Chapter 13
Simple Linear Regression

The Regression Equation

1. Overview of regression analysis

   - Regression analysis is generally used when both the independent and the dependent variables are continuous. (But modifications exist to handle categorical independent variables and dichotomous dependent variables.)

Type of Analysis	Independent Variable	Dependent Variable
ANOVA	Categorical	Continuous
Regression	Continuous or Categorical	Continuous
Categorical Analysis (Contingency Table Analysis)		Categorical

   - Goals of regression analysis:
     - To describe the relationship between two variables
     - To model responses on a dependent variable
     - To predict a dependent variable using one or more independent variables
     - To statistically control the effects of variables while examining the relationship between the independent and dependent variable

   - Regression analysis is usually performed on observational data. In these cases, we describe, model, predict, and control, but we cannot make any causal claims regarding these relationships

- Terminology in regression analysis
  - As in ANOVA, we will develop a model to explain the data

$$DATA = MODEL + ERROR$$

  - The model assumes greater importance in regression. Unlike ANOVA, we are usually interested in the model parameters
  - The goal of most regression models is to use the information contained in a set of variables to predict a response. As a result, we use slightly different terminology in regression, compared to ANOVA.

ANOVA	REGRESSION
Dependent variable	Dependent variable or Response variable or Outcome variable
Independent variables	Independent variables or Predictor variables

Simple Linear Regression
The Regression Equation

2. Linear Functions

- The goal of simple linear regression is to describe an outcome variable (Y) as a linear function of a predictor variable (X).

- The end result will be a model that defines the equation of a straight line

$$Y = b + aX$$

Where

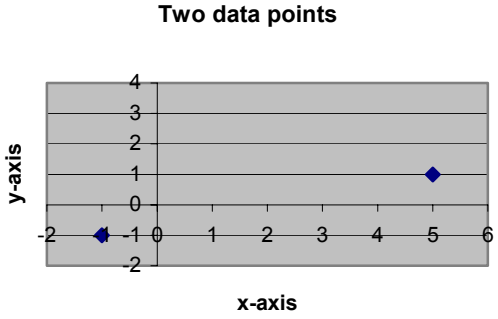
b = the y-intercept

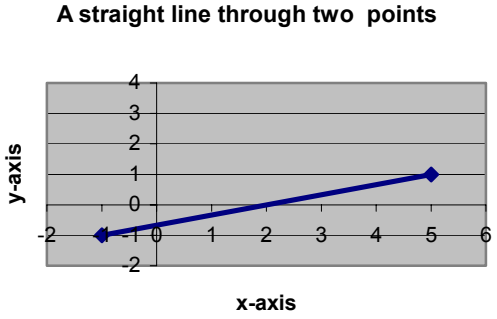
a = the slope

  - Let's consider a simple example: $Y = -1 + \frac{X}{3}$

    - The y-intercept is -1
      $\Rightarrow$ The line crosses the y-axis at $y = -1$

    - The slope of the line is *1/3*
      $\Rightarrow$ The slope is a measure of the steepness of the line
      $\Rightarrow$ The slope is the change in y associated with a 1 unit change in x





- Let's review the method covered in high school algebra for determining the line that falls through 2 points: $(-1,-1)$ & $(5,1)$

  - First, we compute the slope of the line

$$slope = \frac{y_2 - y_1}{x_2 - x_1}$$

$$slope = \frac{1-(-1)}{5-(-1)} = \frac{2}{6} = .333$$

We interpret the slope as the change in y associated with a 1 unit change in x

In this example, for every 1unit increase in x, y will increase by .333

x	-1	0	1	2	3	4	5
y	-1	-.667	-.333	0	.333	.667	1

  - We compute the y-intercept by finding the value of y when x = 0

We can use: the equation for the slope of a line and the (x, y) coordinates of either known point to solve for $(0, y_0)$

Let's use $(5,1)$

$$.333 = \frac{1 - y_0}{5 - 0}$$

$$.333(5) = 1 - y_0$$
$$y_0 = 1 - 1.667$$
$$y_0 = -.667$$

  - Finally, we use the slope and the intercept to write the equation of the line through the 2 points

$$Y = b + aX$$
$$Y = -.667 + .333(X)$$

3. Estimation and interpretation of model parameters

- With real data, the points rarely fall directly on a straight line. Regression is a technique to estimate the slope and the y-intercept from noisy data

- Because not every point will fall on the regression line, there will be error in our model

$$DATA = MODEL + ERROR$$

  - The *DATA*, or the outcome we want to predict is the Y variable
  - The *MODEL* is the equation of the regression line, $b_0 + b_1 X_1$

    b_0 = the population value of the intercept

    b_1 = the population value of the slope

    X_1 = the predictor variable

  - The *ERROR* is deviation of the observed data from our regression line. We refer to the individual error terms as residuals

  - The full simple linear regression model is given by the following equation:

$$DATA = MODEL + ERROR$$
$$Y = b_0 + b_1 X_1 + \varepsilon$$

- Some key characteristics of this model
  - We can only model linear relationships between the outcome variable and the predictor variable
  - The model can be expanded to include the linear relationships between multiple predictor variables and a single outcome

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k + \varepsilon$$

- Predicted values and residuals

  - With real data, we need to estimate the value of the slope and the intercept. (Details on the estimation process will follow shortly.)

