Chapter 14
Simple Linear Regression
Regression Diagnostics and Remedial Measures

Simple Linear Regression
Regression Diagnostics and Remedial Measures

1. Residuals and regression assumptions

- The regression assumptions can be stated in terms of the residuals

$$\varepsilon \sim NID(0, \sigma^2)$$

  - All observations are independent and randomly selected from the population (or equivalently, the residual terms, $\varepsilon_i s$, are independent)
  - The residuals are normally distributed at each level of X
  - The variance of the residuals is constant across all levels of X

- We must also assume that the regression model is the correct model
  - The relationship between the predictor and outcome variable is linear
  - No relevant variables have been omitted
  - No error in the measurement of predictor variables

- Types of residuals

  - (Unstandardized) residuals, e_i

$$e_i = Y_i - \hat{Y}$$

    - A residual is the deviation of the observed value from the predicted value on the original scale of the data
    - If the regression model fits the data perfectly, then there would be no residuals. In practice, we always have residuals, but the presence of many large residuals can indicate that the model does not fit the data well
    - If the residuals are normally distributed, then we would expect to find
      5% of residuals greater than 2σ from the mean
      1% of residuals greater than 2.5σ from the mean
      .1% of residuals greater than 3σ from the mean
    - It can be difficult to eyeball standard deviations from the mean, so we often turn to standardized residuals

- Standardized residuals, $\tilde{e}_i$

$$\tilde{e}_i = \frac{Y_i - \hat{Y}}{\sigma_e} = \frac{Y_i - \hat{Y}}{\sqrt{MSE}}$$

  - Standardized residuals are z-scores. Why?

    The average of the residuals is zero

    $$\bar{e} = \frac{\sum e_i}{n} = 0$$

    The standard deviation of the residuals is $\sqrt{MSE}$

    $$Var(e) = \frac{\sum (e_i - \bar{e})^2}{n-2} = \frac{\sum e_i^{\ 2}}{n-2} = \frac{SSE}{n-2} = MSE$$

    So a standardized residual would be given by:

    $$\tilde{e}_i = \frac{e_i - \bar{e}}{\sigma_e} = \frac{e_i}{\sqrt{MSE}} = \frac{Y_i - \bar{Y}}{\sqrt{MSE}}$$

  - Because standardized residuals are z-scores, we can easily detect outliers. When examining standardized residuals, we should find:
    5% of $|\tilde{e}_i|s$ greater than 2
    1% of $|\tilde{e}_i|s$ greater than 2.5
    .1% of $|\tilde{e}_i|s$ greater than 3

- Studentized residuals, e_i'
  - MSE is the overall variance of the residuals
  - It turns out that the variance of an individual residual is a bit more complicated. Each residual has its own variance, depending on its distance from $\bar{X}$
  - When residuals are standardized using residual-specific standard deviations, the resulting residual is called a studentized residual.
  - In large samples, it makes little difference whether standardized or studentized are used. However, in small samples, studentized residuals give more accurate results.
  - Because SPSS makes the use of studentized residuals easy, it is good practice to examine studentized residuals rather than standardized residuals

- Obtaining residuals in SPSS

```
REGRESSION
  /DEPENDENT dollars
  /METHOD=ENTER miles
  /SAVE RESID (resid) ZRESID (zresid)  SRESID (sresid) .
```

  - RESID produces unstandardized residuals
  - ZRESID produces standardized residuals
  - SRESID produces studentized residuals

  - Each residual appears in a new data column in the data editor

RESID	ZRESID	SRESID
-.80365	-.34695	-.35921
-1.33272	-.57536	-.61672
-1.60685	-.69370	-.73813
1.50761	.65086	.68389
1.97215	.85140	.88173
-1.33425	-.57601	-.64739
.37854	.16342	.18972
-2.73592	-1.18114	-1.22407
-3.46819	-1.49727	-1.56097
-.13105	-.05658	-.05952
3.39148	1.46415	1.54912
1.61081	.69541	.77910
2.91415	1.25808	1.49866
2.50928	1.08329	1.16348
-2.87139	-1.23962	-1.28850

    - You can see the difference between standardized and studentized residuals is small, but it can make a difference in how the model fit is interpreted

- Because all the regression assumptions can be stated in terms of the residuals, examining residuals and residual plots can be very useful in verifying the assumptions

  - In general, we will rely on residual plots to evaluate the regression assumptions rather than rely on statistical tests of those assumptions

2. Residual plots to detect lack of fit

- There are several reasons why a regression model might not fit the data well including:
  - The relationship between X and Y might not be linear
  - Important variables might be omitted from the model

- To detect non-linearity in the relationship between X and Y, you can:

  - Create a scatterplot of X against Y
    - Look for non-linear relationships between X and Y

  - Plot the residuals against the X values
    - The residuals have linear association between X and Y removed. If X and Y are linearly related, then all that should be remaining for the residuals to capture is random error
    - Thus, any departure from a random scatterplot indicates problems
    - In general, this graph is easier to interpret than the simple scatterplot and an added advantage of this graph (if studentized residuals are used) is that you can easily spot outliers

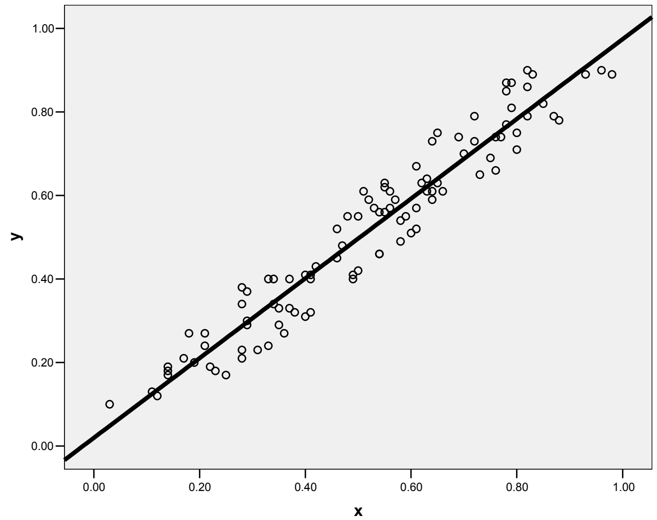
    - In simple linear regression, a plot of e_i *vs* X is identical to a plot of e_i *vs* $\hat{Y}$. Thus, there is no need to examine both of these plots.

      The predicted values are the part of the Ys that have a linear relationship with X, so $\hat{Y}$ and X will always be perfectly correlated when there is only one predictor.

      In multiple regression, different information may be obtained from a plot of e_i *vs* X and from a plot of e_i *vs* $\hat{Y}$.

- Example #1: A good linear regression model (n = 100)
  - A scatterplot of X against Y

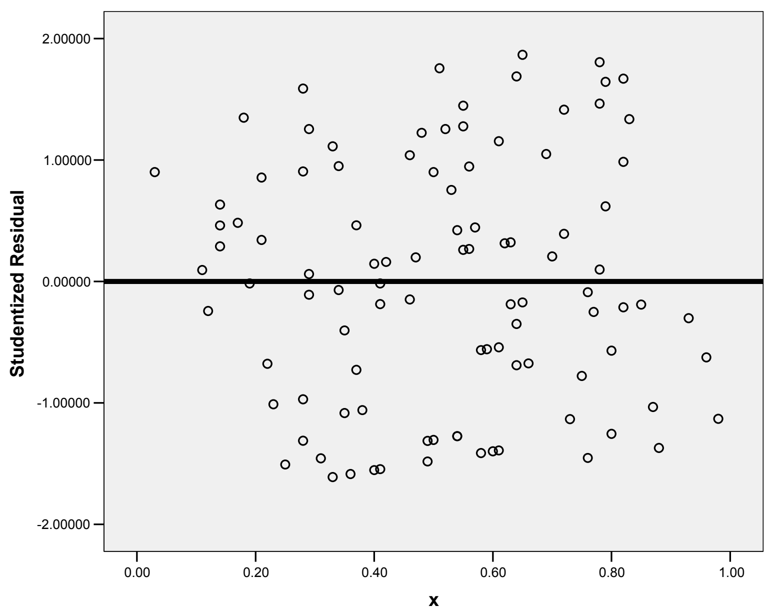
```
GRAPH /SCATTERPLOT(BIVAR)=x WITH y.
```



  - The X-Y relationship looks linear

- Plot the residuals against the X values

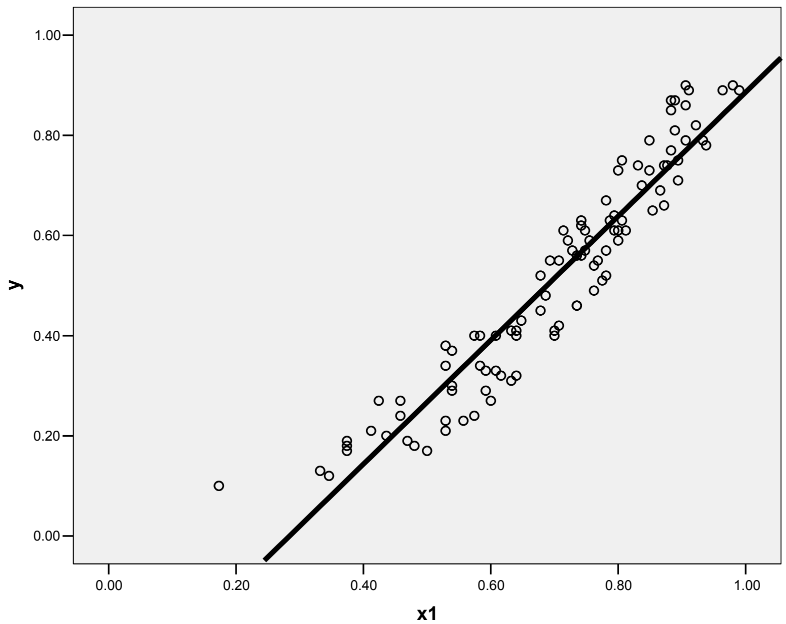
```
GRAPH /SCATTERPLOT(BIVAR)=x WITH sresid.
```



  - The plot looks random so we have evidence that there is no non-linear relationship between X and Y
  - We also see that no outliers are present
  - This graph is as good as it gets!

