

CONTINUATION OF INFERENTIAL STATISTICS

Simple Linear Correlation and Regression

Relationship & Prediction

Page | 1



CHICKEN OR THE EGG (WHICH COMES FIRST?)



Karl Pearson 1857- 1936

When in doubt, use the [Statistics Glossary](#)

<http://www.stats.gla.ac.uk/steps/glossary/>

OBJECTIVES

By now you would have observed that the selection of an appropriate statistical test depends on the objective of the analysis, the number of variables involved, and the type (s) of data.

In this module we will continue with bivariate statistics, and explore the relationship between two quantitative or numerical variables, and determine the extent to which one variable (y) can be predicted from the other (x). Specifically, we will address the following:

- Simple Linear Correlation (¹Pearson's)
- Simple Linear Regression
- Scatter Plots

¹ Another commonly used correlation analysis is Spearman's (a non-parametric test), which is appropriate for ranked (or ordinal) data, and when a normal distribution cannot be assumed.



RECOMMENDED READING

[Correlation, Regression, and Causation:](http://www.bmj.com/cgi/content/full/315/7105/422)
<http://www.bmj.com/cgi/content/full/315/7105/422>

INTRODUCTION

When data are collected in a paired fashion, in other words, when we have a data set with two quantitative or numerical values (X and Y) for each subject, we can perform simple linear correlation analysis to determine if the two variables are significantly correlated, related or associated. This is a very useful test to determine the magnitude (**STRENGTH**) and **DIRECTION** of a relationship between two variables.

Simple linear **correlation** and **regression** methods can generally be applied to the same type of data, and complement each other in providing us with a more comprehensive understanding of the pattern(s) underlying our data. Both variables must be truly numerical².

CAUTION: Linear correlation examines or test for an underlying **LINEAR REALTIONSHIP** only, therefore the absence of a linear relationship, does not mean the absence of a relationship, as there could be an underlying non-linear (e.g. curvilinear) relationship, which will require non-linear statistical methods to detect and describe.

Correlation *does not equal* causation. Potential confounding factors must always be considered.

² There are other types of correlation and regression, not covered in this course.



● In correlation the emphasis is on the degree to which a linear model can describe the relationship between two variables, while in regression, the emphasis is on predicting one variable from the other.

● In regression the interest is directional, as in a regression of Y (the dependent variable) on x (the independent). One variable is predicted (y) and the other is the predictor (x).

● In correlation the focus is on the relationship. Correlation is said to be symmetric, that is, the variables correlate with each other in both directions. **In other words, the correlation of X with Y , is the same as the correlation of Y with X**



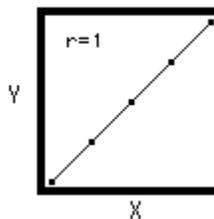
Simple Linear Correlation

The correlation between two variables reflects the degree to which the variables are related. The most common measure of correlation is the Pearson's product moment correlation (called Pearson's correlation for short).

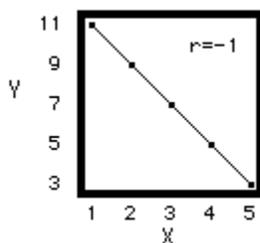
For a population, the Pearson's product moment correlation is represented by the Greek letter rho (ρ), and for a sample, it is designated by the letter "r" (called "Pearson's r"), which is the correlation coefficient.

Pearson's correlation coefficient is a measure of the strength and direction of a linear relationship. It ranges from 0 to 1 (either positive or negative).

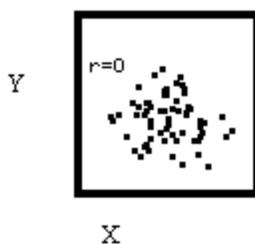
A correlation of +1 means that there is a perfect positive linear relationship between the variables. The scatter-plot below depicts such a relationship. It is a positive relationship (positive slope or gradient) because as one variable increases so does the other. Higher values on one variable are associated with higher values on the other.



A correlation of -1 means that there is a perfect negative (or inverse) linear relationship between the variables. The scatter-plot below depicts a negative or inverse relationship. It is a negative relationship because as one variable increases, the other decreases. Higher values on one variable are associated with lower values on the other.



A correlation of 0 means there is no linear relationship between the two variables. The scatter-plot below depicts a Pearson's correlation of about 0 .



