Investigating the Relationship between Fuel Efficiency and Other Car Characteristics

STAT 5330 Project 1

Team 6. Wenjing Che, Zachary Heese, Siyuan Liu, Ze Wang

1. Executive Summary

The goal of this analysis was to formulate a model that could predict the fuel efficiency of a car given other characteristics. A training dataset where the fuel efficiency of the car was known was used to develop a variety of models derived from different linear regression methods. These models were generated using multiple linear regression, best subset selection, stepwise selection, shrinkage methods, splines. The models were compared for efficiency and adequacy of fit based on cross-validation errors, and a best model was selected. The best model was a variation of multiple linear regression that included some higher order relationships between some variables and fuel efficiency smoothly. Then cross-validation was conducted again to do model assessment. The best model with the best parameter was then used to predict what the fuel efficiency was for a set of car models in a test dataset, in which the fuel efficiency was unknown.

2. Introduction

The U.S. Department of Energy and Environmental Protection Agency collaborated to collect and present data on the fuel economy of various passenger car and truck models. One of the datasets they publish includes the fuel efficiency (*FE*) of car models along with other characteristics of each model. These characteristics include information on the engine displacement (*EngDispl*), number of cylinders (*NumCyl*), transmission (*Transmission*), air aspiration method (*AirAspirationMethod*), number of gears (*NumGears*), transmission lockup (*TransLockup*), transmission creeper gear (*TransCreeperGear*), drivetrain (*DriveDesc*), number of intake/exhaust valves (I*ntakeValvePerCyl, ExhaustValvesPerCyl*) , carline class (*CarlineClassDresc*), valve timing (*VarValveTiming*), valve lift (*ValValveLifting*), and model year(*ModelYear*).

Upon initial inspection of this dataset, it was clear that there were some entries which were not suitable for our analysis. Entries were excluded that had zero intake or exhaust valves. It was clear this was a data collection error on the EPA and DoE's part, because it would be impossible for a model to have zero intake or exhaust valves and function as an automobile. In addition, there was one entry in the training dataset in which the number of cylinders equaled 16, and there were no entries in the test dataset in which the number of cylinders equaled 16. As a result, we removed this entry since it would not contribute positively to our model predicting FE in the testing dataset.

This analysis attempted to find a model which could accurately predict the fuel efficiency of a model of a car given the other characteristics collected. The importance of this analysis was pretty evident, as conclusions drawn from this analysis could be used to help formulate future car designs with fuel efficiency in mind, especially as finite resources such as gasoline become more scarce.

3. Approach to the Problem/Methods

3.1 Logarithmic Transformation

In addition, a logarithmic transformation of the dependent variable, fuel efficiency, was used for the duration of the analysis.

3.2 Multivariate Linear Regression

An initial multiple linear model was created utilizing all the variables in the dataset. Multivariate linear regression is a linear approach for modeling the relationship between a scalar response variable y and more than one predictors (3.1). Multivariate linear model was used as the basic model in this research.

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1}x_1 + \widehat{\beta_2}x_2 + \widehat{\beta_3}x_3 + \cdots \qquad (3.1)$$

There are two reasons why linear regression is not satisfied. The first is prediction accuracy. The least squares estimates often have low bias but large variance. Prediction accuracy can sometimes be improved by shrinking or setting some coefficients to zero estimates. The second reason is interpretation. With a large number of predictors, a smaller subset that exhibit the strongest effect would be desired.[1]

3.3 Hybrid Stepwise Selection

To address potential problems raised by multivariate linear regression, another model was created using stepwise selection. A hybrid iteration method selects a variable of the maximum

[1] Sanjeev Kulkarni, Gilbert Harman, *An Elementary Introduction to Statistical Learning Theory* Second Edition, Page 57.

classification ability in accordance with both direction selection and incorporates the predictor into the model by stepwise increase. This hybrid stepwise method grouped dummy variables together by category in that when a category was determined to be insignificant, all the dummy variables related to that category were removed, and the same for when a category was determined necessary to be added to the equation.

3.4 Best Subset Selection

The other way of subset selection used was best-subset selection. Best subset regression finds for each the subset of different size that performs the best. However, for this method, dummy variables were treated as their own categories, so individual dummy variables could be added and removed. In addition, the cross-validation process and the best subset selection itself were very computationally taxing.

