Linear Regression and Correlation: The Regression Equation*

Susan Dean
Barbara Illowsky, Ph.D.

This work is produced by OpenStax-CNX and licensed under the
Creative Commons Attribution License 3.0†

Abstract

Linear Regression and Correlation: The Regression Equation is a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean. Contributions from Roberta Bloom include instructions for finding and graphing the regression equation and scatterplot using the LinRegTTest on the TI-83,83+,84+ calculators.

Data rarely fit a straight line exactly. Usually, you must be satisfied with rough predictions. Typically, you have a set of data whose scatter plot appears to **"fit"** a straight line. This is called a **Line of Best Fit or Least Squares Line**.

1 Optional Collaborative Classroom Activity

If you know a person's pinky (smallest) finger length, do you think you could predict that person's height? Collect data from your class (pinky finger length, in inches). The independent variable, x, is pinky finger length and the dependent variable, y, is height.

For each set of data, plot the points on graph paper. Make your graph big enough and **use a ruler**. Then "by eye" draw a line that appears to "fit" the data. For your line, pick two convenient points and use them to find the slope of the line. Find the y-intercept of the line by extending your lines so they cross the y-axis. Using the slopes and the y-intercepts, write your equation of "best fit". Do you think everyone will have the same equation? Why or why not?

Using your equation, what is the predicted height for a pinky length of 2.5 inches?

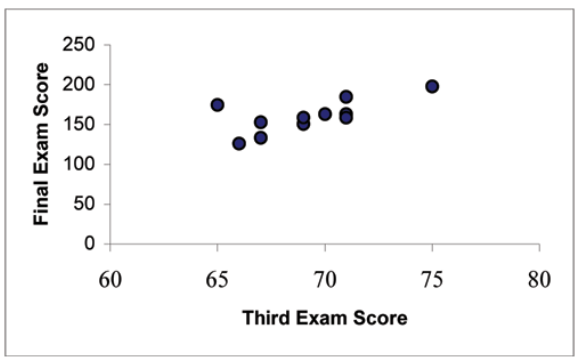
Example 1

A random sample of 11 statistics students produced the following data where x is the third exam score, out of 80, and y is the final exam score, out of 200. Can you predict the final exam score of a random student if you know the third exam score?

*Version 1.15: Jun 26, 2012 11:18 am -0500
†http://creativecommons.org/licenses/by/3.0/

x (third exam score)	y (final exam score)
65	175
67	133
71	185
71	163
66	126
75	198
67	153
70	163
71	159
69	151
69	159

(a)



(b)

Figure 1: (a) Table showing the scores on the final exam based on scores from the third exam. (b) Scatter plot showing the scores on the final exam based on scores from the third exam.

The third exam score, x, is the independent variable and the final exam score, y, is the dependent variable. We will plot a regression line that best "fits" the data. If each of you were to fit a line "by eye", you would draw different lines. We can use what is called a **least-squares regression line** to obtain the best fit line.

Consider the following diagram. Each point of data is of the the form $(x, y)$and each point of the line of best fit using least-squares linear regression has the form $\left(x, \hat{y}\right)$.

The $\hat{y}$ is read **"y hat"** and is the **estimated value of** y. It is the value of y obtained using the regression line. It is not generally equal to y from data.

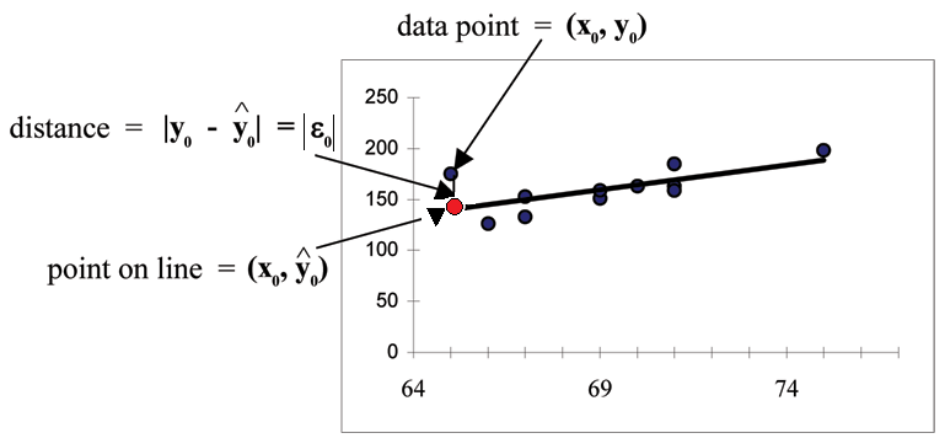


Figure 2

The term $y_0 - \hat{y}_0 = \epsilon_0$ is called the **"error" or residual**. It is not an error in the sense of a mistake. The **absolute value of a residual** measures the vertical distance between the actual value of y and the estimated value of y. In other words, it measures the vertical distance between the actual data point and the predicted point on the line.

