ST 516 Midterm Project-Spring 2020

By
Aditi Khamkar
Yu Deng
Yash Chaturvedi
Suryansh Purwar
Linfeng Wu

submitted to
Dr. Dan Harris
Department of Statistics
ST 516 - Experimental Statistics For Engineers II

1. Executive Summary

The objective of this study report is to understand the effect of various properties of concrete recipes and perform an independent analysis to build a model or set of models to predict compressive strength from these mixture components. It is believed that the compressive strength is a highly non-linear relationship between age and ingredients. We chose three statistical models to get the response of the components which influence the strength of the concrete. The methods are:

- Multivariate Linear Regression
- Multivariate Regression with only interaction terms
- Multivariate Regression with Second Order and Interaction terms

Further, in these models, we used different regression techniques. The techniques which we used are:

- Box-cox transformation of the Response
- Reducing the model by removing insignificant variables
- Best Subset Selection
- Ridge Regression
- Lasso Regression
- Tree-Based Models with boosting

With final discussion on observed values, we chose reduced second order model with 34 predictors and optimized by box-cox transformed response, supported by regularization techniques. The R-squared value came out to be **0.8091** with a mean squared error of **5.066599**. Compressive Strength of the concrete increases as the age and proportion of contents increase. The most important predictor that influences the compressive strength is the interaction of amount of Cement and Age.

2. Introduction

The large design build construction firm is keen on understanding the various properties of concrete recipes and predict total compressive strength. After a detailed analysis, the company has already had observational data from the analysis of concrete strength and want to perform an independent analysis using the data to build a model or set of models to predict compressive strength from the mixture components.

The company specifically wants to follow 3 working objectives: First, to prioritize the ability to accurately predict the compressive strength of concrete. Second, to explicitly state which components influence the strength along with an estimate of the relationship between those components and strength. Finally, to specify any important optimal values for the various components of the mixtures to optimize concrete strength.

3. Data

Aiming to help a construction firm understand and predict the effect of each concrete component on the product compressive strength, studies have been conducted on the previous production and a set of observational data has been achieved. The input observation data includes seven different concrete components (cement, fly ash, etc.) and the specific age of the concrete in days, only output data in this set is the compressive strength which was determined from the laboratory. All data recorded are quantitative and none of the attribute value was missed. The data set consists of 1030 observations of the nine variables in total, the correlation scatterplot and the colored map of raw data are shown in figures 1 and 2 separately below. It can be observed from these two plots that no obvious correlation exists.