$$Y = \hat{b}_0 + \hat{b}_1 X_1 + \varepsilon$$

$\hat{b}_0$ = the estimated value of the intercept

$\hat{b}_1$ = the estimated value of the slope

  - Based on the model, we have a "best guess" as to the participant's response on the outcome variable

$$\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_{1i}$$

    - In other words, we use the equation of the line we have developed to estimate how each participant responded on the outcome variable
    - $\hat{Y}_i$ is called the predicted value or fitted value for the i^{th} participant

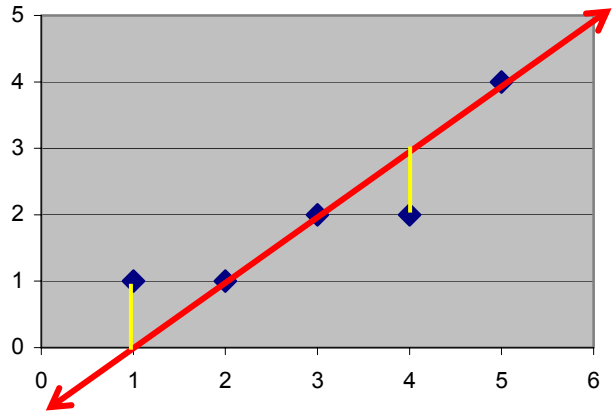
  - If the actual response of the participant deviates from our predicted value, then we have some *ERROR* in the model. We define the residual to be the deviation of the observed value from the predicted value.

$$DATA = MODEL + ERROR$$
$$Y_i = \left(\hat{b}_0 + \hat{b}_1 X_{1i}\right) + e_i$$
$$Y_i = \hat{Y}_i + e_i$$
$$e_i = Y_i - \hat{Y}_i$$

  - If we want to know if our model is a "good" model, we can examine the residuals.
    - If we have many large residuals, then there are many observations that are not predicted well by the model. We say that the model has a poor fit.
    - If most of the residuals are small, then our model is very good at explaining responses on the *Y* variable. This model would have a good fit.

- Let's consider a simple example to illustrate these points

Y	X_1
1	1
1	2
2	3
2	4
4	5



- We notice that a straight line can be drawn that goes directly through three of the 5 observed data points. Let's use this line as our best guess line

$$\tilde{Y} = -1 + X$$

- Now we can calculate predicted values and residuals

Y	X	$\hat{Y}$	e
1	1	0	1
1	2	1	0
2	3	2	0
2	4	3	-1
4	5	4	0

- In the previous example, we "eyeballed" a regression line. We would like to have a better method of estimating the regression line. Let's consider desirable two properties of a good regression line

  i. The sum of the residuals should be zero
  $$\sum (y_i - \hat{y}) = 0$$
     - If we have this property, then the average residual would be zero
     - In other words, the average deviation from the predicted line would be zero

  ii. Overall, we would like the residuals to be as small as possible
     - We already require the residuals to sum to zero, by property (i).
     - So, let's require the sum of the squared residuals to be as small as possible. This approach has the added benefit of "penalizing" large residuals more than small residuals
     $$\sum (y_i - \hat{y}_i)^2 = minimum$$

  - Estimating a regression line using these two properties is called the ordinary least squares (OLS) estimation procedure
  - Estimates of the intercept and slope are called the ordinary least squares (OLS) estimates

  - To solve for these estimates, we can use the following procedure
     - We want to minimize $SSE = \sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - b_0 - b_1 X_i)^2$
     - We take the derivatives of *SSE* with respect to b_0 and b_1, set each equal to zero, and solve for b_0 and b_1
     $$\frac{\partial SSE}{\partial b_0} = 0 \qquad \text{and} \qquad \frac{\partial SSE}{\partial b_1} = 0$$

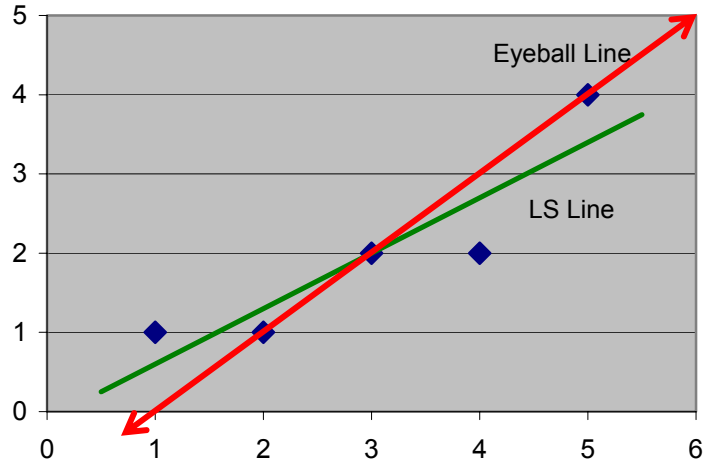
     - We'll skip the details and jump to the final estimates

$$\hat{b}_1 = \frac{SS_{XY}}{SS_{XX}} \qquad\qquad \hat{b}_0 = \overline{Y} - \beta_1 \overline{X}$$

