Regression lab- http://r-statistics.co/Linear-Regression.html https://www.guru99.com/r-simple-multiple-linear-regression.html https://www.tutorialspoint.com/r/r_multiple_regression.htm

https://stattrek.com/regression/linear-regression.aspx?tutorial=reg

https://stattrek.com/multiple-regression/excel.aspx?tutorial=reg

99.com/r-simple-multiple-lihttps://www.gurunear-regression.html#5

What is Linear Regression?

In a cause and effect relationship, the **independent variable** is the cause, and the **dependent variable** is the effect. **Least squares linear regression** is a method for predicting the value of a dependent variable *Y*, based on the value of an independent variable *X*.

For the next few lessons, we focus on the case where there is only one independent variable. This is called simple regression. Toward the end of the tutorial, we will cover multiple regression, which handles two or more independent variables.

Tip: The next lesson presents a simple linear regression example that shows how to apply the material covered in this lesson. Since this lesson is a little dense, you may benefit by also reading the next lesson.

Prerequisites for Regression

Simple linear regression is appropriate when the following conditions are satisfied.

- The dependent variable *Y* has a linear relationship to the independent variable *X*. To check this, make sure that the XY scatterplot is linear and that the residual plot shows a random pattern. (Don't worry. We'll cover residual plots in a future lesson.)
- For each value of X, the probability distribution of Y has the same standard deviation σ. When this condition is satisfied, the

variability of the residuals will be relatively constant across all values of X, which is easily checked in a residual plot.

- For any given value of X,
  - The Y values are independent, as indicated by a random pattern on the residual plot.
  - The Y values are roughly normally distributed (i.e., symmetric and unimodal). A little skewness is ok if the sample size is large. A histogram or a dotplot will show the shape of the distribution.

The Least Squares Regression Line

Linear regression finds the straight line, called the **least squares regression line** or LSRL, that best represents observations in a bivariate data set. Suppose *Y* is a dependent variable, and *X* is an independent variable. The population regression line is:

$$Y = B_0 + B_1X$$

where B_0 is a constant, B_1 is the regression coefficient, X is the value of the independent variable, and Y is the value of the dependent variable.

Given a random sample of observations, the population regression line is estimated by:

$$\hat{y} = b_0 + b_1x$$

where b_0 is a constant, b_1 is the regression coefficient, x is the value of the independent variable, and ŷ is the *predicted* value of the dependent variable.

How to Define a Regression Line

Normally, you will use a computational tool - a software package (e.g., Excel) or a graphing calculator - to find b_0 and b_1. You enter the *X* and *Y* values into your program or calculator, and the tool solves for each parameter.

In the unlikely event that you find yourself on a desert island without a computer or a graphing calculator, you can solve for b_0 and b_1 "by hand". Here are the equations.

$$b_1 = \Sigma [(x_i - x)(y_i - y)] / \Sigma [(x_i - x)^2]$$

$$b_1 = r * (s_y / s_x)$$

$$b_0 = y - b_1 * x$$

where b_0 is the constant in the regression equation, b_1 is the regression coefficient, r is the correlation between x and y, x_i is the *X* value of observation *i*, y_i is the *Y* value of observation *i*, x is the mean of *X*, y is the mean of *Y*, s_x is the standard deviation of *X*, and s_y is the standard deviation of *Y*.

Properties of the Regression Line

When the regression parameters (b_0 and b_1) are defined as described above, the regression line has the following properties.

- The line minimizes the sum of squared differences between observed values (the *y* values) and predicted values (the ŷ values computed from the regression equation).
- The regression line passes through the mean of the *X* values (x) and through the mean of the *Y* values (y).
- The regression constant (b_0) is equal to the y intercept of the regression line.
- The regression coefficient (b_1) is the average change in the dependent variable (*Y*) for a 1-unit change in the independent variable (*X*). It is the slope of the regression line.

The least squares regression line is the only straight line that has all of these properties.

The Coefficient of Determination

The **coefficient of determination** (denoted by R^2) is a key output of regression analysis. It is interpreted as the proportion of the variance in the dependent variable that is predictable from the independent variable.

- The coefficient of determination ranges from 0 to 1.
- An R^2 of 0 means that the dependent variable cannot be predicted from the independent variable.
- An R^2 of 1 means the dependent variable can be predicted without error from the independent variable.
- An R^2 between 0 and 1 indicates the extent to which the dependent variable is predictable. An R^2 of 0.10 means that 10 percent of the variance in *Y* is predictable from *X*; an R^2 of 0.20 means that 20 percent is predictable; and so on.

The formula for computing the coefficient of determination for a linear regression model with one independent variable is given below.

Coefficient of determination. The coefficient of determination (R^2) for a linear regression model with one independent variable is:

$$R^2 = \{ (1 / N) * \Sigma [(x_i - x) * (y_i - y)] / (\sigma_x * \sigma_y) \}^2$$

where N is the number of observations used to fit the model, Σ is the summation symbol, x_i is the x value for observation i, x is the mean x value, y_i is the y value for observation i, y is the mean y value, σ_x is the standard deviation of x, and σ_y is the standard deviation of y.

If you know the linear correlation (r) between two variables, then the coefficient of determination (R^2) is easily computed using the following formula: $R^2 = r^2$.

Standard Error

The **standard error** about the regression line (often denoted by SE) is a measure of the average amount that the regression equation over- or under-predicts. The higher the coefficient of determination, the lower the standard error; and the more accurate predictions are likely to be.

Test Your Understanding

Problem 1

A researcher uses a regression equation to predict home heating bills (dollar cost), based on home size (square feet). The correlation between predicted bills and home size is 0.70. What is the correct interpretation of this finding?

(A) 70% of the variability in home heating bills can be explained by home size.
(B) 49% of the variability in home heating bills can be explained by home size.
(C) For each added square foot of home size, heating bills increased by 70 cents.
(D) For each added square foot of home size, heating bills increased by 49 cents.
(E) None of the above.

Solution

The correct answer is (B). The coefficient of determination measures the proportion of variation in the dependent variable that is predictable from the independent variable. The coefficient of determination is equal to R^2; in this case, $(0.70)^2$ or 0.49. Therefore, 49% of the variability in heating bills can be explained by home size

Linear Regression Example

In this lesson, we apply regression analysis to some fictitious data, and we show how to interpret the results of our analysis.

Note: Regression computations are usually handled by a software package or a graphing calculator. For this example, however, we will do the computations "manually", since the gory details have educational value.

Problem Statement

Last year, five randomly selected students took a math aptitude test before they began their statistics course. The Statistics Department has three questions.

- What linear regression equation best predicts statistics performance, based on math aptitude scores?
- If a student made an 80 on the aptitude test, what grade would we expect her to make in statistics?
- How well does the regression equation fit the data?

How to Find the Regression Equation

In the table below, the x_i column shows scores on the aptitude test. Similarly, the y_i column shows statistics grades. The last two columns show deviations scores - the difference between the student's score and the average score on each test. The last two rows show sums and mean scores that we will use to conduct the regression analysis.

Student	x_i	y_i	(x_i-x)	(y_i-y)
1	95	85	17	8
2	85	95	7	18
3	80	70	2	-7
4	70	65	-8	-12
5	60	70	-18	-7
Sum	390	385		
Mean	78	77		

And for each student, we also need to compute the squares of the deviation scores (the last two columns in the table below).

Student	x_i	y_i	$(x_i-x)^2$	$(y_i-y)^2$
1	95	85	289	64
2	85	95	49	324
3	80	70	4	49

4	70	65	64	144
5	60	70	324	49
Sum	390	385	730	630
Mean	78	77		

And finally, for each student, we need to compute the product of the deviation scores.

Student	x_i	y_i	$(x_i-x)(y_i-y)$
1	95	85	136
2	85	95	126
3	80	70	-14
4	70	65	96
5	60	70	126
Sum	390	385	470
Mean	78	77	

The regression equation is a linear equation of the form: $\hat{y} = b_0 + b_1x$. To conduct a regression analysis, we need to solve for b_0 and b_1. Computations are shown below. Notice that all of our inputs for the regression analysis come from the above three tables.

First, we solve for the regression coefficient (b_1):

$$b_1 = \Sigma [(x_i - x)(y_i - y)] / \Sigma [(x_i - x)^2]$$

$$b_1 = 470/730$$

$$b_1 = 0.644$$

Once we know the value of the regression coefficient (b_1), we can solve for the regression slope (b_0):

$$b_0 = y - b_1 * x$$

$$b_0 = 77 - (0.644)(78)$$

$$b_0 = 26.768$$

Therefore, the regression equation is: $\hat{y} = 26.768 + 0.644x$.

How to Use the Regression Equation

Once you have the regression equation, using it is a snap. Choose a value for the independent variable (*x*), perform the computation, and you have an estimated value ($\hat{y}$) for the dependent variable.

In our example, the independent variable is the student's score on the aptitude test. The dependent variable is the student's statistics grade. If a student made an 80 on the aptitude test, the estimated statistics grade ($\hat{y}$) would be:

$$\hat{y} = b_0 + b_1x$$

$$\hat{y} = 26.768 + 0.644x = 26.768 + 0.644 * 80$$

$$\hat{y} = 26.768 + 51.52 = 78.288$$

Warning: When you use a regression equation, do not use values for the independent variable that are outside the range of values used to create the equation. That is called **extrapolation**, and it can produce unreasonable estimates.

