Stat 512 Project Cover Page

1. Project Topic: Predicting College Basketball Win Rate 2. Group number: **4**

3. List of group members: **Ryan Newman, Jerry Huang, Nikhil Venkatachalam, Parth Gandhi, Le Rui Tay**

4. Project YouTube link: https://youtu.be/8UHvg9MwNHw

5. Project background introduction (why this is an important question, what has been done on the question, what are your major research questions in the project, etc.):

Our research questions are about factors that affect win rate for college basketball teams, which is important for teams that want to win more. We covered 5 domains: Conference location, Offensive Statistics, Defensive Statistics, Type of Shots, and quality of school respectively. Ryan: My question is to determine whether conference has a significant interaction effect with ADJOE and whether conference has significant interaction effect with ADJDE. Jerry: EFG_O (Effective field goal rate) has the same impact as the ORB (offensive rebound rate) on the win rate. Nikhil: Turnover percentage committed has a different impact than effective field goal percentage allowed on win rate. Parth: Rebounds, Turnovers, and Free Throws have no impact on the win rate in a college basketball game. Lerui: The quality of school pertains to specifically the monetary resources being pumped into fund the team. My research question is whether the Athletically Related Student Aid of the basketball team is equally as important as the total salary in affecting the win rate of the team.

6. Project result highlight (what are the major findings of your project, what do you consider the most contribution of this project): To win more, teams should prioritize EFG_O, EFG_D, and reducing their turnovers.

7. Project data introduction (the exact data resource, a table summarizes variable notation and definition, such as the one on the first page in the homework).

College Basketball Dataset Variables: West (1 if team is in a Western Conference, 0 if in Eastern Conference. This was created from original data), G (Games played), W (Games won), ADJOE (Adjusted Offensive Efficiency, estimates points scored per 100 possessions), ADJDE (Adjusted Defensive Efficiency, estimates points allowed per 100 possessions), FTRD (Free throw rate allowed), FTR (Free throw rate), EFG_O (Effective Field Goal Percentage Shot), EFG_D (Effective Field Goal Percentage Allowed), ORB (Offensive Rebound Rate), TORD (Turnover percentage committed, steal rate), TOR (Turnover percentage allowed), 2P_0 (Two-Point Shooting Accuracy), 3P_0 (Three-Point Shooting Accuracy), DRB (Offensive Rebound Rate Allowed), ARSA(Athletically Related Student Aid), REC(Recruiting Expenses), OPE(Men's Basketball Game Expenses), CSAL(Men's Basketball Coaching Staff Salaries).

$Y = \frac{W}{G} * 100\%,\ X_1: ADJOE, X_2: ADJDE,\ X_3: FTRD,\ X_4: West,\ X_5: FTR, X_6: EFG\_O,\ X_7: EFG\_D,$
$X_8: ORB,\ X_9: TORD,\ X_{10}: TOR,\ X_{11}: 2P\_O,\ X_{12}: 3P\_O,\ X_{13}: DRB,\ X_{14}: ARSA,\ X_{15}: REC,\ X_{16}: OPE,\ X_{17}: CSAL$

8. Briefly describe what you learn from the project and what is the most challenging part. We learned how to apply linear regression models to perform a statistical analysis on data. The most challenging part was making sure we remedied any assumption violations.

9. In one sentence, what is your advice for the future student to deliver a high-quality project in the course. Study the slides with the R code on it because it is very helpful for the project and exams.

Research Project

Ryan Newman's Research

Data Setup/Cleaning

This research question is focused on the relationship between conference and ADJOE and ADJDE when using them to predict win rate. ADJOE and ADJDE are chosen because we want to see if offense or defense matters more or less depending on the conference a team is in. Conference is reduced to a binary categorical by creating a new variable "west" which is 1 if the conference is mostly western schools, and 0 otherwise. The initial linear model is Y~ADJOE+ADJDE+FTR+FTRD+west+west*ADJOE+west*ADJDE+west*FTR+west*FTRD. [A.1.1]

Data Exploration

Check for multicollinearity by plotting ADJOE vs. ADJDE and FTR vs. FTRD and multicollinearity is not indicated by these. [A.1.2] Plotting residuals vs. predicted values indicates potential violation of the constant variance assumption. Based on the Breusch-Pagan test, this assumption is violated. Based on the Shapiro test, the normality assumption is violated. The residual plots indicate some outliers. [A.1.3] dffits confirms there are influential points on single fitted values and dfbetas confirm there are influential points on the betas for all variables in the initial model. However, there are no influential points on all fitted values, per Cook's Distance. [A.1.4] VIF indicates no excessive multicollinearity. Based on added-variable plots, FTR, west, and west and the interaction it has with FTR, ADJOE, ADJDE, and FTRD aren't needed given all other predictors are in the model. [A.1.5] Because of assumption violations, Box-Cox transformation on Y is done and $\lambda = 0.93939$ is obtained. Then, the model is fit on transformY and based on the BP test, constant variance is still violated and with a lower p-value, making it worse. Also, the MLE plot shows $\lambda = 1$ was within the best lambda values so choose to not transform Y. [A.1.6]

Model Selection

After selecting, diagnose again, and apply remedial methods. When using the best subset for model selection, AIC and SBC are used because of assumption violations and PRESS because we are interested in predictive power. For AIC and PRESS, the full model is the best, 2nd best is the model with everything except west, and the 3rd best is the model with everything except FTR. For SBC, the best model has everything except west, 2nd best is full model, 3rd best is model without FTR and west. Stepwise model selection says to just remove interaction between west and ADJOE. Because of research goals, west, interaction between west and ADJOE, west and ADJDE, must stay in. Based on this and best subsets algorithm, the selected model will be the initial model with FTR terms dropped, because this model performs 3rd best for AIC, SBC, PRESS, and meets research goals. Plus, the avPlot indicated that FTR and interaction between FTR and west were not needed given other predictors are considered. [A.1.7]

Model Diagnostics

After plotting residuals for the selected model, it looks like there may be a constant variance assumption violation and based on the BP test, there is a violation. The Shapiro test also shows the normality of residuals assumption is violated. [A.1.8] To address this, a Box-Cox transformation is done and $\lambda = 0.93939$ is obtained. Based on the BP test for the selected model with transformY, the constant variance assumption is violated and the p-value is also lower. $\lambda = 1$ was within the best lambdas so don't transform Y. [A.1.9] AvPlots indicate that west, and its interaction between ADJOE, ADJDE, and FTRD aren't needed given all other predictors are included. [A.1.10] dffits shows there are many influential points on single fitted values and dfbetas show there are many influential points on the coefficients for all variables in the selected model. Cook's distance shows there are no influential points on all fitted values. VIF shows there is no excessive multicollinearity. [A.1.11]

Remedial Methods

To address influential points, robust regression is done to dampen their effect without having to remove them. [A.1.12] After this, residuals are plotted and based on the BP test, there is still a constant variance violation. To address this, do WLS. [A.1.13] After WLS and doing the BP test, it is found the constant variance assumption isn't violated. After doing the Shapiro test, the normality assumption violation still persists. [A.1.14]

Cross-validation

Doing cross-validation to evaluate the predictive power of the final model obtains: RMSE = 11.59519, Rsquared = 0.5914609, MAE = 9.304987. As more data is collected as college basketball seasons are finished each year, predictive power will improve. [A.1.16]

Test the hypotheses [A.1.15]

Let X1 = ADJOE, X2 = ADJDE, X3 = FTRD, X4 = west. To test significance of interaction between west and ADJOE, $H_0: \beta_{1,4} = 0$ given all other predictors are in the model. Reduced model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_{2,4} X_2 X_4 + \beta_{3,4} X_3 X_4$

and $H_A: \beta_{1,4} \neq 0$ given all other predictors are in the model. Full model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_{1,4} X_1 X_4 + \beta_{2,4} X_2 X_4 + \beta_{3,4} X_3 X_4$

```
> anova(reducedmod, fullmod)
Analysis of Variance Table

Model 1: Y ~ ADJOE + ADJDE + FTRD + west + west * ADJOE + west * FTRD
Model 2: Y ~ ADJOE + ADJDE + FTRD + west + west * ADJOE + west * ADJDE +
    west * FTRD
  Res.Df    RSS Df  Sum of Sq  F Pr(>F)
1   3516 5508.2
2   3515 5508.2  1 4.8706e-05  0 0.9956
```

The p-value (0.99), shows there is no significant interaction effect between conference and ADJOE. The fact that the p-value is extremely high shows that this interaction effect is very insignificant. This means that when predicting a team's win rate using this model, being in a western conference or not has no effect on linear impact of ADJOE on win rate. The interaction between conference and ADJOE doesn't matter and shouldn't be considered. Next, test significance of interaction between west and ADJDE. $H_0: \beta_{2,4} = 0$ given all other predictors are in the model.

