

Answer on Question #41402 – Math - Statistics and Probability

An economist want to estimate a multiple regression equation in which the (3) amount saved by the i th family depends on the family's income, whether the family is graduate or non graduate, and whether the family is headed by a male or female. Explain how a regression equation of this sort can be estimated. What is the dependent variable ? What are the independent variables? What assumption must be made ?

Solution

$$y = ax + bz + c + \varepsilon.$$

There are n observations:

y_i is i -th family's income (a random variable that can take any real value), $i=1, \dots, n$,

x_i is education status (a random variable that can take either graduate or non graduate) , $i=1, \dots, n$,

z_i is family status (a random variable that can take either male or female which is the head of the family) , $i=1, \dots, n$.

Besides,

ε_i is an error, c is intercept, $i=1, \dots, n$.

Education status and family status are independent variables, in this problem they are categorical variables. Family's income is dependent variable. To use categorical variables in regression analysis, we convert them in dummy variables to code them as 0 or 1, for example, $x_i = 1$ if graduate and $x_i = 0$ if non-graduate. You can research model for case $x_i = 1$.

Leave one dummy variable out from each categorical variable. Do not use both the original categorical variable and the dummy variable. Using categorical variables, we should discuss whether interactions are present (check multicollinearity: if it is present then you can get rid of one of offenders) and heteroscedasity of the error variances. These cases will come to problems in further analysis.