- Example #2: A nonlinear relationship between *X* and *Y* ($n = 100$)
  - A scatterplot of *X* against *Y*

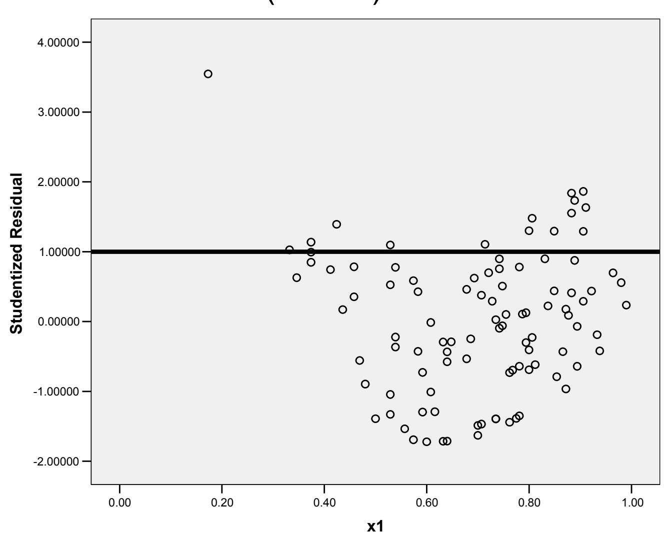
```
GRAPH /SCATTERPLOT(BIVAR)=x1 WITH y.
```



  - The *X*-*Y* relationship looks mostly linear

- Plot the residuals against the *X* values

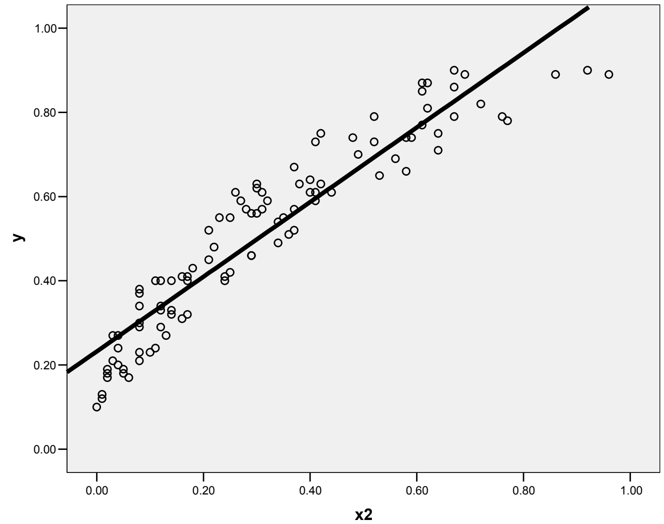
```
GRAPH /SCATTERPLOT(BIVAR)=x1 WITH sresid.
```



  - This graph has a slight U-shape, suggesting the possibility of a non-linear relationship between *X* and *Y*
  - We also see one outlier

- Example #3: A second nonlinear relationship between *X* and *Y* ($n = 100$)
  - A scatterplot of *X* against *Y*

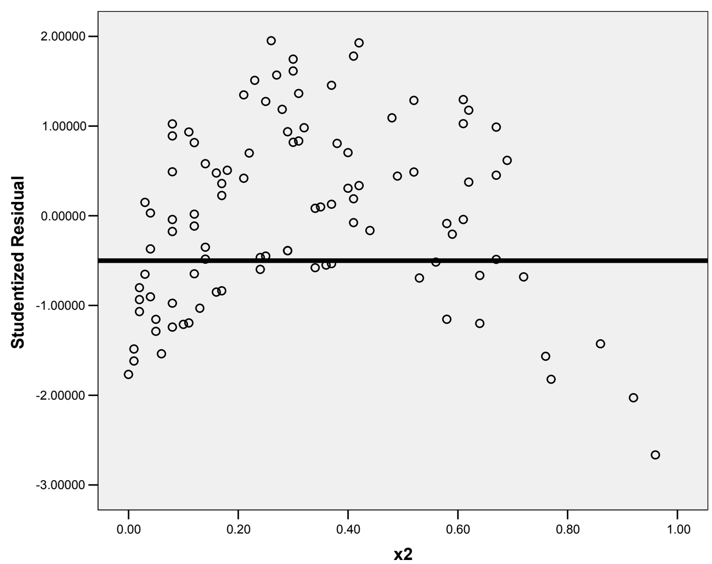
```
GRAPH /SCATTERPLOT(BIVAR)=x2 WITH y.
```



  - The *X*-*Y* looks slightly curvilinear in this case

- Plot the residuals against the *X* values

```
GRAPH /SCATTERPLOT(BIVAR)=x2 WITH sresid.
```



  - This graph has a strong U-shape, indicating a non-linear relationship between *X* and *Y*
  - Notice that it is easier to detect the non-linearity in the residual plot than in the scatterplot

- You can not determine lack-of-fit/non-linearity from the significance tests on the regression parameters

```
REGRESSION
  /STATISTICS COEFF OUTS R ANOVA ZPP
  /DEPENDENT y
  /METHOD=ENTER x2
```

  - In this case, we find evidence for a strong linear relationship between *X2* and *Y*, $b = .887, t(98) = 18.195, p < .001 \ [r = .94]$

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	Correlations Zero-order	Correlations Partial	Correlations Part
1	(Constant)	.232	.013		18.195	.000			
	X2	.887	.032	.941	27.458	.000	.941	.941	.941

a. Dependent Variable: Y

  - This linear relationship between *X2* and *Y* accounts for 88.5% of the variance in *Y*.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.941[a]	.885	.884	.07585

a. Predictors: (Constant), X2

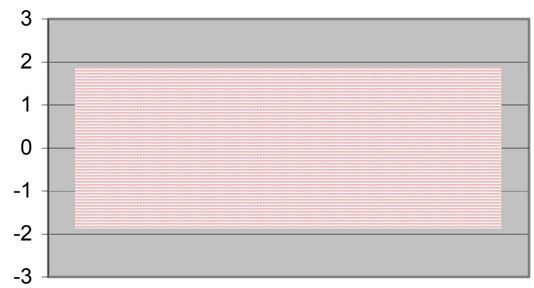
  - Yet from the residual plot, we know that this linear model is incorrect and does not fit the data well

  - Despite the level of significance and the large percentage of the variance accounted for, we should not report this erroneous model

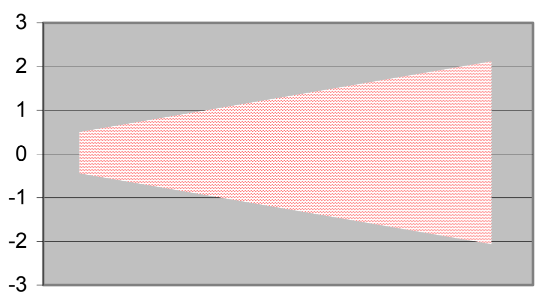
- Detecting the omission of an important variable by looking at the residuals is very difficult!

3. Residual plots to detect homogeneity of variance

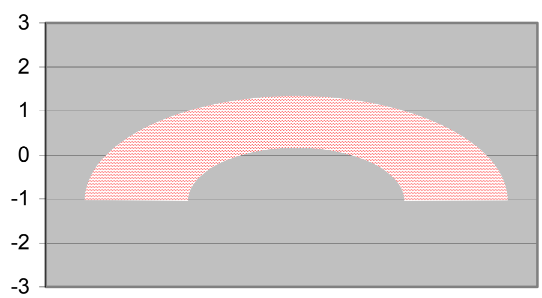
- We assume that the variance of the residuals is constant across all levels of predictor variable(s)
- To examine if the residuals are homoscedastic, we can plot the residuals against the predicted values
  - If the residuals are homoscedastic, then their variability should be constant over the range



GOOD



BAD



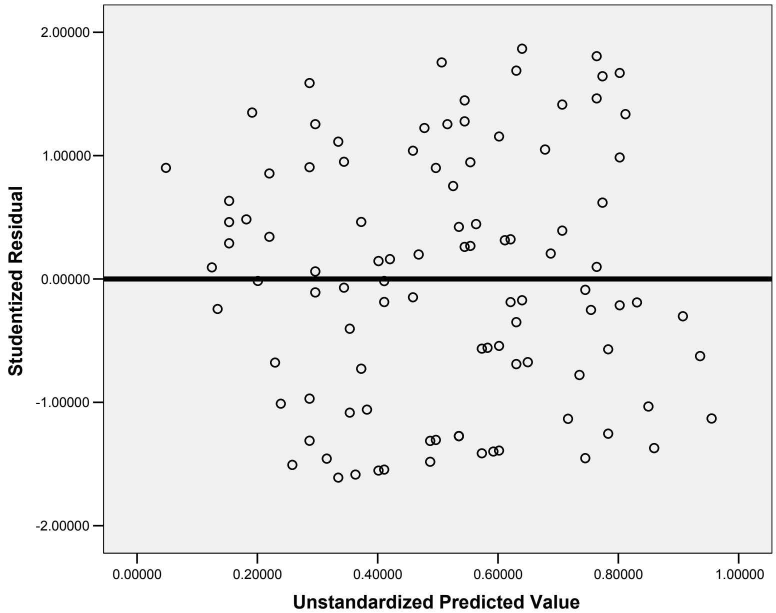
BAD

  - As previously mentioned, plotting residuals against fitted values ($\hat{Y}$) or against the predictor (X) produces the same plots when there is only one X variable. In multiple regression, a plot of the residuals against fitted values ($\hat{Y}$) is generally preferred, but in this case it makes no difference

  - The raw residuals and the standardized residuals do not take into account the fact the variance of each residual is different (and depends on its distance from the mean of X). For plots to examine homogeneity, it is particularly important to use the studentized residuals

- Example #1: A homoscedastic model (n = 100)

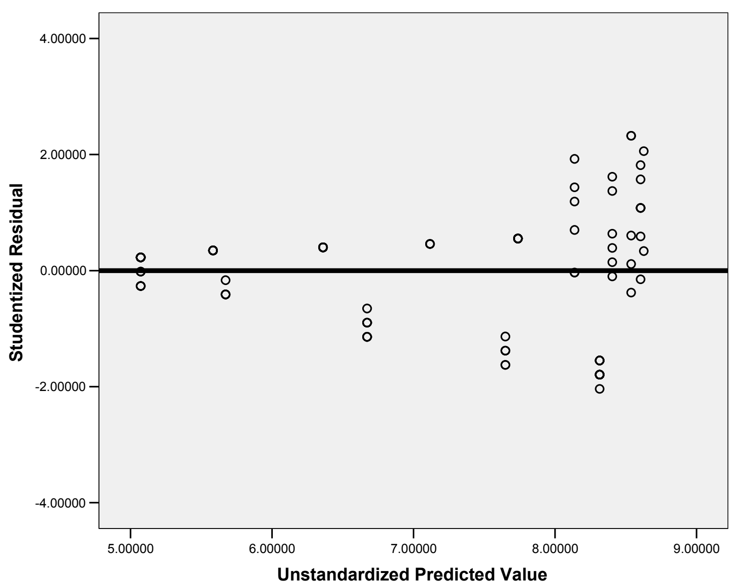
```
GRAPH /SCATTERPLOT(BIVAR)=sresid WITH pred.
```



  - The band of residuals is constant across the entire length of the observed predicted values

- Example #2: A heteroscedastic model (n = 100)

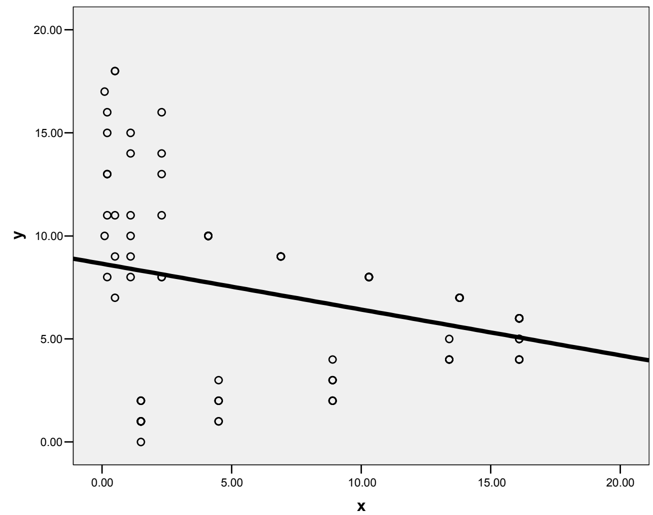
```
GRAPH /SCATTERPLOT(BIVAR)=sresid WITH pred.
```



  - This pattern where the variance increases as Y increases is a common form of heteroscedasticity.