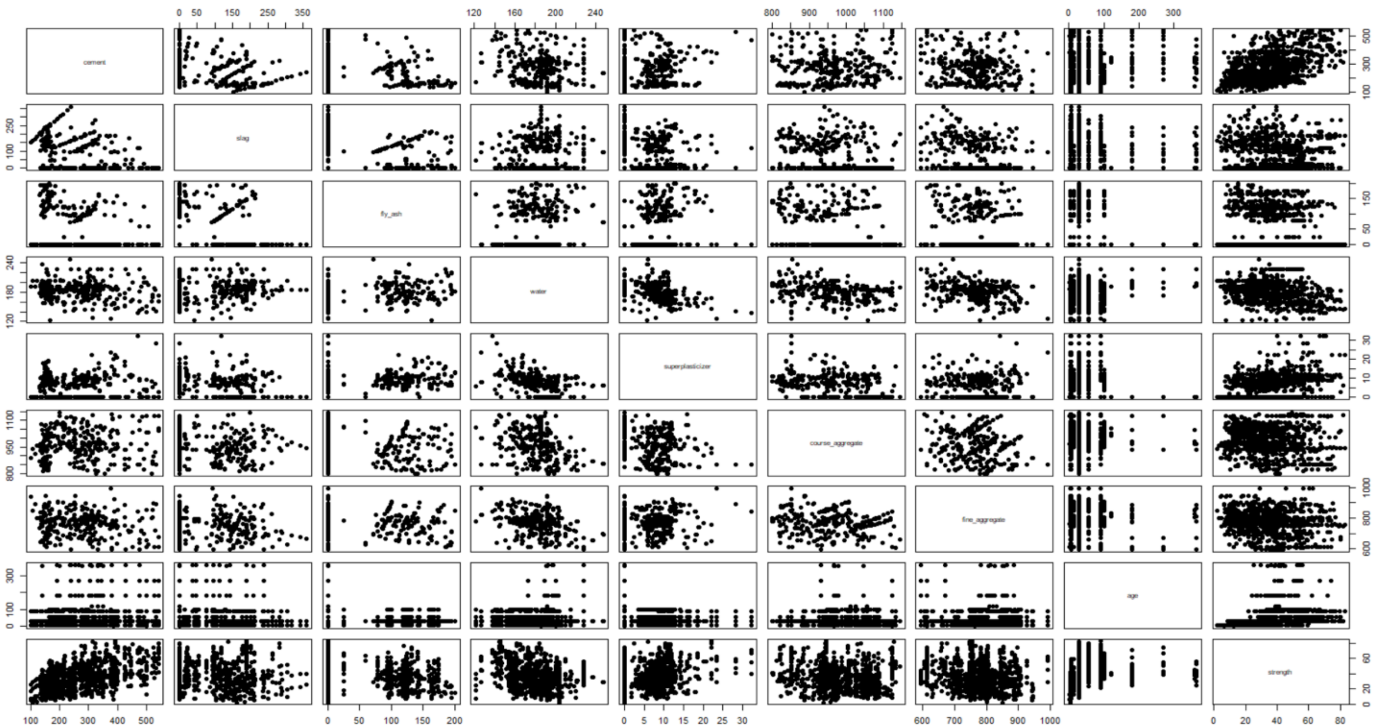

Figure: Pairwise Scatter Plot of the data

4. Methods

We started the model establishment with a **simple additive model** where all predictors were included. As a result, only coarse aggregate and fine aggregate were insignificant. Thus, we reduced those two components and ran the linear regression again. Based on the **Box-cox transform**, we calculated the lambda value for a simple additive model at **0.707**, then refitted the simple additive model with the transformed response. Principal components regression was also adopted in our analysis. For the simple additive model, **8 principal components** were selected and an R-squared value of **0.6155** was achieved. Besides, we applied a tree-based regression method. The tree-based regression and pruned tree regression has cement, water, age, and slag variables as important variables; however, it does not give a good fit. **Bagging** improves the fit and explains the % Var of **92.27** with eight variables. **Random forest regression** with 5 variables results in % Var explained is **92.45** and important variables are the same as bagging. Boosting resulted in a better fit with age and cement being important parameters.

Subsequently, we add **interaction terms** in the model, the R square value increased significantly from **0.6125 to 0.8279**. However, the residual vs. fitted plot was not satisfying enough. Thus, we reduced the aggregate predictors again. The plot of the interaction reduced additive model indicates that the model is more precise now at a cost of decreased R-square value.

After that, we build a **completed second order model** to test the data. While the R-square value and analysis plots show that the performance of the completed second order model doesn't exceed that of the interaction model. PCR analysis was added for this model and several principal component numbers were tested. For model constructed with **20 principal components**, the R-square value of **0.78** was obtained which is even lower than the ordinary least square result. However, if all **44 principal components** were used to construct the model it will give the R-squared of **0.8106** which is slightly better than the original one. The basic model and pruned regression tree model result indicates that they are likely to overfit the data, leading to the poor test set performance visible in figure 7.3. **Bagging** has a mean of squared residuals as **21.79855** and the percentage of variance explained is **92.18**. **Random forest regression** for 12 predictor variables explained % Var of **92.37**, on the other hand, all 44 predictor variables resulted in a % Var of 92.13. **Boosting** performed better by fitting well among all tree-based regressions which is shown in figure 7.4. The most important variable is cement-age interaction. Water-finite aggregate and slag-superplasticizer interactions are also better than other variables (see figure 7.5).s

5. Results

Table.1: Shows R-squared values and Mean Square Errors of the concrete strength of performed models