Real-world data will almost never give rise to correlation coefficients³ of 0 , 1 , or -1 . The degree of linear relationship is generally described as⁴:

Correlation Coefficient Value	Direction and Strength of Correlation
-1.0	Perfectly negative
-0.8	Strongly negative
-0.5	Moderately negative
-0.2	Weakly negative
0.0	No association
$+0.2$	Weakly positive
$+0.5$	Moderately positive
$+0.8$	Strongly positive
$+1.0$	Perfectly positive

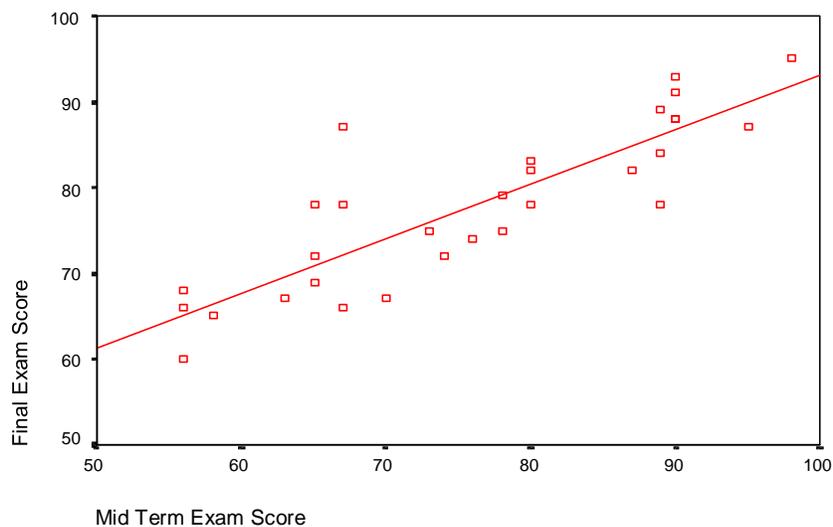
Note.—The sign of the correlation coefficient (ie, positive or negative) defines the direction of the relationship. The absolute value indicates the strength of the correlation.

³ Please refer to the notes/literature on Effect Size.

⁴ Zou, K. H., Tuncali, K., & Silverman, S. G. (2003). Correlation and simple linear regression. *Radiology*, 227(3), 617-628.



● The scatter-plot below depicts a strong positive relationship ($r = 0.86$) between mid-term exam score and final exam score. A plausible interpretation is that as mid-term score increases so does final exam score. In other words, higher mid-term exam scores are associated with higher final exam scores.



$r = 0.86$

● **CAUTION**

Note that negative and positive (in this context) simply refer to the direction of the relationship, and in no way refer to any judgment (bad/good or desirable/undesirable) about the outcome/relationship.



VERY IMPORTANT

OKAY: While you will almost always use a statistical package (such as SPSS) for statistical analysis in the real world, let's take a look at one version of the formula for calculation of the Pearson's correlation coefficient (r).

VERY IMPORTANT

Please refer to the separate PowerPoint presentation for the formula and explanation of the calculations. This will be required for the assignment.

After which, hopefully, you will say:

I ♥
STATS

Let's continue with a demonstration based on SPSS:

● **NOTE: THE NULL HYPOTHESIS FOR PEARSON'S CORRELATION IS:**



There is no statistically significant relationship between the two variables (name them). **That is, $r = 0$.**

● For this exercise we will focus on the variables “self-esteem” and “optimism” (see data below).

● For Pearson's correlation, we will explore the relationship (correlation) between these two variables, and for simple linear regression, we will explore the extent to which “self-esteem” (x) predicts “optimism” (y).

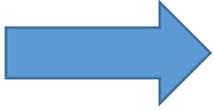
NOTE: In order to meaningfully perform statistical analysis and write a report, you must be conversant with the variables or constructs being measured. In this case, the two constructs are self-esteem and optimism. Both are constructs as they are multidimensional in nature, and each requires the use of a scale for valid and reliable measurement. A scale is a set of items intended to measure an underlying concept such as self-esteem or optimism. Self-esteem is defined as one's perception of self-worth, whereas, optimism is the extent to which a person perceives a positive outlook on life.

SO LET'S DO PEARSON'S CORRELATION FIRST.

 We will use the data below on self-esteem and optimism.

These data were obtained from a convenience sample of 20 clients diagnosed with a substance used disorder.

Self-esteem score	Optimism score
6.0	5.0
9.0	9.0
14.0	10.0
7.0	9.0
7.0	8.0
12.0	9.0
10.0	8.0
8.0	7.0
8.0	8.0
9.0	8.0
6.0	5.0
9.0	8.0
10.0	7.0
10.0	8.0
8.0	6.0
10.0	9.0
7.0	8.0
14.0	10.0
7.0	9.0
6.0	6.0



REMEMBER TO FOLLOW THESE STEPS WHEN ORGANIZING YOUR THOUGHTS AND PERFORMING INFERENCE STATISTICS.

NOTE: You are required to follow these steps for part 2 of the assignment.