In this example, the aptitude test scores used to create the regression equation ranged from 60 to 95. Therefore, only use values inside that range to estimate statistics grades. Using values outside that range (less than 60 or greater than 95) is problematic.

How to Find the Coefficient of Determination

Whenever you use a regression equation, **you should ask how well the equation fits the data**. One way to assess fit is to check the coefficient of determination, which can be computed from the following formula.

$$\mathbf{R^2 = \{ (1 / N) * \Sigma [(x_i - x) * (y_i - y)] / (\sigma_x * \sigma_y) \}^2}$$

where N is the number of observations used to fit the model, Σ is the summation symbol, x_i is the x value for observation i, x is the mean x value, y_i is the y value for observation i, y is the mean y value, σ_x is the standard deviation of x, and σ_y is the standard deviation of y.

Computations for the sample problem of this lesson are shown below. We begin by computing the standard deviation of x (σ_x):

$$\sigma_x = \text{sqrt} [\Sigma (x_i - x)^2 / N]$$

$$\sigma_x = \text{sqrt}(730/5) = \text{sqrt}(146) = 12.083$$

Next, we find the **standard deviation of y, (σ_y):**

$$\boldsymbol{\sigma_y = \text{sqrt} [\Sigma (y_i - y)^2 / N]}$$

$$\sigma_y = \text{sqrt}(630/5) = \text{sqrt}(126) = 11.225$$

And finally, we compute the **coefficient of determination (R^2):**

$$\boldsymbol{R^2 = \{ (1 / N) * \Sigma [(x_i - x) * (y_i - y)] / (\sigma_x * \sigma_y) \}^2}$$

$$R^2 = [(1/5) * 470 / (12.083 * 11.225)]^2$$

$$R^2 = (94 / 135.632)^2 = (0.693)^2 = 0.48$$

A coefficient of determination equal to 0.48 indicates that about 48% of the variation in statistics grades (the dependent variable) can be explained by the relationship to math aptitude scores (the independent variable). This would be considered a good fit to the data, in the sense that it would substantially improve an educator's ability to predict student performance in statistics class.

Residual Analysis in Regression

Because a linear regression model is not always appropriate for the data, you should assess the appropriateness of the model by defining residuals and examining residual plots.

Residuals

The difference between the observed value of the dependent variable (y) and the predicted value ($\hat{y}$) is called the **residual** (e). Each data point has one residual.

Residual = Observed value - Predicted value

$$e = y - \hat{y}$$

Both the sum and the mean of the residuals are equal to zero. That is, $\Sigma\, e = 0$ and e = 0.

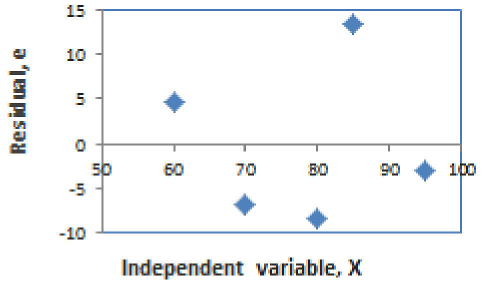
Residual Plots

A **residual plot** is a graph that shows the residuals on the vertical axis and the independent variable on the horizontal axis. If the points in a residual plot are **randomly dispersed** around the horizontal axis, a linear regression model is **appropriate** for the data; otherwise, a **nonlinear model is more appropriate.**

The table below shows inputs and outputs from a simple linear regression analysis.

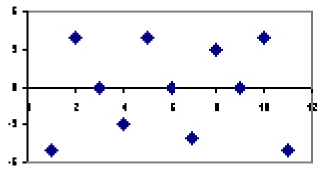
x	y	ŷ	e
60	70	65.411	4.589
70	65	71.849	-6.849
80	70	78.288	-8.288
85	95	81.507	13.493
95	85	87.945	-2.945

And the chart below displays the residual (e) and independent variable (X) as a residual plot.

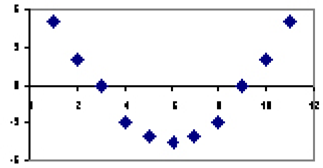


The residual plot shows a fairly random pattern - the first residual is positive, the next two are negative, the fourth is positive, and the last residual is negative. This random pattern indicates that a linear model provides a decent fit to the data.

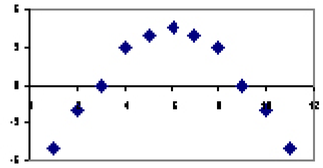
Below, the residual plots show three typical patterns. The first plot shows a random pattern, indicating a good fit for a linear model.



Random pattern



Non-random: U-shaped



Non-random: Inverted U

The other plot patterns are non-random (U-shaped and inverted U), suggesting a better fit for a nonlinear model.

Introduction to Multiple Regression

Simple linear regression is a technique for predicting the value of a dependent variable, based on the value of a single independent variable. Sometimes, you only need one relevant independent variable to make an accurate prediction.

Often, however, the prediction is better when you use two or more independent variables. Multiple regression is a technique for predicting the value of a dependent variable, based on the values of two or more independent variables.

The Regression Equation

This is a tutorial about *linear* regression, so our focus is on *linear* relationships between variables. The regression equation that expresses the linear relationships between a single dependent variable and one or more independent variables is:

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \ldots + b_{k-1}x_{k-1} + b_kx_k$$

In this equation, **$\hat{y}$ is the *predicted* value of the dependent variable**. Values of the ***k* independent variables** are denoted by x_1, x_2, x_3, ... , x_k.

And finally, we have the *b*'s - b_0, b_1, b_2, ... , b_k. The b's are constants, called **regression coefficients**. Values are assigned to the *b*'s based on the **principle of least squares.**

What is the Principle of Least Squares?

In multiple regression, the deviation of the actual value for a dependent variable from its predicted value is called the residual. The residual (e) for a single observation *i* is:

$$\mathbf{e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1x_{1i} + b_2x_{2i} + \ldots + b_kx_{ki})}$$

Assume that the set of data consists of *n* observations. **The principle of least squares requires that the sum of squared residuals for all *n* observations be minimized. That is, we want the following value to be as small as possible:**

$$\mathbf{\Sigma [y_i - (b_0 + b_1x_{1i} + b_2x_{2i} + \ldots + b_kx_{ki}]^2}$$

Regression analysis requires that the values of b_0, b_1, ... , b_k be defined to minimize the sum of the squared residuals. When we assign values to regression coefficients in this way, we are following the principle of least squares.

Normal Equations for Simple Regression

Finding the right values for regression coefficients (i.e., values that satisfy a least squares criterion) involves solving a set of linear equations. **These equations can be derived using calculus, and they are called normal equations.**

To illustrate the use of normal equations, let's look at simple linear regression - regression with one dependent variable (y) and one independent variable (x). With simple linear regression, the regression equation is:

$$\hat{y} = b_0 + b_1x$$

The normal equations for simple linear regression are:

$$\boldsymbol{\Sigma\, y_i = nb_0 + b_1(\Sigma x_i)}$$

$$\boldsymbol{\Sigma\, x_iy_i = b_0(\Sigma x_i) + b_1(\Sigma x_i^2)}$$

Here, we have two equations with two unknowns. The unknowns are the regression coefficients b_0 and b_1. Using ordinary algebra, we can solve for b_0 and b_1. The result is:

$$b_1 = \Sigma [(x_i - x)(y_i - y)] / \Sigma [(x_i - x)^2]$$

$$b_0 = y - b_1 * x$$

where x is the mean x score, and y is the mean y score. Note that these are the same equations that we presented in a previous lesson, when we introduced the topic of simple linear regression.

The use of normal equations to assign values to regression coefficients becomes more complicated when there are two or more independent variables. We'll tackle that challenge in the next lesson.

Test Your Understanding

Problem 1

Which of the following statements are true?

I. A regression equation with *k* independent variables has *k* regression coefficients.
II. Regression coefficients (b_o, b_1, b_2, etc.) are variables in the regression equation.
III. The principle of least squares calls for minimizing the sum of the squared residuals.

(A) I only.
(B) II only.
(C) III only.
(D) All of the above.
(E) None of the above.

Solution

The correct answer is (C). The principle of least squares defines regression coefficients that minimize the sum of the squared residuals. A regression equation with *k* independent variables has *k* + 1 regression coefficients. For example, if there were two independent variables, there would be three regression coefficients - b_o, b_1, and b_2. And finally, regression coefficients are constants - not variables.

Regression Coefficients

With simple linear regression, there is one dependent variable and one independent variable. The regression equation is:

$$\hat{y} = b_0 + b_1x$$

In the previous lesson, we developed a least-squares solution for the regression coefficients of simple linear regression:

$$b_1 = \Sigma \, [\, (x_i - x)(y_i - y) \,] \, / \, \Sigma \, [\, (x_i - x)^2]$$

$$b_0 = y - b_1 * x$$

where $\hat{y}$ is the predicted value of the dependent variable, b_0 and b_1 are regression coefficients, x_i is the value of the independent variable for observation *i*, y_i is the value of the dependent variable for observation *i*, x is the mean x score, and y is the mean y score.

In this lesson, we describe a least-squares solution for the regression coefficients of *multiple* regression.

The Multiple Regression Challenge

With simple linear regression, there are only two regression coefficients - b_0 and b_1. There are only two normal equations. Finding a least-squares solution involves solving two equations with two unknowns - a task that is easily managed with ordinary algebra.

With multiple regression, things get more complicated. There are *k* independent variables and *k* + 1 regression coefficients. There are *k* + 1 normal equations. Finding a least-squares solution involves solving *k* + 1 equations with *k* + 1 unknowns. This can be done with ordinary algebra, but it is unwieldy.

