***Lecture 17***

Regression Analysis: Multiple Regression

In real life, a response variable (y) is affected not by one but many independent variables. Thus a more flexible model is the multiple regression model given by:

The parameters of the multiple regression are estimated by minimizing the residual sum of squares.

Statistical software or Excel can be used to estimate the multiple regression model.

Price of Orion car. Here we want to model the price ($) of a used car of ‘Orion’ brand using age of car and number of miles driven as independent variables.

|  |  |  |  |
| --- | --- | --- | --- |
| **Car** | **Age (years)** | **Miles** | **Price ($)** |
| 1 | 5 | 57000 | 8500 |
| 2 | 4 | 40000 | 10300 |
| 3 | 6 | 77000 | 7000 |
| 4 | 5 | 60000 | 8200 |
| 5 | 5 | 49000 | 8900 |
| 6 | 5 | 47000 | 9800 |
| 7 | 6 | 58000 | 6600 |
| 8 | 6 | 39000 | 9500 |
| 9 | 2 | 8000 | 16900 |
| 10 | 7 | 69000 | 7000 |
| 11 | 7 | 89000 | 4800 |

The software output for this regression is as follows:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **SUMMARY OUTPUT** | |  |  |  |  |  |
| *Regression Statistics* | |  |  |  |  |  |
| Multiple R | 0.967530 |  |  |  |  |  |
| R Square | **0.936115** |  |  |  |  |  |
| Adjusted R Square | 0.920144 |  |  |  |  |  |
| Standard Error | 880.505444 |  |  |  |  |  |
| Observations | 11 |  |  |  |  |  |
| **ANOVA** |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |
| Regression | 2 | 90883135.85 | 45441568 | 58.61236 | 1.67E-05 |  |
| Residual | 8 | 6202318.693 | 775289.8 |  |  |  |
| Total | 10 | 97085454.55 |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* |
| Intercept | **18303.5208** | 1134.76186 | 16.12983 | 2.19E-07 | 15686.76 | 20920.29 |
| Age(years) | **-950.4270** | 387.4188755 | -2.45323 | 0.039736 | -1843.82 | -57.0375 |
| Miles | **-0.0821** | 0.025520666 | -3.21889 | 0.01226 | -0.141 | -0.0233 |

The estimated model in equation form is:

Where y: Price of car ($) Age of car (years), = Miles driven

Interpretation of Parameter Estimates:

Intercept 18303.5: For an Orion car which is brand new (0 year age) and is driven 0 miles, the estimated price is $18303.5. In many cases, intercept is not meaningful to interpret since x variables at zero values are not sensible.

Coefficient of Age -950.4: Keeping number of miles driven fixed, if age of car increases by one year, its price decreases by $950.4 on average.

Coefficient of Miles driven -0.08215: Keeping age of car fixed, if number of miles driven increases by one, car’s price decreases by $0.082 on average (which corresponds to a decrease of $82.1 for every 1000 miles driven.

**Coefficient of determination R2:** Proportion of total variation in dependent variable (y) that is explained by independent variables of the model.

**SST = SSR + SSE**

For the Orion data R2 = 0.936. This shows that 93.6% variation in prices of car is explained by age of the car and number of miles driven through this model.

**Statistical Significance of Coefficients:**

For the hypothesis: (Age of car is not a useful predictor of car price)

Against (Age of car is a useful predictor of car price)

The test statistic is:

Here represents the sample estimate of the parameters. is the value of parameter under the null hypothesis.

The statistics has a student’s **T** distribution with **n – (k+1)** degrees of freedom (n = number of observations or sample size, k +1 = # model parameters including intercept=3 here)

T statistic is – 2.45 t( 0.025, 11-3 df) = , Thus null hypothesis is rejected and we conclude that indeed the age of car is a useful predictor of its price.

**Ex:** Test the hypothesis that there is a negative relationship between age and price of car.

For the hypothesis: (No or positive relationship)

Against (Age increases car price decreases)

T statistic is – 2.45, t( 0.05, 8 df) = , Thus null hypothesis is rejected, and we conclude that there is indeed negative relationship between the age and price of car.

[Alternatively, the p-value of the test = Given two tail p-value /2 = 0.03974/2=0.01987<0.05. Hence the null hypothesis is rejected in favor of alternative at 5% sig level].

**Prediction from the model:** Suppose we want to predict the price of an Orion which is 4 years old and which is already driven 50,000 miles.

= $10,396.9

**Estimation of multiple regression in Excel:**

Go to Data Tab> Data Analysis > Regression

A screenshot of a computer

Description automatically generated

Input the y range and x range (the x variables must be in adjacent columns). Click labels if variable name row is also selected.

A screenshot of a computer

Description automatically generated

Note: Analysis Tool pack must be installed in Excel. To do this within Excel

**Files > Options > Add-Ins >Analysis Tool Pack > Go > Analysis ToolPak > OK**

Then the Analysis Tool Pack named ‘Data Analysis’ is visible in the Data tab.

**#Multiple regression in R**

orion=read.csv(file.choose()) # choose orion1.csv data

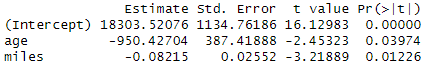
attach(orion)

head(orion)

model1=lm(price~age+miles,data=orion)

summary(model1)

round((summary(model1)$coefficients), 5) # to preset outcome with 5 decimals (avoid scientific notation)



Note: In any multiple regression problem, we can expect from two students to be able to do these things

**F Test for overall significance in multiple regression:**

vs

This F test with DF (**k** and **n – (k+1**) and is reported by Excel (in ANOVA section) and all statistical software. [Note: the manual calculation of F test is not required].

For the Orion case, **F = 58.61, p-value = 0.0000167**, null hypothesis is rejected, and we conclude that at least one variable (age or miles or both) has a significant impact on price.

(1)Estimate parameters from software. Alternatively, software output or estimated eq can be provided in textbook form along with relevant statistics e.g., standard error, R sq etc.

(2) Interpret each coefficient in practical terms.

(3)Interpret R sq in practical terms.

(4) Make predictions given relevant predictor values.

(5) Test hypothesis on individual coefficients.

The overall F test (of the hypothesis that all parameters (except intercept) are zero and its p-value can also be asked.

**Anderson Ex 12,14, pdf p-696.**