Simple Linear Regression

By David R. Larsen

Simple linear regression is a tool for fitting a linear line to a set of data. It is used when you want to predict the value of the "dependent variable" ***Y*** by knowing the value of the "independent variable" ***X***. Figure 1 is an example of a data set with a regression line fit.

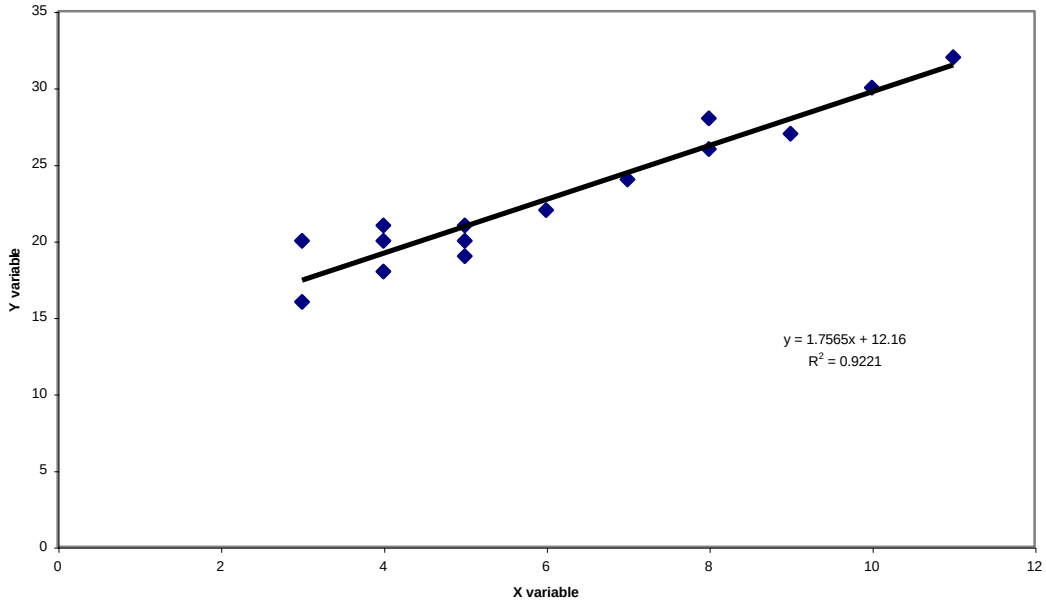


Figure 1. Example data set with regression line fit to data.

The line in the graph can be described as:

$$y = b_0 + b_1 x$$

where y is the dependent variable (also plotted on the y axis of the graph), ***x*** is the independent variable (plotted on the ***x*** axis of the graph). The parameters that are estimated are b_0 and b_1. These parameters can be estimated using the following equations:

$$b_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

where X_i and Y_i are the individual observation and n is the number of observations.

Table 1. Formulas for the ANOVA table that test the significance of the regression

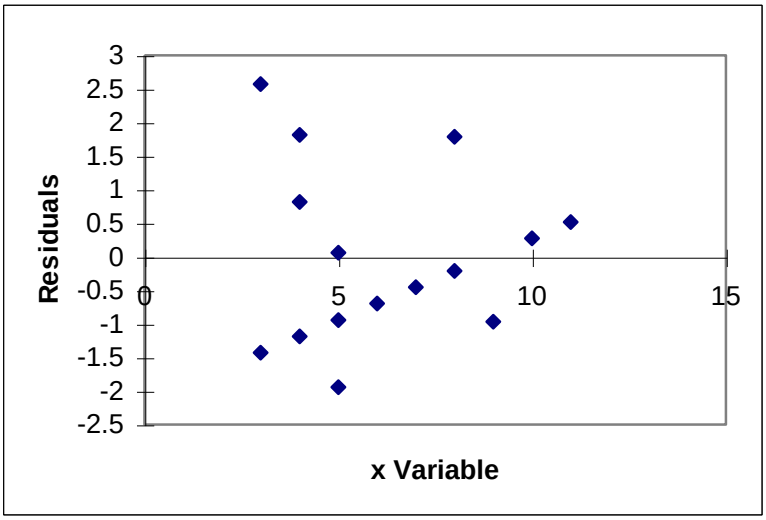
Source	Sum of Squares (SS)	*df*	Mean Square (MS)	*F*
Total	$\sum_{i=1}^{n}(y_i-\bar{y})^2$	$n-1$		
Regression	$\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$	1	$\frac{SS_{regression}}{df_{regression}}$	
Residual or Error	$SS_{Total}-SS_{regression}$	$n-2$	$\frac{SS_{residual}}{df_{residual}}$	$\frac{MS_{regression}}{MS_{residual}}$

The results of a regression are often summarized using an analysis of variance table. The usual configuration for the table is as follows:

The F test is a test to determine if the regression explains more of the variation than the mean. Another statistic that is commonly used to describe a regression is the coefficient of determination R^2. This statistic is the proportion of the observed data explained by the regression. This statistic is a value that ranges from 0 to 1 with 0 being no agreement between the regression and the data and 1 being perfect agreement between the data and the regression.

$$R^2 = \frac{SS_{regression}}{SS_{Total}}$$

Another important method of explaining the results of a regression is to plot the residuals against the independent variable. This analysis can be used to indicate that the model is miss-specified and transformation required.



Author: Dr. David R. Larsen

Figure 2. Residual plot of the data in Figure 1.

Also See:

Chapter 16 - Simple linear Regression pages 317-330 in:

Zar, J. H. 1999. Biostatistical Analysis. Prentice-Hall, Inc. Englewood Cliffs, New Jersey. 718 pp.