To handle the complications of multiple regression, we will use matrix algebra.

The Regression Equation in Matrix Form

With multiple regression, there is one dependent variable and *k* dependent variables. The regression equation is:

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \ldots + b_{k-1}x_{k-1} + b_kx_k$$

where $\hat{y}$ is the predicted value of the dependent variable, b_k are regression coefficients, and x_k is the value of independent variable *k*.

To express the regression equation in matrix form, we need to define three matrices: **Y**, **b**, and **X**.

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}$$

$$\begin{bmatrix} . \\ . \\ . \\ y_n \end{bmatrix} \qquad \begin{bmatrix} . \\ . \\ . \\ b_k \end{bmatrix}$$

$$\mathbf{X} = \begin{bmatrix} 1 & X_{1,\,1} & X_{1,\,2} & \ldots & X_{1,\,k} \\ 1 & X_{2,\,1} & X_{2,\,2} & \ldots & X_{2,\,k} \\ . & . & . & \ldots & . \\ . & . & . & \ldots & . \\ . & . & . & \ldots & . \\ 1 & X_{n,\,1} & X_{n,\,2} & \ldots & X_{n,\,k} \end{bmatrix}$$

Here, the dataset consists of *n* records. Each record includes scores for 1 dependent variable and *k* independent variables. **Y** is an *n* x *1* vector that holds predicted values of the dependent variable; and **b** is a *k* + *1* x *1* vector that holds estimated regression coefficients. Matrix **X** has a column of 1's plus *k* columns of values for each independent variable in the regression equation.

Given these matrices, the multiple regression equation can be expressed concisely as:

$$\mathbf{Y} = \mathbf{Xb}$$

It is sort of cool that this simple expression describes the regression equation for 1, 2, 3, or *any* number of independent variables.

Normal Equations in Matrix Form

Just as the regression equation can be expressed compactly in matrix form, so can the normal equations. The least squares normal equations can be expressed as:

$$\mathbf{X'Y} = \mathbf{X'Xb} \quad \text{or} \quad \mathbf{X'Xb} = \mathbf{X'Y}$$

Here, matrix **X'** is the transpose of matrix **X**. To solve for regression coefficients, simply pre-multiply by the inverse of **X'X**:

$$\mathbf{(X'X)^{-1}X'Xb} = \mathbf{(X'X)^{-1}X'Y}$$

$$\mathbf{b} = \mathbf{(X'X)^{-1}X'Y}$$

where $\mathbf{(X'X)^{-1}X'X} = \mathbf{I}$, the identity matrix.

In the real world, you will probably never compute regression coefficients by hand. Generally, you will use software, like SAS, SPSS, mini-tab, or excel. In the problem below, however, we will compute regression coefficients manually; so you will understand what is going on.

Test Your Understanding

Problem 1

Consider the table below. It shows three performance measures for five students.

Student	Test score	IQ	Study hours
1	100	110	40
2	90	120	30
3	80	100	20
4	70	90	0
5	60	80	10

Using least squares regression, develop a regression equation to predict test score, based on (1) IQ and (2) the number of hours that the student studied.

Solution

For this problem, we have some raw data; and we want to use this raw data to define a least-squares regression equation:

$$\hat{y} = b_0 + b_1x_1 + b_2x_2$$

where $\hat{y}$ is the predicted test score; b_0, b_1, and b_2 are regression coefficients; x_1 is an IQ score; and x_2 is the number of hours that the student studied.

On the right side of the equation, the only unknowns are the regression coefficients. To define the regression coefficients, we use the following equation:

$$\mathbf{b} = \mathbf{(X'X)^{-1}X'Y}$$

To solve this equation, we need to complete the following steps:

- Define **X**.
- Define **X'**.
- Compute **X'X**.
- Find the inverse of **X'X**.
- Define **Y**.

Let's begin with matrix **X**. Matrix **X** has a column of 1's plus two columns of values for each independent variable. So, this is matrix **X** and its transpose **X'**:

$$\mathbf{X} = \begin{bmatrix} 111040 \\ 112030 \\ 110020 \\ 1 \quad 90 \quad 0 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 80 & 10 \end{bmatrix}$$

$$\mathbf{X'} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1101201009080 \\ 40 & 30 & 20 & 0 & 10 \end{bmatrix}$$

Given **X'** and **X**, it is a simple matter to compute **X'X**.

$$\mathbf{X'X} = \begin{bmatrix} 5 & 500 & 100 \\ 50051{,}00010{,}800 \\ 10010{,}800 & 3{,}000 \end{bmatrix}$$

Finding the inverse of **X'X** takes a little more effort. A way to find the inverse is described on this site at https://stattrek.com/matrix-algebra/how-to-find-inverse.aspx. Ultimately, we find:

$$\mathbf{(X'X)}^{-1} = \begin{bmatrix} 101/5 & -7/30 & 1/6 \\ -7/30 & 1/360 & -1/450 \\ 1/6 & -1/450 & 1/360 \end{bmatrix}$$

Next, we define **Y**, the vector of dependent variable scores. For this problem, it is the vector of test scores.

$$\mathbf{Y} = \begin{bmatrix} 100 \\ 90 \\ 80 \\ 70 \\ 60 \end{bmatrix}$$

With all of the essential matrices defined, we are ready to compute the least squares regression coefficients.

$$\mathbf{b} = \mathbf{(X'X)}^{-1}\mathbf{X'Y}$$

$$\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 20 \\ 0.5 \\ 0.5 \end{bmatrix}$$

To conclude, here is our least-squares regression equation:

$$\hat{y} = 20 + 0.5x_1 + 0.5x_2$$

where $\hat{y}$ is the predicted test score; x_1 is an IQ score; and x_2 is the number of hours that the student studied. The regression coefficients are $b_0 = 20$, $b_1 = 0.5$, and $b_2 = 0.5$.

Last lesson Next lesson

Regression Analysis With Excel

In the real world, you will probably never conduct multiple regression analysis by hand. Most likely, you will use computer software (SAS, SPSS, Minitab, Excel, etc.).

Excel is a widely-available software application that supports multiple regression. In this lesson, we use Excel to demonstrate multiple regression analysis. (Other software packages produce outputs similar to Excel.)

Sample Problem With Excel

Consider the table below. It shows three performance measures for 10 students.

Student	Test score	IQ	Study hours
1	100	125	30
2	95	104	40
3	92	110	25

4	90	105	20
5	85	100	20
6	80	100	20
7	78	95	15
8	75	95	10
9	72	85	0
10	65	90	5

In this lesson, using data from the table, we are going to complete the following tasks:

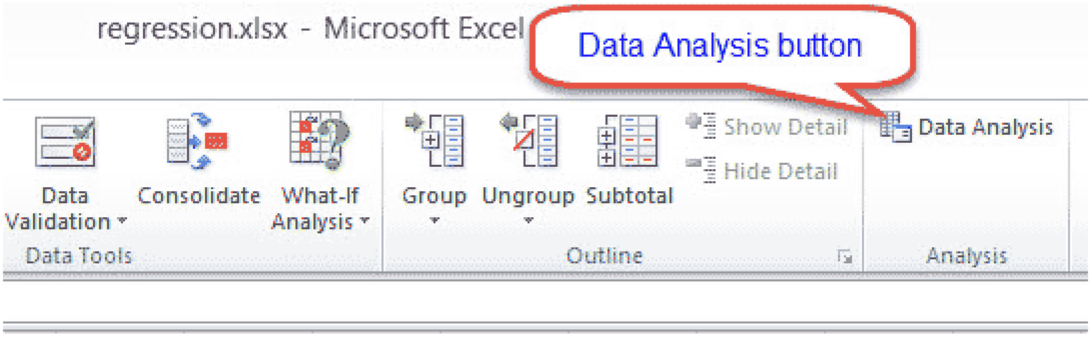
- Develop a least-squares regression equation to predict test score, based on (1) IQ and (2) the number of hours that the student studied.
- Assess how well the regression equation predicts test score, the dependent variable.
- Assess the contribution of each independent variable (i.e., IQ and study hours) to the prediction.

These are common tasks in regression analysis. With the right software, they are easy to accomplish. We'll walk you step by step through each task, starting with setting up Excel.

How to Enable Excel

When you open Excel, the module for regression analysis may or may not be enabled. So, before you do anything else, you need to determine whether Excel is enabled. Here's how to do that:

- Open Excel.
- Click the Data tab.
- If you see the Data Analysis button in the upper right corner, the Analysis TookPak is enabled and you are ready to go.



If the Data Analysis button is not visible, the Analysis ToolPak is not enabled. In that case, do the following:

- Click the File tab.
- Select Options to open the Excel Options dialog box.
- Click the Add-Ins item, from the left column. This opens the View and Manage Microsoft Office Add-ins screen.
- From the Manage drop-down box, choose Excel Add-Ins and click the Go button. This opens the Add-Ins dialog box.
- From the Add-Ins dialog, check the box beside Analysis ToolPak and click Go.

