Linear Regression and Correlation: Correlation Coefficient and Coefficient of Determination*

Susan Dean
Barbara Illowsky, Ph.D.

Abstract

Linear Regression and Correlation: The Correlation Coefficient and Coefficient of Determination is a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean with contributions from Roberta Bloom. The name has been changed from Correlation Coefficient.

1 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 a numerical measure of the strength of association between the independent variable x and the dependent variable y.

The correlation coefficient is calculated as

$$r = \frac{n \cdot \Sigma x \cdot y - (\Sigma x) \cdot (\Sigma y)}{\sqrt{\left[n \cdot \Sigma x^2 - (\Sigma x)^2\right] \cdot \left[n \cdot \Sigma y^2 - (\Sigma y)^2\right]}} \tag{1}$$

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)**.

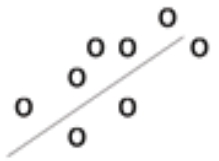
*Version 1.12: Jun 26, 2012 12:10 pm +0000
†http://creativecommons.org/licenses/by/3.0/

- 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

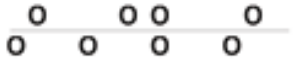
- A positive value of r means that when x increases, y tends to increase and when x decreases, y tends to decrease **(positive correlation)**.
- A negative value of r means that when x increases, y tends to decrease and when x decreases, y tends to increase **(negative correlation)**.
- The sign of r is the same as the sign of the slope, b, of the best fit line.

NOTE: Strong correlation does not suggest that x causes y or y causes x. We say **"correlation does not imply causation."** For example, every person who learned math in the 17th century is dead. However, learning math does not necessarily cause death!



(a) Positive Correlation

(b) Negative Correlation



(c) Zero Correlation

Figure 1: (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).

2 The Coefficient of Determination

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

- r^2, when expressed as a percent, represents the percent of variation in the dependent variable y that can be explained by variation in the independent variable x using the regression (best fit) line.
- 1-r^2, when expressed as a percent, 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: $\hat{y} = -173.51 + 4.83$x
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.)

**With contributions from Roberta Bloom.

Glossary

Definition 1: 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:

$$r = \frac{n\sum \text{xy} - (\sum x)(\sum y)}{\sqrt{\left[n\sum x^2 - (\sum x)^2\right]\left[n\sum y^2 - (\sum y)^2\right]}}, \tag{2}$$

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