Where $$SS_{XX} = \sum_{i-1}^{n} (X_i - \overline{X})^2$$

$$SS_{XY} = \sum_{i-1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})$$

- Now, let's return to our example and examine the least squares regression line



- Let's compare the least squares regression line to our eyeball regression line

$$\tilde{Y} = -1 + X$$
$$\hat{Y} = -.1 + .7X$$

	Data		Eyeball			Least Squares		
	Y	X	$\tilde{Y}$	$\tilde{e}$	$\tilde{e}^2$	$\hat{Y}$	$\hat{e}$	$\hat{e}^2$
	1	1	0	1	2	0.6	.4	.16
	1	2	1	0	0	1.3	-.3	.09
	2	3	2	0	0	2.0	0	0
	2	4	3	-1	2	2.7	-.7	.49
	4	5	4	0	0	3.4	.6	.36
$\sum e_i$				0			0	
$\sum e_i^2$					4			1.1

  - For both models, we satisfy the condition that the residuals sum to zero
  - But the least squares regression line produces the model with the smallest squared residuals

- Note that other regression lines are possible
  - We could minimize the absolute value of the residuals
  - We could minimize the shortest distance to the regression line

4. Inference on the model parameters

- We have learned how to estimate the model parameters, but also want to perform statistical tests on those parameters

$$\hat{b}_1 = \frac{SS_{XY}}{SS_{XX}} \qquad \hat{b}_0 = \overline{Y} - \hat{b}_1 \overline{X}$$

- First, let's estimate the amount of random error in the model, σ^2
  - Intuitively, the greater the amount of random error in a sample, the more difficult it will be to estimate the model parameters.
  - The random error in the model is captured in the residuals, e_i
    - We need to calculate the variance of the residuals
    - Recall a variance is the average squared deviation from the mean
    - When applied to residuals, we obtain

$$Var(\varepsilon_i) = \frac{\sum (\varepsilon_i - \overline{\varepsilon})^2}{N-2}$$

But we know $\overline{\varepsilon} = 0$

$$\hat{Var}(\varepsilon_i) = \hat{\sigma}_\varepsilon^2 = \frac{\sum (\hat{\varepsilon}_i)^2}{N-2} = \frac{\sum (Y_i - \hat{Y}_i)^2}{N-2} = \frac{SSResidual}{N-2}$$

- Why use *N-2*?
  ⇒ A general heuristic is to use *N – (number of parameters fitted)*
  ⇒ In this case, we have estimated two parameter: the slope and the intercept
  ⇒ Recall that for *Var(X)*, we divided by *N-1.* We only estimated one parameter (the grand mean)
  ⇒ This heuristic also applied for ANOVA.

- And so we are left with

$$\hat{Var}(\varepsilon_i) = \frac{\sum(\hat{\varepsilon}_i)^2}{N-2} = \frac{SSresid}{N - \#\ of\ \ parameters} = MSresid$$

  And we are justified using $MSresid$ as the error term for tests involving the regression model

  - Interpreting *MSresid*:
    - Residuals measure deviation from regression line (the predicted values)
    - The variance of the residuals captures the average squared deviation from the regression line
    - So we can interpret $\sqrt{MSresid}$ as a measure of average deviation from the regression line. SPSS labels $\sqrt{MSresid}$ as "standard error of the estimate"

- Now that we have an estimate of the error variance, we can proceed with statistical tests of the model parameters

- We can perform a t-test using our familiar t-test formula

$$t \sim \frac{estimate}{standard\ error\ of\ \ the\ estimate}$$

  - We know how to calculate the estimates of the slope and the intercept. All we need are standard errors of the estimates

- Inferences about the slope, $\hat{b}_1$
  - Deriving the sampling distribution of $\hat{b}_1$ tedious. We'll skip the details (see an advanced regression textbook, if interested) and the end result is:

$$std.\ error\ (\hat{b}_1) = \sqrt{\frac{MSresid}{\sum (X_i - \overline{X})^2}}$$

  - Thus, we can conduct the following statistical test:

$$H_0 : b_1 = 0$$
$$H_1 : b_1 \neq 0$$

$$t(N-2) \sim \frac{\hat{b}_1}{standard\ error\ (\hat{b}_1)}$$

  - We can also easily compute confidence intervals around $\hat{b}_1$

$$estimate \pm t_{\alpha/2,df} * standard\ error\ of\ estimate$$
$$\hat{b}_1 \pm t_{\alpha/2,df} * \sqrt{\frac{MSresid}{\sum (X_i - \overline{X})^2}}$$

  - Conclusions
    - If the test is significant, then we conclude that there is a significant linear relationship between X and Y

      For every one-unit change in X, there is a $\hat{b}_1$ unit change in Y

    - If the test is not significant, then there is no significant <u>linear</u> relationship between X and Y

      Utilizing the linear relationship between X and Y does not significantly improve our ability to predict Y, compared to using the grand mean.