Reduced model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_{1,4} X_1 X_4 + \beta_{3,4} X_3 X_4$

$H_A: \beta_{2,4} \neq 0$ given all other predictors are in the model

Full model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_{1,4} X_1 X_4 + \beta_{2,4} X_2 X_4 + \beta_{3,4} X_3 X_4$

```
> anova(reducedmod, fullmod)
Analysis of Variance Table

Model 1: Y ~ ADJOE + ADJDE + FTRD + west + west * ADJOE + west * FTRD
Model 2: Y ~ ADJOE + ADJDE + FTRD + west + west * ADJOE + west * ADJDE +
    west * FTRD
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1   3516 5510.9
2   3515 5508.2  1    2.7168 1.7337  0.188
```

For the 2nd hypothesis test, the p-value shows there is no significant interaction effect between conference and ADJDE. So, when predicting a team's win rate with this model, whether a team is in a western conference or not has no effect on the linear impact of ADJDE on win rate. Interaction between conference and ADJDE doesn't matter and shouldn't be considered.

Bootstrapping

Bootstrapping is done to calculate a confidence interval for the true value of the linear impact for all variables in the final model. Bootstrapping helps because the normality assumption was violated. Bootstrap CI for $\beta_0: (29.18, 57.21)$. Bootstrap CI for $\beta_1: (1.137, 1.284)$This Bootstrap CI is very precise. All the values being positive makes sense, because if a team has a strong offense, they are likely to win more games. Bootstrap CI for $\beta_2: (-1.119, -0.969)$This is also very precise. All the values in

the interval are negative which makes sense, because the more points allowed, a team will lose more because they are letting their opponent score more. The Bootstrap CI for β_3: $(-0.3307, -0.1733)$ All the values in this are negative which makes sense, because if a team allows a lot of free throws, that means they allow their opponents to score more, resulting in more losses. Bootstrap CI for β_4: $(-18.123, 35.373)$ is not precise. Bootstrap CI for interaction between ADJOE and conference suggests the true value may be 0, meaning no impact, and this is supported by the hypothesis testing earlier $\beta_{1,4}$: $(-0.1694, 0.1045)$. The Bootstrap CI for interaction between ADJDE and conference suggests the true value may be 0, meaning no impact, and this is supported by the hypothesis testing earlier$\beta_{2,4}$: $(-0.259, 0.0277)$. Bootstrap CI for $\beta_{3,4}$: $(0.0725, 0.3593)$. This suggests being in the western conference reduces the negative impact of FTRD on winrate because all the values in the interval are positive, and the FTRD has all negative values. This suggests that giving up free throws matters less in the western conference. [A.1.17]

Conclusion

Overall, the research question conclusion is that in this model, the impact of ADJOE and ADJDE isn't significantly affected by whether or not a school is in a western or eastern conference. Conference has no effect on ADJOE and ADJDE's linear impact on win rate. It doesn't matter what conference a team is in, their ADJOE and ADJDE will have the same impact on win rate. The main limitation is that this conclusion is only for the specific selected linear model. In the future, you would have to study all possible models.

Jerry Huang's Research

Data Cleaning

The dataset was carefully examined for duplicates, missing values, and inconsistencies. No duplicate rows were found, and no rows contained empty values. However, one row with implausible data—indicating 19 games played but 20 games won, resulting in a win rate exceeding 100%—was identified and removed. Additionally, some rows showed a win rate of zero. To facilitate smoother transformations during later analysis, a small value of 0.0001 was added to all win rate entries.

Data Exploration.

A histogram of the data revealed a normal distribution for the variables, suggesting no major deviations from expected patterns. Box plots highlighted some outliers among teams, potentially representing exceptional cases. [A.2.1] These outliers, while notable, did not fall significantly outside the expected range, supporting their inclusion for analysis.

Initial Model fitting: $Y = \beta_0 + \beta_1 X_5 + \beta_2 X_6 + \beta_3 X_8 + \beta_4 X_{10} + \epsilon$

```
Call:
lm(formula = Y ~ X5 + X6 + X8 + X10, data = cbb)

Residuals:
    Min      1Q  Median      3Q     Max
-38.756  -8.117   0.021   7.915  38.504

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -122.36318    4.76693 -25.669  < 2e-16 ***
X5             0.25025    0.03821   6.549 6.63e-11 ***
X6             3.33722    0.06930  48.158  < 2e-16 ***
X8             1.60022    0.05029  31.822  < 2e-16 ***
X10           -2.56732    0.10248 -25.051  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.71 on 3517 degrees of freedom
Multiple R-squared:  0.5844,    Adjusted R-squared:  0.584
F-statistic:  1237 on 4 and 3517 DF,  p-value: < 2.2e-16
```

```
Analysis of Variance Table

Response: Y
           Df Sum Sq Mean Sq F value    Pr(>F)
X5          1  15157   15157  110.61 < 2.2e-16 ***
X6          1 455187  455187 3321.70 < 2.2e-16 ***
X8          1 121441  121441  886.21 < 2.2e-16 ***
X10         1  85995   85995  627.55 < 2.2e-16 ***
Residuals 3517 481949     137
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 2.1 model summary and anova table

Diagnostics

Correlation matrices [A.2.2] and scatter plots [A.2.3] of all variables showed minimal to moderate collinearity, enabling clearer insights into variable relationships. The model summary demonstrated that all predictors had low p-values, indicating their statistical significance and importance to the model. A low F-statistic supported the global significance of the predictors, rejecting the null hypothesis that all coefficients are zero. However, the low R-squared value suggested that the model's overall fit might be poor.

Residual plots [A.2.4] showed that residuals were mostly evenly scattered around a flat zero line, indicating constant variance. Slight curvature was observed for some variables, such as FTR and ORB, though these effects were minor. The Breusch-Pagan(BP) test [A.2.5] yielded a p-value of 0.07, which was insufficient to reject the null hypothesis of constant variance at the 0.05 significance level. However, addressing potential non-constant variance remains advisable. The Shapiro-Wilk test [A.2.6] for normality gave a p-value of 0.4646, affirming that the residuals followed a normal distribution. Added-variable plots [A.2.8] indicated that including each predictor contributed linearly to the model.

Outlier detection methods provided mixed results [A.2.9-A.2.12]. No outliers were identified using studentized deleted residuals or Cook's distance. However, hat leverage values flagged 233 outliers, DFFITS highlighted additional points, and DFBetas identified influential cases. Altogether, 746 outliers were detected across methods. Removing such a large proportion of the data was deemed impractical, so the full dataset was retained for subsequent analysis.

Variance Inflation Factor (VIF) values [A.2.13] were small, confirming no significant multicollinearity issues among predictors.

Transformation

A Box-Cox transformation suggested a lambda of 0.939 [A.2.15], close to 1, which implied no strong need for transformation. Given the lack of significant non-constant variance based on the BP test [A.2.16], and the possibility that a transformation could introduce variance issues, no transformation was applied.

Model Selection

Using the best subset selection algorithm [A.2.17], the full model was determined to have the highest adjusted R-squared value, as well as the lowest Cp, AICp, SBCp, and PRESSp values. This solidified the decision to use the full model without requiring further diagnostics or adjustments.

Cross-Validation

Ten-fold cross-validation [A.2.18] yielded a Root Mean Squared Error (RMSE) of 11.71798, indicating the predictive accuracy of the model. Despite some variance concerns, this performance supported the choice of the original model.

Advanced Remedial Methods

Alternative approaches to improve the model were explored. Weighted Least Squares (WLS) regression was tested to address non-constant variance, but it did not significantly improve standard error, R-squared, or F-statistics [A.2.19]. Robust regression was also applied to handle outliers but resulted in worse residual standard errors and higher standard errors for coefficients [A.2.20]. These findings affirmed that the original model was the best choice.

Bootstrapping methods provided estimates [A.2.21] consistent with the initial model, with all initial estimates falling within bootstrapped confidence intervals. While this reinforced the robustness of the original model, bootstrapping results could serve as an additional validation tool.

Hypothesis Testing

Primary research question: The effective offensive field goal rate EFG_O (X_6) has the same impact on win rate as the offensive rebound rate ORB (X_8).

Using the previously built model: $Y = \beta_0 + \beta_1 X_5 + \beta_2 X_6 + \beta_3 X_8 + \beta_4 X_{10}$

$H_o: \beta_2 = \beta_3 \qquad H_a: \beta_2 \neq \beta_3$

Reduced model: $Y = \beta_0 + \beta_1 X_5 + \beta_2 (X_6 + X_8) + \beta_4 X_{10}$

Full model: $Y = \beta_0 + \beta_1 X_5 + \beta_2 X_6 + \beta_3 X_8 + \beta_4 X_{10}$

A General Linear Test (GLT) was performed to compare the full model against a reduced model where the coefficients of EFG_O and ORB were constrained to be equal.