3.5 Shrinkage

Another two models were generated using the shrinkage method. The shrinkage method grouped all the dummy variables together and shrieked the regression coefficients by imposing a penalty on their size (see ***Appendix***).

The models were constructed by constraining or regularizing the coefficient estimates, or equivalently, shrinking the coefficient estimates towards zero. Two approaches used in this project were Ridge Regression and Lasso Regression.

3.6 Polynomial Regression

Once the generic linear model was determined to be the best model, individual terms were evaluated to see if there were curvilinear relationships between the predictor and response variables. ANOVA was used here to determine the power of higher-order terms. The higher-order relationships were added to the model to generate a final subset of independent variables.

3.7 Spline

Spline method was an extension of the polynomial regression. A spline is a special function defined piecewise by polynomials. Different polynomial regressions were implemented in according region of the predictor variable divided by knots to have a smoother fitting lines.

Polynomial Splines and Natural Splines were used in this project. Compared to polynomial splines, more constraints were added to boundaries which was more smoother.

3.8 Cross-Validation

In this project, the 10-fold cross-validation was implemented to each method mentioned above. This method directly estimates the expected prediction error[2]. Train dataset was divided into 10 roughly equal-sized folds. The kth part was fitted to the model to the other 9 parts of the data, prediction error (MSE_k) of the fitted model was obtained when the kth part of the data was predicted. We do this for k = 1, 2, . . . ,10 and CV.RMSE was obtained by the equation (3.4).

$$RMSE.CV_{10} = \sqrt{\frac{1}{10}\sum_{k=1}^{10} MSE_k} \tag{3.4}$$

Finally the method with the least RMSE.CV was selected as the best one.
Cross-validation was also conducted to determine parameter for the best method.

4. Results and Findings

4.1 Modeling Process

In the initial analysis, it was determined that the response variable, fuel efficiency (*FE*), was not normally distributed. A histogram showed the data was skewed right, and the Q-Q plot confirmed that this data set was not normally distributed, since it tailed away from the line at either end. The boxplot for *FE* showed the same skewness as well (**Figure 1)**. Since most linear modeling techniques used in this analysis assumed that the response variable was normally distributed, a transformation was employed to ensure normality. As shown in **Figure 2**, it was clear that taking the log of the dependent variable produced a normally distributed dependent variable. All linear regression techniques used in this analysis aimed to create a model for ***log(FE)***.

A multivariate model was the initial modeling attempt. 10-fold cross-validation was conducted to estimate the prediction error and the result was 3.253. This model included 14 predictors in which 9 of them were categorical variables. Therefore this model ended up with 48 predictors. Because of the sheer amount of predictive variables included, we attempted to find a model that used fewer

[2] Sanjeev Kulkarni, Gilbert Harman, *An Elementary Introduction to Statistical Learning Theory* Second Edition, Page 241

predictive variables. Subset selection and shrinkage methods were tried to improve the simple multivariate regression model.

For the subset selection, best subset and stepwise selection were utilized. 10-fold cross-validation was carried out again to obtain the estimate for prediction under these two model. Hybrid stepwise selection method was carried out to reduce the number of predictors. The cross-validation RMSE for hybrid stepwise selection was 3.264. Two variables were removed (*NumGears* and *TransCreeperGear)* after applying hybrid stepwise selection method to the train dataset.

Further attempts to shrink the model utilized the lasso and ridge regression techniques. These techniques did not attempt to remove variables from the equation, but rather tried to reduce the estimated coefficients for these variables by imposing a penalty for their size. The lasso technique, which used the absolute value of the coefficients, yielded a cross-validation RMSE of 3.2520. The ridge technique, which used the squared values, yielded a cross-validation RMSE of 3.3116. Taking into account the complexity of the models and the cross-validation RMSE values, the hybrid stepwise selection method was selected for further analysis.

Using the model generated by the hybrid stepwise selection method, higher-order relationships between log(FE) and the quantitative variables were investigated. Two variables that appeared to have a higher-order relationship with log(FE) were Engine Displacement (*EngDispl)* and the Number of Cylinders (*NumCyl*). A quick glance at each scatterplot confirmed that there was a distinct higher order relationship between each variable and *log(FE)* (**Figures 3, Figure 4**).