      There may still exist a significant non-linear relationship between X and Y

- Inferences about the intercept, b_0
  - b_0 tells us the predicted value of Y when $X = 0$
  - The test of b_0 is automatically computed and displayed, but be careful not to misinterpret its significance!
  - Only rarely are we interested in the value of the intercept
  - Again, we'll skip the details concerning the derivation of the sampling distribution of $\hat{b}_0$ (see an advanced regression textbook, if interested) and the end result is:

$$std.\ error\ (\hat{b}_0) = \sqrt{MSresid}\sqrt{\frac{\sum X_i^2}{N\sum (X_i - \bar{X})^2}}$$

  - Thus, we can conduct the following statistical test:

$$H_0 : b_0 = 0$$
$$H_1 : b_0 \neq 0$$

$$t(N-2) \sim \frac{\hat{b}_0}{standard\ error\ (\hat{b}_0)}$$

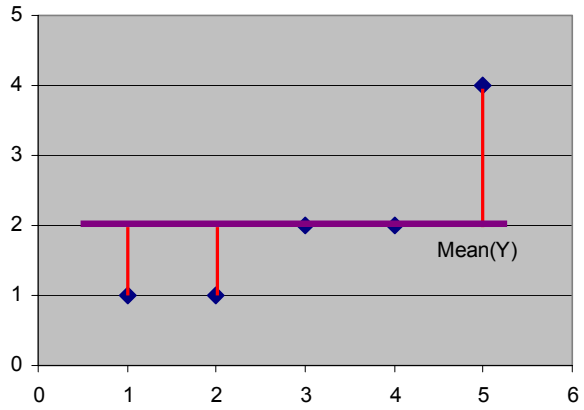
  - We can also easily compute confidence intervals around $\hat{b}_0$

$$estimate \pm t_{\alpha/2,df} * standard\ error\ of\ estimate$$
$$\hat{b}_0 \pm t_{\alpha/2,df} * \sqrt{MSresid}\sqrt{\frac{\sum X_i^2}{N\sum (X_i - \bar{X})^2}}$$

5. Sums of Squares in Regression and the ANOVA table

- Total Sums of Squares (SST)
  - In ANOVA, the total sums of squares were the sum of the squared deviations from the grand mean
  - We will use this same definition in regression. SST is the sum of the squared deviations from the grand mean of Y

$$SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

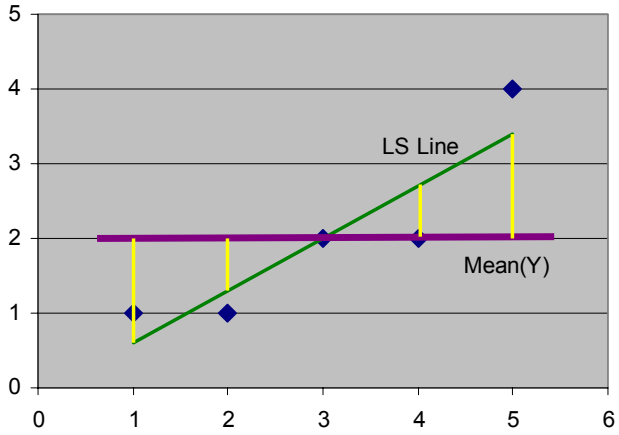


- Sums of Squares Regression
  - In ANOVA, we had a sum of squares for the model. This SS captured the improvement in our prediction of Y based on all the terms in the model
  - In regression, we can also examine how much we improve our prediction (compared to the grand mean) by using the regression line to predict new observations
    - If we had not conducted a regression, then our "best guess" for a new value of Y would be the mean of Y, $\bar{Y}$
    - But we can use the regression line to make better predictions of new observations

$$\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_{1i}$$

- The deviation of the regression best guess (the predicted value) from the grand mean is the SS Regression.

$$SSReg = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2$$

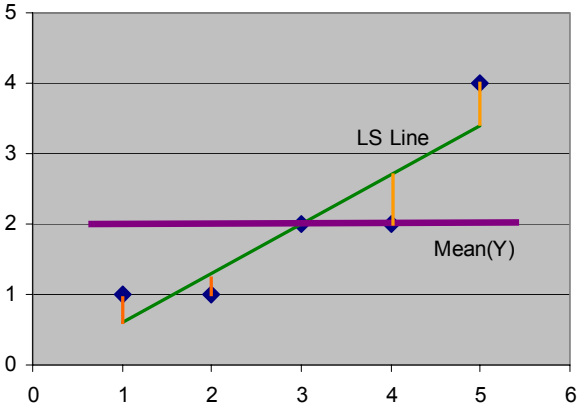


- Sums of Squares Error / Residual
  - The residuals are the deviations of the predicted values from the observed values

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

  - The SS Residual is the amount of the total SS that we cannot predict from the regression model

$$SSResid = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$



- Sums of Squares partitioning
  - We have three SS components and we can partition them in the following manner

$$SST = SSreg + SSresid$$

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 + \sum(Y_i - \hat{Y}_i)^2$$

  - In ANOVA, we had a similar partition

$$SST = SSmodel + SSerror$$

    - It turns out that ANOVA is a special case of regression. If we set up a regression with categorical predictors, then we will find

$$SSreg = SSmodel$$
$$SSresid = SSerror$$

    - Every analysis we conducted in ANOVA, can be conducted in regression. But regression provides a much more general statistical framework (and thus is frequently called the "general linear model").