```
Analysis of Variance Table

Model 1: Y ~ X5 + X + X10
Model 2: Y ~ X5 + X6 + X8 + X10
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   3518 543700
2   3517 481949  1     61751 450.62 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 2.2 Anova table reflecting the comparison between the reduced and full model in the hypothesis

The GLT yielded a p-value of 2.2×10^{-16}. At the 0.05 significance level, the null hypothesis was rejected, leading to the conclusion that X_6 (effective offensive field goal EFG_O) does not have the same impact on win rate as ORB. Furthermore, the coefficient of EFG_O was much higher than that of ORB, suggesting that EFG_O likely has a greater influence on win rate. The results are further substantiated by the bootstrapped confidence intervals [A.2.21] for the coefficients of EFG_O (β_2) and ORB (β_3). The confidence interval for β_2 ranges from 3.198 to 3.465, while the confidence interval for β_3 ranges from 1.507 to 1.698. Notably, these intervals do not overlap, indicating a clear distinction between the effects of EFG_O and ORB on win rate. Furthermore, the point estimates of β_2 = 3.337 and β_3 = 1.600 do not fall within each other's confidence intervals, reinforcing the conclusion that their impacts are statistically and practically different. This conclusion is statistically significant and aligns with the fundamental dynamics of basketball.

From a practical standpoint, the result makes intuitive sense. Basketball is fundamentally a scoring game, and EFG_O directly measures scoring efficiency, which is critical to winning. Even if a team excels in offensive rebounds (ORB), these efforts only create opportunities to score; they do not directly translate into points unless coupled with efficient scoring. Therefore, the finding that EFG_O has a stronger impact underscores the importance of converting scoring opportunities into points, rather than merely creating additional chances through rebounding.

Moreover, the disparity in the coefficients highlights a strategic insight for basketball teams: while rebounding remains valuable, optimizing scoring efficiency should be prioritized for improving win rates. This result reinforces common basketball strategies that emphasize high-percentage shots and effective shot selection over simply focusing on rebounds. Overall, the model's findings are both statistically robust and contextually meaningful, providing valuable insights into the factors that influence success in basketball games.

Nikhil Venkatachalam's Research

Data Setup/Cleaning

Initially, all libraries were imported if installed, and installed then imported if not already. Additionally, the seed was set for the purposes of data replication, and the dataset was imported. Then, in order to safeguard against potential data discrepancies between other team members who had already observed minor errors within the data, any rows with a win rate above 100 were automatically discarded, and 0.001 was added to the win rate of remaining entries to make data analysis easier. Additionally, the West variable was created in order to maintain consistency with the overarching project.

Data Exploration

Histograms were made for the variables used (using the square root rule for determining the number of bins to be used in the histograms), and as far as could be ascertained, the distributions of each variable were mostly normal, meaning no special modifications needed to be made for their inclusion in analysis [A.3.1]. This conclusion seemed to be supported by a further side-by-side boxplot analysis, where the distributions seemed to be normal and, while outliers were present within the data, these outliers were not considered abnormal enough for any data cleaning or variable modification [A.3.2]. Therefore, the initial model fitted is as follows:

$$Y = \beta_0 + \beta_1 X_3 + \beta_2 X_7 + \beta_3 X_9 + \beta_4 X_{13} + \epsilon$$

Where, as stated, Y = $\frac{W}{G} * 100\%$, $X_3 : FTRD$, $X_7 : EFG\_D$, $X_9 : TORD$, $X_{13} : DRB$. A model summary and ANOVA table for said model can be observed:

```
Call:
lm(formula = Y ~ FTRD + EFG_D + TORD + DRB, data = cbb)

Residuals:
    Min      1Q  Median      3Q     Max
-40.554  -8.458   0.220   8.462  43.526

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 236.79257    4.19930   56.39   <2e-16 ***
FTRD         -0.72870    0.03606  -20.21   <2e-16 ***
EFG_D        -3.07578    0.07360  -41.79   <2e-16 ***
TORD          2.62977    0.10316   25.49   <2e-16 ***
DRB          -1.85297    0.07064  -26.23   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.41 on 3517 degrees of freedom
Multiple R-squared:  0.5327,    Adjusted R-squared:  0.5322
F-statistic:  1002 on 4 and 3517 DF,  p-value: < 2.2e-16
```

```
Analysis of Variance Table

Response: Y
            Df Sum Sq Mean Sq F value    Pr(>F)
FTRD         1  89435   89435  580.43 < 2.2e-16 ***
EFG_D        1 360835  360835 2341.81 < 2.2e-16 ***
TORD         1  61531   61531  399.33 < 2.2e-16 ***
DRB          1 106017  106017  688.05 < 2.2e-16 ***
Residuals 3517 541913     154
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Data Diagnostics

Looking further at the model summary, all the predictors for the model have low p-values (<2e-16), thus indicating that all the predictors are statistically significant. Additionally, the F-statistic (1002) is also comparatively low, corroborating the p-values and indicating that all coefficients are not zero. It is,

however, important to keep in mind that the R-squared (0.5327) and adjusted R-squared (0.5322) values are low, indicating a poor fit of the model. A correlation matrix was produced for all variables involved, indicating moderate collinearity at most among different pairs of variables [A.3.3]. A scatterplot matrix was also produced along the same lines, and this production corroborated the previous results [A.3.4].

Residual plots were produced, which seemed to indicate constant variance for each variable, given that residuals were mostly evenly scattered along a flat-zero line [A.3.5]. However, a Breusch-Pagan (BP) test was conducted, and the p-value was returned as 0.0009817 (well below the 0.05 significance level), meaning that there was heteroscedasticity observed, meaning that the variance of the residuals is in fact not constant [A.3.6]. This test would be of concern, and would thus place transformation of the model under consideration. A Shapiro-Wilk test was also conducted, returning a p-value of 0.1774 (above the 0.05 significance level), showing that the residuals follow a normal distribution [A.3.7]. Additionally, added-variable plots done on each variable concluded that each variable contributes linearly to the model [A.3.8].

To determine the existence and importance of considering outliers, several methods were utilized. First, the threshold for which outliers would be considered significant for each method was determined [A.3.9]. Then, the methods (studentized deleted residuals, hat leverage values, DFFITS, Cook's distance, DFBETAS, and Mahalanobis distance) were all used to track outliers within the model, while graphing each data point plus outlier threshold [A.3.10-A.3.15]. All in all, while select outliers can be observed, the amount observed is considered to simultaneously not impact the model while being too many to track and remove. Thus, the model was retained as-is. The Variance Inflation Factor (VIF) values were also small, being close to 1, which indicates low multicollinearity among predictors [A.3.16].

Assessment of Need for Transformation

A Box-Cox transformation was done on the original model, suggesting a lambda of 1 [A.3.17]. This indicates that a transformation of the model would in fact have no effect, and that the model is fine without transformation. Though the previous BP test result was concerning, a lambda of 1 suggested from the Box-Cox transformation is a very strong mandate to leave the model unchanged. Therefore, the model was not transformed.

Model Selection

The best subset selection algorithm revealed that the full model had the highest adjusted R-squared value, along with the lowest values for Cp, AICp, SBCp, and PRESSp [A.3.18]. This confirmed the choice to use the full model without the need for extended focus on other additional diagnostics or modifications that could be made.

Cross-Validation of the Model

Ten-fold cross-validation produced a Root Mean Squared Error (RMSE) of 12.41292, demonstrating the model's predictive accuracy [A.3.19]. Although there were some heavy, substantiated concerns about variance, this performance reinforced the decision to retain the original model.

Exploring Advanced Remedial Methods

The presence of heteroscedasticity and some potential influential points indicated by previous tests underscored a need to take alternative approaches to improve the model. Therefore, Weighted Least Squares (WLS) regression was tested to address the non-constant variance problem observed [A.3.20]. The improvement is noticeable, as the WLS model, which includes weights, demonstrates a significantly lower residual standard error (3.156 compared to 12.41 in the original model), indicating a more

accurate fit to the data. The WLS model also shows a much higher R-squared value (0.9264 vs. 0.5327 in the original model), suggesting that it explains a far greater proportion of the variance in the dependent variable Y. Additionally, the coefficients in the WLS model have smaller standard errors, leading to more precise estimates. These improvements point to a better model performance, precisely due to adjustments for heteroscedasticity. This can be corroborated by a BP test on the new model, which shows a p-value of 0.7109, indicating very little heteroscedasticity [A.3.21].

Robust regression was also tested to see if it could eliminate outliers [A.3.22]. However, this model testing resulted in a worse residual standard error and higher standard errors for coefficients. Therefore, this model can be safely disregarded. Thus, given that the WLS model exhibited significant positive change in terms of the effectiveness of the model, that is the model that will be used henceforth to conduct hypothesis testing.

Hypothesis Testing

The hypothesis shall be tested as such:

Primary research question: The effective field goal percentage allowed EFG_D (X_7) has the same impact on win rate as the turnover percentage committed (steal rate) TORD (X_9).

Using the previously built model: $Y = \beta_0 + \beta_1 X_3 + \beta_2 X_7 + \beta_3 X_9 + \beta_4 X_{13}$

$H_o: \beta_2 = \beta_3 \qquad H_a: \beta_2 \neq \beta_3$

Reduced model: $Y = \beta_0 + \beta_1 X_3 + \beta_2 (X_7 + X_9) + \beta_4 X_{13}$

Full model: $Y = \beta_0 + \beta_1 X_3 + \beta_2 X_7 + \beta_3 X_9 + \beta_4 X_{13}$

In this testing, a General Linear Test (GLT) was performed to compare the full WLS model against a reduced WLS model (with the same weights) where the coefficients of EFG_D and TORD were constrained to be equal. The model summary of the reduced model and the ANOVA table detailing the results of the GLT are as follows:

```
Call:
lm(formula = Y ~ FTRD + EFG_TORD + DRB, data = cbb, weights = weights)

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-144.13   -5.32   -1.35    3.05  338.30

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 190.99308    4.11740   46.39   <2e-16 ***
FTRD         -0.57515    0.04109  -14.00   <2e-16 ***
EFG_TORD     -1.09829    0.06269  -17.52   <2e-16 ***
DRB          -1.38805    0.07408  -18.74   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.79 on 3518 degrees of freedom
Multiple R-squared:  0.2919,    Adjusted R-squared:  0.2913
F-statistic: 483.4 on 3 and 3518 DF,  p-value: < 2.2e-16
```

```
Analysis of Variance Table

Model 1: Y ~ FTRD + EFG_TORD + DRB
Model 2: Y ~ FTRD + EFG_D + TORD + DRB
  Res.Df    RSS Df Sum of Sq     F    Pr(>F)
1   3518 337209
2   3517  35033  1    302176 30336 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Conclusion

The results indicate a highly significant difference between the reduced model, where the coefficients of EFG_D and TORD are constrained to be equal, and the full model, where these coefficients are allowed

to differ. With a p-value smaller than 2.2e-16 and an F-statistic of 30,336, we reject the null hypothesis that the two variables have the same impact. This suggests that the full model provides a significantly better fit, as shown by the large reduction in residual sum of squares (from 337,209 in the reduced model to 35,033 in the full model). In fact, constraining these variables to have equal effects significantly harms the model's ability to explain the variance in the outcome, as indicated by the large improvement in the residual sum of squares. Therefore, it is statistically clear that the relationship between the predictors EFG_D and TORD and the response variable should be modeled separately.

We can look back at the full model to better explore this relationship. While both EFG_D and TORD have significant impacts on the outcome, EFG_D has a slightly stronger effect in terms of magnitude compared to TORD. This implies that, all else being equal, reducing the opponent's effective field goal percentage would have a greater effect on improving team performance than increasing the rate of forcing turnovers. Thus, the results suggest that contesting shots and limiting opponents' shooting efficiency (EFG_D) should be a priority for improving team performance, as this has a slightly greater impact than forcing turnovers (TORD). Coaches should focus on strategies that reduce the opponent's effective field goal percentage, like better closeouts and tighter perimeter defense. While forcing turnovers is important, defensive schemes should not sacrifice solid shot defense for aggressive turnover tactics. A balanced defensive approach that emphasizes shot contesting while still working to force turnovers could maximize team success. Prioritizing shot defense is likely to yield more significant performance improvements.

Parth Gandhi's Research

Materials: "**STAT512Project_Parth.Rmd**", "**CollegeBBData.csv**", "**CollegeBBDataExcel.xlsx**"

Research Qn: Rebounds, Turnovers, and Free Throws have no impact on the win rate in a college basketball game.

Data Cleaning

The dependent variable (Win Rate) was calculated as "WinPercent".

The dataset was meticulously reviewed to identify duplicates, missing values, and inconsistencies. One row containing implausible data—reporting 19 games played but 20 games won, leading to a win rate exceeding 100%—was detected and subsequently removed. To address the issue of rows with a win rate of zero and to ensure smoother transformations during subsequent analyses, a small constant value of 0.0001 was added to all win rate entries. To ensure uniformity, we perform mapping of variables as seen in Section A.4.1. A more streamlined dataframe was also created as shown in Section A.4.1.

Data Exploration

The histogram and boxplot of the variables, as presented in Section A.4.2, indicate that all variables exhibit an approximately normal distribution with no significant deviations. However, the boxplots reveal the presence of some outliers. Presented below are the model specifications, summary of the model and the corresponding ANOVA Table.

Initial model: $Y = \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \beta_3 X_8 + \beta_4 X_9 + \beta_5 X_5 + \epsilon$

```
Call:
lm(formula = y ~ x11 + x12 + x8 + x9 + x5, data = clgdata)

Residuals:
    Min      1Q  Median      3Q     Max
-43.214  -8.404   0.582   8.161  44.401

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -228.79793    4.39354 -52.076   <2e-16 ***
x11            2.84797    0.06523  43.658   <2e-16 ***
x12            1.93254    0.08213  23.532   <2e-16 ***
x8             1.34222    0.05216  25.732   <2e-16 ***
x9             1.71437    0.09373  18.290   <2e-16 ***
x5             0.09429    0.03943   2.391   0.0168 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.11 on 3516 degrees of freedom
Multiple R-squared:  0.5552,    Adjusted R-squared:  0.5546
F-statistic: 877.7 on 5 and 3516 DF,  p-value: < 2.2e-16
```

```
Analysis of Variance Table

Response: y
           Df Sum Sq Mean Sq  F value  Pr(>F)
x11         1 400590  400590 2730.344 < 2e-16 ***
x12         1  51924   51924  353.901 < 2e-16 ***
x8          1 140357  140357  956.644 < 2e-16 ***
x9          1  50160   50160  341.883 < 2e-16 ***
x5          1    839     839    5.718 0.01684 *
Residuals 3516 515860     147
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
```

Diagnostics

To examine the impact of correlation and multicollinearity in our model, we refer to Sections A.4.3 to A.4.6. Sections A.4.3 and A.4.4 present the correlation matrix and scatter plots, respectively, which indicate weak to moderate correlations among the predictors. Comparing the Type I and Type II ANOVA tables (in Section A.4.5) suggests the presence of a minor degree of multicollinearity among the predictors. However, the Variance Inflation Factors (VIFs) provided in Section A.4.6 confirm that multicollinearity is not a significant issue, as all VIF values are well below the threshold of 10.

From the residual plots in Section A.4.7, we observe that for all five predictors, the residuals are evenly scattered around the zero line, indicating linearity. The variance appears approximately constant across the predictors, suggesting no significant evidence of heteroscedasticity. In Section A.4.8, the Breusch-Pagan test yields a p-value of 0.1053, which exceeds the 0.05 threshold, confirming that there is no evidence of non-constant variance. Similarly, in Section A.4.9, the Shapiro-Wilk test produces a p-value of 0.1559, which is also above the 0.05 threshold, indicating that the residuals satisfy the assumption of normality. This conclusion is further validated by the Q-Q plot in Section A.4.10, where the residuals closely follow the diagonal reference line.

To analyze the outliers and influential points, we refer to Sections A.4.11 to A.4.14. In A.4.11, the studentized deleted residuals identify the outliers in the model. However, given the size of the dataset, their overall impact on the model appears to be negligible. Sections A.4.12, A.4.13, and A.4.14 present the results of the DFFITS, DFBETAS, and Cook's Distance tests, respectively. While these tests reveal the presence of several influential points, their overall effect on the model is minimal.

