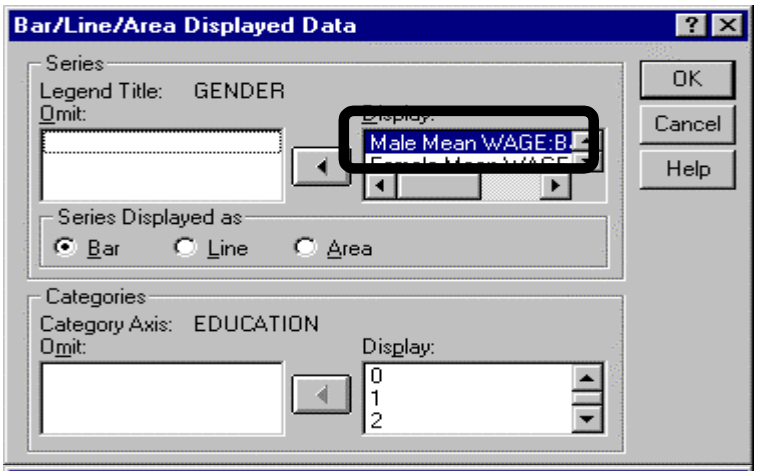
Ch 11. Section 2.f. Using the SERIES menu: Changing the series that are displayed

To show the use of this menu, we go back to our "Mixed" chart (shown below), which we made in section 11.2.d.



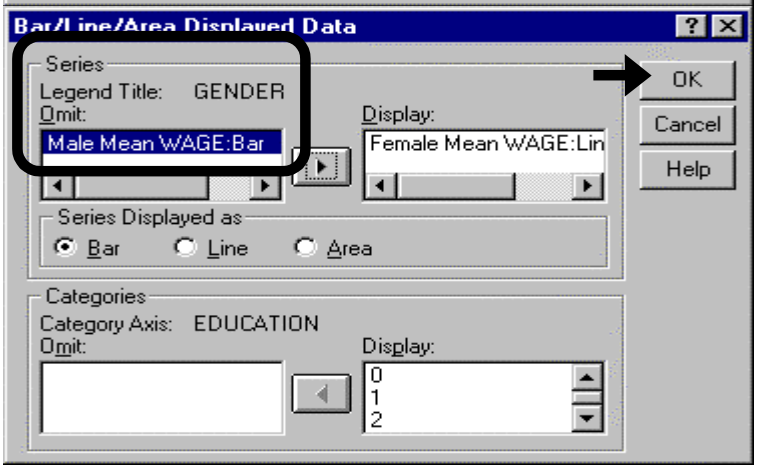
Go to SERIES/DISPLAYED.

Select the series you want to omit.



Move it into the box "Omit" in the area "Series."

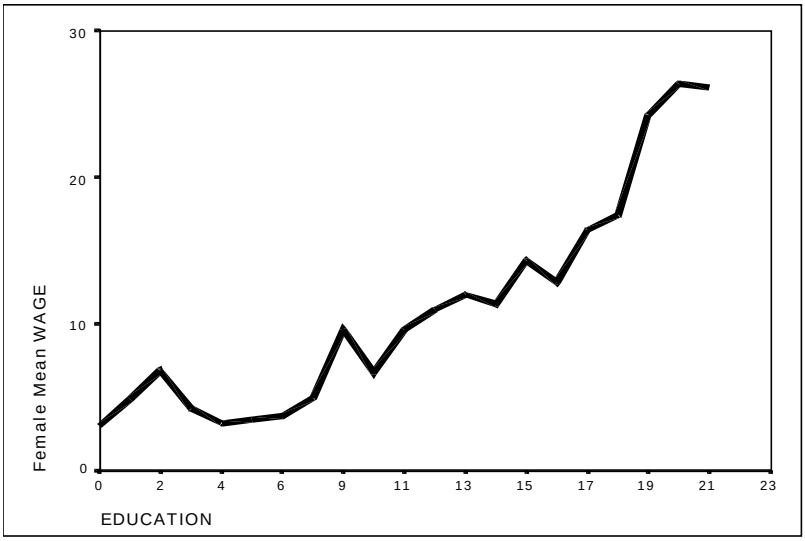
Click on "OK."



Note: Look at the area "Categories." Using the two boxes there ("Omit" and "Display"), you can choose to omit certain values of the category axis (the X-axis). Here the X-axis variable is *education*. If you want to omit everyone with an education level of zero, then move the number 0 from the box "Display" to the box "Omit."

Male Mean Wage is no longer displayed.

Note: Until now, we were showing features that used the menu choices GALLERY and SERIES. At this stage, we advise you to practice what you have learned in sections 11.2.a. - 11.2.f.



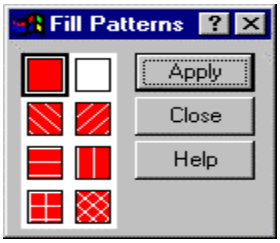
Ch 11. Section 2.g. Changing the patterns of bars, areas, and slices

Now we are focusing on using the FORMAT menu.

We go back to the chart from section 11.2.d to illustrate this topic.

For the series/point whose pattern you want to change, click on the relevant bar(s)/area(s)/slice(s).

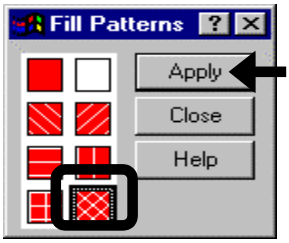
Then go to FORMAT/PATTERNS.



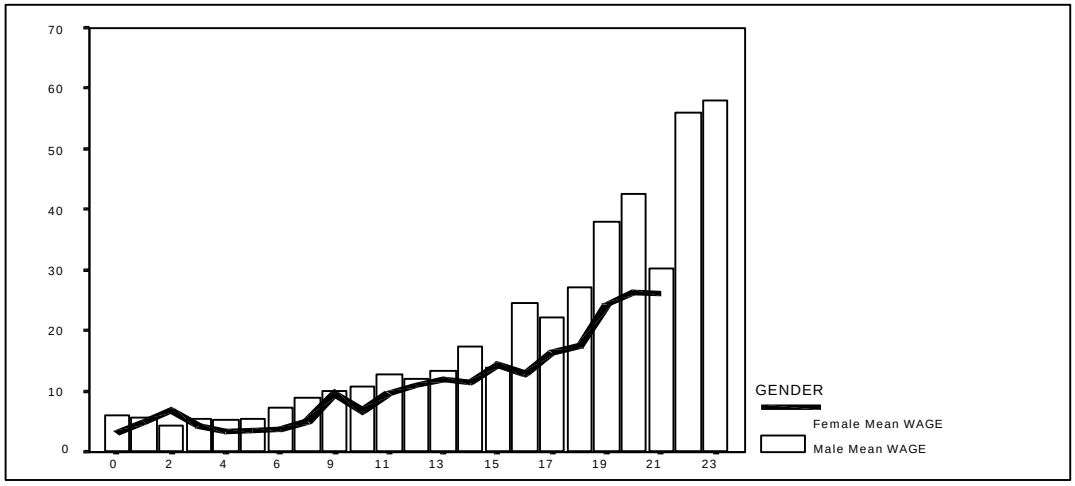
Select the pattern you desire.

Click on “Apply."

The series will be displayed with the pattern you have just chosen.



See the style of the bars in the chart below. Of course, this format may need to be modified slightly in order to achieve the look that you desire. Some experimentation in pattern choice is essential if you have a black-and-white printer.



Ch 11. Section 2.h. Changing the color of bars, lines, areas, etc.

For the series/point whose color you want to change, click on a bar/area/line/slice.

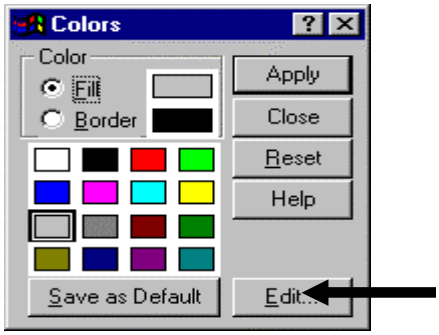
Then go to FORMAT/ COLORS.

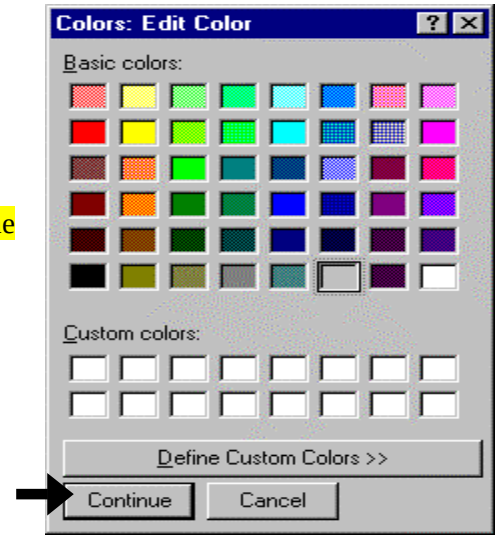
Select from one of the displayed colors.

If you want some other color, click on “Edit."
A new dialog box opens up. You can choose a color from this wide range of options.

If you want to add some custom colors to this, simply press on "Define Custom Colors" and pick the exact color you wish to use.

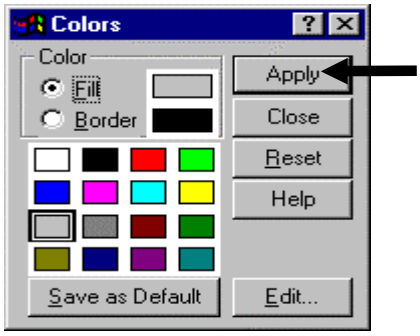
Click on “Continue."





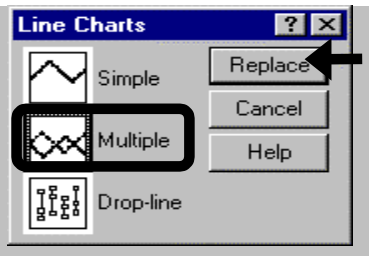
Click on “Apply."

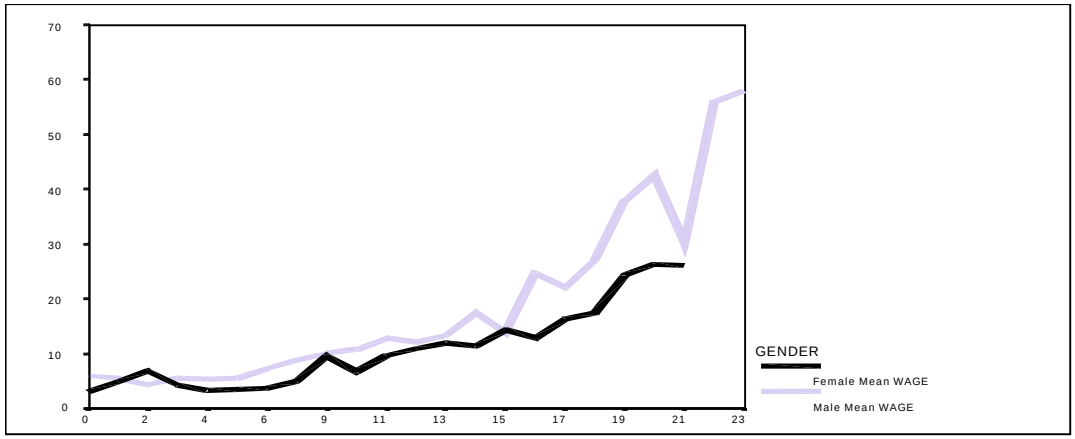
Note: The result is not shown here.



Ch 11. Section 2.i. Changing the style and width of lines

This option is valid only for line charts. Our chart (from section 11.2.g) has one series displayed as a line. If you choose, you can display both series as lines by using GALLERY/ LINE and choosing the options as shown.





Now we want to change the style and width of the lines and their markers.

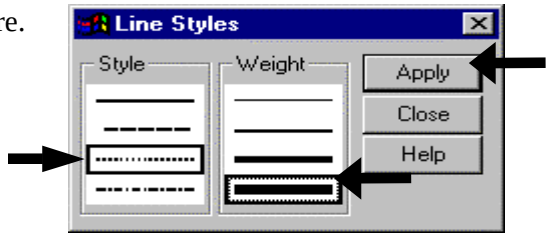
Select the line (series) you want to format by clicking on it. In our example, we chose *Male Mean Wage*.

Go to FORMAT/LINE STYLES.

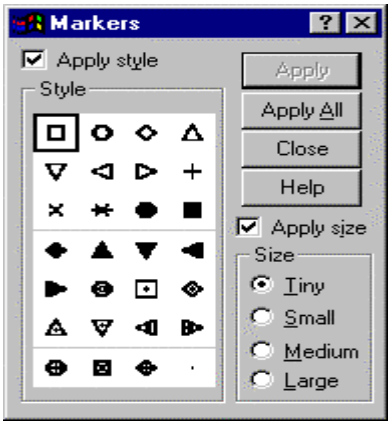
Select the style and width you desire.

Click on “Apply."

It is best to change the data markers at this time.



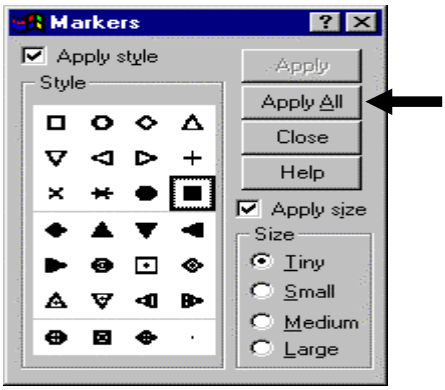
Go to FORMAT/ MARKERS.

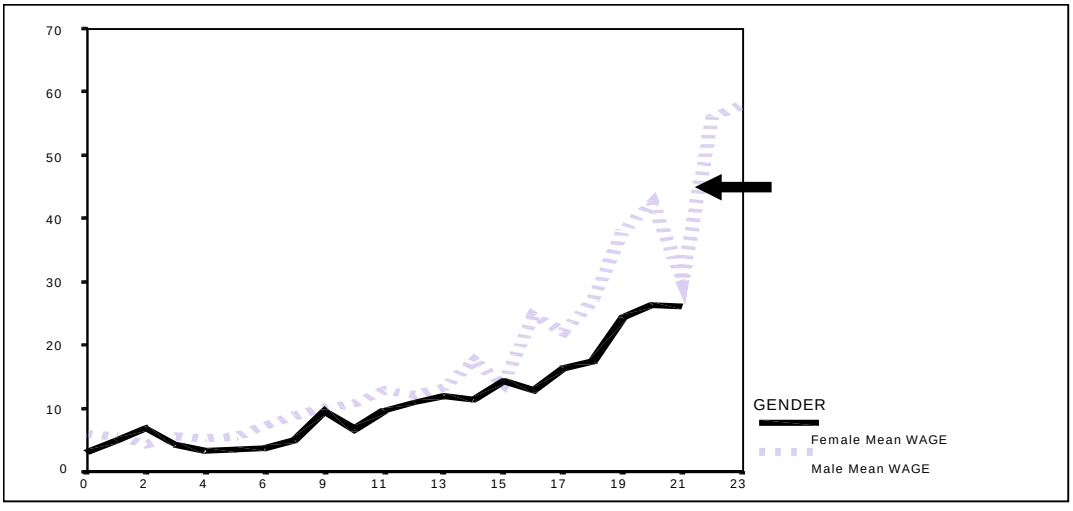


Select the style you want. Click on “Apply” or “Apply All."

You may want to change the size of the markers. A small size can be difficult to see, especially if your lines are thick and the printer does not have high resolution printing capability.

The line chart changes - the width of the line has increased and its style is now "broken line."

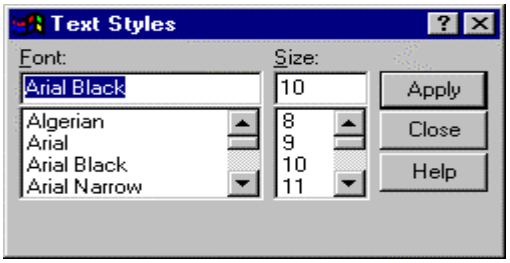




Ch 11. Section 2.j. Changing the format of the text in labels, titles, or legends

Select the text-bearing label/title/legend by clicking on it. As an example, we have clicked on the legend.

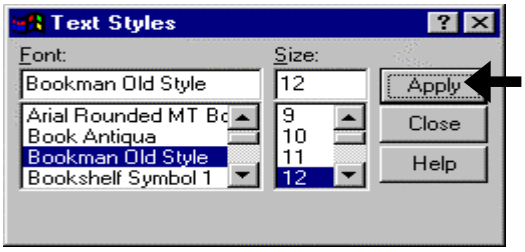
Go to FORMAT/TEXT.



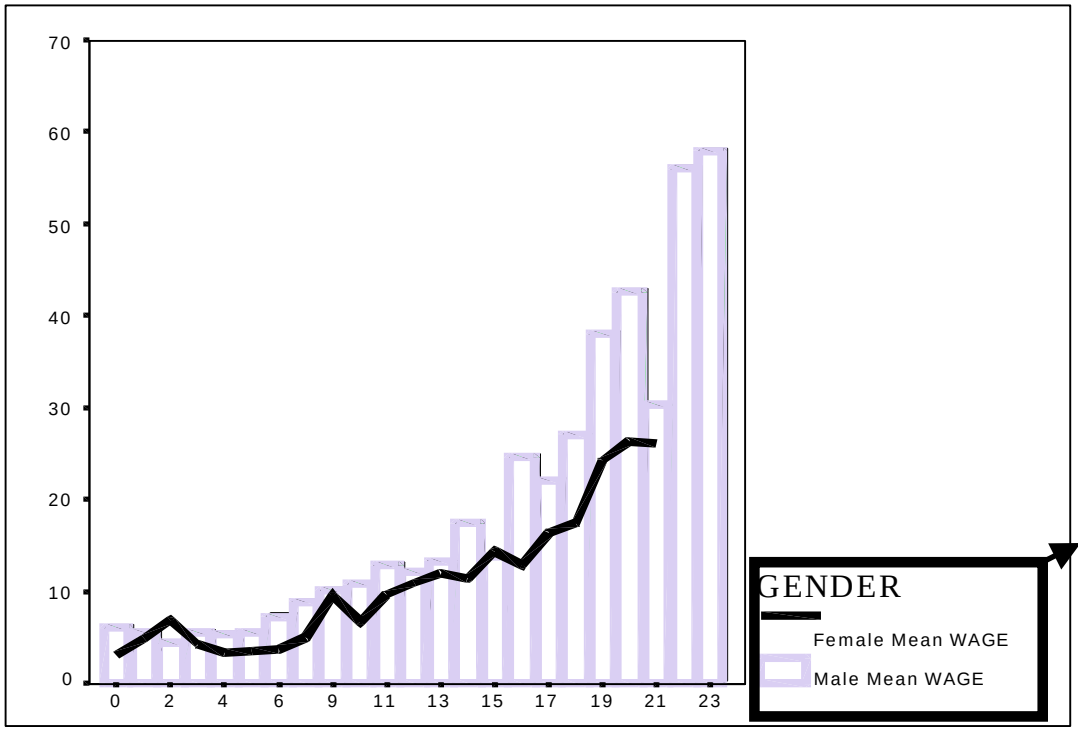
Select the font format you want - both type and size.

Click on "OK."

The font for the legend text has changed from "Arial Black, size 10" to "Bookman Old Style, size 12."



.

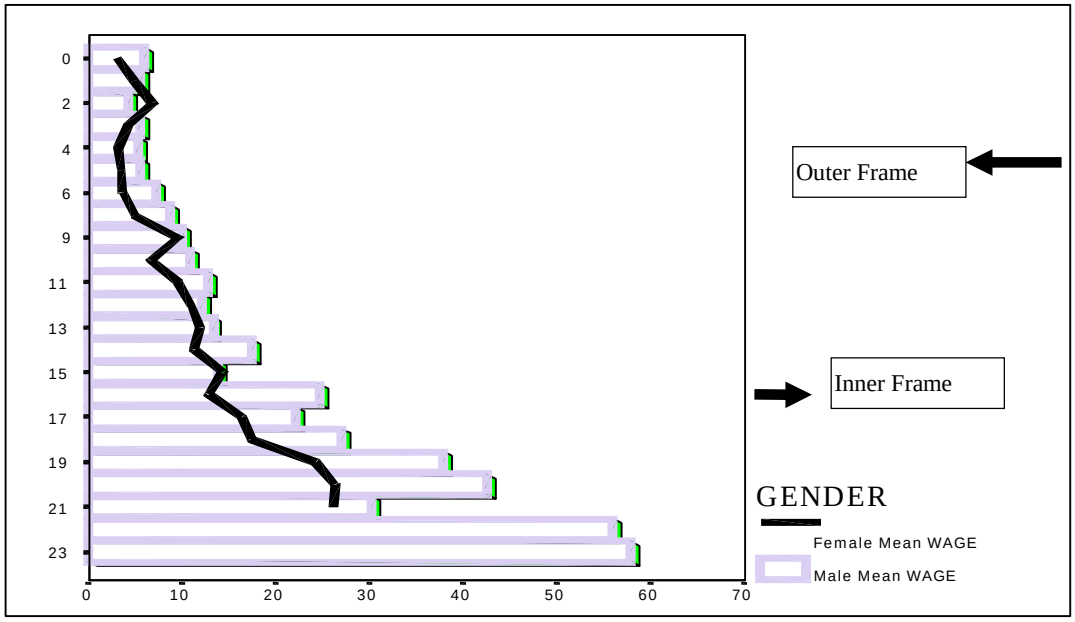


Ch 11. Section 2.k. Flipping the axes

You may want to switch the horizontal (X) and vertical (Y) axes[139].

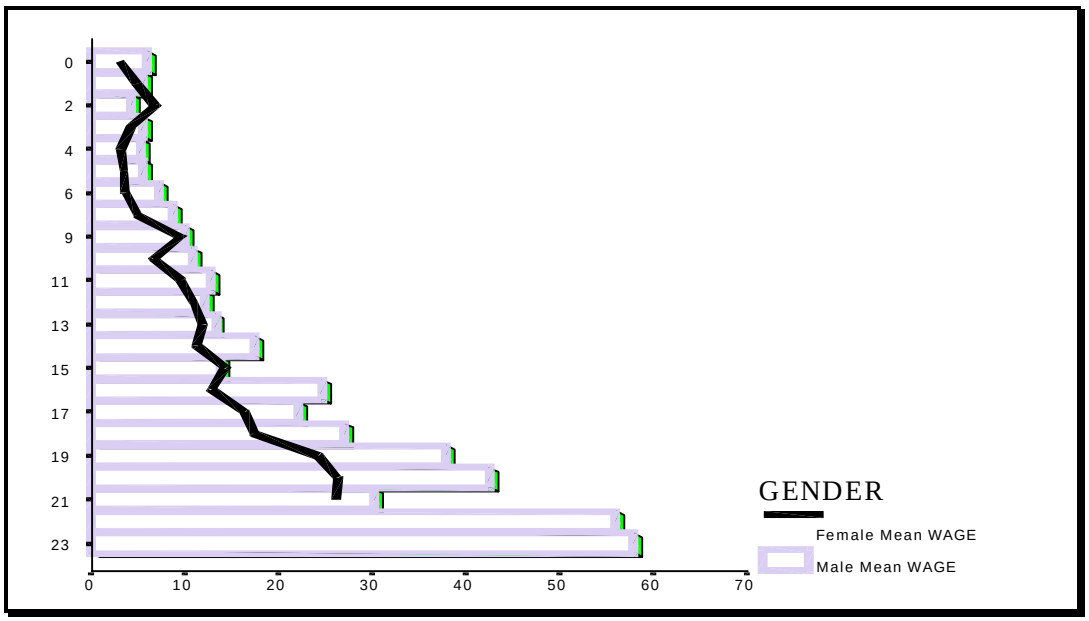
[139] If you made the chart incorrectly, or if you want to make a vertical bar chart.

To do so, go to FORMAT/SWAP AXIS. The above chart flips its axis and changes into the following chart.



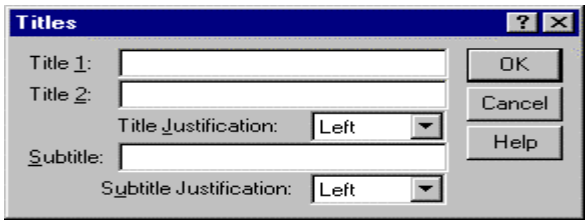
Ch 11. Section 2.l. Border and Frames

You can add or remove two types of frames - "Outer" and "Inner," as shown in the previous chart. To remove/add the inner (*outer*) frame, click on CHART/INNER (*OUTER*) FRAME. Compare this to the chart on the previous page. The inner frame is gone!



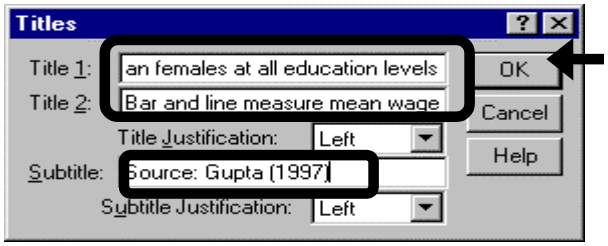
Ch 11. Section 2.m. Titles and subtitles

To add/edit titles and sub-titles, go to CHART/ TITLES.



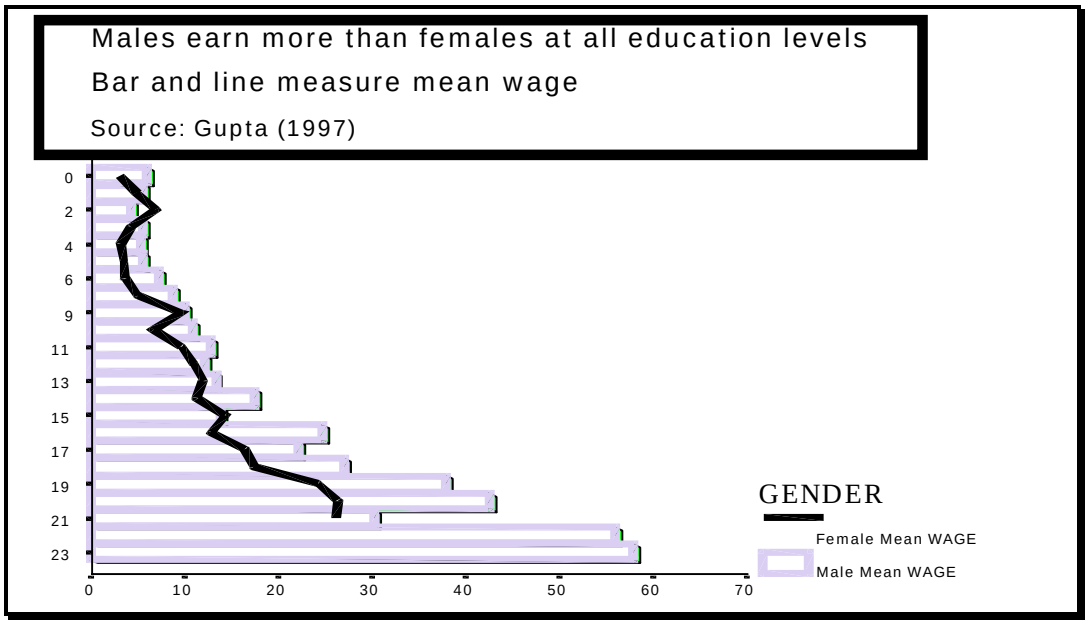
Enter the text you wish to have displayed in your title(s) and subtitle.

Click on "OK."



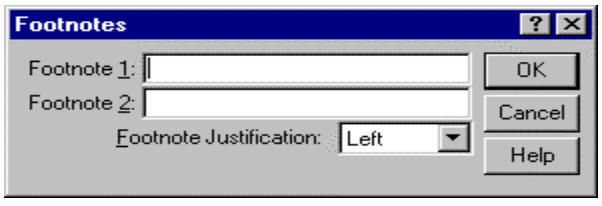
Note: Ideally, you should include the titles while making the graph. Most graph procedures have this option under the box that opens when you press the button "Titles."

The next chart has the two title lines and the subtitle line. Compare this to the previous chart that contained no titles or subtitles.

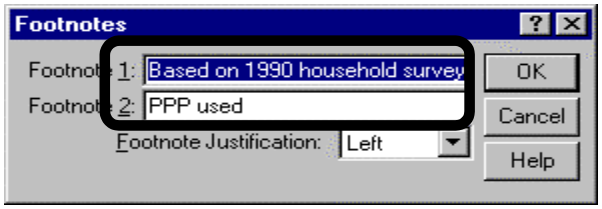


Ch 11. Section 2.n. Footnotes

To add/edit footnotes, go to CHART/FOOTNOTES.



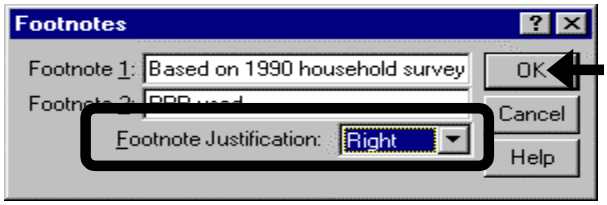
Enter the text you want in your footnote(s).



Note: Footnotes are extremely important. The information you can place in a footnote includes the data source, the name of the person who conducted the analysis, important caveats/qualifications, etc.

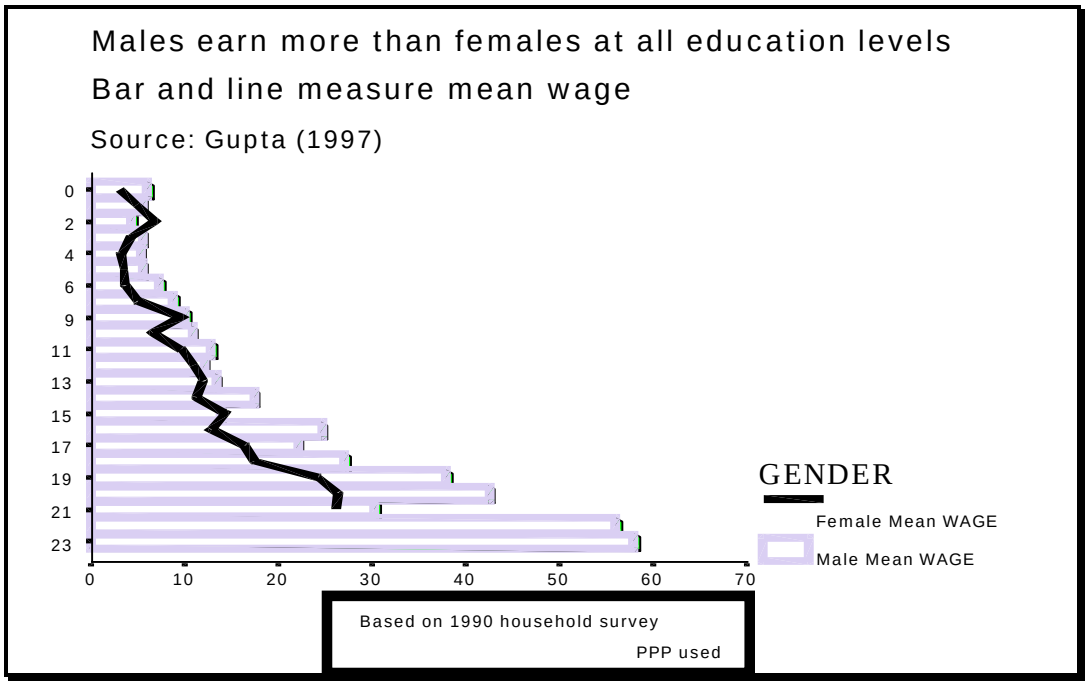
If desired, you can change the "Footnote Justification." Here we have chosen "Right" as the justification[140].

Click on "OK."



Note: Ideally, you should include the footnotes while making the graph. Most graph procedures have this option under the box that opens up when you press the button "Titles."

The two footnotes are inserted at the bottom. Because we asked for "Right-Justification," the footnotes are aligned to the right side.



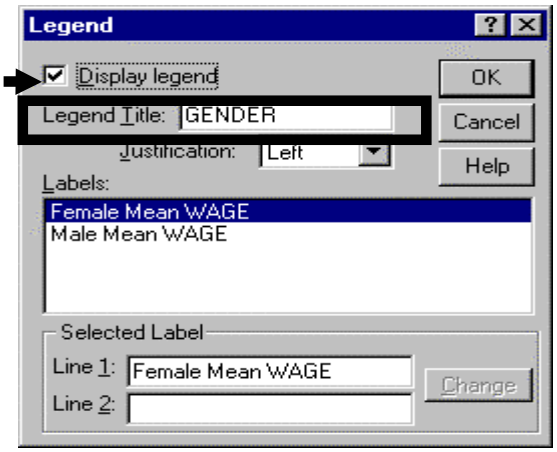
[140] "Justification" is the same as "Alignment."

Ch 11. Section 2.o. Legend entries

Go to CHART/ LEGEND.

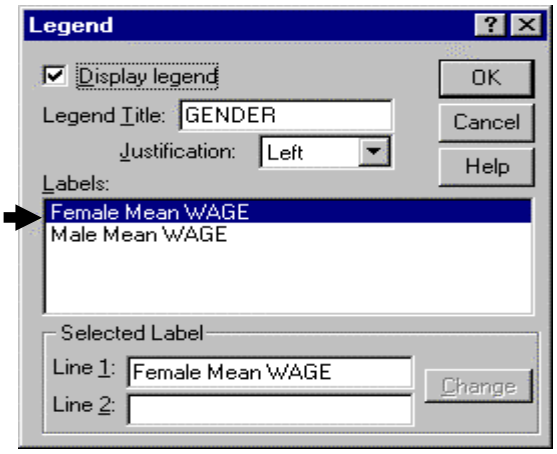
You will usually want to choose the option “Display Legend."

To change the legend title, click in the box “Legend Title” and type in the new title.

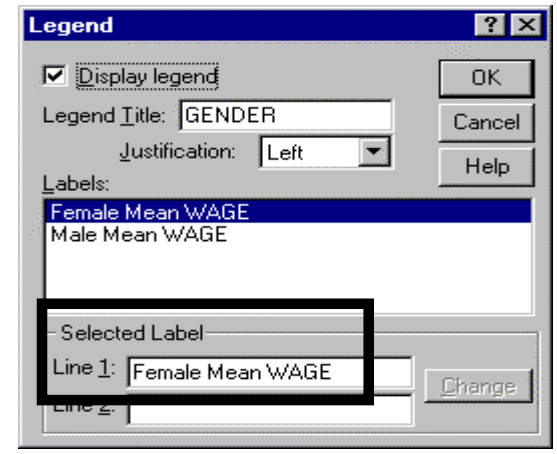


We want to change the labels. Specifically, we want to replace the all-caps "WAGE" with "Wage."

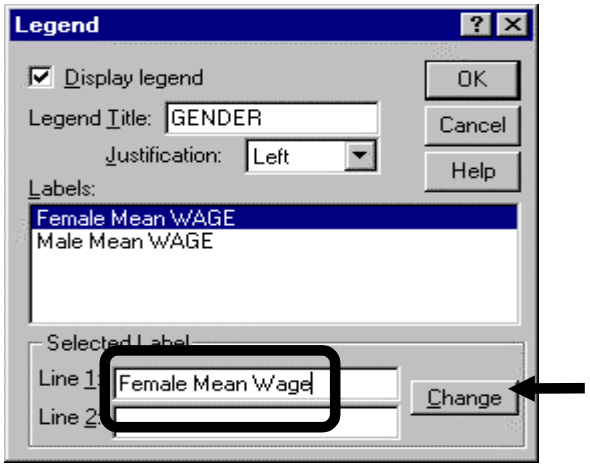
To do this for *Female Mean Wage*, click on the series name.



Change the label in the box “Selected Label."

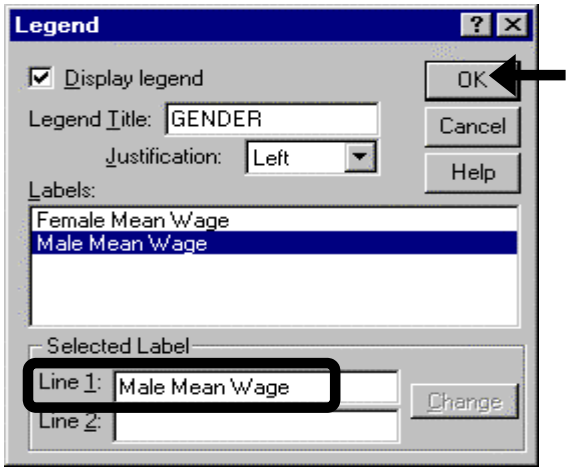


Click on the button “Change."

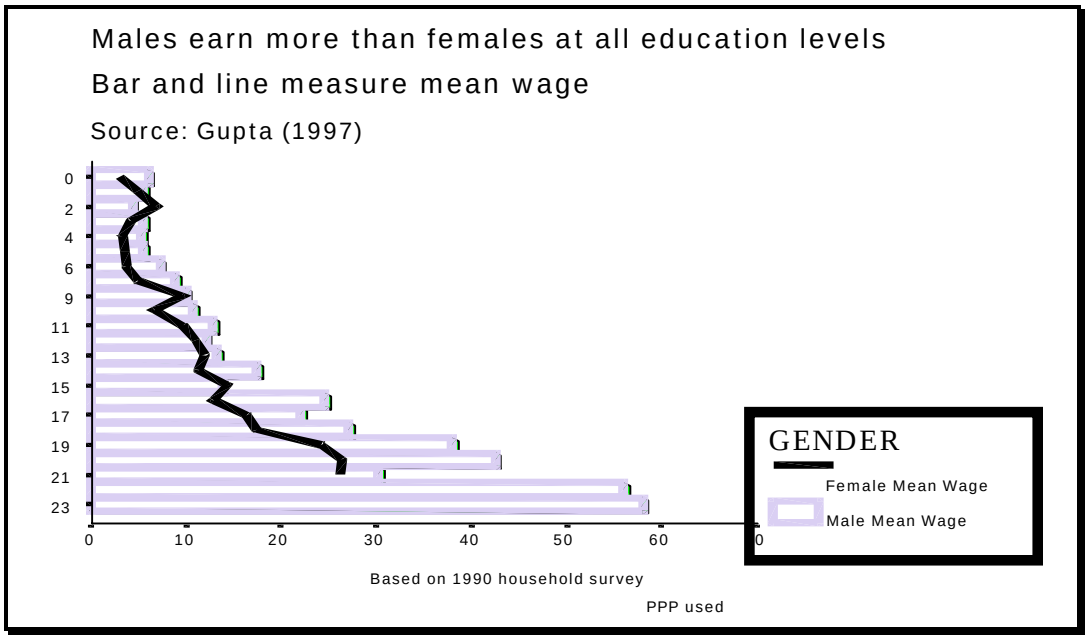


Now you must do the same for the other series.

Click on “OK."

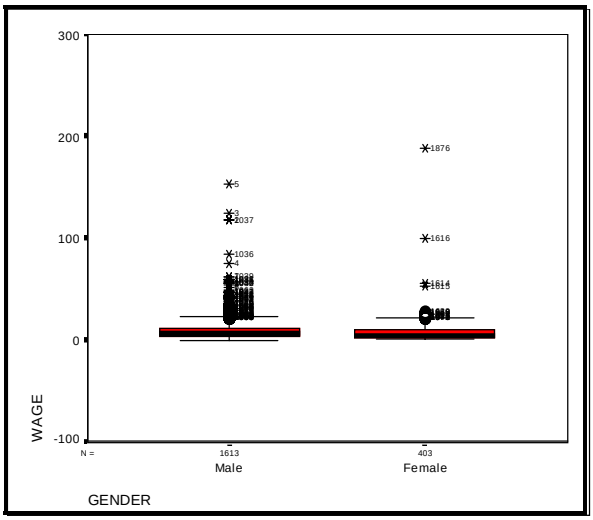


The labels of the legend entries have been changed. "WAGE" has been replaced by "wage."



Ch 11. Section 2.p. Axis formatting

A poorly formatted axis can destroy the depictive power of the chart. A simple example is that of scaling. Outliers tend to dramatically increase the displayed range of a chart, thus hiding patterns. See the boxplot below-- the quartiles cannot be seen clearly. Please experiment with different types of charts. We have found that axis formatting is very useful with error bars, boxplots and histograms.



Go to CHART/ AXIS.

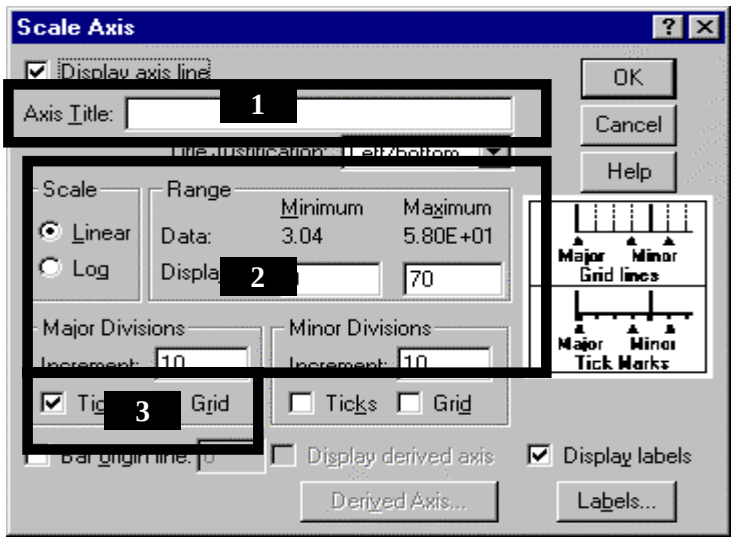
The dialog box has three important areas.

Area 1 allows you to enter/edit the name of the axis.

In Area 2, you choose the range to be displayed and the gaps between the labels on an axis (i.e. - should the axis show 0, 1, 2, 3,..., 100 or should it show 0, 5, 10, 15, ..., 100).

Area 3 permits you to choose whether to display/hide tick marks and/or gridlines.

In the next few sub-sections, we show how to format different aspects of the axis.