Model	**Technique**	**R-Squared**	**MSE**
Simple Linear Model	Full	0.6125	107.1972
	Reduced	0.6118	107.6148
	Box-Cox	0.6083	13.63576
	Best-Subset	0.6155	109.0982
	Ridge/Lasso	0.6068/0.6067	109.73/109.7466
	Principal Component Regression	0.6125	
	Bagging	0.9227	
	Random Forest	0.9245	
Interaction Model	Full	0.8279	36.78846
	Reduced	0.768	60.73244
	Box-Cox	0.8294	45.33518
	Best-Subset	0.7567	76.276
	Ridge/Lasso	0.7304/0.7308	75.253/75.133
Second Order Model	Full	0.8021	52.8096
	Reduced	0.7819	59.22072
	Box-Cox	0.8026	5.002225
	Best-Subset	0.8105	59.70802
	Ridge/Lasso	0.762/0.780	61.687/61.381
	Principal Component Regression	0.8106	
	Bagging	0.9227	
	Random Forest	0.9245	
	Boosting	0.9911	
	Reduced (Optimized)	**0.8091**	**5.066599**

Results convey that few of the models have a great R-squared values but that does not mean that it is the best model. The best model found to be the second order reduced model optimized by box-cox transformation and this is supported by tree-based model.

The Final Estimated Equation Came out to be:

$$\begin{aligned}
&\boldsymbol{Compressive\ strength} \\
&= -0.0002023Age^2 - 0.0001223Fine\ Aggregate^2 - 0.009059Superplasticizer^2 \\
&- 0.002432Water^2 + 0.0001206Fly\ Ash^2 - 0.00004316Cement^2 \\
&+ 0.00009306(Fine\ Aggregate * Age) \\
&- 0.00007158(Course\ Aggregate * Fine\ Aggregate) \\
&+ 0.0001308\ (Superplasticizer * Age) - 0.003091(Superplasticizer * Fine\ Aggregate) \\
&- 0.002405(Superplasticizer * Course\ Aggregate) \\
&- 0.001204(Water * Fine\ Aggregate) - 0.01030(Water * Course\ Aggregate) \\
&- 0.007739(Water * Superplasticizer) + 0.0002386(Fly\ ash * Age) \\
&+ 0.00007653(Fly\ ash * Fine\ Aggregate) + 0.0001009(Fly\ ash * Course\ Aggregate) \\
&- 0.005520(Fly\ ash * Superplasticizer) - 0.0009725(Fly\ ash * Water) \\
&+ 0.0001302(Slag * Age) - 0.002940(Slag * Superplasticizer) \\
&- 0.0007437(Slag * Water) + 0.0002162(Slag * Fly\ ash) + 0.00004632(Cement * Age) \\
&- 0.00007131(Cement * Fine\ Aggregate) - 0.003010(Cement * Superplasticizer) \\
&- 0.0009266(Cement * Water) + 0.0001176(Cement * Fly\ ash) + 0.5084Fine\ Aggregate \\
&+ 0.2501Course\ Aggregate + 7.7745Superplasticizer + 3.170Water + 0.1649Slag \\
&+ 0.2934Cement - 663.2
\end{aligned}$$

All resulting and supported values have been provided in the table above. The plots below depict the fitting of the models.