For each variable, in order to determine the power of the polynomial terms to be added, five models were created, each with a higher power polynomial term introduced in order. Here we assumed that "*A*" represented all other predictors expect for EngDispl and NumCyl. Then the following five models were generated for the purpose of power detection

Model 1. $Log(FE)\sim A+EngDispl+NumCyl$

Model 2. $Log(FE)\sim A+EngDispl+EngDispl^2+NumCyl$

Model 3. $Log(FE \sim A+EngDispl+EngDispl^2+EngDispl^3+NumCyl$

Model 4. $Log(FE)\sim A+EngDispl+EngDispl^2+EngDispl^3+EngDispl^4+NumCyl$

Model 5. $Log(FE) \sim A + EngDispl + EngDispl^2 + EngDispl^3 + EngDispl^4 + EngDispl^5 + NumCyl$

These models were compared using an ANOVA test. For Engine Displacement *(EngDispl)*, these ANOVA results are displayed in **Table 1.** A second order polynomial term was added to the equation because an F-Test for significance showed that the model with the second order term was significant. **Table 2** summarized the ANOVA test results for the Number of Cylinders (*NumCyl*). Similarly, because the third-order term was not significant, a second order polynomial term was added to the equation.

After investigating higher-order relationships, a polynomial regression was conducted by adding the quadratic terms of Engine Displacement (*EngDispl*) and Number of Cylinders (*NumCyl*). This polynomial model yielded a cross-validation RMSE of 3.1739, which was obviously lower than previous models. Since polynomials are limited by their global nature, splines that allowed for local polynomial representations was considered in the further analysis.

A spline was applied to this model that included cubic terms of *NumCyl* and *EngDispl*. Polynomial splines and natural splines were tried on the model and the results were compared. The polynomial splines, with 4 knots in each fifth quantile, had an abnormally large and unstable cross-validation RMSE. The natural splines, with the same knot distribution, had a cross-validation RMSE of 3.107, which was the lowest found amongst all techniques.

The comparison among CV.RMSE for each method was shown in **Table 3**. This allowed us to determine that the natural spline model should be used in the prediction process. Then cross-validation was used for the hunting of the best knots number. Uniform knots of 3 to 7 were candidates and 4 knots turned out to be the best choice with the least cross-validation RMSE (**Figure 5**).

4.2 Model Check

In order to predict values using this model generated, assumptions must be checked to make sure the model was an accurate fit for the data set. It was clear that there was a linear relationship between the model and fuel efficiency. Looking at a residual plot, as seen in **Figure 6**, it was clear that the residuals were independent and homoscedastic. Finally, looking at a Q-Q plot of the residuals, as shown in **Figure 7**, it was clear the residuals were normally distributed for the most part, which made it acceptable to predict future values using this model.

This model was then used to predict values of fuel efficiency in the testing dataset, where the fuel efficiency was unknown. These FE values were written to a csv, which was prepared for final submission.

4.3 Conclusion

Since the lowest cross-validation RMSE, the natural spline method with four uniform knots was selected as the final model to predict *log(FE)* in the test dataset. The predicted values were transformed into FE using the exponential function and saved in the submission.csv.

5. Approaches that didn't work so well

5.1 Best Subset Selection

Since the train dataset included 48 predictors, it would take more than five hours to get the final cross-validation result. Obviously, the best subset selection was not a proper approach to investigate of the relationship between fuel efficiency and other car characteristics in this train dataset. Even though the smallest RMSE was obtained finally (**Figure 8**), it was 3.316 and the selected model included all variables. Therefore it was postponed with no further exploration.

