**Chapter 14: Multiple Regression**

A regression model that includes two or more independent variables is called multiple regression model. The population model is given by the equation:

*y = A + B*1*x*1 + *B*2*x*2 + *B*3*x*3*+ … + Bkx*k *+ ε*

If all the x variables are raised to the power of one then the model becomes a first-order multiple regression model. If a model has k independent variables and a sample size of n, then the degrees of freedom is *df = n -* *k –* 1

**Multiple Regression Model Assumptions**

Assumption 1: The mean of the probability distribution of εis zero, that is ***E*(ε) = 0**

Assumption 2: The errors associated with different sets of values of independent variables are independent. Furthermore, these errors are normally distributed and have a constant standard deviation, which is denoted by σε

Assumption 3: The independent variables are not linearly related. However, they can have a nonlinear relationship. When independent variables are highly linearly correlated, it is referred to as ***multicollinearity***.

Assumption 4: There is no linear association between the random error term ε and each independent variable *xi*

**Standard Deviation of Sample Errors:**

Population standard deviation of errors is denoted by the symbol σε. The sample standard deviation of errors is denoted by the symbol *s*e.

This value can be generated using Stata. In the results, *s*e is usually termed Root MSE, where MSE means Mean Square Error.

**Coefficient of Multiple Determination**

Similar to Simple Regression models, the coefficient of determination in a Multiple Regression Model, $R^{2}$, measures the proportion of Total Sum of Squares (SST) that is explained by the regression model. It measures how well the independent variables explain the dependent variable. Recall that SSR is the part of SST that is explained by the regression model. (SSE is the part of SST that is not explained by the regression model).

In multiple regression models, the value of $R^{2}$ keeps increasing with the addition of more independent variables, even if the additional variables do not have significant influence on the dependent variable. However, increasing values of $R^{2}$ misleads us to think that the model is improving with the additional independent variable.

To overcome this short-coming of $R^{2}$ we use an alternative coefficient of determination measure, the Adjusted $R^{2}$ or $\overbar{R}^{2}$. The formula is

One thing to note is that while $R^{2}$ can only be a value between 0 and 1, $\overbar{R}^{2}$ can be negative.

Interpreting Regression Results in STATA:

**Example 1: Simple Linear Regression Model**

Last year five randomly selected students took a match aptitude test before they began their statistics course. Their results are presented in the table below:

|  |  |  |
| --- | --- | --- |
| Student Serial | Aptitude Test Score  | Grade in Statistics class |
| 1 | 95 | 85 |
| 2 | 85 | 95 |
| 3 | 80 | 70 |
| 4 | 70 | 65 |
| 5 | 60 | 70 |

Q1. Present descriptive statistics of the dependent and the independent variables.

. sum statistics

 Variable | Obs Mean Std. Dev. Min Max

-------------+--------------------------------------------------------

 statistics | 5 77 12.5499 65 95

. sum aptitude

 Variable | Obs Mean Std. Dev. Min Max

-------------+--------------------------------------------------------

 aptitude | 5 78 13.50926 60 95

Q2. What linear regression equation best predicts statistics performance, based on math aptitude scores?

. reg statistics aptitude

 Source | SS df MS Number of obs = 5

-------------+------------------------------ F( 1, 3) = 2.77

 Model | 302.60274 1 302.60274 Prob > F = 0.1945

 Residual | 327.39726 3 109.13242 R-squared = 0.4803

-------------+------------------------------ Adj R-squared = 0.3071

 Total | 630 4 157.5 Root MSE = 10.447

------------------------------------------------------------------------------

 statistics | Coef. Std. Err. t P>|t| [95% Conf. Interval]

-------------+----------------------------------------------------------------

 aptitude | .6438356 .3866477 1.67 0.194 -.58665 1.874321

 \_cons | 26.78082 30.51824 0.88 0.445 -70.34184 123.9035

------------------------------------------------------------------------------

$$\hat{y}=26.78+0.644aptitude$$

Q3. Fit the linear regression line through a scatter plot of the observations.

. graph twoway (lfit statistics aptitude ) (scatter statistics aptitude)



Q4. If a student made an 80 on the aptitude test, what grade would we expect her to make in statistics?

$$\hat{y}=26.78+0.644aptitude\hat{y}=26.78+0.644(80)\hat{y}=78.3$$

Q5. How well does the regression equation fit the data?

R-squared = 0.4803

48.03% of the variation in statistics grades is explained by the scores of the aptitude test.

**Example 2: Multiple Regression Model**

A researcher wanted to find the effect of driving experience and the number of driving violations on auto insurance premiums. A sample of 12 drivers was selected.