- In this case, the unequal heteroscedasticity is also apparent from the X-Y scatterplot. But in general, violations of the variance assumption are easier to spot in the residual plots

```
GRAPH /SCATTERPLOT(BIVAR)=y WITH x.
```



- As in the case of looking for non-linearity, examining the regression model provides no clues that the model assumptions have been violated

```
REGRESSION
  /STATISTICS COEFF OUTS R ANOVA ZPP
  /DEPENDENT y
  /METHOD=ENTER x.
```

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.303[a]	.092	.083	4.11810

a. Predictors: (Constant), X

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations		
		B	Std. Error	Beta			Zero-order	Partial	Part
1	(Constant)	8.650	.649		13.336	.000			
	X	-.222	.070	-.303	-3.153	.002	-.303	-.303	-.303

a. Dependent Variable: Y

4. Residual plots to detect non-normality

- As for ANOVA, symmetry is more important than normality
- There are a number of techniques that we can use to check normality of the residuals. In general, these are the same techniques we used to check normality in ANOVA
  - Boxplots or histograms of residuals
  - A normal P-P plot of the residuals
  - Coefficients of skewness/kurtosis may also be used
- Normality is difficult to check and can be influenced by other violations of assumptions. A good strategy is to check and address all other assumptions first, and then turn to checking normality

- These tests are not foolproof
  - Technically, we assume that the residuals are normally distributed at each level of the predictor variable(s)
  - It is possible (but unlikely) that the distribution of residuals might be left-skewed for some values of X and right skewed for other values so that, on average, the residuals appear normal.
  - If you are concerned about this possibility and if you have a very large sample, you could divide the *Xs* into *a* equal categories, and check normality separately for each of the *a* subsamples (you would want at least 30-50 observations per group). In general, this is not necessary.

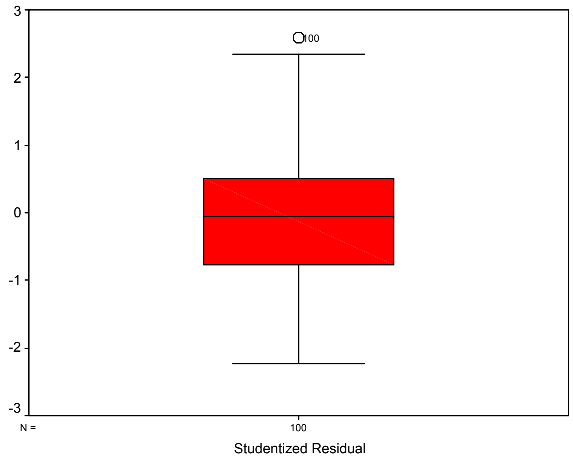
- Example #1: Normally distributed residuals ($N = 100$)

```
EXAMINE VARIABLES=sresid
 /PLOT BOXPLOT HISTOGRAM NPPLOT.
```

Descriptives

		Statistic	Std. Error
Studentized Residual	Mean	.0002928	.10048947
	5% Trimmed Mean	-.0129241	
	Median	-.0584096	
	Variance	1.010	
	Std. Deviation	1.004895	
	Minimum	-2.22678	
	Maximum	2.57839	
	Range	4.80518	
	Interquartile Range	1.2754269	
	Skewness	.211	.241
	Kurtosis	-.182	.478

  - The mean is approximately equal to the median
  - The coefficients of skewness and kurtosis are relatively small

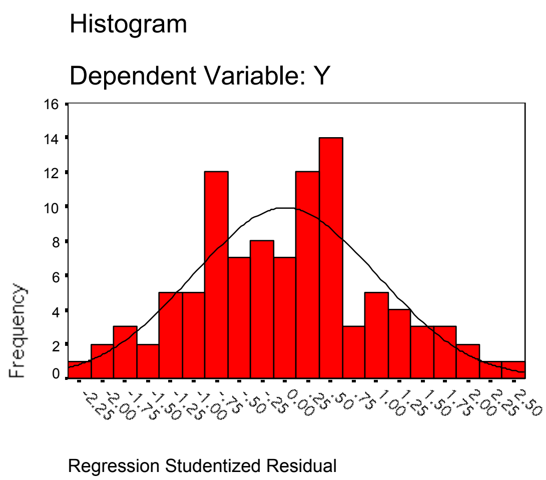


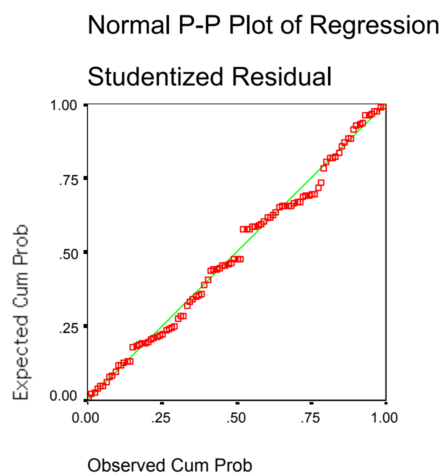
Tests of Normality

	Shapiro-Wilk		
	Statistic	df	Sig.
Studentized Residual	.990	100	.648

- Plots can also be obtained directly from the regression command

```
REGRESSION /DEPENDENT y
  /METHOD=ENTER z
  /RESIDUALS HIST(SRESID) NORM(SRESID)
  /SAVE sRESID (sresid).
```

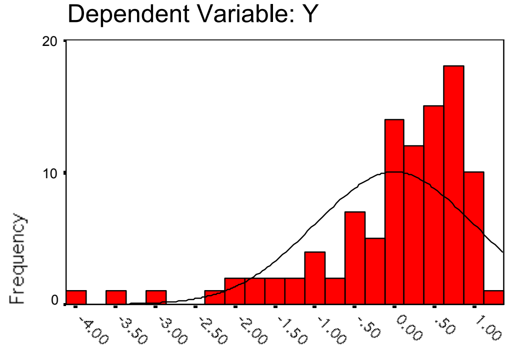


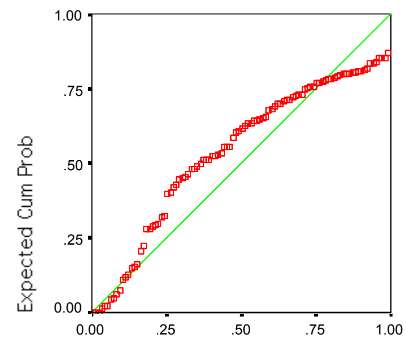


- The histogram and P-P plot are as good as they get. There are no problems with the normality assumption.

- Example #2: Non-normally distributed residuals (N = 100)

```
REGRESSION
  /DEPENDENT y
  /METHOD=ENTER z1
  /RESIDUALS HIST(ZRESID1) NORM(ZRESID1)
  /SAVE ZRESID (zresid1).
```






```
EXAMINE VARIABLES=zresid1
  /PLOT BOXPLOT HISTOGRAM NPPLOT
  /STATISTICS DESCRIPTIVES.
```

Descriptives

		Statistic	Std. Error
Standardized Residual	Mean	.0000000	.09949367
	5% Trimmed Mean	.1054551	
	Median	.2856435	
	Variance	.990	
	Std. Deviation	.99493668	
	Minimum	-4.06265	
	Maximum	1.13547	
	Range	5.19812	
	Interquartile Range	1.1329148	
	Skewness	-1.769	.241
	Kurtosis	3.747	.478

Tests of Normality

	Shapiro-Wilk		
	Statistic	df	Sig.
Standardized Residual	.835	100	.000

- All signs point to non-normal, non-symmetrical residuals. There is a violation of the normality assumption in this case.

5. Identifying outliers and influential observations

- Observations with large residuals are called outliers
- But remember, when the residuals are normally distributed, we expect a small percentage of residuals to be large

  We expect 5% of $|e_i|s$ greater than 2

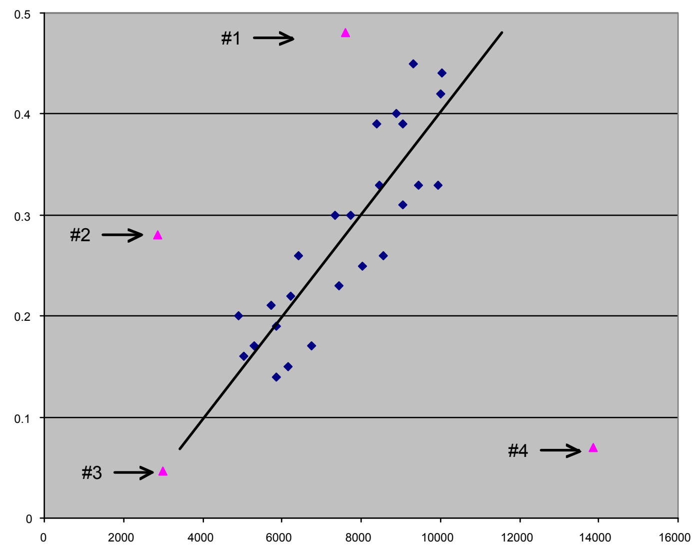
  We expect 1% of $|e_i|s$ greater than 2.5

  We expect .1% of $|e_i|s$ greater than 3

# of observations	Expected number of residuals >2	>2.5	>3
50	2.5	0.5	.005
100	5	1	0.1
200	10	2	0.2
500	25	5	0.5
1000	50	10	1

  - Many people use $|e_i| > 2$ as a check for outliers, but this criterion results in too many observations being identified as outliers. In large samples, we expect a large number of observations to have residuals greater than 2
  - A more reasonable cut-off for outliers is to use $|e_i| > 2.5$ or even $|e_i| > 3$

- There are multiple kinds of outliers



  - #1 is an Y outlier
  - #2 is an X outlier
  - #3 and #4 are outliers for both X and Y

- When we examine extreme observations, we want to know:
  - Is it an outlier? (i.e., Does it differ from the rest of the observed data?)
  - Is it an influential observation?
    (i.e., Does it have an impact on the regression equation?)

- Clearly, each of the values highlighted on the graph is an outlier, but how will each influence estimation of the regression line?
  - Outlier #1
    - Influence on the intercept:
    - Influence on the slope:
  - Outlier #2
    - Influence on the intercept:
    - Influence on the slope:
  - Outlier #3
    - Influence on the intercept:
    - Influence on the slope:
  - Outlier #4
    - Influence on the intercept:
    - Influence on the slope:

- Not all outliers are equally influential. It is not enough to identify outliers; we must also consider the influence each may have (particularly on the estimation of the slope)

- Methods of identifying outliers and influential points:
  - Examination of the studentized residuals
  - A scatterplot of studentized residuals with X
  - Examination of the studentized deletion residuals
  - Examination of leverage values
  - Examination of Cook's distance (Cook's D)

- Studentized Deletion Residuals

  - A <u>deletion residual</u> is the difference between the observed Y_i and the predicted $\hat{Y}_{(i)}$ value based on a model with the i^{th} observation deleted

$$d_i = Y_i - \hat{Y}_{i(i)}$$

  - The deletion residual is a measure of how much the i^{th} observation influences the overall regression equation
  - If the i^{th} observation has no influence on the regression line then $Y_i = \hat{Y}_{i(i)}$ and $d_i = 0$
  - The greater the influence of the observation, the greater the deletion residual
  - Note that we cannot determine if the observation influences the estimation of the intercept or of the slope. We can only tell that it has an influence on at least one of the parameters in the regression equation.

  - The size of the deletion residuals will be determined, in part, by the scale of the Y values. In order to create deletion residuals that do not depend on the scale of Y, we can divide d_i by its standard deviation to obtain a studentized deletion residual

$$\tilde{d}_i = \frac{Y_i - \hat{Y}_{i(i)}}{s(d_i)}$$

  - Studentized deletion residuals can be interpreted like z-scores (or more precisely, like t-scores)

- Leverage values
  - It can be shown (proof omitted) that the predicted value for the i^{th} observation can be written as a linear combination of the observed Y values

$$\hat{Y}_i = h_1 Y_1 + h_2 Y_2 + ... + h_i Y_i + ... + h_n Y_n$$

    Where $h_1, h_2, ..., h_n$ are known as <u>leverage values</u> or leverage weights $0 \le h_i \le 0$

  - The leverage values are computed by only using the X value(s).
  - A large h_i indicates that Y_i is particularly important in determining $\hat{Y}_j$.
  - But because the $h_i s$ are computed by only using the X value(s), h_i measures the role of the X value(s) in determining how important Y_i is in affecting $\hat{Y}_j$.
  - Thus, leverage values are helpful in identifying outlying X observations that influence $\hat{Y}$
  - To identify large leverage values, we compare h_i to the average leverage value. The standard rule of thumb is if the h_i is twice as large as the average leverage value, then X observation(s) for the i^{th} participant should be examined

    The average leverage value is:

$$\bar{h} = \frac{p}{n}$$

    Where p = the number of parameters in the regression model (2 for simple linear regression)
    n = the number of participants

    And so the rule-of -thumb cutoff value is:

$$h_i > \frac{2p}{n}$$

  - Other common cut-off values include
    - $h_i > .5$
    - Look for a large gap in the distribution of $h_i s$

- Cook's Distance (1979)
  - Cook's D is another measure of the influence an outlying observation has on the regression coefficients. It combines residuals and leverage values into a single number.

$$D_i = \frac{e_i^2}{p * MSE}\left[\frac{h_i}{(1-h_i)^2}\right]$$

Where: e_i is the (unstandardized) residual for the i^{th} observation
p is the number of parameters in the regression model
h_i is the leverage for the i^{th} observation

  - D_i for each observation depends on two factors:
    - The residual: Larger residuals lead to larger $D_i s$
    - The leverage: Larger leverage values lead to larger $D_i s$

  - The i^{th} observation can be influential (have a large D_i) by
    - Having a large e_i and only a moderate h_i
    - Having a moderate e_i and a large h_i
    - Having a large e_i and a large h_i

  - A D_i is considered to be large (indicating an influential observation) if it falls at or above the 50th percentile of the F-distribution

$F_{crit}(\alpha = .50, dfn, dfe)$
dfn = # of parameters in the model = p (2 for simple linear regression)
dfe = degrees of freedom for error = $N - p$

    - For example, with a simple linear regression model (*p = 2*) with 45 observations (*dfe = 45-2=43*)

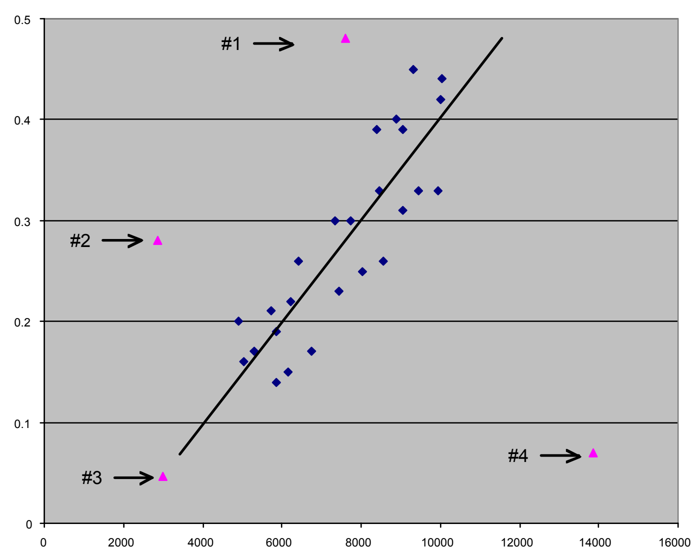
$$D_{crit} = F(\alpha = .50, dfn, dfe) = F(\alpha = .50, 2, 43) = .704$$

In this case, observations with Cook's D values greater than .704 should be investigated as possibly being influential

- Other methods of identifying outliers and influential observations exist to measure the influence of the i^{th} observation:
  - on each regression coefficient (DFBETAS)
  - on the predicted values (DFFITS)

- These methods of outliers and influence often work well, but can be ineffective at times. Ideally, the different procedures would identify the same cases, but this does not always happen. The use of these procedures requires thought and good judgment on the part of the analyst.

- Once influential points are identified:
  - Check to make sure there has not been a data coding or data entry error.
  - Conduct a sensitivity analysis to see how much your conclusions would change if the outlying points were dropped.
  - Never drop data points without telling your audience why those observations were omitted. In general, it is not advisable to drop observations from your analysis
  - The presence of many outliers may indicate an improper model
    - Perhaps the relationship is not linear
    - Perhaps the outliers are due to a variable omitted from the model



- Baseline example: No outliers included

```
REGRESSION
  /STATISTICS COEFF OUTS R ANOVA ZPP
  /DEPENDENT y
  /METHOD=ENTER x
  /CASEWISE PLOT(SRESID) ALL
 /SAVE PRED (pred) SRESID (sresid).
```

  - The regression model

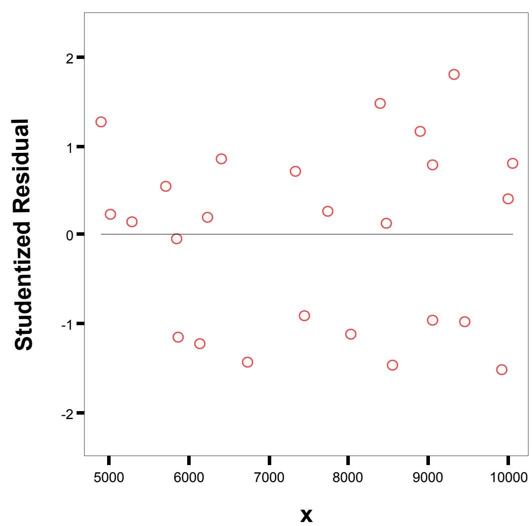
    $Y = -.104 + .0000506X$ $\qquad r_{XY} = .876$ $\qquad R^2 = .767$

  - Examining residuals

Casewise Diagnostics[a]

Case Number	Stud. Residual	Y	Predicted Value	Residual
1	.277	.30	.2870	.0130
2	1.805	.45	.3676	.0824
3	.786	.39	.3539	.0361
4	-1.518	.33	.3978	-.0678
5	1.170	.40	.3461	.0539
6	-.977	.33	.3744	-.0444
7	.131	.33	.3239	.0061
8	.414	.42	.4015	.0185
9	.805	.44	.4042	.0358
10	.712	.30	.2668	.0332
11	-.905	.23	.2723	-.0423
12	-.956	.31	.3539	-.0439
13	-1.116	.25	.3021	-.0521
14	1.491	.39	.3207	.0693
15	.196	.22	.2110	.0090
16	-1.155	.14	.1927	-.0527
17	.155	.17	.1631	.0069
18	-1.226	.15	.2063	-.0563
19	1.272	.20	.1440	.0560
20	-.035	.19	.1916	-.0016
21	.235	.16	.1496	.0104
22	.866	.26	.2200	.0400
23	.547	.21	.1851	.0249
24	-1.427	.17	.2363	-.0663
25	-1.465	.26	.3280	-.0680

a. Dependent Variable: Y



Look for Studentized Residuals larger than 2.5

- Examining influence statistics

```
REGRESSION
  /DEPENDENT y
  /METHOD=ENTER x
  /SAVE COOK (cook) LEVER (level) SDRESID (sdresid).
List var = ID cook level sdresid.
```

ID	COOK	LEVEL	SDRESID
1.00	.00162	.00029	.27175
2.00	.15078	.04468	1.90591
3.00	.02389	.03175	.77952
4.00	.15820	.08079	-1.56462
5.00	.04787	.02541	1.17947
6.00	.04820	.05181	-.97548
7.00	.00046	.01122	.12792
8.00	.01235	.08593	.40647
9.00	.04834	.08969	.79907
10.00	.01084	.00102	.70417
11.00	.01722	.00035	-.90124
12.00	.03537	.03179	-.95448
13.00	.02788	.00283	-1.12251
14.00	.05806	.00963	1.53438
15.00	.00139	.02770	.19142
16.00	.06141	.04435	-1.16348
17.00	.00162	.07952	.15130
18.00	.05796	.03156	-1.24063
19.00	.14004	.10758	1.29016
20.00	.00006	.04544	-.03471
21.00	.00443	.09888	.22963
22.00	.02431	.02094	.86075
23.00	.01523	.05234	.53871
24.00	.05487	.01111	-1.46223
25.00	.06055	.01339	-1.50504

- Critical values

  Cook's D: $D_{crit} = F(\alpha = .50, 2, 23) = .714$

  Leverage: $h_{crit} > \frac{2p}{N} = \frac{4}{25} = .16$

  Studentized Deletion Residuals: $\tilde{d}_{crit} > 2.5$

- In this case, we do not identify any outliers or influential observations

- Example #1: Outlier #1 included

```
REGRESSION
  /STATISTICS COEFF OUTS R ANOVA ZPP
  /DEPENDENT y
  /METHOD=ENTER x
  /CASEWISE PLOT(SRESID) ALL
 /SAVE PRED (pred) SRESID (sresid).
```

  - The regression model

    $Y = -.0967 + .0000506X$ $\qquad r_{XY} = .809$ $\qquad R^2 = .654$

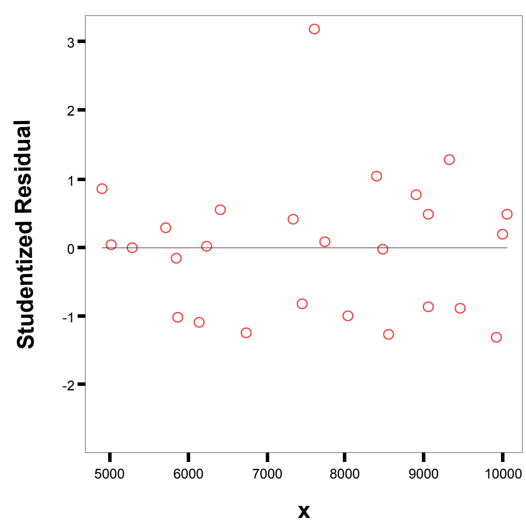
    (Slope is unchanged)

  - Examining residuals

Casewise Diagnostics[a]

Case Number	Stud. Residual	Y	Predicted Value	Residual
1	.087	.30	.2947	.0053
2	1.268	.45	.3753	.0747
3	.479	.39	.3616	.0284
4	-1.309	.33	.4055	-.0755
5	.777	.40	.3538	.0462
6	-.888	.33	.3821	-.0521
7	-.027	.33	.3316	-.0016
8	.187	.42	.4092	.0108
9	.490	.44	.4119	.0281
10	.424	.30	.2744	.0256
11	-.829	.23	.2800	-.0500
12	-.871	.31	.3616	-.0516
13	-.993	.25	.3098	-.0598
14	1.027	.39	.3284	.0616
15	.022	.22	.2187	.0013
16	-1.025	.14	.2004	-.0604
17	-.013	.17	.1708	-.0008
18	-1.080	.15	.2140	-.0640
19	.850	.20	.1517	.0483
20	-.158	.19	.1993	-.0093
21	.047	.16	.1573	.0027
22	.542	.26	.2277	.0323
23	.293	.21	.1928	.0172
24	-1.234	.17	.2440	-.0740
25	-1.264	.26	.3357	-.0757
26	3.189	.48	.2877	.1923

a. Dependent Variable: Y



Look for Studentized Residuals larger than 2.5
Observation #26 looks problematic

- Examining influence statistics

```
REGRESSION
  /DEPENDENT y
  /METHOD=ENTER x
  /SAVE COOK (cook) LEVER (level) SDRESID (sdresid).
List var = ID cook level sdresid.
```