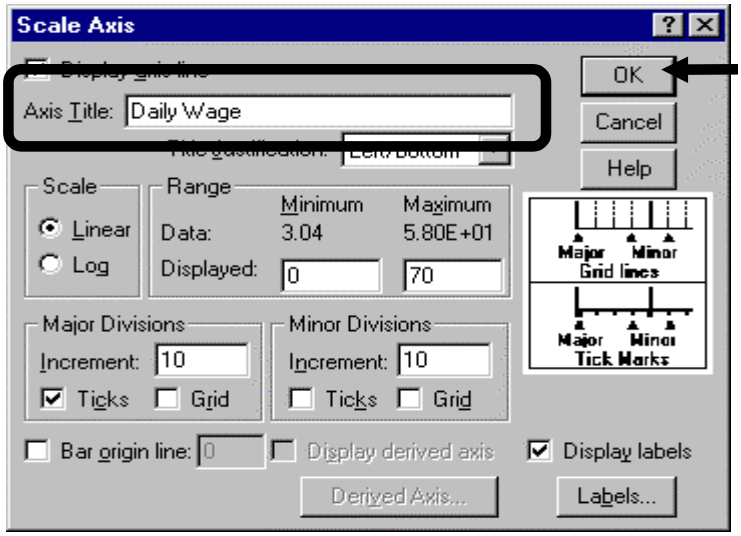


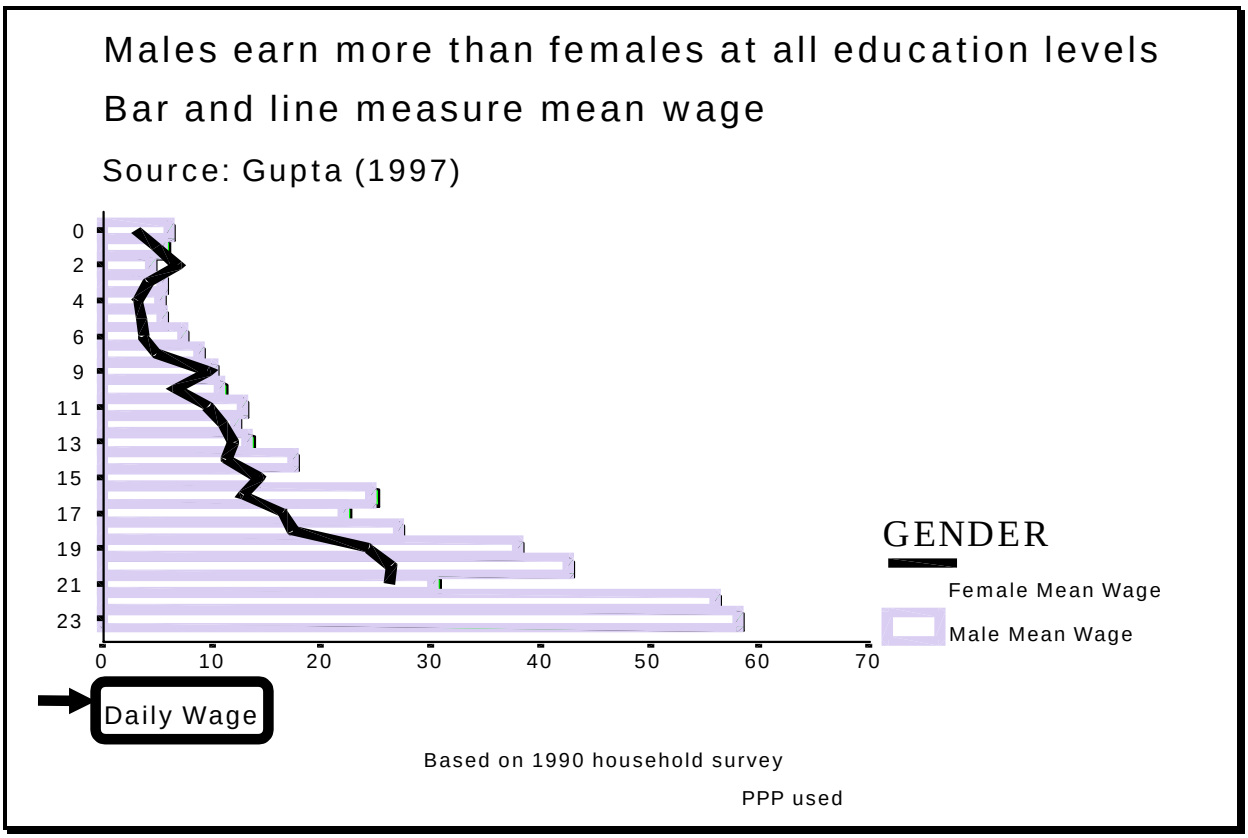
Ch 11. Section 2.q. Adding/editing axis title

Go to CHART/ AXIS.

In the box “Axis Title," enter/edit the title of the axis.

Click on “OK."



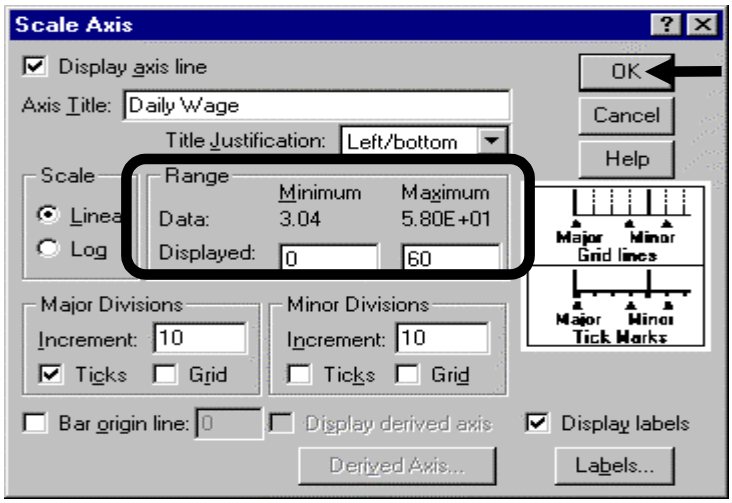


Ch 11. Section 2.r. Changing the scale of the axis

Go to CHART/AXIS.

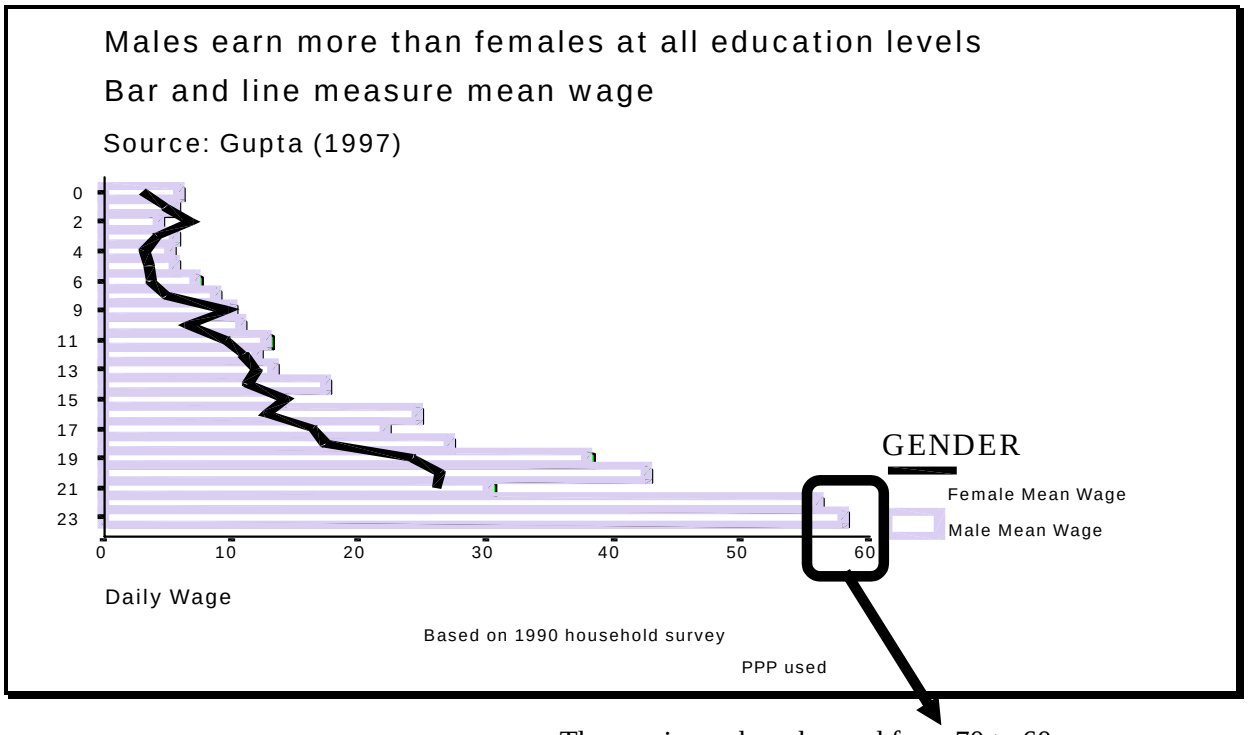
In the options area “Range," change the range as you see fit. We have changed the range from 0-70 to 0-60 by changing the “Maximum” from 70 to 60.

Click on “OK."



Note: Whenever you do this, make sure that the entry in the area “Major Divisions” is less than the gap between the numbers in the boxes "Maximum" and "Minimum."

We advise you to experiment with axis formatting. It is the most important topic in this chapter. Using the proper scale, increments, and labels for the axis is essential for accurately interpreting a graph.



The maximum has changed from 70 to 60

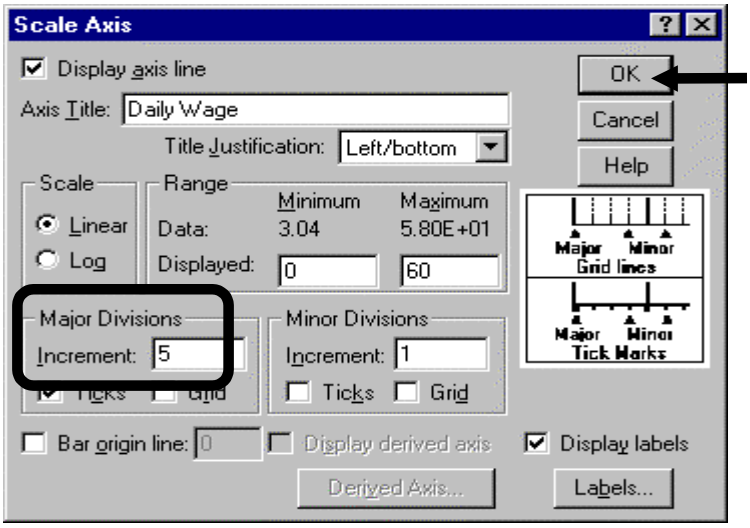
Ch 11. Section 2.s. Changing the increments in which values are displayed on an axis

Instead of seeing an indicator for every 10th dollar level, you may prefer to see indicators for every 5th dollar level.

Go to CHART/AXIS.

In the box "Major Divisions," change the number of divisions.

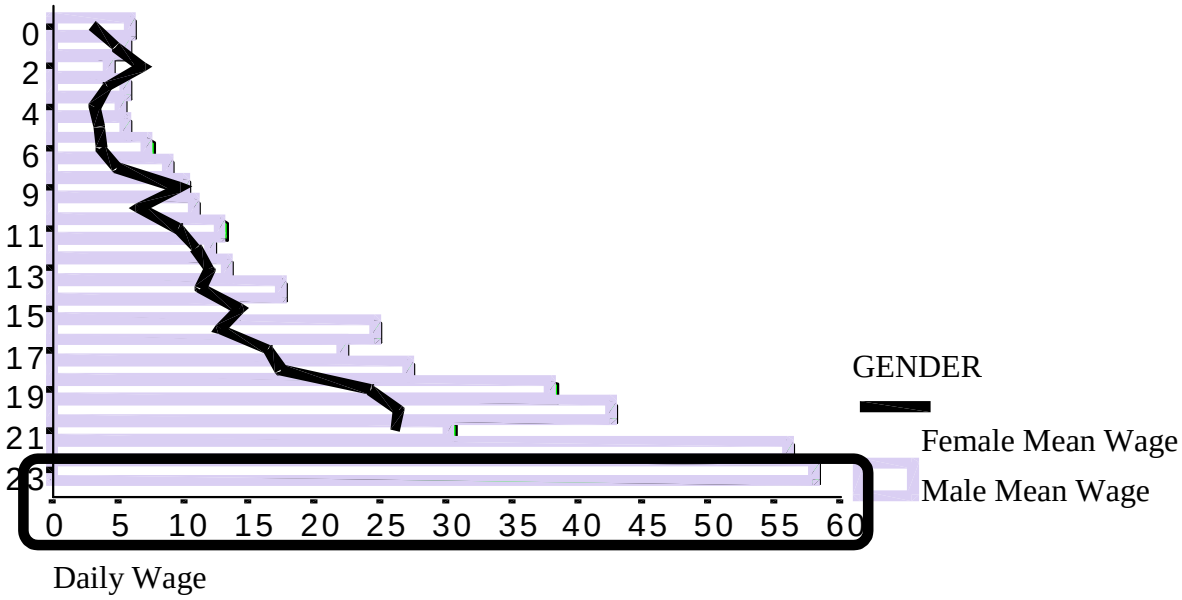
Click on "OK."



Males earn more than females at all education levels

Bar and line measure mean wage

Source: Gupta (1997)



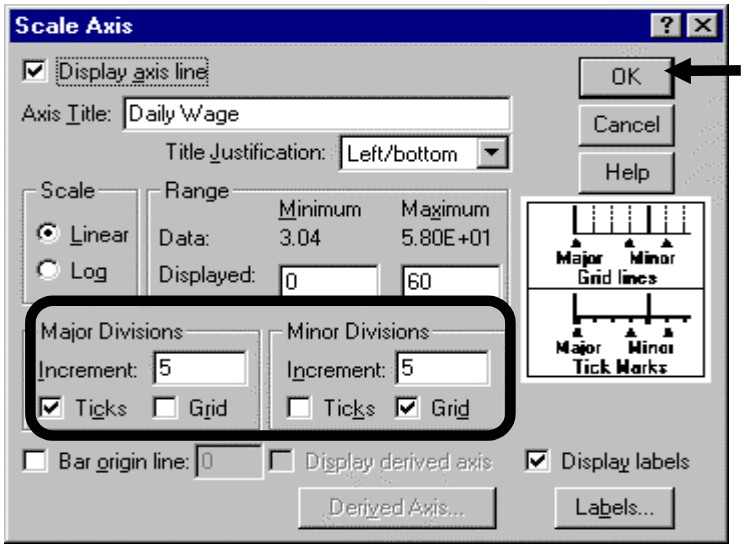
Ch 11. Section 2.t. Gridlines

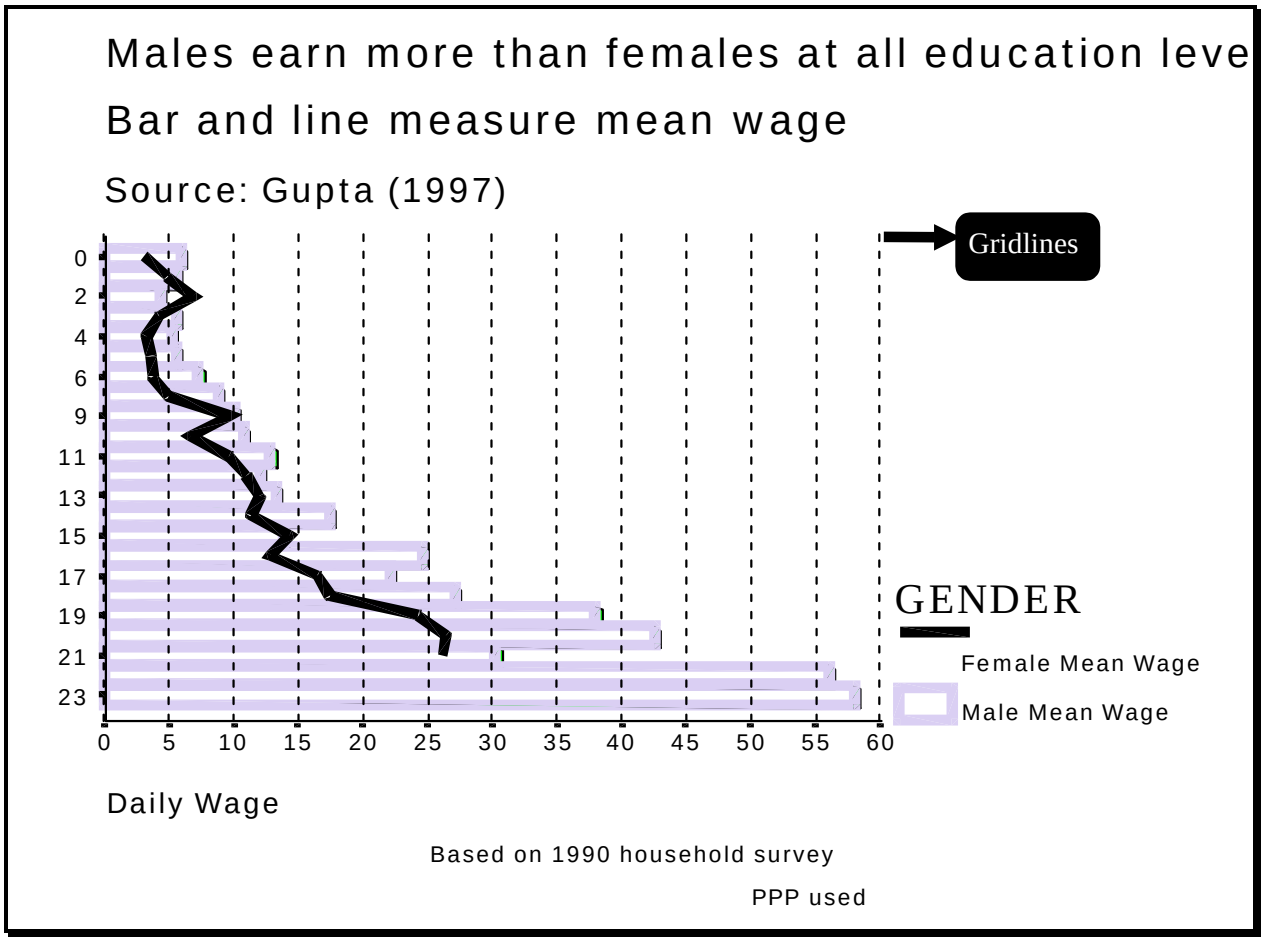
Gridlines are useful for reading values from complex charts.

Go to CHART/AXIS.

Select “Grid” in the area “Major Division.”

Click on “OK."



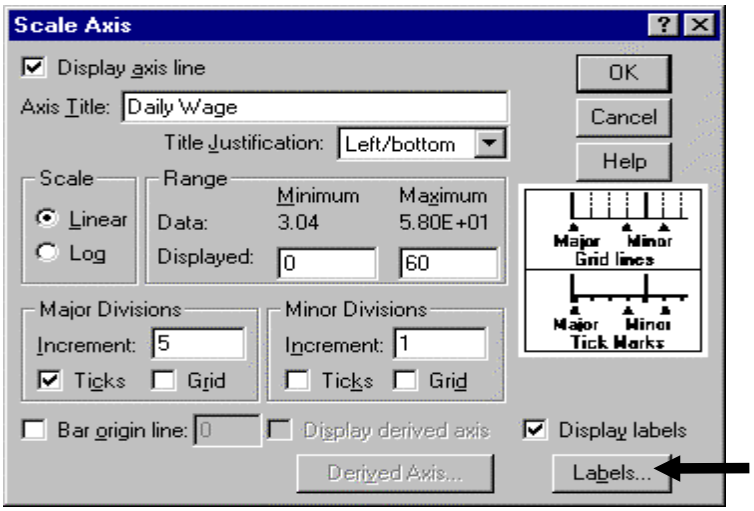


Ch 11. Section 2.u. Formatting the labels displayed on an axis

Go to CHART/AXIS.

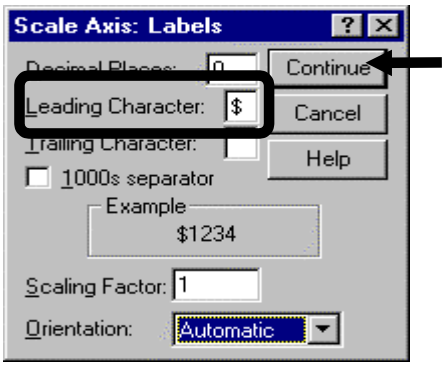
Click on the button “Labels."

Note: Labels are very helpful to the readers of your output. The labels can be used to provide information on the scales of the variables displayed, the units of measurement, etc.



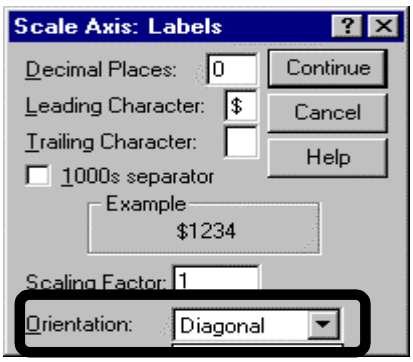
A new dialog box opens.

Make any changes you desire. We have asked for a dollar sign ($) to precede all the labels for the axis.



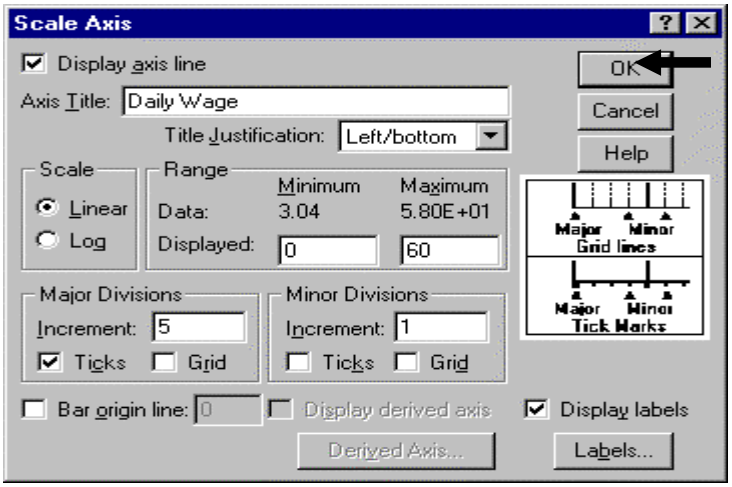
We have also asked for “Diagonal Orientation.” This is useful if you fear that the labels will be too close to one another.

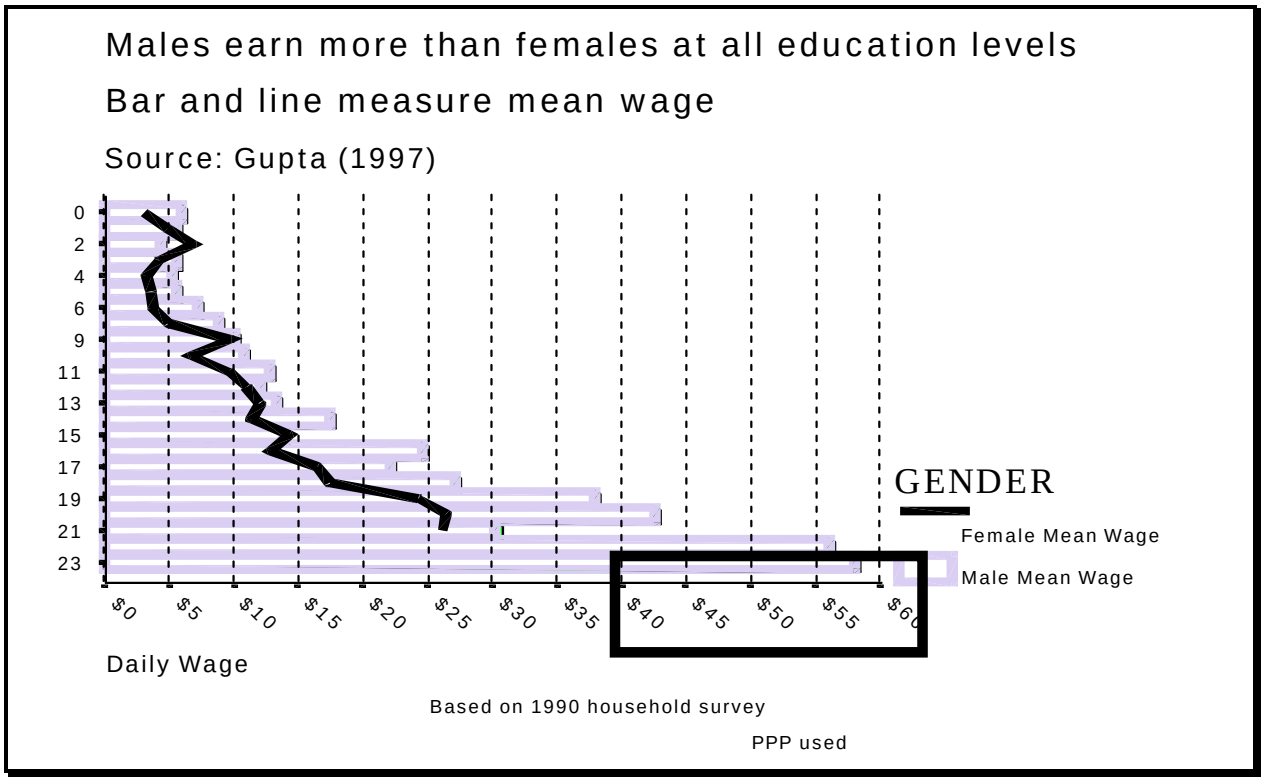
Click on “Continue.”



Click on "OK."

Once you are satisfied with the formatting, go to EDIT/COPY TABLE and open Word (or any other software you are using) and choose EDIT/PASTE. To save on storage space, you may want to choose EDIT/PASTE SPECIAL/PICTURE.





To take quizzes on topics within each chapter, go to http://www.spss.org/wwwroot/spssquiz.asp

Ch 12. READING ASCII TEXT DATA

The most frustrating stage of a research project should not be the reading of the data. The "ASCII Text" format (often called simply "ASCII" or "Text" format) can make this stage extremely frustrating and time consuming. Though we teach you how to read ASCII text data into SPSS, please keep the following issues in mind:

- If the suppliers of data can provide you with data in a simpler format (for instance, SPSS, dbase, Excel), then ask them to do so!
- The Windows point-and-click method for reading in ASCII text data is tedious and slow and can be very painful if you make mistakes or have to re-do the process by adding more variables. Using programming code is a much better method to achieve the same goal. In fact, programming is also better for defining variables (see section 1.2) and some other procedures. However, because this book is for beginners, we avoid going into the intricacies of programming techniques and code.
- An excellent option for quickly converting data from different file formats into SPSS format is through the use of data conversion software like STATTRANSFER (web site: www.stattransfer.com) or DBMSCOPY (web site: www.dbmscopy.com).

- SPSS 9.0 has an easier procedure for reading ASCII text data. We do not discuss the procedure because once you learn the procedures shown here, the procedures in SPSS 9.0 will be easy to pick up. If we get feedback on the need to show how to read ASCII text data in SPSS 9 or 10, we will place the instructions on the web site www.spss.org.

In sections 12.1.a and 12.1.b we explain what ASCII Text data is and the differences that exist among different ASCII Text data formats.

Then, in sections 12.2-12.4, we describe in detail the steps involved in reading ASCII text data.

Ch 12. Section 1 Understanding ASCII text data

Most large data sets, especially those available to the public via CD-ROM or the Web, are in ASCII Text format because of the relatively small disk space needed to store and distribute ASCII Text data and the conformity of ASCII with standardization guidelines across operating systems, hardware, etc.

As the name "Text" suggests, ASCII Text data is data written in the form of text, as in a Word or WordPerfect file. In contrast to ASCII, data are organized into rows and columns in Excel and variables and observations in SPSS.

When ASCII data are being entered, the data-entry person types the variables next to each other, separating data from different variables by a standard character[141] called a "delimiter," or by column positions. We now delve deeper into understanding the two broad types of ASCII text

[141] The standard "delimiters" are tab, comma, or space.

formats. These are "fixed field" (also called "fixed column") and "free field" (also called "delimited").

Ch 12. Section 1.a. Fixed-Field/Fixed-Column

In this format, the data are identified by position. When you obtain such data, you will need a "code book" that indicates the position each variable occupies in the file.

Assume there are three variables - ID, First Name, and Last Name. The case to be entered is "ID=9812348," "First Name = VIJAY," and "Last Name = GUPTA." The code book says that the variable "ID" is to be entered in the position 1 to 9, "first name" in 10 to 14, and "last name" in 15 to 21.

When you read the data into SPSS, you must provide the program with information on the positions of the variables. That is, reading our sample file would involve, at the minimum, the provision of the following information: "ID=1 to 9," "First Name =10 to 14," and "Last Name =15 to 21." This is shown in the next text box.

Location	*1 2 3 4 5 6 7 8 9*	*10 11 12 13 14*	*15 16 17 18 19 20 21*
Actual data	0 0 9 8 1 2 4 8	V I J A Y	G U P T A

Ch 12. Section 1.b. Delimited/Freefield

In this ASCII text format, spaces, tabs, or commas separate the data. That is, after entering the data for one of the variables, the data input person inserts a delimiter to indicate the beginning of the next variable. The next text box shows this for the three common delimiters: space, tab, and comma.

Space delimited (.prn):	00981234821 VIJAY GUPTA		
Tab delimited: (.dat):	00981234821	VIJAY	GUPTA
Comma delimited (.csv):	00981234821,VIJAY,GUPTA		

In SPSS versions 6.x- 7.5, you need not specify the type of delimiter (space, tab, or comma). In version 8.0, you are given the option of choosing the delimiter. If this line confuses you, please re-read sections 12.1.a and 12.1.b.

Ch 12. Section 2 Reading data stored in ASCII tab-delimited format

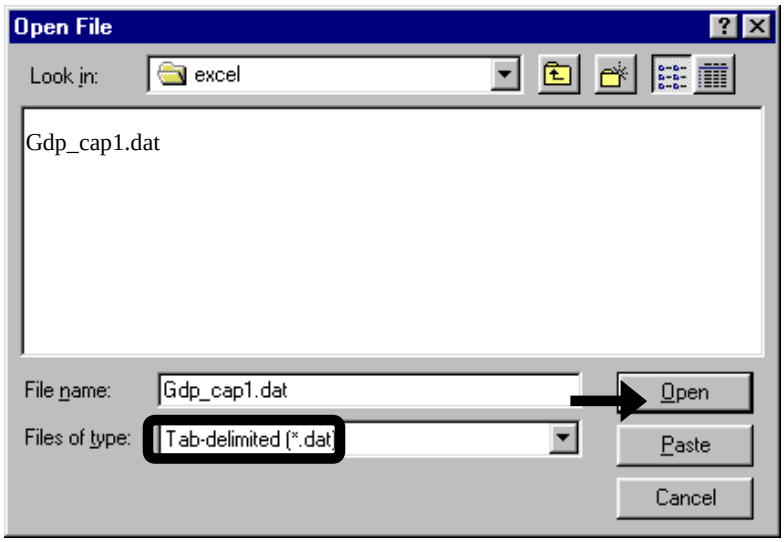
If the delimiters are tabs in SPSS versions 6.x to 7.5, then opening the file is easy[142].

Go to FILE/OPEN.

Click on the arrow next to "Files of type" and choose the type "Tab-delimited."[143]

Select the file and press "Open."

The data will be read into the data editor. Go to FILE/SAVE AS and save the data as a SPSS file with the extension (.sav).



Note: The above process may not work.

If you feel that there are problems, then use the procedure shown in section 12.3.

[142] So, request that the data provider supply the data in tab-delimited format.

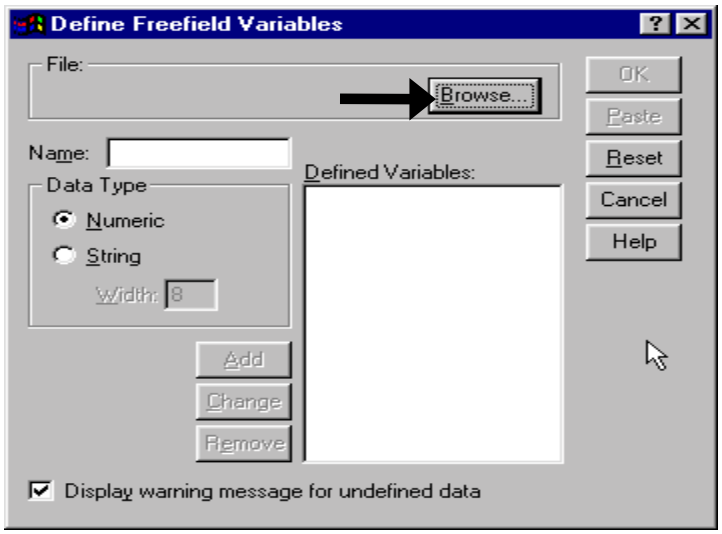
[143] ASCII file extensions can be misleading. ASCII data files typically have the extension ".dat," ".prn," ".csv," ".asc," or ".txt." But the ASCII file may have a different extension. Find out the exact name of the file and its extension from the data supplier.

Ch 12. Section 3 Reading data stored in ASCII delimited (or Freefield) format other than tab-delimited

Go to FILE/READ ASCII DATA/FREEFIELD.

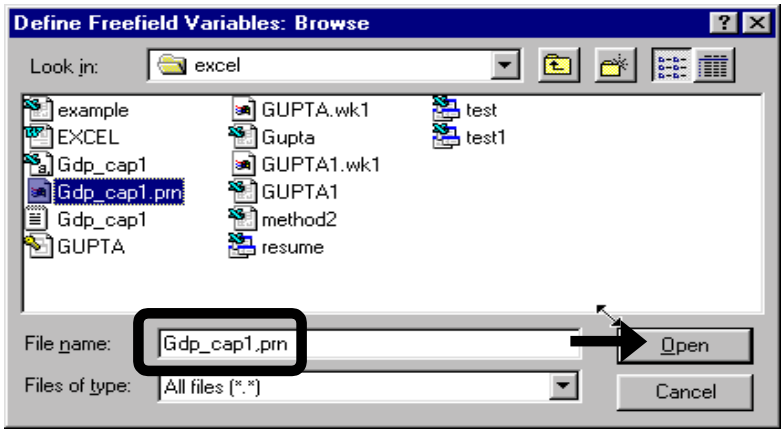
You first must specify the location and name of the file from which data is to be read. To do so, click on the button "Browse."

Note: The Browse button acts like the option FILE/OPEN, as will become clear in the next picture.



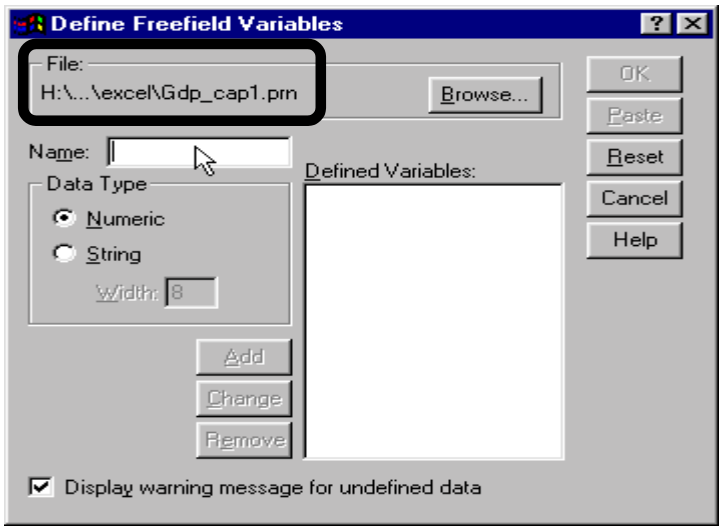
Select the correct path and choose the ASCII text file you wish to read.

Click on "Open."



SPSS now knows the location of the data file.

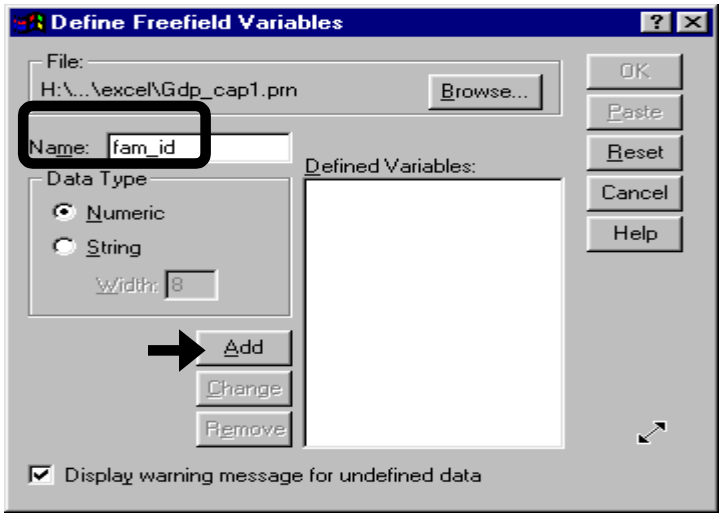
Now you must enter information on the variables you want to read from the chosen file.



To do so, click in the box "Name" and enter the name of the first variable.

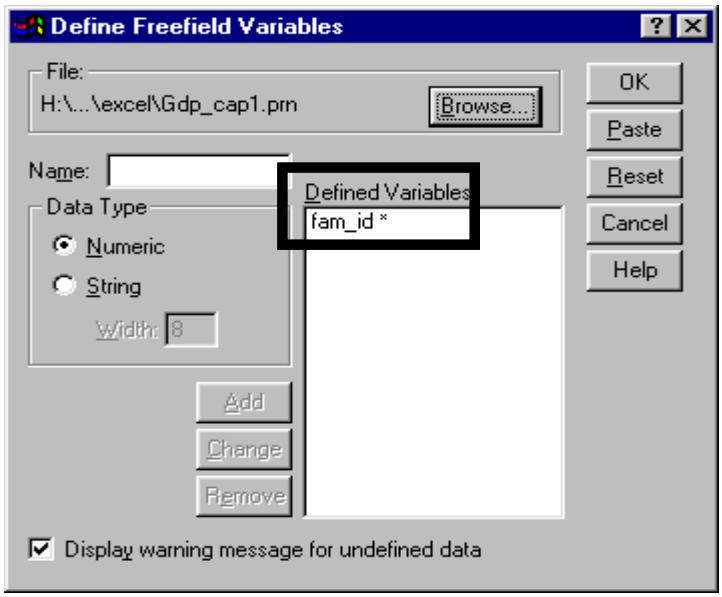
Click on the button "Add."

Note: Newer versions of SPSS may have more options in the dialog box



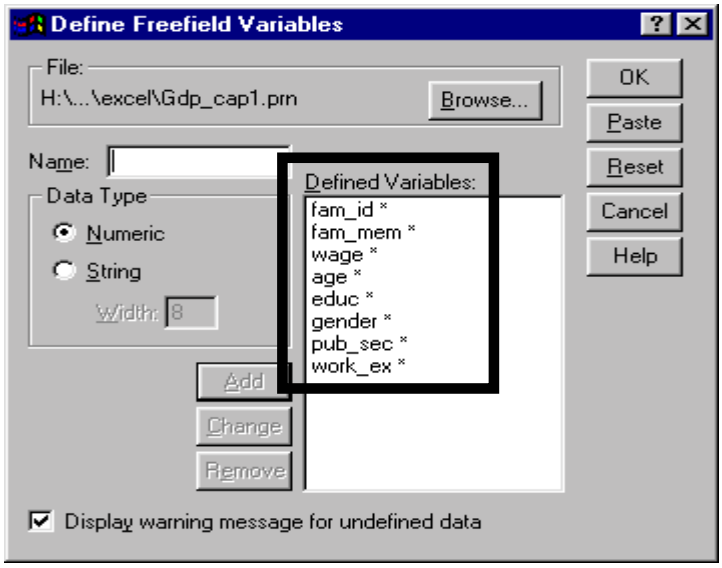
Information on the first variable is displayed in the box "Defined Variables."

You have now told SPSS to "Read in the variable *fam_id*, to read it as a numeric variable, and to read it from the file H:\ …\ excel \ Gdp_cap1.prn."



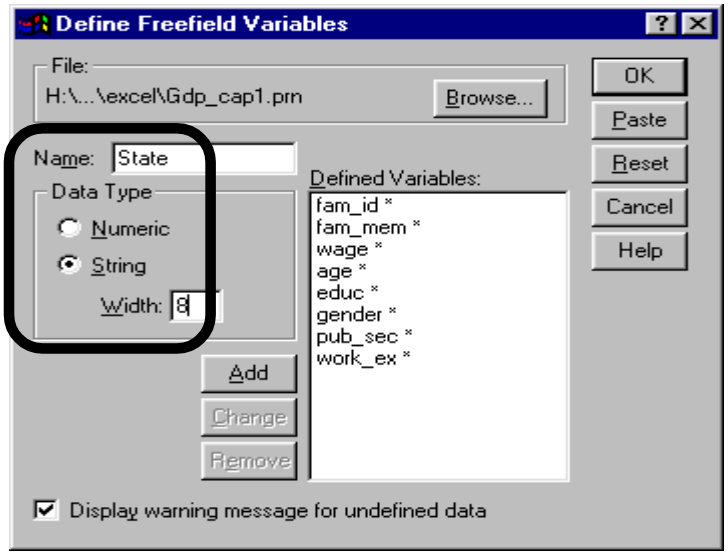
Do the same for all of the variables.

Note: The asterisk (*) after a variable name indicates that it should be read as a numeric variable. See the next page for the suffix-indicator for string/text variables.



If a variable (e.g. - *state*) is of data type "Text," then choose the option "String" in the area "Data Type" after typing in the name of the variable.

Click on "Add."

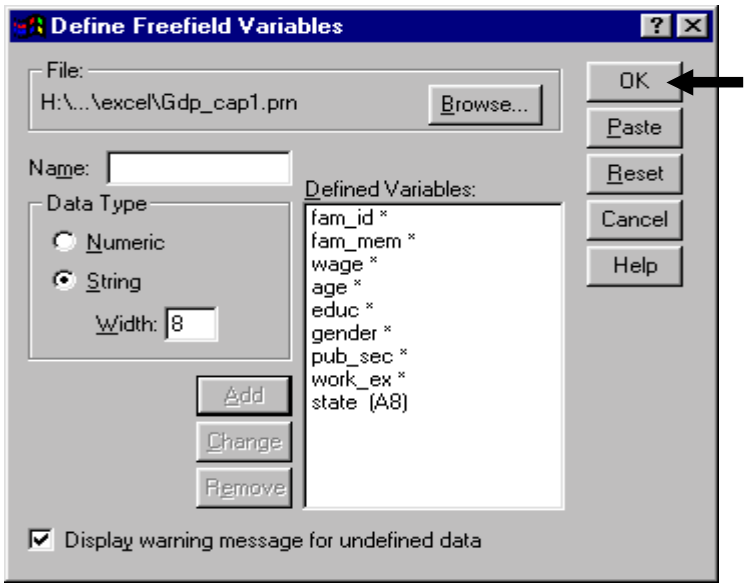


All the variables you want to read are defined and displayed in the box "Defined Variables."

Note the suffix (A8) after the variable *state*. The suffix (A8) indicates that *state* is a string/text variable.

Click on "OK."

The data will be read into the data editor. Go to FILE/SAVE AS and save the data as a SPSS file with the extension ".sav."



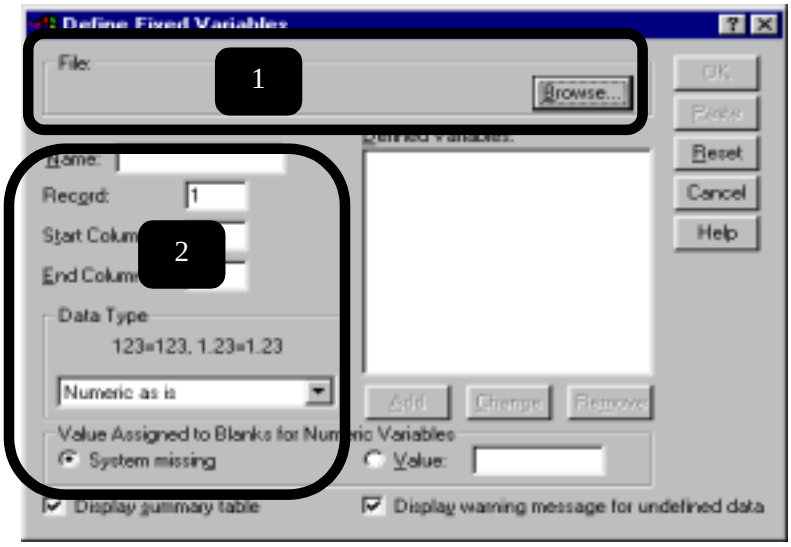
Ch 12. Section 4 Reading data stored in fixed width (or column) format

If you are obtaining data from a large data set with a large number of variables, the data will be in fixed-width text format. These files may have hundreds of variables, especially if they are supplied on a CD-ROM. You usually need only a few and you will have to tell SPSS which variables to read.

