BIA652D Multivariate Data Analysis

Factor Analysis of HATCO Dataset

GROUP5 Homework 5

MEHRNOOSH OGHBAIE

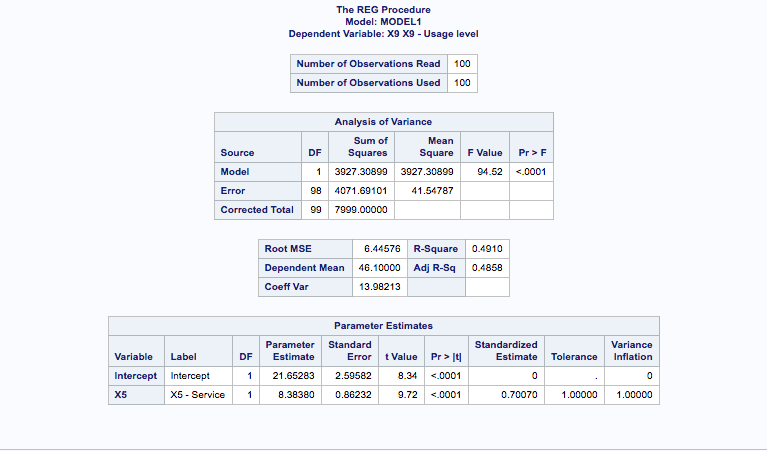
2016

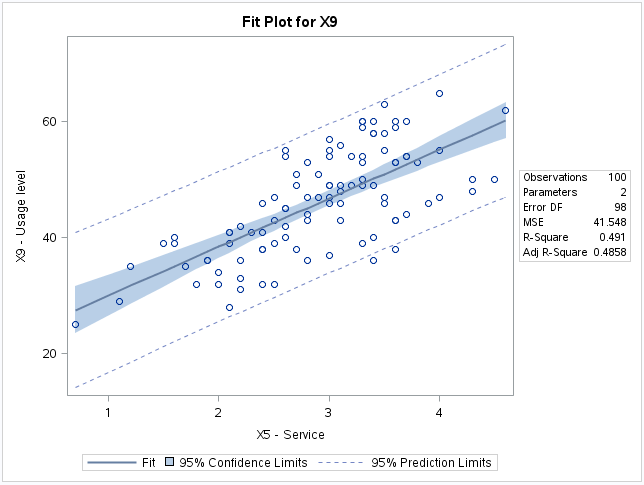
# **Summary of Multiple regression Analysis**

Here is the report on Multiples regression analysis on nine variables from HATCO database. X1-X7 and X9 are all numerical. X8 is a two levels categorical variable.

In the first part of the assignment, correlation between different values of HATCO was calculated. Following variables are ordered based on their correlation with X9: X5, X1, X3, X6, X4, X7, X2.

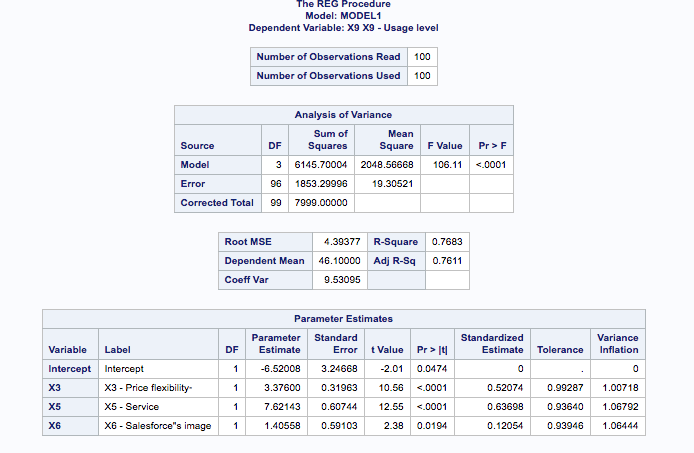
Analyzing the simple regression of the best single independent shows that P-values are pretty significant. The residuals are all between (-3,3). Cook’s d shows the most significant records to be 3,17,71. Q-Q plots of residuals don’t seem to be very normal. Residuals are not equally distributed for X5. There are some outliers beyond the prediction limit.