ID	COOK	LEVEL	SDRESID
1.00	.00015	.00029	.08550
2.00	.07291	.04468	1.28518
3.00	.00868	.03175	.47159
4.00	.11600	.08079	-1.32978
5.00	.02059	.02541	.77023
6.00	.03909	.05181	-.88363
7.00	.00002	.01122	-.02646
8.00	.00249	.08593	.18329
9.00	.01765	.08969	.48210
10.00	.00370	.00102	.41665
11.00	.01386	.00035	-.82313
12.00	.02864	.03179	-.86611
13.00	.02122	.00283	-.99226
14.00	.02665	.00963	1.02828
15.00	.00002	.02770	.02160
16.00	.04745	.04435	-1.02631
17.00	.00001	.07952	-.01313
18.00	.04390	.03156	-1.08375
19.00	.06178	.10758	.84490
20.00	.00115	.04544	-.15495
21.00	.00018	.09888	.04601
22.00	.00926	.02094	.53355
23.00	.00428	.05234	.28711
24.00	.03972	.01111	-1.24846
25.00	.04367	.01339	-1.28045
26.00	.20343	.00000	**4.11310**

- Critical values

  Cook's D: $D_{crit} = F(\alpha = .50, 2, 24) = .695$

  Leverage: $h_{crit} > \frac{2p}{N} = \frac{4}{26} = .154$

  Studentized Deletion Residuals: $\tilde{d}_{crit} > 2.5$

- Observation #26
  - Has large residual and deletion residual
  - Has OK Cook's D and leverage

- Example #2: Only outlier #2 included

```
REGRESSION
  /STATISTICS COEFF OUTS R ANOVA ZPP
  /DEPENDENT y
  /METHOD=ENTER x
  /CASEWISE PLOT(SRESID) ALL
 /SAVE PRED (pred) SRESID (sresid).
```

  - The regression model

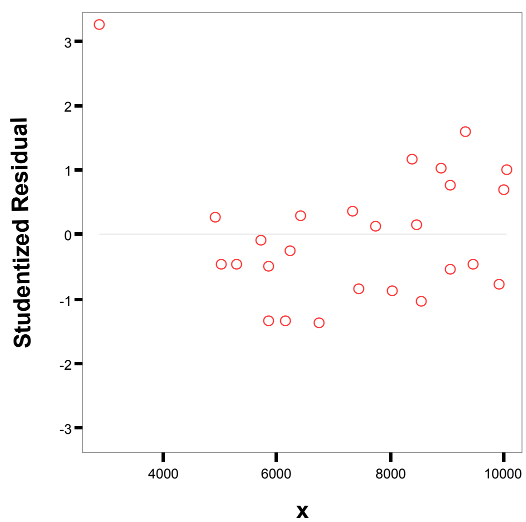
    $Y = -.00443 + .0000384X$ $\qquad r_{XY} = .762$ $\qquad R^2 = .564$

  - Examining residuals

Casewise Diagnostics[a]

Case Number	Stud. Residual	Y	Predicted Value	Residual
1	.126	.30	.2923	.0077
2	1.610	.45	.3534	.0966
3	.779	.39	.3430	.0470
4	-.784	.33	.3763	-.0463
5	1.040	.40	.3371	.0629
6	-.477	.33	.3585	-.0285
7	.160	.33	.3203	.0097
8	.695	.42	.3791	.0409
9	1.002	.44	.3811	.0589
10	.376	.30	.2769	.0231
11	-.833	.23	.2811	-.0511
12	-.547	.31	.3430	-.0330
13	-.877	.25	.3037	-.0537
14	1.184	.39	.3178	.0722
15	-.241	.22	.2347	-.0147
16	-1.336	.14	.2208	-.0808
17	-.475	.17	.1983	-.0283
18	-1.336	.15	.2311	-.0811
19	.273	.20	.1839	.0161
20	-.496	.19	.2200	-.0300
21	-.475	.16	.1881	-.0281
22	.304	.26	.2415	.0185
23	-.084	.21	.2151	-.0051
24	-1.371	.17	.2538	-.0838
25	-1.040	.26	.3233	-.0633
27	3.261	.28	.1059	.1741

a. Dependent Variable: Y



Look for Studentized Residuals larger than 2.5

Observation #27 looks problematic

- Examining influence statistics

```
REGRESSION
  /DEPENDENT y
  /METHOD=ENTER x
  /SAVE COOK (cook) LEVER (level) SDRESID (sdresid).
List var = ID cook level sdresid.
```

ID	COOK	LEVEL	SDRESID
1.00	.00033	.00116	.12296
2.00	.11241	.04134	1.66884
3.00	.02246	.03043	.77271
4.00	.03788	.07116	-.77795
5.00	.03662	.02499	1.04164
6.00	.01065	.04729	-.46875
7.00	.00068	.01244	.15652
8.00	.03099	.07536	.68703
9.00	.06647	.07842	1.00235
10.00	.00284	.00007	.36942
11.00	.01389	.00001	-.82777
12.00	.01107	.03046	-.53874
13.00	.01720	.00431	-.87309
14.00	.03644	.01097	1.19431
15.00	.00167	.01578	-.23622
16.00	.06241	.02692	-1.35925
17.00	.01110	.05117	-.46709
18.00	.05370	.01833	-1.35897
19.00	.00458	.07090	.26774
20.00	.00870	.02765	-.48791
21.00	.01299	.06476	-.46726
22.00	.00242	.01138	.29784
23.00	.00027	.03236	-.08220
24.00	.04294	.00525	-1.39763
25.00	.03021	.01441	-1.04229
27.00	**1.97793**	**.23268**	**4.27763**

- Critical values

  Cook's D: $D_{crit} = F(\alpha = .50, 2, 24) = .695$

  Leverage: $h_{crit} > \frac{2p}{N} = \frac{4}{26} = .154$

  Studentized Deletion Residuals: $\tilde{d}_{crit} > 2.5$

- Observation #27
  - Has large residual, deletion residual, Cook's D, and leverage

- Example #3: Only outlier #3 included

```
REGRESSION
  /STATISTICS COEFF OUTS R ANOVA ZPP
  /DEPENDENT y
  /METHOD=ENTER x
  /CASEWISE PLOT(SRESID) ALL
 /SAVE PRED (pred) SRESID (sresid).
```

  - The regression model

    $Y = -.105 + .0000506X$ $\qquad r_{XY} = .900$ $\qquad R^2 = .811$

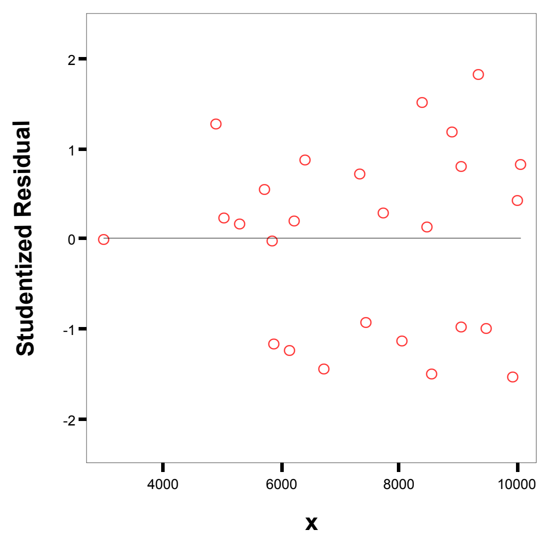
    (No change in slope or intercept)

  - Examining residuals

Casewise Diagnostics[a]

Case Number	Stud. Residual	Y	Predicted Value	Residual
1	.283	.30	.2870	.0130
2	1.839	.45	.3677	.0823
3	.802	.39	.3539	.0361
4	-1.542	.33	.3979	-.0679
5	1.193	.40	.3461	.0539
6	-.995	.33	.3744	-.0444
7	.133	.33	.3239	.0061
8	.419	.42	.4016	.0184
9	.816	.44	.4042	.0358
10	.727	.30	.2667	.0333
11	-.923	.23	.2723	-.0423
12	-.976	.31	.3539	-.0439
13	-1.140	.25	.3021	-.0521
14	1.523	.39	.3207	.0693
15	.199	.22	.2110	.0090
16	-1.167	.14	.1926	-.0526
17	.157	.17	.1630	.0070
18	-1.242	.15	.2063	-.0563
19	1.273	.20	.1439	.0561
20	-.035	.19	.1916	-.0016
21	.236	.16	.1496	.0104
22	.880	.26	.2200	.0400
23	.554	.21	.1851	.0249
24	-1.452	.17	.2363	-.0663
25	-1.497	.26	.3280	-.0680
28	-.007	.05	.0473	-.0003

a. Dependent Variable: Y



Look for Studentized Residuals larger than 2.5

All observations are OK

- Examining influence statistics

```
REGRESSION
  /DEPENDENT y
  /METHOD=ENTER x
  /SAVE COOK (cook) LEVER (level) SDRESID (sdresid).
List var = ID cook level sdresid.
```

ID	COOK	LEVEL	SDRESID
1.00	.00166	.00114	.27793
2.00	.14731	.04166	1.94253
3.00	.02385	.03063	.79550
4.00	.14727	.07179	-1.59008
5.00	.04836	.02514	1.20442
6.00	.04665	.04767	-.99473
7.00	.00048	.01248	.13066
8.00	.01137	.07603	.41213
9.00	.04440	.07913	.81044
10.00	.01059	.00008	.71940
11.00	.01705	.00001	-.92025
12.00	.03536	.03067	-.97485
13.00	.02903	.00430	-1.14763
14.00	.06035	.01100	1.56861
15.00	.00114	.01611	.19514
16.00	.04801	.02743	-1.17608
17.00	.00122	.05206	.15340
18.00	.04678	.01870	-1.25721
19.00	.10074	.07207	1.29087
20.00	.00004	.02818	-.03419
21.00	.00325	.06584	.23152
22.00	.02042	.01164	.87563
23.00	.01179	.03296	.54551
24.00	.04835	.00539	-1.48820
25.00	.06259	.01447	-1.53863
28.00	.00001	**.22342**	-.00664

- Critical values
  - Cook's D: $D_{crit} = F(\alpha = .50,2,24) = .695$
  - Leverage: $h_{crit} > \frac{2p}{N} = \frac{4}{26} = .154$
  - Studentized Deletion Residuals: $\tilde{d}_{crit} > 2.5$

- Observation #28
  - Has large leverage
  - Has OK residual, deletion residual, and Cook's D

- Example #4: Only outlier #4 included

```
REGRESSION
  /STATISTICS COEFF OUTS R ANOVA ZPP
  /DEPENDENT y
  /METHOD=ENTER x
  /CASEWISE PLOT(SRESID) ALL
 /SAVE PRED (pred) SRESID (sresid).
```

  - The regression model

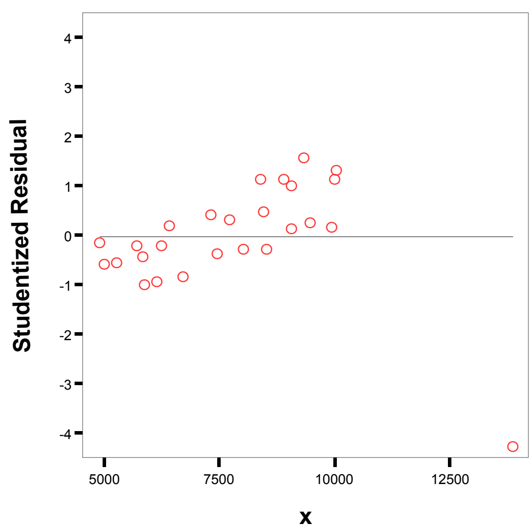
    $Y = .113 + .0000203X$ $\quad r_{XY} = .403$ $\quad R^2 = .162$