It is your job to provide a name for each variable, its data type, and the exact location (start location/column and end location/column) where this variable is stored in the ASCII text file[144] (See section 12.1.a.).

[144] Remember that with the other kind of ASCII text format (freefield or delimited), all you had to enter were variable names (and, if you choose, the data type). See section 12.3.

Go to FILE/READ ASCII DATA/FIXED.

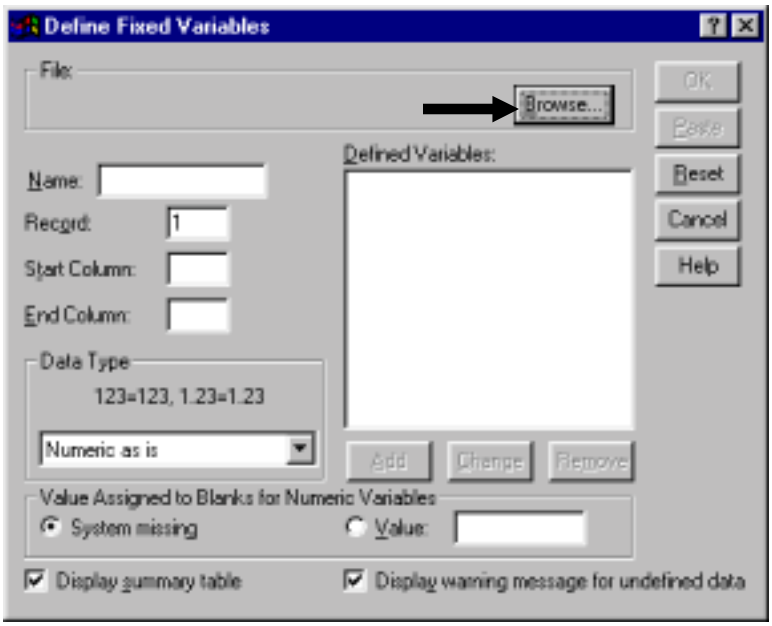


The following dialog box opens.

Area 1 is used to locate the file with the ASCII text fixed-width data.

Area 2 is where you enter the specifications of the variables you want to read, such as their names, locations, and formats. These specifications should be provided by the data supplier in a code book.

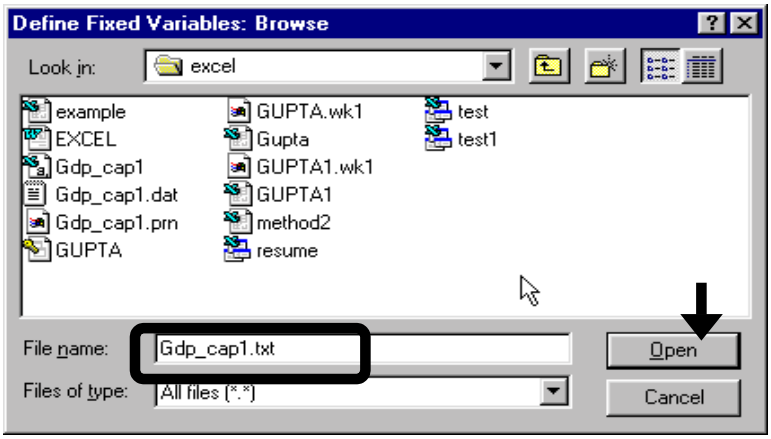
You first must tell SPSS the location of the ASCII text file.



To do so, click on the button "Browse."

Note: The Browse button acts like the option FILE/OPEN. This will become clear in the next paragraph.

Select the path and file name of the ASCII text file.



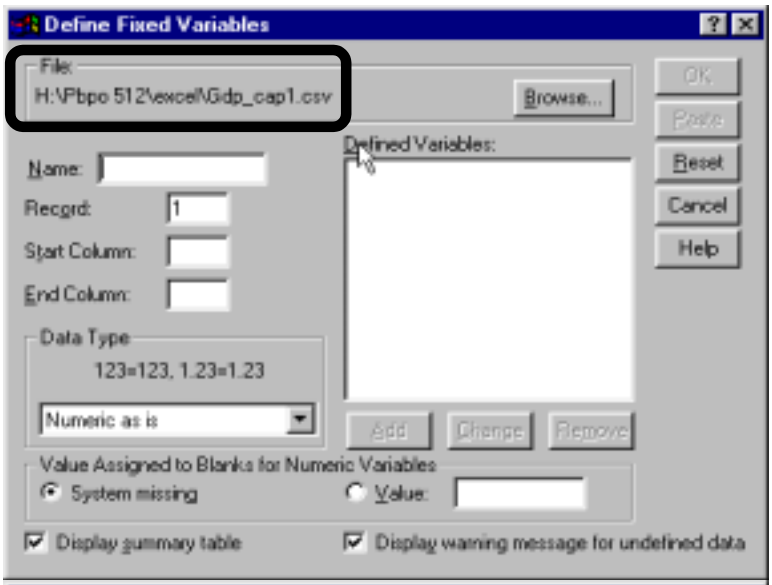
Click on "Open."

Reminder: ASCII file extensions can be misleading. Usually ASCII data files have the extension ".dat," ".prn," ".csv," ".asc," or ".txt." But the ASCII file may have a different extension. Find out the exact name of the file and its extension from the data

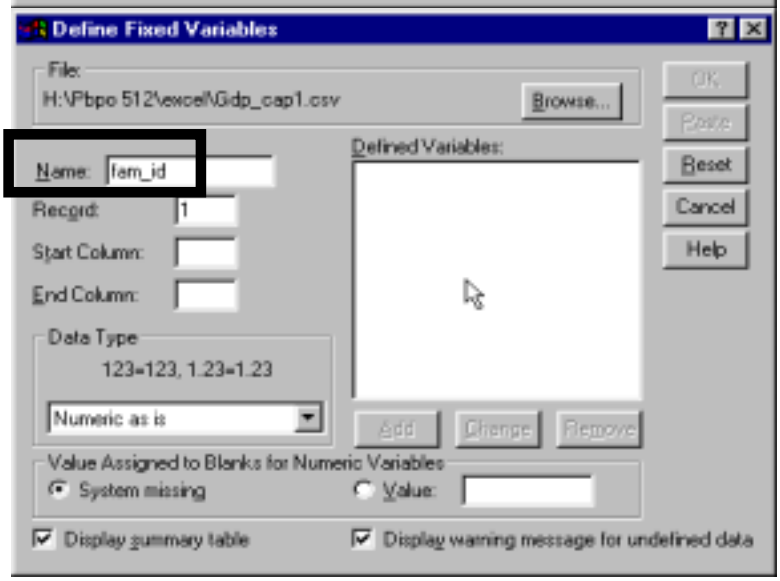
supplier.

Now SPSS knows the file that contains the data.

The next step is identifying the variables, their locations, and their formats.



Click in the box "Name" and enter the name for the first variable *fam_id*. **The code book supplied with the ASCII text file will include this information**[145].



Note: The code book should have the following information for all the variables:

- File name, extension, and location
- The name of each variable
- The data type of each variable
- The location of each variable. This may be presented as either, (a) "start" and "end" column or (b) "start" column and length. In the latter case, obtain the end column by using the formula:

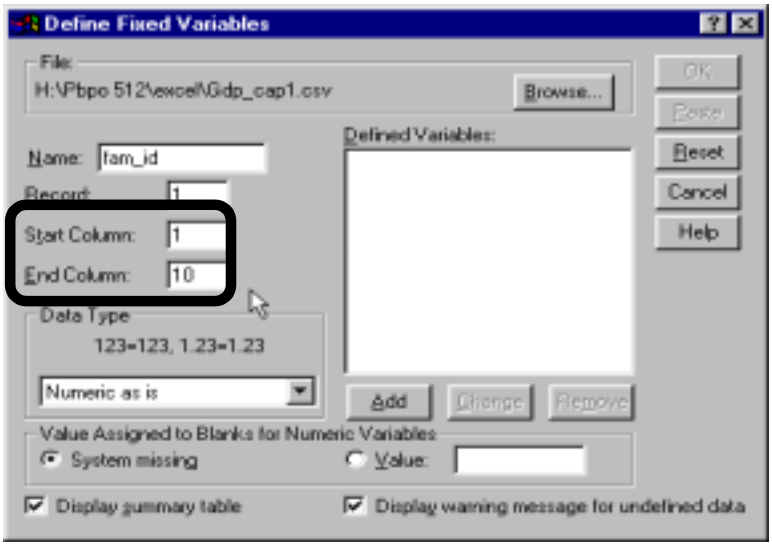
End column = (start column) + length - 1

[145] The code book may be another file on the CD-ROM that contains the ASCII text file.

In the boxes, "Start Column" and "End Column," enter the location of the variable *fam_id* in the ASCII text file.

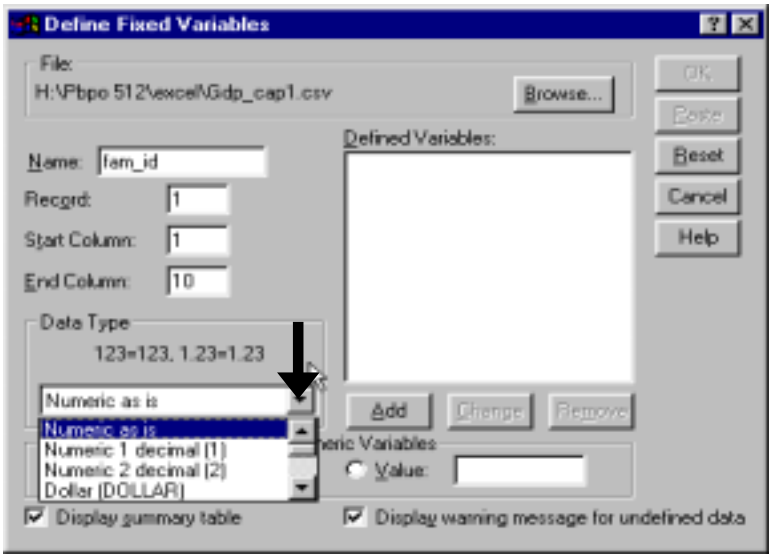
The variable is ten columns long, from column 1 to column 10.

Note: Refer to section 12.1.a to review the intuitive meaning of the "start" and "end" column TERMS.



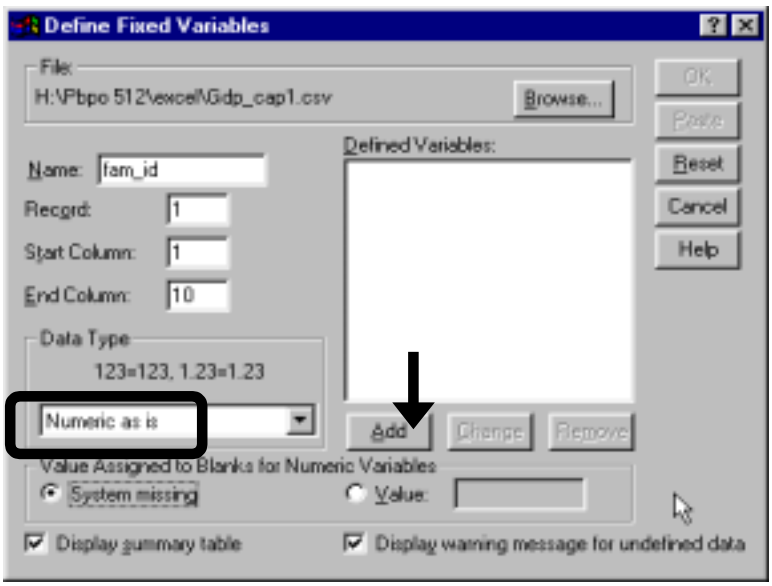
Click on the arrow next to "Data Type." Now you must choose the data type of the variable *fam_id*.

Note: This is not an essential step unless the variable is of text data type because the default data type is "Numeric as is."

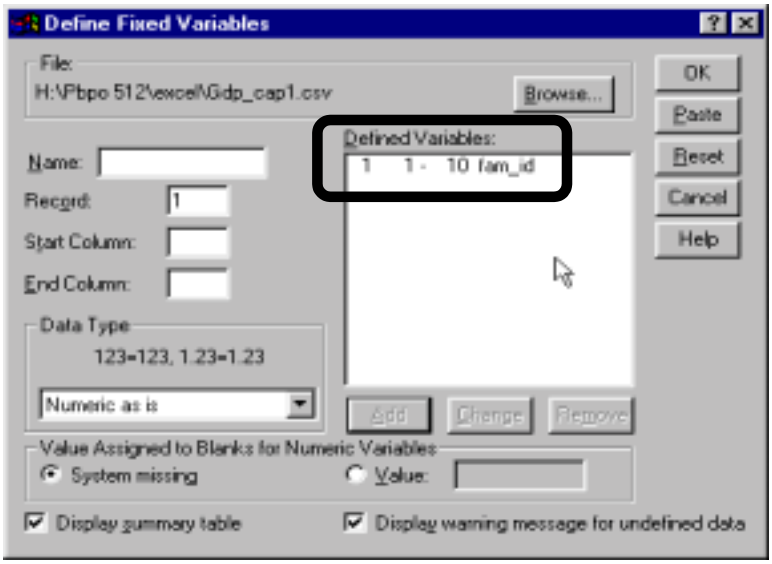


Select the data type "Numeric as is." This format works well with all numeric data.

Click on the button "Add."

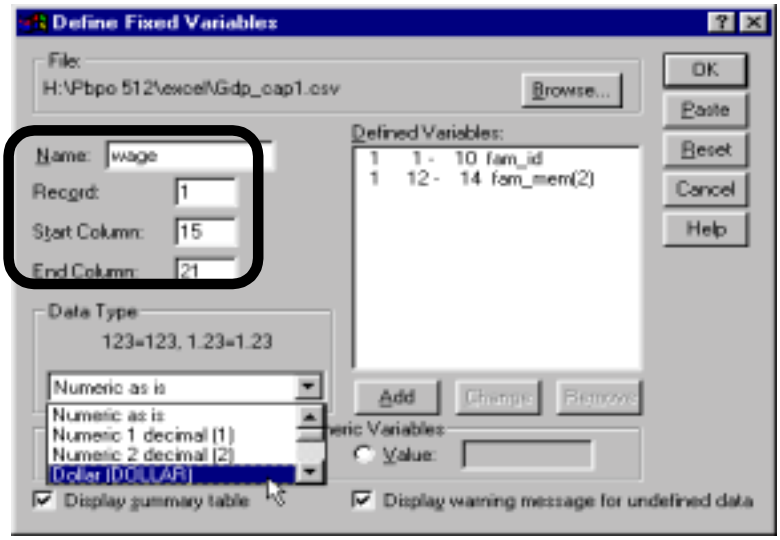


The attributes required to read the variable *fam_id* are displayed in the box "Defined Variables."

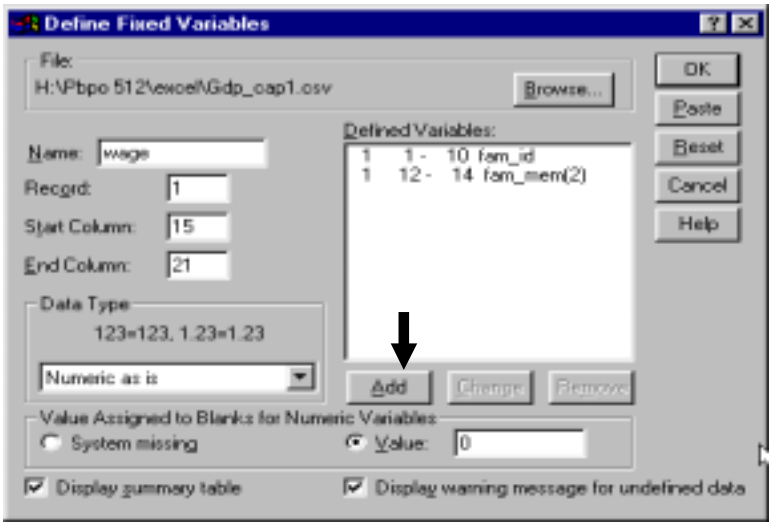


Enter the variable name *wage* and its location in the ASCII file[146].

Press the button "Data Type" and choose the type "Numeric as is."

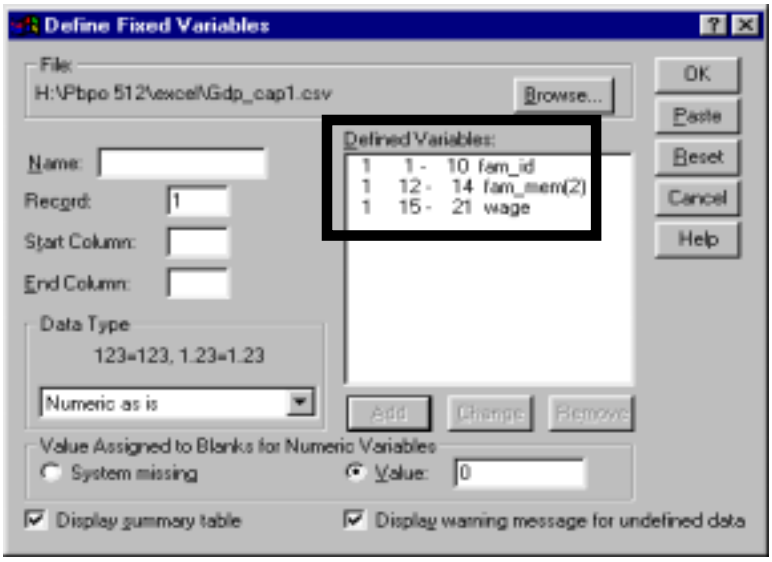


Click on the button "Add."



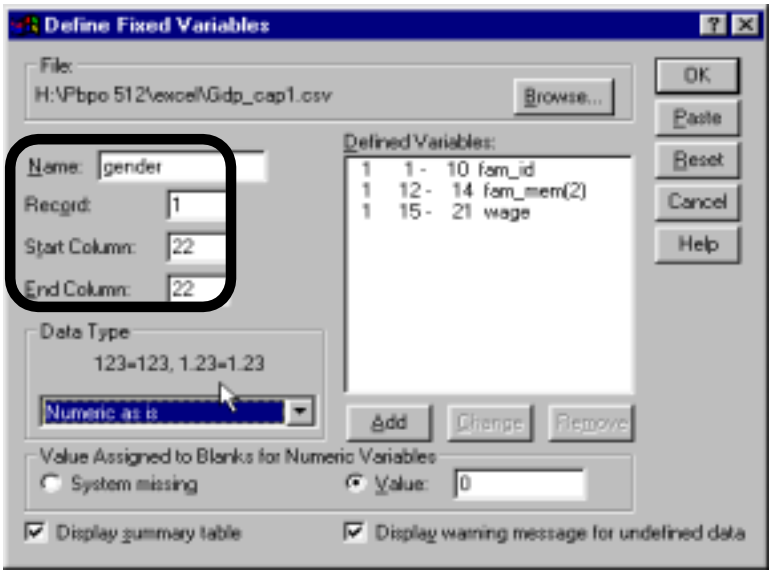
[146] We have skipped the steps for *fam_mem*. These steps are the same as for the other variables.

The first three variables have been defined.

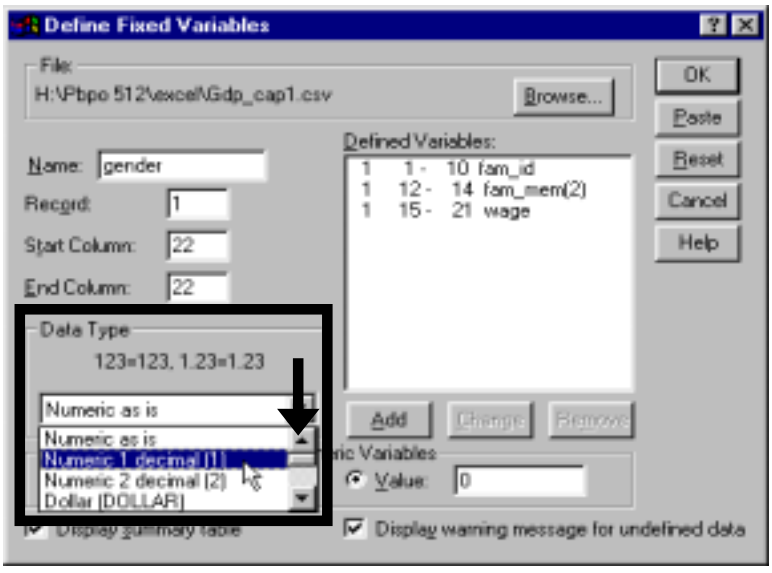


Let us demonstrate how to use a different data type.

Type the variable's name (*gender)* and its location.



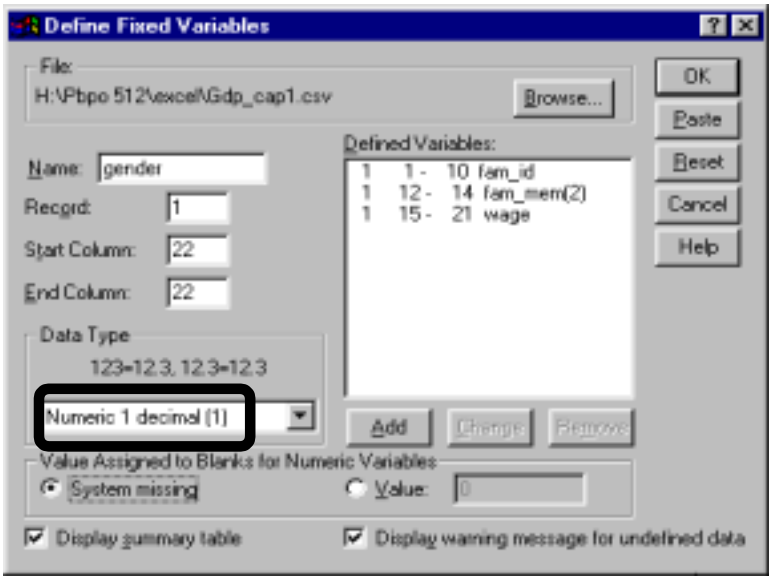
Click on the arrow next to "Data Type."



Note: Do not take lightly the role of defining the data type. If you use the incorrect data type, then several problems will arise. You will be forced to re-do the work of reading in the data, and if you did not realize that you had made a mistake in defining the data type, you may make horrendous mistakes in your statistical project. SPSS will probably not inform you of the mistakes.

Select the data type "Numeric decimal (1)." Such a specification is appropriate for *gender* because it is a dummy variable and can take only two single-digit numbers - 0 or 1.

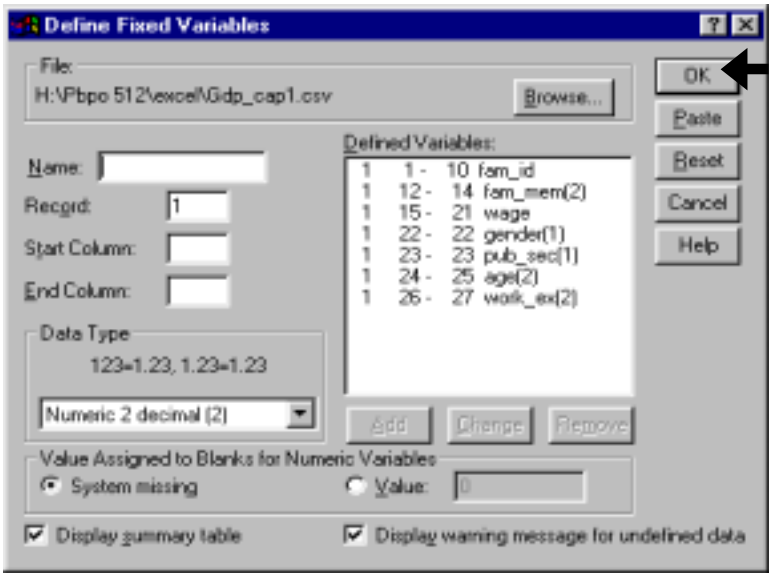
Click on "Add."



Similarly, define all the other variables you wish to read.

Press "OK." The data will be read.

The data will be read into the data editor. Go to FILE/SAVE AS and save the data as a SPSS file with the extension ".sav."



Note: Compare the mapping you have defined above (within the box "Defined Variables") with the map in the code book. If it does not match, then click on the incorrect entry in the dialog box, click on the button "change," and make the required changes.

To take quizzes on topics within each chapter, go to http://www.spss.org/wwwroot/spssquiz.asp

Ch 13. MERGING: ADDING CASES & VARIABLES

Merging two files is a difficult process and one that is prone to error. These errors can severely affect your analysis, with an inaccurate merge possibly creating a data set that is radically different from that which your project requires.

There are two types of merges:

1. Adding more observations - rarely used. This is useful only when data for a new year or cross-section becomes available on a sample that is being followed over time (panel data). See section 13.1.
2. Adding new variables- **used often**. Let's assume you have been working on earnings data from a Department of Labor survey. The social security number uniquely identifies each respondent. You have also obtained an excellent survey by the department of education, again with the social security number as the identifying variable. This survey has information on schooling history that would really enhance your earnings survey data. You can use merging, with the social security number as the key variable, to add the schooling variables to the data from the earnings survey. See section 13.2. *Don't worry if the example confuses you. Its relevance will become apparent as you go through section 13.2.*

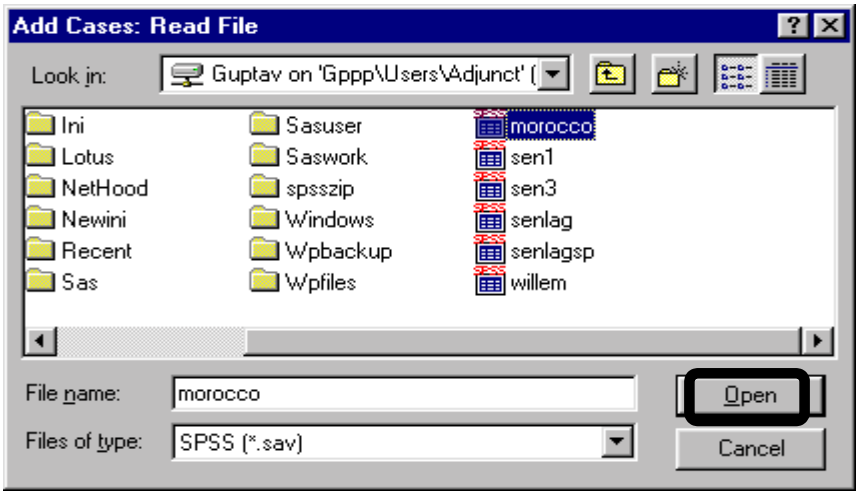
Ch 13. Section 1 Adding new observations

Assume your original data file includes data from a survey of three counties. Now, survey data from a fourth county (Morocco) is available. You want to expand the original file by adding data from the new survey. However, you do not want to add/change the number of variables in the existing data file. Essentially, you are appending new observations to the existing variables in the file that does not have data on Morocco.

Go to DATA/MERGE/ADD CASES.

Select the file with the new data to be added.

Click on "Open."



The new dialog box contains information on all the variables in the original data file [marked with the suffix (*)] and the one from which you want to add data [marked with the suffix (+)].

Variables that have the same name in both files are automatically matched and placed in the box "Variables in New Working Data File."

Other variable names may not correspond. For example, *gender* in the working data file (*) corresponds to *sex* in the file from which you want to add data (+). You will have to tell SPSS that *gender* and *sex* are a pair.

Before doing so, you might want to change the name of the variable *sex* to *gender*.

Click on the variable name *sex*.

Click on the button "Rename."

Rename the variable.

Click on "Continue."

Now we must make pairs. Click on *gender*. Then press the control key and, keeping it pressed, click on *sex*.

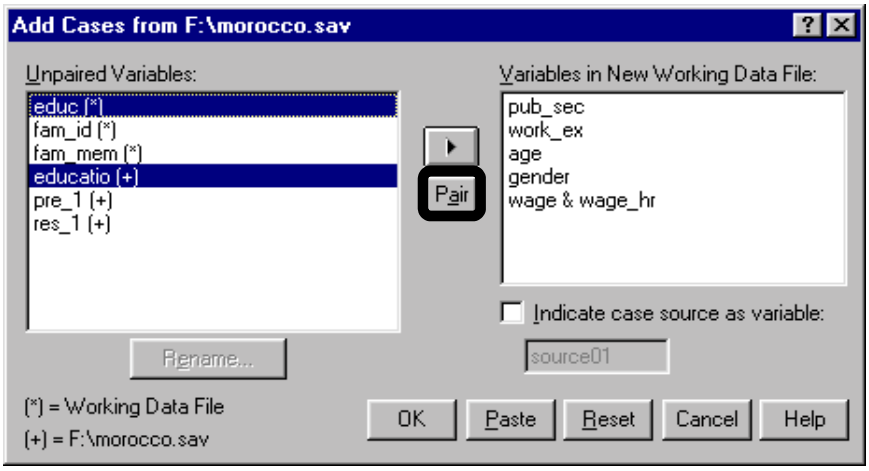
Click on the button "Pair."

This moves the pair into the box on the right side, "Variables in New Working Data File."

Similarly, choose the next pair, *wage* and *wage_hr*. Click on the button "Pair." This moves the pair into the box on the right side "Variables in New Working Data File."

Select the next pair, *educ* and *educatio*.

Click on the button “Pair.”



Click on “OK.”

Data from the chosen variables (those which are paired and are in the box "Variables in the New Working Data Set”) will be added from the file "morrocco.sav" to the working file.

The working file will have all its original data plus the new data. It will not contain any of the unpaired variables.

Ch 13. Section 2 Adding new variables (merging)

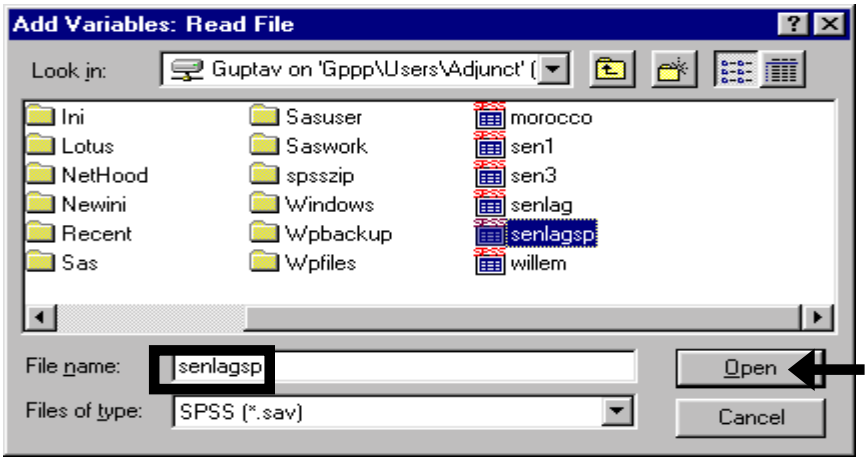
A more complex and difficult process is the addition of new variables from one data set to another. Let's assume you have data from a national **Labor** survey on only eight variables. The Department of **Education** informs you that they have information on several other variables based on a national **Education** survey. In both surveys, an individual is identified by his or her unique social security number (a national ID number in the U.S.). You want to include these new variables in your data set with an assurance that you are not matching incorrect observations across the variables. Basically, you want to make sure that the education variables you add are added to the same respondent's row or case in the Labor survey data.

For this you will require a concordance key variable pair to match observations across the files. In our example, the variable *fam_id* is to be matched with the variable *lnfp* for the merge. Those were the names given to the social security number by the survey authorities.

Go to DATA/MERGE/ADD VARIABLES.

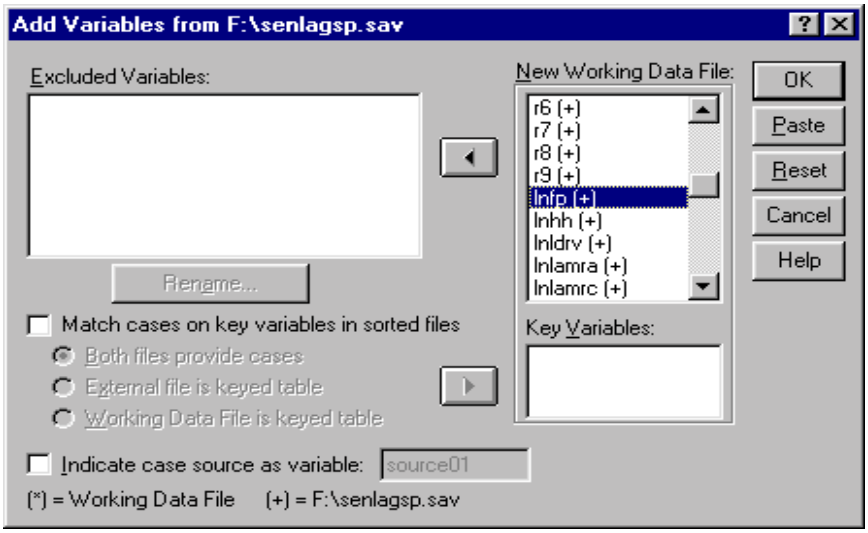
Select the file with the variables to be added.

Click on "Open."



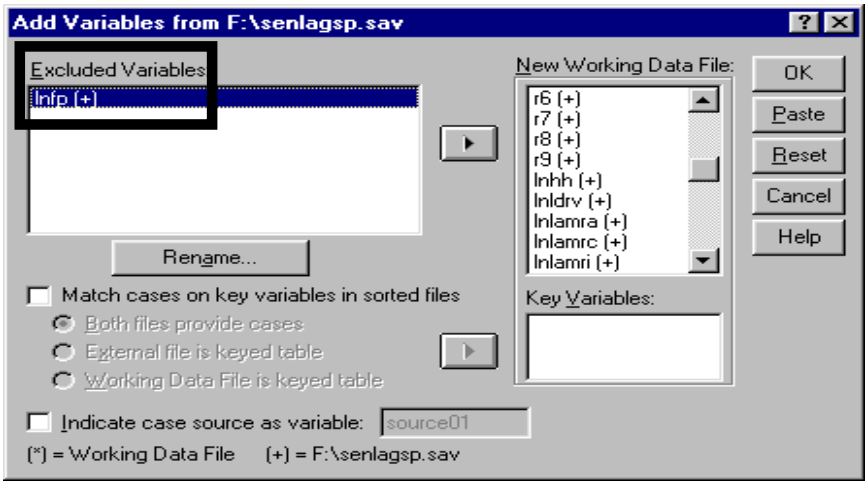
The suffix (*) indicates that a variable is from the original data file, the file to which we are adding new variables.

The suffix (+) indicates that the variable is from the data file from which we are taking variables.



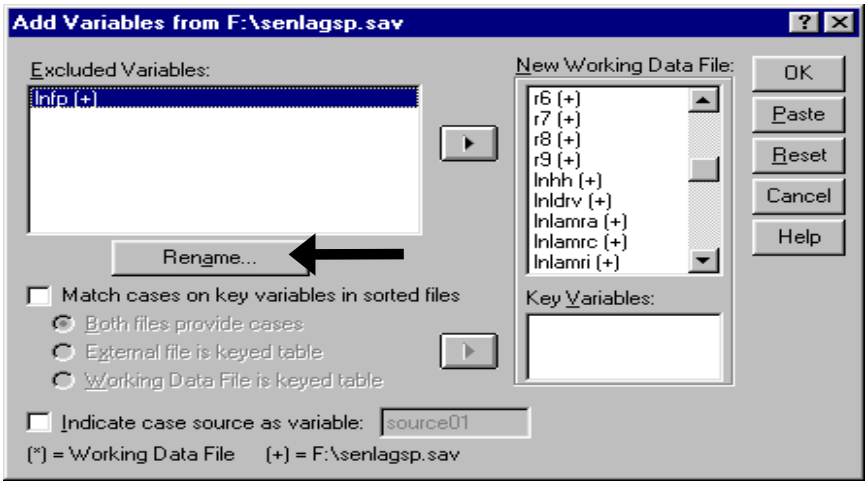
We must first identify the key variables. These are the variables whose correspondence dictates the merge.

Because the key variable has a different name in one file (*lnfp*) than the other (*fam_id*), we must first change the name of the variable *lnfp* to *fam_id*. If they have the same name, then you should skip the renaming.



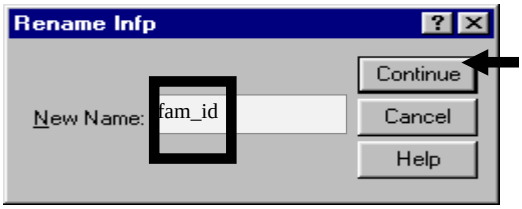
Move the variable *lnfp*(+) into the box "Excluded Variables." It will eventually be moved into the box "Key Variables" after we perform a name concordance.

Click on the button "Rename."

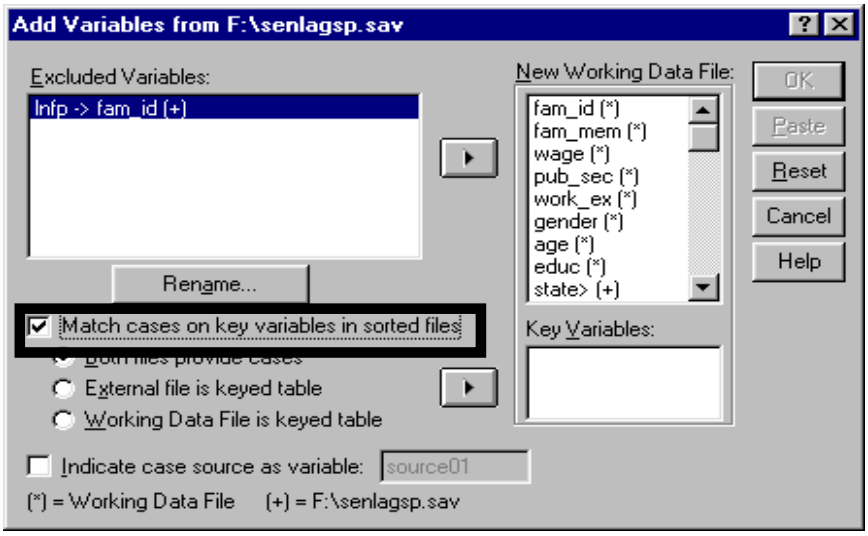


Type in the new name.

Click on "Continue."

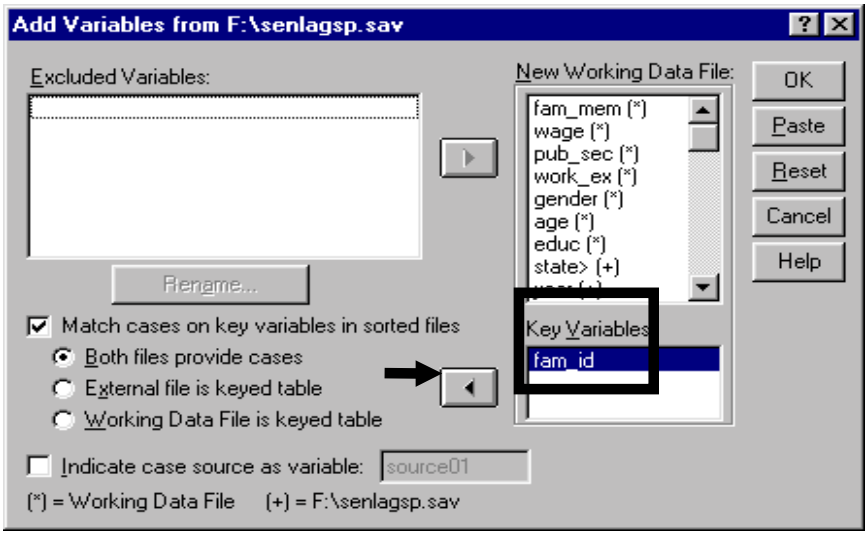


Click on the box to the left of "Match cases on key variables in sorted files."



Click on the variable *fam_id* and move it into the box "Key Variables."

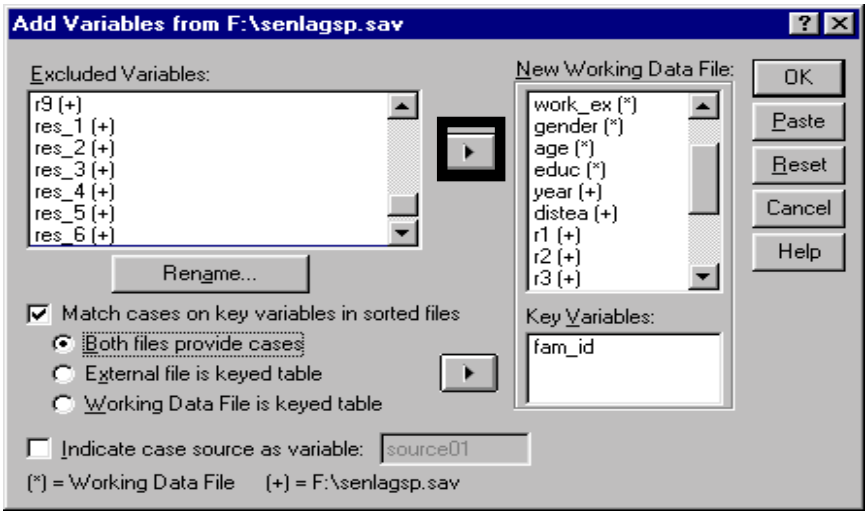
Click on "Both Files Provide Cases." This implies that every unique value of the keyed variable from both the files will be included in the new file[147].



Move all the variables you do not want in your analysis from the box "New Working Data File" into the box "Excluded Variables."

Click on "OK."

A two-way merge will be performed.



Section 13.2.a shows a one-way merge and section 13.2.b compares it to a two-way merge.

Ch 13. Section 2.a. One-way merging

In one-way merging, the cases in the merged file are identical to those in one of the two data sets being merged. So, if you perform a one-way merge of a data set "A" and the working file[148], the merged data set will have only cases from the working data set (see the next picture) or from "A" (see the second picture on the next page).