6. References

1. James, Gareth, et al. *An introduction to statistical learning: with applications in R*. Springer, 2017.

2. An Elementary Introduction to Statistical Learning Theory

Sanjeev Kulkarni, Gilbert Harman *An Elementary Introduction to Statistical Learning Theory Second Edition*

3. U.S. Department of Energy and U.S. Environmental Protection Agency. *The official U.S. government source for fuel economy information.* http://fueleconomy.gov

Appendix

The Ridge

Ridge regression estimated coefficients by minimizing a penalized residual sum of squares. Using a tuning parameter λ and $\lambda \sum_{j=1}^{p} \beta_j^2$ shrinkage penalty to shrinkage $\beta_0, \dots \beta_p$.

$$\sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = RSS + \lambda \sum_{j=1}^{p} \beta_j^2$$

The Lasso

By replacing the β_j^2 by β_j, when λ is big enough, some of β_i would be forced to be zero, therefore Lasso could also select variables.

$$\sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j| = RSS + \lambda \sum_{j=1}^{p} |\beta_j|$$

Table 1. ANOVA for *EngDispl*

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)	
1107	8.5363					
1106	8.3823	1	0.153953	20.4075	6.932e-06	***
1105	8.3768	1	0.005501	0.7293	0.39331	
1104	8.3436	1	0.033237	4.4057	0.03605	*
1103	8.3285	1	0.015052	1.9953	0.15807	

Table 2. ANOVA for *NumCyl*

Res.Df	RSS	Df	Sumof Sq	F	Pr(>F)	
1106	7.8266					
1105	7.706	1	0.120562	17.4718	3.147e-05	***
1104	7.706	1	0.000027	0.0040	0.9498065	
1103	7.6198	1	0.086144	12.4839	0.0004275	***
1102	7.5977	1	0.022168	3.2126	0.0733490	.

Table 3. Comparison among models

Technique	Cross Validation RMSE
Multivariate Linear Regression	3.2533
Hybrid Stepwise Selection	3.2638
Best-Subset Selection	3.3100
Lasso Regression	3.2500
Ridge Regression	3.3116
Multivariate w/ Polynomial Terms	3.1739
Polynomial Splines	Abnormally large and unstable
Natural Cubic Splines	**3.1074**