  - Examining residuals

Casewise Diagnostics[a]

Case Number	Stud. Residual	Y	Predicted Value	Residual
1	.319	.30	.2699	.0301
2	1.580	.45	.3022	.1478
3	.994	.39	.2967	.0933
4	.169	.33	.3143	.0157
5	1.132	.40	.2936	.1064
6	.269	.33	.3049	.0251
7	.480	.33	.2847	.0453
8	1.128	.42	.3158	.1042
9	1.334	.44	.3169	.1231
10	.405	.30	.2617	.0383
11	-.359	.23	.2639	-.0339
12	.142	.31	.2967	.0133
13	-.274	.25	.2759	-.0259
14	1.129	.39	.2834	.1066
15	-.207	.22	.2393	-.0193
16	-.992	.14	.2320	-.0920
17	-.547	.17	.2201	-.0501
18	-.938	.15	.2375	-.0875
19	-.137	.20	.2124	-.0124
20	-.448	.19	.2315	-.0415
21	-.602	.16	.2147	-.0547
22	.182	.26	.2429	.0171
23	-.205	.21	.2289	-.0189
24	-.845	.17	.2495	-.0795
25	-.279	.26	.2863	-.0263
29	-4.287	.07	.3944	-.3244

a. Dependent Variable: Y



Look for Studentized Residuals larger than 2.5

Observation #29 is clearly problematic

- Examining influence statistics

```
REGRESSION
  /DEPENDENT y
  /METHOD=ENTER x
  /SAVE COOK (cook) LEVER (level) SDRESID (sdresid).
List var = ID cook level sdresid.
```

ID	COOK	LEVEL	SDRESID
1.00	.00204	.00010	.31270
2.00	.07917	.02119	1.63404
3.00	.02744	.01416	.99364
4.00	.00125	.04155	.16580
5.00	.03319	.01081	1.13871
6.00	.00245	.02513	.26350
7.00	.00508	.00375	.47244
8.00	.05754	.04450	1.13456
9.00	.08286	.04667	1.35774
10.00	.00350	.00241	.39819
11.00	.00268	.00148	-.35230
12.00	.00056	.01418	.13866
13.00	.00151	.00036	-.26839
14.00	.02759	.00302	1.13603
15.00	.00144	.02454	-.20276
16.00	.04004	.03687	-.99116
17.00	.01675	.06205	-.53933
18.00	.03104	.02744	-.93564
19.00	.00129	.08174	-.13450
20.00	.00827	.03766	-.44047
21.00	.02337	.07567	-.59412
22.00	.00102	.01940	.17853
23.00	.00185	.04266	-.20071
24.00	.01885	.01163	-.84022
25.00	.00175	.00476	-.27318
29.00	**5.74625**	**.34626**	**-8.67294**

- Critical values

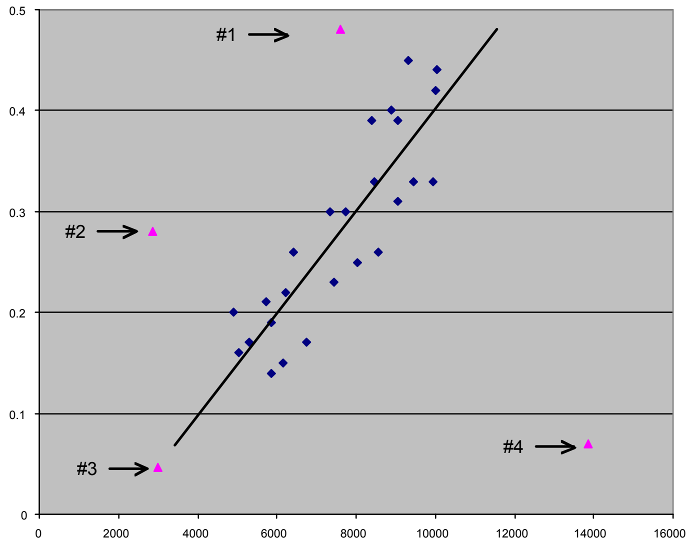
  Cook's D: $D_{crit} = F(\alpha = .50, 2, 24) = .695$

  Leverage: $h_{crit} > \frac{2p}{N} = \frac{4}{26} = .154$

  Studentized Deletion Residuals: $\tilde{d}_{crit} > 2.5$

- Observation #29
  - Has large residual, deletion residual, Cook's D, and leverage

- Summary and comparison:



Obs	Regression Equation	r_{XY}	Problematic? $\tilde{e}_i$	d_i	D_i	h_i
Baseline	$Y = -.104 + .0000506X$	$r_{XY} = .876$	No	No	No	No
#1	$Y = -.097 + .0000506X$	$r_{XY} = .809$	Yes	Yes	No	No
#2	$Y = -.004 + .0000384X$	$r_{XY} = .762$	Yes	Yes	Yes	Yes
#3	$Y = -.105 + .0000506X$	$r_{XY} = .900$	No	No	No	Yes
#4	$Y = .113 + .0000203X$	$r_{XY} = .403$	Yes	Yes	Yes	Yes

  - Using a combination of all the methods, we (properly) identify outliers #2 and #4 as problematic. Outlier #1 may or may not be problematic, depending on our purposes.

6. Remedial Measures: An overview of alternative regression models

- When regression assumptions are violated, you have two options
    - o Explore transformations of X or Y so that the simple linear regression model can be used appropriately
    - o Abandon the simple linear regression model and use a more appropriate model.

- More complex regression models are beyond the scope of this course, but let me highlight some possible alternative models that could be explored.

- Polynomial regression
    - o When the regression function is not linear, a model that has non-linear terms may better fit the data.

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + ... + \beta_k X^k + \varepsilon$$

- o In these models, we are fitting/estimating non-linear regression lines that correspond with polynomial curves. This approach is very similar to the trend contrasts that we conducted in ANOVA except:
    - For polynomial regression, the predictor variable (IV) is continuous; in ANOVA it is categorical.
    - In polynomial regression, we obtain the actual equation of the (polynomial) line that best fits the data.

- Weighted least squares regression
    - o The regression question we have been using is known as ordinary least squares (OLS) regression. In the OLS framework, we solve for the model parameters by minimizing the squared residuals (squared deviations from the predicted line).

$$SSE = \sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - b_0 - b_1 X_i)^2$$

- When we minimize the residuals, solve for the parameters, and compute p-values, we need the residuals to have equal variance across the range of X values. If this equal variance assumption is violated, then the OLS regression parameters will be biased.

- In OLS regression, each observation is treated equally. But if some observations are more precise than others (i.e., they have smaller variance), it make sense that they should be given more weight than the less precise values.

- In weighted least squares regression, each observation is weighted by the inverse of its variance

$$w_i = \frac{1}{\sigma_i^2}$$

$$SSE = \sum w_i e_i^2 = \sum w_i (Y_i - \hat{Y}_i)^2 = \sum w_i (Y_i - b_0 - b_1 X_i)^2$$

  - Observations with a large variance are given a small weight; observations with a small variance are given a big weight.

- Issues with weighted least squares regression
  - We do not know the variances of the residuals; this value must be estimated. The process of estimating these variances is not trivial – particularly in small datasets.
  - R^2 is uninterpretable for weighted least squares regression (but that does not stop most programs from printing it out!).

- Robust regression
  - When assumptions are violated and/or outliers are present, OLS regression parameters may be biased. Robust regression is a series of approaches that computes estimates of regression parameters using techniques that are robust to violations of OLS assumptions

  - Least absolute residual (LAR) regression estimates regression parameters by minimizing the sum of the absolute deviations of the Y observations from the regression line:

$$\sum |e_i| = \sum |Y_i - \hat{Y}|_i = \sum |Y_i - b_0 - b_1 X_i|$$

    - This approach reduces the influence of outliers

  - Least median of squares (LMS) regression estimates regression parameters by minimizing the median of the squared deviations of the Y observations from the regression line:

$$SSE = Median(e_i^2) = median(Y_i - \hat{Y})^2 = median(Y_i - b_0 - b_1 X_i)^2$$

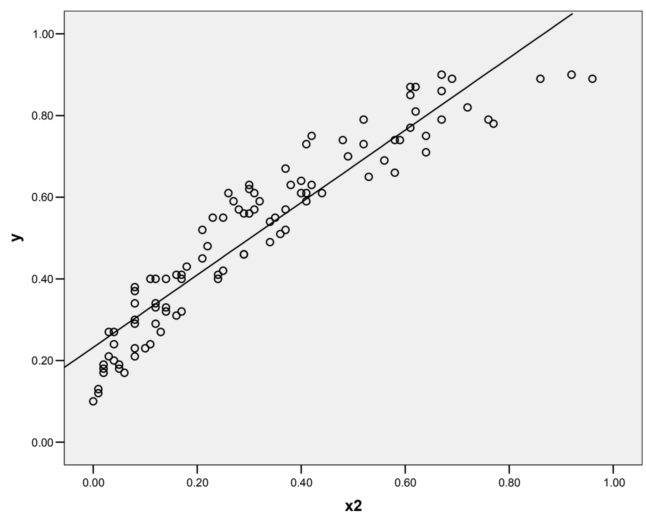
    - This approach also reduces the influence of outliers

  - Iterative reweighed least squares (IRLS) regression is a form of weighted least squares regression where the weights for each observation are M-estimators of the residuals (Huber and Tukey-Bisquare estimators are the most commonly used M-estimators).
    - This approach reduces the influence of outliers

  - Disadvantages of these approaches include:
    - They are not commonly included in statistical software
    - They are robust to outliers, but they require that other regression assumptions be satisfied.
    - They are not commonly used in psychology and, thus, psychologists may regard these methods skeptically

- Non-parametric regression
    - Non-parametric regression techniques do not estimate model parameters. For a regression analysis, we are usually interested in estimating a slope, thus non-parametric methods are of limited utility from an inferential perspective.
    - In general, non-parametric techniques can be used to explore the shape of the regression line. These methods provide a smoothed curve that fits the data, but do not provide an equation for this line or allow for inference on this line.