Page | 10

1. Clearly state the variables (independent and dependent, and the type of data) – for research purposes, you should orient your variables as to independent (self-esteem) and dependent (optimism), but for the actual analysis this does not matter. Note: This applies to Pearson's correlation only, as it is symmetric.
2. Write the research objective: The objective of this analysis is to determine if there is a statistically significant relationship between self-esteem and optimism scores.
3. Write the null hypothesis: There is no statistically significant relationship between self-esteem and optimism scores. That is, $r = 0$
4. Write the alternative/research hypothesis: There is a statistically significant relationship between self-esteem and optimism scores. That is, $r \neq 0$
5. State the alpha level: Either .05 or .01, but generally .05 for the social and behavioral sciences.
6. Select an appropriate statistical test and provide a justification for its use: Pearson's correlation is appropriate because we are exploring for possible linear relationship between two numerical variables.
7. Conduct/perform the analysis with SPSS (see procedure detailed below, page 11)
8. Interpret the output, decide whether to reject or accept the null hypothesis, and write a brief practical conclusion (see CONCLUSION below, page 13).

SPSS Procedure

Performing linear correlation analysis with SPSS

While in SPSS, open the data set, then click on **Analyze** on the menu bar, and then choose **Correlate**. From the resulting menu, choose **Bivariate** You will see a dialog box. Choose the two variables (**self-esteem** and **optimism**) for the analysis by moving them from the box on the left to the box on the right (under “**Variables**”). Ensure that the box next to “**Pearson**” is checked, then click **OK** to run the analysis.

INTERPRETING THE PEARSON’S CORRELATION OUTPUT

		esteem	Optimism
Esteem	Pearson Correlation	1.000	.687**
	Sig. (2-tailed)		.001
	N	20	20
Optimism	Pearson Correlation	.687**	1.000
	Sig. (2-tailed)	.001	
	N	20	20

** . Correlation is significant at the 0.01 level (2-tailed).



● A convenient way of summarizing the correlation analysis is to organize the information as a **correlation matrix** (see table above, page 11).

● In each cell of the correlation matrix, there are three (3) items of information.

● The top number is the **Pearson's correlation** coefficient (which is .687 in this case), the number below this is the **p-value or level of significance** for the correlation (see **Sig. (2-tailed)** which is .001 in this case), and the bottom number is the sample size (**N**) on which the correlation is based.

● Note the redundancy in the table. Each correlation appears twice in the matrix. Also the **Pearson's correlation coefficient of 1**, represents the correlation of each variable with itself.

OKAY: So let's now decide whether to reject or accept the NULL HYPOTHESIS, and comment on the strength and direction of the relationship.

NOTE:

● If the p-value or level of significance – **Sig. (2-tailed)** is less than .05, we can conclude that the correlation coefficient is significantly different from zero (that is the null hypothesis is rejected). In other words, the evidence supports the alternative hypothesis that there is a statistically significant relationship between the two variables.

● Sometimes the p-value (level of significance) is shown as .000 and should be reported as .001 (we must always allow for some error in inferential statistics).

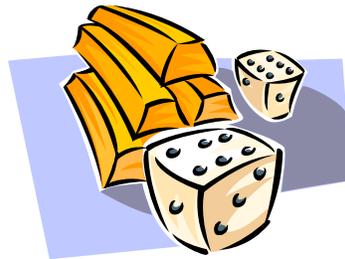


CONCLUSION: In this case the level of significance is .001, which is less than the alpha level (.05), and the Pearson's correlation coefficient (r) is .687. We will therefore reject the null hypothesis and conclude that there is a statistically significant moderate positive relationship between self-esteem and optimism scores. That is, as self-esteem score increases so does optimism score. In other words, higher self-esteem scores are associated with higher optimism scores (and vice versa).

I ♥
STATS



NOTE: Next we will look at Simple Linear Regression



- As stated above, in regression the emphasis is on predicting one variable from the other.
- **Regression is directional - one variable is predicted, and the other is the predictor.**
- Linear regression assumes⁵ that the relationship between the two variables can be represented by a straight line (hence linear).

⁵ See other assumptions required for linear regression.

● However, most health and behavioral phenomena are complex and non-linear, and will require other, more sophisticated statistical techniques/models. Nonetheless, the concepts of regression and correlation can be understood from this exercise.



There are different kinds of lines that might be fitted to a particular data set. The most basic regression line is straight, representing a linear relationship between two variables.



