MS -Business Intelligence & Analytics

Fall 2015

**BIA – 652 C**

*October 17, 2015*

**Mohit Ravi Ghatikar**

**CWID - 10405877**

Multivariate Data Analytics – Homework 3

**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: 10/17/2015\_\_\_

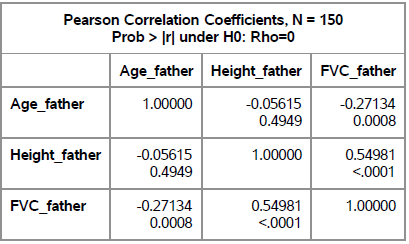
**Problem 7.2**

Fit the regression plane for the fathers using FFVC as the dependent variable and age and height as the independent variables.

**Solution:**

The objective of the problem is to perform Multiple Linear Regression Analysis on FVC\_father with Age\_father and Height\_father as the dependent variables.

Before performing the analysis, we find the Co-relation Matrix of the three variables.



In an ideal case, there should not be any Co-relation between the two independent variables. The r value for Height\_father & Age\_father is -0.05 with a p-value of 0.49. This is statistically insignificant and we can ignore the co-relation. Hence there is no multicollinearity. We can also see that there is a Co-relation between FVC\_father – Age\_father and FVC\_father – Height\_father. The p-values for both are less than 0.05 and hence they are statistically significant.

Next we run the Regression Model. The analysis is described below:

1. The Model is statistically significant because we obtain the F-value of 41.4 whose p-value if less than 0.05.
2. The R-square value is 0.36 and adjusted R-square is 0.35. About 35% of the model can be explained by regression.
3. The Parameter estimates of Intercept, Age\_father & Height\_father are statistically significant. Their T-values are less than 0.05.
4. The Variation Inflation Factor for Age\_father & Height\_father is around 1. This shows that there is hardly any Multicollinearity.
5. The Regression Equation is :

FVC\_father = - 453.9 – 2.77 \* Age\_father + 15.31 \* Height\_father..

**Problem 7.5**

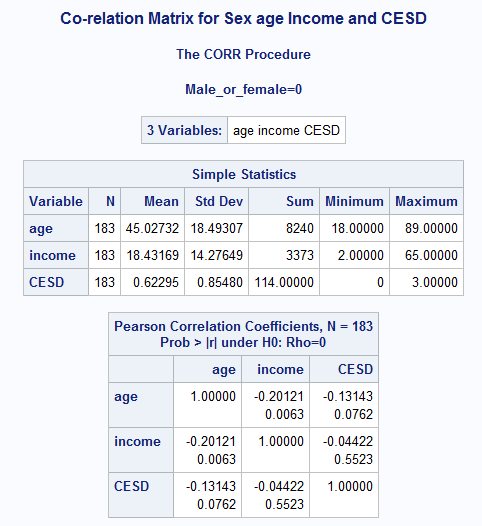
From the depression data set described in Table 3.4, predict the reported level of depression as given by CESD, using INCOME, SEX, and AGE as independent variables. Analyze the residuals and decide whether or not it is reasonable to assume that they follow a normal distribution.

**Solution:**

The objective of the problem is to perform Multiple Linear Regression Analysis on CESD with Income, Sex and Age as the dependent variables. Then analyze if the Residuals follow a normal distribution.

We create a dummy variable for Sex since it’s a categorical variable. We assign the Value ‘1’ for Male and ‘0’ for female.

The Co-relation Matrix segregated by Sex is given below:



1. We can see that the Co-relation between age & Income for both Males and Females are statistically insignificant since the p-values are less than 0.05.
2. For Males, the Co-relation between CESD and age is statistically insignificant and the Co-relation CESD and income is statistically significant. This indicates that there might be no co-relation between CESD and age for males.
3. For Females, Co-relation between CESD and age is statistically insignificant and the Co-relation CESD and income is also statistically insignificant. This indicates that there might be no co-relation between CESD and the independent variables for females.

Next we run the Regression Model. The analysis is described below:

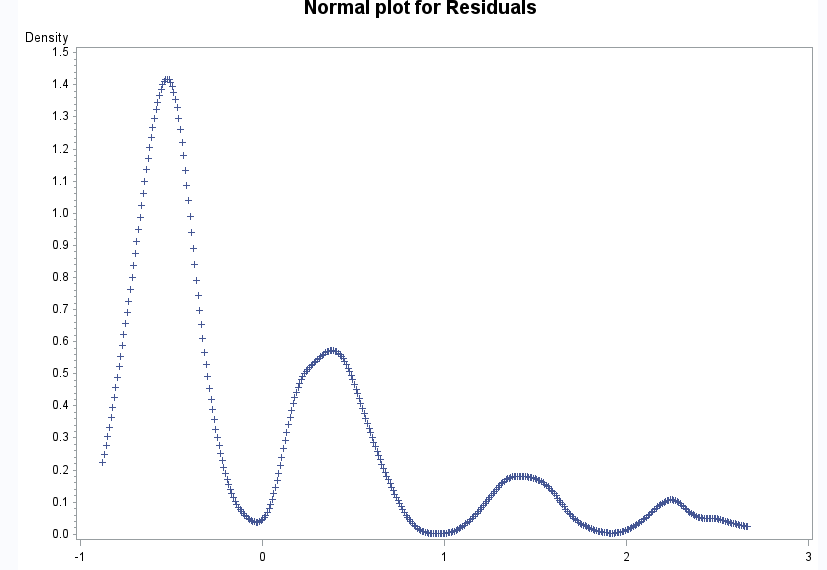
1. The Model is statistically significant because we obtain the F-value of 3.93 whose p-value if less than 0.05.
2. The R-square value is 0.03 and adjusted R-square is 0.02. About 2% of the model can be explained by regression.
3. The Parameter estimates of Intercept, Age & income are statistically significant. Their T-values are less than 0.05. The parameter estimate for the Sex variable is statistically insignificant since its p-value is greater than 0.05.
4. The Variation Inflation Factor for Age, income and sex is around 1. This shows that there is hardly any Multicollinearity.
5. The Regression Equation is :

For Males: CESC = 1.05 – 0.0067 \* Age – 0.0068 \* Income – 0.117

= 0.933 – 0.0067 \* Age – 0.0068 \* Income

For Females: CESC = 1.05 – 0.0067 \* Age – 0.0068 \* Income

The Normal Plot for the Residuals is:



As we can see from the above plot, The Residuals are not normally distributed. Instead they follow a multimodal distribution.

**Problem 8.11**

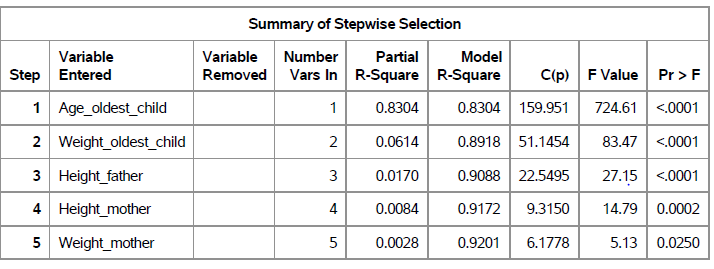
Using the methods described in this chapter and the family lung function data described in Appendix A, and choosing from among the variables OCAGE, OCWEIGHT, MHEIGHT, MWEIGHT, FHEIGHT, and FWEIGHT, select the variables that best predict height in the oldest child. Show your analysis.

**Solution:**

We need to predict the height in the oldest child from the following set of variables and select the best variable that predicts the height:-

1. Age of oldest child
2. Weight of oldest child
3. Height of mother
4. Weight of mother
5. Height of father
6. Weight of father.

We run a Stepwise Selection procedure. The analysis is presented below:



We can see that Age of the oldest Child is the best predictor for height of the oldest child since it has the highest Value of R square Value.

Next we run a simple Linear Regression with Height\_oldest\_child as the dependent variable and Age\_oldest\_child as the independent variable. The Analysis is present below:

1. The Model is statistically significant because we obtain the F-value of 724.61 whose p-value if less than 0.05.
2. The R-square value is 0.83 and adjusted R-square is 0.82. About 82% of the model can be explained by regression.
3. The Parameter estimates of Intercept and Age\_oldest\_child are statistically significant since their T-values are less than 0.05.
4. The Regression Equation is :

Height\_oldest\_child = 35.09 + 1.97 \* Age\_oldest\_child.

**Problem 8.15**

From among the candidate variables given in Problem 8.11, find the subset of three variables that best predicts height in the oldest child, separately for boys and girls. Are the two sets the same? Find the best subset of three variables for the group as a whole. Does adding OCSEX into the regression equation improve the fit?

**Solution:**

We run a forward selection procedure to select the best three predictors. They are Age\_oldest\_child, Weight\_oldest\_child & Height\_father. The three best predictors are same for both boys and girls.

The Regression Analysis is as follows:

1. The Model and the parameters estimates are statistically significant for both boys and girls.
2. The R-square value is 0.93 and adjusted R-square is 0.92 for boys. About 92% of the model can be explained by regression.
3. The R-square value is 0.879 and adjusted R-square is 0.874 for girls. About 87% of the model can be explained by regression.
4. The Variance Inflation factors for the parameter estimates in both Boys and girls is less than 4. There is a slight co-relation among the independent variables but the model as whole remains a good fit.
5. The Regression Equation is for boys is:

Height\_oldest\_child = 17.16 + 1.35 \* Age\_oldest\_child + 0.07 \* Weight\_oldest\_child + 0.27 \* Height\_father.

The Regression Equation is for girls is:

Height\_oldest\_child = 11.23+ 1.21 \* Age\_oldest\_child + 0.06 \* Weight\_oldest\_child + 0.38 \* Height\_father.

By adding the sex of the oldest child in the regression equation, the adjusted R2 improves. But we cannot say if this improves the model or not.