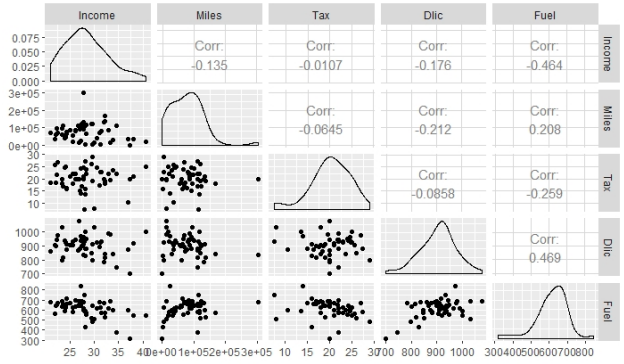
1 Introduction

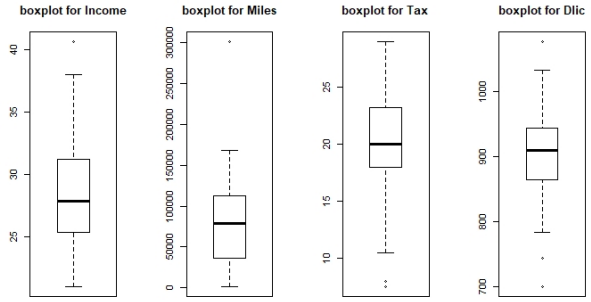
In this statistical analysis, the aim is to formulate a general linear model, which is able to interpret and predict relationship between the response variable, fuel consumption, and the predictor variables in the dataset. Furthermore, by using classical and robust methods, detecting outliers of the dataset is another purpose of this study.

2 Descriptive Statistics

Multicollinearity and a significant volume of outliers are major aspects of disturbing in statistical analysis and interpretation. In figure 1, the predictor variables of the dataset seem to have no Multicollinearity. While there are some outliers identified for each variables and then we need to take a consideration into further analysis and interpretation.



(a) Figure 1: Correlation with scatter plots



(b) Figure 2: boxplot of the predictors

3 Ordinary Least Squares Estimates

3.1 AIC-based Methods

Selecting significant predictor variables of the dataset is the first procedure in a multivariate regression analysis. In this case, there are 4 predictor variables; which are names as "Income," "Miles," "Tax," and "Dlic." By AIC-based methods, all of the methods have the same result that three predictor variables are significant in the regression model. The equation of the regression model is in equation 1.

$$Fuel = 3.840e^{02} - 7.137Income + 4.016e^{-04}Miles + 5.353e^{-01}Dlic \tag{1}$$

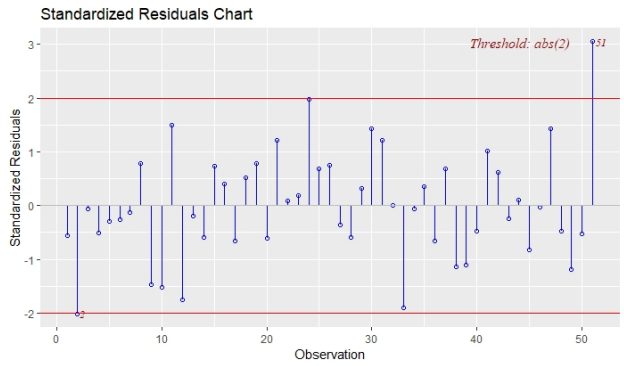
3.2 F-test-based Manual Methods

Apart AIC-based selection methods, another classical predictor variables selection, F-test selection, is a strategy to formulate regression model with significant predictor variables. Based on the results from the methods, predictor variables, "Tax," has to be pooled in the linear regression model. The equation of the linear regression model after pooled the "Tax" variables is in the equation (2).

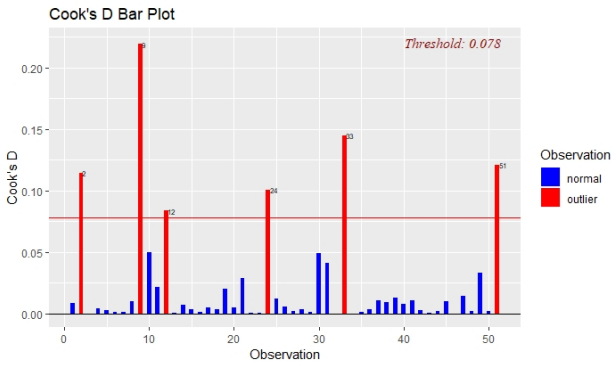
$$Fuel = 2.669e^{02} - 6.995Income + 4.349e^{-04}Miles + 5.644e^{-01}Dlic \tag{2}$$