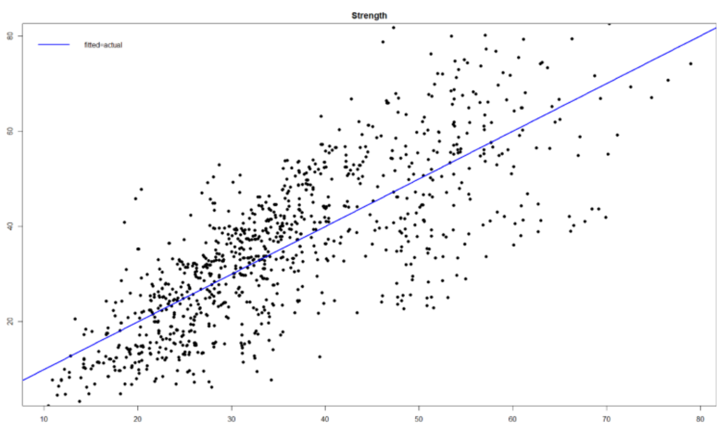

Figure: Casual Regression 1st Order

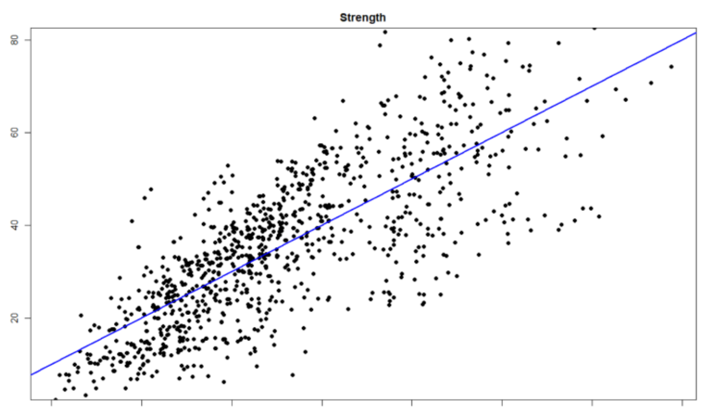

Figure: Reduced 1st order

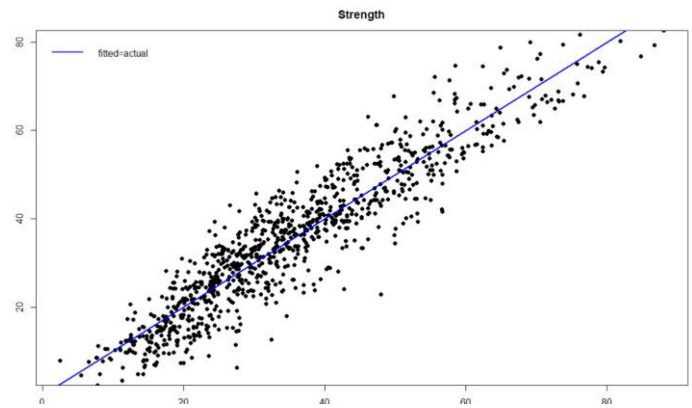

Figure: Interaction Model Terms

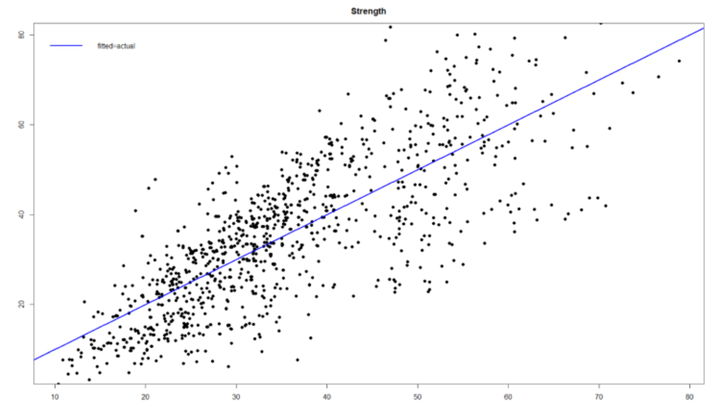

Figure: Reduced Interaction Model

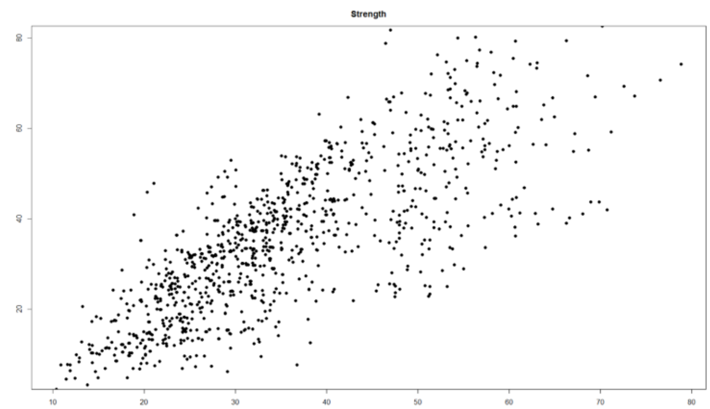

Figure: Complete Second Order

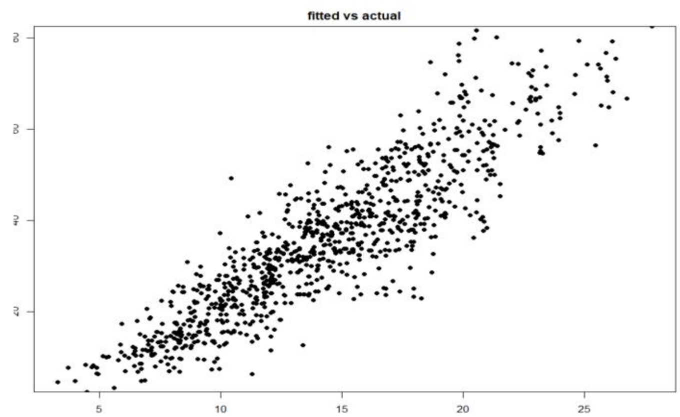

Figure: Reduced Second Order Model

6. **Conclusion**

The project involved the analysis of concrete data to accurately predict the compressive strength of concrete as a function of eight variables and to know which components influence the strength.

The Reduced Second Order Model optimized with box-cox has accurate prediction and interpretability of statistical inference among all regression models. This model can completely predict the values of the coefficients, with reasonably high prediction accuracy. Moreover, the low VIF values for the coefficient estimates indicated that we could rely on our coefficient