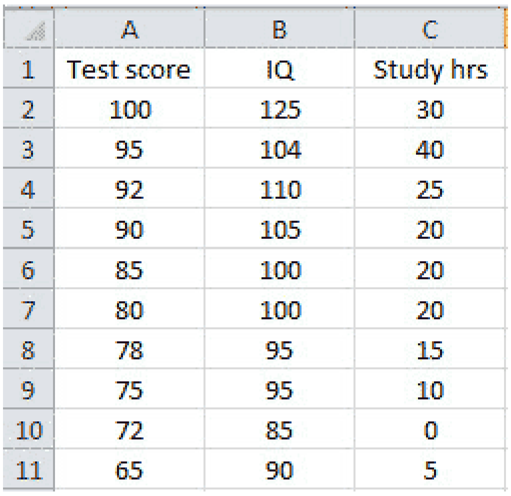
This enables the Analysis ToolPak. Now, when you click the Data tab, you will see a Data Analysis button in the upper right corner under the Data tab.

Data Entry With Excel

Data entry with Excel is easy. There are three main steps:

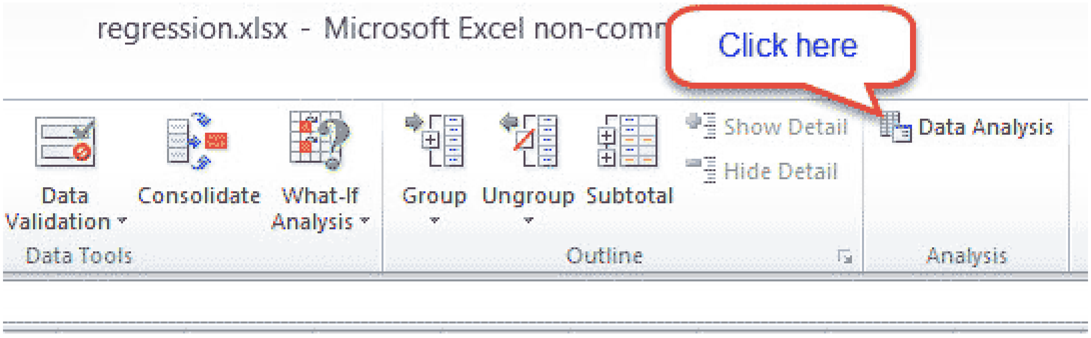
- Enter data on spreadsheet.
- Identify independent and dependent variables.
- Specify desired analyses.

To illustrate the process, we'll walk through each step, using data from our sample problem. First, we want to enter data on an Excel spreadsheet. This means listing data for each variable in adjacent columns, as shown below:

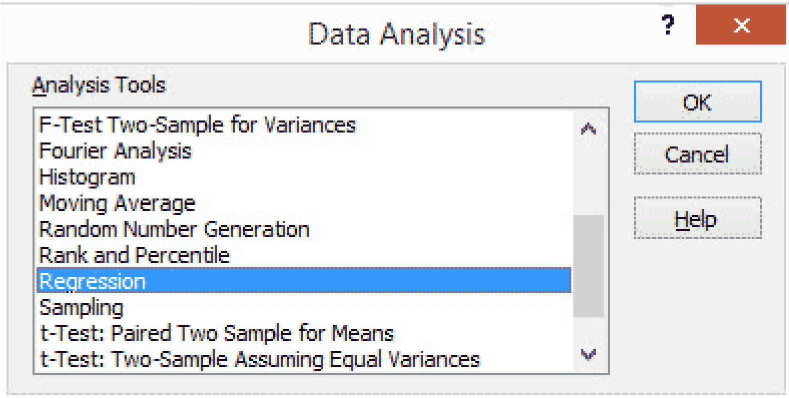


	A	B	C
1	Test score	IQ	Study hrs
2	100	125	30
3	95	104	40
4	92	110	25
5	90	105	20
6	85	100	20
7	80	100	20
8	78	95	15
9	75	95	10
10	72	85	0
11	65	90	5

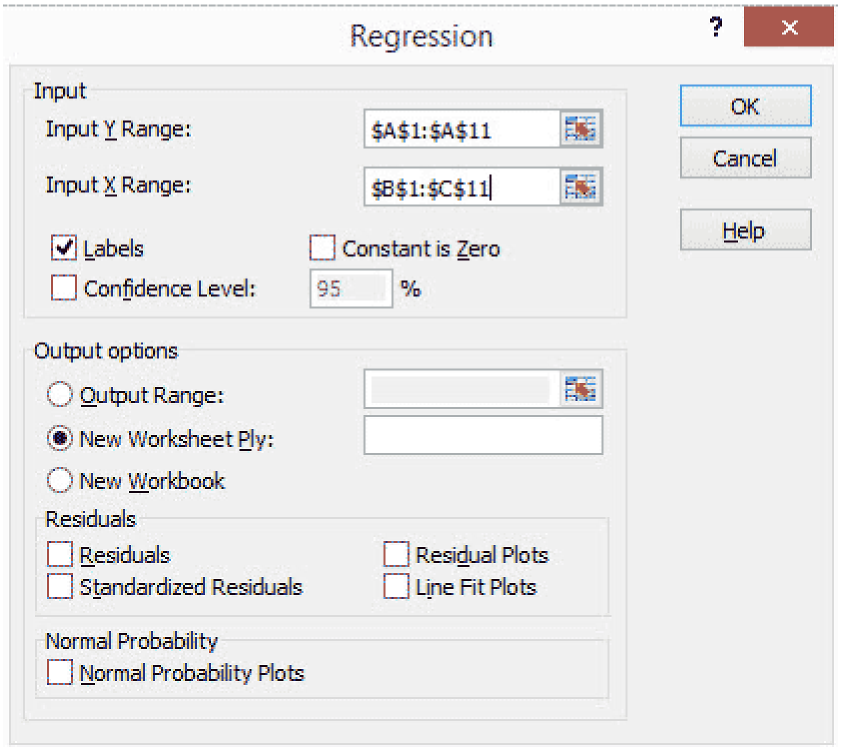
Next, we want to identify the independent and dependent variables. Begin by clicking the Data tab and the Data Analysis button.



This will open the Data Analysis dialog box. From the drop-down list, select "Regression" and click OK.



Excel will display the Regression dialog box. This is where you identify data fields for the independent and dependent variables. In the Input Y Range, enter coordinates for the dependent variable. In the Input X Range, enter coordinates for the independent variable(s). If you include column labels in these input ranges, check the Labels box. In the example below, we have included labels, so the Labels box is checked.



Input

Input Y Range: $A$1:$A$11

Input X Range: $B$1:$C$11

☑ Labels ☐ Constant is Zero

☐ Confidence Level: 95 %

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Residuals

☐ Residuals ☐ Residual Plots

☐ Standardized Residuals ☐ Line Fit Plots

Normal Probability

☐ Normal Probability Plots

By default, Excel will produce a standard set of outputs. For this sample problem, that's all we need; so click OK to generate standard regression outputs.

Note: If desired, you can request additional outputs in the form of residual plots and normal probability plots. To produce the plots, check the appropriate box(es) under Output options on the Regression dialog box.

Data Analysis With Excel

Excel provides everything we need to address the tasks we defined for this sample problem. Recall that we wanted to do three things:

- Develop a least-squares regression equation to predict test score, based on (1) IQ and (2) the number of hours that the student studied.
- Assess how well the regression equation predicts test score, the dependent variable.

- Assess the contribution of each independent variable (i.e., IQ and study hours) to the prediction.

Let's review the output produced by Excel and see how it addresses each task.

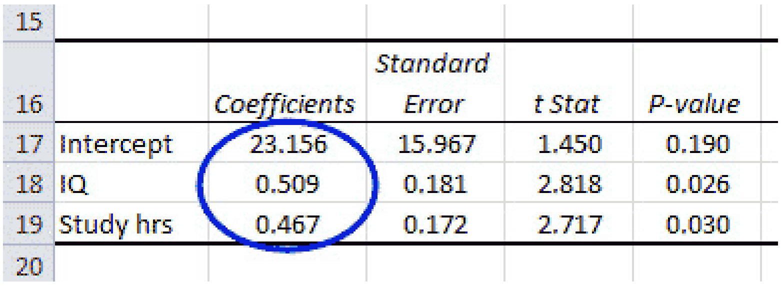
Regression Equation

The first task in our analysis is to define a linear, least-squares regression equation to predict test score, based on IQ and study hours. Since we have two independent variables, the equation takes the following form:

$$\hat{y} = b_0 + b_1x_1 + b_2x_2$$

In this equation, $\hat{y}$ is the *predicted* test score. The independent variables are IQ and study hours, which are denoted by x_1 and x_2, respectively. The regression coefficients are b_0, b_1, and b_2. On the right side of the equation, the only unknowns are the regression coefficients; so to specify the equation, we need to assign values to the coefficients.

In the previous lesson, we showed how to assign values to regression coefficients, using matrix algebra - a time-consuming, labor-intensive process by hand. Excel does all the hard work behind the scenes, and displays the result in a regression coefficients table:



15					
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*
17	Intercept	23.156	15.967	1.450	0.190
18	IQ	0.509	0.181	2.818	0.026
19	Study hrs	0.467	0.172	2.717	0.030
20					

Here, we see that the regression intercept (b_0) is 23.156, the regression coefficient for IQ (b_1) is 0.509, and the regression coefficient for study hours (b_2) is 0.467. So the least-squares regression equation can be re-written as:

$$\hat{y} = 23.156 + 0.505 * IQ + 0.467 * Hours$$

This is the only linear equation that satisfies a least-squares criterion. That means this equation fits the data from which it was created better than any other linear equation.

Coefficient of Multiple Determination

The fact that our equation fits the data better than any other linear equation does not guarantee that it fits the data well. We still need to ask: How well does our equation fit the data?