- Where there are sums of squares, there is an ANOVA table.
  - Based on the SS decomposition, we can construct an ANOVA table

Source	SS	df	MS	F
Regression	$SSReg = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$	(# of parameters) -1	$\frac{SSreg}{df}$	$\frac{MSreg}{MSresid}$
Residual	$SSResid = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$	N – (# of parameters)	$\frac{SSresid}{df}$	
Total	$SST = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$	N-1		

- The *Regression* test examines all of the slope parameters in the model simultaneously. Do these parameters significantly improve our ability to predict Y, compared to using the grand mean to predict Y?

$$H_0 : b_1 = b_2 = ... = b_k = 0$$
$$H_1 : \text{Not all } b_j\text{'s} = 0$$

- For simple linear regression, we only have one slope parameter. This test becomes a test of the slope of b_1

$$H_0 : b_1 = 0$$
$$H_1 : b_1 \neq 0$$

  - In other words, for simple linear regression, the *Regression* F-test will be identical to the t-test of the b_1 parameter
  - This relationship will not hold for multiple regression, when more than one predictor is entered into the model

- Calculating a measure of variance in Y accounted for by X

  - SS Total is a measure of the total variability in Y

$$SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \qquad Var(Y) = \frac{SST}{N-1}$$

  - The SS Regression is the part of the total variability that we can explain using our regression line

  - As a result, we can consider the following ratio, R^2 to be a measure of the proportion of the sample variance in Y that is explained by X

$$R^2 = \frac{SSReg}{SSTotal}$$

    - R^2 is analogous to η^2 in ANOVA

- But in ANOVA, we preferred a measure variance accounted for in the <u>population</u> (ω^2) rather than in the sample (η^2).

- The regression equivalent of ω^2 is called the *Adjusted* R^2.

  - Any variable (even a completely random variable) is unlikely to have *SSReg* exactly equal to zero. Thus, any variable we use will explain some of the variance in the sample

  - *Adjusted* R^2 corrects for this overestimation by penalizing R^2 for the number of variables in the regression equation

- What happens if we take the square root of R^2?

$$R = \sqrt{\frac{SSReg}{SSTotal}}$$

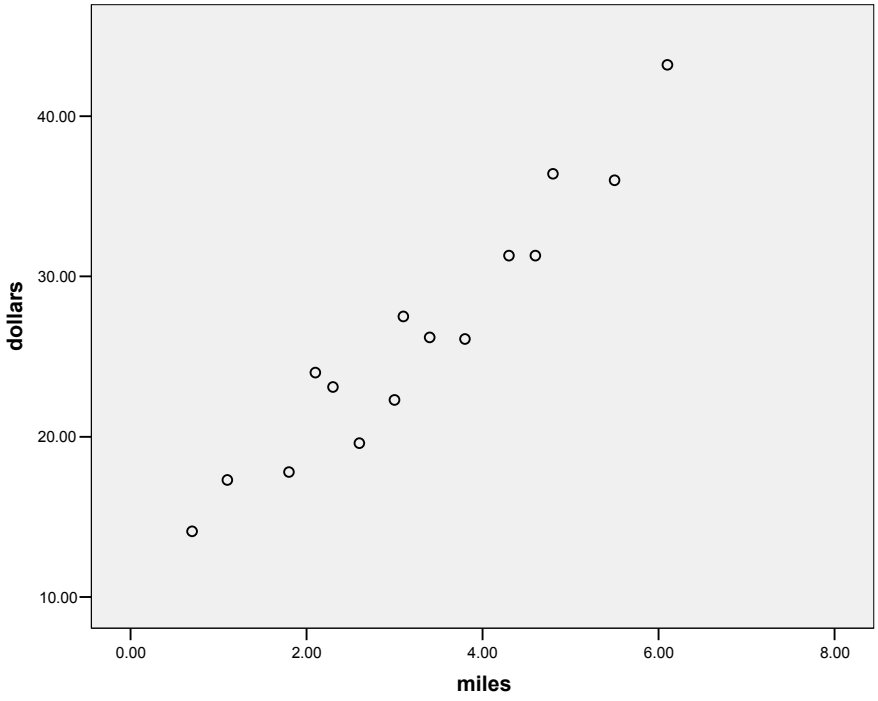
- R is interpreted as the overall correlation between all the predictor variables and the outcome variable
- When only one predictor is in the model, R is the correlation between X and Y, r_{XY}

6. An example

- Predicting the amount of damage caused by a fire from the distance of the fire from the nearest fire station

Fire Damage Data			
Distance from Station (Miles)	Fire Damage (Thousands of Dollars)	Distance from Station (Miles)	Fire Damage (Thousands of Dollars)
3.40	26.20	2.60	19.60
1.80	17.80	4.30	31.30
4.60	31.30	2.10	24.00
2.30	23.10	1.10	17.30
3.10	27.50	6.10	43.20
5.50	36.00	4.80	36.40
0.70	14.10	3.80	26.10
3.00	22.30		

- Always plot the data first!!!



- In SPSS, we use the "*Regression*" command to obtain a regression analysis

```
REGRESSION
  /DEPENDENT dollars
  /METHOD=ENTER miles.
```

This box tells us that "MILES" was entered as the only predictor

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	MILES[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: DOLLARS

This box gives us measures of the variance accounted for by the model

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.961[a]	.923	.918	2.31635

a. Predictors: (Constant), MILES

$\sqrt{MSE}$

Here is our old friend the ANOVA table

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	841.766	1	841.766	156.886	.000[a]
	Residual	69.751	13	5.365		
	Total	911.517	14			

a. Predictors: (Constant), MILES

b. Dependent Variable: DOLLARS

These are the tests of the intercept and the slope

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	10.278	1.420		7.237	.000
	MILES	4.919	.393	.961	12.525	.000

a. Dependent Variable: DOLLARS

- From this table, we read that $\hat{b}_0 = 10.278$ and that $\hat{b}_1 = 4.919$. Using this information we can write the regression equation

$$\hat{Y} = 10.278 + 4.919 * X$$

- To test the slope:

$$H_0 : b_1 = 0$$
$$H_1 : b_1 \neq 0$$

  We find a significant linear relationship between the distance from the fire, and the amount of damage caused by the fire, $t(13) = 12.53, p < .01$.

