Bivariate Correlation and Regression

Two variables are functionally related when variations in the values of one are systematically associated with variations in the values of the other.

The word "***correlation***" is often used as a vague synonym for "***association***".

While a numerical value can be derived to show the amount of the correlation between two variables, this is more aptly done by plotting the two variables in a scatterplot.

When constructing scatterplots it is important to remember the direction of the association … often we know that one variable influences another and not vice versa.

Relationships between variables have both ***direction*** and ***strength***.

- Direction – positive, negative, or indeterminate
- Strength – the magnitude of the association

Correlation Analysis:

Involves assessing the strength of the relationship between two interval or ratio variables.

Areal association between any two variables can be assessed numerically through the use of the:

Pearson product-moment coefficient of correlation, *r*

Properties of ***r***, include:

1. It is a bound measure that falls between -1 and 1.

    r > 1 indicates positive association
    r < 1 indicates negative association

2. Extreme values of ***r***, near -1 or 1 occur only in the case of a perfect linear association.

3. The values of ***r*** does not change when the unit of measurement of x, y or both change.

4. ***r*** is dimensionless

5. Correlation only measures the strength of the linear association between two variables … does not account for curvilinear relationships.

Correlations are graphically portrayed in ***scatterplots*** and tabled in ***matrices***.

The correlation coefficient is given by:

$$r = \frac{COV_{1,2}}{s_1 \ s_2}$$

Or the ratio of the ***covariance*** … the extent to which y and x vary together about their common means … to the product of the standard deviations of y and x.

When conducting a correlation analysis, there is an inherent hypothesis that is set and tested. The goal is to ascertain whether the magnitude of ***r*** is large enough to be significant.

$$H_o : \rho = 0$$
$$H_a : \rho \neq 0$$

Just like most inferences, this is also tested though a *t-test* and the values are normally given in the correlation matrix table.

Correlation and Causation:

"Measures of strength of association should not be thought of as something that proves ***causation***, but only as a measure of covariation in the measurements."

1. Variation in either characteristic may be caused (directly or indirectly) by variation in the other – x may cause y, but y does not cause x … ***independent*** vs. ***dependent*** variables. Knowledge of the system you are studying will suggest the direction of the cause and effect.

2. Covariation of the two variables may be due to a common cause or causes affecting each variable in the same way, or in opposite ways. Example: days of sunshine and per capita federal income tax levels.

3. The causal relationship between the two characteristics may be a result of inter-dependent relationships. Example: some other variable has an affect on both x and y.

4. The association may be due to chance alone. Example: sample versus population … relationship may be spurious.

5. The strength of the association may be affected by one or two extreme values of either variable … ***outliers***.

6. The strength of the association may be affected by a lack of independence among the observations. Mostly results from spatial and temporal ***autocorrelation*** (i.e., one event in space or time is related to the other). Example: price of one home affects another.

7. An erroneous (non-linear) functional form is specified. May need to transform the data, re-specify the form, or run some other non-parametric statistic.

Linear Regression

The correlation coefficient estimates the degree of closeness of the linear relationship between two variables.

Often, however, the most interesting questions about these variables are:

- **How much does one variable change for a given change in the other?**
- **How accurately can the value of one variable be predicted from the knowledge of the other?**

These questions can be answered with the aid of **regression analysis**.

Distinctions between variables arise in regression analysis, with a **dependent variable, Y**, and an **independent variable, X.**

The purpose of regression is to suggest a possible explanation of the variation in the dependent variable, **Y**, by demonstrating a systematic covariation in some logically related variable, **X**.

Does not have to be causal.

Also does not imply **X** is the only thing influencing **Y**.

A functional relationship is linear when pairs of **X** and **Y** values fall into a pattern that is best described by the linear algebraic model:

$$\hat{Y} = a + bX_i$$

where ***a*** is the intercept and ***b*** is the slope of the line, and **Y** is the expected value of **Y** (called Y hat) for a given value of **X**. This is normally called the **estimating equation**.

The values of ***a*** and ***b*** are estimates, and when the values of **X** are substituted into the equation, the solution of the equation provides estimates of **Y** for given values of **X**.