To answer this question, researchers look at the coefficient of multiple determination (R^2). The coefficient of multiple determination measures the proportion of variation in the dependent variable that can be predicted from the set of independent variables in the regression equation. When the regression equation fits the data well, R^2 will be large (i.e., close to 1); and vice versa.

The coefficient of multiple determination can be defined in terms of sums of squares:

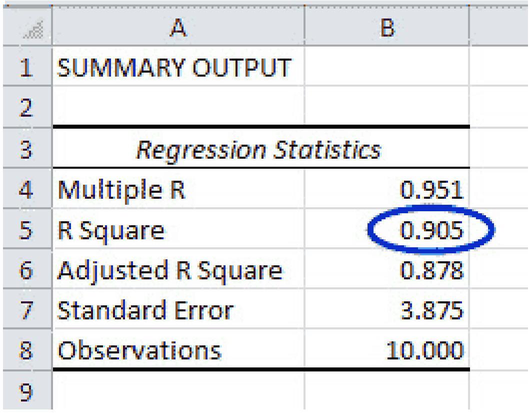
$$SSR = \Sigma (\hat{y} - y)^2$$

$$SSTO = \Sigma (y - y)^2$$

$$R^2 = SSR / SSTO$$

where SSR is the sum of squares due to regression, SSTO is the total sum of squares, $\hat{y}$ is the predicted value of the dependent variable, y is the dependent variable mean, and y is the dependent variable raw score.

Luckily, you will never have to compute the coefficient of multiple determination by hand. It is a standard output of Excel (and most other analysis packages), as shown below.



	A	B
1	SUMMARY OUTPUT	
2		
3	*Regression Statistics*	
4	Multiple R	0.951
5	R Square	0.905
6	Adjusted R Square	0.878
7	Standard Error	3.875
8	Observations	10.000
9		

A quick glance at the output suggests that the regression equation fits the data pretty well. The coefficient of muliple determination is 0.905. For our sample problem, this means 90.5% of test score variation can be explained by IQ and by hours spent in study.

An Alternative View of R^2

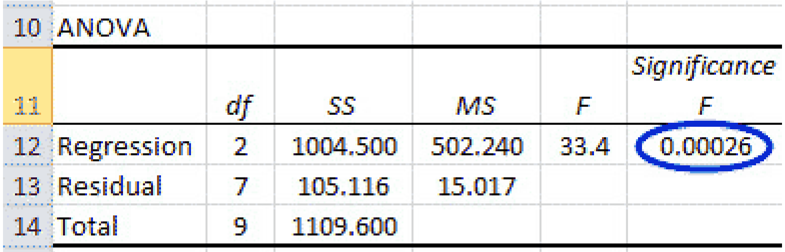
The coefficient of multiple correlation (R^2) is the square of the correlation between actual and predicted values of the dependent variable. Thus,

$$R^2 = r^2_{y,\,\hat{y}}$$

where y is the dependent variable raw score, ŷ is the predicted value of the dependent variable, and $r_{y,\,\hat{y}}$ is the correlation between y and ŷ.

ANOVA Table

Another way to evaluate the regression equation would be to assess the statistical significance of the regression sum of squares. For that, we examine the ANOVA table produced by Excel:



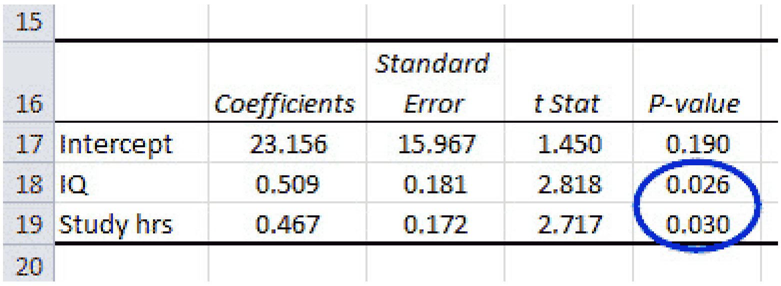
10	ANOVA					
11		*df*	*SS*	*MS*	*F*	*Significance F*
12	Regression	2	1004.500	502.240	33.4	0.00026
13	Residual	7	105.116	15.017		
14	Total	9	1109.600			

This table tests the statistical significance of the independent variables as predictors of the dependent variable. The last column of the table shows the results of an overall F test. The F statistic (33.4) is big, and the p value (0.00026) is small. This indicates that one or both independent variables has explanatory power beyond what would be expected by chance.

Like the coefficient of multiple correlation, the overall F test found in the ANOVA table suggests that the regression equation fits the data well.

Significance of Regression Coefficients

With multiple regression, there is more than one independent variable; so it is natural to ask whether a particular independent variable contributes significantly to the regression *after effects of other variables are taken into account*. The answer to this question can be found in the regression coefficients table:



15					
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*
17	Intercept	23.156	15.967	1.450	0.190
18	IQ	0.509	0.181	2.818	0.026
19	Study hrs	0.467	0.172	2.717	0.030
20					

The regression coefficients table shows the following information for each coefficient: its value, its standard error, a t-statistic, and the significance of the t-statistic. In this example, the t-statistics for IQ and study hours are both statistically significant at the 0.05 level. This means that IQ contributes significantly to the regression after effects of study hours are taken into account. And study hours contribute significantly to the regression after effects of IQ are taken into account.

Note: This analysis omits any consideration of multicollinearity, a topic we will cover in the next lesson. Be aware, however, that it is best practice to assess multicollinearity in the independent variables *before* testing significance of regression coefficients.

Final Thoughts

This lesson was all about multiple regression analysis. We used Excel, but the analysis would be much the same with other software packages. All major software packages (SAS, SPSS, Minitab, etc.) produce three key outputs:

- Regression coefficients, based on a least-squares criterion.
- Measures of goodness of fit, like a coefficient of multiple determination and/or an overall F test.
- Significance tests for individual regression coefficients.

If you can interpret these regression outputs from Excel, you should have no trouble interpreting the same outputs from other packages.

Stat Trek

Teach yourself statistics

Linear Regression

Introduction

Simple Regression

Multiple Regression

- Dummy Variables
- Interaction Effects
- Discriminant Analysis

Multicollinearity and Regression Analysis

In regression, multicollinearity refers to the extent to which independent variables are correlated. Multicollinearity exists when:

- One independent variable is correlated with another independent variable.
- One independent variable is correlated with a linear combination of two or more independent variables.

In this lesson, we'll examine the impact of multicollinearity on regresssion analysis.

The Multicollinearity Problem

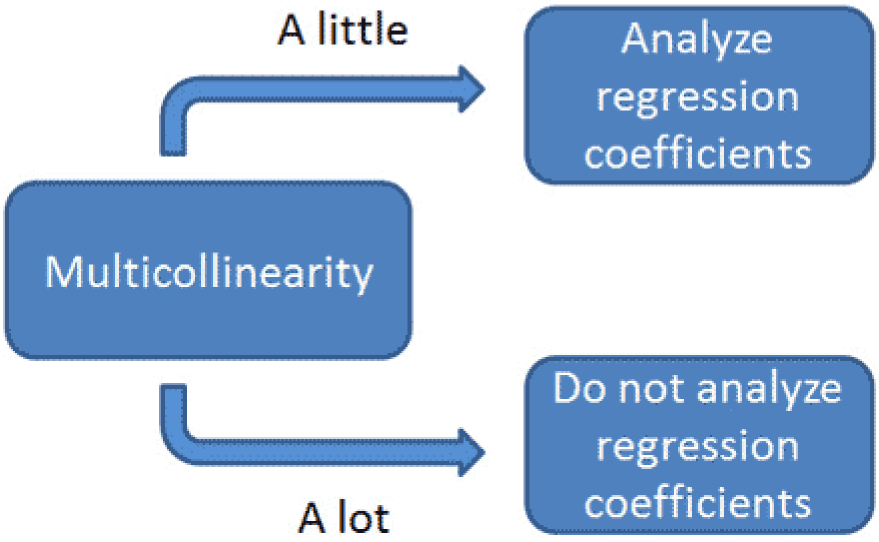
As part of regression analysis, researchers examine regression coefficients to assess the relative influence of independent variables. They look at the magnitude of coefficients, and they test the statistical significance of coefficients.

If the coefficient for a particular variable is significantly greater than zero, researchers judge that the variable contributes to the predictive ability of the regression equation. In this way, it is possible to distinguish variables that are more useful for prediction from those that are less useful.

This kind of analysis makes sense when multicollinearity is small. But it is problematic when multicollinearity is great. Here's why:

- When one independent variable is *perfectly* correlated with another independent variable (or with a combination of two or more other independent variables), a unique least-squares solution for regression coefficients does not exist.
- When one independent variable is *highly* correlated with another independent variable (or with a combination of two or more other independent variables), the marginal contribution of that independent variable is influenced by other independent variables. As a result:
    - Estimates for regression coefficients can be unreliable.
    - Tests of significance for regression coefficients can be misleading.

With this in mind, the analysis of regression coefficients should be contingent on the extent of multicollinearity. This means that the analysis of regression coefficients should be preceded by an analysis of multicollinearity.


A little
Analyze regression coefficients
Multicollinearity
Do not analyze regression coefficients
A lot

If the set of independent variables is characterized by a little bit of multicollinearity, the analysis of regression coefficients should be straightforward. If there is a lot of multicollinearity, the analysis will be hard to interpret and can be skipped.

Note: Multicollinearity makes it hard to assess the relative importance of independent variables, but it does not affect the usefulness of the regression equation for prediction. Even when multicollinearity is great, the least-squares regression equation can be highly predictive. So, if you are only interested in prediction, multicollinearity is not a problem.

How to Measure Multicollinearity

There are two popular ways to measure multicollinearity: (1) compute a coefficient of multiple determination for each independent variable, or (2) compute a variance inflation factor for each independent variable.

Coefficient of Multiple Determination

In the previous lesson, we described how the coefficient of multiple determination (R^2) measures the proportion of variance in the dependent variable that is explained by all of the independent variables.

If we ignore the dependent variable, we can compute a coefficient of multiple determination (R^2_k) for each of the *k* independent variables. We do this by regressing the *kth* independent variable on all of the other independent variables. That is, we treat X_k as the dependent variable and use the other independent variables to predict X_k.

How do we interpret R^2_k? If R^2_k equals zero, variable *k* is not correlated with any other independent variable; and multicollinearity is not a problem for variable *k*. As a rule of thumb, most analysts feel that multicollinearity is a potential problem when R^2_k is greater than 0.75; and, a serious problem when R^2_k is greater than 0.9.

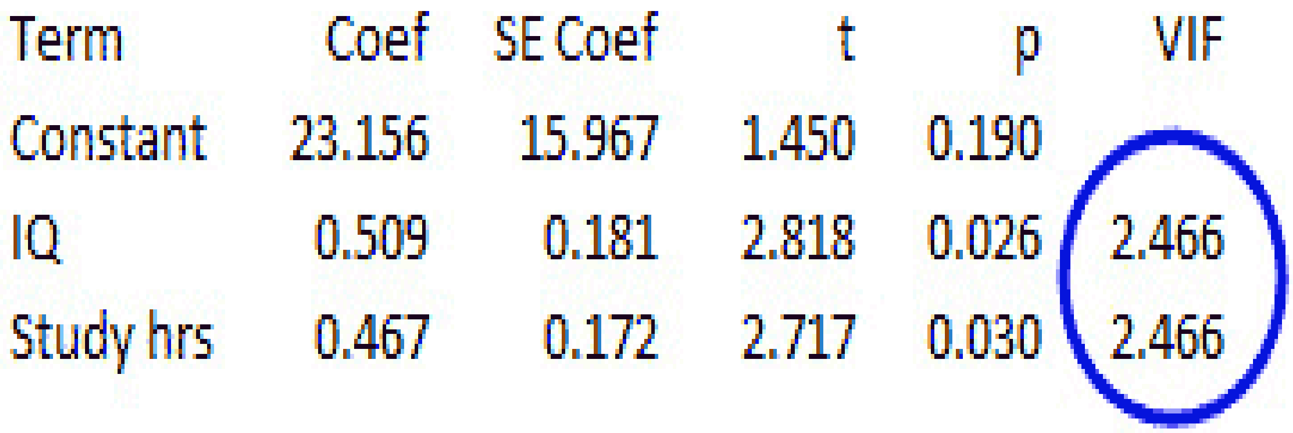
Variance Inflation Factor

The variance inflation factor is another way to express exactly the same information found in the coefficient of multiple correlation. A variance inflation factor is computed for each independent variable, using the following formula:

$$VIF_k = 1 / (1 - R^2_k)$$

where VIF_k is the variance inflation factor for variable *k*, and R^2_k is the coefficient of multiple determination for variable *k*.

In many statistical packages (e.g., SAS, SPSS, Minitab), the variance inflation factor is available as an optional regression output. In MiniTab, for example, the variance inflation factor can be displayed as part of the regression coefficient table.



Term	Coef	SE Coef	t	p	VIF
Constant	23.156	15.967	1.450	0.190	
IQ	0.509	0.181	2.818	0.026	2.466
Study hrs	0.467	0.172	2.717	0.030	2.466

The interpretation of the variance inflation factor mirrors the interpretation of the coefficient of multiple determination. If $\text{VIF}_k = 1$, variable *k* is not correlated with any other independent variable. As a rule of thumb, multicollinearity is a potential problem when VIF_k is greater than 4; and, a serious problem when it is greater than 10. The output above shows a VIF of 2.466, which indicates some multicollinearity but not enough to worry about.

Bottom line: If R^2_k is greater than 0.9 or VIF_k is greater than 10, it is likely that regression coefficients are poorly estimated. And significance tests on those coefficients may be misleading.

How to Deal with Multicollinearity

If you only want to predict the value of a dependent variable, you may not have to worry about multicollinearity. Multiple regression can produce a regression equation that will work for you, even when independent variables are highly correlated.

The problem arises when you want to assess the relative importance of an independent variable with a high R^2_k (or, equivalently, a high VIF_k). In this situation, try the following:

- Redesign the study to avoid multicollinearity. If you are working on a true experiment, the experimenter controls treatment levels. Choose treatment levels to minimize or eliminate correlations between independent variables.
- Increase sample size. Other things being equal, a bigger sample means reduced sampling error. The increased precision may overcome potential problems from multicollinearity.

- Remove one or more of the highly-correlated independent variables. Then, define a new regression equation, based on the remaining variables. Because the removed variables were redundant, the new equation should be nearly as predictive as the old equation; and coefficients should be easier to interpret because multicolinearity is reduced.
- Define a new variable equal to a linear combination of the highly-correlated variables. Then, define a new regression equation, using the new variable in place of the old highly-correlated variables.

Note: Multicollinearity only affects variables that are highly correlated. If the variable you are interested in has a small $R^2{}_j$, statistical analysis of its regression coefficient will be reliable and informative. That analysis will be valid, even when other variables exhibit high multicollinearity.

Test Your Understanding

In this section, two problems illustrate the role of multicollinearity in regression analysis. In Problem 1, we see what happens when multicollinearity is small; and in Problem 2, we see what happens when multicollinearity is big.

Problem 1

Consider the table below. It shows three performance measures for 10 students.

Student	Test score	IQ	Study hours
1	100	125	30
2	95	104	40
3	92	110	25
4	90	105	20
5	85	100	20
6	80	100	20
7	78	95	15
8	75	95	10
9	72	85	0
10	65	90	5

In the previous lesson, we used data from the table to develop a least-squares regression equation to predict test score. We also conducted statistical tests to assess the contribution of each independent variable (i.e., IQ and study hours) to the prediction.

For this problem,

- Measure multicollinearity, when IQ and Study Hours are independent variables.
- Discuss the impact of multicollinearity for interpreting statistical tests on IQ and Study Hours.

Solution

In this lesson, we described two ways to measure multicollinearity:

- Compute a coefficient of multiple determination (R^2_k) for each independent variable.
- Compute a variance inflation factor (VIF_k) for each independent variable.

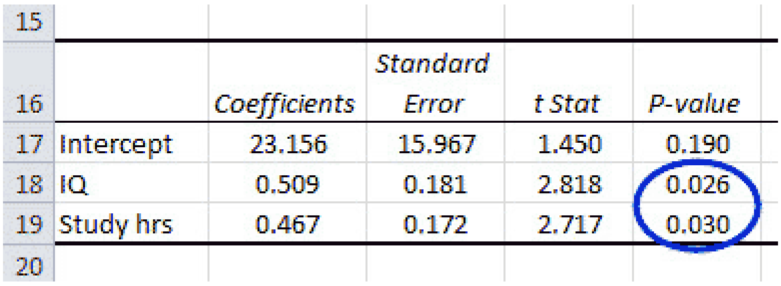
The two approaches are equivalent; so, in practice, you only need to do one or the other, but not both. In the previous lesson, we showed how to compute a coefficient of multiple determination with Excel, and how to derive a variance inflation factor from the coefficient of multiple determination.

Here are the variance inflation factors and the coefficients of multiple determination for the present problem.

Variable *k*	VIF_k	R^2_k
IQ	2.466	0.595
Study hours	2.466	0.595

We have rules of thumb to interpret VIF_k and R^2_k. Multicollinearity makes it hard to interpret the statistical significance of the regression coefficient for variable *k* when VIF_k is greater than 4 or when R^2_k is greater than 0.75. Since neither condition is evident in this problem, we can safely accept the results of statistical tests on regression coefficients.