*y =* the monthly auto insurance premium (in dollars) paid by a driver

*x*1 = the driving experience (in years) of a driver

*x*2 = the number of driving violations committed by a driver during the past three years

We are to estimate the regression model: *y =* *A +* *B*1*x*1 + *B*2*x*2 + ε

. reg premium experience violations

 Source | SS df MS Number of obs = 12

-------------+------------------------------ F( 2, 9) = 60.88

 Model | 17961.2895 2 8980.64477 Prob > F = 0.0000

 Residual | 1327.71045 9 147.523383 R-squared = 0.9312

-------------+------------------------------ Adj R-squared = 0.9159

 Total | 19289 11 1753.54545 Root MSE = 12.146

------------------------------------------------------------------------------

 premium | Coef. Std. Err. t P>|t| [95% Conf. Interval]

-------------+----------------------------------------------------------------

 experience | -2.747272 .9770062 -2.81 0.020 -4.957414 -.5371308

 violations | 16.10612 2.61332 6.16 0.000 10.19438 22.01786

 \_cons | 110.2761 14.61865 7.54 0.000 77.20639 143.3458

------------------------------------------------------------------------------

Using the software output, we are going to answer the following questions:

1. Explain the meaning of the estimated regression coefficients.

From the STATA output above, we can form the regression equation as follows:

$$\hat{y}=110.2761-2.747experience+16.106violations$$

The value of $a=110.2761$. This means that a person with 0 years of driving experience and 0 violations in the past three years would be charged a premium of $110.2761 per month.

The value of $b\_{1}=-2.747$ indicating that with each additional year of experience, the premium reduces by $2.474, keeping the number of violations in the past three years fixed.

The value of $b\_{2}=16.106$ indicating that with each additional violation committed in the past three years, the premium increases by $16.106, keeping the number of years of driving experience fixed.

1. Comment on the statistical significance of the coefficients.

By intuition, we expect the relationship between years of driving experience and monthly insurance premiums to be negatively related. As per our sample, the coefficient of driving experience in the regression model is $b\_{1}=-2.747<0$. To determine the statistical significance of the coefficient, we need to evaluate whether there is enough evidence present to conclude that the relationship between monthly premiums and years of driving experience is negative in the population. (STATA always run a two-tailed test by default)

$$H\_{0}: B\_{1}=0$$

$$H\_{1}: B\_{1}\ne 0$$

The test is t-distributed, where the t-statistic is -2.81. The p-value for this two-tailed test is 0.020. If we select the value of alpha to be 5%, then p-value < $α$ and we reject the null hypothesis, concluding that there is sufficient evidence present to conclude that the relationship between the two variables is different from zero in the population (the t-statistic being negative indicates that the relationship is negative). However, if we choose a 1% level of significance, then we do not reject the null hypothesis. Hence, there is a moderately strong indication that the relationship between the variables is different from 0 in the population. Hence it is difficult to conclude whether the variable is indeed statistically significant.

By intuition, we expect the relationship between number of driving violations (in the past three years) and monthly insurance premiums to be positively related. As per our sample, the coefficient of number of violations (in the past three years) in the regression model is$ b\_{2}=16.106>0$ . To determine the statistical significance of the coefficient, we need to evaluate whether there is enough evidence present to conclude that the relationship between monthly premiums and number of violations in the past three years is positive in the population.

$$H\_{0}: B\_{2}=0$$

$$H\_{1}: B\_{2}\ne 0$$

The test is t-distributed, where the t-statistic is 6.16. The p-value for this two-tailed test is 0.000. If we select the value of alpha to be 1%, then p-value < $α$ and we reject the null hypothesis, concluding that there is sufficient evidence present to conclude that the relationship between the two variables is different from zero in the population (the t-statistic being positive indicates that the relationship is positive in the population). Hence, there is overwhelming evidence about the relationship being different from 0 and the variable, number of violations in the past three years, is statistically significant in explaining variations in the dependent variable.

1. What are the values of the coefficient of multiple determination, and the adjusted coefficient of multiple determination?

$R^{2}=0.9321$ and $Adjusted R^{2}=0.9159$

1. What is the predicted auto insurance premium paid per month by a driver with seven years of driving experience and three driving violations committed in the past three years?

$$\hat{y}=110.2761-2.747experience+16.106violations\hat{y}=110.2761-2.747\left(7\right)+16.106\left(3\right)\hat{y}=139.37$$

A driver with 7 years of driving experience and three violations in the past year is predicted to pay a premium of $139.37.

1. What is the point estimate of the expected (or mean) auto insurance premium paid per month by all drivers with 12 years of driving experience and 4 driving violations committed in the past three years?

$$\hat{y}=110.2761-2.747experience+16.106violations\hat{y}=110.2761-2.747\left(12\right)+16.106\left(4\right)\hat{y}=141.74$$

1. Determine a 95% confidence interval for *B*1 (the coefficient of experience) for the multiple regression of auto insurance premium on driving experience and the number of driving violations.

According to the STATA output the 95% confidence interval for the coefficient of experience is between -4.957 to -0.537. This means with 95% confidence the interval -4.957 to -0.537 contains the true population coefficient for experience.