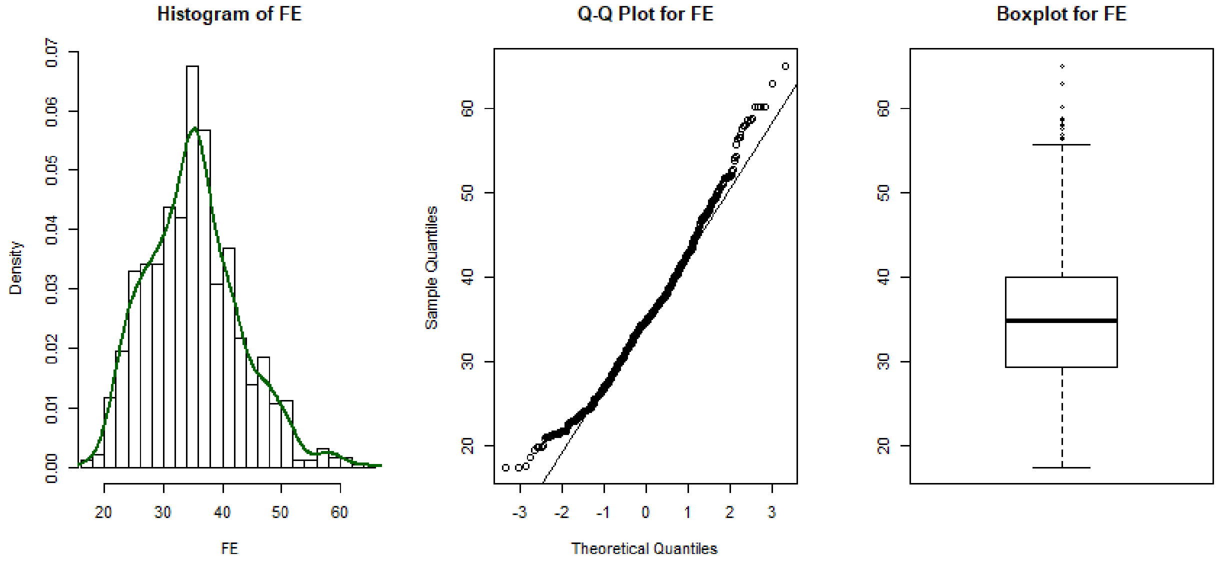

Figure 1. Histogram, Q-Q, and Boxplots of Fuel Efficiency

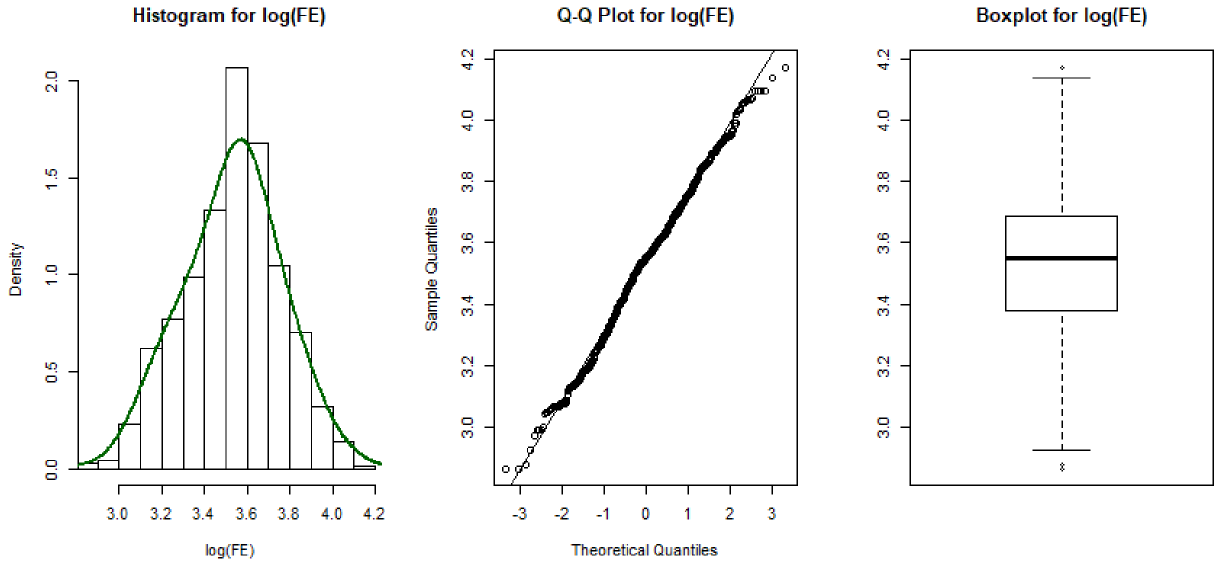

Figure 2. Histogram and Q-Q plots of *log(FE)*