estimates. The boosting method supported the result from Reduced Second Order Model. To summarize, the R-squared value is **0.8091** with the mean squared error as **5.066599.**

7. Appendix

Figure 7.1 Correlation Matrix

Multivariate

Correlations

	cement	slag	fly_ash	water	superplasticizer	course_aggregate	fine_aggregate	age	strength
cement	1.0000	-0.2752	-0.3975	-0.0816	0.0924	-0.1093	-0.2227	0.0819	0.4978
slag	-0.2752	1.0000	-0.3236	0.1073	0.0433	-0.2840	-0.2816	-0.0442	0.1348
fly_ash	-0.3975	-0.3236	1.0000	-0.2570	0.3775	-0.0100	0.0791	-0.1544	-0.1058
water	-0.0816	0.1073	-0.2570	1.0000	-0.6575	-0.1823	-0.4507	0.2776	-0.2896
superplasticizer	0.0924	0.0433	0.3775	-0.6575	1.0000	-0.2660	0.2227	-0.1927	0.3661
course_aggregate	-0.1093	-0.2840	-0.0100	-0.1823	-0.2660	1.0000	-0.1785	-0.0030	-0.1649
fine_aggregate	-0.2227	-0.2816	0.0791	-0.4507	0.2227	-0.1785	1.0000	-0.1561	-0.1672
age	0.0819	-0.0442	-0.1544	0.2776	-0.1927	-0.0030	-0.1561	1.0000	0.3289
strength	0.4978	0.1348	-0.1058	-0.2896	0.3661	-0.1649	-0.1672	0.3289	1.0000

