Correlation Analysis

Many studies use correlation analysis to explore the degree association between study variables. Especially in social science research, linear correlation analysis is a tool for representing the closeness of one related variable to another. The linear correlation coefficient, denoted by ρ (*Greek letter rho*), is a measure of the strength of the linear relationship existing between two variables, say X and Y, that is independent of their respective scales of measurement.

Linear Correlation: Meaning and Coefficient

Correlation is meant for exploring the degree of relationship between two variables in consideration. Correlation Coefficient is the measure to quantify such degree of relationship of the variables. Generally, two correlation coefficient are used in applications, namely: ***Pearson's Product Moment Correlation Coefficient*** and ***Spearman's Rank Correlation Coefficient***.

The linear correlation Coefficient possesses the following interesting properties:

1. A linear correlation coefficient can only assume values between $-1 \leq \rho \leq 1$.
2. The sign of ρ describes the direction of the linear relationship between X and Y. A positive value for ρ means that the line slopes upward to the right, and so as X increases, the value of Y increases. On the other hand, a negative value for ρ means that the line slopes downward to the right, and so as X increases, the value of Y decreases.
3. If $\rho = 0$, then there is no linear correlation between X and Y. However, this does not mean a lack of association. It is possible to obtain a zero correlation even if the two variables are related, though their relationship is nonlinear.
4. When ρ is -1 or 1, there is perfect linear relationship between X and Y and all the points (x, y) fall on a straight line. A ρ that is close to 1 or -1 indicates a strong linear relationship.
5. A strong linear relationship does not necessarily imply that X causes Y or Ycauses X. It is possible that a third variables may have caused the change in both X and Y, producing the observed relationship.

Take Note!!!

This is an important point that we should always remember when studying not just relationships, but also comparing two populations, say by using a *t*-test. Unless we collected our data using a well-designated experiment where we able to randomize the treatments and substantially control the extraneous variables, we need to use the more complex "causal" models to study causality. Otherwise, we just describe the observed relationship or the observed difference between means.

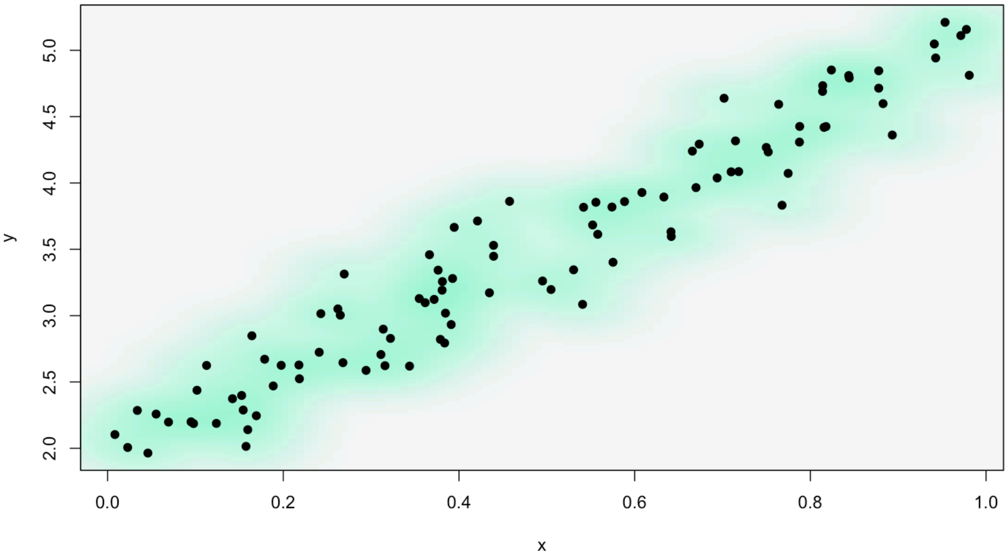

Figure 1: Direct and Strong Correlation ($\rho\ = 0.93$)

