Assignment Topic

Student Name

University Name

Instructor Name

Date

**Multiple Regression Analysis**

 Multiple regression analysis is performed when we are dealing with a dependent variable and two or more independent variables. The regression model is given below

Y = b0 + b1 \* X1 + b2 \* X2 + ….. bn \* Xn

Where b0, b1, …., bn are regression coefficients

 The dependent variable taken into consideration should be continuous and measured under interval scale and the independent variables taken into consideration may be combination of continuous or nominal variables. If the independent variable is nominal variable and has three categories, then we need to create two dummy variables

 In order to perform the multiple regression analysis, we first need to check when the variables taken into consideration satisfies the following assumptions

* Linear relationship
* Multivariate normality
* No or little multicollinearity
* No auto-correlation and
* Homoscedasticity

 Multicollinearity is an issue whether we see that there exists a significant relationship among the independent variables included in the study. In cases of perfect multicollinearity, OLS estimators are not even defined. An exact linear relationship between two or more (explanatory) variables; more than one exact linear relationship between two or more explanatory variables. In perfect collinearity there is an exact linear relationship between two or more variables, whereas in imperfect collinearity this relationship is not exact but an approximate one.

Heteroscedasticity is violation of the assumption that “Variances among the groups are equal” and it normally occurs when the variance of the error terms differ across observations. Heteroscedasticity has serious consequences for the OLS estimator. Although the OLS estimator remains unbiased, the estimated SE is wrong. Because of this, confidence intervals and hypotheses tests cannot be relied on. In addition, the OLS estimator is no longer BLUE. If the form of the Heteroscedasticity is known, it can be corrected (via appropriate transformation of the data) and the resulting estimator, generalized least squares (GLS), can be shown to be BLUE. The effects of Heteroscedasticity are:

* OLS is still unbiased and consistent
* The standard errors of the estimates are biased if we have Heteroscedasticity

**Objective**

The main objective of this study is to determine the relationship between the dependent variable Annual Cost of the cars and the independent variables: City Mileage, Highway Mileage, Engine Displacement and Number of Cylinders. In this study, we will also be interested in determining if the multiple regression model account for how much of the errors in the model

**Multicollinearity**

When two or more independent variables are highly correlated with each other, then there exists multicollinearity. When these variables are included in constructing the regression model, then some of these variables turn insignificant predictors of dependent variable which actually is not true. This is due to the effect of multicollinearity. Moderate multicollinearity may not be problematic. However, severe multicollinearity is a problem because it can increase the variance of the coefficient estimates and make the estimates very sensitive to minor changes in the model. The result is that the coefficient estimates are unstable and difficult to interpret. Multicollinearity saps the statistical power of the analysis, can cause the coefficients to switch signs, and makes it more difficult to specify the correct model. Variance Inflation Factor can be used to check multicollinearity. A VIF of 5 or greater indicates a reason to be concerned about multicollinearity.

**Regression Analysis**

In this study, we also include a dummy variable that represents the location of car purchased, where 1 represents whether the car is purchased in city and 0 represents that the car is purchased in out of city circle

The regression output is given below

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ***Regression Statistics*** |  |  |  |  |  |
| Multiple R | 0.9808 |  |  |  |  |  |
| R Square | 0.9620 |  |  |  |  |  |
| Adjusted R Square | 0.9540 |  |  |  |  |  |
| Standard Error | 199.2039 |  |  |  |  |  |
| Observations | 30 |  |  |  |  |  |
|  |  |  |  |  |  |  |
| **ANOVA** |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |
| Regression | 5 | 24085043.8220 | 4817008.7644 | 121.3897 | 0.0000 |  |
| Residual | 24 | 952372.8446 | 39682.2019 |  |  |  |
| Total | 29 | 25037416.6667 |   |   |   |  |
|  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* |
| Intercept | 4200.4603 | 356.1597 | 11.7938 | 0.0000 | 3465.3828 | 4935.5378 |
| City Mileage | -32.9735 | 19.2154 | -1.7160 | 0.0990 | -72.6321 | 6.6852 |
| Highway Mileage | -62.1202 | 16.1185 | -3.8540 | 0.0008 | -95.3872 | -28.8531 |
| Engine Displacement | -133.2240 | 71.0561 | -1.8749 | 0.0730 | -279.8766 | 13.4287 |
| Number of Cylinders | 203.5851 | 42.0692 | 4.8393 | 0.0001 | 116.7585 | 290.4117 |
| Location | 59.7818 | 81.7053 | 0.7317 | 0.4715 | -108.8496 | 228.4132 |

|  |  |
| --- | --- |
| **Durbin-Watson Calculations** |  |
|  |  |
| Sum of Squared Difference of Residuals | 2401209.36 |
| Sum of Squared Residuals | 952372.8446 |
|  |  |
| **Durbin-Watson Statistic** | **2.521291292** |

|  |  |
| --- | --- |
| Independent Variables | VIF |
| City Mileage | 9.646205888 |
| Highway Mileage | 8.543887185 |
| Engine Displacement | 9.253561821 |
| Number of Cylinders | 9.401642532 |
| Location | 1.172004717 |

In the results above, Highway Mileage, and number of cylinders are significant while city mileage and engine displacement is not significant. However, three of the VIFs are very high because they are well over 5. These values suggest that the coefficients are poorly estimated and we should be wary of their p-values.

**Results summary**

The coefficient of determination is 0.9620. This indicates that 96.20% of the variation in he dependent variable is explained by the regression model and the remaining 3.8% left unexplained. The value of f test statistic is 121.3897 and its corresponding p – value is 0.000< 0.05, indicating that the estimated regression model is good fit in predicting the dependent variable

**References**

James T. McClave, P. George Benson, Terry Sincich, ‘Statistics for Business and Economics. Pearson; 11 edition (January 13, 2010). ISBN-13: 978-0321640116

Douglas Lind, William Marchal, Samuel Wathen, ‘Basic Statistics for Business and Economics, McGraw-Hill/Irwin; 7 edition (January 11, 2010). ISBN-13: 978-0077384470

Donald R. Cooper, Pamela S. Schindler, ‘Business Research Methods’, 11th Edition (January 1, 2011).