  For every 1 mile from the fire station, the fire caused an additional $4,919 in damage

- Note that the t-test for $\hat{\beta}_1$ is identical to the *Regression* test on the ANOVA table because we only have one predictor in this case.

- In this case, the test of the intercept is not meaningful

- You can also easily obtain 95% confidence intervals around the parameter estimates

```
REGRESSION
  /STATISTICS coeff r anova ci
  /DEPENDENT dollars
  /METHOD=ENTER miles  .
```

  - *COEFF*, *R* and *ANOVA* are defaults
    - *COEFF* prints the estimates of b_0 and b_1
    - *R* prints R^2 and *Adjusted* R^2
    - *ANOVA* prints the regression ANOVA table

- Adding CI to the STATISTICS command will print the confidence intervals for all model parameters

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	10.278	1.420		7.237	.000	7.210	13.346
	MILES	4.919	.393	.961	12.525	.000	4.071	5.768

a. Dependent Variable: DOLLARS

$\hat{b}_1 = 4.919$ $\quad t(13) = 12.53, p < .001$

95% CI for $\hat{b}_1$: $\quad \hat{b}_1 \pm t_{\alpha/2,df} * std.error(\hat{b}_1)$

$4.919 + 2.16(.393)$

$(4.07, 5.77)$

7. Estimation and prediction

- One of the goals of regression analysis is to allow us to estimate or predict new values of Y based on observed X values. There are two kinds of Y values we may want to predict

  - Case I: We may want to estimate the mean value of Y, $\hat{Y}$, for a specific value of X

    - In this case, we are attempting to estimate the mean result of many events at a single value of X

    - For example, what is the average damage caused by (all) fires that are 5.8 miles from a fire station?

- Case II: We may also want to predict a particular value of Y, $\hat{Y}_i$, for a specific value of X

  - In this case, we are attempting to predict the outcome of a single event at a single value of X

  - For example, what would be the predicted damage caused by a (single) fire that is 5.8 miles from a fire station?

- In either case, we can use our regression equation to obtain an estimated mean value or particular value of Y

$$\hat{Y} = 10.278 + 4.919 * X$$

  - For a fire 5.8 miles from a station, we substitute $X = 5.8$ into the regression equation

$$\hat{Y} = 10.278 + 4.919 * 5.8$$
$$\hat{Y} = 38.81$$

- The difference in these two uses of the regression model lies in the accuracy (variance) of our estimate of the prediction

- Case I: Variance of the estimate the mean value of Y, $\hat{Y}$, at X_p

  - When we attempt to estimate a mean value, there is one source of variability: the variability due to the regression line

    - We know the equation of the regression line:

$$\hat{Y} = \hat{b}_0 + \hat{b}_1 X$$

$$Var(\hat{Y}) = Var(\hat{b}_0 + \hat{b}_1 X)$$

    - Skipping a few details, we arrive at the following equation

$$\hat{\sigma}^2_{\hat{Y}} = Var(\hat{Y}) = MSE\left[\frac{1}{N} + \frac{(X_p - \overline{X})^2}{S_{XX}}\right]$$

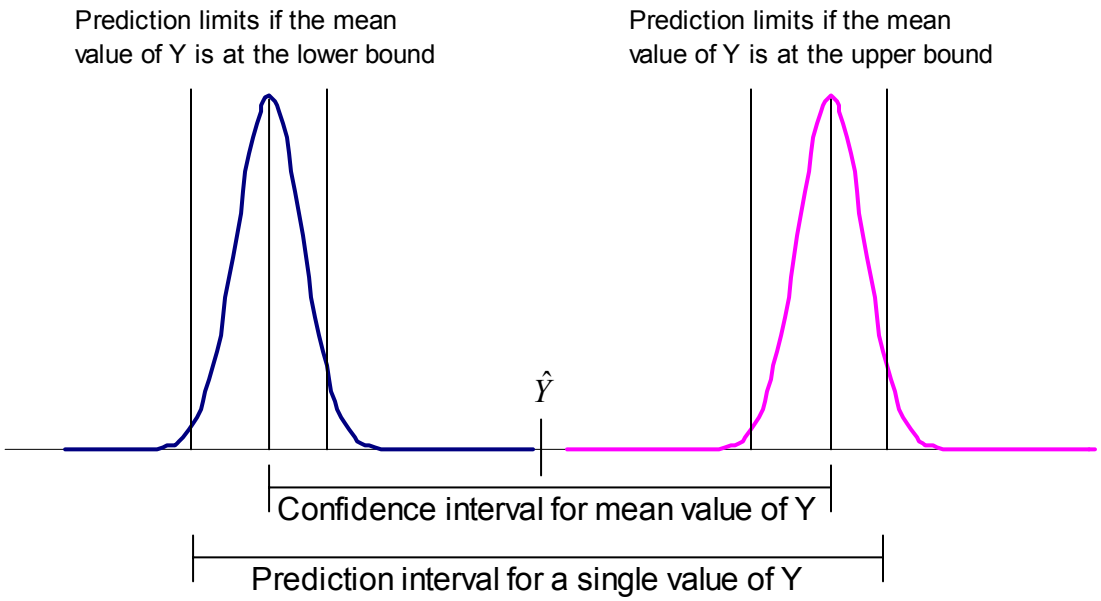
- And thus, the equation for the confidence interval of the estimate of the mean value of Y, $\hat{Y}$ is

$$\hat{Y} \pm t_{\alpha/2,N-2}\hat{\sigma}_{\hat{Y}}$$

$$\hat{Y} \pm t_{\alpha/2,N-2}\sqrt{MSE\left[\frac{1}{N}+\frac{(X_p-\bar{X})^2}{S_{XX}}\right]}$$