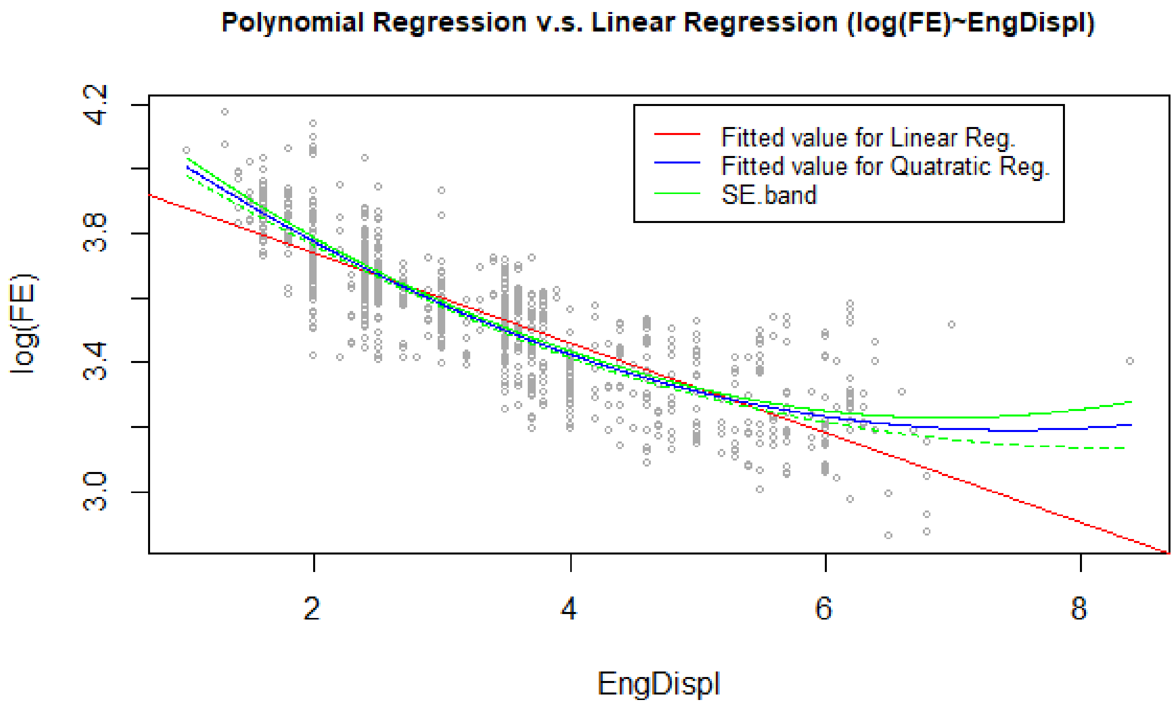

Figure 3: Scatterplot of log(FE) vs Engine Displacement

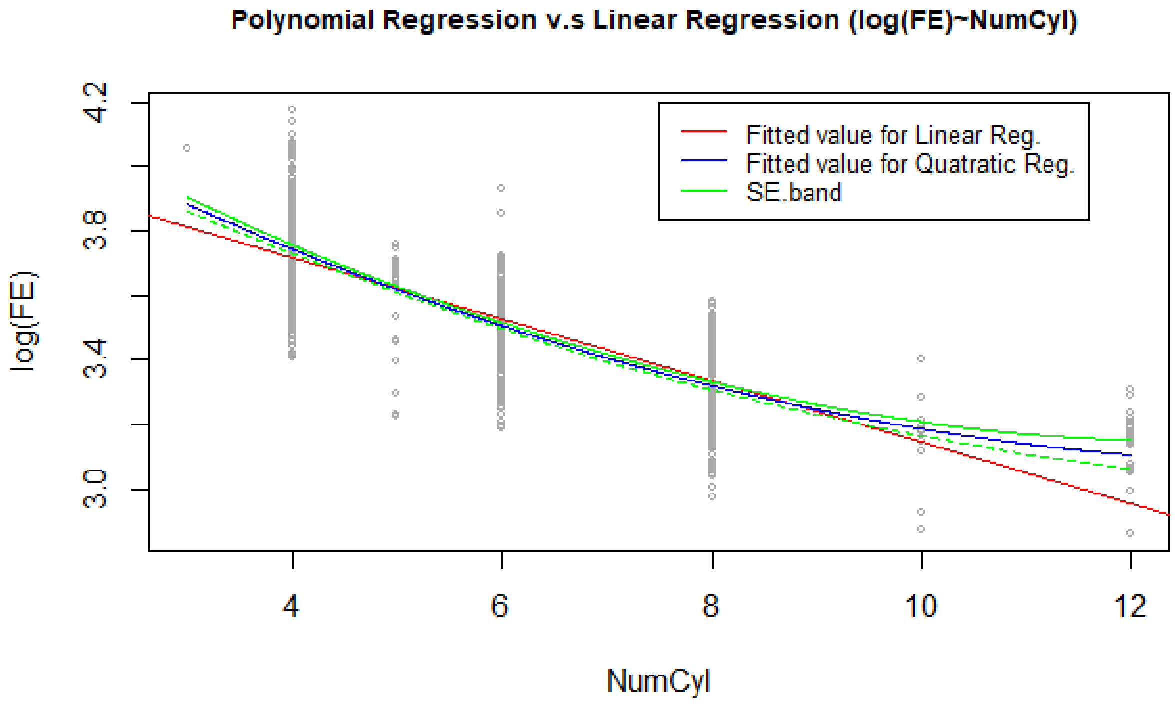

Figure 4: Scatterplot of log(FE) vs Number of Cylinders