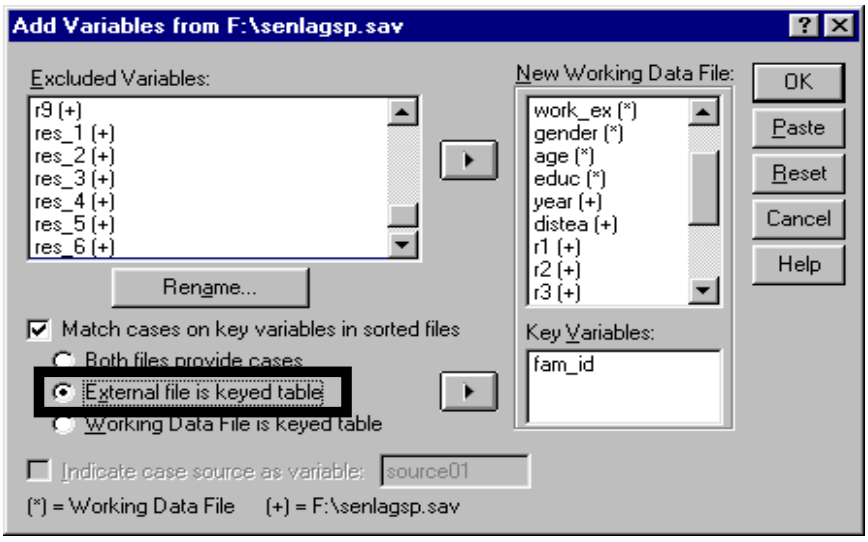
[147] So, if *fam_id* = 121245 has an observation in the original file only, it will still be included - the observations for the variables from the other file will be empty in the new merged data file.

[148] The data set currently open and displayed in SPSS is the "working data set." The file (not opened) from which you want to add variables is the "external file."

Example 1: Cases based on working file

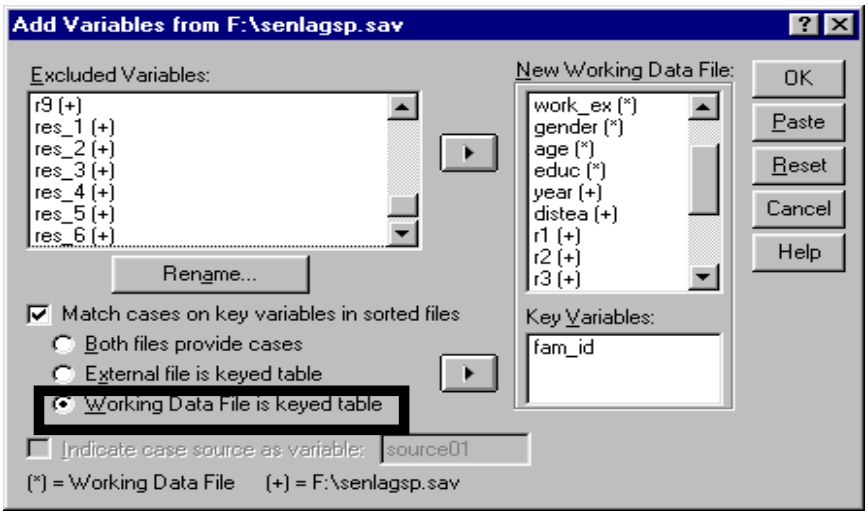
Click on "External File is keyed table." This implies that every unique value of the keyed variable from the working data file will be included, and only those observations from the external data file will be included that have an observation in the working file.

So, if *fam_id* = 121245 and has an observation in the external file only, it will be excluded in the new merged data file.



Example 2: Cases based on external file

In contrast, by clicking on "Working Data File is keyed table," the opposite can be done: one-way merging in which only those cases that are in the external data file are picked up from the working data file.



Ch 13. Section 2.b. Comparing the three kinds of merges: a simple example

Let us take a simple example in which each file has three variables and two observations. The "keying" variable is *fam_id*.

Working data file data (open in the data editor).

fam_id	*wage*	*age*
111111	12	23
555555	25	35

The external file from which data must be added (not open but available on a drive).

fam_id	*educ*	*gender*
999999	18	0
555555	16	1

Note that the respondent "555555" is in both files but "111111" is only in the original file and "999999" is only in the external file.

1. A **two-way merge** using *fam_id* as the key variable will merge all the data (all three *fam_id* cases are included):

Fam_id	*educ*	*gender*	*wage*	*age*
111111			12	23
999999	18	0		
555555	16	1	25	35

2. A **one-way merge using the external data file as the keyed file** will include only those observations that occur in the non-keyed file, which is in the working data file. Here, those have the *fam_id* "111111" and "555555." *Fam_id* "99999" is excluded.

fam_id	*educ*	*gender*	*wage*	*age*
111111			12	23
555555	16	1	25	35

3. A **one-way merge using the working data file as the keyed file** will include only those observations that occur in the non-keyed file, which is in the external data file. Here, those have the *fam_id* "999999" and "555555." *Fam_id* "111111" is excluded.

fam_id	*educ*	*gender*	*wage*	*age*
999999	18	0		
555555	16	1	25	35

Ch 14. NON-PARAMETRIC TESTING

In chapters 3-10 we used procedures that (for the most part) allowed for powerful hypothesis testing. We used tests like the Z, T, F, and Chi-square. The T and F were used repeatedly. In essence, the F was used to determine whether the entire "model" (e.g. - a regression as a whole) was statistically significant and therefore trustworthy. The T was used to test whether specific coefficients/parameters could be said to be equal to a hypothesized number (usually the number zero) in a manner that was statistically reliable or significant. For maximum likelihood methods (like the Logit) the Chi-Square, Wald, and other statistics were used. The use of these tests allowed for the the drawing of conclusions from statistical results.

What is important to remember is that these tests all assume that underlying distribution of variables (and/or estimated variables like the residuals in a regression) follow some "parametric" distribution - the usual assumption is that the variables are distributed as a "normal" distribution. We placed a great emphasis on checking whether a variable was distributed normally (see section 3.2). Unfortunately, most researchers fail to acknowledge the need to check for this assumption.

We leave the decision of how much importance to give the assumption of a "parametric" distribution (whether normal or some other distribution) to you and your professor/boss. However, if you feel that the assumption is not being met and you want to be honest in your research, you should avoid using "parametric" methodologies[149] and use "non-parametric" assumptions instead. The latter does not assume that the variables have any specific distributional properties. Unfortunately, non-parametric tests are usually less powerful than parametric tests.

We have already shown the use of non-parametric tests. These have been placed in the sections appropriate for them:

- 3.2.e (Kolmogirov-Smirnov),
- 4.3.c (Related Samples Test for differences in the distributions of two or more variables),
- 5.3.b (Spearman's Correlation), and
- 5.5.d (Independent Samples Test for independence of the distributions of sub-sets of a continuous variable defined by categories of another variable)

In this chapter we show some more non-parametric tests. Section 14.1 teaches the Binomial test and section 14.2 teaches the Chi-Square test. These test if the distribution of the proportions of the values in a variable conform to hypothesized distribution of proportions for these values. Section 14.3 teaches the Runs test; it checks whether a variable is distributed randomly.

Ch 14. Section 1 Binomial test

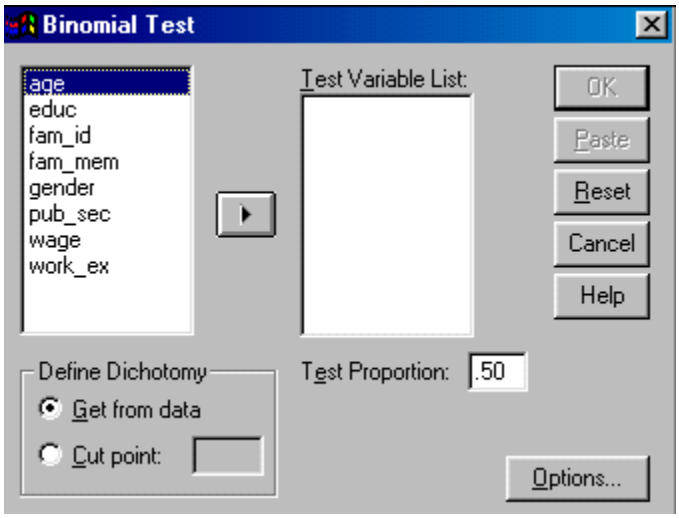
Let's assume we have a variable whose distribution is binomial. That is, the variable can take on only one of two possible values, X and Z.

[149] Almost all of the methods used in this book are parametric - the T-tests, ANOVA, regression, Logit, etc. Note that the Logit presumes that the model fits a Logistic distribution, not a Normal distribution.

The standard example is a coin toss - the outcomes are distributed as binomial. There are two and only two possible outcomes (heads or tails) and if one occurs on a toss then the other cannot also occur on the same toss. The probability of a "tails" outcome and the probability of a "heads" outcome are the relevant parameters of the distribution[150]. Once these are known, you can calculate the mean, standard deviation, etc. Check your textbook for details.

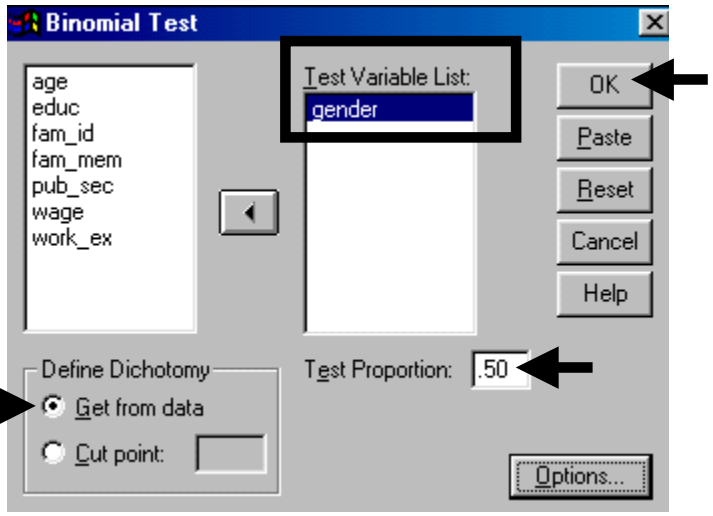
A variable like *gender* is distributed binomially[151]. We want to test the parameters of the distribution – the probabilities of the variable *gender* taking on the value 0 (or "female") versus the probability of it taking on the value 1 (or "male").

Go to STATISTICS/NON-PARAMETRIC/BINOMIAL .



Place the variable *gender* into the area "Test Variable List" (note: you can place more than one variable into the list).

Look at the area "Define Dichotomy." We have chosen "Get from data." This implies that the two possible outcomes are defined in the data (i.e. - in the values of the variable *gender*). They are: 0 (for female) and 1 (for male).



Look at the box "Test Proportion." We have chosen the default of 0.50. We are asking for a test that checks if the "Test Proportion" of .5 equals the probability of *gender* being equal to 0 ("female") for any one observation. As the probabilities have to add to 1, it follows that we are testing if the probability of *gender* being equal to 1 ("male") for any one observation =1- 0.50 = 0.50.

Click on "OK."

[150] The sample size (number of tosses) is also a parameter but it is not an estimated parameter.

[151] Some may disagree. After all, each observation is not a coin toss. However, you can interpret the categories as "outcome of conception/birth."

Interpretation: The two groups are female and male. The observed proportion (or probability) of group 1 (females) is .81 (or 81%). The proportion in the null hypothesis is the "Test Prop" of .50. The Sig value is below .1, .05, and .01. So, we can say with 90, 95, and 99% confidence that the proportion of group 1 values (females) in the data is not equal to the hypothesized proportion of .50.

Binomial Test

	GENDER		
	Group 1	Group 2	Total
Category	Female	Male	
N	1626	390	2016
Observed Prop.	.81	.19	1.00
Test Prop.	.50		
Asymp. Sig. (2-tailed)	.000[a]		

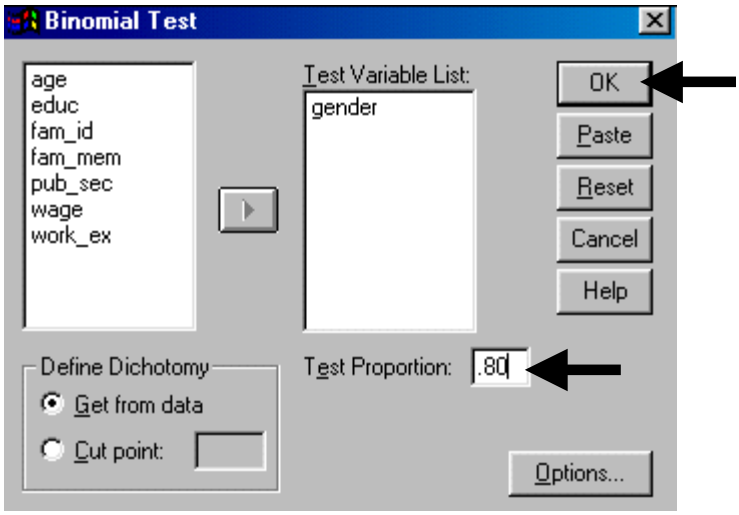
a. Based on Z Approximation.

Example 2: Setting the Test Proportion

We repeat the same procedure, but with a different "Test Proportion."

We use the proportion of .80.

Click on "OK" after entering the hypothesis value of ".80" into the box "Test Proportion."



The observed proportion of group 1 (females) is .8065. Because the Sig value is greater than .1, we can infer that "we cannot reject the hypothesis that the proportion (probability) of group 1 (females) does equal .80." In simpler terms, we can say with 95% confidence that .8 may be the real proportion of group 1.

Binomial Test

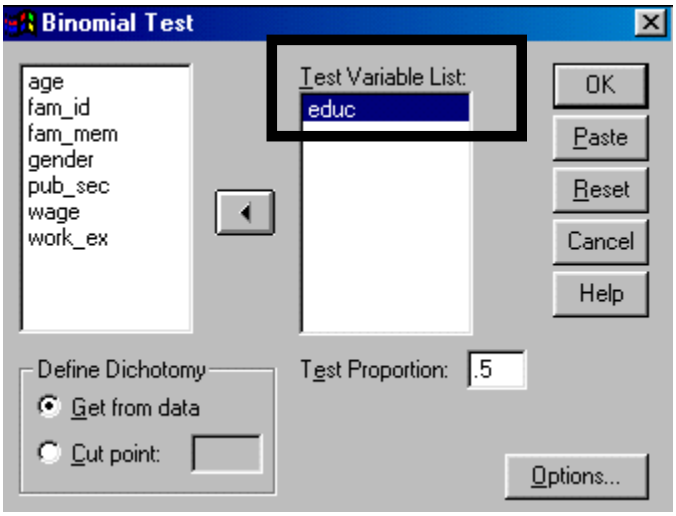
		Category	N	Observed Prop.	Test Prop.	Asymp. Sig. (1-tailed)
GENDER	Group 1	Female	1626	.806548	.8	.240[a]
	Group 2	Male	390	.2		
	Total		2016	1.0		

a. Based on Z Approximation.

Example3: Using a continuous or ordered variable to define the dichotomy of outcomes

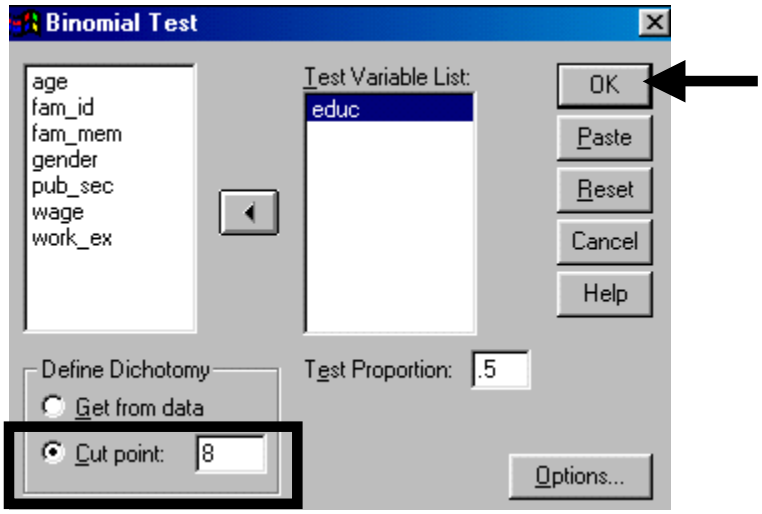
What if we want to use the groups "*education* > 8" and "*education* ≤ 8?"

Choose the variable *education* as the "Test Variable."



Education (in our data set) could take on the values 0-23. We want to define the outcomes as "low" *education* (8 or less) versus "high" (greater than 8). We want to use this definition for defining the dichotomy of outcomes. To do so, click on "Cut point" and enter the value "8" into the box.

Click on "OK."



Interpretation: The observed proportion of cases that fall in the group of "*education* less than or equal to 8" is .67. The "Test Proportion" is .50. Because the Sig value is below .01, we can reject the null hypothesis that the "proportion of cases in group 1 = .50" with 95% confidence.

Binomial Test

	EDUCATION		
	Group 1	Group 2	Total
Category	<= 8	> 8	
N	1346	670	2016
Observed Prop.	.67	.33	1.00
Test Prop.	.50		
Asymp. Sig. (2-tailed)	.000[a]		

a. Based on Z Approximation.

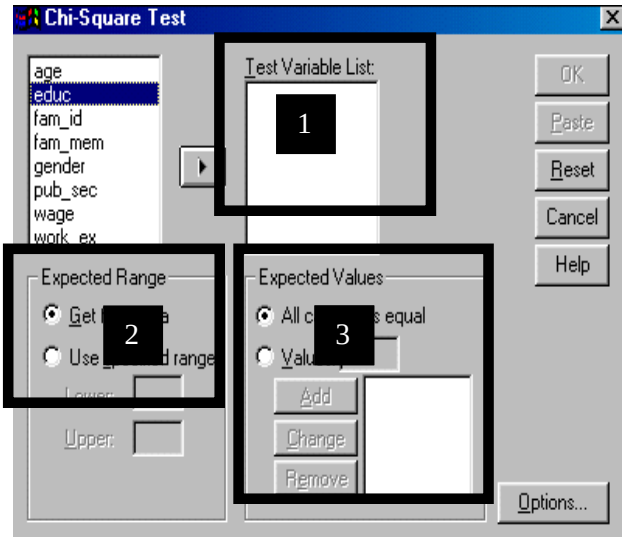
Ch 14. Section 2 Chi-square

Let's assume you have a variable that is categorical or ranked ordinal. You want to test whether the relative frequencies of the values of the variable are similar to a hypothesized distribution of relative frequencies (or you can imagine you are testing observed "proportions" versus hypothesized "proportions"). You do not know what the distribution type is, nor do you care. All you are interested in is testing whether the "measured" relative frequencies/proportions are similar to the "expected" relative frequencies/proportions.

Example 1: A basic example

For example, assume you want to check whether the proportions of all values of *education* (measured in terms of years of schooling) are the same. A histogram of this hypothesized distribution would be a perfect rectangle.

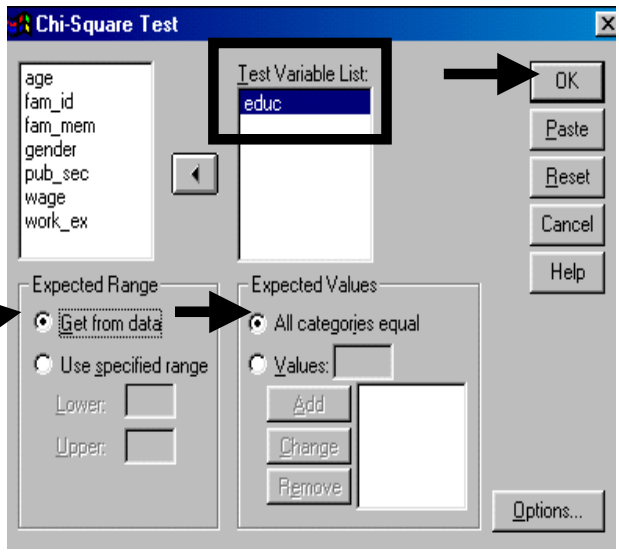
Go to STATISTICS / NON-PARAMETRIC TESTS / CHI-SQUARE TEST. The following dialog box opens. In area 1 you choose the variable(s) whose proportions you want to check (note: do not place a continuous variable here). In area 2 you define the range of values whose proportions you wish to check (essentially, you are telling SPSS to constrain itself to a sub-set of only those values you choose here). In area 3 you define the hypothesis of expected proportions.



Choose the variable(s) whose distribution of "proportions" you want to check and move it into the box "Test Variable List."

Now we will test for the entire range of values for *education*. To do that, choose the option "Get from data" in the area "Expected Range." Note: If you have defined a value as missing using the method shown in section 1.2, then SPSS will not use that value.

We will use a very simple hypothesis for the first example. We are testing whether the proportions of all the observed values of *education* are equal. To do this, choose the option "All categories equal" in the area "Expected Values." Click on "OK."



The first output table compares the observed occurrence (or frequency) of each *education* value (the values are given in the first column as 0, 1, 2, …,23). In this table:

- The second column gives the actual number of observations with the respective *education* level.
- The third column ("Expected N") gives the number of observations for each *education* level, as expected under the null hypothesis of all frequencies being equal.

EDUCATION

	Observed N	Expected N	Residual
0	151	87.7	63.3
1	680	87.7	592.3
2	15	87.7	-72.7
3	47	87.7	-40.7
4	46	87.7	-41.7
5	71	87.7	-16.7
6	286	87.7	198.3
8	50	87.7	-37.7
9	53	87.7	-34.7
10	54	87.7	-33.7
11	172	87.7	84.3
12	61	87.7	-26.7
13	46	87.7	-41.7
14	124	87.7	36.3
15	25	87.7	-62.7
16	34	87.7	-53.7
17	19	87.7	-68.7
18	51	87.7	-36.7
19	8	87.7	-79.7
20	8	87.7	-79.7
21	6	87.7	-81.7
22	7	87.7	-80.7
23	2	87.7	-85.7
Total	2016		

The estimated Chi-square statistic is significant at the 99% level (because Asymp. Sig. < .01), so the null hypothesis can be rejected. In simple terms: "The values of *education* do not have the same frequencies - the variable does not have a uniform distribution."

Test Statistics

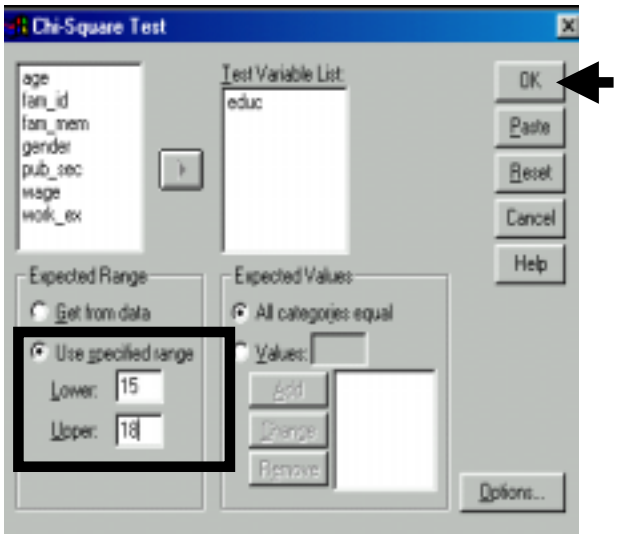
	EDUCATION
Chi-Square[a]	5292.090
df	22
Asymp. Sig.	.000

a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 87.7.

Example 2: Testing over a limited range of values

We can also constrain the range of values over which we test. Continuing the previous example, we want to test if the values of education in the range 15 to 18 have the same frequencies. The only difference from the previous example is that in the area "Expected Range," choose the option "Use specified range" and enter the range as shown. The end-points are inclusive.

Click on "OK."



The interpretation of this table is the same as in the first example. Notice that the "Category" column only has the values you chose. As a result of that step, the "Expected N" has been scaled down to one appropriate for the small sample size of the constrained range of *education*.

Frequencies

	EDUCATION			
	Category	Observed N	Expected N	Residual
1	15	25	32.3	-7.3
2	16	34	32.3	1.8
3	17	19	32.3	-13.3
4	18	51	32.3	18.8
Total		129		

The estimated Chi-square statistic is significant at the 99% level (because Asymp. Sig. < .01),[152] so the null hypothesis can be rejected. In simple terms: "The values of *education* within the range 15 to 18 do not have the same frequencies - the variable within the range of values 15 to 18 does not have a uniform distribution."

Test Statistics

	EDUCATION
Chi-Square[a]	18.070
df	3
Asymp. Sig.	.000

a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 32.3.

Example 3: Testing a complex hypothesis

[152] To repeat the criterion for significance again:

- If Sig <.01, then significant at the 99% level
- If Sig <.05, then significant at the 95% level
- If Sig <.1, then significant at the 90% level
- If Sig >.1, then not significant

The previous examples tested a very simple hypothesis - "All the frequencies are equal." This example shows the use of a more complex and realistic hypothesis.

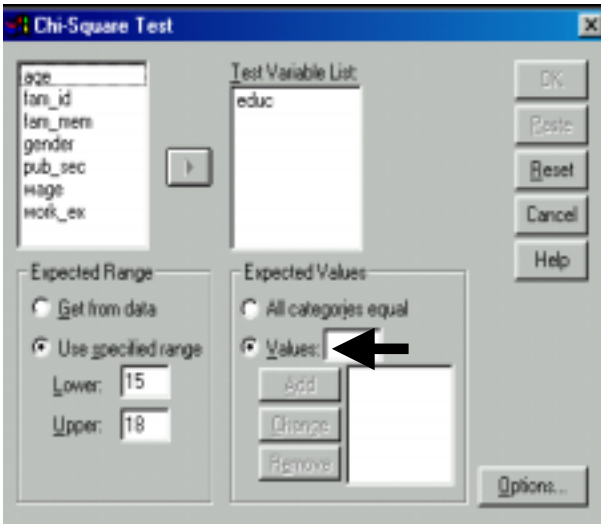
We want to test if:

The *education* value of	**Is this proportion of cases** (out of the cases defined by only those education values in column 1)
15	0.1 or 10%
16	0.2 or 20%
17	0.4 or 40%
18	0.3 or 30%

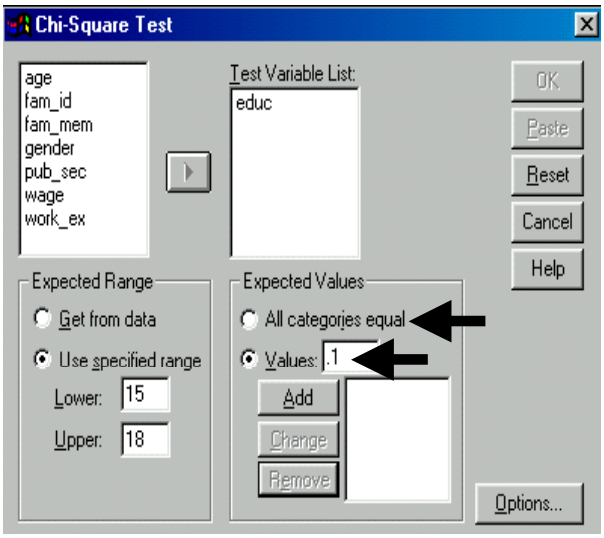
Note: This is the way you will most often be using the test.

We continue using the same dialog box as in the previous example.

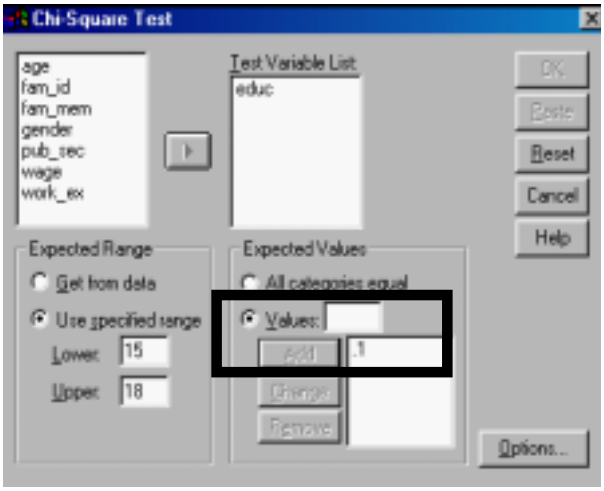
Now we are using a complex set of criteria for the test. To enter this set, choose the option "Values" within the area "Expected Values."



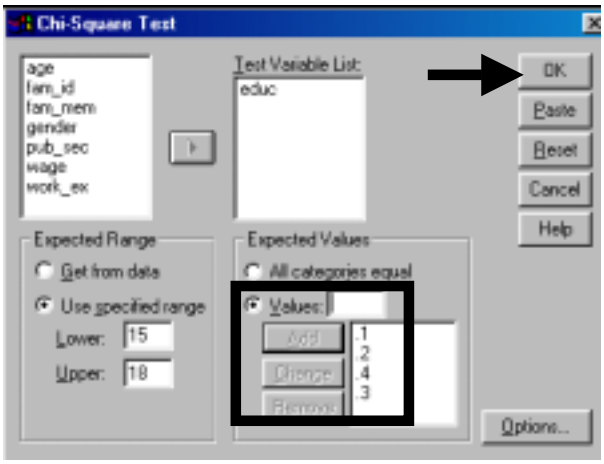
The first item in the hypothesis is that "value of 15 → Proportion of .1." Enter this proportion into the box "Values." Click on the button "Add."



The first item in the hypothesis has been added. SPSS will link this to the "Lower" end point (15 in this example).



Do the same for the other three hypothesis items. SPSS will assign them, in order, to the ascending values 16, 17, and 18. Click on "OK."



The interpretation is the same as in the last example. The only difference is that the "Expected N" are no longer equal. Instead, they are based on the items in the hypothesis.

Frequencies

	EDUCATION			
	Category	Observed N	Expected N	Residual
1	15	25	12.9	12.1
2	16	34	25.8	8.2
3	17	19	51.6	-32.6
4	18	51	38.7	12.3
Total		129		

The estimated Chi-square statistic is significant at the 99% level (because Asymp. Sig. < .01), so the null hypothesis can be rejected. In simple terms, "The values of *education* within the range 15 to 18 do not have the relative frequency distribution presented in the hypothesis (15→ .1, 16→ .2, 17→.4 and 18→.3)."

Test Statistics

	EDUCATION
Chi-Square[a]	38.461
df	3
Asymp. Sig.	.000

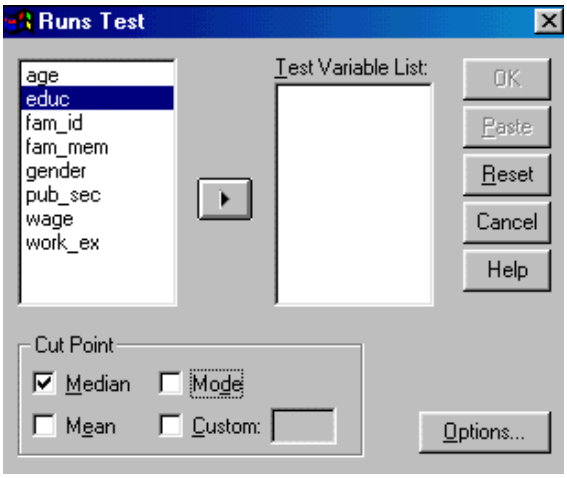
a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 12.9.

Ch 14. Section 3 The Runs Test - checking whether a variable is really "randomly distributed"

The "Runs Test" checks whether the values of a continuous variable are actually "random" as is presumed in all "random samples." The null hypothesis is that the variable is not random.

An excellent application of this test is to determine whether the residuals from a regression are distributed randomly or not. If not, then a classical assumption of linear regression has been violated. See section 7.2 also.

Go to STATISTICS/NON-PARAMETRIC TESTING/RUNS.



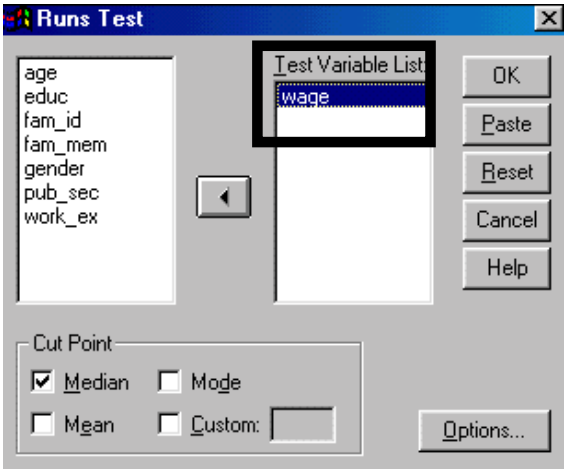
The runs test is valid only for continuous quantitative data. The test is conducted in two parts:

1. The variable is split into two groups on the basis of a "Cut Point" measure (which may be the mean, median, mode, or any other value you choose).

2. The test uses this dichotomy of groups.

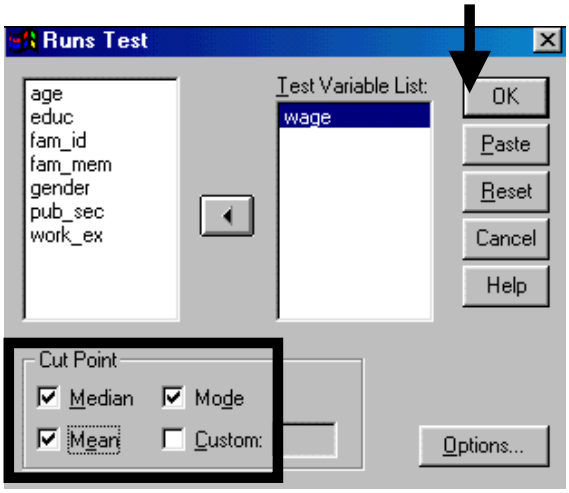
Choose the variable(s) you want to test and place it (or them) in the box "Test Variable List."

Note: You can choose more than one continuous quantitative variable.



Select the criteria for the "Cut Point." You can select more than one criterion. A separate runs test will be run for each criterion.

Choose the options shown and click on "OK."



Using the median
(the usual method)

Runs Test

	WAGE
Test Value[a]	5.9500
Cases < Test Value	1008
Cases >= Test Value	1008
Total Cases	2016
Number of Runs	432
Z	-25.708
Asymp. Sig. (2-tailed)	.000

a. Median

Using the mean

Runs Test 2

	WAGE
Test Value[a]	9.0484
Cases < Test Value	1368
Cases >= Test Value	648
Total Cases	2016
Number of Runs	96
Z	-40.062
Asymp. Sig. (2-tailed)	.000

a. Mean

Using the mode

Runs Test 3

	WAGE
Test Value[a]	3.75
Cases < Test Value	513
Cases >= Test Value	1503
Total Cases	2016
Number of Runs	504
Z	-15.381
Asymp. Sig. (2-tailed)	.000

a. Mode

Interpretation: The "Test Value" in each output table corresponds to the statistic/value used as the "Cut Point." The median=5.95, mean=9.04, and mode=3.75.

Look at the rows "Asymp., Sig., and (2-tailed)." All the tests show that the null can be rejected. We can therefore say that "Runs Tests using all three measures of central tendency (median, mean, and mode) indicated that wage comes from a random sample."

To take quizzes on topics within each chapter, go to http://www.spss.org/wwwroot/spssquiz.asp

Ch 15. SETTING SYSTEM DEFAULTS

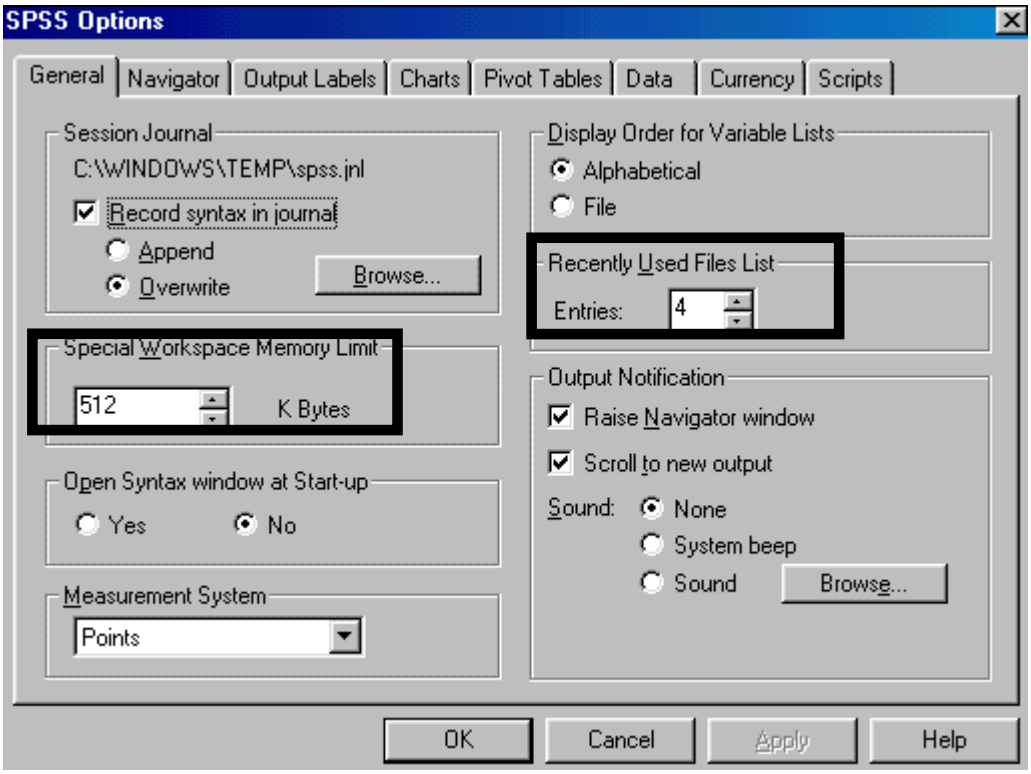
In most software packages the user is able to set some system options/defaults. If you haven't used this feature in Excel or Word, go to TOOLS/OPTIONS and try the default settings. In SPSS, you can change some of the default settings. For the most part, these settings define the default format of output and the manner in which data are shown.

In section 15.1 we show how to set general system options. The most important settings are those for the default format of output tables (called "Pivot Tables" in SPSS) and the labels on output.

Section 15.2 shows how to change the manner in which data/text is shown on screen.

Ch 15. Section 1 General settings

Go to EDIT/OPTIONS. The settings you choose here set the default environment for SPSS on your machine (and perhaps the entire school/office network - check with your system administrator). The best way to learn about these settings is to "play around" with different options. In the following section, we briefly demonstrate the most important settings that you should customize.



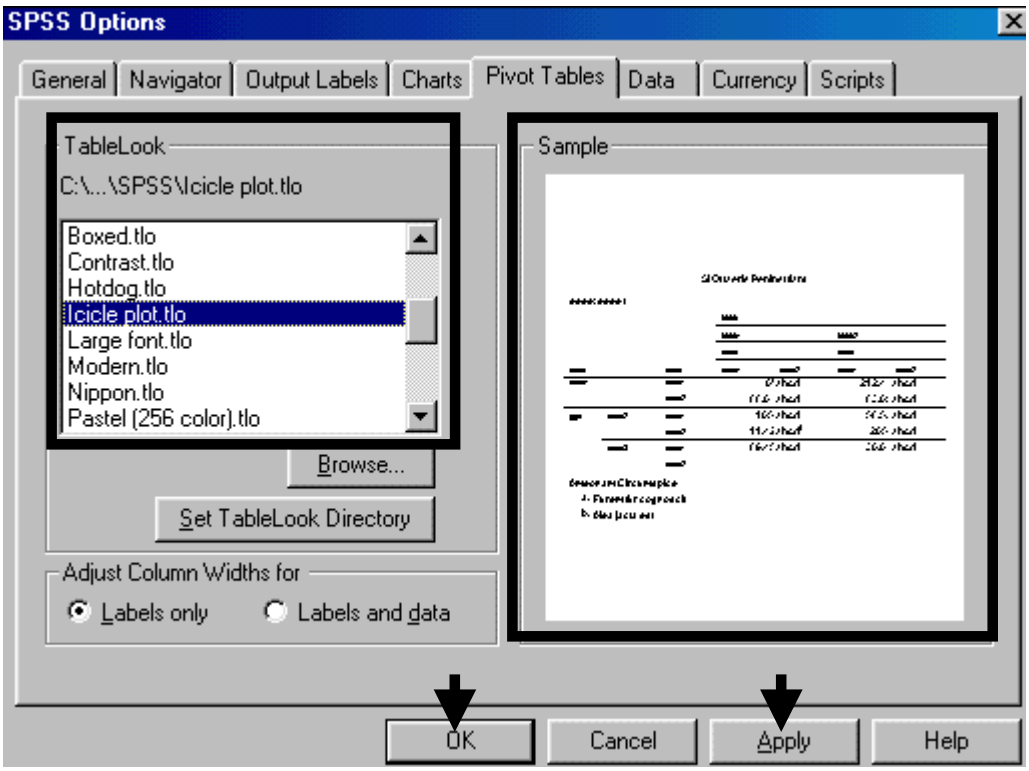
We would suggest choosing the options as shown above. You may want to change:

- The “Recently Used Files List” to a number such as 8 or 10. When you open the menu FILE, the files you used recently are shown at the bottom. When you choose to see 8 files, then the last 8 files will be shown. You can go to any of those files by clicking on them.
- “Special Workspace Memory Limit” may be increased by a factor of 1.5 to 2 if you find that SPSS is crashing often. It's always a good idea to ask your system administrator before making a change.

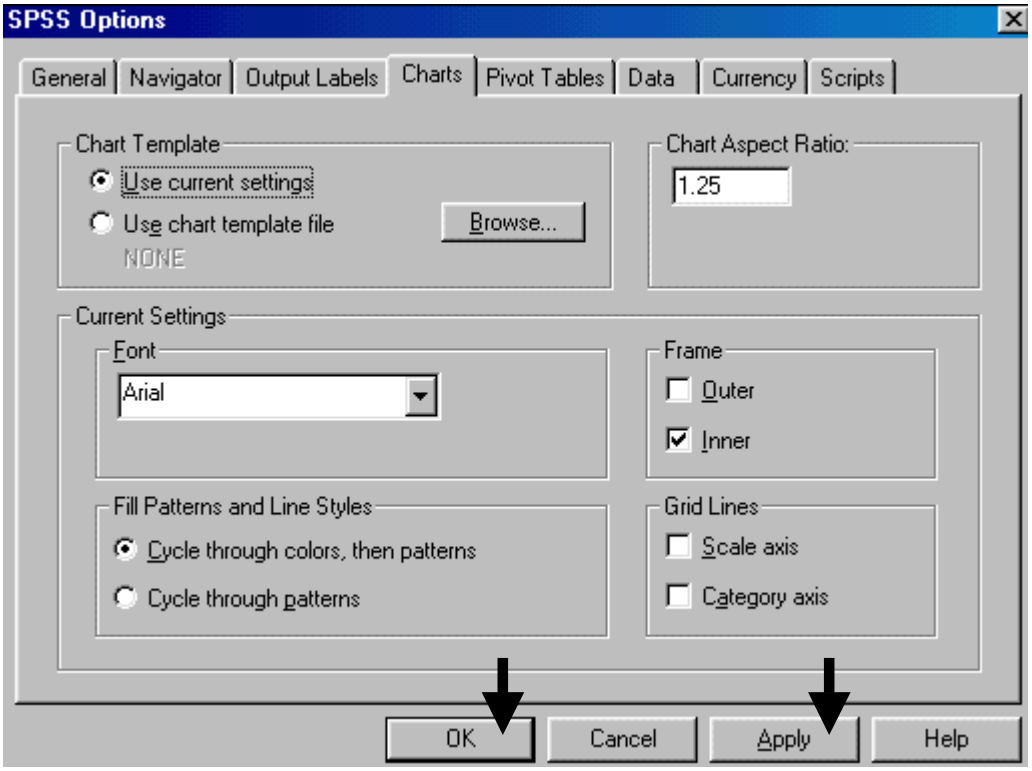
Click on the tab “Pivot Tables.” (See next picture). The box on the left shows table formatting styles called “Table Looks." Each item on the list corresponds to one look. The “look” defines several formatting features:

- Font type, size, style (bold, italic, color, etc.)
- Cell shadings
- Border width and type
- Other features

When you click on the name of a “look” on the left side, a sample appears on the right side. Choose the “look” you prefer and press “Apply” and “OK.” See section 11.1 for more on table formatting and changing the "look" of individual tables.