- Case II: Variance of the prediction of a particular value of Y, $\hat{Y}_i$, at X_p

  - When we attempt to predict a single value, there are now two sources of variability: the variability due to the regression line and variability of Y around its mean



  - The variance for the prediction interval of a single value must include these two forms of variability

$$\hat{\sigma}^2_{\hat{Y}i} = \sigma^2_{\varepsilon} + \sigma^2_{\hat{Y}}$$

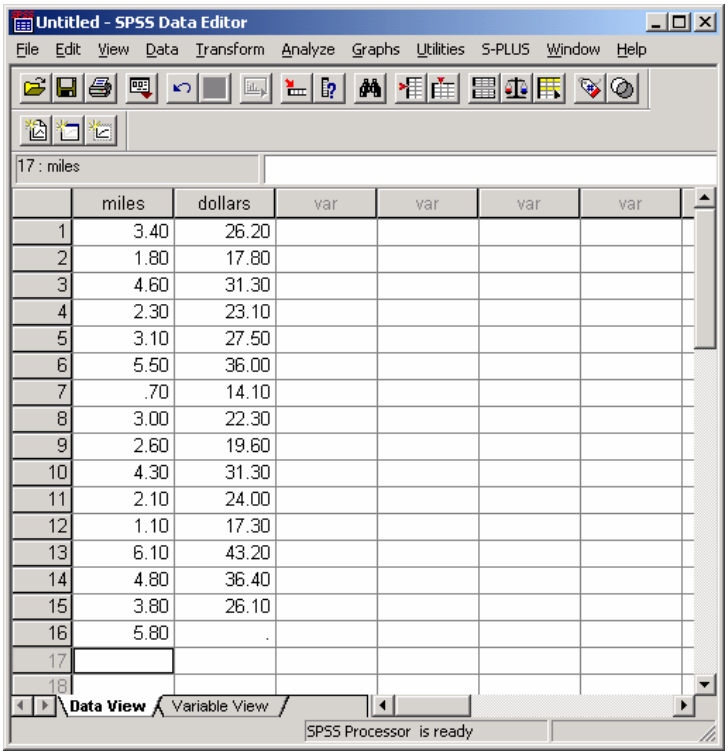
$$\hat{\sigma}^2_{\hat{Y}i} = MSE\left[1+\frac{1}{N}+\frac{(X_p-\bar{X})^2}{S_{XX}}\right]$$

- And thus, the equation for the prediction interval of the estimate of the mean value of Y, $\hat{Y}$ is

$$\hat{Y} \pm t_{\alpha/2,N-2}\hat{\sigma}_{\hat{Y}i}$$

$$\hat{Y} \pm t_{\alpha/2,N-2}\sqrt{MSE\left[1+\frac{1}{N}+\frac{(X_p-\overline{X})^2}{S_{XX}}\right]}$$

- Luckily, we can get SPSS to perform most of the intermediate calculations for us, but we need to be sneaky
  - Add a new line to the data file with a missing value for Y and $X = X_p$



  - Ask SPSS to save the predicted value and the standard error of the predicted value when you run the regression

```
REGRESSION
  /MISSING LISTWISE
  /DEPENDENT dollars
  /METHOD=ENTER miles
  /SAVE PRED (pred) SEPRED (sepred) .
```

- We will have two new variables in the data file

  PRED $\hat{Y}$ for the X value

  SEPRED $\hat{\sigma}_{\hat{Y}}$ for the X value

DOLLARS	MILES	PRED	SEPRED
26.20	3.40	27.00365	.59993
17.80	1.80	19.13272	.83401
31.30	4.60	32.90685	.79149
23.10	2.30	21.59239	.71122
27.50	3.10	25.52785	.60224
36.00	5.50	37.33425	1.05731
14.10	.70	13.72146	1.17663
22.30	3.00	25.03592	.60810
19.60	2.60	23.06819	.65500
31.30	4.30	31.43105	.71985
24.00	2.10	20.60852	.75662
17.30	1.10	15.68919	1.04439
43.20	6.10	40.28585	1.25871
36.40	4.80	33.89072	.84503
26.10	3.80	28.97139	.63199
.	5.80	38.81005	1.15640

- For $X_p = 5.8$

  $\hat{Y} = 38.81$ $\quad\quad$ $\hat{\sigma}_{\hat{Y}} = 1.156$

o Use the formulas to compute the confidence and prediction intervals

- To calculate a 95% confidence interval around the mean value, $\hat{Y}$

$$\hat{Y} \pm t_{\alpha/2,N-2}\hat{\sigma}_{\hat{Y}}$$
$$38.81 \pm t_{.025,13}(1.156)$$
$$38.81 \pm (2.16)(1.156)$$
$$(36.31, 41.31)$$