If the observed data point lies above the line, the residual is positive, and the line underestimates the actual data value for y. If the observed data point lies below the line, the residual is negative, and the line overestimates that actual data value for y.

In the diagram above, $y_0 - \hat{y}_0 = \epsilon_0$ is the residual for the point shown. Here the point lies above the line and the residual is positive.

ϵ = the Greek letter **epsilon**

For each data point, you can calculate the residuals or errors, $y_i - \hat{y}_i = \epsilon_i$ for $i = 1, 2, 3, ..., 11$.

Each $|\epsilon|$ is a vertical distance.

For the example about the third exam scores and the final exam scores for the 11 statistics students, there are 11 data points. Therefore, there are 11 ϵ values. If you square each ϵ and add, you get

$(\epsilon_1)^2 + (\epsilon_2)^2 + ... + (\epsilon_{11})^2 = \sum_{i=1}^{11} \epsilon^2$

This is called the **Sum of Squared Errors (SSE)**.

Using calculus, you can determine the values of a and b that make the **SSE** a minimum. When you make the **SSE** a minimum, you have determined the points that are on the line of best fit. It turns out that the line of best fit has the equation:

$$\hat{y} = a + \mathrm{bx} \tag{1}$$

where $a = \overline{y} - b \cdot \overline{x}$ and $b = \frac{\Sigma(x-\overline{x})\cdot(y-\overline{y})}{\Sigma(x-\overline{x})^2}$.

$\overline{x}$ and $\overline{y}$ are the sample means of the x values and the y values, respectively. The best fit line always passes through the point $(\overline{x}, \overline{y})$.

The slope b can be written as $b = r \cdot \left(\frac{s_y}{s_x}\right)$ where s_y = the standard deviation of the y values and s_x = the standard deviation of the x values. r is the correlation coefficient which is discussed in the next section.

Least Squares Criteria for Best Fit

The process of fitting the best fit line is called **linear regression**. The idea behind finding the best fit line is based on the assumption that the data are scattered about a straight line. The criteria for the best fit line is that the sum of the squared errors (SSE) is minimized, that is made as small as possible. Any other line you might choose would have a higher SSE than the best fit line. This best fit line is called the **least squares regression line** .

NOTE: Computer spreadsheets, statistical software, and many calculators can quickly calculate the best fit line and create the graphs. The calculations tend to be tedious if done by hand. Instructions to use the TI-83, TI-83+, and TI-84+ calculators to find the best fit line and create a scatterplot are shown at the end of this section.

THIRD EXAM vs FINAL EXAM EXAMPLE:

The graph of the line of best fit for the third exam/final exam example is shown below:

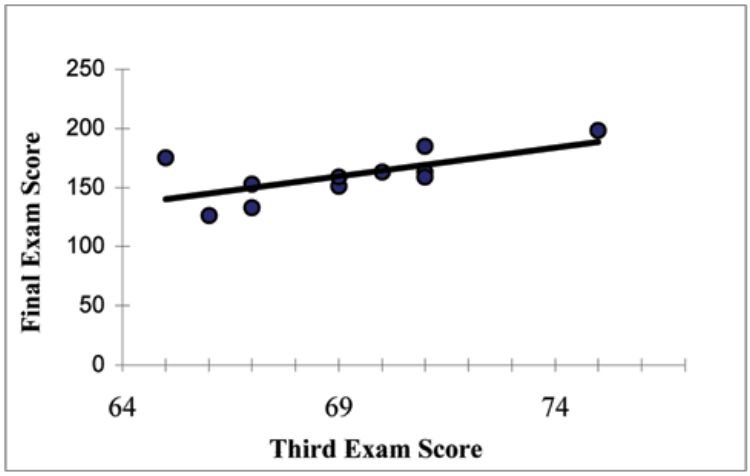


Figure 3

The least squares regression line (best fit line) for the third exam/final exam example has the equation:

$$\hat{y} = -173.51 + 4.83\text{x} \tag{2}$$

NOTE:

Remember, it is always important to plot a scatter diagram first. If the scatter plot indicates that there is a linear relationship between the variables, then it is reasonable to use a best fit line to make predictions for y given x within the domain of x-values in the sample data, **but not necessarily for x-values outside that domain.**

You could use the line to predict the final exam score for a student who earned a grade of 73 on the third exam.

