University of Guelph
Numeracy Project

About Bivariate Correlations and Linear Regression

UNIVERSITY
of GUELPH

TABLE OF CONTENTS

About Bivariate Correlations and Linear Regression

What is BIVARIATE CORRELATION?

- Bivariate correlation refers to an expression that indicates the direction and magnitude of the relationship between two continuous variables

What is LINEAR REGRESSION?

- Simple linear regression is the analysis of the relationship among two variables for the purpose of understanding how one variable may predict the other variable

Bivariate Correlations

Background

- Correlation and regressions focus on the differences between individual respondents. More specifically, they indicate how one variable is related to the other variable

  We refer to these variables as X and Y:

  > X is the independent or predictor variable
  > Y is the dependent or outcome variable

  Unlike other statistical tests, we are not interested in comparing the means of X and Y. Rather, we want to determine if there is a relationship between the two variables --- and if so, we want to know the nature of the relationship

Three Types of Relationships between Two Variables

1)   Positive relationship
   As one of the variables increases, the other variable increases as well.
   Similarly, as one variable decreases, so does the other. In other
   words, the variables move in the same direction.

2)   Negative relationship
   As one variable increases, the other variable decreases. In other
   words, the variables move in opposite directions.

3)   Zero relationship
   There is no relationship between the two variables

Strength of Bivariate Correlations

●   The strength of a correlation is usually described as strong, moderate,
   weak, or zero

The following depicts positive correlation:

| r=0 | r=.28 | r=.55 | r=.86 |
| zero correlation | weak correlation | moderate correlation | strong correlation |
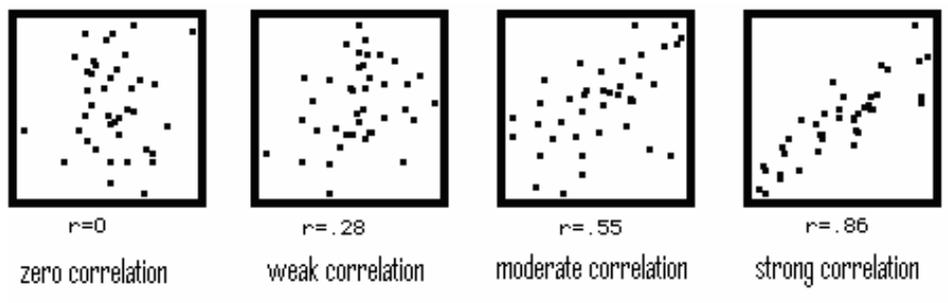
The following depicts negative correlation:

| r=0 | r=-.28 | r=-.55 | r=-.85 |
| zero correlation | weak correlation | moderate correlation | strong correlation |
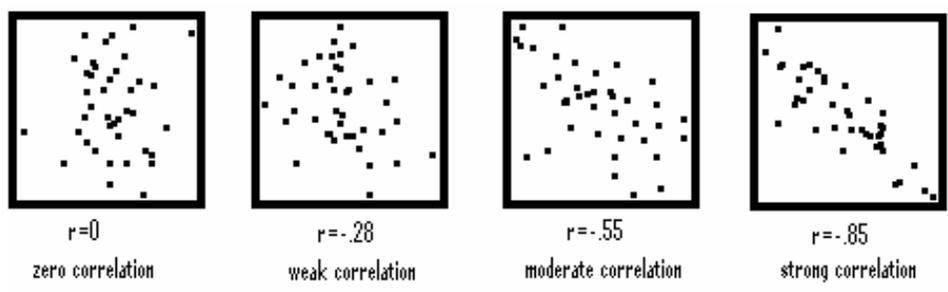
Types of Bivariate Correlations

Pearson's Product Moment Correlation

●   Pearson's product moment correlation coefficient --- represented by r
    --- is the most common type of correction

    It is often referred to as:

          Pearson's r   or   correlation coefficient

●   Pearson's r evaluates the possibility that two interval- or ratio-level
    variables are related in a linear way.

●   Pearson's r measures the strength of the relationship between two
    variables. It varies from -1 (a perfect negative relationship) to +1 (a
    perfect positive relationship).

    The formula for calculating Pearson's r is:

$$r \ = \ \frac{N(\sum XY) - (\sum X)(\sum Y)}{\{\ [N(\sum X^2)-(\sum X)^2]\ [N(\sum Y^2) - (\sum Y)^2]\ \}^{1/2}}$$

    where N is the sample size
    $\sum X$ is the sum of the X values
    $\sum Y$ is the sum of the Y values
    $\sum XY$ is the sum of the X values multiplied by the Y Values
    $\sum X^2$ is the sum of the X values squared
    $\sum Y^2$ is the sum of the Y values squared.