The proximity of the actual values of **Y** to a straight line derived through the regression equation provides an indication of the strength of the relationship, similar to what is observed in correlation.

Ex:

A perfect linear relationship between distance and time.

The regression equation would therefore describe the relationship between the distance (**Y**) and the time (**X**) and object travels.

The regression coefficient, ***b***, is **the rate of change constant** and represents the **slope** of the linear function between the two variables.

In this example, ***b*** is equal to one. Or in other words, the distance increases one mile per hour for each minute increase in time.

The constant ***a***, is the value of **Y** when **X** is equal to zero.

The constant is called the **Y-intercept**, and represents the point at which the linear function passes through the **Y-axis**.

In this example when time is equal to 0, distance also equals 0.

Ex:

Each of the four regression lines has different values for ***a*** and ***b***.

The four different values of **a** are the result of four different intersections of the lines with **Y** axes.

The different values of ***b***, reflect the steepness of their individual slopes. The higher the value of ***b***, the steeper the slope. Remember that ***b*** indicates the rate of change in the dependent variable (**Y**) given a one unit change in **X**.

Finally, the sign of ***b*** expresses the direction of the relationship between **Y** and **X**; when ***b*** is positive, an increase in **X** is accompanied by an increase in **Y**; when ***b*** is negative, **Y** decreases as **X** increases.

Ordinarily a regression analysis begins with a scatterplot.

As an example look at the values of mean annual runoff as a function of mean annual precipitation. The pattern in the scatterplot shows a linear relationship but it is not perfect around the equation line. This demonstrates that there is some variation from the expected perfect linear relationship.

The linear model obtained through regression is the line that produces the **minimum sum of squares of error** and is called **the least squares method**.

This done by calculating the parameters, ***a*** and ***b***, by attempting to minimize the error about the regression line:

$$\sum_{i=1}^{n} (Y_i - \hat{Y})^2$$

How good does a regression line fit the data?

In virtually all research you should expect some deviations of your data about the regression line. In addition, there are usually several, often interrelated, causes for any observable effect (Ex: yields of a crop are simply a function of rainfall).

Therefore, you should expect variations around some simple regression line to be similar to the deviations around the mean frequency of a distribution.

If every place were identical with respect to some phenomenon, **Y**, regardless of their values of **X**, there would be no deviations, and every observation would equal the mean (**Y**).

On the other hand, if the correlation between **Y** and **X** is ± 1.0 then all paired values would fall on the regression line. Then all variation in **Y** would be explained by the variation in **X** and there would be no residual variation.

In the real world, the components of variation are:

The components of variation can be broken down to that due to regression (RSS):

$$\text{Regression Sum of Squares} = \sum (\hat{Y}_i - \bar{Y})^2$$

and that due to error, or residual sum of squares- or the variation not explained by the regression (ESS):

$$\text{Residual sum of squares} = \sum (Y_i - \hat{Y})^2$$

So the **total variation** (TSS) between two variables can be summarized as that part to due to regression (RSS) and that due to error (ESS).

The ratio of the explained variation (RSS) to the total variation is referred to as **the coefficient of determination** and is symbolized as $\mathbf{R^2}$.

This statistic, may also be computed by squaring the correlation coefficient, **r.**

$\mathbf{R^2}$ is a measure of the proportion of the total variation in the data that is accounted for by the regression.

When $\mathbf{R^2}$ = 1.0 it is a perfect linear relationship.

<u>Multiple Regression</u>:

A simple extension of simple linear regression where several explainatory variables are used to help explain or predict a single response variable.

Multiple regression is not represented by an individual regression equation, but by a model with multiple weighted variables:

$$\hat{Y} = a + b_1 X_1 + b_2 X_2 + b_3 X_3 + \ldots + b_n X_n$$

Statistical Assumptions in Linear Models or Equations

1) No specification error or linearity

a) relationship between X and Y is linear
b) no relevant IVs have been excluded
c) no relevant IVs have been included

Or the theoretical model in the equation is correct. Look for curvilinear relationships and transform or include other variables.

2) No measurement error

a) variables X and Y are measured correctly
b) variables X and Y are coded correctly

3) Error or residual term assumptions