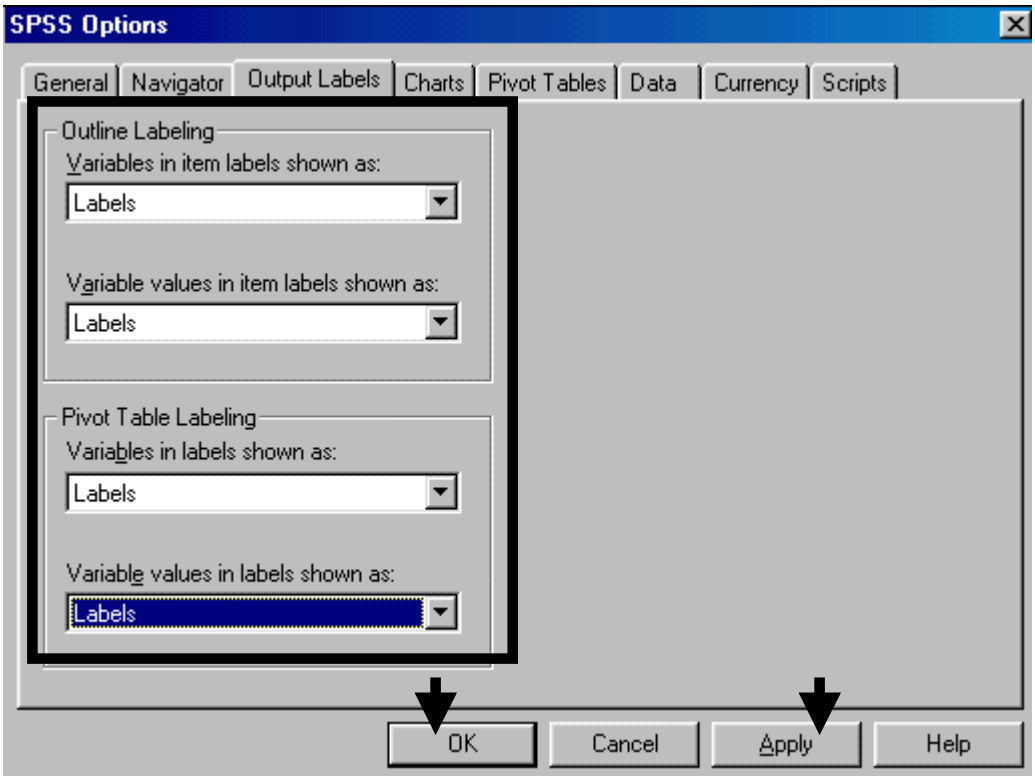


Click on the tab “Charts” and choose the settings you like. (See next picture). Experiment until you get the right combination of font, frames, grid lines, etc. When you are finished, press “Apply” and “OK.”

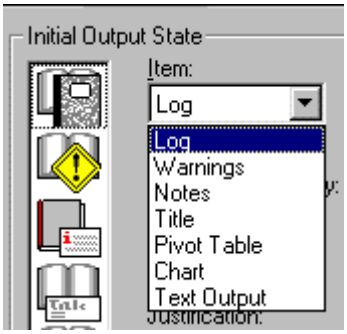


Choosing to see labels instead of variable names and values

The most important option is the choice of labels to depict variables and values of categorical variables in output tables and charts. Click on "Output Labels" and choose "Labels" for all the options. Press "Apply" and "OK." (See next picture).

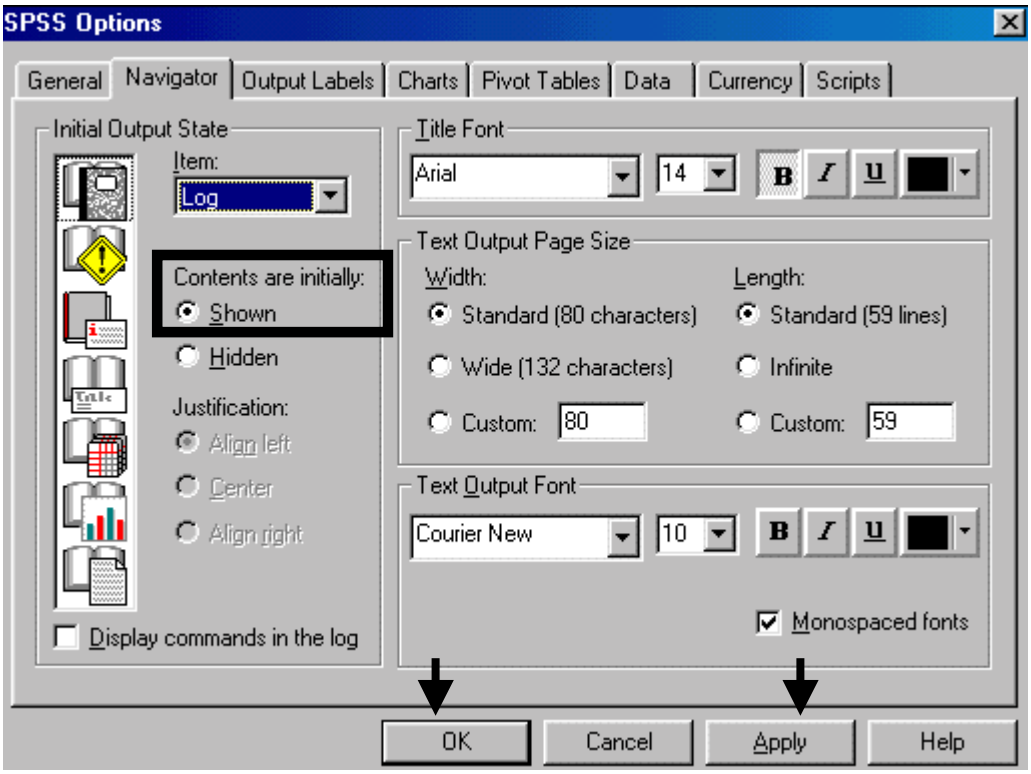


Finally, click on the tab "Navigator." This is the Output Window, the window that includes all the output tables and charts. Click on "Item." You will see 7 items. (See next picture).



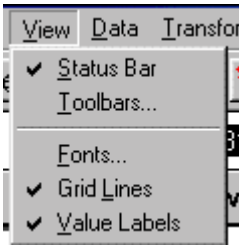
If they are not all accompanied by the option "Shown," then simply:

- Choose the relevant item (e.g. -"Warnings") from the item list.
- Choose the option "Shown" in the area "Contents are initially." (See next picture).



Ch 15. Section 2 Choosing the default view of the data and screen

Go to VIEW.



You can choose to view/hide:

- Gridlines (leave this checked).
- Value Labels (if you have defined the values 0 and 1 of the variable *gender* as "male" and "female" respectively, then your data sheet will show "male" or "female" instead of 0 or 1).
- Status bar (in any software program, it is the raised gray lower border to the application. It contains information on the status of processing and the cursor. For example, "Getting data…" whenever you load new data into SPSS. On the right side it tells you if SPLIT FILE (chapter 10) is on or off, if SELECT CASE (section 1.7) is on or off, and if WEIGHING is on or off.

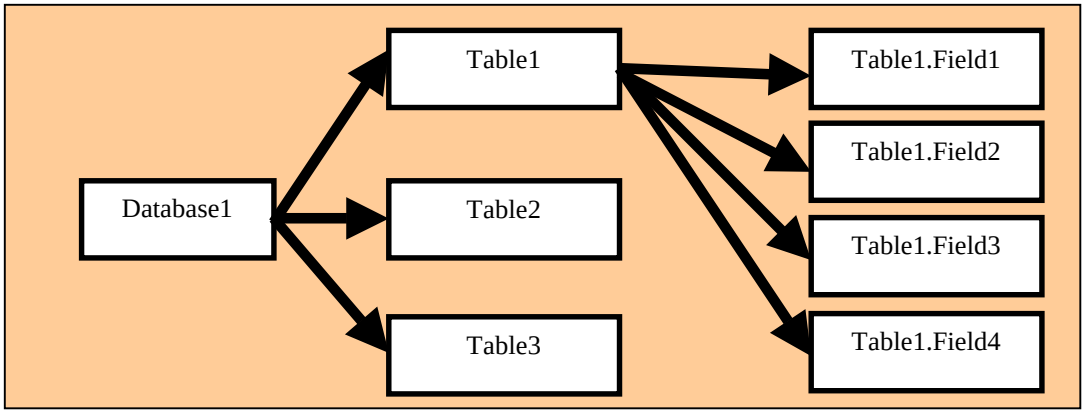
You can also choose the font in which data are shown on screen. The font you choose does not affect the font in output tables and charts.

To take quizzes on topics within each chapter, go to http://www.spss.org/wwwroot/spssquiz.asp

Ch 16. READING DATA FROM DATABASE FORMATS

In the business world, data are often stored in systems called Relational Database Management Systems (RDBMS). These include applications like Oracle, SQL Server, Sybase, Wang, FoxPro, and Access. The applications have a few common features:

- The data are stored in the same structure. This structure essentially consists of three parts - the database, individual tables within the database, and individual fields within each table. The best intuitive analogy is an Excel workbook - the entire file (also called workbook) is analogous to the database file, each sheet within the workbook is analogous to a table, and each column to a field. For this reason, Excel can be treated as a database if the data are stored strictly in columns.

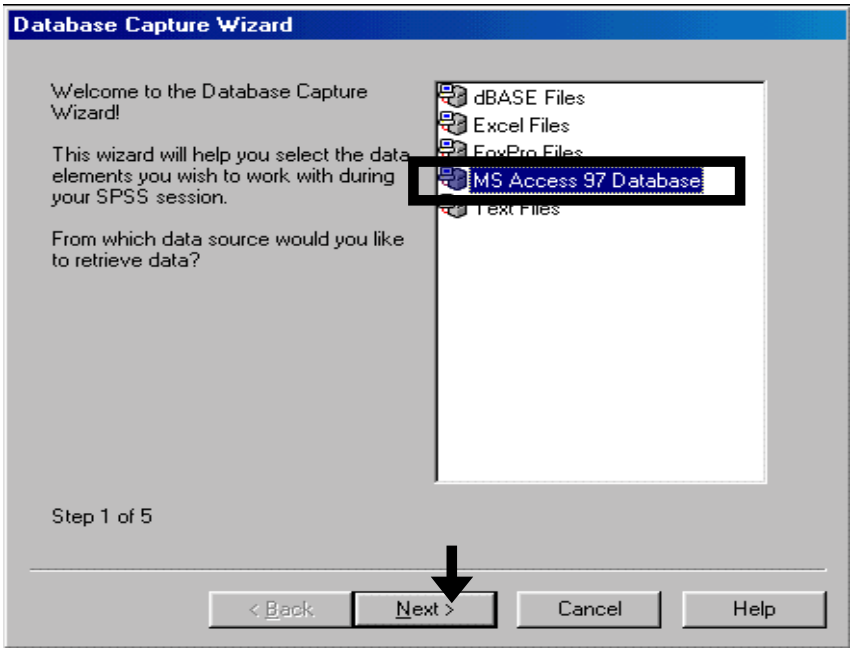


- A common programming language (called SQL) can be used to manipulate data and run procedures in all these programs. For example, in Excel, look at the option DATA/GET EXTERNAL DATA.

For the purpose of learning how to read data into SPSS, you need not learn the details about database structures or language. The important inference from the two points above is that, irrespective of the source application, the commonality of data storage features permits one process to be used for accessing data from any of the applications. If you learn the process for one application, you can do it for the others. We provide an example using Access.

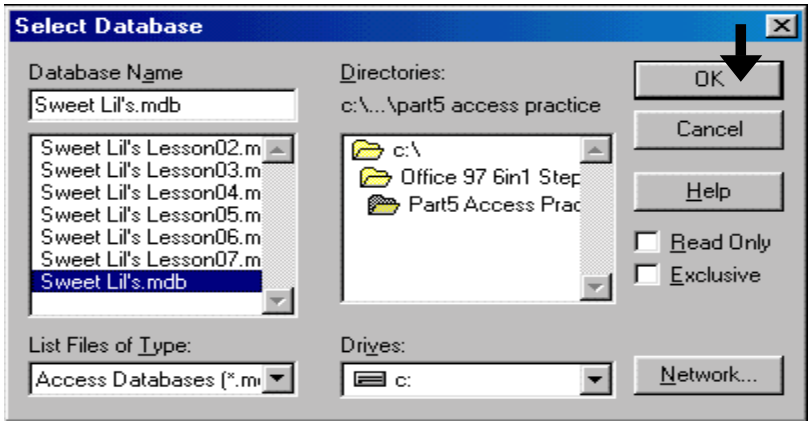
Note: In SPSS versions 9 and 10 you will see some more features for reading data. You can ignore them; the procedures shown in this book should be sufficient.

Assume that you want to read in data from five fields in two tables in an Access database, i.e. - a SPSS file with five variables selected from a database with many more. Go to FILE/DATABASE CAPTURE.



As the dialog box indicates on the bottom-left, the process includes five steps. In step 1, choose the "Data Source" that includes the database in which you are interested. Basically, it is asking you to name the application in whose format the data is stored. Here we see five applications[153].

Choose "MS Access…" and press "Next."



Locate the file name and press "OK."

Note: You may have to click on "Network" and enter the location, password, etc. in order to access the file. If the data managers are afraid that unauthorized users may harm the original data, they may have made it "Read Only." Find out from them if this is the case. If so, then choose the option "Read Only." This allows you to read from the file but will not allow you to write to it.

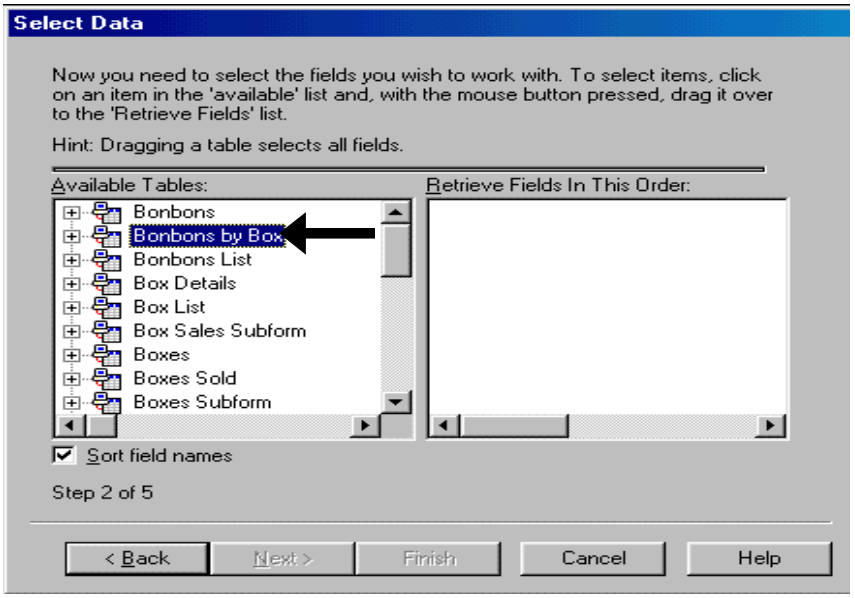
Click on "OK."

[153] What about Oracle, SQL Server, etc? The reason why these five options are shown is that the system on the computer we worked on had "Drivers" for these five formats. You can buy (and maybe even download for free from web sites like cnet.com) and install drivers for other applications. If you are curious, look at the option "ODBC Drivers" under your computer's "Control Panel." The easier way would be to ask your IT guru to do the install.

Now you have reached step 2. The left half of the next dialog box shows all of the available tables.

Remember: each database is analogous to an Excel Workbook and each table is analogous to one Excel sheet within the workbook.

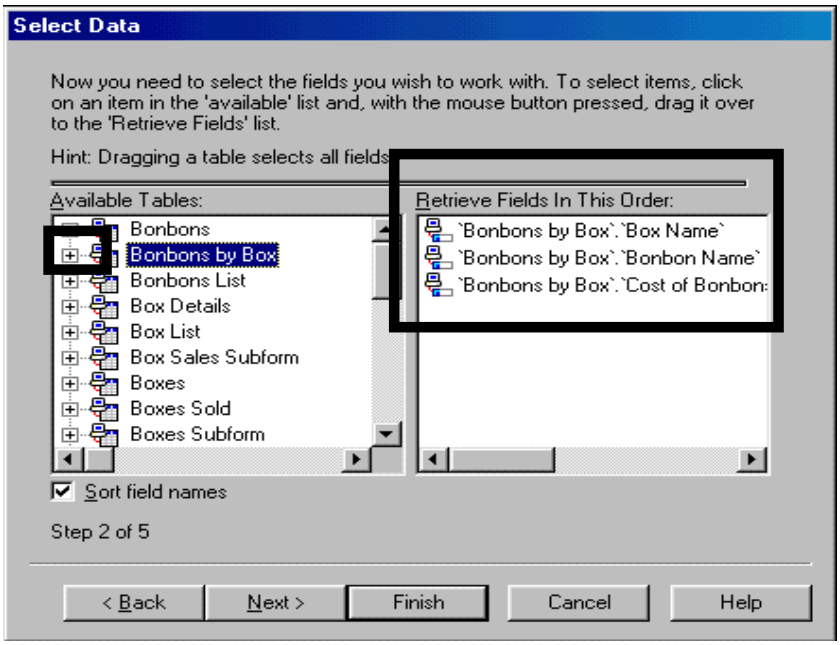
Click on the first table from which you want to extract data.



Now you must select the fields within the table you selected. You can do this in one of two ways:

- Click on the "+" button next to the table name. All the fields inside the table will be shown on the left half. Drag any fields you want to extract into the right half.

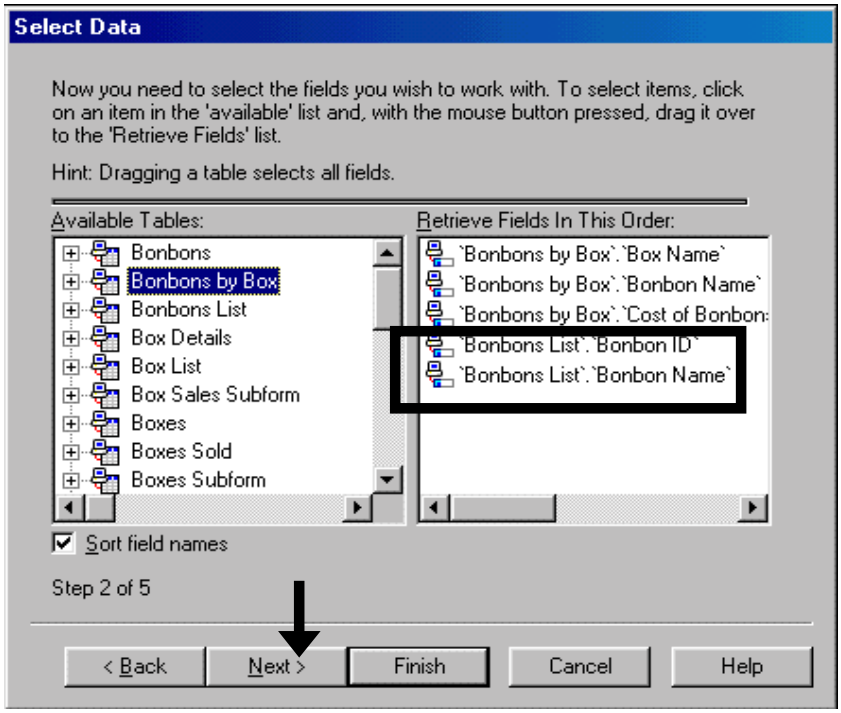
- Drag and move the table name. In this case, all the fields will be selected. This is what we have done in the next figure.



We also need two fields from another table. Repeat the same process as above.

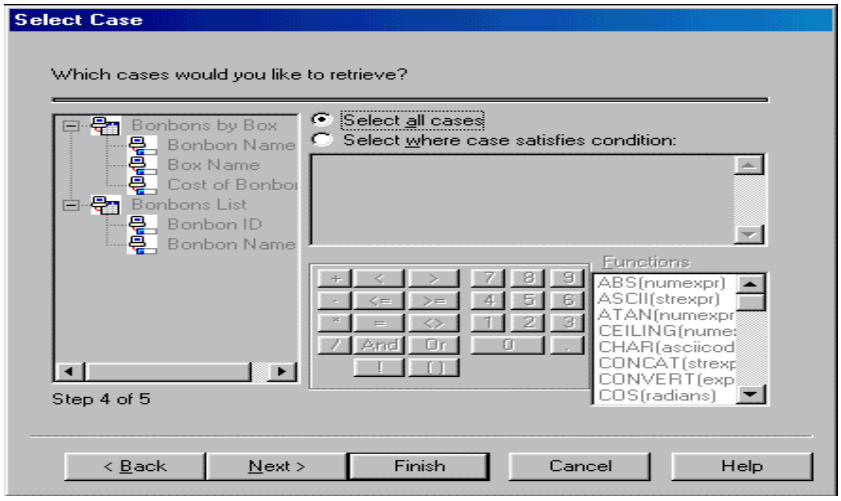
On the right side, be aware of the notation for each item: it is "table name.field name."

Click on "Next."



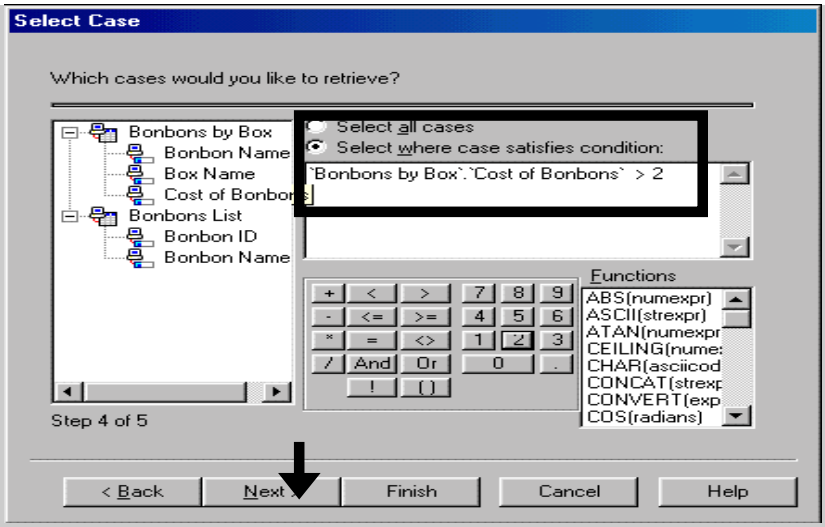
If you want to restrict the cases you select, click on "Select where case satisfies…"

Otherwise click on "Next."



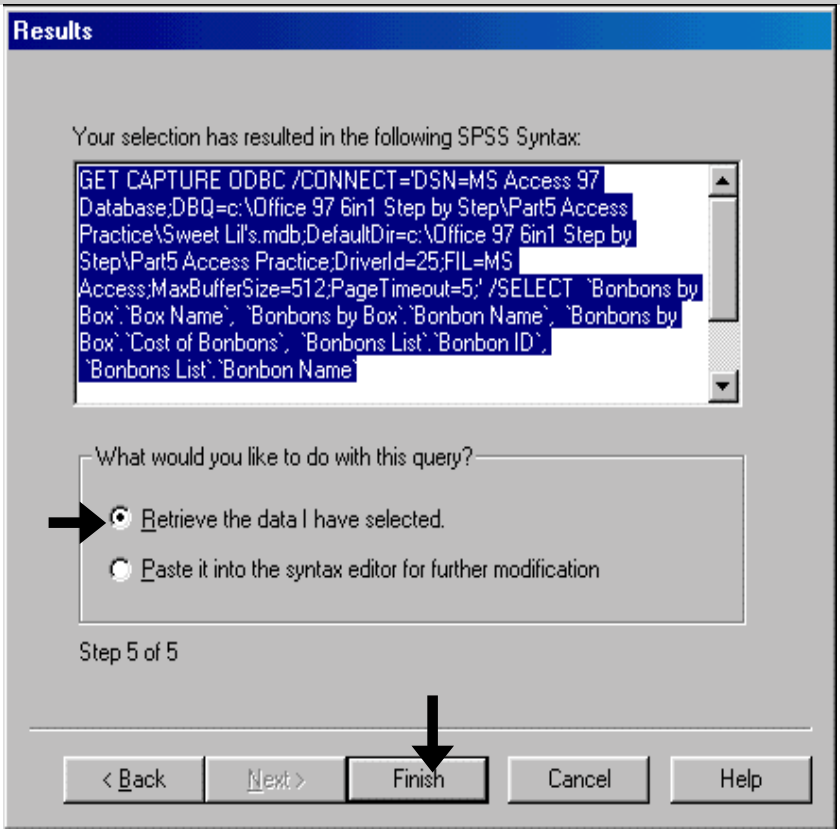
If you selected to restrict the cases, enter the criteria (similar to the process shown in section 1.7).

Click "Next."



Choose "Finish." The data will be read into SPSS. Save the file as a SPSS file with the extension ".sav."

For those who are curious, the code you see is in the language "SQL" (Standard Query Language). It can be used (with some minor changes) to retrieve data from the Access database to SPSS, Excel, Oracle, etc.



To take quizzes on topics within each chapter, go to http://www.spss.org/wwwroot/spssquiz.asp

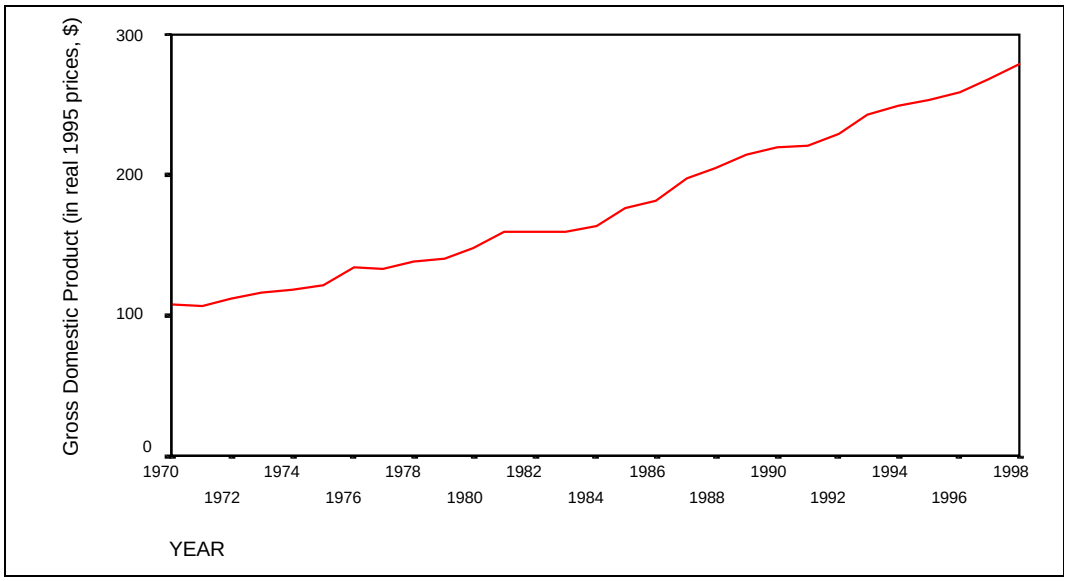
Ch 17. TIME SERIES ANALYSIS

This chapter leads up to the construction and estimation of an ARIMA model, the preferred model specification technique to use when the data set is a Time Series.

A typical Time Series is "*US Income*, *Investment*, and *Consumption* from 1970-98." The data are usually arranged by ascending time sequence. Year, quarter (four months), month, etc. may define a time period. The reasons for using ARIMA (and not a simple OLS Linear Regression model) when any of the regression variables is a time series are:

- The fact that the value of a variable in a period (e.g. - 1985) is typically related to "lagged" (or previous) values of the same variable[154]. In such a scenario, the "lagged" value(s) of the dependent variable can function as independent variable(s)[155]. Omitting them may cause an Omitted Variable Bias. The "AR" in "ARIMA" refers to the specification of this "Auto-Regressive" component. Section 17.2 shows an example, which is reproduced below.

As the graph below depicts, the value for any single year is a function of the value for the previous year(s) and some increment thereof.

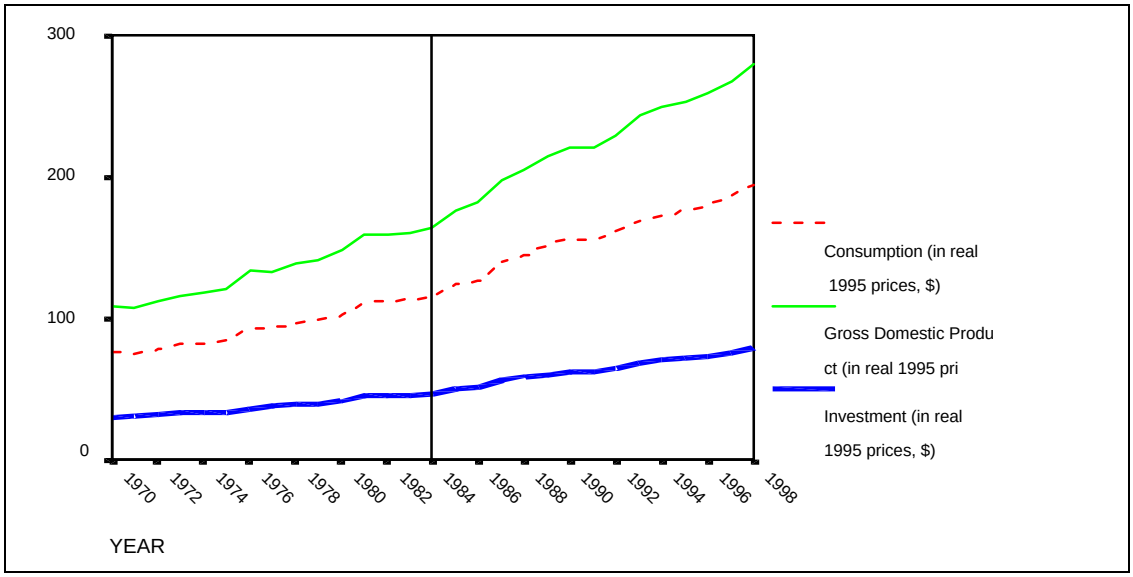


- The value at any period in a time series is related to its values in previous time periods. From general knowledge, you would know that the value of a variable such as national income (*GDP*), even when adjusted for inflation, has been increasing over time. This means that for any sub-set defined by a period (e.g. - 1970-84, 195-96), the attributes of the variable are

[154] For example, *US Income* in 1985 is related to the levels in 1980, 81, 82, 83. Income in 1956 is related to the *incomes* in previous years.

[155] For example,
$GDP_t = a + b*GDP_{t-1} + p*GDP_{t-2} + r*GDP_{t-3} + c*Inv_t + more\ldots$

changing. The classical assumptions that underlie a Linear Regression demand "stationarity" (or "randomness") of a variable[156]. That is, "irrespective of the Sub-set of observations chosen, the expected mean and variance should be constant and cross-observation relations should be zero (so, the relation between observation 'n' and 'n-2' should be zero." However, for a time series, these attributes change over time, and thus the series is "non-stationary." Look at the next chart - you can see that the mean for any of the variables for the period until 1984 is lower than that for the period after 1984. As such, the variable cannot be used in a Linear Regression.



- The question of lagged influence also arises across variables. For example, *investment* in 1980, 81, 82, 83, 84, and 85 may influence the *level of income* (*GDP*) in 1985. The cross-correlation function, shown in section 17.3, helps to determine if any lagged values of an independent variable must be used in the regression. Hence, we may end up using three variables for investment - *this period's investment, last period's investment, and investment from two periods prior*.[157]

- The presence of a "Moving Average" relation across the residuals at each time period. Often, the residuals from year T are a function of T-1, T-2, etc. A detailed description of "Moving

[156] Stationarity implies "random," and non-stationarity the opposite. If a variable were truly random, then its value in 1985 should not be dependent on its own historical values. Essentially, time series data are in conflict with the classical assumptions primarily because each variable that is non-stationary is not obeying a key classical assumption: "each variable is distributed randomly."

[157] What about collinearity between these? Would not that cause a problem in the regression? The answers are:

- Rarely is more than one transformation used in the same model
- Once other transformations have taken place, there may be no such collinearity
- In any case, collinearity is a lesser problem than mis-specification
- Lastly, SPSS uses a Maximum Likelihood Estimation method (and not Linear Regression) to estimate the ARIMA.

Average" process is beyond the scope of this book.

- Autocorrelation between the residuals. Section 17.6 shows a method to correct for first-order autocorrelation. For higher-order autocorrelation, consult your textbooks for methods of detection and correction.

- Co-integration is a complex method of correcting for non-stationarity. A detailed discussion is beyond the scope of this book; but in section 17.7 we provide an intuitive grasp of co-integration.

This edition ignores seasonality.

Regarding Unit Roots, Non-Stationarity, Cointegration, DF Test, PACF, ARIMA, and other complex tests: could this be **much ado about nothing**?" A cynical view of Time Series analysis would suggest as much. In practice, most macroeconomists don't even test for non-stationarity. They simply transform everything into differenced forms, maybe using logs, and run a simple OLS! From our experience, what you will learn in this chapter should suffice for most non-Ph.D. Time Series analysis.

Graphical analysis is essential for time series. The first graph one should obtain is the "pattern of variables across time." Essentially, this involves a multiple-line graph with time on the X-axis. Section 17.1 shows how to make "Sequence" charts and makes simple inferences from the charts as to the implications for a linear regression model.

Section 17.2 tests for non-stationarity using the Partial Autocorrelation Function (PACF) charts. SPSS does not conduct formal tests for Unit Roots like the Dickey Fuller test but, in our experience, the PACF is usually sufficient for testing for non-stationarity and Unit Roots. We also show the ACF (Autocorrelation function). Together, the PACF and ACF provide an indication of the "integration-order" (differencing required to make a variable stationary) and the "Moving Average." If a variable is non-stationary, it cannot be used in a regression. Instead, a non-stationary transformation of the variable must be used (if this is unclear, wait until the end of section 17.2.)

Section 17.3 shows how to determine whether any lagged values of an independent variable must be used in the regression. The method used is the cross-correlation function (CCF).

After testing for non-stationarity, one may have to create new variables for use in a regression. The PACF will tell us about the type of transformation required. Section 17.4 shows how to create these new "transformed" variables.

After creating the new variables, you are ready for running a regression on the time series data. The generic method for such regressions is called ARIMA for "Autoregressive etc." It allows the incorporation of an autoregressive component (i.e. - the lagged value of the dependent variable as an independent variable), differencing (for overcoming the obstacle of non-stationarity), and moving average correction. Section 17.5 shows an example of a simple ARIMA model.

Even after correcting for non-stationarity in each variable, the model as a whole may still suffer from the problem of autocorrelation among residuals. Section 17.6 shows a procedure that allows for automatic correction of first-order autocorrelation and also allows for incorporation of an autoregressive component.

Co-integration is a complex method of correcting for non-stationarity. A detailed discussion is beyond the scope of this book, but in section 17.7 we provide an intuitive grasp of co-integration.

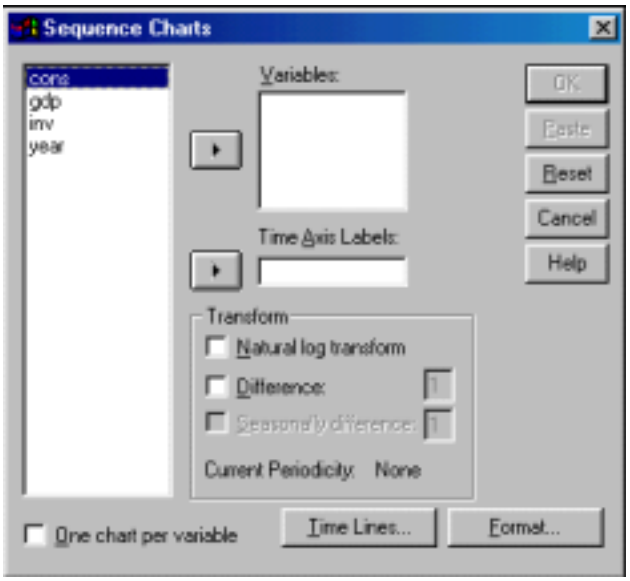
This "autocorrelation" is different from the "autocorrelation in the PACF (section 17.2). There, the autocorrelation being measured is for individual variables – the relation of a variable's value at time "T" to previous values.

Ch 17. Section 1 Sequence charts (line charts with time on the X-axis)

Ch 17. Section 1.a. Graphs of the 'level" (original, untransformed) variables

Go to GRAPHS/SEQUENCE.

Though you can use GRAPHS/LINE to make similar graphs, you should use GRAPHS/SEQUENCE because of the Transform options in the latter (you will see these options in action a bit later).

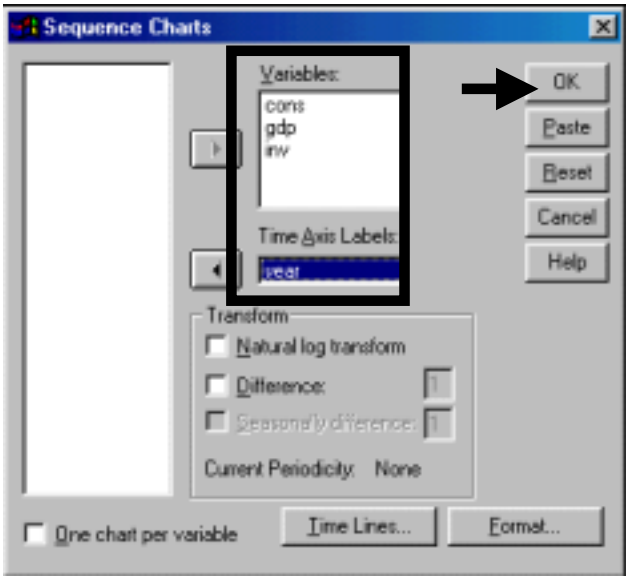


In the box "Variables," place the variables whose values you want to trace over time.

In the box "Time Axis Labels," place the variable whose values define the time dimension (the dimension may be year, month, week, etc.).

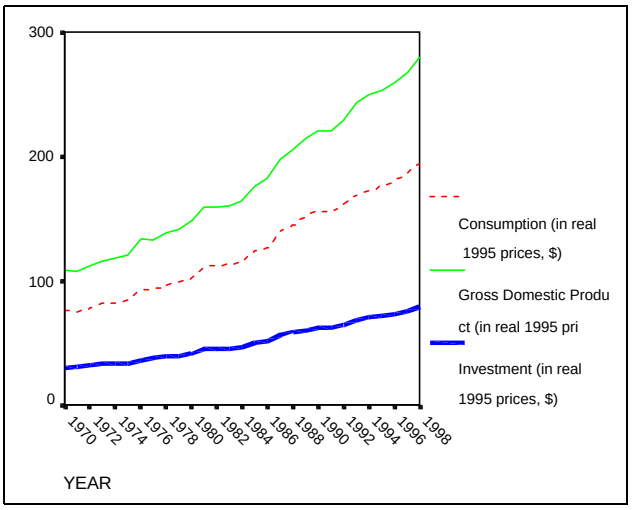
Usually, these are all the steps you will need. Click on "OK."

A multiple-line graph is created.

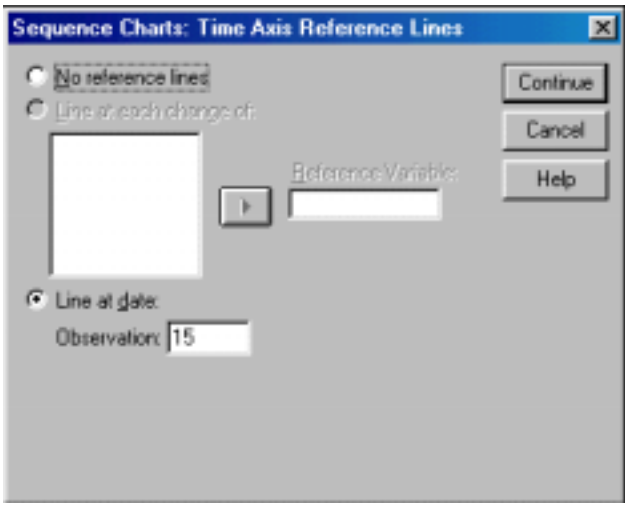


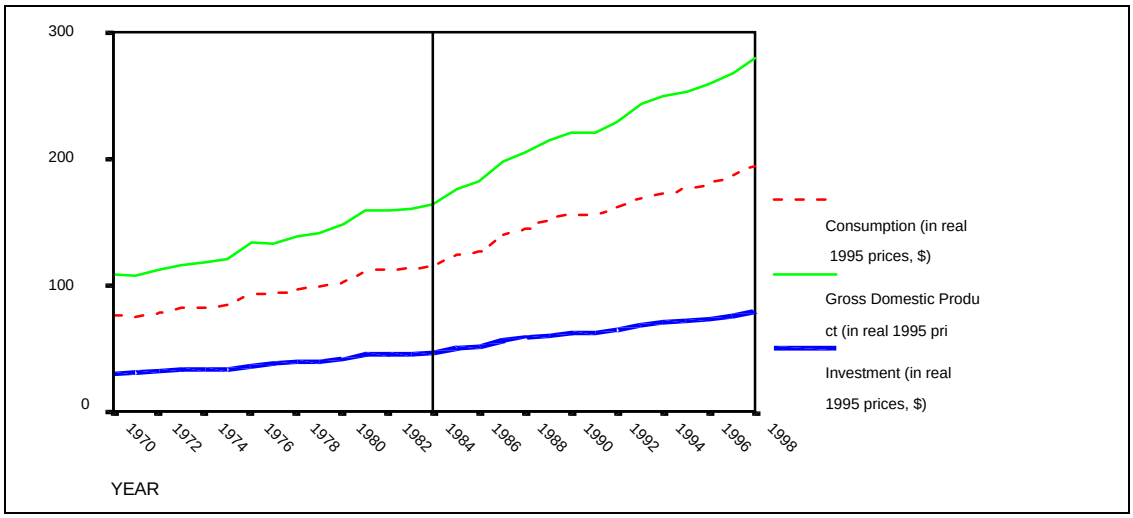
Note: SPSS created all three lines as thin unbroken lines. We changed the format of each line so as to distinguish them better. To do the same, refer to section 11.2.

All the variables increase over time. In a nutshell, this is the main reason why time series variables cannot be used directly in a regression- the value of GDP in 1985 is not truly random because it is based on historical values.