3.3 Classical Methods to detect influential outliers

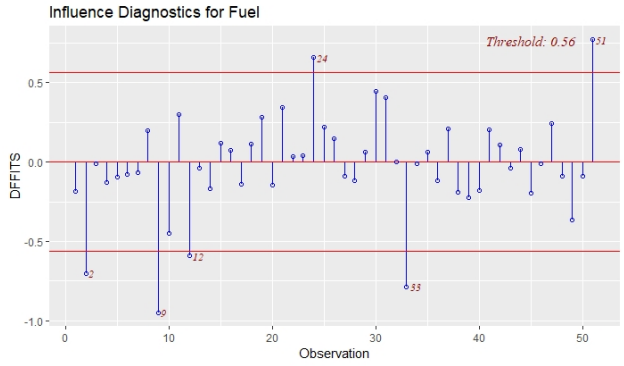
As a single outlier can cause serious misinterpretation in statistical analysis, detecting and consideration of possible outlier is an important aspect when analyzing a dataset. In this study, we use several functions for the sake of detecting influential outliers from "olsrr" package in R software. Based on the results of the plots in figure 3,4,5, and 6, observation 2 and 51 are suspected to be common influential outliers in this dataset.



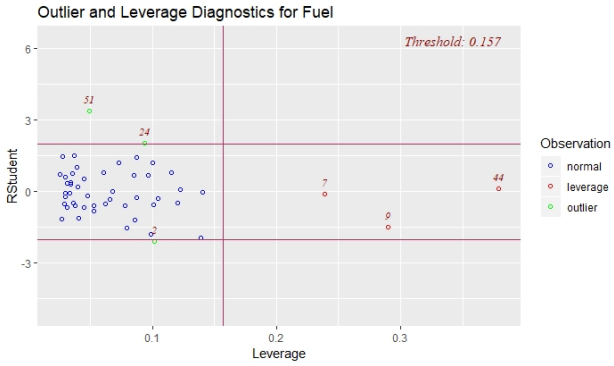
(c) Figure 3: Standardized Residual



(d) Figure 4: Cook' Distance Bar



(e) Figure 5: DFFITS plot



(f) Figure 6: Studentized Residuals vs Leverage Plot

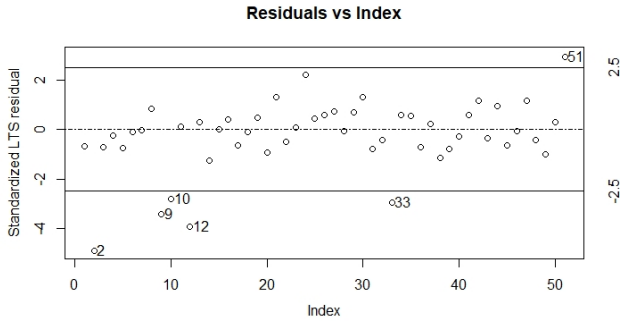
4 Robust Regression

4.1 Least Trimmed Squares Estimate

A robust regression method is able to find a fitted linear model based on majority of the data, in addition, the method is useful for detecting influential outliers. In this study, the robust Least Trimmed Squares (LTS) method uses to obtain a fitted linear model with detecting influential outliers. As the degree of robustness for an estimator is highly dependent on breakdown value, the first step to perform the robust method is to select a breakdown value. In this study, we use two breakdown values; one 20% and another is 50% which is the maximal breakdown value. The result of the LTS robust method with 20% breakdown value is in the equation 3 and with the maximal (50%) breakdown value in the equation 4. Since the LTS robust method with the maximal breakdown value is too penalized, we use 20% breakdown value to detect the influential outliers and the plots are in figure 7 and 8.

$$Fuel = 5.614e^{02} - 5.461Income - 7.828Tax + 4.045e^{-01}Dlic \tag{3}$$

$$Fuel = 7.389e^{02} - 5.161Income - 6.820Tax \tag{4}$$



(g) Figure 7: Residuals vs Index (20%)

(h) Figure 8: Diagnostic plot (20%)

According to the plots of residuals vs index (figure 7) and the regression diagnostic plot (figure 8), the vertical outliers of the LTS robust method are observation 12, 10, 33, and 51. Also, the bad leverage points of the robust method are observation 2 and 9. Since observation 2 and 51 were identified as influential outliers in the previous section, we may expect to more accurate statistical analysis after remove the two observations.

5 Conclusion

Due to various statistical analysis, there are possible to have several different statistical outcomes depending on analyzing methods and restrictions. However, as we have seen in this study, critical influential outliers can be identified regardless of statistical methods.