●   The coefficient of determination, R, indicates what proportion of Y's
    variance can be explained by X.

●   In the case of two-variable (X and Y) regression, R is the square of
    the Pearson correlation coefficient.

Spearman's rho

● Spearman's rho is an alternative way of deriving a correlation coefficient. It is used when the sample size is small (i.e., less than 30 pairs of measurements) and when both variables are ordinal.

● It differs from Pearson's product-moment correlation only in that the calculations are done after the values have been converted into ranks.

● When converting to ranks, the smallest value of X becomes 1, the 2nd smallest value becomes 2, the third smallest becomes 3, and so on.

● The same formulas are used as in Pearson's r to find the correlation coefficient, coefficient of determination, and regression.

Seven Features of Correlations

1) A correlation of +1 indicates that there is a perfect positive linear relationship between the two measurement variables. This means that as one variable increases, so does the other.

2) A correlation of -1 indicates that there is a perfect negative relationship between the two measurement variables. This means that as one variable increases, the other variable decreases.

3) A correlation of zero indicates that there is no relationship between the two measurement variables. This would be depicted in a scatter plot as a horizontal or vertical line parallel to the x-axis.

4) A positive relationship indicates that both variables increase or decrease together

5) A negative relationship indicates that as one variable increases, the other decreases.

6) Correlations are unaffected if the units of measurement are changed.

For instance, the correlation between height and weight would be the same regardless if height was measured in centimetres or feet, or if weight was measured in pounds or kilograms.

7)   Correlations do not indicate causation.

From a correlation, we can only determine the direction and strength of the relationship. We cannot say that changes in one variable cause changes in the other variable. The reason is that there are many uncontrolled extraneous factors that could be affecting the variables of interest.

Linear Regression

●   Simple linear regression is the analysis of the relationship among two variables for the purpose of understanding how one variable may predict the other.

●   Regression is defined by an equation that describes how changes in one variable (Y) occur as a result of changes in another variable (X).

●   This equation also allows us to predict what changes will occur in variable Y as a result of changes in variable X. The regression equation is:

$$Y' = a + bX$$

Y' is the predicted value of Y given the value of X
a is the Y intercept or constant (i.e. the value of Y when X=0)
b is the regression coefficient or the slope of the straight line that best fits the data.

●   How do we find these values?

1)   The equation to determine the value of b is:

$$b = \frac{\Sigma(X - X)(Y\text{-}Y)}{\Sigma(X - X)}$$

2)   To determine the value of a use the regression equation:
$$a = Y - bX$$

- The regression equation is useful in that it allows us to predict the value of the Y variable from a value of the X variable.

- The regression equation provides an approximate prediction of the change in the outcome variable due to changes in the predictor variable. Therefore, the exact number is not necessarily given and this represents a degree of error.

- Error could occur because the relationship is not a perfect correlation, meaning that there are extraneous variables that account for some changes in the outcome variable.

Glossary

Correlation: an expression that indicates the direction and magnitude of the relationship between continuous variables.

Negative relationship: as one measurement variable increases, the other increases.

Pearson's product-moment correlation:

the most commonly used correlation procedure, which evaluates the possibility that two interval- or ratio-level variables are related in a linear way

Positive relationship: as one measurement variable increases, so does the other.

r: Pearson's correlation coefficient; it tells us the strength and the direction of the relationship.

R^2: coefficient of determination, indicates what proportion of y's variance can be explained by x

Regression: analysis of relationships among variables for the purpose of understanding how one variable may predict another.

Regression equation: allows us to predict the value of the Y variable from a value of the X variable. Represented as $Y' = a + bX$

Regression line: a straight line that comes as close to the greatest number of points on a scatter-plot as possible

Spearman's rho: alternative way of deriving a correlation coefficient, which is often used when there is a small sample size (i.e. less than 30 pairs of measurements) and especially when both variables are ordinal.

Zero relationship: there is no relationship between the two variables.

References

Cohen, R. J., & Swerdlik, M. E. (2005). Psychological testing and
 assessment: An introduction to tests and measurements. (6th ed.) NY:
 McGraw-Hill.

Gardner, R. C. (2001). Psychological statistics using SPSS for
 Windows. New Jersey: Prentice Hall.

Hamilton, L. C. (1996). Data analysis for social scientists.
 Scarborough: Duxbury Press.

Ottawa Senators. (2007). Stats and Standings. Retrieved on May
 10, 2007 from http://www2.ottawasenators.com/eng/stats/
 redirect.cfm?sectionID=hockey/hockeyGameResults.cfm

Utts, J. (2005). Seeing through the statistics. (3rd ed.). Toronto:
 Thomson Brooks/Cole.