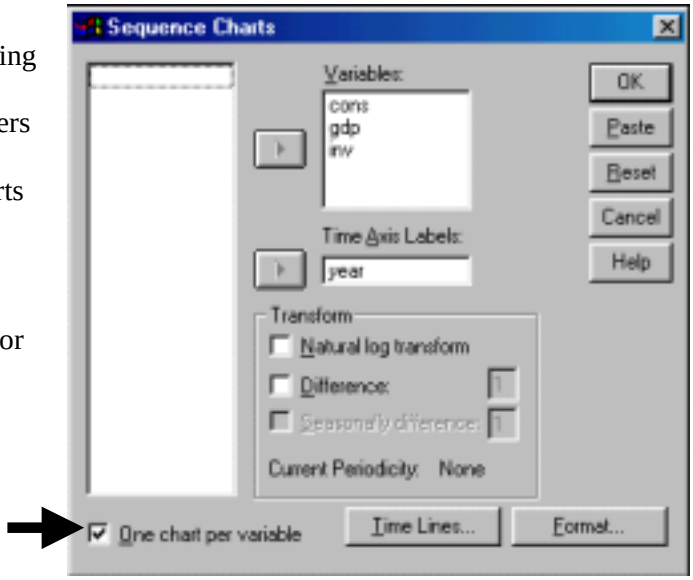


Adding some more information can enrich the above graphs. Let's assume you wanted to break the time dimension into two groups (e.g. - "before policy" and "after policy," or "pre-oil crisis" and "post oil-crisis"). You can do so using the option "Time Lines." We have asked for a "Line at date" corresponding to the 15th time period (in our example this corresponds to 1983). The result is the following graph.

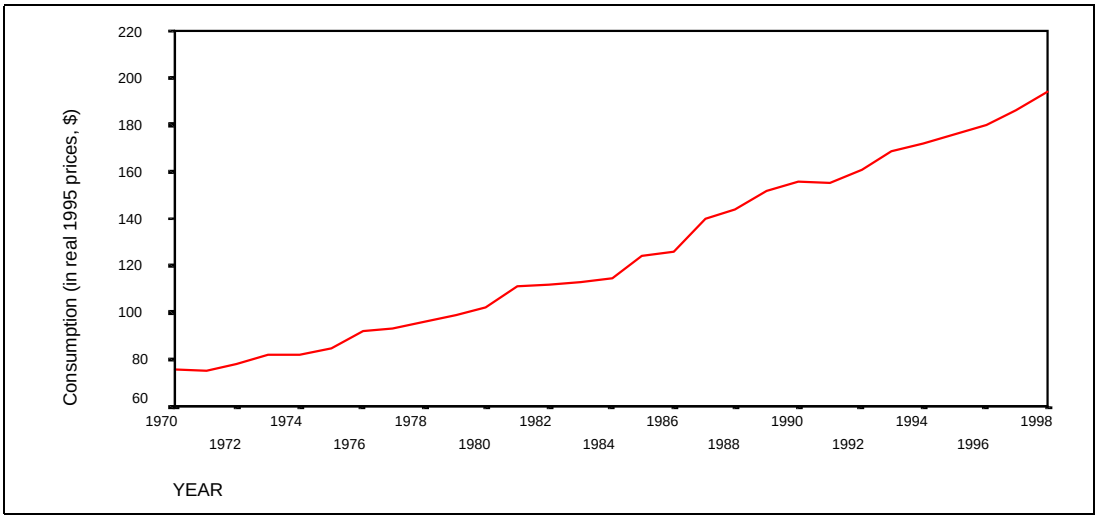


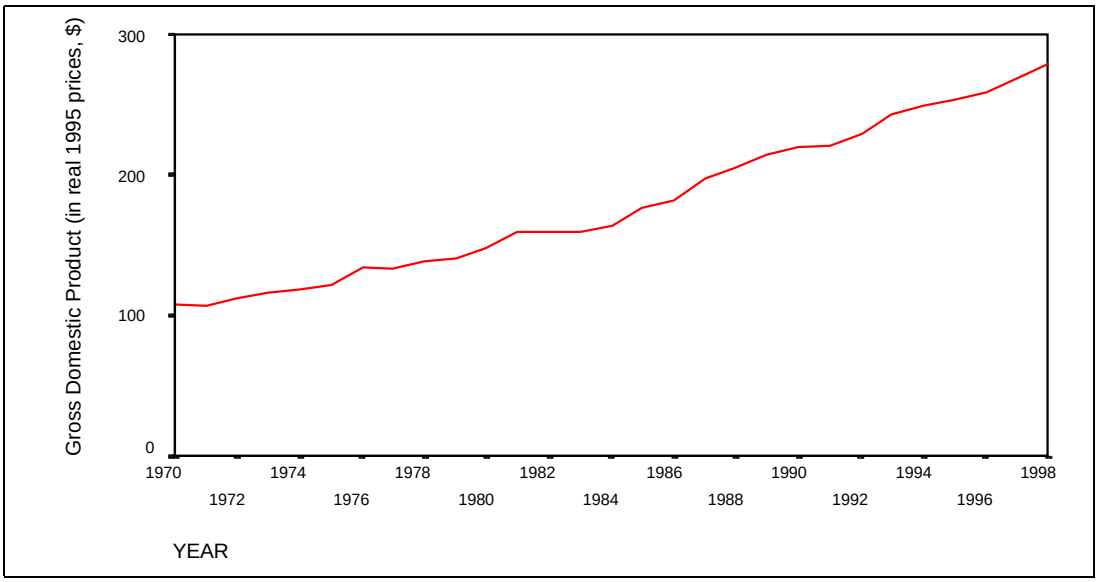


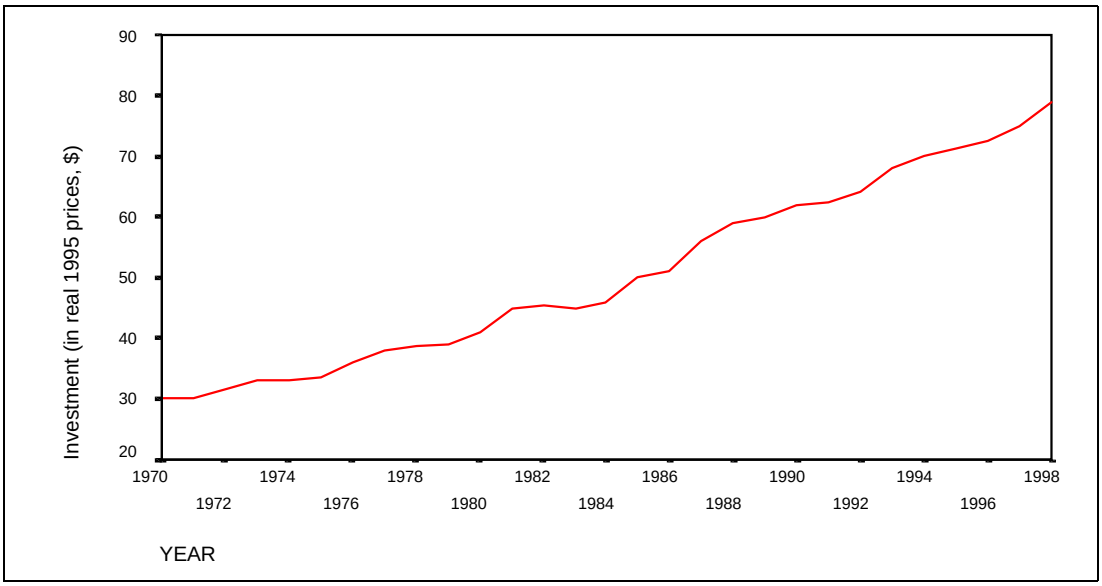
If you have several variables, one chart that displays all of them may be confusing (especially if their units and scales are different, e.g. - some are in dollars, others in yen, or some in millions, others in billions). SPSS can make separate charts for every variable. Choose the option "One chart per variable."



Now separate charts will be produced for each variable.




Gross Domestic Product (in real 1995 prices, $)
300
200
100
0
1970
1972
1974
1976
1978
1980
1982
1984
1986
1988
1990
1992
1994
1996
1998
YEAR


Investment (in real 1995 prices, $)
90
80
70
60
50
40
30
20
1970
1972
1974
1976
1978
1980
1982
1984
1986
1988
1990
1992
1994
1996
1998
YEAR

Ch 17. Section 1.b. Graphs of transformed variables (differenced, logs)

The sequence charts above all show a definite time trend and, using an admittedly simplistic interpretation, non-stationarity. Two methods are usually used to correct for non-stationarity[158]:

- Using the "differenced" version of the variables.

  A "first-differenced" version is the variable obtained by subtracting the value of the variable at time period "T-1" from the value of the variable in time period "T." That is, instead of the observation for 1985 being the "level of GDP in 1985," it is "the difference in the value of GDP in 1985 and the previous period (1984)."

- Transforming the variables into a log format by calculating the log of each variable.

  The logs may reduce the problem because the log scale flattens out the more pronounced patterns.

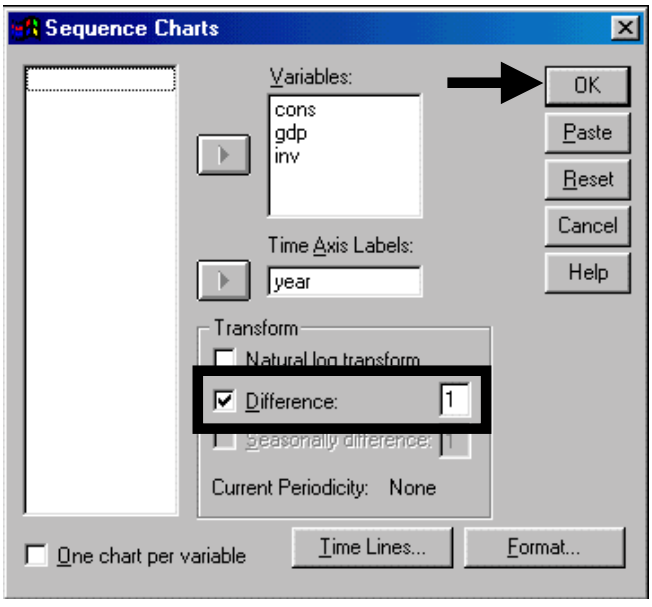
Example 1: Differencing

The first method ("differencing") is more effective than the second method ("logs"). We show examples of each and the results.

As in the example before, go to GRAPHS/SEQUENCE and enter the dialog box as shown. We want to graph the "first-difference" transformations. To do that, choose the option "Difference" and enter the value "1" in the area "Transform."

Click on "OK."

Note: The "second-difference" transformation is graphed by entering the value "2" instead of "1." The second difference for 1985 GDP is the value of 1985 GDP minus the value from two time periods back (1983 in this example).



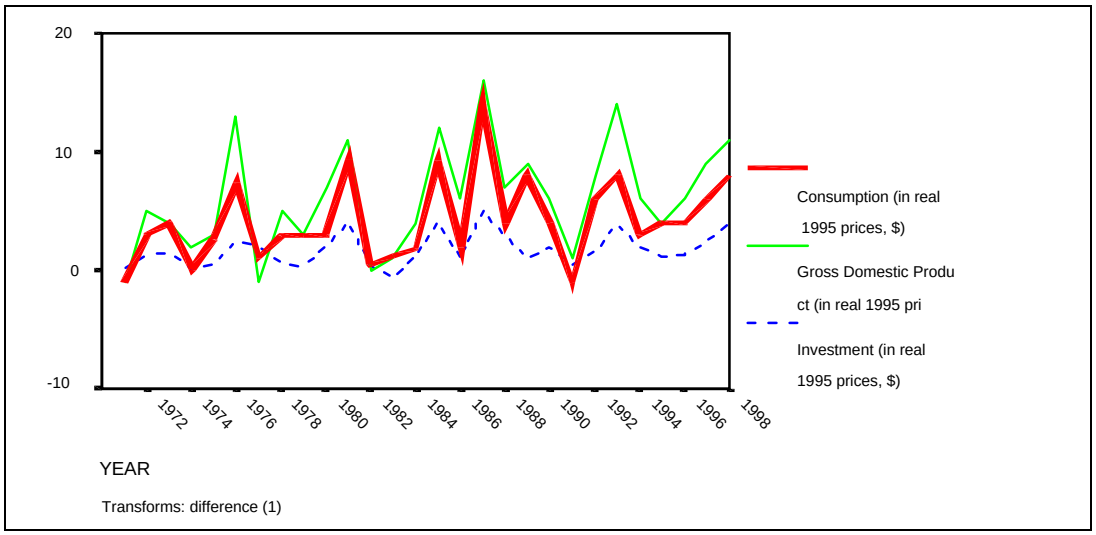
As the graph below shows, the first-

Has the problem of non-stationarity gone? To

[158] Note: Non-stationarity is similar to the Unit Root problem (this admittedly simplistic logic will be used by us in this book).

differences have a zigzag pattern. This indicates a higher possibility of their being "random" as compared to the original "level" variables.

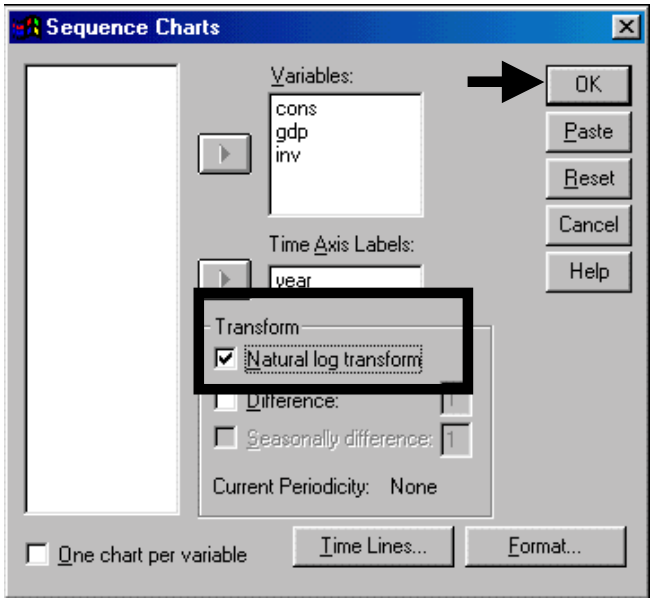
answer this, we must use Partial Auto Correlation Functions (see section 17.2).



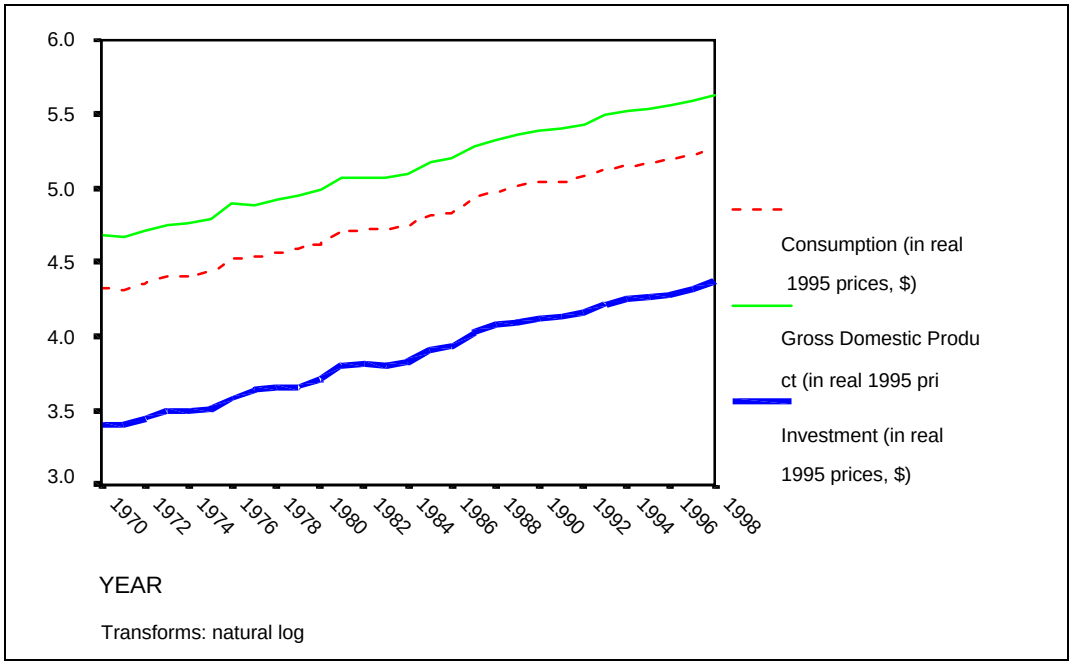
Example 2: Logs

Let us also show an example of using log transformations.

As in the example before, go to GRAPHS/SEQUENCE and enter the dialog box as shown. We want to graph the log transformations. To do that, choose the option "Natural log transform" in the area "Transform." Click on "OK."



The next graph shows that the transformation may have flattened the time trend somewhat, but the pattern over time is definitely not "random." 1985's value is dependent on previous values. If we know the values for 1983 and 1984, we can say that 1985 will be in the same range of values and will probably be higher. If the variables were truly random, we would not be able to make such a claim.



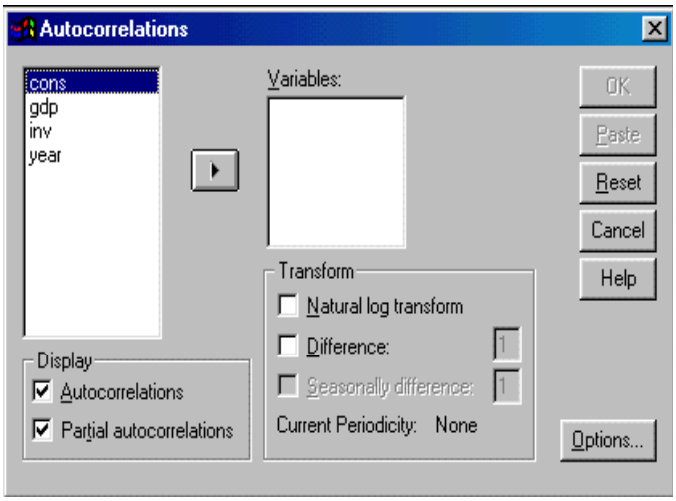
Ch 17. Section 2 "Formal" checking for unit roots / non-stationarity

The real "Formal Tests" include Dickey-Fuller and others. To many macroeconomists, the biggest weakness of SPSS with regards to Time Series analysis is its inability to conduct (directly[159]) formal tests for establishing the presence and nature of non-stationarity. In SPSS's defense, the PACF functions are usually sufficient for determining the presence and nature of non-stationarity. The typical project does not require the use of the more formal tests.

[159] If the test is required, then SPSS can do it indirectly. You must create some new variables, run a regression, and use some special diagnostic tables to interpret the result. All that is beyond the scope of this book.

Ch 17. Section 2.a. Checking the "level" (original, untransformed) variables

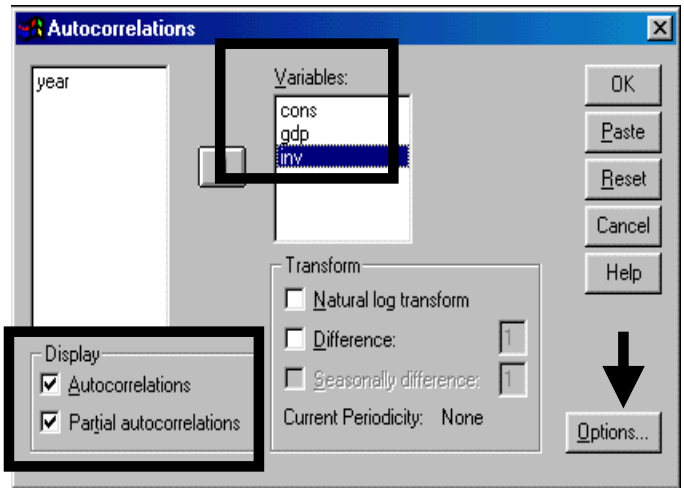
Go to GRAPHS/TIME SERIES / AUTOCORRETIONS.

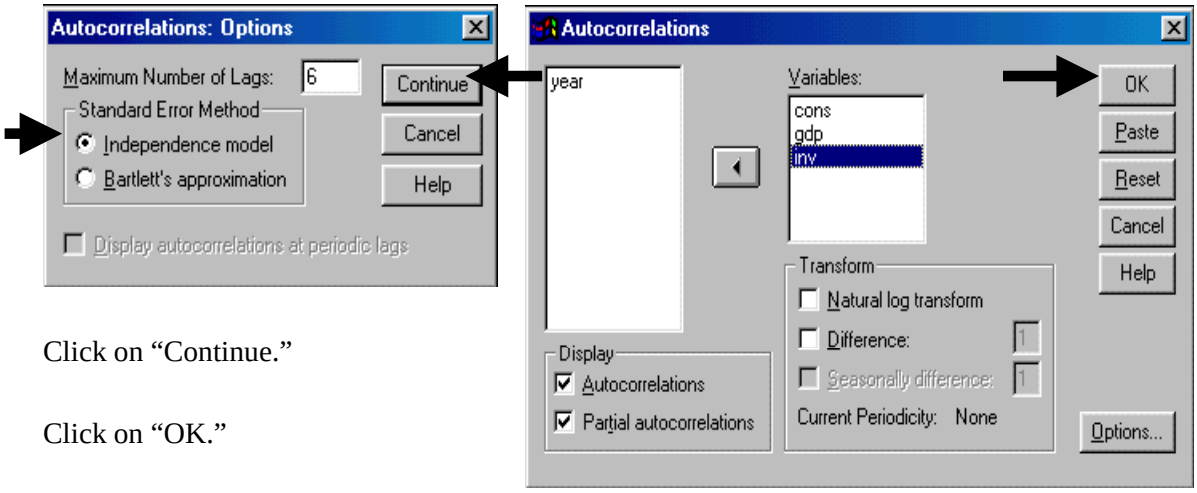


In the box "Variables," place the variables that you wish to check for the presence of non-stationarity (and the nature of this non-stationarity).

Choose "Display" and "Partial Autocorrelation" (to check for non-stationarity and the autoregressive component) and the "Autocorrelation."

Click on "Options." Choose the number of lags (with annual data a number like 6 is fine). Choose the "Independence model" method for calculation of the "Standard Errors."





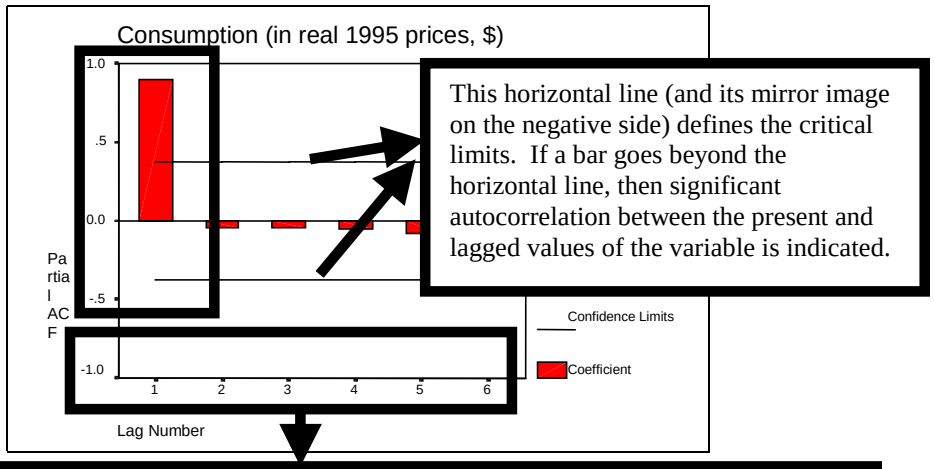
Click on "Continue."

Click on "OK."

Interpretation of the charts produced:

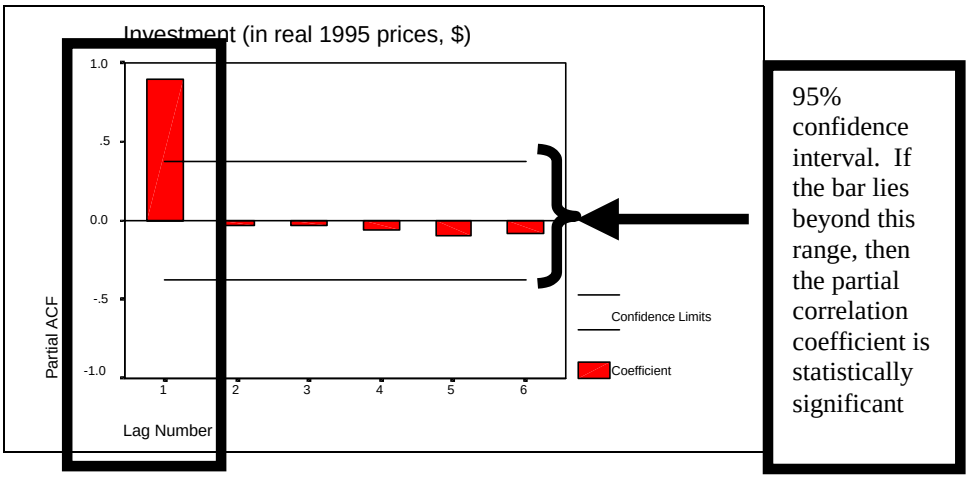
All three variables exhibit non-stationarity **as at least one of the vertical bars is higher than the horizontal line(s) that indicate the cut-off points for statistical significance**. (See next chart to see the bars and the line).

Furthermore, the non-stationarity is of the order "1" **as only the first-lagged bar is significantly higher than the cut-off line. So a first-differenced transformation will probably get rid of the non-stationarity problem** (as was hinted at in the previous section on sequence charts).

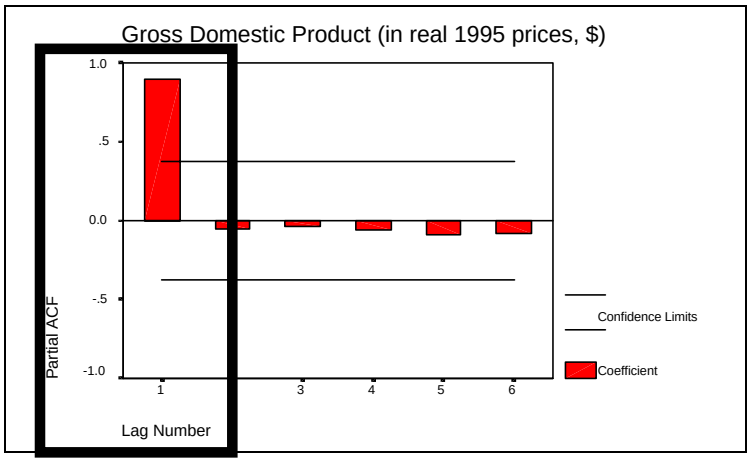


The numbers 1, 2,…6 imply the "Partial Correlation between the value of the variable today (that is at time "T") with the value at time "T-1," "T-2,"…"T-6" respectively.

The first-lag partial auto-correlation is above the critical limit. This indicates the presence of non-stationarity and suggests first-order differencing as the remedy.



The first-lag partial auto-correlation is above the critical limit. This indicates the presence of non-stationarity and suggests first-order differencing as the remedy.



The first-lag partial auto-correlation is above the critical limit. This indicates the presence of non-stationarity and suggests first-order differencing as the remedy.

Autoregression

The second interpretation of the GDP PACF is as follows: In running a regression using *GDP* (or some transformation of it) as the dependent variable, include the 1-lag of the same transformation of *GDP* as an independent variable.

Implications for model specification

So, the model was originally specified as:

GDP_t = a + other variables
becomes (because of the autocorrelation of GDP with its first lag)
GDP_t = a + b*GDP_{t-1} + other variables

This is called an ARIMA(1,0,0) model. ARIMA (1 lag used for autoregression, differencing levels required=0, moving average correction=0)

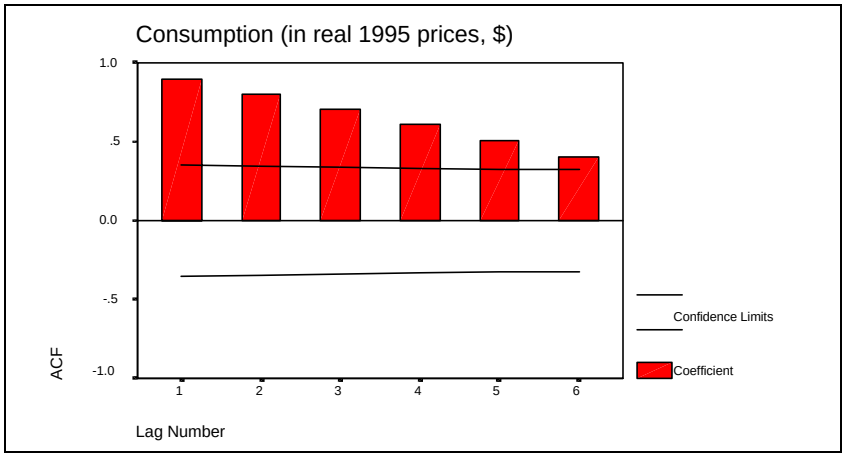
What if the PACF showed that the first three lags were significant? Then the model would become:

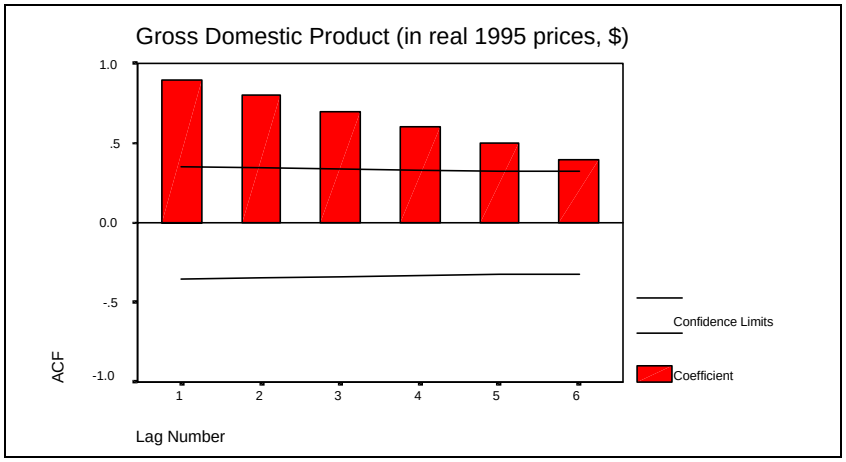
GDP_t = a + b*GDP_{t-1} + p*GDP_{t-2} + r*GDP_{t-3} + other variables

This would be an ARIMA(3,0,0) model. ARIMA(3 lags used for autoregression, differencing levels required=0, moving average correction=0)

The ACF

Two of the autocorrelation function (ACF) charts are shown below. What is the difference between the PACF and ACF? A simple intuitive explanation: "the PACF for lag 3 shows the correlation between the current period and 3 periods back, disregarding the influence of 1 and 2 periods back. The ACF for lag 3 shows the combined impact of lags 1, 2 and 3." In our experience, the PACF gives a clear indication of the presence of "non-stationary" and the level of differencing required. The ACF (along with the PACF) are used to determine the Moving Average process. The Moving Average process is beyond the scope of this book.





Ch 17. Section 2.b. The removal of non-stationarity using differencing and taking of logs

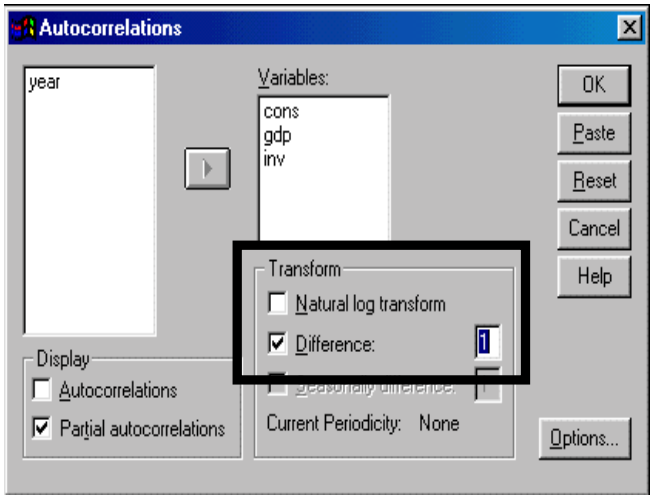
The partial auto-correlation charts have shown that first-differenced transformations do not have the problem of non-stationarity. So, we should use the first difference transformations (that is, new variables created using "first-differencing" in any regression).

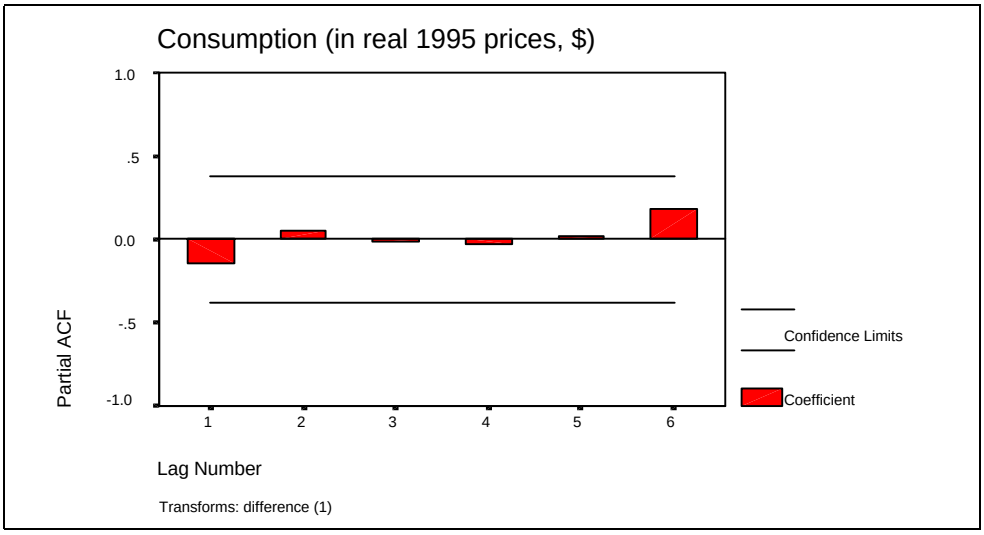
Example 1: First Differenced variables

This time, follow all the procedures from example 1, but choose the option "Difference" in the area "Transform."

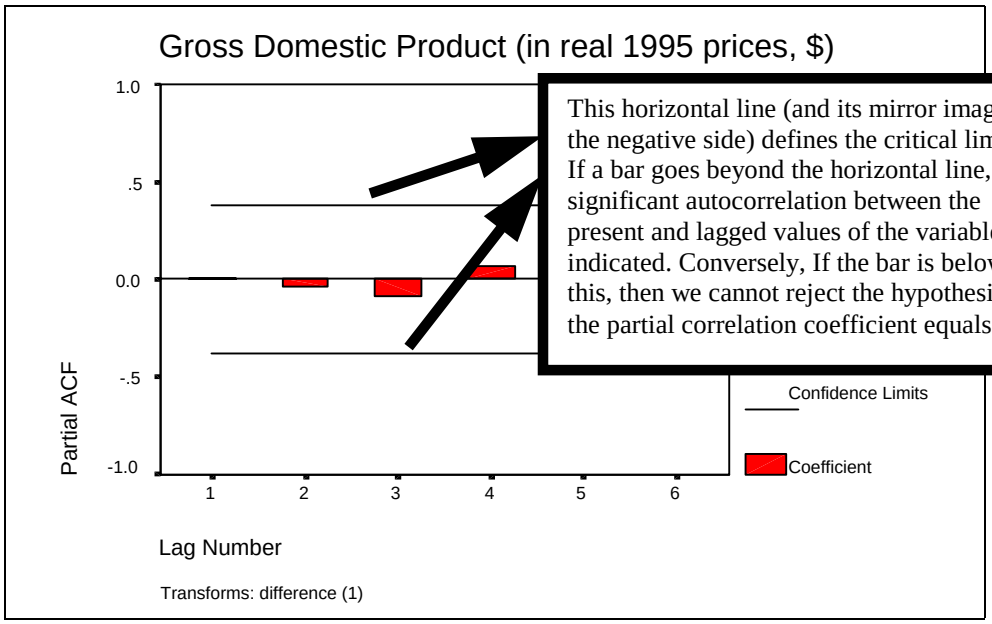
Click on "OK."

Note: In all the graphs below the variable being used is not the "level" but the first difference.

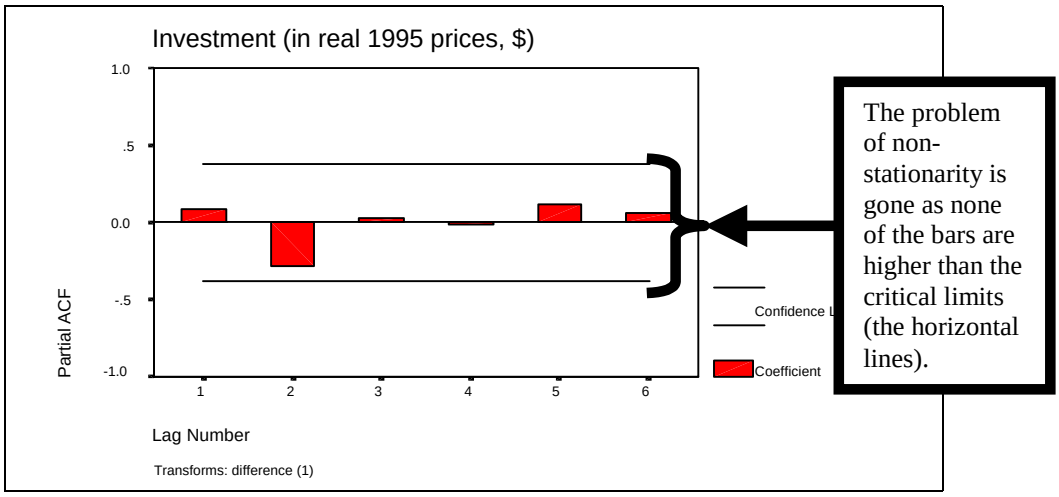




Interpretation: None of the partial-autocorrelation coefficients are above the critical limit. This indicates the absence of non-stationarity and strongly indicates the use of first-order differenced transformations of this variable in any regression analysis.



Interpretation: None of the partial-autocorrelation coefficients are above the critical limit. This indicates the absence of non-stationarity and strongly indicates the use of first-order differenced transformations of this variable in any regression analysis.



Interpretation: None of the partial-autocorrelation coefficients are above the critical limit. This indicates the absence of non-stationarity and strongly indicates the use of first-order differenced transformations of this variable in any regression analysis.

Implications for model specification

First order differencing eliminated the non-stationarity. This indicates that the regression model should be re-specified in terms of differences. The equation in 17.2.a. was an ARIMA (1,0,0) model:

$GDP_t = a + b*GDP_{t-1} + c*Inv_t + d*Cons_t$

after differencing, the "correct" model is:

$(GDP_t - GDP_{t-1}) = a + b* (GDP_{t-1} - GDP_{t-2}) + c*(Inv_t - Inv_{t-1}) + d*(Cons_t - Cons_{t-1})$

ARIMA (1,1,0) model. ARIMA(1 lag used for autoregression, differencing levels required=1, moving average correction=0)

What if the PACF showed that second order differencing was required?

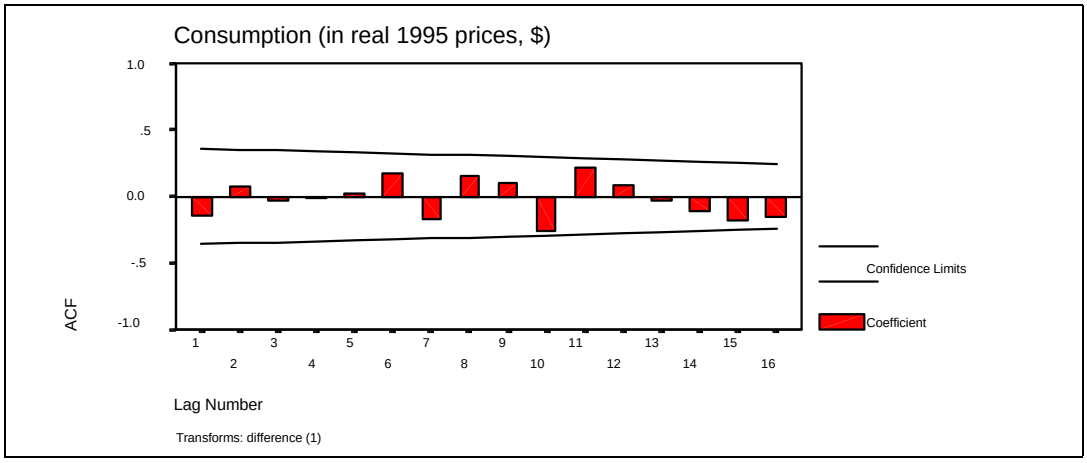
$(GDP_t - GDP_{t-2}) = a + b* (GDP_{t-1} - GDP_{t-3}) + c*(Inv_t - Inv_{t-2}) + d*(Cons_t - Cons_{t-2})$

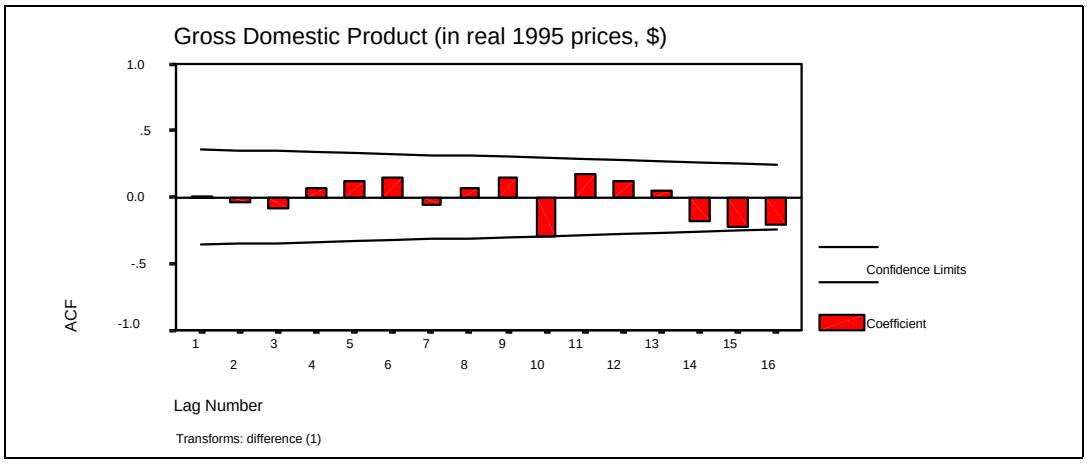
This would be an ARIMA(1,2,0) model. ARIMA(3 lags used for autoregression, differencing levels required=2, moving average correction=0)

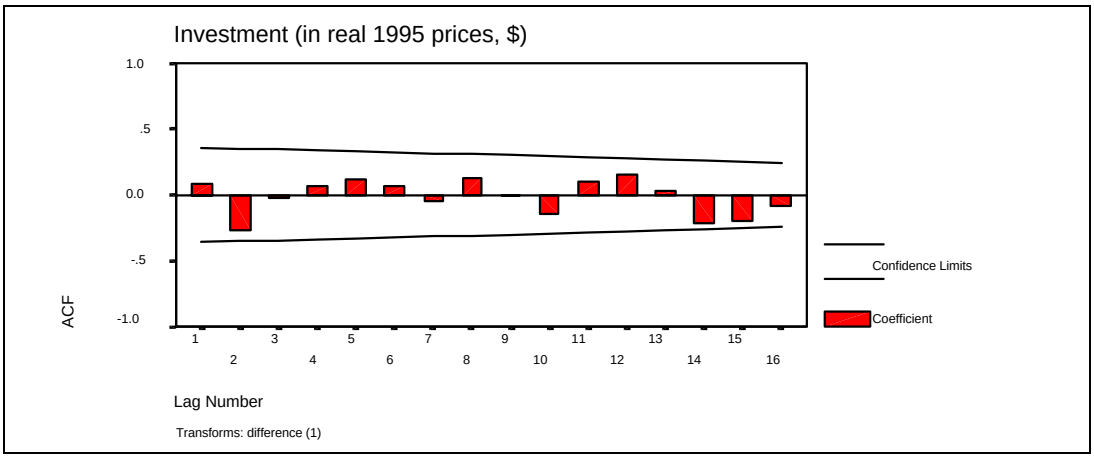
Note: Each entity inside a bracket is the first difference as GDP at time "t" is being differenced with GDP from time "t-1," a 1-period difference.