We actually conducted those tests for this problem in the previous lesson. For convenience, key results are reproduced below:



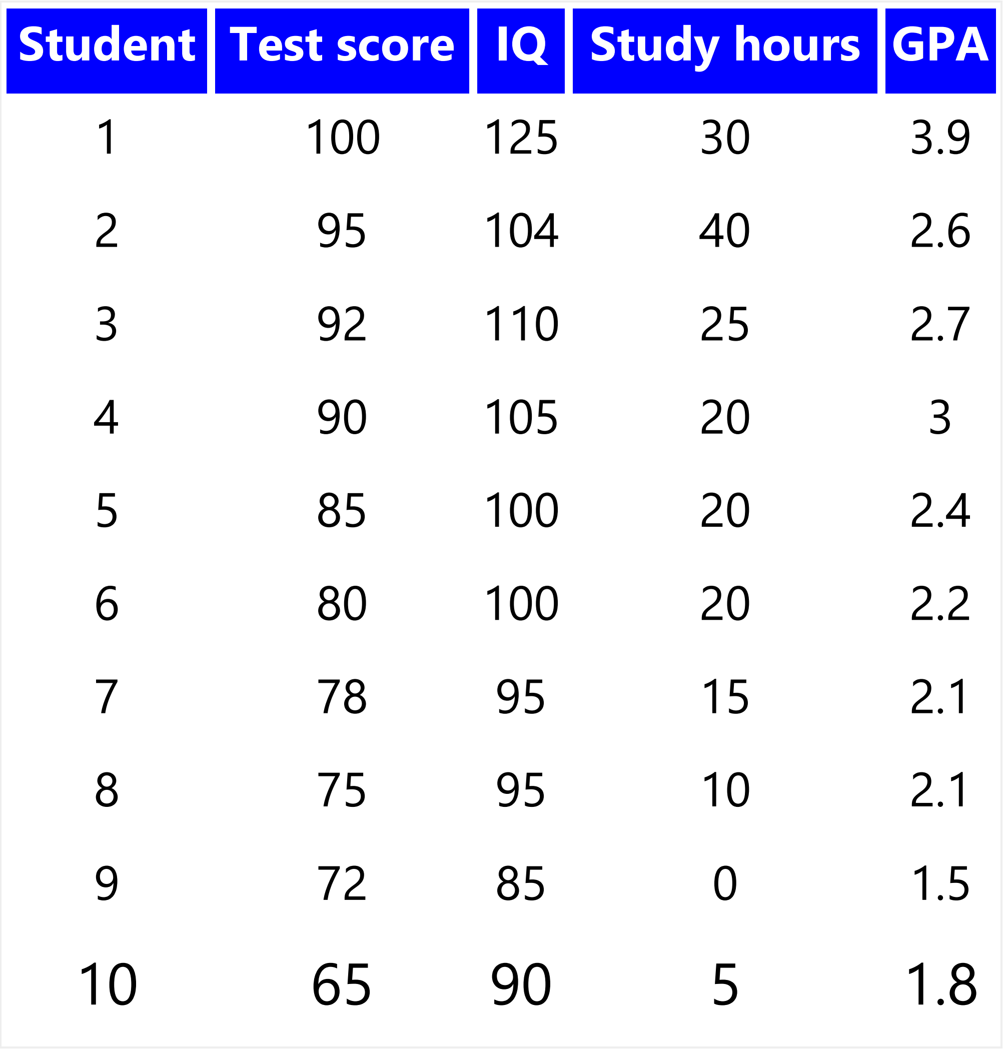
15					
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*
17	Intercept	23.156	15.967	1.450	0.190
18	IQ	0.509	0.181	2.818	0.026
19	Study hrs	0.467	0.172	2.717	0.030
20					

The p-values for IQ and for Study Hours are statistically significant at the 0.05 level. We can trust these findings, because multicollinearity is within acceptable levels.

It is also interesting to look at the effectiveness of the regression equation to predict the dependent variable, Test Score. As part of the analysis in the previous lesson, we found that the coefficient of determination (R^2) for the regression equation was 0.905. This means about 90% of the variance in Test Score was accounted for by IQ and Study Hours.

Problem 2

Problem 2 is identical to Problem 1, except that we've added a new independent variable - grade point average (GPA). Values for each variable appear in the table below.



Student	Test score	IQ	Study hours	GPA
1	100	125	30	3.9
2	95	104	40	2.6
3	92	110	25	2.7
4	90	105	20	3
5	85	100	20	2.4
6	80	100	20	2.2
7	78	95	15	2.1
8	75	95	10	2.1
9	72	85	0	1.5
10	65	90	5	1.8

Assume that you want predict Test Score, based on three independent variables - IQ, Study Hours, and GPA. As part of multiple regression analysis, you will assess the relative importance of each independent variable.

Before you make that assessment, you need to understand multicollinearity among the independent variables.

For this problem, do the following:

- Measure multicollinearity, based on IQ, Study Hours, and GPA.
- Discuss how multicollinearity affects your ability to interpret statistical tests on IQ, Study Hours, and GPA.

Solution

By now, we know that there are two ways to measure multicollinearity:

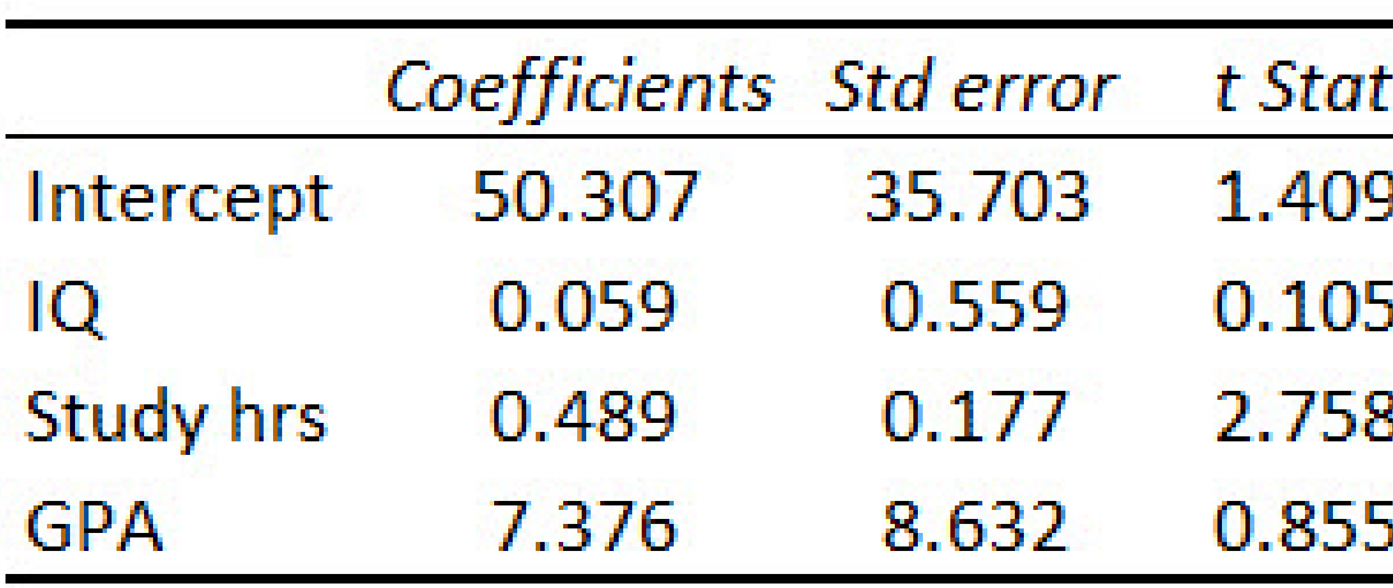
- Compute a coefficient of multiple determination (R^2_k) for each independent variable.
- Compute a variance inflation factor (VIF_k) for each independent variable.

We used the MiniTab formula to compute variance inflation factors, and we used Excel to compute coefficients of multiple determination. Here are the results of our analysis.

Variable *k*	**VIF_k**	**R^2_k**
IQ	22.64	0.956
Study hours	2.52	0.603
GPA	19.66	0.949

We have rules of thumb to interpret VIF_k and R^2_k. Multicollinearity makes it hard to interpret the statistical significance of the regression coefficient for variable *k* when VIF_k is greater than 4 or when R^2_k is greater than 0.75. Based on these guidelines, we would conclude that multicollinearity is not a problem for Study Hours, but it is a problem for IQ and GPA.

Here is the regression coefficients table for this problem. The table shows the following information for each coefficient: its value, its standard error, a t-statistic, and the significance of the t-statistic. Based on what you know about multicollinearity, how would you interpret results reported in this table?



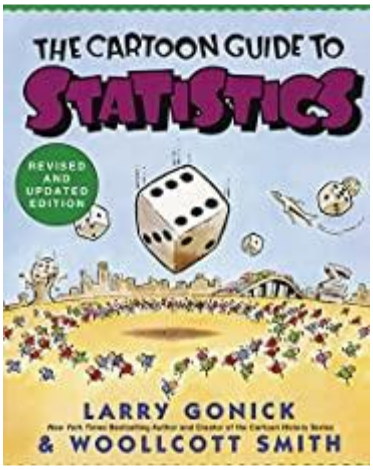
	Coefficients	*Std error*	*t Stat*
Intercept	50.307	35.703	1.409
IQ	0.059	0.559	0.105
Study hrs	0.489	0.177	2.758
GPA	7.376	8.632	0.855

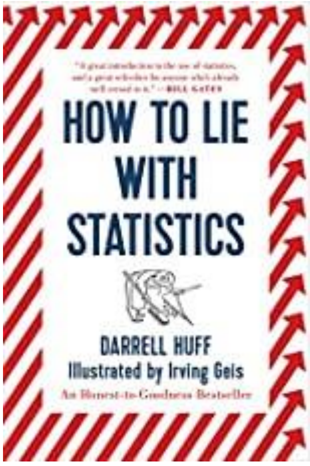
The p-value for Study Hours is statistically significant at the 0.05 level. We can trust this finding, because multicollinearity is within acceptable levels for the Study Hours variable. The p-values for IQ and GPA are not statistically significant. However, multicollinearity is very high for these variables; so their tests of significance are suspect. Despite the non-significant test results, we cannot say with confidence that neither IQ nor GPA are poor predictors of test score.

In fact, based on the analysis in Problem 1, we would conclude that IQ is a good predictor of test score. In this problem, the effect of IQ is strongly correlated with the other two predictors (R^2_{IQ} = 0.956). Similarly, the effect of GPA is strongly correlated with IQ and Study Hours (R^2_{IQ} = 0.949). This may explain the insignificant p-values for IQ and GPA in this problem. Here's the moral: Insignificant p-values may be misleading when the variables being tested suffer from multicollinearity.