Figure 7.2 Correlation Color Map

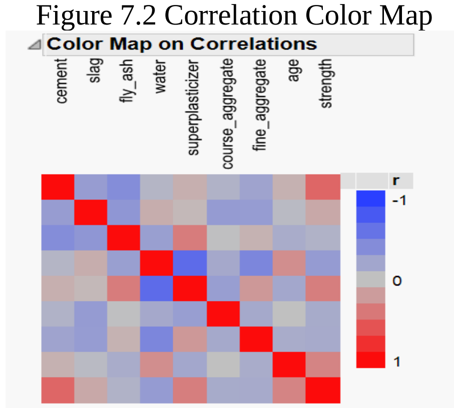

Figure 7.3 OLS and Regression Tree Prediction Plot

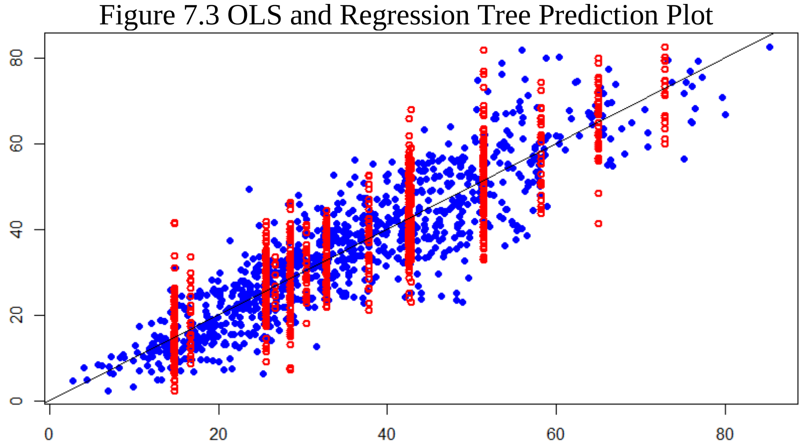

Figure 7.4 Boosting Prediction Plot

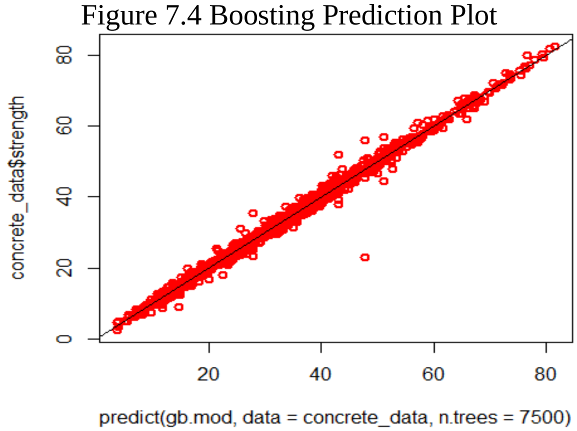

Figure 7.5 Summary of Relative Influential Variable by Boosting

```
> summary(gb.mod,cBars=10)
                                var    rel.inf
ce_a                           ce_a 44.3030663
w_fi                           w_fi  9.0562223
sl_su                         sl_su  7.5622636
ce_su                         ce_su  5.4332350
w_co                           w_co  3.5056111
ce_fi                         ce_fi  3.0803086
cement                       cement  2.9966399
sl_a                           sl_a  2.9201889
water                         water  2.1477434
ce_sl                         ce_sl  2.1027542
su_a                           su_a  1.7231007
ce_co                         ce_co  1.7119581
fine_aggregate       fine_aggregate  1.4948820
co_fi                         co_fi  1.3332332
w_a                             w_a  1.2086229
fi_a                           fi_a  1.1337575
course_aggregate course_aggregate  1.0913220
co_a                           co_a  0.9187646
ce_w                           ce_w  0.8590526
slag                           slag  0.6753212
fl_su                         fl_su  0.5198301
fly_ash                     fly_ash  0.4714618
fl_a                           fl_a  0.4277949
sl_fi                         sl_fi  0.3761782
sl_fl                         sl_fl  0.3585060
w_su                           w_su  0.3528702
ce_fl                         ce_fl  0.3291086
superplasticizer superplasticizer  0.3237850
sl_co                         sl_co  0.2406845
su_fi                         su_fi  0.2389116
sl_w                           sl_w  0.2347933
fl_co                         fl_co  0.2038323
su_co                         su_co  0.1917737
fl_w                           fl_w  0.1735729
fl_fi                         fl_fi  0.1539099
age                             age  0.1449390
```