The ACF

The autocorrelation function (ACF) charts are shown below. What is the difference between the PACF and ACF? A simple intuitive explanation: "the PACF for lag 3, shows the correlation between the current period and 3 periods back, disregarding the influence of 1 and 2 periods back. The ACF for lag 3 shows the combined impact of lags 1, 2, and 3." In our experience, the PACF gives a good clear indication of the presence of "non-stationarity" and the level of differencing required. The ACF (along with the PACF) are used to determine the Moving Average process. The Moving Average process is beyond the scope of this book.

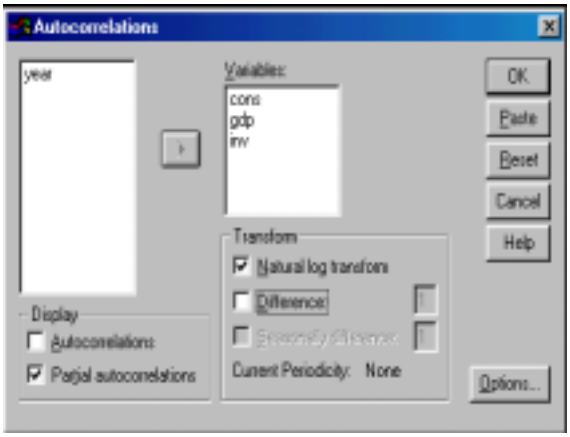






Example 2: Using a log transformation

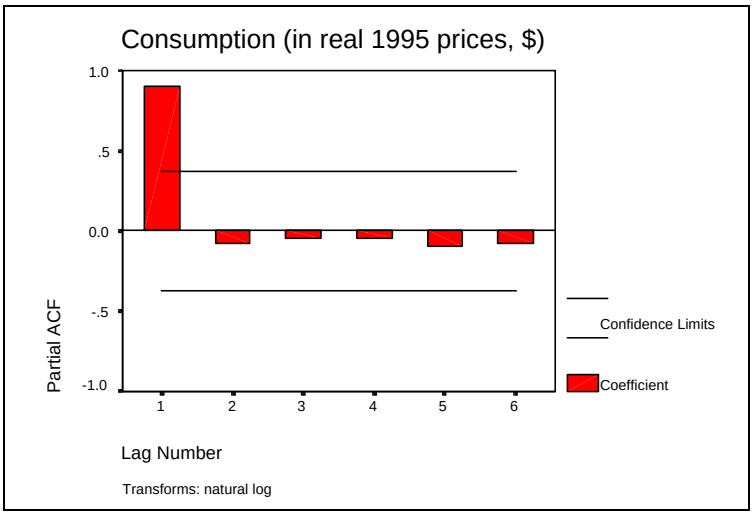
This time, follow all the procedures from example 1, but choose the option "Natural Log Transform" instead of "Difference."



Note: You can use both together.

Click on "OK."

We show only one graph.



The first-lag partial autocorrelation is above the critical limit. This indicates the presence of non-stationarity and hints in disfavor of the use of logs as the remedy.

Ch 17. Section 3 Determining lagged effects of other variables

Investment may take a few years to have a major impact on the GDP. Say, the investment is in transportation infrastructure. The building of the infrastructure may take several years - each year some money is invested into the infrastructure. But the infrastructure cannot be used until it is completed. So, although investments may take place in 1981, 1982 and 1983, the major impact on GDP will not be felt until 1984 when the infrastructure is being used. In such a situation, investment from previous periods is having an impact on today's GDP. If we run a regression in which today's GDP is the dependent variable, then we have to first correctly specify the investments of which period(s) should be explanatory variables. Cross-correlations help with this.

The question whose answer we seek is: Should the model be:

$$Y_t = a + b\, Y_{t-1} + \mathbf{X_t}$$

or,

$$Y_t = a + b\, Y_{t-1} + \mathbf{X_{t-1}}$$

or,

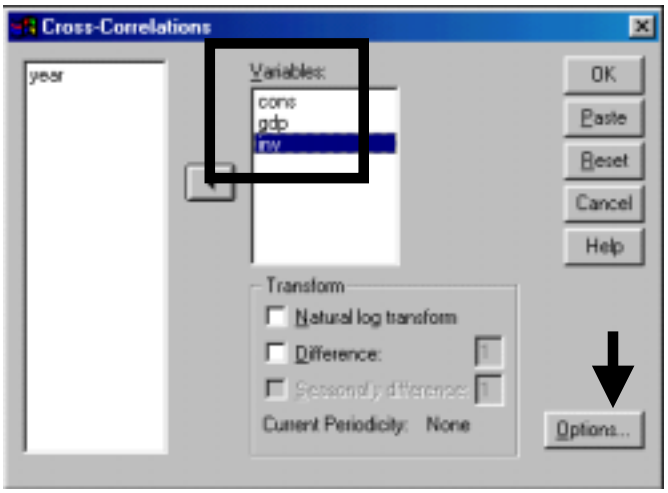
$$Y_t = a + b\, Y_{t-1} + \mathbf{X_{t-2}}$$

or,

$$Y_t = a + b\, Y_{t-1} + X_t + \mathbf{X_{t-1}} + \mathbf{X_{t-2}}$$

Cross-correlations help us decide which lagged transformation(s) of an independent variable should be used.

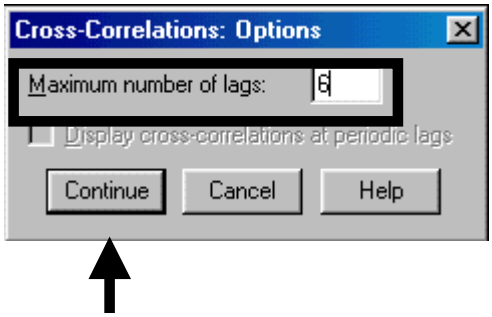
Go to GRAPHS/TIME SERIES /CROSS CORRELATIONS.



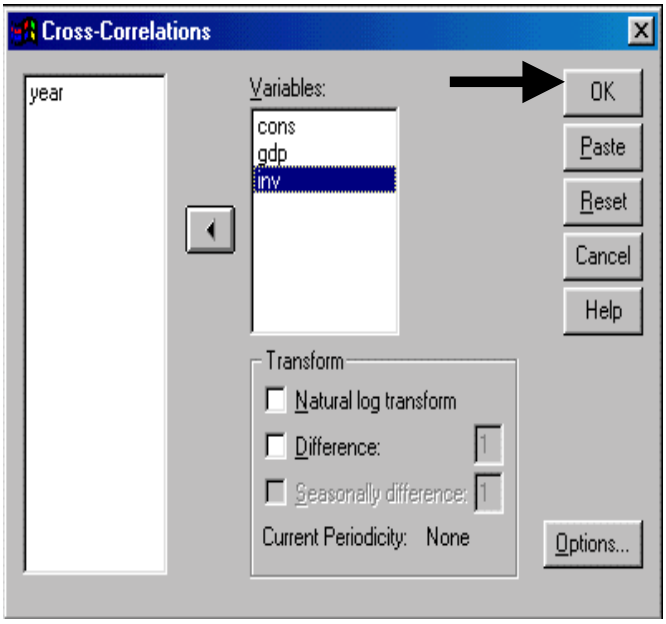
Into the area "Variable," place the variables between whose pairs you want to check for cross-correlation. Note: Each CCF graph will show the cross-correlation function for one pair.

The difference between "cross" and simple "(auto) partial" correlation is that the former looks at the correlation between the value of a variable at time T with the values of another variable (or Investment) at times T-1, T-2, etc. In contrast, the autocorrelation function looks at the correlation between the value of a variable at time T with the values of the same variable at times T-1, T-2, etc.

Click on the button "Options." Choose the maximum number of lags for which you wish to check for cross-correlations. This number will depend on the number of observations (if sample size is 30, choosing a lag of 29 is not useful) and some research, belief, and experience regarding the lags. For example, if *investment* flows take up to 4 years to have an impact on *GDP*, then you may choose a number greater than 4 but not too large. If the impact of *consumption* is felt in the current and following year, then a smaller number of lags can be chosen. Click on "Continue."

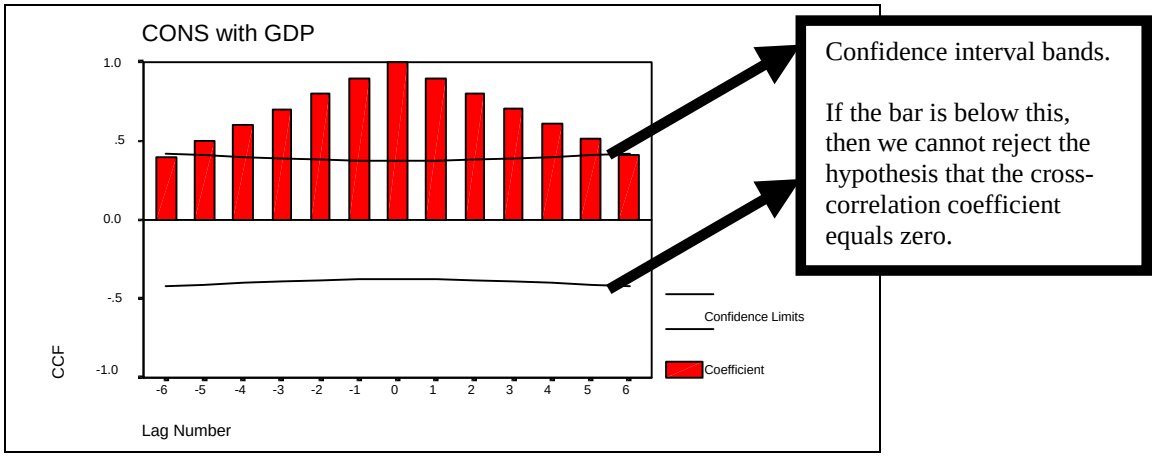


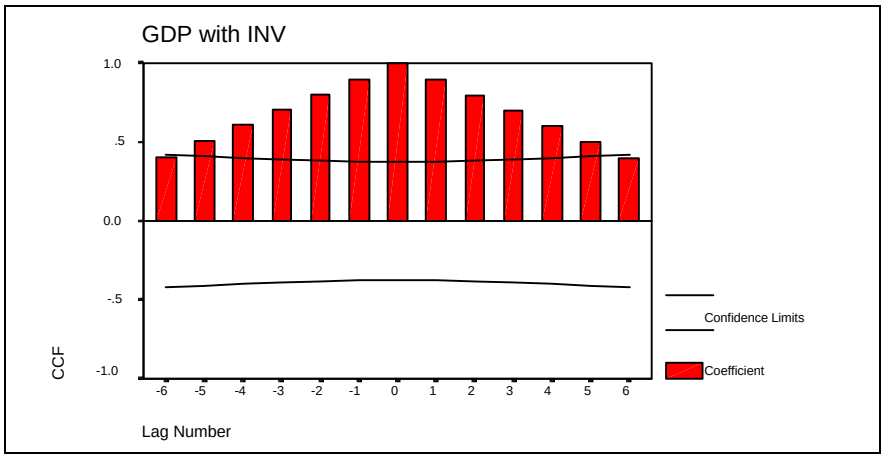
Click on "OK."



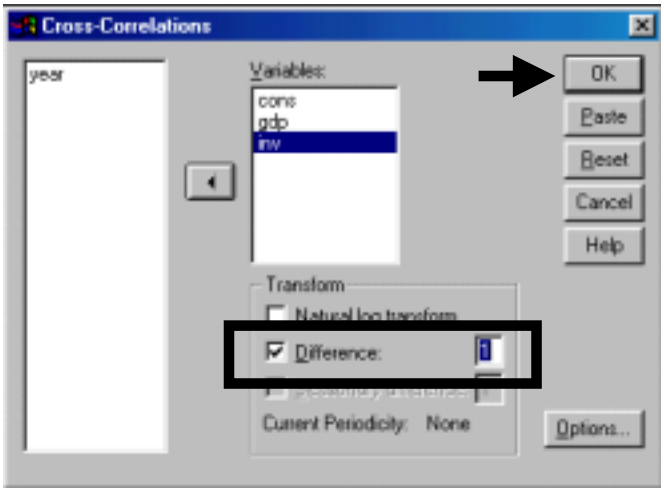
Three charts are produced - one for each pair derived from the three chosen variables. Our aim is to see the cross-correlations between the dependent variable (*GDP*) and the independent variables. So we have deleted the chart that showed the correlation between *CONS* and *INV*, the two independent variables. So you see only two charts.

In both of them, several of the bars are above the horizontal confidence limit line, indicating cross-correlation with 5 lags and 5 leads. If we presume this is correct, then the regression will have to incorporate [at the minimum] five terms for lagged *consumption*, 5 for lagged *investment*, and one each for today's *consumption* and *investment.*





We know from the PACF that the "level" variables are non-stationary, but their "first-differenced" transformations are not. As such, we will use only the "first-differenced" versions in any model.

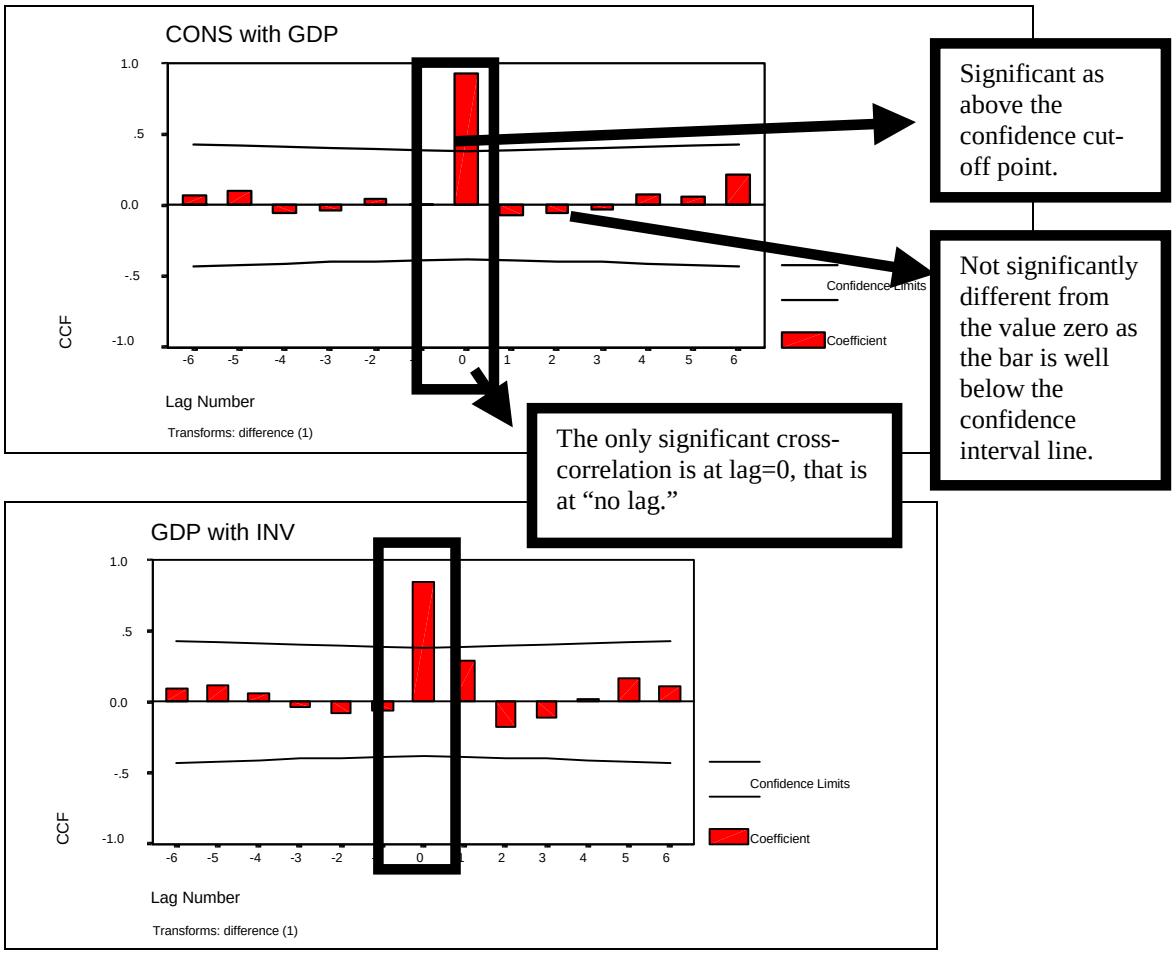


Consequently, we need to know about cross-correlations **only** when the "first-differenced" variables are used.

To do so, repeat other steps from the previous example, and then choose "Difference" and type "1" into the box to the right of it. Choose the same options and variables as before.

Click on "OK."

Now only the cross-correlation for lag zero is significant. This implies that, for both consumption and investment, the transformations that should be used are the "first-differenced, no-lag."



Now we are ready to create the new variables deemed important by the ACF, PACF, and CCF. Once they have been created in 17.4, the ARIMA regression can be run (17.5).

Implications for model specification

The result is saying to use first differenced transformations (from the ACF/PACF) with an autoregressive component of 1 lag (from the PACF), and with no lagged *cons* or *inv*.

The equation we had obtained after 17.2 was:

$(GDP_t - GDP_{t-1}) = a + b^* (GDP_{t-1} - GDP_{t-2}) + c^*(Inv_t - Inv_{t-1}) + d^*(Cons_t - Cons_{t-1})$

Because the cross-correlations for the first differenced observations showed a correlation only at the 0 lag, there is no change to the model.

But, if the cross-correlation between GDP and the first and third lags of investment were significant, then the model would be:

$(GDP_t - GDP_{t-1}) = a + b^* (GDP_{t-1} - GDP_{t-2}) + \mathbf{c^*(Inv_t - Inv_{t-1}) + e^*(Inv_{t-1} - Inv_{t-2}) + f^*(Inv_{t-3} - Inv_{t-4})} + d^*(Cons_t - Cons_{t-1})$

It is an ARIMA (1,1,0) model. ARIMA (1 lag used for autoregression, differencing levels required=1, moving average correction=0). The cross-correlation does not figure explicitly into the ARIMA. You have to create new variables (see next section).

Where, $\mathbf{(Inv_{t-1} - Inv_{t-2})}$ is a 1-lagged first difference, and

$\mathbf{(Inv_{t-3} - Inv_{t-4})}$ is a 3-lagged first difference, and

$\mathbf{(GDP_{t-1} - GDP_{t-2})}$ is a 1-lagged first difference autoregressive component (because it is the lag of the dependent variable; in a sense, we are regressing GDP on itself—thus the term "auto." And,

$\mathbf{(Inv_t - Inv_{t-1})}$ is a 0-lagged first difference, and

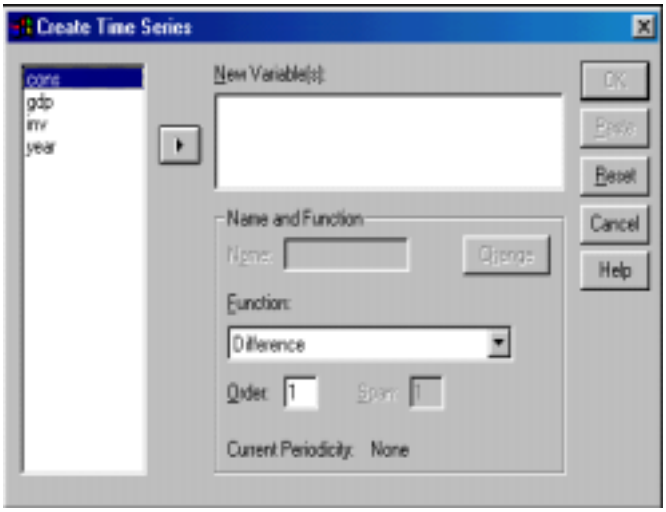
$\mathbf{(Cons_t - Cons_{t-1})}$ is a 0-lagged first difference.

Ch 17. Section 4 Creating new variables (using time series specific formulae: difference, lag, etc

The Sequence charts and PACF charts show us that the easiest and most efficient way to get rid of the problem of unit roots/non-stationarity was to use the "first-differenced" transformations of the original variables. This section explains how to create these new variables.

In chapter 2 we learned how to create new variables using COMPUTE, RECODE, and some other procedures. There is a simpler way to create time series transformed variables. Using TRANSFORM/CREATE TIME SERIES, several variables can be transformed by a similar mathematical function. This makes sense in Time series analysis because you will often want to obtain the "differenced" transformations on many of the variables in your data.

Go to TRANSFORM/CREATE TIME SERIES.

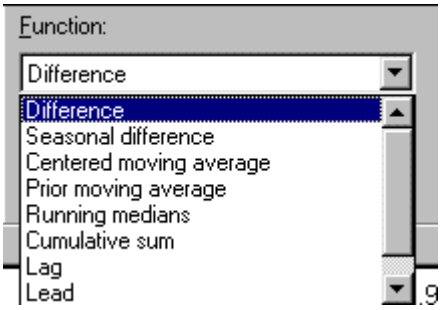


The area "New Variable(s)" contains the formula for each new variable.

The area "Function" allows you to choose the function used for creating the variables listed in "New Variable(s)."

The box "Order" allows you to customize the time gaps used for differencing/lagging/leading, etc

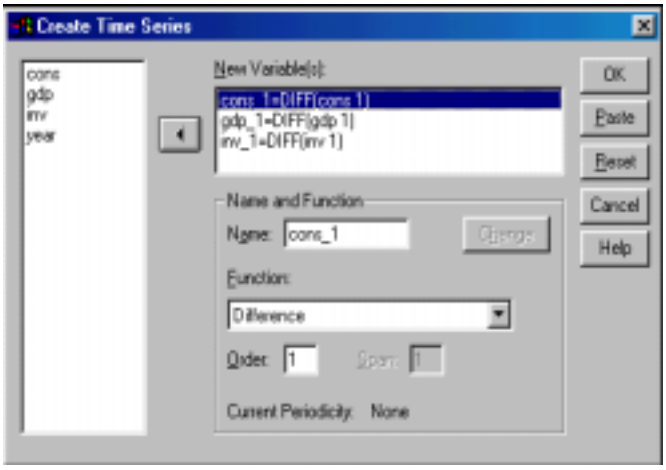
Let us look a bit closer at the options in the area function.



Choose the function "Difference."

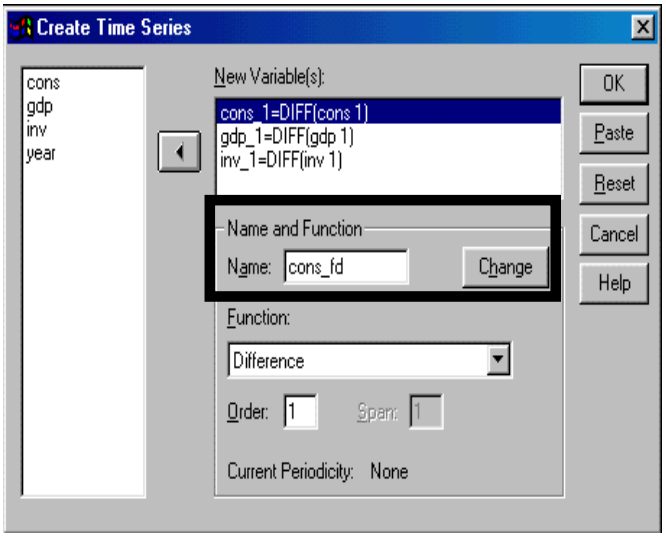
The most used functions are "Difference" and "Lag." The former will create a new variable whose value for 1985 is the value of GDP in 1985 minus the value of GDP in 1984. The latter ("Lag") will create a new variable whose value for 1985 is the value of GDP for 1984.

We want to obtain the first difference of the three variables of which we made PACFs and sequence charts (in the previous two sections). Move these variables into the box "New Variable(s)."



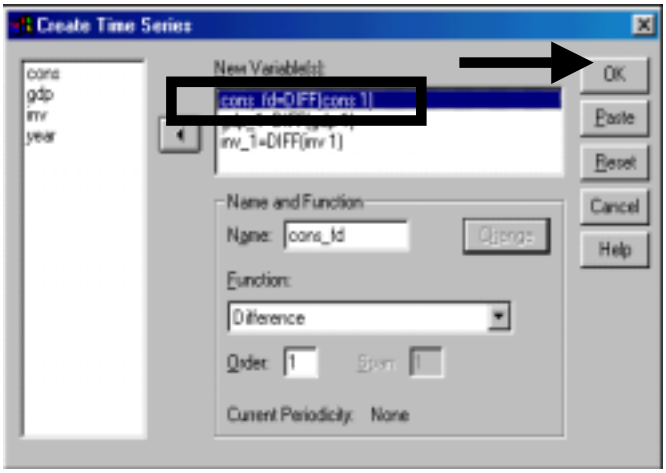
Look at each item on the list. Each item is an equation. The left-hand side variable ("cons_1," automatically named by SPSS) is the new variable. It is equal to the difference formula ("=DIFF()") applied to the original variable ("=DIFF(cons 1)") with a 1st-order of differencing. (The "1" after "cons" in the DIFF() formula indicates that first order differencing is taking place).

You may want to choose more intuitive names for the new variables. Assume that instead of "cons_1" you choose "cons_fd" (where the "fd" stands for "first-difference"). To do that, type the new name into the box "Name" and click on "Change."



The name has been changed.

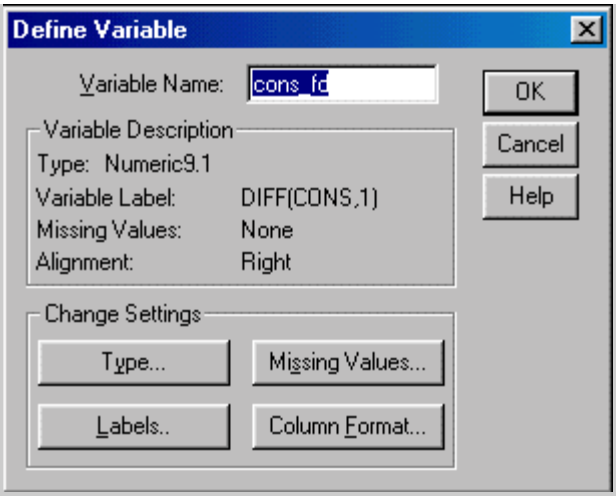
You can do the same for the other variables also. After doing that click on "OK."



Note: Let's assume you want the "second difference" for one variable, the "lag" for the second, and the "third difference for the third. The way to do this is to click on the original function in the box "New variable(s)." Go to the area "Function" and change the function to "Difference" or "lag" as appropriate. Enter the order of differencing (so, for third differencing, enter the number "3" into the box "Order"). Click on "Change." Repeat for the next variable.

After creating the new variables, always go to DATA/DEFINE and define <u>each</u> new variable. At the minimum, define a good "Value Label."

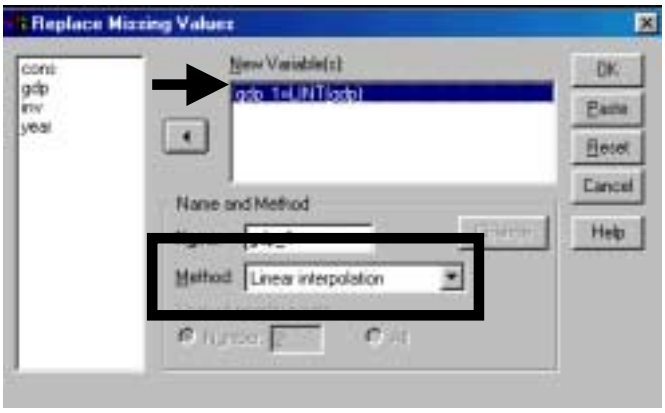
The new variables that are created using functions like "Difference" or "Lag" will have less observations than that of the original variables. With "First-Order" you will lose 1 observation, with Second-Order 2 observations, etc. See the picture below for an illustration.



	gdp	cons	inv	cons_fd	gdp_fd	inv_fd
1	108.0	76.0	30.0	.	.	.
2	107.0	75.0	30.2	-1.0	-1.0	.2
3	112.0	78.0	31.5	3.0	5.0	1.3

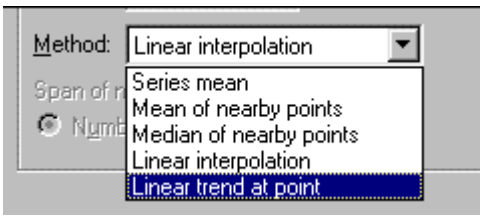
Ch 17. Section 4.a. Replacing Missing values

You can use "Linear Interpolation/Trend" to replace the missing values in a time series variable, if you believe that the variable follows a linear trend in the time vicinity of the missing values.



To do so, go to DATA/REPLACE MISSING VALUES. Choose the Method for replacing values and move the variable with the missing values into the area "New Variables."

Alternate methods are shown on the right.



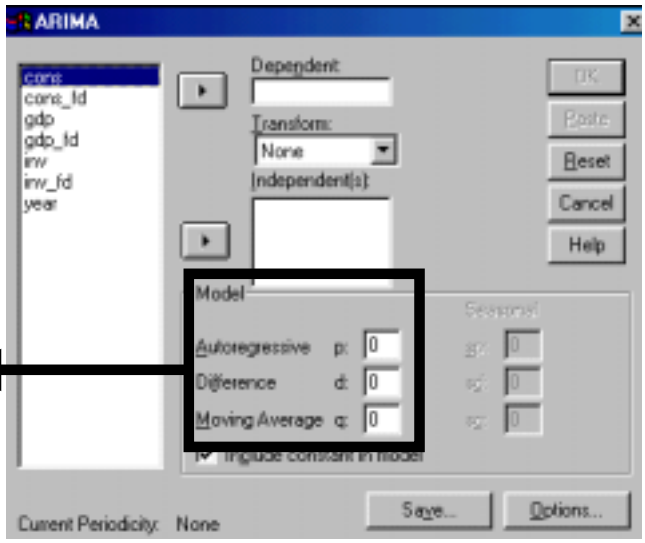
Press "OK."

Ch 17. Section 5 ARIMA

Now we are ready to conduct a linear regression using our time series data set. The most often used method is the ARIMA ("Auto Regressive Integrated Moving Average"). The method permits one to run a linear regression with the added provisos:

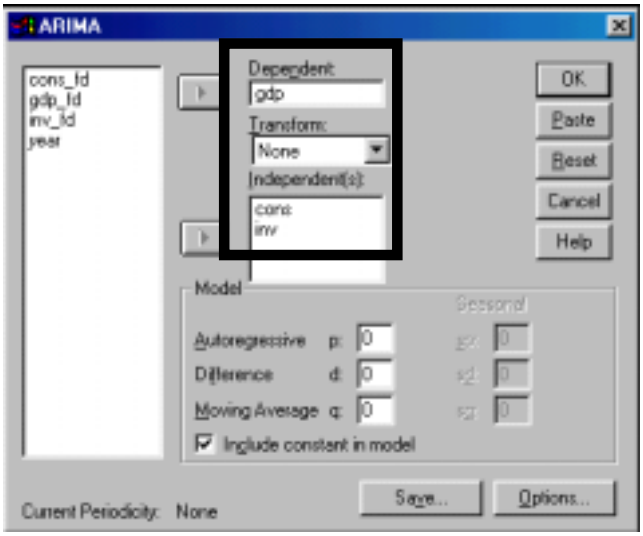
- Autoregressive (or lagged) components can be included as independent variables. For example, for a regression in which (differenced) GDP is the dependent variable and we believe that the differenced GDP from one period back is a causal factor behind the value of this period's difference, then a one order autoregressive component should be included.
- The regression can use "differenced" versions of the original variables.
- A moving average correction can be included. (Note: Moving Average is beyond the scope of this book).

Go to STATISTICS/TIME SERIES /ARIMA.



ARIMA is often written as ARIMA (p,d,q) with the "p," "d," and "q" being specified.

Choose the dependent and independent variables as shown on the right.



In 17.3 you learned how to determine if lagged values of the independent variables should be used in the ARIMA. Suppose you found that you needed a lag (say "1-lag of first differenced cons"). To incorporate this lagged variable, you must first create the lagged variable using methods shown in section 17.4. Then, use this new variable as an independent variable in the ARIMA.

Note: If you want to estimate the coefficient of the "Time Trend," then include year as an independent variable.

Choose the level of differencing. We have chosen "1" (based on the results in sections 17.1.b and 17.2.b). Now the regression will use the differenced versions of each variable in the equation.

Note: Do not use the differenced variables you created in the previous section. Use the original "level" variables as the independent variables. **We think SPSS does the conversion automatically.**

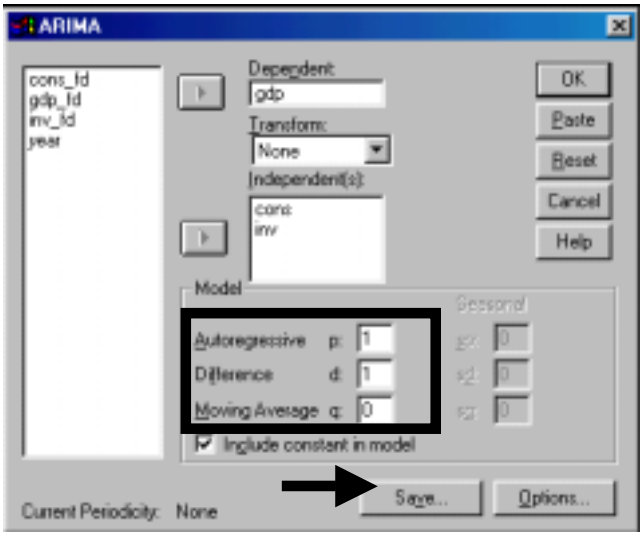
Choose the autoregressive order. Choosing one implies that a 1-period lagged variable of the differenced GDP will also be used as an independent variable.

Click on "Save."

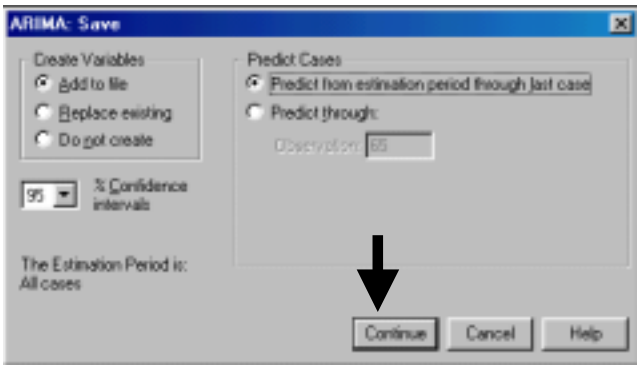
Choose as shown.

Click on "Continue."

The predicted variables will be predictions of the original "level" variables, not of the differenced or lagged variables.

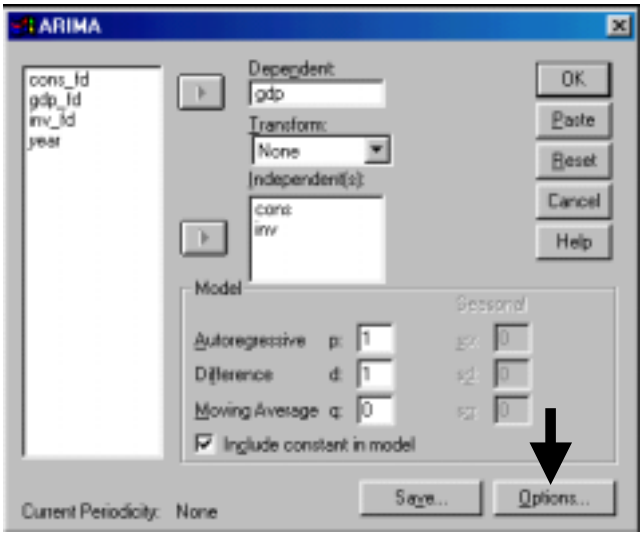


Note: Moving Average correction is beyond the scope of this book.

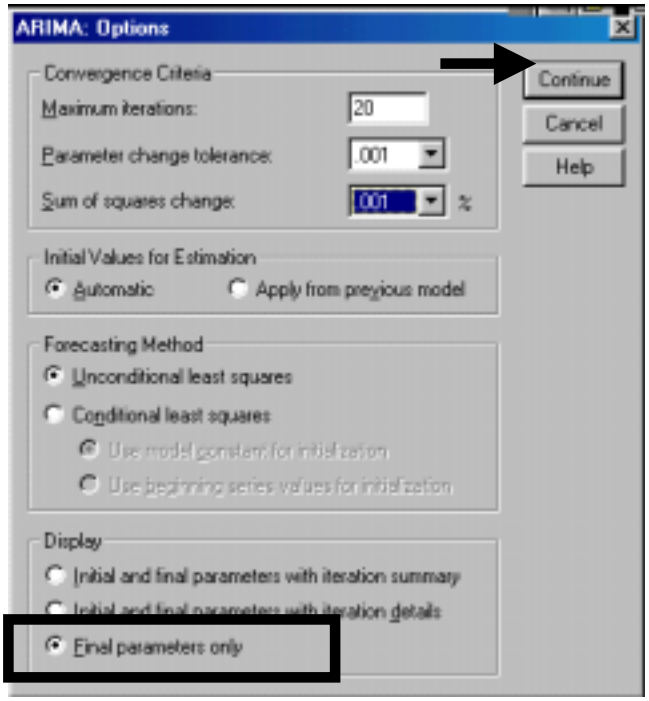


If you want (and have been taught the theory and use of Moving Averages, then enter a value for the moving average).

Click on "Options."



Choose as shown.



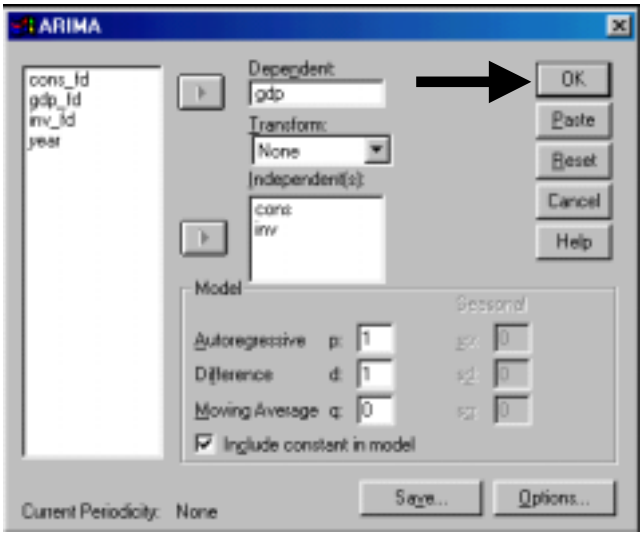
SPSS uses Maximum Likelihood Estimation (MLE) method for estimation of the ARIMA. MLE runs an algorithm several times, using as the starting point the solution obtained in the previous iteration/run. Basically SPSS is maximizing the value of a function by choosing the set of coefficient estimates that would maximize the function. Each time, it uses the estimates obtained in the previous iteration/run. We have asked SPSS to run the iteration a maximum of 20 times. Usually a solution would be found within 20 iterations.

The "Parameter change tolerance" and "Sum of squares change" provide criteria at which SPSS should stop running more iterations. You can make these criteria stricter by choosing lower numbers for these two. (Explanations of these criteria are beyond the scope of this book).

Note: Choose "Final parameters only" in the area "Display." You are interested only in the result of the final iteration/run of the algorithm.

Click on "Continue."

Click on "OK."



Note: Maximum Likelihood Methods are used.

Note: The ARIMA in SPSS does not have an option for correcting for autocorrelation between residuals. But, because SPSS uses Maximum Likelihood Estimation rather than Linear Regression for the ARIMA, violations of the classical assumptions can probably be ignored. Remember that the classical assumptions are relevant for linear regressions estimated using the Least Squares approach.

```
MODEL:  MOD_23

Split group number: 1  Series length: 29
No missing data.
Melard's algorithm will be used for estimation.

Conclusion of estimation phase.
Estimation terminated at iteration number 2 because:
Sum of squares decreased by less than .001 percent.

FINAL PARAMETERS:

Number of residuals   28
Standard error        1.4625021
Log likelihood        -48.504027
AIC                   105.00805
SBC                   110.33687

            Analysis of Variance:

               DF  Adj. Sum of Squares    Residual Variance

Residuals      24            51.888236            2.1389124

           Variables in the Model:

                     B         SEB      T-RATIO   APPROX. PROB.

AR1          -.5096389   .17882940   -2.8498611      .00884025
CONS         1.0577897   .17338624    6.1007712      .00000266
INV           .6612879   .38730423    1.7074120      .10065195
CONSTANT      .5064671   .37684807    1.3439557      .19153092

The following new variables are being created:
```

Telling us that a solution was found.

Use these only if part of your coursework. The Log Likelihood or, more specifically, the "-2 Log Likelihood," can be used to compare across models. Consult your textbook for details.

This is the autoregressive component

Significant if less than .1 (for 90% and .05 for

```
Name        Label

FIT_2       Fit for GDP from ARIMA, MOD_23 CON
ERR_2       Error for GDP from ARIMA, MOD_23 CON
LCL_2       95% LCL for GDP from ARIMA, MOD_23 CON
UCL_2       95% UCL for GDP from ARIMA, MOD_23 CON
SEP_2       SE of fit for GDP from ARIMA, MOD_23 CON
```

Note: The new variables are predictions (and upper and lower confidence bounds) of the original variables and not of their differenced transformations.

Interpretation of the coefficients: we have to reassure ourselves that our interpretation below is correct

- AR1: for every 1 unit increase in the change of GDP between two and 1 periods back (that is, for example, in *GDP* of 1984 - 1983) the effect on the change in *GDP* between the last period and the current period (that is, for the same example, in *GDP* of 1985 - 1984) is "-.50." If the difference in *GDP* between 1983 and 1984 increases, then the difference between 1984 and 1985 decreases.

- CONS: for every 1 unit increase in the change of *consumption* between the last and current periods (that is, for example, in *consumption* of 1985 - 1984), the effect on the change in *GDP* between the last period and the current period (that is, for the same example, in *GDP* of 1985 - 1984) is "1.05." If the difference in *consumption* between 1983 and 1984 increases, then the difference between 1984 and 1985 decreases.

- INV: note that the T is barely significant at the 90% level (as the Sig value = .10). For every 1 unit increase in the change of *investment* between the last and current period (that is, for example, in *investment* of 1985 - 1984), the effect on the change in *GDP* between the last period and the current period (that is, for the same example, in *GDP* of 1985 - 1984) is "1.05." If the difference in *investment* between 1983 and 1984 increases, then the difference between 1984 and 1985 decreases.