    - Previously, we examined the following data and concluded that the relationship between *X2* and *Y* was non-linear (see p. 14-8)

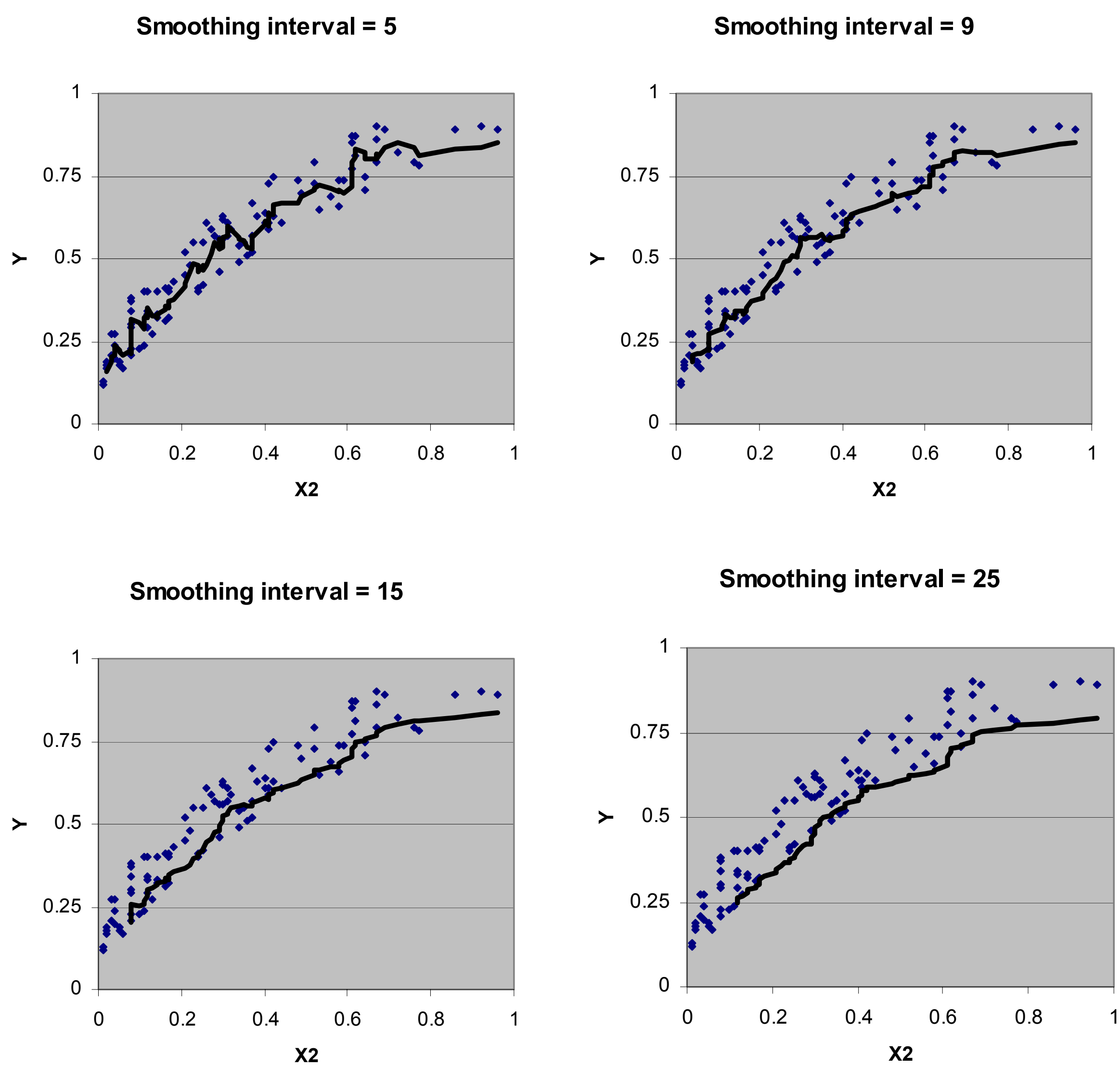


        - Let's examine this data with non-parametric techniques to explore the relationship between *X2* and *Y*.

- Method of moving averages
    - The method of moving averages can be used to obtain a smoothed curve that fits the data.
    - To use this method, you must specify a window (*w*) – the number of observations you will average across. First, average the Y values associated with the first *w* responses (the *w* smallest *X* values), and plot that point. Next, you discard the first value, add the next point along the X axis, compute the average of this set of *w* Y values, and plot that point. Continue moving the down the X axis, until you have used all the points. Then draw a line to connect the smoothed average points.

- For example, if $w = 3$, then you take the three smallest X values, average the Y values associated with these points, and plot that point. Next, take the 2nd, 3rd, and 4th smallest X values and repeat the process . . .

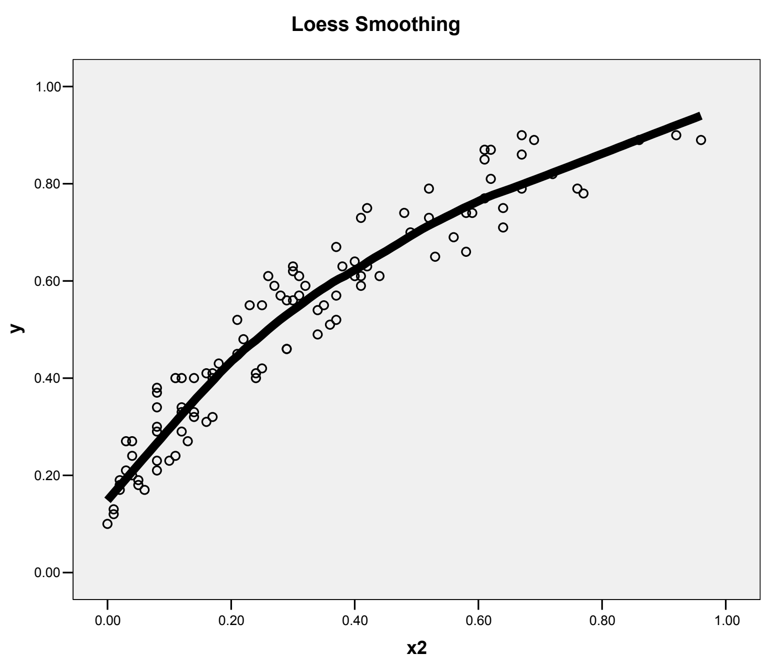
- An example of the method of moving averages, comparing different w values:



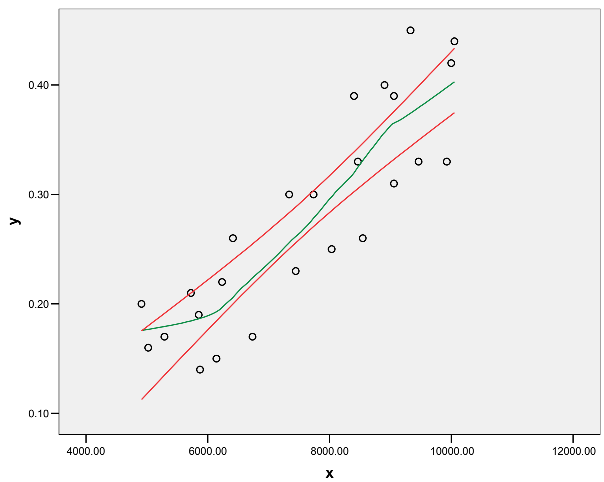
    - If the window is too small, it does not smooth the data enough; if the window is too large, it can smooth too much and you lose the shape of the data.
    - In this case, $w = 15$ looks about right.

- Issues regarding the method of moving averages:
  - Averages are affected by outliers, so it can be preferable to use the method of moving medians
  - The method of moving averages is particularly useful for time-series data
  - If the data are unequally spaced and/or have gaps along the X axis, the method of moving averages can provide some wacky results.
  - In EXCEL, you can use the method of moving averages and you can specify *w*.

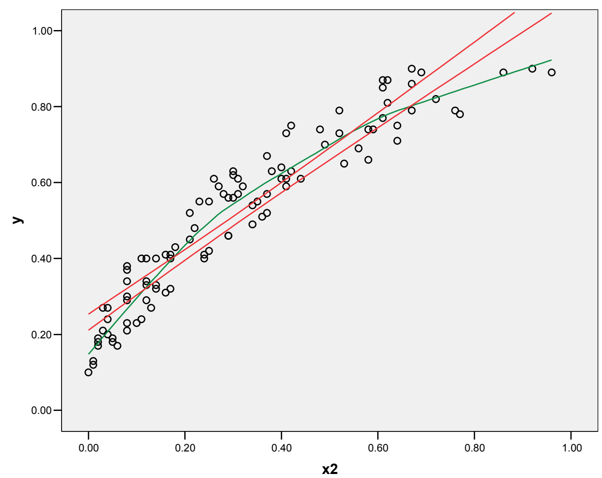
- Loess smoothing
  - Loess stands for "locally weighted scatterplot smoothing" (the w got dropped somewhere along the way).
  - Loess smoothing is a more sophisticated method of smoothing than the method of moving averages.
    - In each "neighborhood", a regression line is estimated and the fitted line is used for the smoothed line
    - This regression is weighted to give cases further from the middle X value less weight.
    - This process of fitting a linear regression line is repeated in each neighborhood so that observations with large residuals in the previous iteration receive less weight.



- One particularly useful application of loess smoothing is to confirm a fitted regression function
  - Fit a regression function and graph 95% confidence bands for the fitted line
  - Fit a loess smoothed curve through the data.
  - If the loess curve stays within the confidence bands, the fit of the regression line is good. If the loess curve strays from the confidence bands, the fit of the regression line is not good.



Good Fit



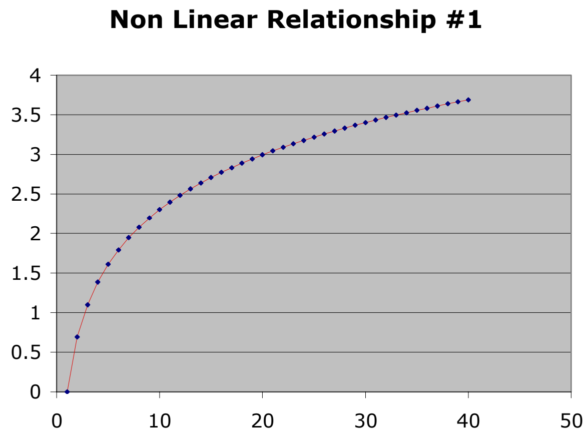
Poor Fit

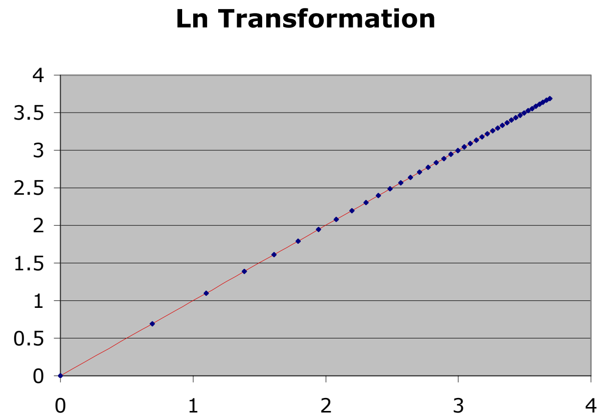
7. Remedial Measures: Transformations

- If the data do not satisfy the regression assumptions, a transformation applied to either the X variable or to the Y variable may make the simple linear regression model appropriate for the transformed data.