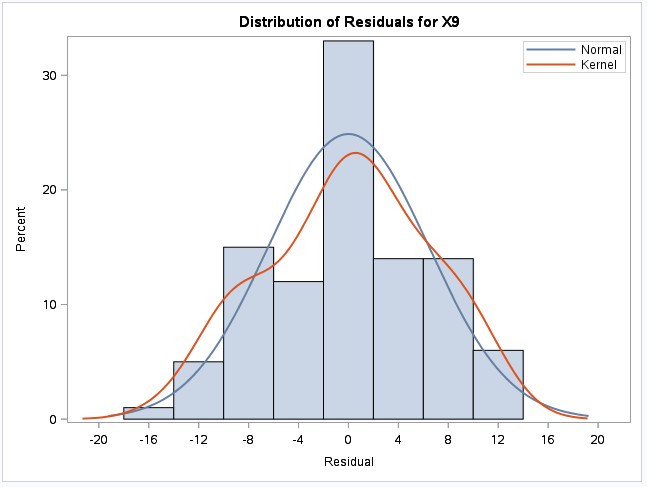
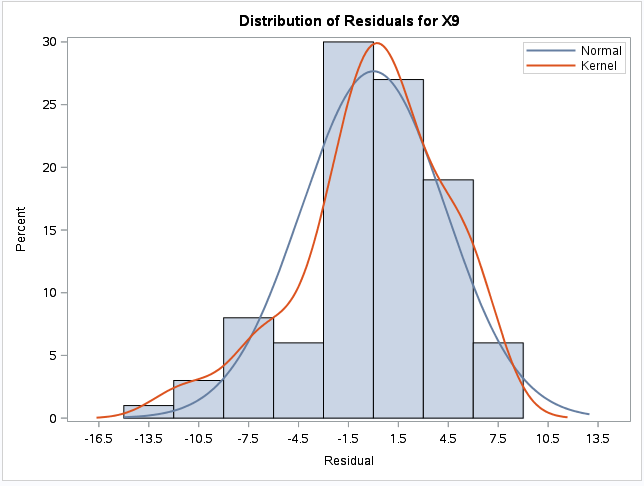




In the next part of the assignment, stepwise regression analysis was done on 7 variables:

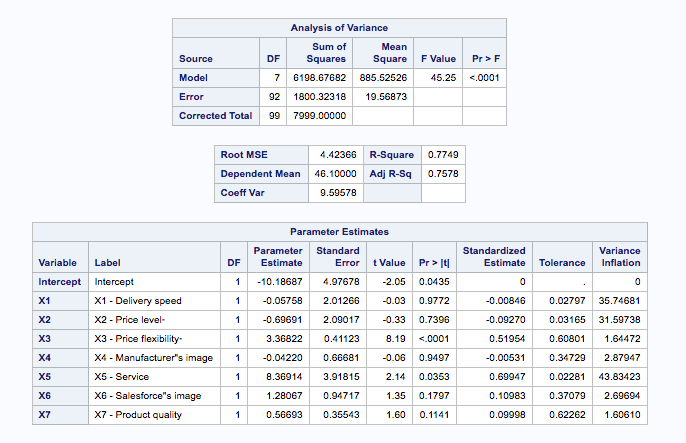
X1, X2, X3, X4, X5, X6, X7. X5 was first variable added as the variable with highest correlation. The p-value was significant. In the second step X3 was added with significant p-value, and finally X6 with fairly significant p-value was added. All residuals are between (-3,3). Cook’s D for X9 shows that records 7, 11, 14 and 100 look to be more influential in the model. The stepwise model seems to be slightly better than single variable model. The residuals are closer to normal in the middle of quantile, but not very normally distributed when go further from the center of quantile. Residuals for X9 are not equally distributed along variables X3, X5 and X6 especially in the lower part of residual axis <-10. Rstudent/leverage have few outliers, leverages and mix of outliers& leverages points. VIF for all variables is around 1 and no problem in terms of multi-collinearity.





The picture in the left shows the distribution of X9 in single regression over X5, and the picture in the right shows the X9 distribution based on multiple regression model over (X5, X3, X6). The Adj R-Square improved from 0.4858 in single regression model to 0.7611 in the multiple regression models.

The confirmatory full regression analysis shows significant relationships between X9 and X1, X3 and X5 with p-value less than 0.05. Adj R-square decreased in compare to previous model from 0.7611 to 0.7578. Residuals don’t seem to be equally distributed across the variables. Cook’s D shows that records 7, 14, 22, 100 were the most influential one in this model. R-student shows few records lower than -2. There are couple of outliers and leverages but no mix of them. Q-Q plot doesn’t seem to be very normal. X5, X1 and X2 are showing high VIF, there is multi-collinearity between variables.



Finally, stepwise analysis including X8 shows higher Adj R-squared than other models. Studentized residual shows lower than -3 residual for record 7. Cook’s d for X9 tells us that record 7, 11, 14 and 100 are the most influential ones. The Q-Q norm plot shows improvement compare to previous one. The model is fairly normal. VIF are around 1. There’s no multi-collinearity between variables.