Transformation

The Box-Cox plot in Section A.4.15 indicates a smooth curve with a peak near a lambda of 1 (Section A.4.16). This suggests that a transformation is not strongly needed. Considering the BP test in Section A.4.8, which shows no significant non-constant variance, and the potential risk of introducing variance instability through transformation, the original model will be retained.

Model Selection

Using the best subset selection algorithm in Section A.4.17, the full model was identified as having the highest R-squared and adjusted R-squared values, along with the lowest Cp, AICp, and PRESSp values. Although the SBCp value for this model was the second lowest, it was deemed sufficiently optimal. Therefore, it was concluded that the full model would be utilized without modifications.

Advanced Remedial Methods

Since our analysis above confirmed that the model did not exhibit multicollinearity, there was no need to perform Ridge Regression. Similarly, as no significant issue of heteroscedasticity was identified, a Weighted Least Squares test was not conducted.

However, the presence of influential points necessitated the use of robust regression, the results of which are presented in Section A.4.18. The residual standard errors increased with robust regression, and also output higher standard errors for the coefficients. These results reinforce that the original model remains the most appropriate choice.

Cross-Validation

To ensure the model is not overfitted and provides a realistic estimate of its performance, 10-fold cross-validation was conducted, as detailed in Section A.4.19. This process yielded a Root Mean Squared Error (RMSE) of 12.11805, demonstrating the model's predictive accuracy and supporting the selection of the original model.

Hypothesis Testing

Primary research question:Rebounds, Turnovers, and Free Throws have no impact on the win rate in a college basketball game.

Main Model: $Y = \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \beta_3 X_8 + \beta_4 X_9 + \beta_5 X_5 + \epsilon \qquad n = 3522$

$H_o: \beta_3 = \beta_4 = \beta_5 = 0 \qquad H_a: Not\ H_o (At\ least\ one\ of\ \beta_3, \beta_4, \beta_5\ is\ not\ equal\ to\ 0)$

$Reduced\ Model: Y = \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} \qquad dfE_R = 3522 - 3 = 3519$

$Full\ Model: Y = \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \beta_3 X_8 + \beta_4 X_9 + \beta_5 X_5$

$dfE_f = 3522 - 6 = 3516$ **Refer to A.4.20 and A.4.21 for summary & ANOVA of reduced model**

```
Analysis of Variance Table

Model 1: y ~ x11 + x12
Model 2: y ~ x11 + x12 + x8 + x9 + x5
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   3519 707216
2   3516 515860  3    191356 434.75 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
```

$$F_s = \frac{\frac{SSE(R) - SSE(F)}{df_R - df_F}}{SSE(F)/df_F} = \frac{191356/3}{515860/3516}$$

$$= 434.74825 \approx 434.75$$

$$Critical\ Value = F(0.95, 3, 3516)$$

$$= 2.6074$$

```
[1] "Critical Value 2.60743521209573"
```

Since $F_s > critical\ value$, we reject the null hypothesis. There is a good chance that at least one of β_3, β_4, *or* β_5 is not equal to 0.

Conclusion

Rebounds, turnovers, and free throws play a significant role in determining the outcome of college basketball games. This relationship is supported by the summary of the full model, where the coefficients of x8, x9, and x5, representing offensive rebound rate, turnover percentage committed, and free throw rate, respectively, show a positive relationship with the win rate. Among these, turnover percentage appears to have the most substantial impact on the win rate, followed by offensive rebound rate and free throw rate. The findings are further substantiated by the added-variable plots in A.4.22.

Practically, this aligns with the dynamics of basketball, where reducing turnovers not only prevents the opposing team from gaining possession but also maintains offensive momentum. Similarly, securing offensive rebounds creates additional scoring opportunities, while maintaining a strong free throw rate ensures efficiency in capitalizing on scoring chances.

Le Rui Tay's Research

Data Cleaning

The dataset was carefully examined for duplicates, missing values, and inconsistencies. No duplicate rows were found, and no rows contained empty values. Due to additional sources used, string matching has been used to make sure that the data between sources are aligned. The independent variables were scaled due to the large magnitude of the original data.

Data Exploration

Based on scatter plots between the Y variable and the individual independent variables, there isn't a clear linear relationship as seen in A.5.1. By drawing out the boxplot, we see that there might be a number of outliers worth considering.

The initial model will be the full model of $Y = X_{14} + X_{15} + X_{16} + X_{17}$.

```
Call:
lm(formula = Y ~ X14 + X15 + X16 + X17, data = final_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.42528 -0.11208  0.02406  0.12191  0.38893

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.575621   0.016360  35.184   <2e-16 ***
X14         -0.007753   0.019023  -0.408    0.684
X15         -0.011922   0.024003  -0.497    0.620
X16          0.002430   0.022116   0.110    0.913
X17          0.036675   0.028811   1.273    0.206
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1716 on 105 degrees of freedom
Multiple R-squared:  0.02651,   Adjusted R-squared:  -0.01058
F-statistic: 0.7148 on 4 and 105 DF,  p-value: 0.5837
```

Table 5.1 Model summary of the initial model

It can be seen from the model summary that the model has a very poor performance due to the low R^2 value of 0.02651. The high p-values for all the independent variables shows that the coefficients are not significant. The p-value in the F-statistics show that the model is not statistically significant, meaning the predictors collectively do not explain much variation in Y.

Diagnostics

We start off by plotting the residual plot. We can see from A.5.2 that the residuals are scattered around the 0 line. We can see a constant variance of the residual as well as no strong outliers.

We start by checking for multicollinearity issues. We use VIF as seen in A.5.3 and it turns out that VIF value for each of the predictors is low, signally that there is no multicollinearity between the predictors. Looking at the pairwise correlation in A.5.4, X17 has high correlation with X16 and X15 while X15 has high correlation with X16.

Looking at whether the residuals have constant variance, we use bptest. From bptest, the high p-value of 0.6871 suggests the residuals have constant variance(A.5.5).

Onto whether the residuals follow normal distribution, we use shapiro test as well as qqnorm diagram. The shapiro test gives a p-value larger than 0.05, showing that the residuals follow a normal distribution(A.5.6). The qqnorm diagram further evidences the normal distribution of residuals.

Lastly, we check for any potential influential points. Looking at the Student deleted residuals and comparing it to the Bonferroni procedure of t(1-0.05/(2*110),110-5-1)=t(1-0.05/(2*110), 104), we realise that none of the points are above 3.621544.(A.5.7) This means there are no outliers from SDR. Looking at Cook's distance, the data point with the highest distance is 93. However, by comparing with the F(5,105) distribution, it can be seen it lies at the 10th percentile. This signifies that there are no outliers to consider. Both DFFITS and DFBETAS also show no possible outliers.

Transformation

Since there is no violation of constant variance and normal distribution, no transformation is needed.

Model Selection

The best subset selection algorithm revealed that the Y=X17 had the highest adjusted R-squared value, along with the lowest values for AICp and SBCp(A.5.8). Additionally, by employing stepwise regression, it shows Y=X17 is the best. However, due to the nature of our research question, we decided to keep X1 as well. Thus, the model selected is Y=X14+X17.

Cross-Validation

After doing K-fold Cross validation, we get RMSE=0.1703072, Rsquared= 0.1146159, MAE=0.1403273.(A.5.9)

Advanced Remedial Methods

We check for multicollinearity issues. As seen in A.5.10 and it turns out that VIF value for each of the predictors is low again, signally that there is no multicollinearity between the predictors and thus no remedy is needed.

Next is the variance of the residuals. After subjecting the residuals to the bptest, the high p-value of 0.5766. As such, we cannot reject the null hypothesis and conclude the residuals could have constant variance(A.5.11).

We use the shapiro test to test for the distribution of the residuals. From A.5.12, the high p-value of 0.8595 shows that the residuals could follow a normal distribution.

However, when we check for influential points, ,we realised there are some potential influential points due to the hat matrix leverage value being larger than (2p/n)= 2*3/110 = 0.05454545. In total, there are 8 points as shown in A.5.13. As such, to remedy, we use robust regression to limit the influence of the points. While the RMSE is higher for the robust regression, we believe it is a good tradeoff to limit the impact of the influential points.

Hypothesis Testing

We aim to check if Men's Basketball Athletic Student Aid(X14) and Men's Basketball Coaching Staff(X17) have equal importance. We can do a GLT test between a reduced model and a full model.