Figure 7.6 Casual Linear Regression

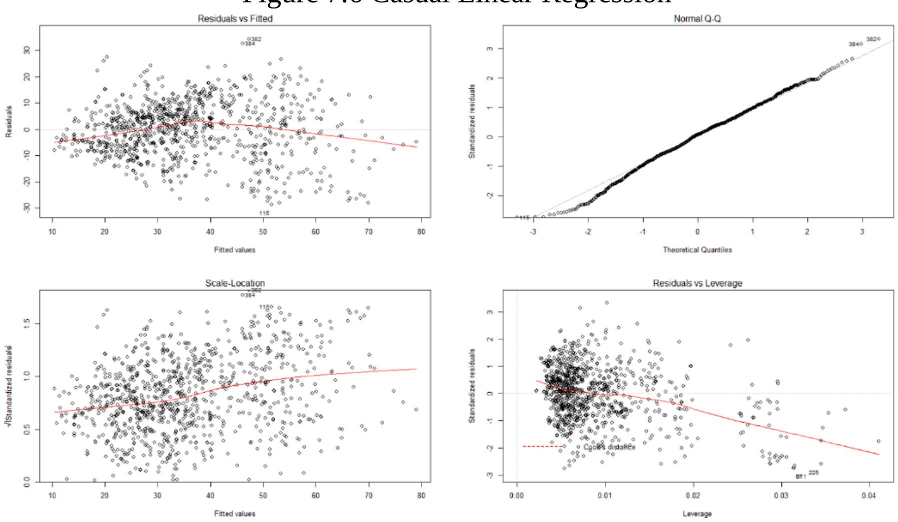

Figure 7.7 Reduced Linear Regression

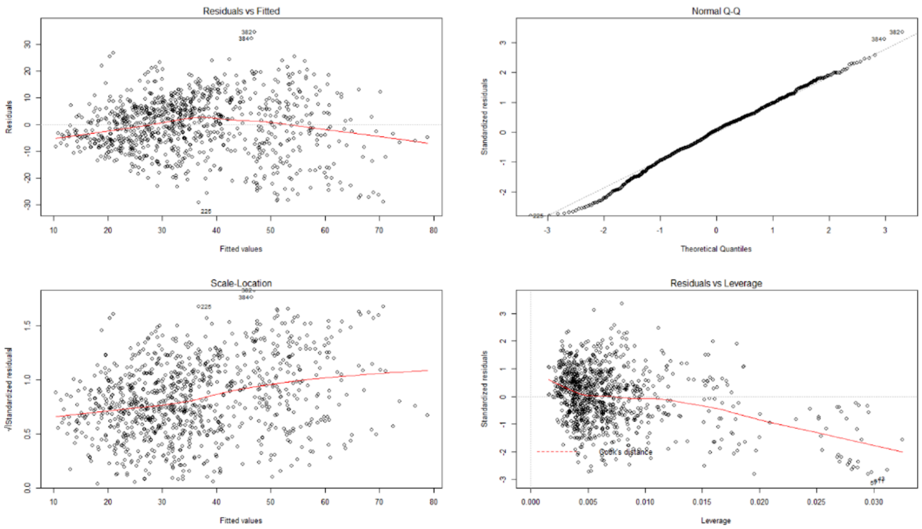

Figure 7.8 Linear Box-Cox Optimized

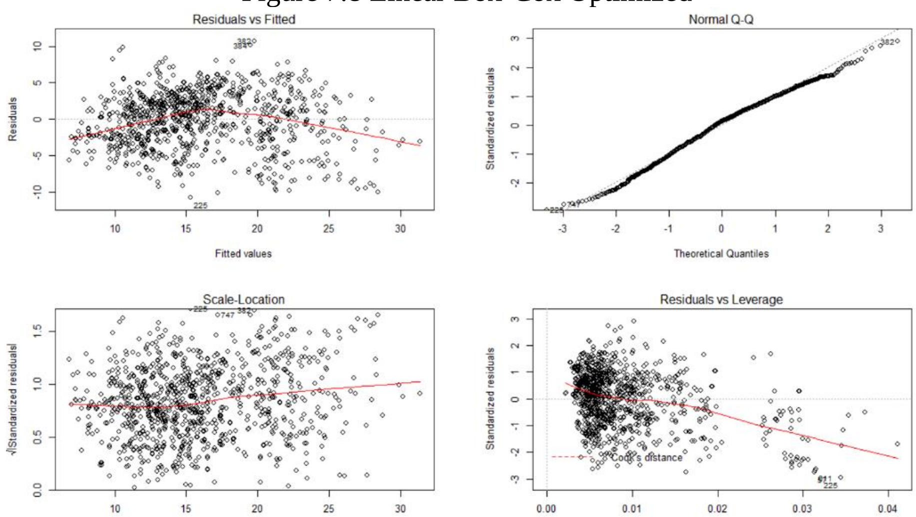