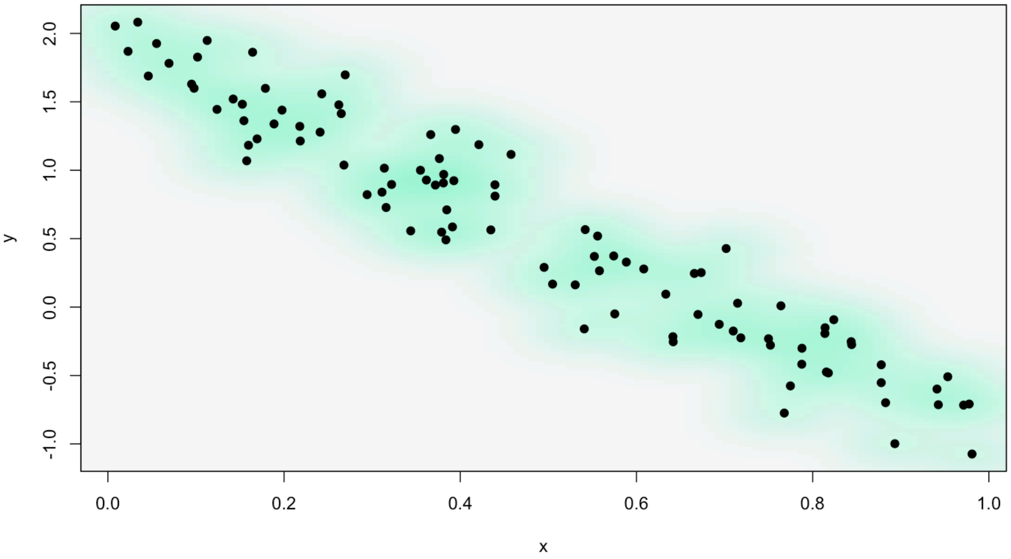

Figure 2: Inverse and Strong Correlation ($\rho\ = -0.93$)

A point estimator of ρ is the Pearson Product Moment Correlation Coefficient, which we denote by (R or r), and for Spearman's Rank Correlation Coefficient we denote it by (R_s or r_s). Their definition are stated in definition 1 and 2. The values of Pearson Product Moment Correlation Coefficient and Spearman's Rank Correlation Coefficient is also between -1 and 1.

Definition 1:

Product Moment Correlation Coefficient (R or r) is a scale to measure the strength of linear association between variables. As it measures the degree of linear association of variables, *interval or ratio* variables should be in consideration with a condition that the variables considered should fall in *normal distribution*.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{\left[n(\sum x_i^2 - (\sum x_i)^2\right]\left[n(\sum y_i^2 - (\sum y_i)^2\right]}}$$

The numerator is the sum of the cross-product (SS_{XY}): $\sum (x_i - \bar{x})(y_i - \bar{y})$

The denominator are the sum of squares for variable X (SS_{XX}): $\sum (x_i - \bar{x})^2$ and the sum of squares for variable Y (SS_{YY}): $\sum (y_i - \bar{y})^2$.

Exercise 1:

Show that

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{\left[n(\sum x_i^2 - (\sum x_i)^2\right]\left[n(\sum y_i^2 - (\sum y_i)^2\right]}}$$

Definition 2:
Spearman's Rank Correlation Coefficient is a measure of association which requires that both variables be measured in at least an ordinal scale so that the objects or individuals under study may be ranked in two ordered series. The assumption of normality is not required in this test.

$$\rho = 1 - \frac{6\sum {d_i}^2}{n(n^2 - 1)}$$

We can also perform a test of hypothesis for ρ. The null hypothesis states that the underlying linear correlation coefficient, ρ, is equal to some hypothesized value, ρ_o. In most applications, the hypothesized value, ρ, is 0. In such case, we may be able to conclude that there is significant linear association between the variables under study.

Table 1: Test of Hypotheses for ρ

Null Hypothesis, H_o	**Alternative Hypothesis, H_a**	**Test Statistic**	**Region of Rejection**
$\rho = \rho_o$	$\rho < \rho_o$	$t = \frac{(r - \rho_o)\sqrt{n-2}}{\sqrt{1-r^2}}$	$t < -t_{\alpha,df=n-2}$
	$\rho > \rho_o$		$t > t_{\alpha,df=n-2}$
	$\rho \neq \rho_o$		$t < -t_{\alpha/2,df=n-2}$

Exercise 2:
A. Fill out the table below by providing the missing information.

No.	**No. Of Hours Studying (X)**	**Final Scores (Y)**	X^2	Y^2	XY
1	2	21			
2	4	27			
3	6	29			
4	8	64			
5	10	86			
6	12	92			
Total:					

No.	No. Of Hours Studying (X)	Final Scores (Y)	$X - \bar{X}$	$Y - \bar{Y}$	$(X - \bar{X})^2$	$(Y - \bar{Y})^2$	$(X - \bar{X})(Y - \bar{Y})$
1	2	21					
2	4	27					
3	6	29					
4	8	64					
5	10	86					
6	12	92					
Total:							

B. Calculate the Pearson's Product Moment Correlation Coefficient and test the null hypothesis that there is no significant relationship between the number of hours students spend studying and their final Statistics score using the data provided.

Exercise 3:

Calculate the Spearman's Rank Correlation Coefficient using the data from Exercise 2 under the assumption that the data for the number of hours and final scores are not normal.

No.	No. Of Hours Studying (X)	Final Scores (Y)	Rank (X)	Rank (Y)	d^2
1	2	21			
2	4	27			
3	6	29			
4	8	64			
5	10	86			
6	12	92			
Total:					

ASSIGNMENT (NOT RECORDED):

Please review Chapter 1 of Applied Linear Regression Model by Neter, Wasserman, and Kutner for information on probability and statistics fundamentals.

Prepared by:

KATRINA D. ELIZON
Department of Mathematics and Statistics
PUP College of Science