Wk. 8 Material, Regression Analysis

Topic 4B | Lecture of Wk. 8 | Tutorial held Week 10

Contents

[Introduction to Multiple Regression 2](#_Toc400195944)

[Solving for Coefficient Values via Matrix values 2](#_Toc400195945)

[Assessing Overall Significance of Regresion 3](#_Toc400195946)

[Coefficient of Zero 3](#_Toc400195947)

[Hypothesis Test of overall significance ( F Test for Significance) 3](#_Toc400195948)

[Hypothesis 3](#_Toc400195949)

[Rejection Region 3](#_Toc400195950)

[Test Statistic 3](#_Toc400195951)

[Conclusion 3](#_Toc400195952)

[ANOVA Table for F-Test of Significance 4](#_Toc400195953)

[Significance of Individual Coefficients 5](#_Toc400195954)

[Hypothesis Test 5](#_Toc400195955)

[Hypothesis 5](#_Toc400195956)

[Test Statistic 5](#_Toc400195957)

[Rejection Region 5](#_Toc400195958)

[Confidence Interval 5](#_Toc400195959)

[Prediction Interval 5](#_Toc400195960)

[Excel Output 6](#_Toc400195961)

[Coefficient of Determination 6](#_Toc400195962)

[Multicollinearity 6](#_Toc400195963)

[Klein’s Rule 6](#_Toc400195964)

[Forward Selection and Backward Elimination Method 6](#_Toc400195965)

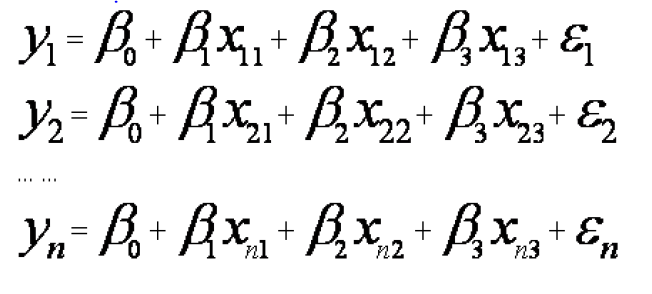
# Introduction to Multiple Regression

Where a response value is *linearly* related to multiple independent values then:

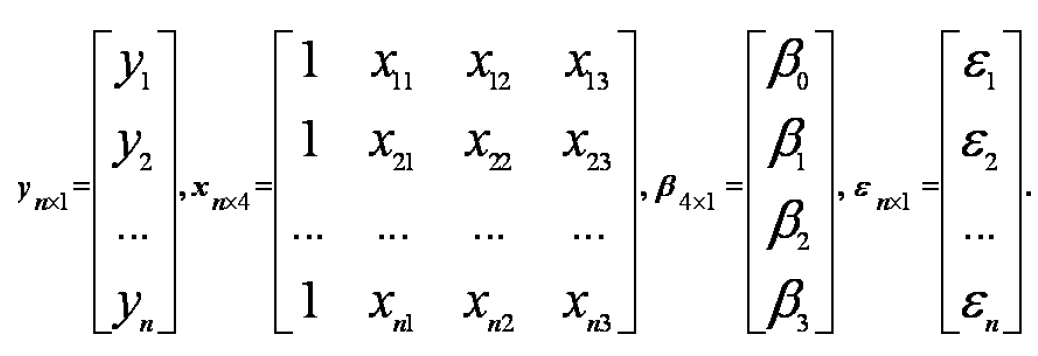
Where represents the residual or error that occurs disjoint from the model.

## Solving for Coefficient Values via Matrix values

The statistical model must satisfy the following n equations:



Thus writing it in matrix form:



The coefficient values can be estimated by Minitab or Excel and the use of matrix mathematics isn’t within the scope of this unit.

# Assessing Overall Significance of Regression[[1]](#footnote-1)

## Coefficient of Zero

If the coefficient () of any term in the regression equation:

Is insignificant it will be zero, that is to say a coefficient of zero will represent a linear regression with no real relationship between & .