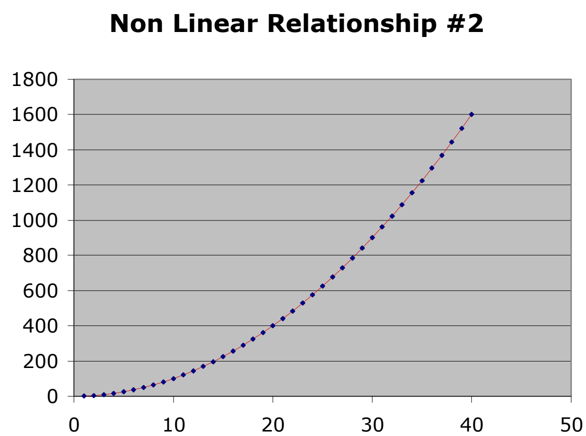
- General rules of thumb:
  - Transformations on X
    - Can be used to linearize a non-linear relationship
  - Transformations on Y
    - Can be used to fix problems of nonnormality and unequal error variances
    - Once normality and homoscedasticity are achieved, it may be necessary to transform X to achieve linearity

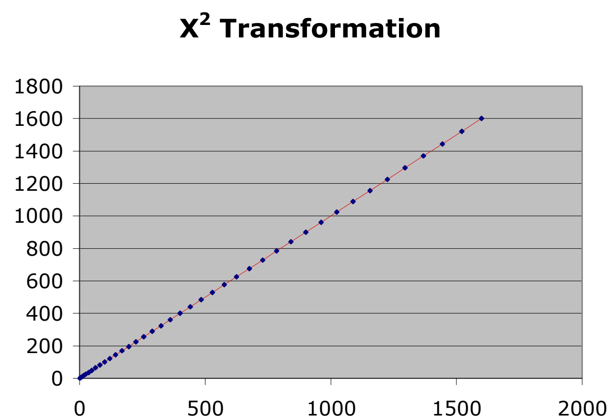
- Prototypical patterns and transformations of X



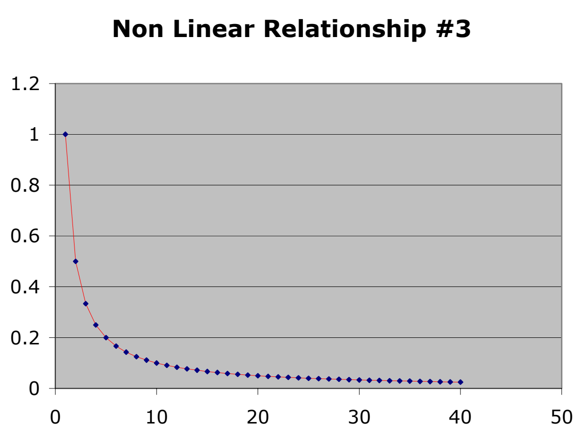


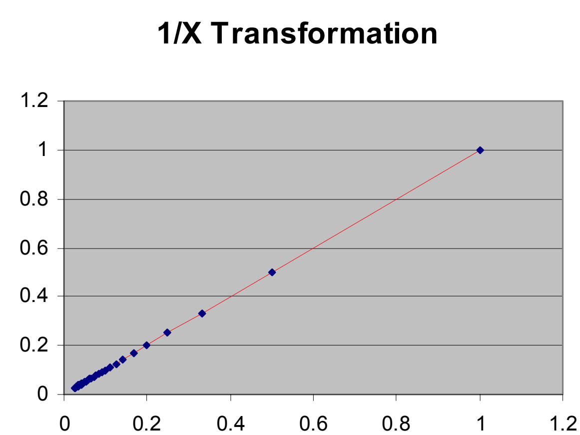
  - Try $X' = \ln(X)$ or $X' = \sqrt{X}$





  - Try $X' = X^2$ or $X' = \exp(X)$





  - Try $X' = 1/X$ or $X' = \exp\left(\frac{1}{X}\right)$

- Transforming Y
    - If the data are non-normal and/or heteroscedasticitic, a transformation on Y may be useful
    - It can be very difficult to determine the most appropriate transformation in Y to fix the data

    - One popular class of transformations is the family of power transformations

$$Y' = Y^{\lambda}$$

$\lambda = 2$	$Y' = Y^2$	
$\lambda = 1$	$Y' = Y$	
$\lambda = .5$	$Y' = \sqrt{Y}$	
$\lambda = 0$	$Y' = \ln(Y)$	by definition
$\lambda = -.5$	$Y' = 1/\sqrt{Y}$	
$\lambda = -1$	$Y' = 1/Y$	
$\lambda = -2$	$Y' = 1/Y^2$	

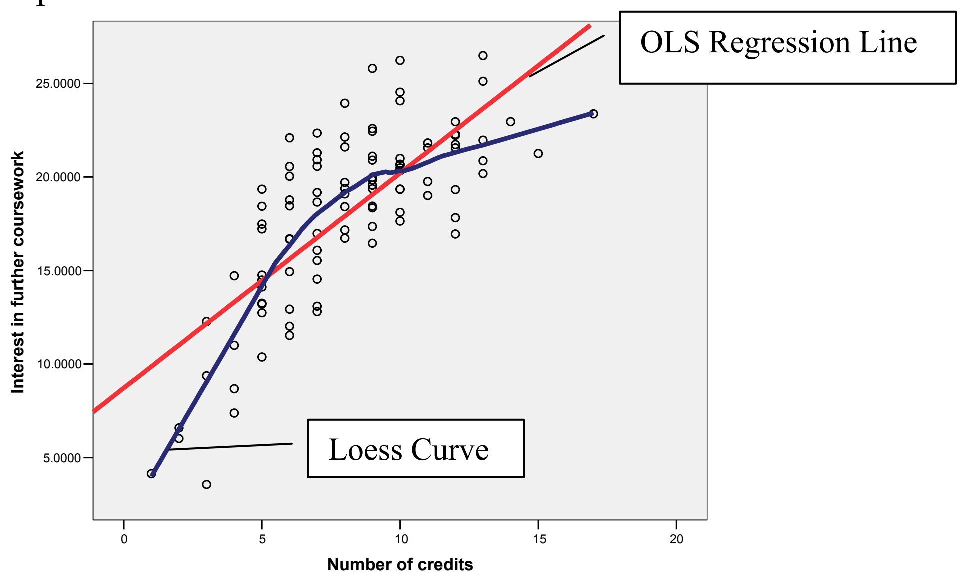
To determine a λ that works:

- Guess (trial and error)
- Use the Box-Cox procedure (unfortunately not implemented in SPSS)

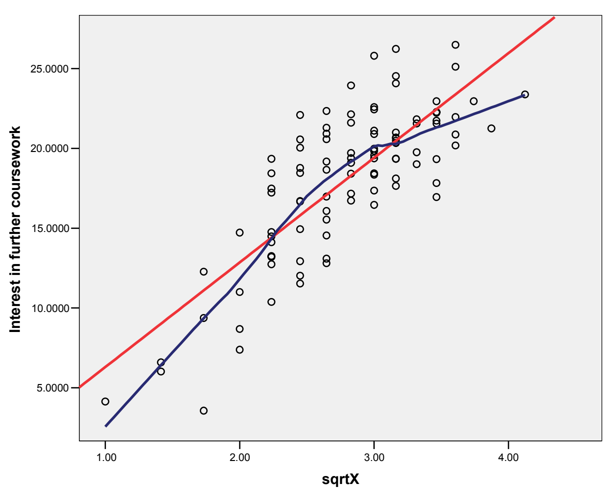
- Warnings and cautions about transformations
    - Do not transform the data because of a small number of outliers
    - After transforming the data, recheck the fit of the regression model using residual analysis
    - Once the data are transformed, and a regression run on the transformed data, b_0 and b_1 apply to the transformed data and not to the original data/scale
    - For psychological data, if the original data are not linear, but the transformed data are, it can often be very difficult to interpret the results

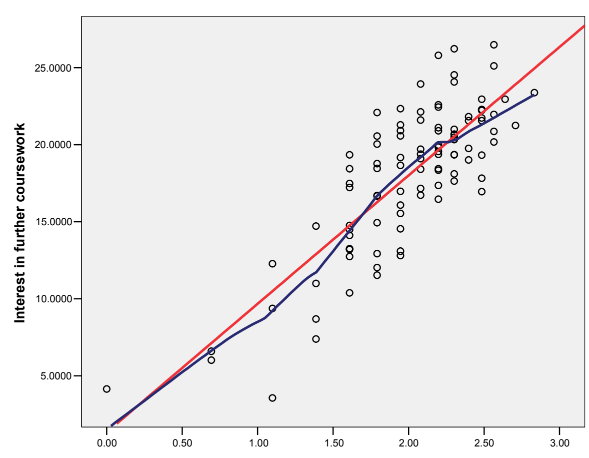
- Transformations: An example
  - Let's examine the relationship between the number of credits taken in a minor and interest in taking further coursework in that discipline.
  - A university collects data on 100 students
    - X = Number of credits completed in the minor
    - Y = Interest in taking another course in the minor discipline

  - First, let's plot the data



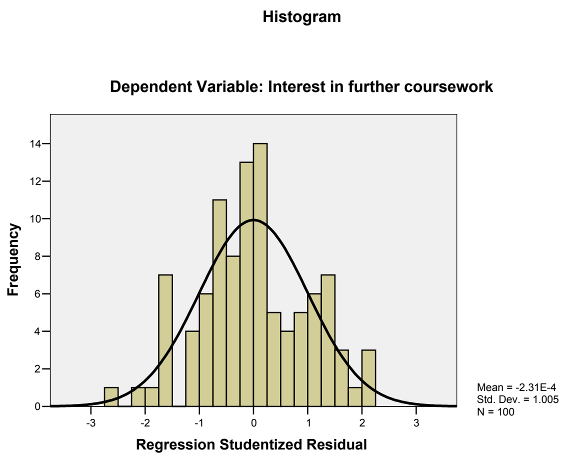
    - This relationship looks non-linear.
    - We can try a square root or a log transformation to achieve linearity.

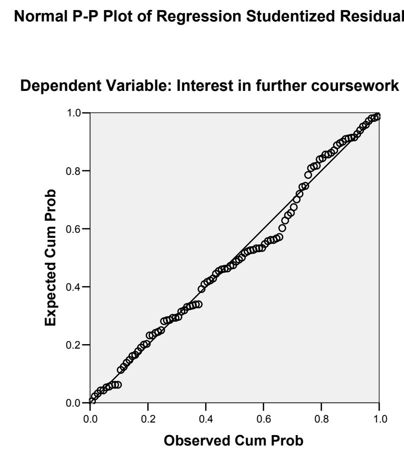




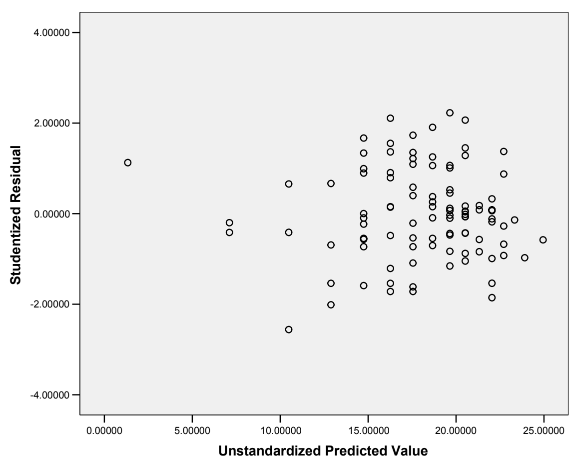
  - The log transformation appears to work, so we should check the remaining assumptions

```
REGRESSION
  /DEPENDENT ybr
  /METHOD=ENTER lnx
  /RESIDUALS HIST(SRESID) NORM(SRESID)
  /SAVE RESID (resid1) ZRESID (zresid1)  SRESID (sresid1) pred (pred1).
```





- The residuals appear to be normally distributed



- We might worry about an outlier, but homoscedasticity seems ok.

- Now, we can analyze the ln-transformed data.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.813[a]	.661	.658	2.7779456

a. Predictors: (Constant), lnx

b. Dependent Variable: Interest in further coursework

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1.337	1.244		1.075	.285
	lnx	8.335	.603	.813	13.828	.000

a. Dependent Variable: Interest in further coursework

  - There is a strong linear relationship between ln of credits taken and interest in taking additional courses in the discipline, $\beta = .81, t(98) = 13.83, p < .01, R^2_{Adjusted} = .66$

- But in this case, the non-linear relationship is interesting (and interpretable). We would be better off with an approach where we could model the non-linearity than with this approach where we try to transform to linearity.
- In this case, polynomial regression may be very useful.

- Note that if we had not graphed or explored our data, we would have missed the non-linear relationship altogether!

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.749[a]	.561	.556	3.1625661

a. Predictors: (Constant), Number of credits

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	8.718	.897		9.721	.000
	Number of credits	1.149	.103	.749	11.187	.000

a. Dependent Variable: Interest in further coursework