

### 3.3: Simple Linear Regression

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.

#### Example 3.3.1

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?

1a: Table showing the scores on the final exam based on scores from the third exam.

$x$ (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

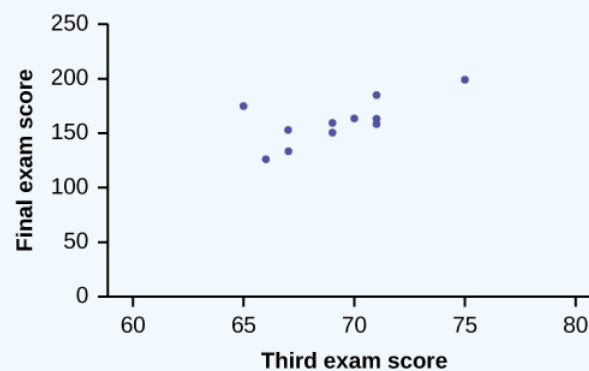


Figure 3.3.1: Scatter plot showing the scores on the final exam based on scores from the third exam.

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

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.

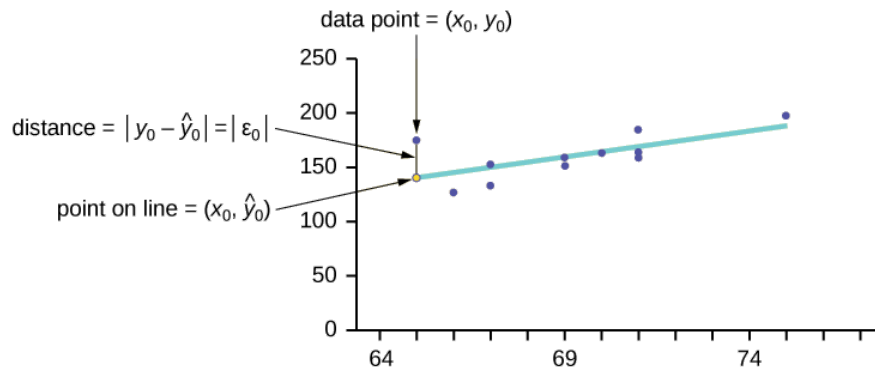


Figure 3.3.2

The term  $y_0 - \hat{y}_0 = \varepsilon_0$  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 in Figure,  $y_0 - \hat{y}_0 = \varepsilon_0$  is the residual for the point shown. Here the point lies above the line and the residual is positive.

$\varepsilon$  = the Greek letter **epsilon**

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

Each  $|\varepsilon|$  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  $\varepsilon$  values. If you square each  $\varepsilon$  and add, you get

$$(\varepsilon_1)^2 + (\varepsilon_2)^2 + \dots + (\varepsilon_{11})^2 = \sum_{i=1}^{11} \varepsilon_i^2 \quad (3.3.1)$$

Equation 3.3.1 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:

$$\hat{y} = a + bx \quad (3.3.2)$$

where

- $a = \bar{y} - b\bar{x}$  and
- $b = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2}$ .

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

The slope  $b$  can be written as  $b = r \left( \frac{s_y}{s_x} \right)$  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 Square 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**.

Computer spreadsheets, statistical software, and many calculators can quickly calculate the best-fit line and create the graphs.

### THIRD EXAM vs FINAL EXAM EXAMPLE:

The graph of the line of best fit for the third-exam/final-exam example is as follows:

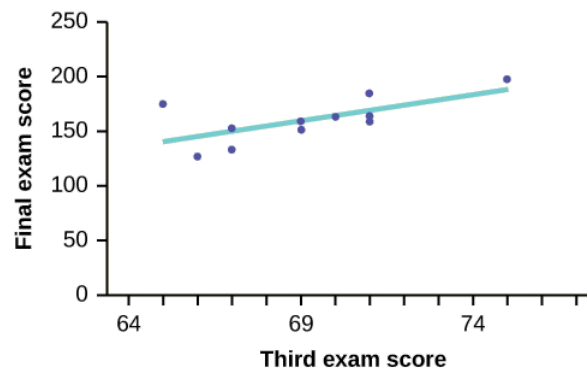


Figure 3.3.3

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

$$\hat{y} = -173.51 + 4.83x \quad (3.3.3)$$

#### REMINDER

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.

#### The Correlation Coefficient $r$

Besides looking at the scatter plot and seeing that a line seems reasonable, how can you tell if the line is a good predictor? Use the correlation coefficient as another indicator (besides the scatterplot) of the strength of the relationship between  $x$  and  $y$ . The **correlation coefficient**,  $r$ , developed by Karl Pearson in the early 1900s, is numerical and provides a measure of strength and direction of the linear association between the independent variable  $x$  and the dependent variable  $y$ .

The correlation coefficient is calculated as

$$r = \frac{n \sum(xy) - (\sum x)(\sum y)}{\sqrt{[n \sum x^2 - (\sum x)^2][n \sum y^2 - (\sum y)^2]}} \quad (3.3.4)$$

where  $n$  = the number of data points.

If you suspect a linear relationship between  $x$  and  $y$ , then  $r$  can measure how strong the linear relationship is.

What the VALUE of  $r$  tells us:

- The value of  $r$  is always between  $-1$  and  $+1$ :  $-1 \leq r \leq 1$ .

The size of the correlation  $r$  indicates the strength of the linear relationship between  $x$  and  $y$ . Values of  $r$  close to  $-1$  or to  $+1$  indicate a stronger linear relationship between  $x$  and  $y$ .

- If  $r = 0$  there is absolutely no linear relationship between  $x$  and  $y$  (**no linear correlation**).
- If  $r = 1$ , there is perfect positive correlation. If  $r = -1$ , there is perfect negative correlation. In both these cases, all of the original data points lie on a straight line. Of course, in the real world, this will not generally happen.

