

LINEAR REGRESSION AND CORRELATION: FACTS ABOUT THE CORRELATION COEFFICIENT FOR LINEAR REGRESSION*

Susan Dean
Barbara Illowsky, Ph.D.

This work is produced by OpenStax-CNX and licensed under the
Creative Commons Attribution License 2.0[†]

Abstract

This module provides an overview of Facts About the Correlation Coefficient for Linear Regression as a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean.

- A positive r means that when x increases, y increases and when x decreases, y decreases (**positive correlation**).
- A negative r means that when x increases, y decreases and when x decreases, y increases (**negative correlation**).
- An r of zero means there is absolutely no linear relationship between x and y (**no correlation**).
- High 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!

*Version 1.7: Jan 17, 2009 2:43 pm -0600

[†]<http://creativecommons.org/licenses/by/2.0/>

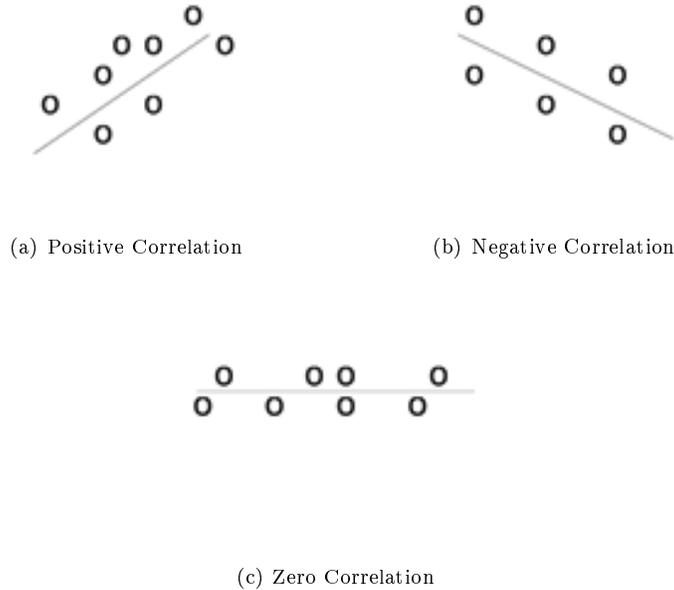


Figure 1: (a) A scatter plot showing data with a positive correlation. (b) A scatter plot showing data with a negative correlation. (c) A scatter plot showing data with zero correlation.

The 95% Critical Values of the Sample Correlation Coefficient Table at the end of this chapter (before the Summary) may be used to give you a good idea of whether the computed value of r is **significant or not**. Compare r to the appropriate critical value in the table. If r is significant, then you may want to use the line for prediction.

Example 1

Suppose you computed $r = 0.801$ using $n = 10$ data points. $df = n - 2 = 10 - 2 = 8$. The critical values associated with $df = 8$ are -0.632 and $+0.632$. If $r <$ negative critical value or $r >$ positive critical value, then r is significant. Since $r = 0.801$ and $0.801 > 0.632$, r is significant and the line may be used for prediction. If you view this example on a number line, it will help you.

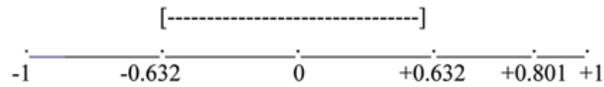


Figure 2: r is not significant between -0.632 and $+0.632$. $r = 0.801 > +0.632$. Therefore, r is significant.

Example 2

Suppose you computed $r = -0.624$ with 14 data points. $df = 14 - 2 = 12$. The critical values are -0.532 and 0.532 . Since $-0.624 < -0.532$, r is significant and the line may be used for prediction



Figure 3: $r = -0.624 < -0.532$. Therefore, r is significant.

Example 3

Suppose you computed $r = 0.776$ and $n = 6$. $df = 6 - 2 = 4$. The critical values are -0.811 and 0.811 . Since $-0.811 < 0.776 < 0.811$, r is not significant and the line should not be used for prediction.



Figure 4: $-0.811 < r = 0.776 < 0.811$. Therefore, r is not significant.

If $r = -1$ or $r = +1$, then all the data points lie exactly on a straight line.

If the line is significant, then **within the range of the x-values**, the line can be used to predict a y value.

As an illustration, consider the third exam/final exam example. The line of best fit is: $\hat{y} = -173.51 + 4.83x$ with $r = 0.6631$

Can the line be used for prediction? **Given a third exam score (x value), can we successfully predict the final exam score (predicted y value).** Test $r = 0.6631$ with its appropriate critical value.

Using the table with $df = 11 - 2 = 9$, the critical values are -0.602 and $+0.602$. Since $0.6631 > 0.602$, r is significant. **Because r is significant and the scatter plot shows a reasonable linear trend, the line can be used to predict final exam scores.**

Example 4

Suppose you computed the following correlation coefficients. Using the table at the end of the chapter, determine if r is significant and the line of best fit associated with each r can be used to predict a y value. If it helps, draw a number line.

- $r = -0.567$ and the sample size, n , is 19. The $df = n - 2 = 17$. The critical value is -0.456 . $-0.567 < -0.456$ so r is significant.
- $r = 0.708$ and the sample size, n , is 9. The $df = n - 2 = 7$. The critical value is 0.666 . $0.708 > 0.666$ so r is significant.
- $r = 0.134$ and the sample size, n , is 14. The $df = 14 - 2 = 12$. The critical value is 0.532 . 0.134 is between -0.532 and 0.532 so r is not significant.
- $r = 0$ and the sample size, n , is 5. No matter what the dfs are, $r = 0$ is between the two critical values so r is not significant.