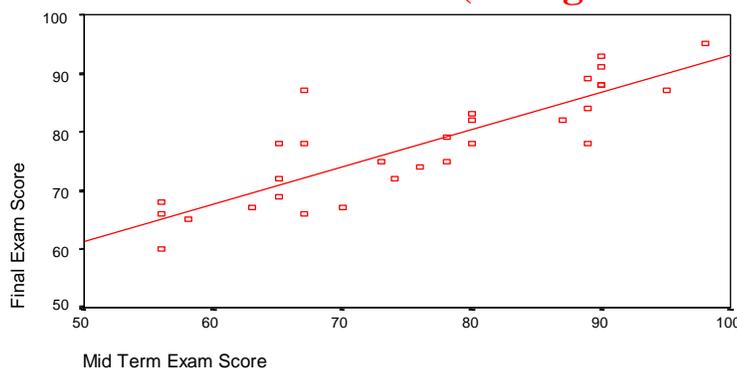
The way to obtain a regression line is to plot the data on the x-y axes, and plot a line through the set of points (scatter-plot) so as to minimize the distance of each point from the line.



Technically speaking, this minimizes the squared residuals (the reason it is also referred to as the “**least squares line**”).



Rarely, if ever, will all the points fall on the line unless there is a perfect relationship between the two variables, which is highly unlikely with real-world data. The resulting line is called the line of "best fit" (see figure below).



$$r = 0.86$$



The simple linear regression line (or model) can be described by the equation⁶:

$Y = bx + a$, where:

Y is the dependent/outcome/criterion variable (plotted on the vertical axis).

X is the independent/explanatory/predictor variable (plotted on the horizontal axis).

b is the slope of the line, and

a is the y-intercept (the point where the line intercepts the y-axis, that is where $x = 0$).

The simple linear regression equation expresses the relationship between **two variables (one independent and one dependent)** algebraically.

- The magnitude of the slope (b) indicates the change in Y that results from every unit change in X.
- In other words, for every increase of 1 unit in X there is a change of b units in Y.
- For example, with a slope of 3, the y value goes up by 3 units every time the x value goes up by 1 unit
- Similarly, for a slope of -3, the y value goes down by 3 units every time the x value goes up by 1.
- With the slope and y-intercept, the formula for the line is completed, and we can proceed to use that formula to make predictions based on the regression model.
- For example, if we calculate that the "best fit" line has a slope of 3 and a y-intercept of 10, then the equation $Y = 3X + 10$ describes that line.
- If we want to know the value of Y when X is 5, we simply plug 5 into the equation and solve for Y, thus: $Y = 3(5) + 10 = 25$. This is our "predicted" value for Y, which means that it's the value for Y on the line at the point where X is 5.

⁶ Multiple regression analysis is used to test two or more independent variables (X) as predictors of the dependent variable (Y).



● To find out how much of the variance in Y is accounted for (or explained) by X, simply square the r value (Pearson's correlation coefficient), resulting in what is called **R-Squared** (R^2), or the **COEFFICIENT OF DETERMINATION**.

● For example, if we had a Pearson's r of 0.8, then the R^2 value will be **$r \times r$** ($.8 \times .8$) = .64 or 64%.

● This means that 64% of the variance in Y (the outcome or dependent variable) is accounted for (or explained) by the X (the independent or predictor variable).

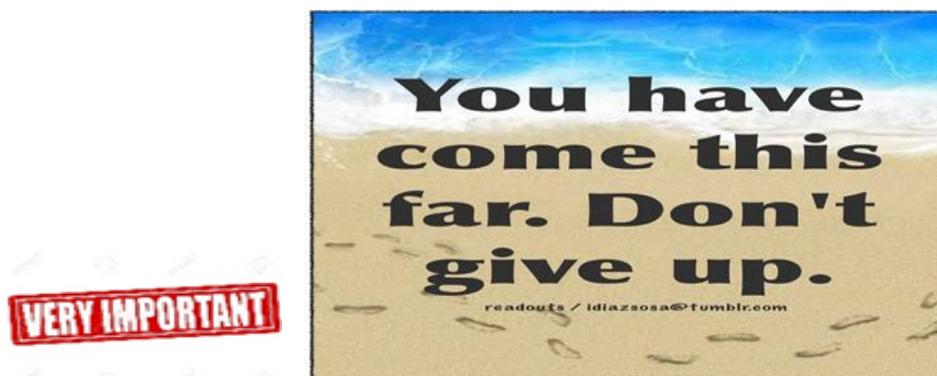
● Of course, since the total variance = 100% (or 1), an R^2 of 64% (.64) indicates that approximately 36% (.36) of the variance in Y is not accounted for (or explained) by X.

$$1 - R^2$$

$$= 1 - .64 = .36 \text{ or } 36\%$$

● The unexplained variance is also referred to as the **residual or error variance**.

● R^2 can be viewed as the **predictive value or level of effectiveness of the regression model**.



OKAY: While you will almost always use a statistical package (such as SPSS) for statistical analysis in the real world, let's take a look at one version of the formula for calculating the terms of the simple linear regression model.

Please refer to the separate PowerPoint presentation for the formula and explanation of the calculations.

NOTE: Regression analysis (neither manual nor SPSS) is required for the assignment.