What the SIGN of  $r$  tells us

- A positive value of  $r$  means that when  $x$  increases,  $y$  tends to increase and when  $x$  decreases,  $y$  tends to decrease (**positive correlation**).
- A negative value of  $r$  means that when  $x$  increases,  $y$  tends to decrease and when  $x$  decreases,  $y$  tends to increase (**negative correlation**).

The sign of  $r$  is the same as the sign of the slope,  $b$ , of the best-fit line.

Strong correlation does not suggest that  $x$  causes  $y$  or  $y$  causes  $x$ . We say "**correlation does not imply causation.**"

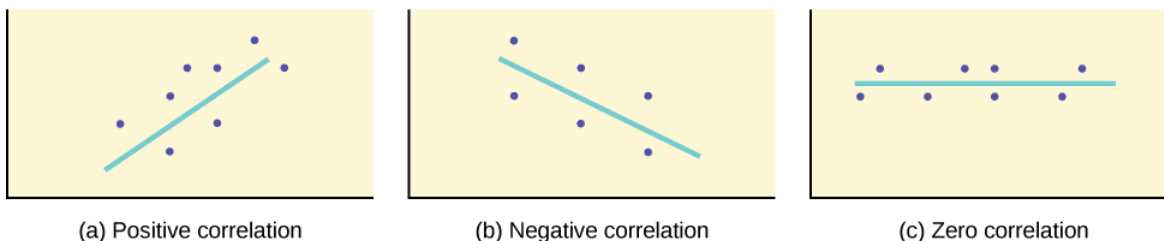


Figure 3.3.5: (a) A scatter plot showing data with a positive correlation.  $0 < r < 1$  (b) A scatter plot showing data with a negative correlation.  $-1 < r < 0$  (c) A scatter plot showing data with zero correlation.  $r = 0$

The formula for  $r$  looks formidable. However, computer spreadsheets, statistical software, and many calculators can quickly calculate  $r$ . The correlation coefficient  $r$  is the bottom item in the output screens for the LinRegTTest on the TI-83, TI-83+, or TI-84+ calculator (see previous section for instructions).

### The Coefficient of Determination

The variable  $r^2$  is called *the coefficient of determination* and is the square of the correlation coefficient, but is usually stated as a percent, rather than in decimal form. It has an interpretation in the context of the data:

- $r^2$ , when expressed as a percent, represents the percent of variation in the dependent (predicted) variable  $y$  that can be explained by variation in the independent (explanatory) variable  $x$  using the regression (best-fit) line.
- $1 - r^2$ , when expressed as a percentage, represents the percent of variation in  $y$  that is NOT explained by variation in  $x$  using the regression line. This can be seen as the scattering of the observed data points about the regression line.

Consider the third exam/final exam example introduced in the previous section

- The line of best fit is:  $\hat{y} = -173.51 + 4.83x$
- The correlation coefficient is  $r = 0.6631$
- The coefficient of determination is  $r^2 = 0.6631^2 = 0.4397$
- **Interpretation of  $r^2$  in the context of this example:**
- Approximately 44% of the variation (0.4397 is approximately 0.44) in the final-exam grades can be explained by the variation in the grades on the third exam, using the best-fit regression line.
- Therefore, approximately 56% of the variation ( $1 - 0.44 = 0.56$ ) in the final exam grades can NOT be explained by the variation in the grades on the third exam, using the best-fit regression line. (This is seen as the scattering of the points about the line.)

### Summary

A regression line, or a line of best fit, can be drawn on a scatter plot and used to predict outcomes for the  $x$  and  $y$  variables in a given data set or sample data. There are several ways to find a regression line, but usually the least-squares regression line is used because it creates a uniform line. Residuals, also called "errors," measure the distance from the actual value of  $y$  and the estimated value of  $y$ . The Sum of Squared Errors, when set to its minimum, calculates the points on the line of best fit. Regression lines can be used to predict values within the given set of data, but should not be used to make predictions for values outside the set of data.

The correlation coefficient  $r$  measures the strength of the linear association between  $x$  and  $y$ . The variable  $r$  has to be between  $-1$  and  $+1$ . When  $r$  is positive, the  $x$  and  $y$  will tend to increase and decrease together. When  $r$  is negative,  $x$  will increase and  $y$  will decrease, or the opposite,  $x$  will decrease and  $y$  will increase. The coefficient of determination  $r^2$ , is equal to the square of the correlation coefficient. When expressed as a percent,  $r^2$  represents the percent of variation in the dependent variable  $y$  that can be explained by variation in the independent variable  $x$  using the regression line.

## WeBWork Problems

## Glossary

### Coefficient of Correlation

a measure developed by Karl Pearson (early 1900s) that gives the strength of association between the independent variable and the dependent variable; the formula is:

$$r = \frac{n \sum xy - (\sum x)(\sum y)}{\sqrt{[n \sum x^2 - (\sum x)^2][n \sum y^2 - (\sum y)^2]}} \quad (3.3.5)$$

where  $n$  is the number of data points. The coefficient cannot be more than 1 or less than  $-1$ . The closer the coefficient is to  $\pm 1$ , the stronger the evidence of a significant linear relationship between  $x$  and  $y$ .

## Contributors and Attributions

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at <http://cnx.org/contents/30189442-699...b91b9de@18.114>.

---

This page titled [3.3: Simple Linear Regression](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [OpenStax](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [12.4: The Regression Equation](#) by OpenStax is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/introductory-statistics>.
- [1.2: Definitions of Statistics, Probability, and Key Terms](#) by OpenStax is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/introductory-statistics>.