And finally, we looked at the effectiveness of the regression equation to predict the dependent variable, Test Score. The coefficient of determination (R^2) for the regression equation was 0.916. This is slightly greater than the coefficient of determination (0.905) found in Problem 1, when we only used IQ and Study Hours as independent variables. Even though two predictors in this problem suffered from multicollinearity, the regression equation was still highly predictive. This illustrates the fact that multicollinearity does not affect the ability of the regression equation to predict a dependent variable; it only affects statistical tests on regression coefficients.

Last lesson Next lesson

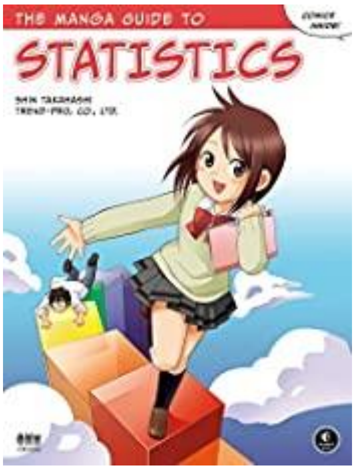



HOW TO LIE WITH STATISTICS
DARRELL HUFF
Illustrated by Irving Geis
An Honest-to-Goodness Bestseller

How to Lie with Statistics

$9.99~~$13.95~~

(1151)



The Manga Guide to Statistics

$18.20~~$19.95~~

(112)

NEW YORK TIMES BESTSELLER
naked statistics
STRIPPING THE DREAD FROM THE DATA
charles wheelan
AUTHOR OF NAKED ECONOMICS

- About
- Contact
- Privacy
- Terms
- Advertising

Casio fx-9860GII Graphing Calculator, Black

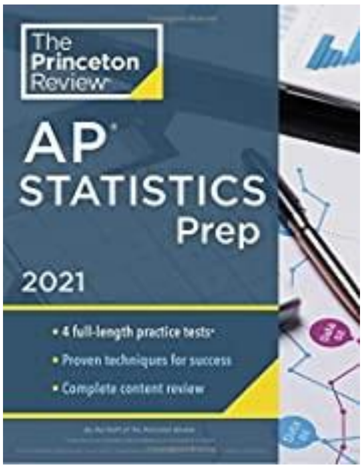
$63.99~~$79.99~~

(834)

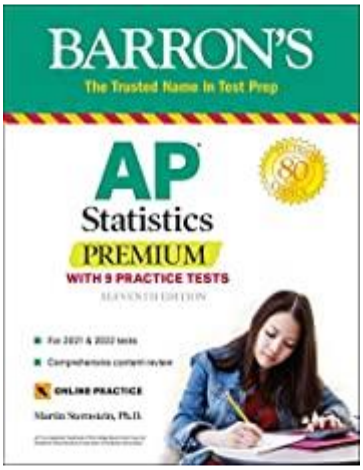
Texas Instruments TI-Nspire CX II Color Graphing Calcul...

$129.99~~$165.00~~

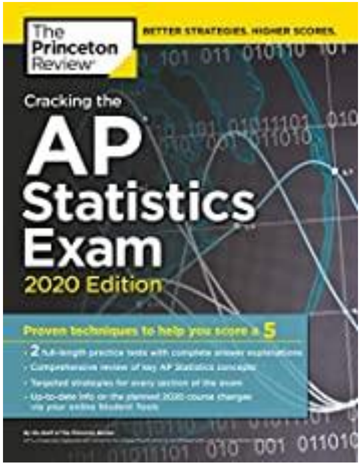
(1100)

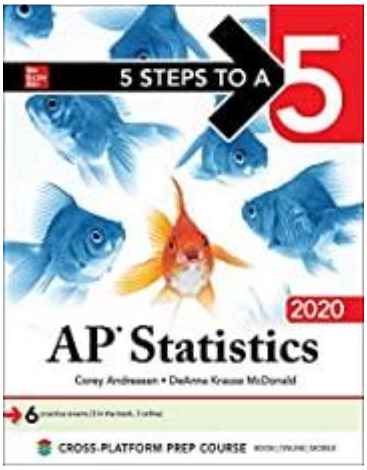


The Princeton Review
AP STATISTICS Prep
2021
• 4 full-length practice tests
• Proven techniques for success
• Complete content review

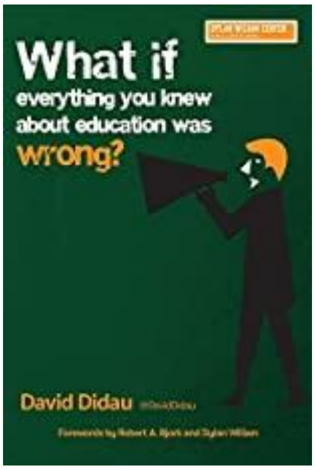


BARRON'S
The Trusted Name in Test Prep
AP Statistics
PREMIUM
WITH 9 PRACTICE TESTS
ELEVENTH EDITION
ONLINE PRACTICE


The Princeton Review
BETTER STRATEGIES. HIGHER SCORES.
Cracking the
AP Statistics Exam
2020 Edition
Proven techniques to help you score a 5


5 STEPS TO A 5
2020
AP Statistics
Corey Andreasen • DeAnna Krause McDonald
CROSS-PLATFORM PREP COURSE

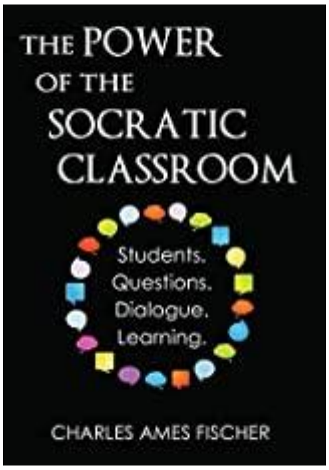
Education and Teaching



What If Everything You Knew About Education Was Wrong?

$19.89~~$29.95~~

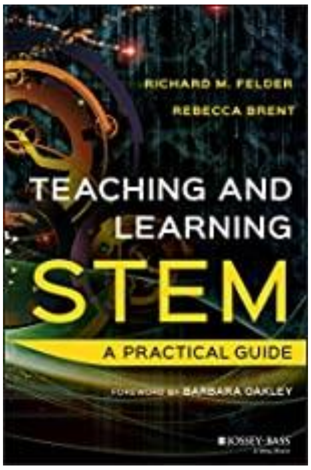
(61)

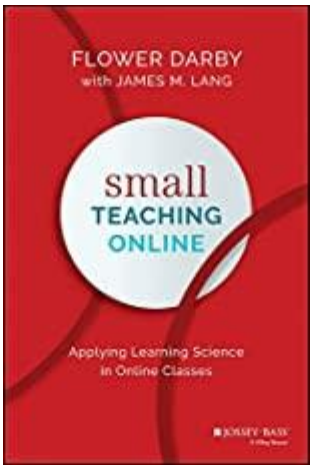


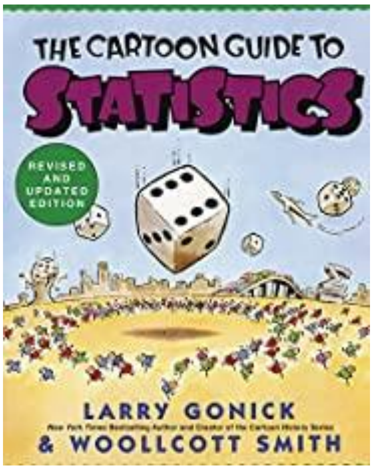
The Power of the Socratic Classroom: Students. Quest...

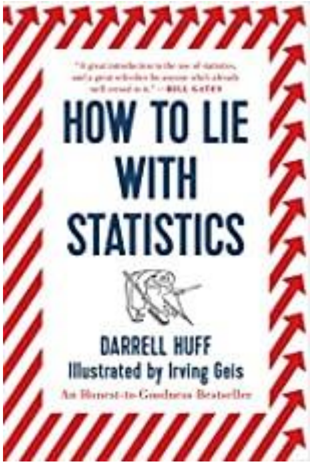
$35.00

(30)




FLOWER DARBY
with JAMES M. LANG
small
TEACHING
ONLINE
Applying Learning Science in Online Classes
JOSSEY-BASS

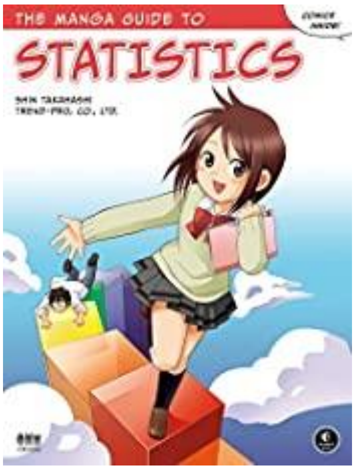

THE CARTOON GUIDE TO
STATISTICS
REVISED AND UPDATED EDITION
LARRY GONICK
& WOOLLCOTT SMITH


HOW TO LIE WITH STATISTICS
DARRELL HUFF
Illustrated by Irving Geis
An Honest-to-Goodness Bestseller

How to Lie with Statistics

$9.99~~$13.95~~

(1153)



The Manga Guide to Statistics

$18.20~~$19.95~~

(112)

NEW YORK TIMES BESTSELLER
naked statistics
STRIPPING THE DREAD FROM THE DATA
charles wheelan
AUTHOR OF NAKED ECONOMICS

Last lesson Next lesson