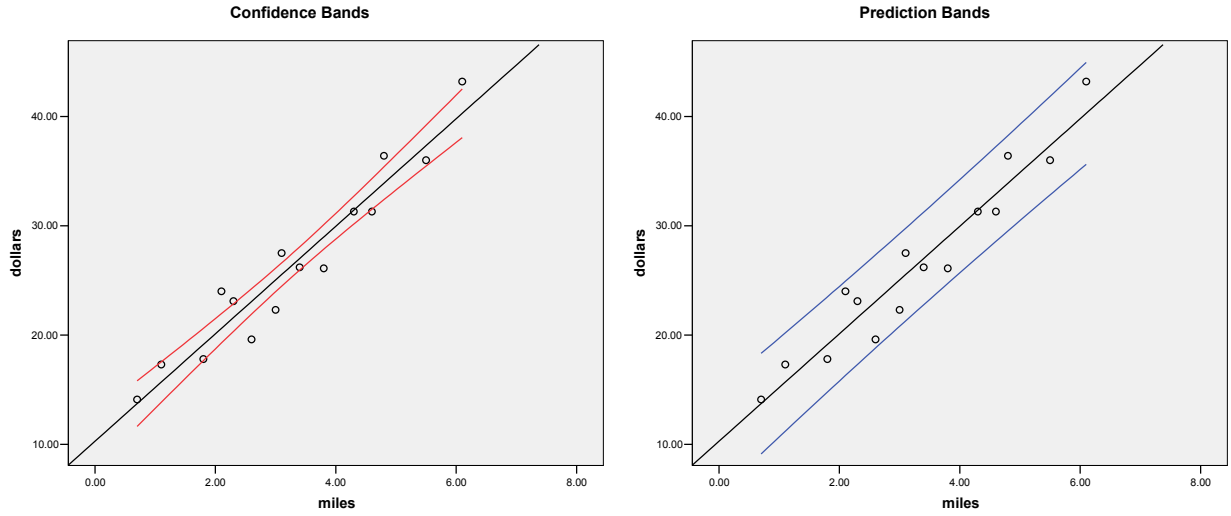
- To calculate a 95% prediction interval around the single value, $\hat{Y}_i$

$$\hat{Y} \pm t_{\alpha/2,N-2}\hat{\sigma}_{\hat{Y}i}$$
$$\hat{Y} \pm t_{\alpha/2,N-2}\sqrt{\hat{\sigma}_{\varepsilon}^2 + \hat{\sigma}_{\hat{Y}}^2}$$

$$38.81 \pm t_{.025,13}\sqrt{5.365 + (1.156)^2}$$
$$38.81 \pm (2.16)(2.589)$$
$$(32.08,\ 45.54)$$

- The regression line can be used for prediction and estimation, but not for extrapolation
- In other words, the regression line is only valid for *X*s within the range of the observed *X*s

- SPSS can be used to graph confidence intervals and prediction intervals



8. Standardized regression coefficients
   - To interpret the slope parameter, we must return to the original scale of the data
   - $b_1 = 156$ suggests that for every one unit change in the X variable, Y changes by 156 units.
   - This dependence on units can make for difficulty in comparing the effects of X on Y across different studies

     - If one researcher measures self-esteem using a 7 point scale and another uses a 4 point scale, they will obtain different estimates of b_1
     - If one researcher measures length in centimeters and another uses inches, they will obtain different estimates of b_1

   - One solution to this problem is to use standardized regression coefficients

$$\beta_1 = b_1 \frac{\sigma_X}{\sigma_Y}$$

- To understand how to interpret standardized regression coefficients, it is helpful to see how they can be obtained directly
  - Transform both Y and X into z-scores, z_Y and z_X

```
compute zmiles = (miles - 3.28)/1.5762.
compute zdollar = (dollars - 26.41)/8.06898.
```

  - Regress z_Y on z_X

```
REGRESSION
  /DEPENDENT zdollar
  /METHOD=ENTER zmiles.
```

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	4.131E-04	.074		.006	.996
	ZMILES	.961	.077	.961	12.525	.000

a. Dependent Variable: ZDOLLAR

$$b_1 = \beta_1 = .961$$

  - Compare this result to the regression on the raw data

```
REGRESSION
  /DEPENDENT dollars
  /METHOD=ENTER miles.
```

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	10.278	1.420		7.237	.000
	MILES	4.919	.393	.961	12.525	.000

a. Dependent Variable: DOLLARS

$$\beta_1 = .961$$

- To interpret standardized beta coefficients, we need to think in terms of z-scores

  - A 1 standard deviation change in X (miles), is associated with a .96 standard deviation change in Y (dollars)

  - For simple linear regression (with only 1 predictor), $\beta_1 = r_{XY}$

  - With more than 1 predictor, standardized coefficients should not be interpreted as correlations. It is possible to have standardized coefficients greater than 1.

9. Additional concerns and observations

- Standard assumptions of regression analysis

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

  - 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

- Additionally, we assume that the regression model is a suitable proxy for the "correct" (but unknown) model:
  - The relationship between X and Y must be linear
  - No important variables have been omitted from the model
  - No outliers or influential observations

- These assumptions can be examined by looking at the residuals