$H_0: \beta_{14} = \beta_{17} = \beta_{new}$

$H_0: \beta_{14} \neq \beta_{17}$

Reduced model: $Y = \beta_{new} * (X14 + X17)$

Full model : $Y = \beta_{14} * (X14) + \beta_{17} * (X17)$

Analysis of Variance Table

Model 1: Y ~ X_new
Model 2: Y ~ X14 + X17

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	108	3.1341				
2	107	3.0988	1	0.035391	1.2221	0.2714

Table 5.2 Anova summary between reduced and full model

Conclusion:

Based on the anova table, it seems that the p-value is more than 0.05. As such, we cannot reject the null hypothesis and conclude that the Men's Basketball Athletic Student Aid(X14) and Men's Basketball Coaching Staff(X17) can have equal importance. As such, the linear impact of athletic student aid and coaches salary on the win rate of a team can be the same. However, when looking at the overall model. It can be see that the new model still has an extremely low adj R^2 value. As such, it can be concluded that the combined variable of Men's Basketball Athletic Student Aid and Men's Basketball Coaching Staff does not explain the variance of the win rate of the basketball team. Money does not play a part in the basketball team's success.

Overall Appendix

Ryan Newman's Appendix

A.1.1

Complete R Analysis Process [can also refer to R markdown file]

```
library(ALSM)
library(boot)
library(car)
library(caret)
library(leaps)
library(lmtest)
library(MASS)
set.seed(123)
```

First load data and call it cbb and create win-rate column. Add a small value for Box-Cox:

```
cbb$Y = cbb$W / cbb$G
cbb$Y = cbb$Y * 100
cbb$Y = cbb$Y + 0.0001
```

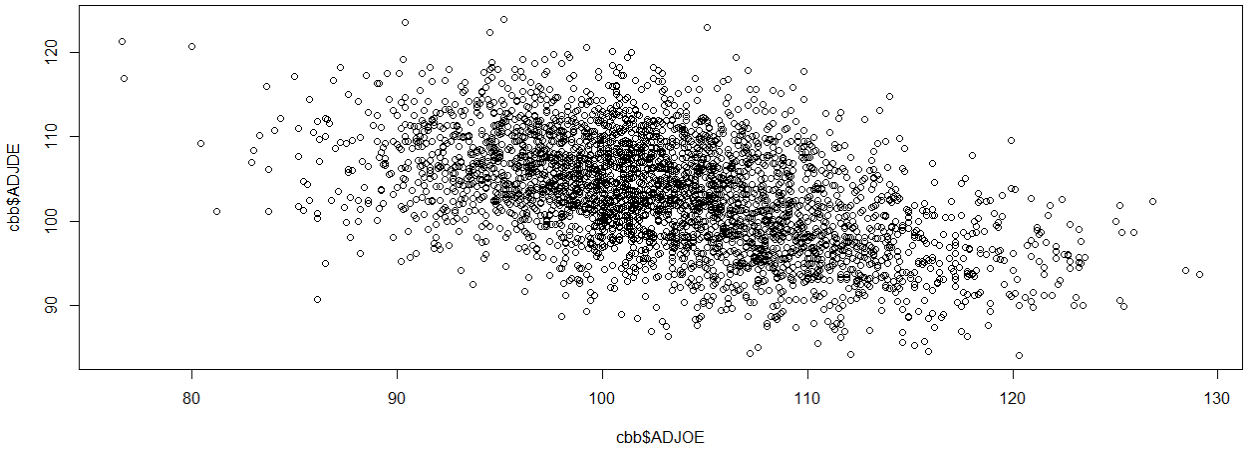
Then we make a column where values are 1 if it is in a Western Conference, 0 for Eastern Conference:

```
cbb$west = ifelse(cbb$CONF %in% c("B12","WCC","P12","BSky","BW","MWC","Sum","Slnd","SWAC",
"WAC"), 1, 0)
```

A.1.2

Checking for potential multicollinearity by computing correlation and plotting:

```
cor(cbb$ADJOE, cbb$ADJDE)
[1] -0.4940387
plot(cbb$ADJOE, cbb$ADJDE)
```

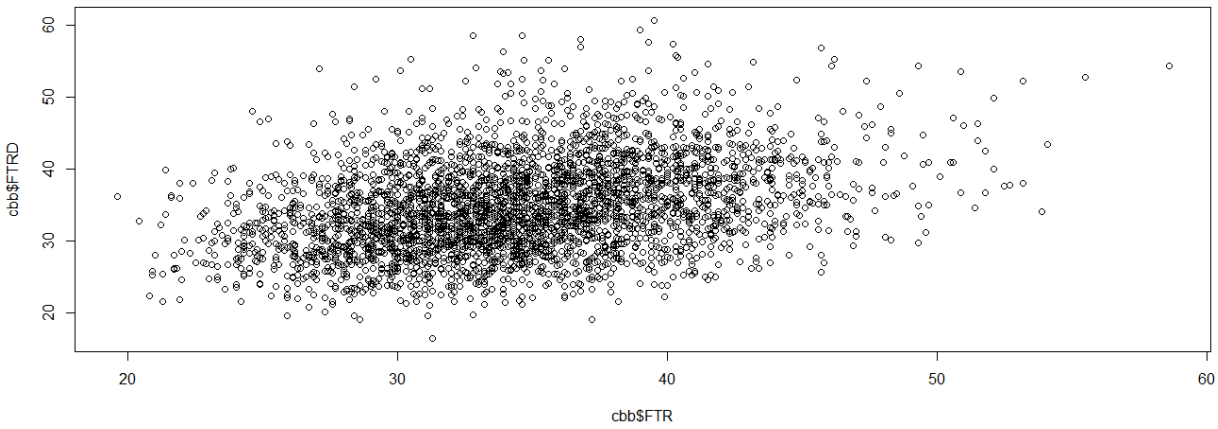


We are interested in offensive and defensive efficiency because those seem like variables that may potentially differ between Eastern and Western Conferences. For example, it is possible that teams in the East may tend to have higher offensive efficiency than teams in the West and vice-versa. The correlation value of -0.494 is moderate and while the plot has a negative sloping pattern, it doesn't look extremely

diagonal, so there likely isn't a multicollinearity issue. Also, the negative slope makes sense in the context of the data because having a high Offensive Efficiency probably indicates that the team may have a lower Defensive Efficiency, but it isn't enough to reduce the Defensive Efficiency by a noticeably large amount, which is why the correlation is only moderately negative.

Free Throw Rate and Free Throw Rate Allowed may also be relevant because it is possible for Western conference teams to have more or less fouling compared to Eastern Conference teams.

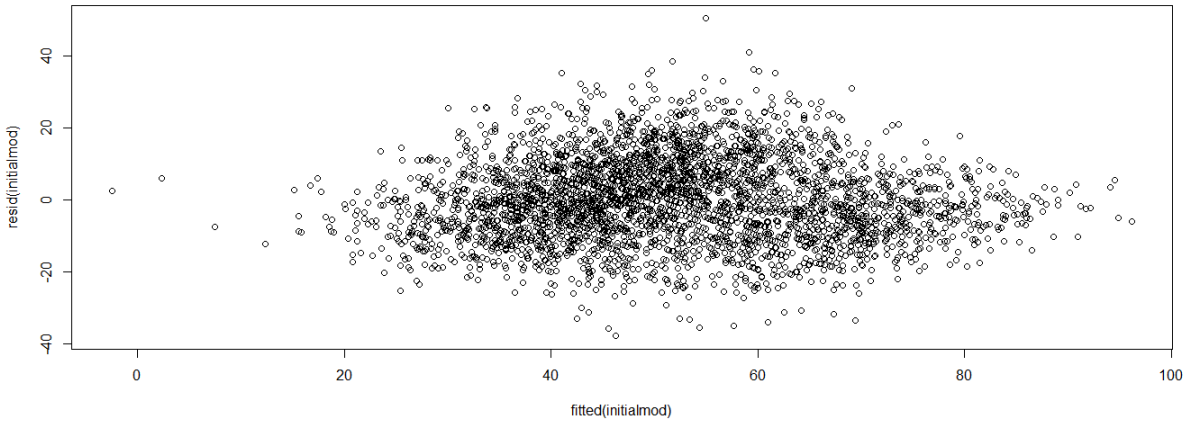
```
> cor(cbb$FTR, cbb$FTRD)
[1] 0.3452293
> plot(cbb$FTR, cbb$FTRD)
```



The correlation value is 0.345 which looks like a low moderate level of correlation and while the plot has a positive, upward sloping pattern, it doesn't look extremely diagonal, so there likely isn't a multicollinearity issue. In the context of the data, this positive correlation is interesting because it indicates that the general trend is that as a team increases its rate of shooting free throws (in other words, is able to draw opponents into committing more fouls against them), the amount of free throws it enables opposing teams to shoot will also increase (it fouls its opponents more). This is quite surprising because intuitively it would seem that a team with a higher free throw rate would be more aware of committing fouls, and thus would have a lower free throw rate allowed, but as we see from the correlation, this is not the case.

