

MS -Business Intelligence & Analytics

Fall 2015

**BIA – 652 C**

*October 21, 2015*

**Mohit Ravi Ghatikar**

**CWID - 10405877**

Multivariate Data Analytics – Homework 4

**Ethics Statement**

I pledge on my honor that I have not given or received any unauthorized assistance on this assignment /examination. I further pledge that I have not copied any material from a book, article, the Internet or any other source except where I have expressly cited the source.

Signature \_Mohit Ravi Ghatikar\_\_\_\_\_

Date: 10/21/2015\_\_\_\_

Question:

Using the Iris dataset in the Sashelp library, develop a regression model for predicting the “Sepal Length in mm” (SepalLength) using the other variables in the dataset.

Does “Iris Species” (Species) play a significant factor in predicting Sepal Length?

Solution:

In the Iris Dataset, we first need to create dummy variables for Species.

Ie. Iris\_spec1=1 when Species='Setosa'

Iris\_spec2=1 when Species = 'Versicolor'

Iris\_spec3=1 when Species = 'Virginica'

The other Independent variables for predicting Sepallength are Sepalwidth, Petallength and Petalwidth.

We first run a Regression Model with Sepal Length as the dependent variable and Sepalwidth, Petallength, Petalwidth, Iris\_spec1 and Iris\_spec2 as the independent variables. We then check the variable with the highest VIF (i.e high co-relation) and remove those variables from the model and run the regression again with the remaining variables.

By doing so we remove Petalwidth and Iris\_spec1 due to high VIF. We also remove Iris\_spec2 because its p-value is insignificant.

We are finally left with Sepalwidth and Petallength as our most significant and the best predictors for Sepallength.

The Regression Equation is:

**Sepallength = 22.49 + 0.59 \* Sepalwidth + 0.47 \* Petallength.**

To decide if Species plays a significant role in predicting Sepal length, we use interaction variables and run a regression model.

Ie. We run a regression model of Sepalwidth, Petallength, Iris\_sepc1 and Iris\_spec1\*Sepalwidth as independent variables and sepallength as independent variable. We find that the interaction variable is insignificant.

Similarly, we run a regression model of Sepalwidth, Petallength, Iris\_sepc1 and Iris\_spec1\*Petallength

Sepalwidth, Petallength, Iris\_sepc2 and Iris\_spec2\*Sepalwidth

And Sepalwidth, Petallength, Iris\_sepc2 and Iris\_spec2\*Petallength.

In all the above four cases, the interaction variables are insignificant.

**Therefore, we can conclude that Iris Species does not play a significant role in predicting Sepallength.**