You should NOT use the line to predict the final exam score for a student who earned a grade of 50 on the third exam, because 50 is not within the domain of the x-values in the sample data, which are between 65 and 75.

UNDERSTANDING SLOPE

The slope of the line, b, describes how changes in the variables are related. It is important to interpret the slope of the line in the context of the situation represented by the data. You should be able to write a sentence interpreting the slope in plain English.

INTERPRETATION OF THE SLOPE: The slope of the best fit line tells us how the dependent variable (y) changes for every one unit increase in the independent (x) variable, on average.

THIRD EXAM vs FINAL EXAM EXAMPLE

Slope: The slope of the line is b = 4.83.

Interpretation: For a one point increase in the score on the third exam, the final exam score increases by 4.83 points, on average.

2 Using the TI-83+ and TI-84+ Calculators

Using the Linear Regression T Test: LinRegTTest

Step 1. In the STAT list editor, enter the X data in list L1 and the Y data in list L2, paired so that the corresponding (x,y) values are next to each other in the lists. (If a particular pair of values is repeated, enter it as many times as it appears in the data.)

Step 2. On the STAT TESTS menu, scroll down with the cursor to select the LinRegTTest. (Be careful to select LinRegTTest as some calculators may also have a different item called LinRegTInt.)

Step 3. On the LinRegTTest input screen enter: Xlist: L1 ; Ylist: L2 ; Freq: 1

Step 4. On the next line, at the prompt β or ρ, highlight "$\neq 0$" and press ENTER

Step 5. Leave the line for "RegEq:" blank

Step 6. Highlight Calculate and press ENTER.

LinRegTTest Input Screen and Output Screen

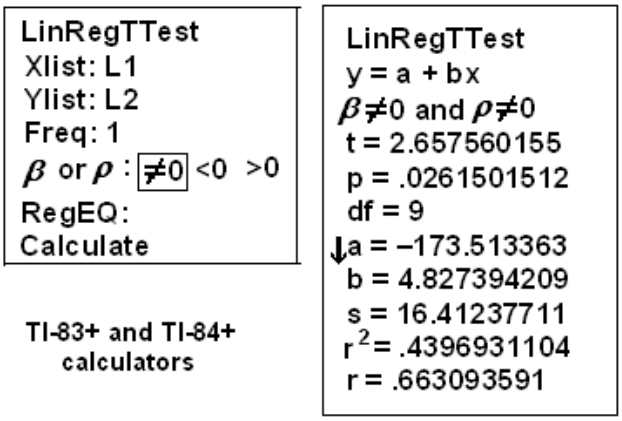


Figure 4

The output screen contains a lot of information. For now we will focus on a few items from the output, and will return later to the other items.

The second line says y=a+bx. Scroll down to find the values a=-173.513, and b=4.8273 ; the equation of the best fit line is $\hat{y} = -173.51 + 4.83x$

The two items at the bottom are r^2 = .43969 and r=.663. For now, just note where to find these values; we will discuss them in the next two sections.

Graphing the Scatterplot and Regression Line

Step 1. We are assuming your X data is already entered in list L1 and your Y data is in list L2
Step 2. Press 2nd STATPLOT ENTER to use Plot 1
Step 3. On the input screen for PLOT 1, highlight **On** and press ENTER
Step 4. For TYPE: highlight the very first icon which is the scatterplot and press ENTER
Step 5. Indicate Xlist: L1 and Ylist: L2
Step 6. For Mark: it does not matter which symbol you highlight.
Step 7. Press the ZOOM key and then the number 9 (for menu item "ZoomStat") ; the calculator will fit the window to the data
Step 8. To graph the best fit line, press the "Y=" key and type the equation -173.5+4.83X into equation Y1. (The X key is immediately left of the STAT key). Press ZOOM 9 again to graph it.
Step 9. Optional: If you want to change the viewing window, press the WINDOW key. Enter your desired window using Xmin, Xmax, Ymin, Ymax

**With contributions from Roberta Bloom