A.1.3

```
initialmod = lm(Y~ADJOE+ADJDE+FTR+FTRD+west+west*ADJOE + west*ADJDE + west*FTR + west*FTRD, cbb)
plot(fitted(initialmod), resid(initialmod))
```



Residual plot indicates violation of constant variance assumption.

```
library(lmtest)
```

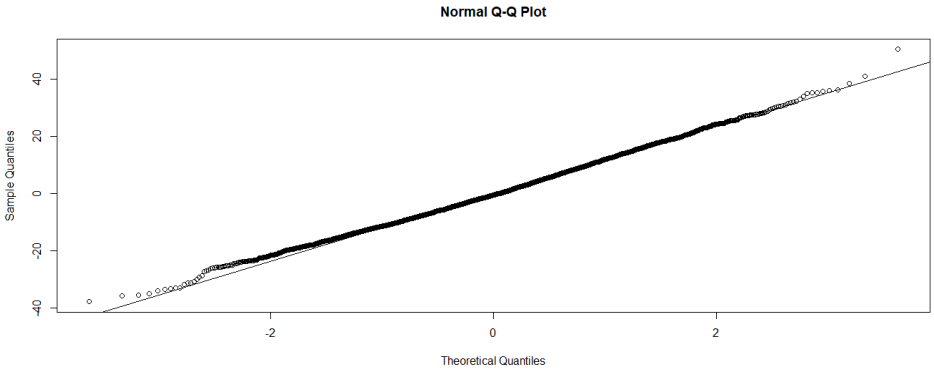
```
> bptest(initialmod)

        studentized Breusch-Pagan test

data:  initialmod
BP = 31.299, df = 9, p-value = 0.000263
```

The BP test indicates violation of constant variance assumption.

The Shapiro test indicates violation of normality assumption:

```
> shapiro.test(resid(initialmod))

        Shapiro-Wilk normality test

data:  resid(initialmod)
W = 0.99749, p-value = 1.608e-05
```

```
qqnorm(resid(initialmod))
qqline(resid(initialmod))
```



A.1.4

Our dataset is large (n = 3523) so a point is influential if the absolute value of DFFITS >

$$2\sqrt{\frac{p}{n}} = 2\sqrt{\frac{10}{3523}} = 0.106555$$

```
mydffits = dffits(initialmod)
mydffits = abs(mydffits)
sort(mydffits, decreasing=TRUE)
```

```
      2530       2773       3209       2512       2130       2796       2292       3010       3020       1954       2196       1494       2686        889
0.33929833 0.28927341 0.28422962 0.25323616 0.24569740 0.23727879 0.23635745 0.22322163 0.22127206 0.21721115 0.21605927 0.21389608 0.21375299 0.19222702
      2855       1309       2772       3001       2997       1915        313       3381       1285       1695       1523       2317       2961       2532
0.18820483 0.18676895 0.18196121 0.18117078 0.17924689 0.17874468 0.17743168 0.16866799 0.16838143 0.16796415 0.16650746 0.16625660 0.16577460 0.16573361
      2716       1425       1938       1346        310       1702       3492       1895       2653       2790       3230       2563       3358       1649
0.16463593 0.16345814 0.16306644 0.16021928 0.15981829 0.15966209 0.15958346 0.15897483 0.15856007 0.15802636 0.15784817 0.15592429 0.15434514 0.15425511
      2873       1442       2551        309       2304        311       1871       2795       2226       2305       1307       2939       1474       2594
0.15419779 0.15021749 0.14693959 0.14642498 0.14608718 0.14562411 0.14468129 0.14449668 0.14386359 0.14277503 0.14261323 0.14212747 0.14209956 0.14204199
       898       2291       2528       3352       1403       2618       2227       1410       2282       2672       2293       3517       1460       1629
0.14024188 0.14006927 0.14005108 0.13977910 0.13936725 0.13879357 0.13865596 0.13812202 0.13735658 0.13552002 0.13531706 0.13473673 0.13422755 0.13321289
      3073       2870        927       3512       2509        306        307       1672       3195       2513        917       1572       1907       3006
0.13306774 0.13292949 0.13214066 0.13211791 0.13184529 0.13171434 0.13128977 0.13112996 0.13110543 0.13097526 0.13026949 0.12975238 0.12926509 0.12920355
      2526       1383       2516       2678       2297        965       3379        382       3150       3221       2085        305       1296       2333
0.12792246 0.12765283 0.12718146 0.12601615 0.12600099 0.12569278 0.12561858 0.12540283 0.12517138 0.12515767 0.12439709 0.12375667 0.12321024 0.12304442
      2501       2504       2639       2642       2326       2934       3218       2912        308        378       2330       2980       3065       3090
0.12295160 0.12218079 0.12196471 0.12195911 0.12193704 0.12191830 0.12190099 0.12160057 0.11975294 0.11828843 0.11734608 0.11692076 0.11637854 0.11621250
      1579       1412       1748       1705       1276        299       1694       3089       3226       2515       1747        607       2793       2708
0.11616748 0.11612873 0.11530098 0.11490178 0.11483435 0.11405630 0.11362042 0.11350187 0.11343588 0.11333882 0.11286706 0.11245757 0.11210939 0.11161675
      2630       1912       2851       1371       1374       3495       2658       2383       1712       2117        375        478       1330       2300
0.11160503 0.11157106 0.11152435 0.11145416 0.11092951 0.11085999 0.11017629 0.11006463 0.10967028 0.10952789 0.10937497 0.10889528 0.10880965 0.10872014
      2522       1489        291       1676        901       1479       3401       1922       1628       1377       1729
0.10862375 0.10842769 0.10839940 0.10830629 0.10825981 0.10796822 0.10786915 0.10732425 0.10679018 0.10671652 0.10668219
```

There are 10*14 + 11 = 151 influential points on a single fitted value.

```
> mycooks = cooks.distance(initialmod)

> qf(0.5, 10, 3523-10)
[1] 0.9343602
```

An influential points will have a cooks distance > 0.9343602

```
> sort(mycooks, decreasing=TRUE)
         2530          2773          3209
0.0114797949 0.0083479103 0.0080594939
```

The largest Cook's distance out of all of them is smaller than the threshold, so there are no influential points on all the fitted values.

For DFBETAS, an influential point will have an absolute value of DFBETAS > $\frac{2}{\sqrt{n}} = \frac{2}{\sqrt{3523}} = 0.0337$

```
mydfb = dfbetas(initialmod)

mydfb = abs(mydfb)
```

All of these numbers are the rows that are influential points for the intercept coefficient:

```
> unique(which(mydfb[, 1] > 0.0337))
  [1]  114  116  122  127  133  141  156  159  173  187  189  212  229  244  248  271  274  283  291  302  324  334  336  437  522  527  556  565  590  594
 [31]  604  607  664  670  690  712  714  749  753  776  804  805  815  834  841  842  849  850  957  980  984 1005 1116 1139 1201 1208 1213 1234 1242 1245
 [61] 1252 1490 1494 1511 1523 1532 1547 1564 1572 1587 1606 1628 1640 1642 1644 1645 1661 1690 1694 1703 1704 1705 1712 1713 1737 1741 1748 1761 1765 1788
 [91] 1829 1832 1833 1851 1856 1884 1892 1909 1910 1918 1931 1934 1947 1986 1990 2024 2091 2136 2177 2179 2206 2226 2227 2230 2234 2271 2279 2282 2308 2325
[121] 2330 2333 2343 2356 2358 2369 2372 2379 2383 2392 2413 2487 2490 2502 2505 2508 2513 2520 2523 2527 2528 2548 2551 2553 2557 2558 2592 2594 2627 2630
[151] 2642 2655 2658 2664 2685 2690 2715 2722 2736 2747 2761 2775 2789 2798 2800 2801 2877 2881 2896 2922 2944 2958 2965 2967 2968 2996 3001 3007 3009 3014
[181] 3037 3043 3118 3122 3124 3141 3158 3190 3211 3223 3342 3358 3398 3401 3403 3431 3436 3444 3458 3480 3491 3501 3508 3519
```

