Linear Regression

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable (or regressor), and the other is considered to be a dependent variable. For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model.

Before attempting to fit a linear model to observed data, a modeler should first determine whether or not there is a relationship between the variables of interest. This does not necessarily imply that one variable causes the other (for example, higher SAT scores do not cause higher college grades), but that there is some significant association between the two variables. A scatterplot can be a helpful tool in determining the strength of the relationship between two variables. If there appears to be no association between the proposed explanatory and dependent variables (i.e., the scatterplot does not indicate any increasing or decreasing trends), then fitting a linear regression model to the data probably will not provide a useful model. A valuable numerical measure of association between two variables is the correlation coefficient, which is a value between -1 and 1 indicating the strength of the association of the observed data for the two variables.

A linear regression model postulates the following equation for the relationship between X and Y:

$Y_i = a + bX_i + u_i$,

where the subscript i runs over observations (i=1, ….n), X is the explanatory variable (regressor) and Y is the dependent variable. The slope of the line is b, and a is the intercept (the value of y when x = 0). We call "a" and "b" coefficients or parameters.

The first part of the equation (a+b X) is the relationship that holds between X and Y on average over the population (this is called the population regression line). The term u_i is the error term that incorporates all the other factors besides X that determine the value of the dependent variable Y.

When we add more explanatory variables (or regressors) we have a multiple regression.

Least-Squares Regression

The most common method for fitting a regression line is the method of least-squares. This method calculates the best-fitting line for the observed data by minimizing the sum of the squares of the vertical deviations from each data point to the line (if a point lies on the fitted line exactly, then its vertical deviation is 0). Because the deviations are first squared, then summed, there are no cancellations between positive and negative values.

Several assumptions need to be satisfied so that the estimated coefficient have the appropriate properties . An important property is that the error term is not correlated with the regressors.

Significance Tests

As in linear regression, one wishes to test the significance of the parameters included. For any of the variables x_j included in a multiple regression model, the null hypothesis states that the coefficient (b_j) is equal to 0. The alternative hypothesis may be one-sided or two-sided, stating that b_j is either less than 0, greater than 0, or simply not equal to 0.

The test statistic t is equal to b_j/s_{bj}, the parameter estimate divided by its standard deviation. This value follows a *"t* distribution "

Instrumental variables regression (IV)

can be performed when the regressors are correlated with the errors. In this case, we need the existence of some auxiliary *instrumental variables* $\mathbf{z}_i$ such that is not correlated to the errors and is correlated to the regressor we need to "instrument" (the one that presented the problem).

Sources:

J. Stock and M. Watson: "Introduction to Econometrics," Addison Wesley, 2003.

http://en.wikipedia.org/wiki/Linear_regression

http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm