MS -Business Intelligence & Analytics

Fall 2015

**BIA – 652 C**

*October 7, 2015*

**Mohit Ravi Ghatikar**

**CWID - 10405877**

Multivariate Data Analytics – Homework 2

**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: 10/07/2015\_\_\_

**Problem 6.2**

From the family lung function data set in Appendix A, perform a regression analysis of weight on height for fathers. Repeat for mothers. Determine the correlation coefficient and the regression equation for fathers and mothers. Test that the coefficients are significantly different from zero for both sexes. Also, find the standardized regression equation and report it. Would you suggest removing the woman who weighs 267 pounds from the data set?

Discuss why the correlation for fathers appears higher than that for mothers.

**Solution:**

**Analysis:**

***Regression Analysis for Father:***

The Regression Analysis is a good model since the p-value for the F-test from the ANOVA table and T-tests for Intercept and Height\_father are less than the threshold value of 5%.

From the Results table, the Regression Equation is:

**Weight\_father = -129.06 + 4.49 \* Height\_father**

The Standardized Regression Equation is:

**Weight\_father = 0 + 0.52 \* Height\_father**

***Regression Analysis for Mother:***

The Regression Analysis isn’t a very good model compared to Regression model of father. The p-value for the F-test from the ANOVA table is less than 5%. The T-test for the slope (Height\_mother) is less than 5% but the T-test for the intercept fails to reject the Null hypothesis (Beta1=0) since the p-value is greater than 5%.

From the Results table, the Regression Equation is:

**Weight\_mother = -108.17 + 3.98 \* Height\_mother**

The Standardized Regression Equation is:

**Weight\_mother = 0 + 0.31 \* Height\_mother**

***Co-relation Coefficient for Father:***

The Co-relation Coefficient for Weight\_father & Height\_Father is : 0.52116

To test if the co-efficient is significantly from zero, we set

Null Hypothesis: Rho = 0

Alternate Hypothesis: Rho is not equal to 0.

The p-value we obtain (<0.001) is less than the threshold value of 5%. Therefore we reject the Null Hypothesis.

***Co-relation Coefficient for Mother:***

The Co-relation Coefficient for Weight\_mother & Height\_mother is : 0.31758

To test if the co-efficient is significantly from zero, we set

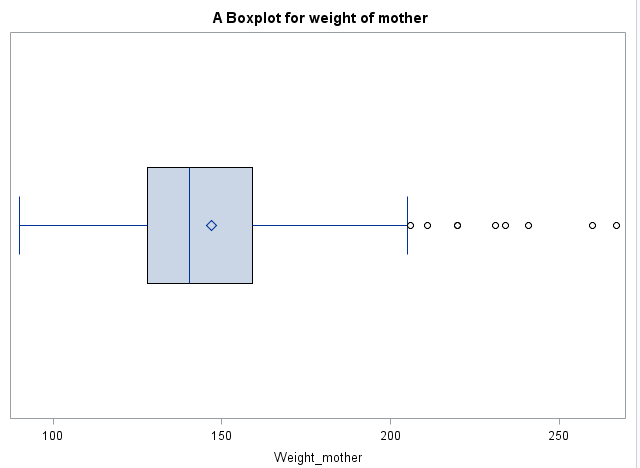
Null Hypothesis: Rho = 0

Alternate Hypothesis: Rho is not equal to 0.

The p-value we obtain (<0.001) is less than the threshold value of 5%. Therefore we reject the Null Hypothesis.

***Decision to remove the 267 pounds data from the Weight of mother dataset:***

***:***



As we can see from the boxplot of Weight\_mother variable, there are several outliers.

Removing the outlier of 267 will not make a major difference in the Regression Model. Removing the outlier will slightly improve the model but will not make it drastically better.

The Co-relation for Father is better than mother because the Regression model of Father is better than the mother. Specifically the R2 value for father is 0.27 or 27% of the model can be explained by regression whereas the R2 value of mother is 0.1 or only 10% of the model can be explained by regression. The rest is explained due to error.

**Problem 6.9**

From the depression data set described in Table 3.4 create a data set containing only the variables AGE and INCOME.

(a) Find the regression of income on age.

(b) Successively add and then delete each of the following points:

AGE INCOME

42 120

80 150

180 15

and repeat the regression each time with the single extra point. How does the regression equation change? Which of the new points are outliers? Which are influential?

**Solution:**

***Regression Analysis for income on age***

The Regression Analysis is a good model since the p-value for the F-test from the ANOVA table and T-tests for Intercept and age are less than the threshold value of 5%.

The Regression equation is **income=27.772 - 0.162\* age.**

We successively add and delete the three points, and repeat the regression each time with the single extra point. The Regression analysis for all three extra points are good models model since the p-value for the F-test from the ANOVA table and T-tests for Intercept and age are less than the threshold value of 5%.

The Regression Equation for additional data points of 42 120 is:

**Income = 28.218 – 0.164 \* age………………(1)**

The Regression Equation for additional data points of 80 150 is:

**Income = 26.031 – 0.112 \* age………………(2)**

The Regression Equation for additional data points of 180 15 is:

**Income = 26.957 – 0.142 \* age………………(3)**

An **influential point** is an outlier that greatly affects the slope or the intercept of the Regression line or significantly changes the R2 value. From the above, Equation (2) is influential or the points (80,150) are influential points since slope has reduced from **0.162 to 0.112** and R2 value decreases from **0.03 to 0.01**.

The Regression equation (3) or the points (180,15) is also influential since the slope and intercept have changed. But it is not as influential as the previous data point.

The Regression equation (1) or the points (42,120) has high leverage but is not influential since the slope is almost the same with a slight variation in intercept. This implies that the outlier has no significant effect on Regression.