(This doesn’t include the intercept value of )

## Hypothesis Test of overall significance ( F Test for Significance)

Before determining whether or not specific coefficients are significant to the linear regression it is best to perform an overall test for overall fit

The *Test Statistic* and degrees of freedom can all be found by way of the ANOVA table.

### Hypothesis

All coefficient values are insignificant and equal to zero

### Rejection Region

fails for:

Where:

### Test Statistic

### Conclusion

At least on coefficient ( is nonzero and the regression is overall significant

All coefficients are zero and the regression has no significance.

## ANOVA Table for F-Test of Significance

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***Source of Variation*** | ***Sum of Squares*** | ***df*** | ***Mean Square*** | ***F-stat*** |
| ***Regression (Explained variation of Data as per the Model)*** |  |  |  |  |
| ***Residual or Error (Random Variation that occurs distinct of the model)*** |  |  |  |  |
| ***Total)*** |  |  |  |  |

The Excel p-value is given by:

The variance of the regression is given by:

The standard error of the regression is:

# Significance of Individual Coefficients[[2]](#footnote-2)

In order to test whether an individual coefficient is significant it is tested as being equal to zero, by default the tests are two tailed because if the null hypothesis can be rejected by a two tailed test then it can also by rejected in a one-tailed test at the same .

## Hypothesis Test

### Hypothesis

### Test Statistic

Where:

, represents the standard error of , (the standard error being the standard deviation)

The value of is not calculated because the calculation is tedious, instead it is usually taken from a Minitab or excel output.

### Rejection Region

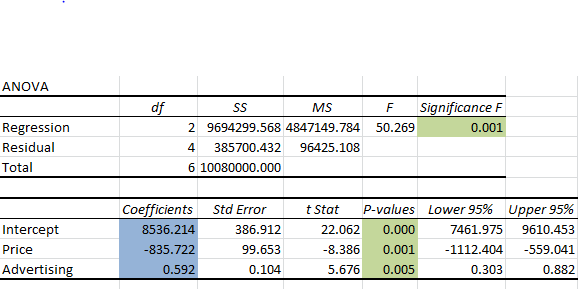
The Null Hypothesis fails for:

Where:

## Confidence Interval

A Confidence Interval for the Coefficient is given by:

## Excel Output



In this layout the upper table represents a typical ANOVA table and the bottom layout represents data regarding coefficient values.

* *Coefficients:* This is the corresponding value of
* *StdError:* This is the value of
* *t Stat*: This is the value of for the significance of , it is given by

# Coefficient of Determination

More predictor values can inflate the value of , which represents the proportion of data explained by the model, the adjusted value makes it relative to the number of parameters

# Multicollinearity

This occurs when other variables are related to one another instead of just x to y, while to some degree this is always going to be somewhat the case, the depth of such concern would depend upon the degree of multicollinearity.

## Klein’s Rule

Klein’s Rule states that multicollinearity is only an issue where the correlation coefficient matrix demonstrats correlations higher that the overall multiple correlation coefficient, i.e. R.

# Forward Selection and Backward Elimination Method

Where a coefficient has a statistically insignificant coefficient (that is the *t-stat* of that coefficient is less that )

The smallest t-value’s can be removed one by one, (Backward elimination)

OR

The coefficients with the largest *t-stat’s* can be selected to be used only, (Forward Selection)

Where:

This is all in an effort to simplify the equation

# Confidence Interval of

## Confidence Interval

A () % Confidence Interval is given by:

This is a confidence Interval for the expected value of .

This is used where

## Prediction Interval

A () % Prediction Interval is given by:

This is a prediction interval for a single value of

1. David Doane & Lori Seward, *Applied Statistics In Business & Economics*, (McGraw Hill Publishing, 2013, 4th ed.) ch. 13.2, p. 553 [↑](#footnote-ref-1)
2. Ibid, Ch. 13.3, p. 557 [↑](#footnote-ref-2)