- CONSTANT: not significant even at 90% as the sig value is well above .1.

Ch 17. Section 6 Correcting for first order autocorrelation among residuals (AUTOREGRESSION)

The residuals from a regression using Time Series data suffer from the problem of Autocorrelation if the residuals in time period "T" are a function of residuals in any other time period. Why is this a problem? Quite simply, it violates the classical regression assumption that the residuals are distributed independently from each other and are random. The most often-occurring type of autocorrelation is first order autocorrelation. Stepping away from our avoidance of using equations, a first order autocorrelation implies:

$Residuals_T = a * Residuals_{T-1} + u_T$
(where u_T is truly random and uncorrelated with previous period values)

The coefficient "a" is called "Rho." In 7.1 we showed how to ask for the Durbin Watson statistic (DW) from a linear regression. The easiest way to estimate "Rho" is by using this statistic.

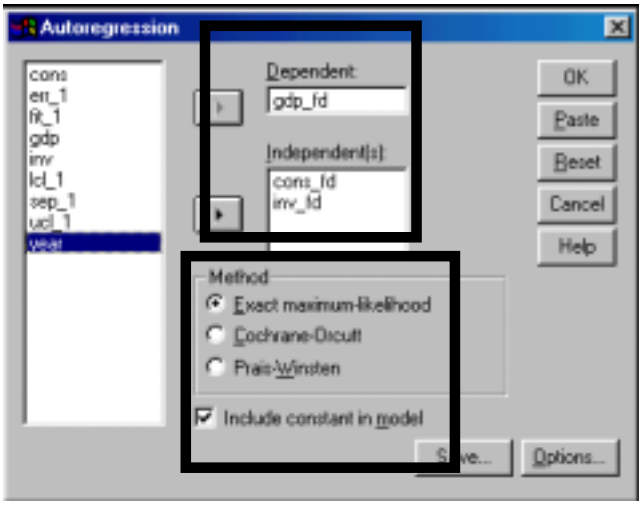
Rho = (DW-2)/2

A second order autocorrelation would be:
$Residuals_T = a * Residuals_{T-1} + a * Residuals_{T-2} + u_T$
(where u_T is truly random and uncorrelated with previous period values)

To correct for this problem, we would advise using the ARIMA process described in 17.5 along with any transformations that are necessitated to correct for autocorrelation. Consult your textbook for these transformations.

Luckily, for first order autocorrelation,[160] SPSS offers an automatic procedure - AUTOREGRESSION. Unfortunately, it is a bit restrictive compared to the ARIMA because a one-lag autoregressive component is automatically added, higher lag autoregressive components cannot be added, and Moving Average correction cannot be incorporated. Still, you may find it useful[161] and we devote the rest of this section to showing the procedure.

Go to STATISTICS/TIME SERIES/ AUTORESGRESSION. Choose the dependent and independent variables. (Note: We have chosen the first differenced transformations of each variable because here, unlike in the ARIMA, SPSS does not provide for automatic differencing). SPSS automatically assigns a 1-lag autoregressive component.



Choose the method for correcting for first order autocorrelation. (Consult your textbook for detailed descriptions of each option).

Click on "Options."

[160] The correlation among the residuals may be of a higher order. The Durbin Watson statistic cannot be used for testing for the presence of such higher order correlations. Consult your textbook for testing methods and for corrective methodology. Unfortunately, as in other econometric procedures, SPSS does not provide for automatic testing. It is incredible that it still sells so much.

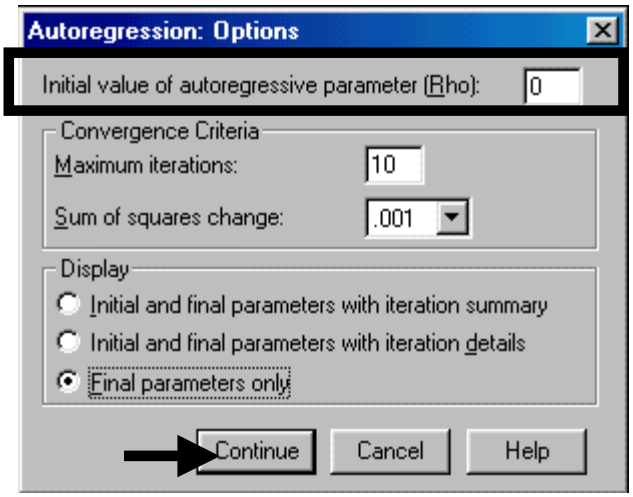
[161] Especially because Moving Average corrections are rarely conducted rigorously.

"Rho" is the coefficient of the first-order correlation. If you have an initial estimate (you may have it if you used Linear Regression and got the Durbin Watson statistic. Rho = ((DW-2)/2)).

Otherwise choose "0."

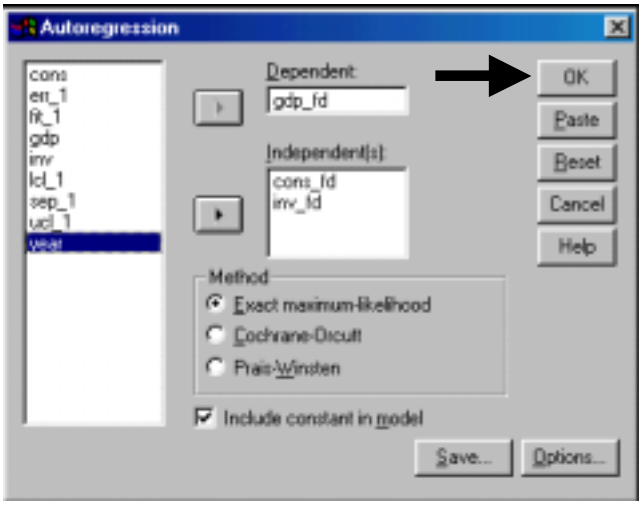
The other options are the same as for the ARIMA. Refer to 17.5 on how to choose them. (Usually, choose the defaults).

Click on "Continue."



Press "OK."

Note: You should click on "Save" and choose options the same way as in the ARIMA (see 17.5).



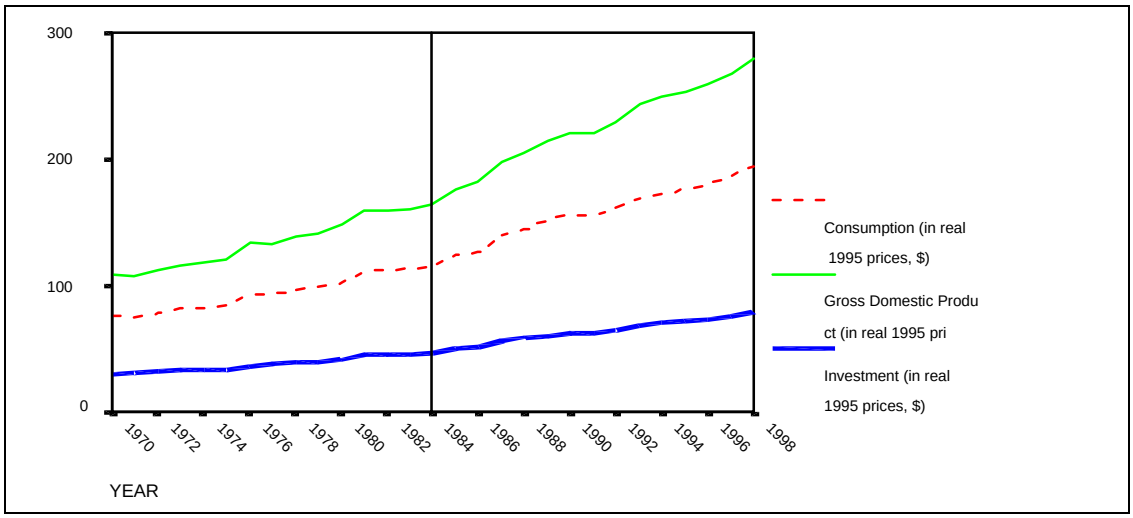
The result is exactly the same as in the previous section (this is an ARIMA(1,1,0) model). The only difference is that the algorithm corrected for first order autocorrelation.

Note: The STATISTICS/TIME SERIES/AUTOREGRESSION procedure is almost a sub-set of STATISTICS/TIME SERIES/ARIMA. As such, we would advise you to use the latter because of its greater flexibility in choosing:

- The degree of autoregression (for example you can use a variable that has two-period lags).
- Capability of ARIMA to incorporate Moving Average considerations.
- The need to create the differenced variables when using AUTOREGRESSION.

Ch 17. Section 7 Co-integration

SPSS does not have procedures for co-integration. In this section, we want to provide an intuitive understanding of co-integration as a method for removing the problem of non-stationarity. Recall that GDP and consumption were non-stationary (section 17.1 and 17.2).



Even though both series are non-stationary, can we find a relation between the two that is stationary?[162] That is, can we find a series that is calculated from GDP and consumption but itself exhibits randomness over time?

For example:
GDP = intercept - 0.7*(CONS) + Residual

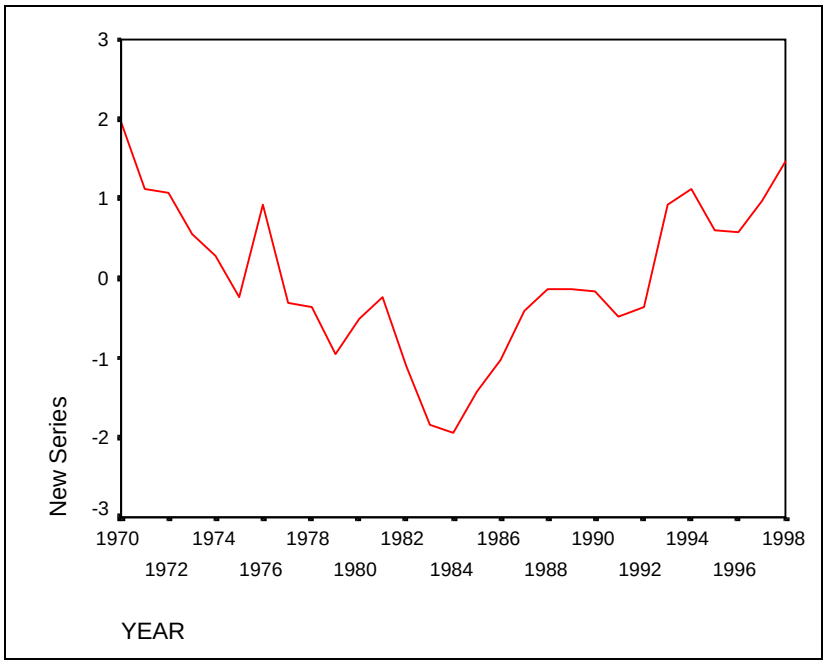
Or, Residual = GDP - intercept – (-0.7*(CONS))

The residual may be stationary.

Note: Our explanation is extremely simplistic. Do not use it on an exam!

A sequence chart of the above variable is presented below:

[162] One condition is that both (or all) the variables should have the same level of integration. That is, the same level of differencing should make them stationary. See 17.2 - we show that the variables all have the level of integration of 1 as the PACF of the first differenced transformations show no non-stationarity. We believe that the term "co-integration" has its roots in this condition – "Co" ("*all variables in the relation*…) + "integration(…*have the same level of integration*").



To take quizzes on topics within each chapter, go to http://www.spss.org/wwwroot/spssquiz.asp

Ch 18. Programming without programming

Never done programming? Afraid of or not interested in learning software code and coding? In SPSS, you can use program without knowing how to program or write code! As you go through this chapter the meaning of this will become clear.

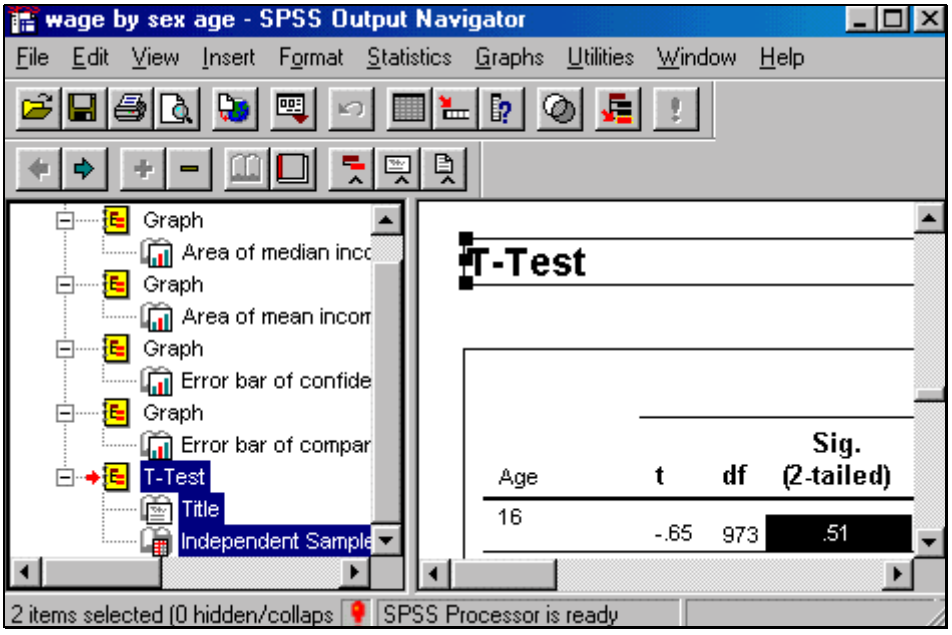
SPSS has two programming languages:

- Script. This language is used mainly for working on output tables and charts. Section 18.1 teaches how to use Scripts.
- Syntax. This language is used for programming SPSS procedures. It is the more important language. Section 18.2 teaches how to use Syntax.

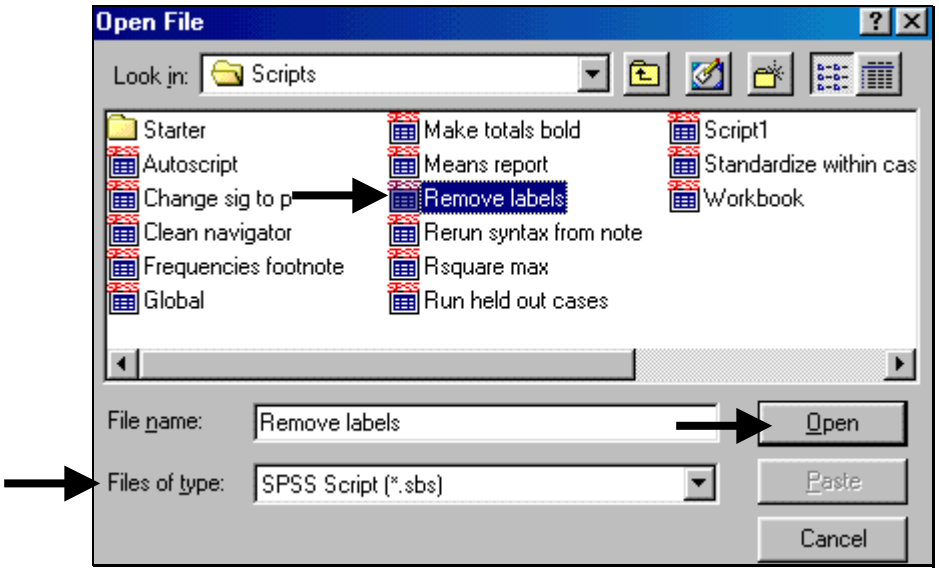
Ch. 18. Section 1. Using SPSS Scripts

Scripts are programs that provide tools that enable the saving of time in working on output tables, charts, etc. To use Scripts you don't have to learn programming-- you can use Scripts written by others. Some scripts are supplied with SPSS and were installed on your computer when you installed SPSS.

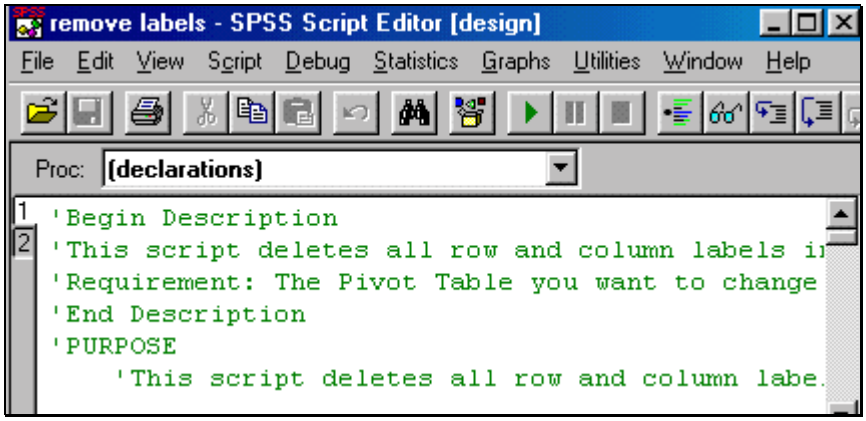
Most Scripts work on SPSS output. So, to learn how to use Scripts, open an output file. (Or use the one supplied with this document-- it is shown in the next picture).



Lets open a Script file. Go to FILE/OPEN. In the area "Files of type" choose the file type "SPSS Script (.sbs)" as shown below. To locate the Script files, go to the folder "SPSS/Scripts/" on the path SPSS is installed[163]. (Or do a search for files ending with the extension ".sbs.")



Click on "Open." The Script file opens in a new window called the "Script Editor" as shown in the picture below.



The Script file is basically a word-processing document in which text is written. The code starts with the line "'Begin Description." The lines of code that start with an apostrophe are comments that provide information to you on what the Script does and what requirements and actions it needs from you. In the next picture I show the entire description and purpose of the Script "Remove Labels."

[163] If you used the default installation then the folder will be "C:\Program Files\SPSS\Scripts."

```
'Begin Description
'This script deletes all row and column labels in the selected Pivot Table.
'Requirement: The Pivot Table you want to change must be selected.
'End Description
'PURPOSE
'This script deletes all row and column labels in the selected Pivot Table
```

Scroll down the page. You will see lines of text that do not have any apostrophe at the starting. These are the lines of functional code. Don't try to learn or worry about understanding the code[164].

```
Option Explicit     'All variables must be declarated before being use
'*********************************************************************
Sub Main
'Declare variables here
    Dim objPivotTable As PivotTable
    Dim objItem As ISpssItem
```

Actually you don't need to change anything or do anything to the code. You just "Run" (that is, execute) it. To run the code, first see if the Script has any "Requirements" from the user. (Look at the lines that begin with an apostrophe-- one of them may start with the word "Requirements.") This Script requires that the user first choose one output table (also called "Pivot" table). So, first go to the "SPSS Output Navigator" window and choose one table. Then go back to the "Script Editor" window and click on the icon with a rightward-pointing arrow. The icon is shown in the next picture.

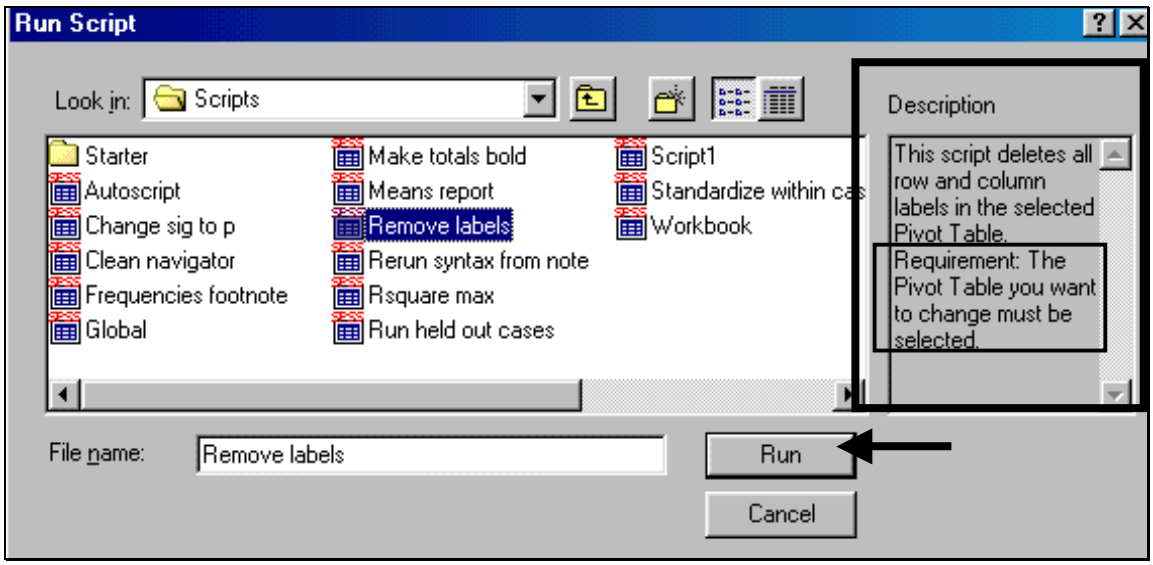
The code will run and perform the operations it is designed to do-- it will delete all row and column labels on all the tables in the open output window.

Another way to run a script-- go to UTILITIES / RUN SCRIPT as shown in the picture below.

The available scripts will be shown in the "RUN SCRIPT" dialog box as shown in the next picture. When you click on the name of a script, its description will be shown under the box

[164] It is based on the SAX BASIC language and is very similar to Visual Basic for Applications, the language for Microsoft Office.

"Description" on the right-hand side. Within the description area, play close attention to the paragraph "Requirements." In this case, you are required to choose one output table (also called a "Pivot" table) and then run the script. So, first choose one table. Then, go to the Script window shown below and click on "RUN." The Script will be executed.



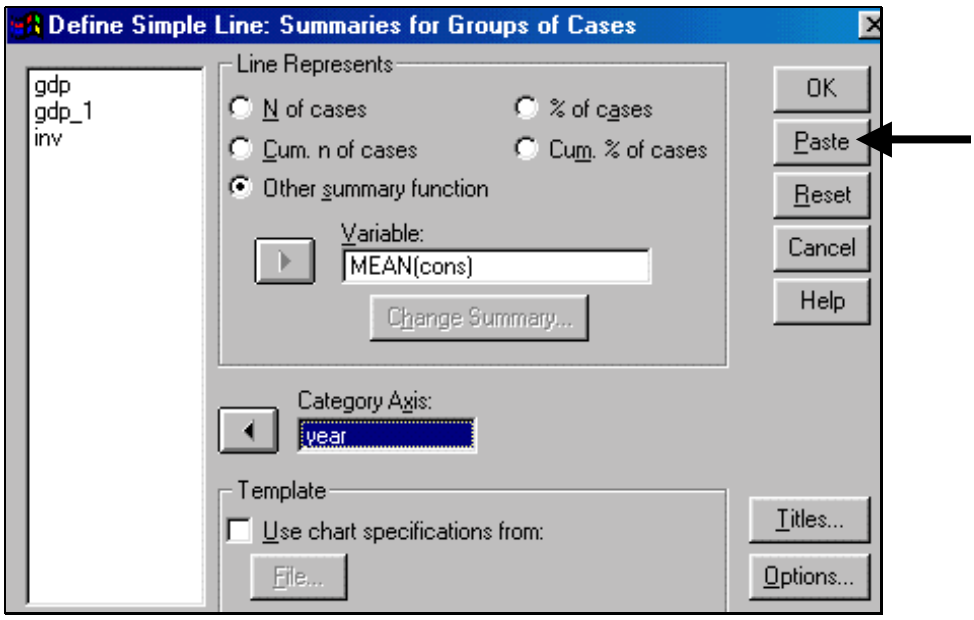
More Scripts: you can look for more Scripts at the SPSS web locations mentioned below. If you find them useful, download them and use them. Early next year, I will be writing some Scripts. Please provide me with some ideas for these Scripts.

- http://www.spss.com/software/spss/scriptexchange/
- http://www.spss.com/software/spss/scriptexchange/#new

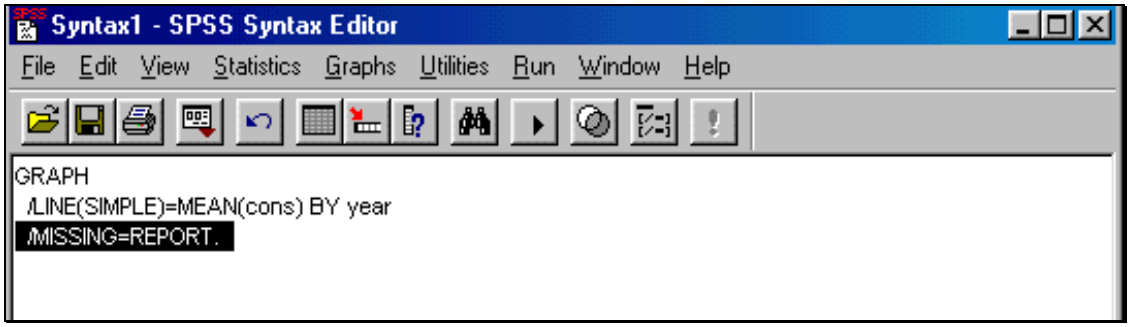
Ch 18. Section 2. Using SPSS Syntax

The use of Syntax is best shown with an example. Please work through the example along with me. Open the sample data file provided in the file you downloaded.

Go to GRAPHS / LINE GRAPH. Choose the option "Simple" and "Summary of individual cases." Click on "Define." The following dialog box opens.



Choose the options as shown above. Now, instead of clicking on the button "OK," click on the button "Paste." Doing this automatically launches the "Syntax Editor" window and automatically writes and pastes the code for the line graph you are planning to construct[165]. The syntax window is shown below.



As shown above, the code for the graph is written on 3 lines.

Just notice a few features:

- Variable names are in small case while all other code is in capital case
- Each line (apart from the first line) starts with the key "?."
- The bunch of code ends with a period.

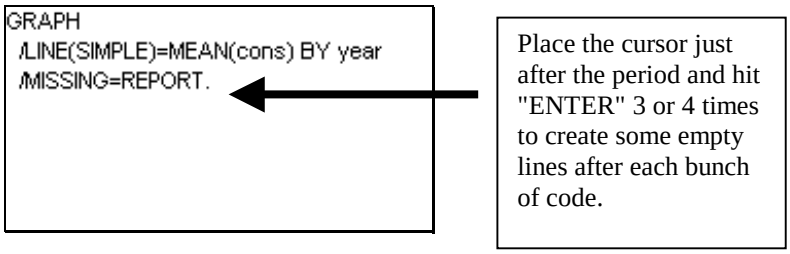
This is all you need to learn about the coding structure.

Two housekeeping steps to make your work more organized:

[165] Note that the graph does not get constructed at this stage.

1. Place the cursor after the end of the code and hit "Enter" a few times. This provides empty space between bunches of code thereby increasing the ease of reading and using the code. (This is shown in the next picture.)

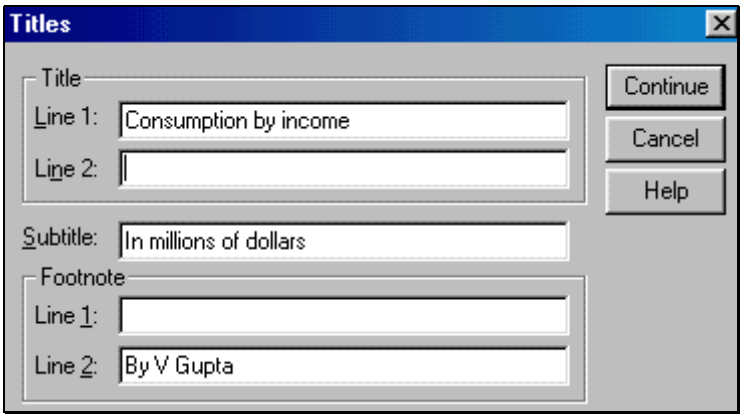
2. Save the Syntax file-- go to FILE/SAVE AS and save it as a Syntax file (with the extension ".sps.")



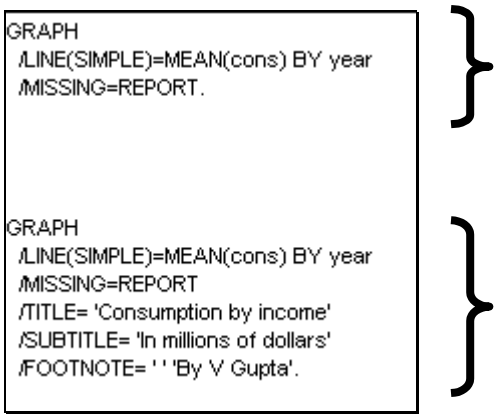
How does one use/execute the code? Using the mouse, highlight the 3 lines of code and go to RUN / ALL (or RUN / SELECTION) or click on the icon with the right-arrow sign (shown in the next picture.)

Try it out-- the graph is constructed.

Let's do another example. Make the same line graph as in the previous example, but this time, in addition, click on the button "Titles" (within the dialog box for line graphs) and enter titles and footnotes as shown in the next picture.



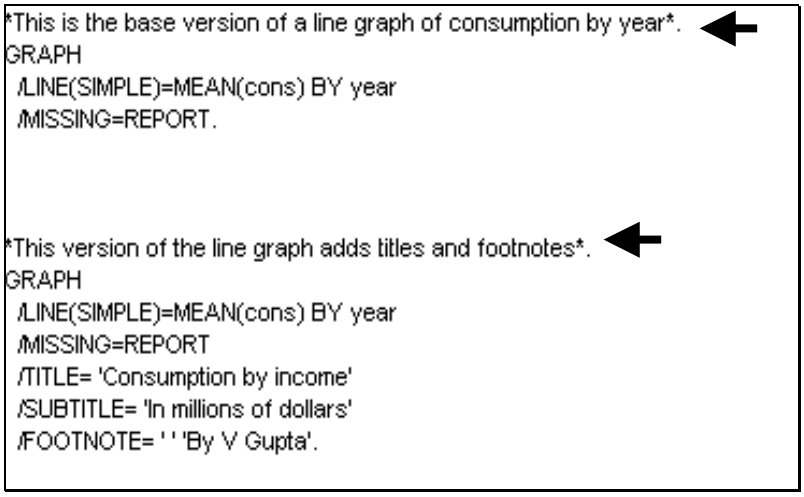
Click on "Continue" and then on "Paste." The code is written and pasted onto the Syntax file automatically by SPSS. The code is shown in the next picture. (It is the second bunch of code.)



This time the code includes new lines that capture the titles and footnotes you wrote. To run the entire code (that is, both the line graphs) choose both bunches of code and go to RUN/ SELECTION. To run only one bunch of code (that is, only one of the line graphs) choose one bunch of code and go to RUN/SELECTION. (Do you now realize the importance of placing empty lines after each bunch of code?)

After you paste every new bunch of code, go to FILE/SAVE.

One more good housekeeping strategy-- write some comments before each bunch of code. This comment may include what the code does, who pasted it, the date, etc. To write a comment, start the line with an asterisk (*) and end the line with an asterisk and then a period (*.). This is shown in the next picture.



Continue writing code this way. Choose a procedure and click on the button "Paste" instead of "OK." And continue…

18.2.a Benefits of using Syntax

As you use Syntax, you will notice its usefulness. Some major advantages of using Syntax are listed below:

1. Getting over any phobia/aversion to using (and maybe writing) software code. In general, becoming more confident with software technology.

2. Documenting all the work done in a project. If you use the simple point-and-click windows interface for your project then you will not have a record of all the procedures you conducted.

3. Because the Syntax allows you to document all your work, checking for errors becomes easier. With experience, you will be able to understand what the Syntax code says-- then checking for errors becomes very easy.

4. The main advantage is the massive saving of time and effort. How does syntax do this? Several ways, a few of which are listed below.

    - Replication of code (including using Word to assist in replication) allows you to save considerable time (in using the point-and-click windows interface) as shown in the example above.

    - Assume you want to run the same 40 procedures on 25 different files (say on data for five countries). If the files have the same variables that you are using in the procedures and the same variable names, then considerable time can be saved by creating the Syntax file using one country's data file and then running the same code on the data files of the other counties. In a follow-up chapter ('Advanced Programming in SPSS") I will show more ways to write time saving Syntax files.

    - Assume you have several files with similar data but with different variable names. Create the syntax file for one data file. Then, for the other data files, just replace the variable names on the original Syntax file!

    - A frustrating situation arises when you have to redo all you work because of data or other issues.[166] SPSS can save an incredible amount of time as also the boredom produced by repeating tasks.

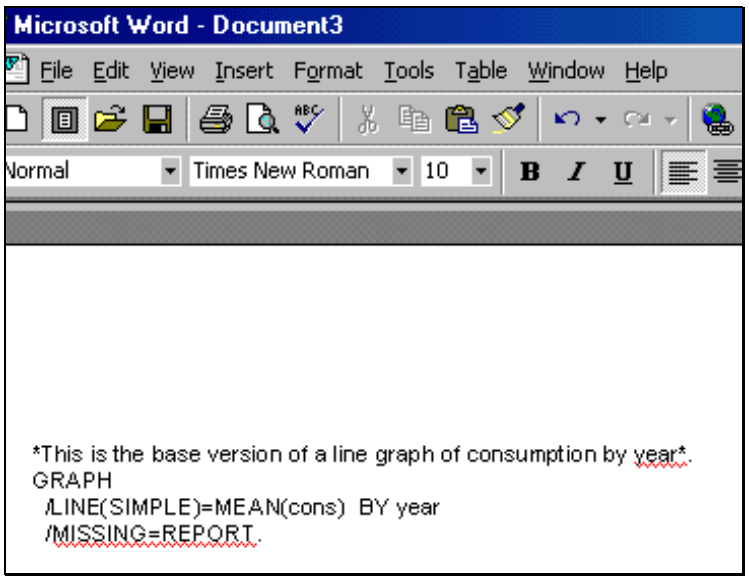
    - After running some procedures, you may want to run them again on only a subset of the data file (see section 1.7 in my book), or separately for sub-groups of the data (see ch 10 in my book). Syntax makes this easy. If you have the syntax file for the procedures you conducted on the entire data file, then the same procedures can be redone for the sub-group(s) of the data by first making the sub-groups(s) and then re-running the code in the Syntax file.

18.2.b Using Word (or WordPerfect) to save time in creating code

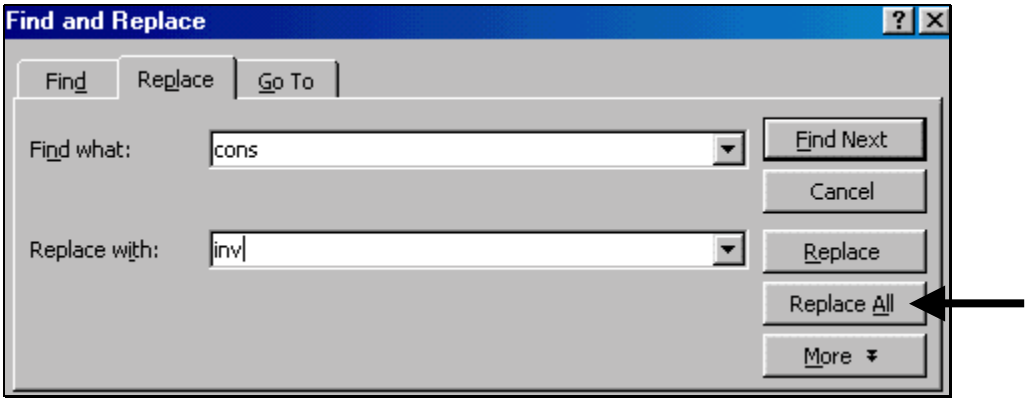
Microsoft Word has an extremely powerful "Find and Replace" facility. You can use this facility to save time in replicating SPSS code. In the earlier examples, I showed how to write code for a graph of "Consumption" and "Year." Let's use Word to replicate the code for graphs of "Investment" by "Year."

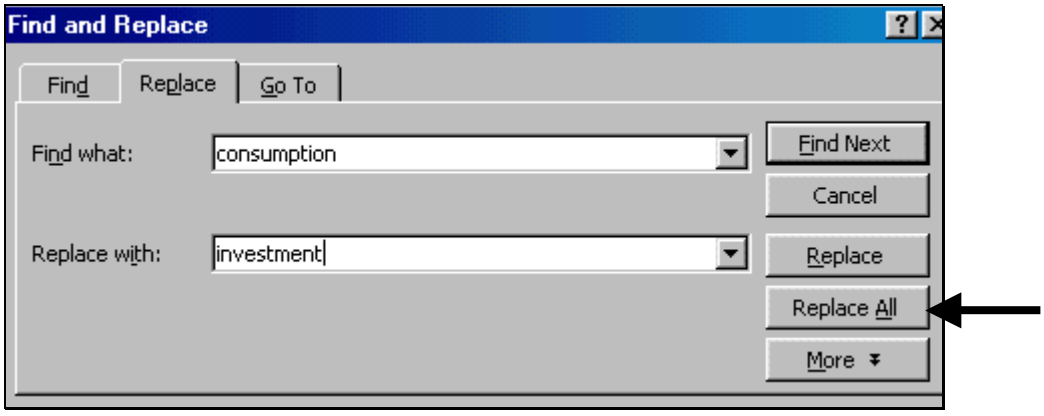
[166] This may happen if you are provided a more accurate version of the data than the one you worked on for a few weeks, or you have to change the choice of the dependent variable in all your regression models, or you used the incorrect data file, or (and this is a frequent occurrence) you forgot (or were not informed of the need to) perform some crucial data step like defining value labels or weighing cases.

Select the two bunches of code you "wrote" earlier. Go to EDIT / COPY. Open Microsoft Word (or WordPerfect). Go to EDIT / PASTE. The SPSS code is pasted onto the Word document as shown in the next picture.



Go to EDIT / REPLACE and choose the options for replacing "cons" and "consumption" by "inv" and "investment," respectively. (See the next two pictures.)





Copy the changed text (all of it) and go to EDIT / COPY. Go back to the SPSS syntax file and go to EDIT / PASTE. To run the two new bunches of code, choose the lines of code and go to RUN / SELECTION.

INDEX

The index is in two parts:

1. Part 1 has the mapping of SPSS menu options to sections in the book.

2. Part 2 is a regular index.

Part 1: Relation between SPSS menu options and the sections in the book

Menu	Sub-Menu	Section that teaches the menu option
FILE	NEW	-
„	OPEN	1.1
„	DATABASE CAPTURE	16
„	READ ASCII DATA	12
„	SAVE	-
„	SAVE AS	-
„	DISPLAY DATA INFO	-
„	APPLY DATA DICTIONARY	-
„	STOP SPSS PROCESSOR	-
EDIT	OPTIONS	15.1
„	ALL OTHER SUB-MENUS	-
VIEW	STATUS BAR	15.2
„	TOOLBARS	15.2
„	FONTS	15.2
„	GRID LINES	15.2
„	VALUE LABELS	15.2
DATA	DEFINE VARIABLE	1.2
„	DEFINE DATES	-
„	TEMPLATES	-
„	INSERT VARIABLE	-

Menu	Sub-Menu	Section that teaches the menu option
DATA	INSERT CASE, GO TO CASE	-
„	SORT CASES	1.5
„	TRANSPOSE	-
„	MERGE FILES	13
„	AGGREGATE	1.4
„	ORTHOGONAL DESIGN	-
„	SPLIT FILE	10
„	SELECT CASES	1.7
„	WEIGHT CASES	1.3
TRANSFORM	COMPUTE	2.2
„	RANDOM NUMBER SEED	-
„	COUNT	2.4
„	RECODE	2.1
„	RANK CASES	-
„	AUTOMATIC RECODE	2.1
„	CREATE TIME SERIES	17.4
„	REPLACE MISSING VALUES	1.8, 17.4.a
STATISTICS / SUMMARIZE	SUMMARIZE / FREQUENCIES	3.2.a
„	DESCRIPTIVES	3.3.a
„	EXPLORE	5.4
„	CROSSTABS	-
„	ALL OTHER	-
STATISTICS / CUSTOM TABLES	BASIC TABLES	6.1
„	GENERAL TABLES	2.3 and 6.2 together
„	TABLES OF FREQUENCIES	6.2
STATISTICS / COMPARE MEANS	MEANS	-
„	ONE SAMPLE T-TEST	3.4.b
„	INDEPENDENT SAMPLES T-TEST	5.5.b

Menu	Sub-Menu	Section that teaches the menu option
STATISTICS / COMPARE MEANS	PAIRED SAMPLES T-TEST	4.3.b
„	ONE-WAY ANOVA	5.5.c
STATISTICS / GENERAL LINEAR MODEL		-
STATISTICS /CORRELATE	BIVARIATE	5.3.a, 5.3.b
„	PARTIAL	5.3.c
„	DISTANCE	-
STATISTICS / REGRESSION	LINEAR	7 (and 8)
„	CURVE ESTIMATION	9.1.a
„	LOGISTIC [LOGIT]	9.1
„	PROBIT	-
„	NON-LINEAR	9.1.b
„	WEIGHT ESTIMATION	8.2.a
„	2-STAGE LEAST SQUARES	8.4
STATISTICS / LOGLINEAR		-
STATISTICS / CLASSIFY	K-MEANS CLUSTER	2.5
„	HIERARCHICAL CLUSTER	-
„	DISCRIMINANT	-
STATISTICS / DATA REDUCTION		-
STATISTICS / SCALE		-
STATISTICS / NON-PARAMETRIC TESTS	CHI-SQUARE	14.2
„	BINOMIAL	14.1

Menu	Sub-Menu	Section that teaches the menu option
STATISTICS / NON-PARAMETRIC TESTS	RUNS	14.3
„	1 SAMPLE K-S	3.2.e
„	2 INDEPENDENT SAMPLES	5.5.d
„	K INDEPENDENT SAMPLES	5.5.d
„	2 RELATED SAMPLES	4.3.c
„	K RELATED SAMPLES	4.3.c
STATISTICS / TIME SERIES	EXPONENTIAL SMOOTHING, X11 ARIMA, SEASONAL DECOMPOSITION	-
„	ARIMA	17.5
„	AUTOREGRESSION	17.6
STATISTICS / SURVIVAL		-
STATISTICS / MULTIPLE SETS	DEFINE SETS	2.3
„	FREQUENCIES	2.3 (see 3.1.a also)
„	CROSSTABS	2.3
GRAPHS	BAR	3.1, 4.1, 5.1
„	LINE	3.1, 5.1
„	AREA	3.1, 5.1
„	PIE	3.1, 4.1, 5.1
„	HIGH-LOW, PARETO, CONTROL	-
„	BOXPLOT	3.3.b, 4.2, 5.1.d
„	ERROR BAR	3.4.a, 4.3.a, 5.5.a
„	SCATTER	5.2
„	HISTOGRAM	3.2.a
„	P-P	3.2.b, 3.2.c, 3.2.d
„	Q-Q	3.2.b, 3.2.c, 3.2.d
„	SEQUENCE	17.1

Menu	Sub-Menu	Section that teaches the menu option
GRAPHS	TIME SERIES/AUTO CORRELATIONS	17.2
„	TIME SERIES/CROSS CORRELATIONS	17.3
„	TIME SERIES/SPECTRAL	-
UTILITIES	VARIABLES	1.2.f
„	FILE INFO	1.2.g
„	DEFINE SETS	1.9
„	USE SETS	1.9
„	RUN SCRIPT	18.1
„	ALL OTHER	-

Based on your feedback, we will create sections on menu options we have ignored in this book. These sections will be available for download from spss.org and vgupta.com. (We do not want to add sections to the book because it is already more than 400 pages.)

Part 2: Regular index