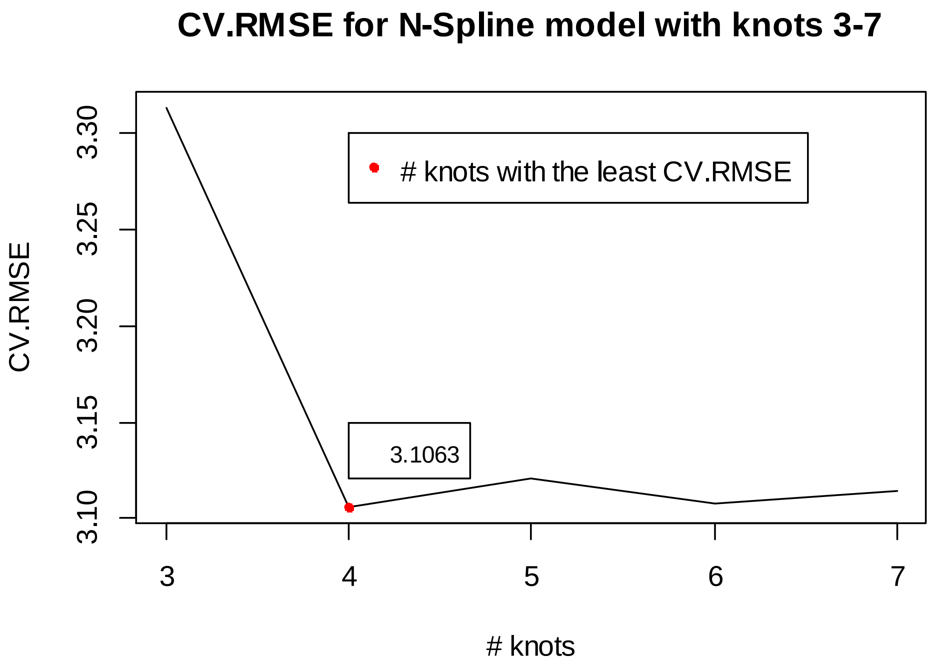

Figure 5. Best # knots hunting

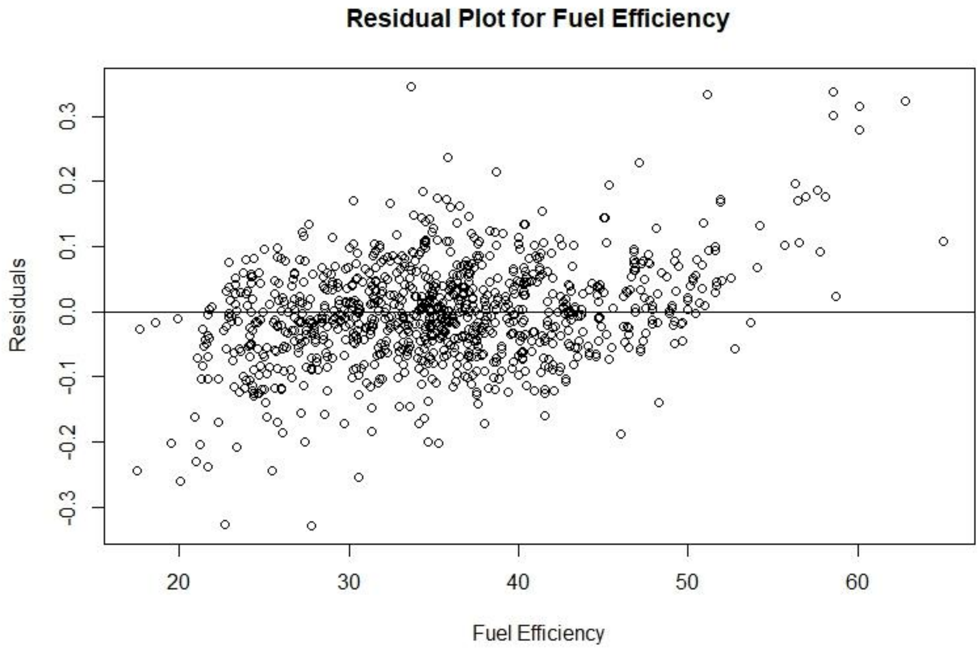

Figure 6. Residual Plot for Spline Model