Figure 7.9 OLS with Interaction Terms Only

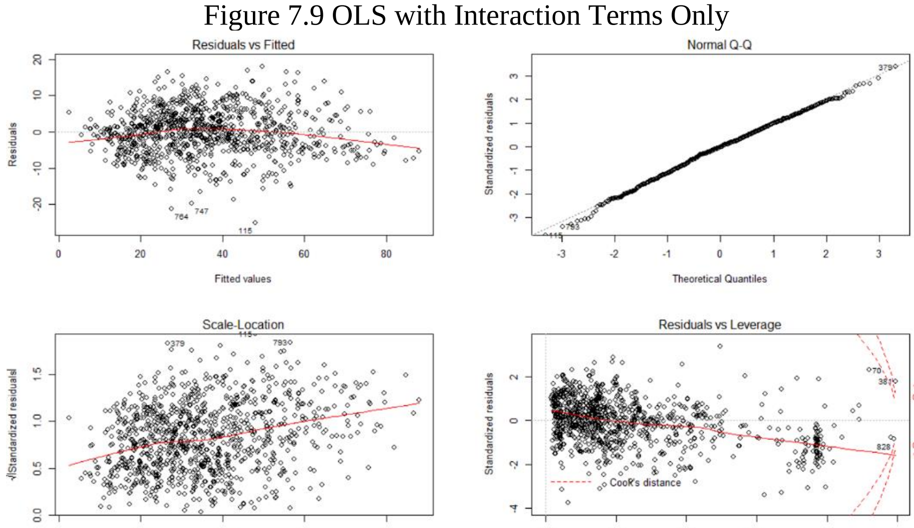

Figure 7.9 Reduced Interaction Only Model

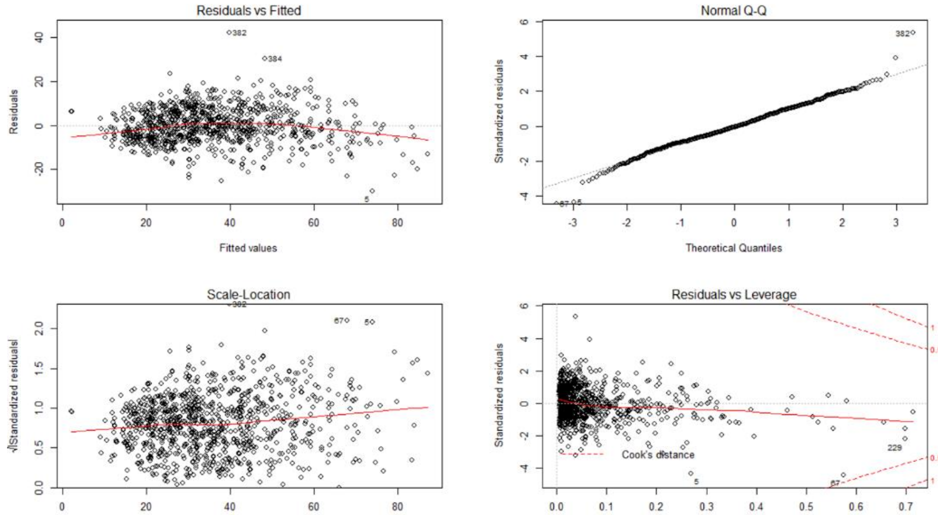

Figure 7.10 Second Order Model