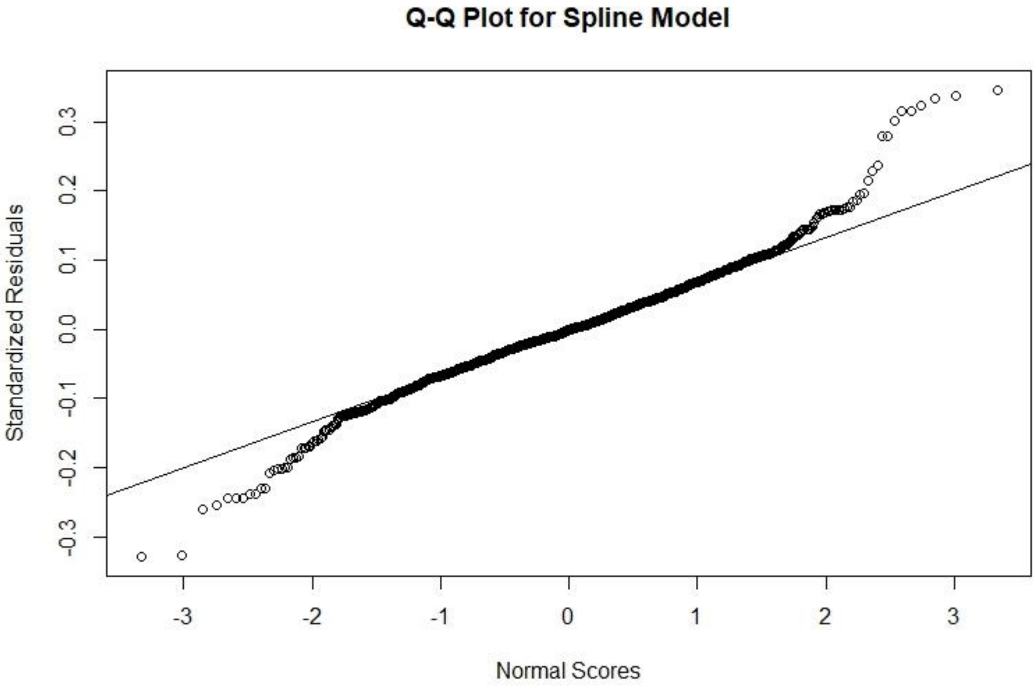

Figure 7. Q-Q Plot for Spline Model Fitted Residuals

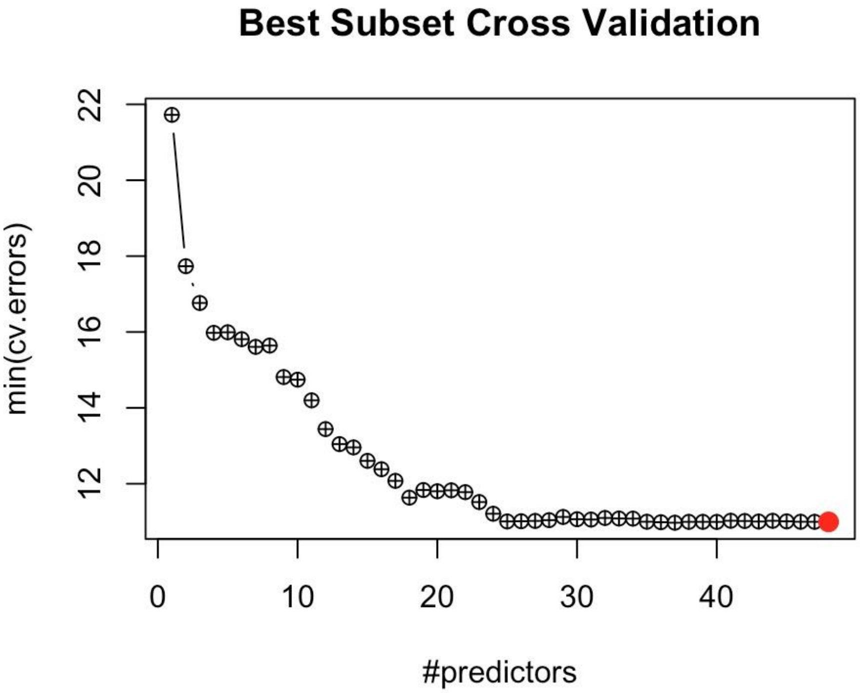

Figure 8. Best Subset Cross Validation Results