All of these numbers are the rows that are influential points for the ADJOE coefficient:

```
> unique(which(mydfb[, 2] > 0.0337))
  [1]   22   36  106  116  122  127  131  135  141  150  156  159  173  187  189  227  229  254  262  274  281  283  289  291  299  323  328  330  336  403
 [31]  437  522  527  590  594  607  664  670  680  749  776  781  785  804  813  814  838  841  849  931  975  980  984  993  998 1103 1116 1139 1149 1201
 [61] 1214 1230 1234 1242 1245 1252 1490 1494 1511 1523 1547 1552 1587 1640 1644 1645 1661 1703 1712 1741 1748 1761 1765 1788 1829 1832 1851 1856 1884 1896
 [91] 1909 1910 1918 1931 1933 1934 1947 1953 1986 1990 2047 2079 2083 2088 2091 2092 2105 2107 2177 2179 2206 2226 2227 2230 2234 2263 2308 2333 2369 2372
[121] 2379 2383 2392 2400 2402 2413 2447 2460 2461 2462 2486 2501 2509 2513 2520 2523 2525 2527 2528 2535 2548 2551 2553 2557 2558 2594 2642 2658 2664 2670
[151] 2673 2678 2685 2690 2714 2715 2722 2736 2747 2754 2775 2789 2800 2825 2869 2877 2896 2934 2945 2968 2996 3003 3007 3009 3037 3043 3118 3122 3124 3126
[181] 3135 3141 3158 3159 3190 3193 3197 3211 3212 3221 3223 3231 3237 3259 3284 3290 3337 3338 3401 3403 3431 3444 3501 3508 3519
```

All of these numbers are the rows that are influential points for the ADJDE coefficient:

```
> unique(which(mydfb[, 3] > 0.0337))
  [1]  114  116  119  122  143  164  173  187  189  244  246  252  271  274  291  302  303  324  330  334  415  437  497  522  527  530  539  556  559  565
 [31]  571  590  604  611  664  690  700  710  712  753  805  815  816  817  818  827  828  834  839  841  842  849  850  858  860  940  957  973 1005 1082
 [61] 1116 1139 1163 1194 1197 1208 1213 1216 1234 1245 1249 1494 1511 1523 1532 1547 1564 1572 1587 1606 1628 1642 1645 1657 1661 1669 1703 1704 1705 1711
 [91] 1712 1737 1741 1761 1765 1821 1829 1832 1833 1851 1856 1884 1892 1918 1928 1934 1950 1967 1979 1986 2024 2046 2104 2113 2121 2136 2227 2230 2231 2234
[121] 2258 2271 2279 2282 2333 2343 2350 2356 2358 2369 2372 2383 2396 2487 2490 2494 2502 2505 2508 2513 2524 2528 2539 2548 2551 2594 2610 2627 2630 2673
[151] 2703 2720 2727 2736 2761 2789 2798 2800 2829 2861 2880 2881 2896 2934 2965 2967 2968 2974 2996 3001 3009 3033 3043 3125 3190 3211 3214 3223 3232 3235
[181] 3239 3398 3424 3431 3436 3444 3457 3458 3491 3501 3508 3519 3523
```

All of these numbers are the rows that are influential points for the FTR coefficient:

```
> unique(which(mydfb[, 4] > 0.0337))
  [1]   87  107  115  121  122  128  129  130  134  141  148  150  158  181  185  264  269  299  332  334  338  418  421  443  497  505  539  548  565  571
 [31]  575  582  590  592  607  619  664  682  711  714  748  781  787  818  827  851  947  965  972  975  984  993  998 1003 1121 1166 1168 1189 1204 1212
 [61] 1213 1214 1227 1240 1248 1249 1252 1477 1488 1489 1494 1523 1532 1539 1540 1559 1560 1564 1575 1620 1628 1636 1660 1689 1694 1712 1713 1714 1715 1737
 [91] 1741 1744 1761 1766 1804 1851 1885 1901 1917 1929 1934 1940 1953 1980 1986 2040 2046 2079 2083 2088 2091 2104 2105 2116 2128 2166 2178 2219 2221 2226
[121] 2227 2237 2248 2280 2281 2282 2325 2326 2330 2331 2350 2367 2392 2447 2448 2462 2490 2501 2508 2509 2513 2524 2533 2538 2548 2550 2551 2552 2553 2557
[151] 2572 2574 2592 2627 2642 2646 2673 2678 2685 2703 2712 2714 2715 2728 2729 2754 2755 2756 2786 2789 2801 2802 2843 2854 2856 2858 2869 2879 2896 2922
[181] 2934 2935 2937 2943 2958 2963 2965 2974 2991 3001 3003 3043 3089 3103 3197 3214 3221 3223 3284 3290 3337 3350 3358 3401 3430 3456 3469 3479 3480 3491
[211] 3509 3511 3513 3514 3520 3523
```

All of these numbers are the rows that are influential points for the FTRD coefficient:

```
> unique(which(mydfb[, 5] > 0.0337))
  [1]   58   59  107  113  130  131  134  141  145  147  158  174  181  199  205  212  229  248  262  264  281  289  292  299  323  327  328  330  331  423
 [31]  497  498  523  527  539  556  575  578  582  590  605  609  612  619  666  710  728  754  781  826  827  837  841  849  856  931  947  963  964  965
 [61]  974  990  993 1006 1121 1132 1163 1172 1213 1215 1216 1237 1248 1249 1251 1252 1487 1489 1494 1523 1528 1540 1543 1547 1551 1552 1560 1572 1628 1646
 [91] 1651 1652 1659 1663 1667 1669 1685 1694 1703 1704 1705 1706 1709 1714 1748 1770 1839 1896 1901 1902 1903 1910 1911 1917 1918 1929 1932 1933 1939 1940
[121] 1953 1967 1979 1986 2040 2079 2087 2088 2091 2098 2104 2168 2169 2206 2221 2226 2227 2234 2277 2280 2282 2326 2333 2367 2379 2383 2396 2462 2487 2508
[151] 2510 2513 2514 2520 2521 2523 2524 2528 2539 2548 2550 2551 2564 2572 2594 2632 2642 2658 2664 2670 2673 2678 2685 2707 2719 2729 2747 2749 2754 2755
[181] 2756 2757 2761 2763 2775 2781 2792 2794 2797 2798 2802 2829 2856 2858 2863 2866 2869 2877 2913 2926 2944 2974 2986 2992 2998 3001 3003 3007 3014 3037
[211] 3045 3089 3093 3103 3117 3141 3156 3159 3181 3210 3214 3220 3221 3228 3232 3234 3235 3244 3258 3284 3337 3342 3350 3358 3383 3436 3444 3496 3497 3523
```

All of these numbers are the rows that are influential points for the West coefficient:

```
> unique(which(mydfb[, 6] > 0.0337))
  [1]  116  122  189  291  302  307  310  311  313  352  362  365  366  375  378  384  395  452  455  459  470  478  522  889  901  917 1034 1045 1050 1245
 [31] 1256 1261 1273 1278 1279 1281 1291 1309 1322 1346 1352 1371 1410 1412 1418 1425 1444 1446 1460 1462 1474 1494 1520 1523 1570 1574 1629 1638 1649 1672
 [61] 1677 1703 1705 1710 1712 1717 1747 1764 1790 1813 1819 1832 1871 1888 1895 1907 1910 1912 1915 1922 1934 1938 1949 1954 1955 1960 1981 1985 1986 1996
 [91] 2071 2117 2129 2130 2191 2196 2198 2199 2212 2223 2227 2230 2291 2293 2304 2305 2317 2333 2353 2370 2372 2383 2392 2431 2445 2512 2513 2515 2516 2522
[121] 2528 2530 2532 2544 2551 2562 2563 2565 2576 2594 2618 2639 2642 2686 2713 2716 2736 2743 2744 2773 2775 2790 2793 2796 2800 2811 2851 2855 2868 2870
[151] 2873 2892 2893 2912 2921 2957 2961 2964 2980 2996 2997 3010 3013 3020 3021 3040 3043 3065 3081 3101 3171 3195 3215 3218 3226 3230 3254 3286 3352 3355
[181] 3408 3431 3467 3478 3490 3491 3492 3512 3517 3519 3521
```

All of these numbers in the output are the rows that are influential points for the coefficient for interaction between West and ADJOE:

unique(which(mydfb[, 7] > 0.0337))

All of these numbers in the output are the rows that are influential points for the coefficient for interaction between West and ADJDE:

unique(which(mydfb[, 8] > 0.0337))

All of these numbers in the output are the rows that are influential points for the coefficient for interaction between West and FTR:

unique(which(mydfb[, 9] > 0.0337))

All of these numbers in the output are the rows that are influential points for the coefficient for interaction between West and FTRD:

unique(which(mydfb[, 10] > 0.0337))

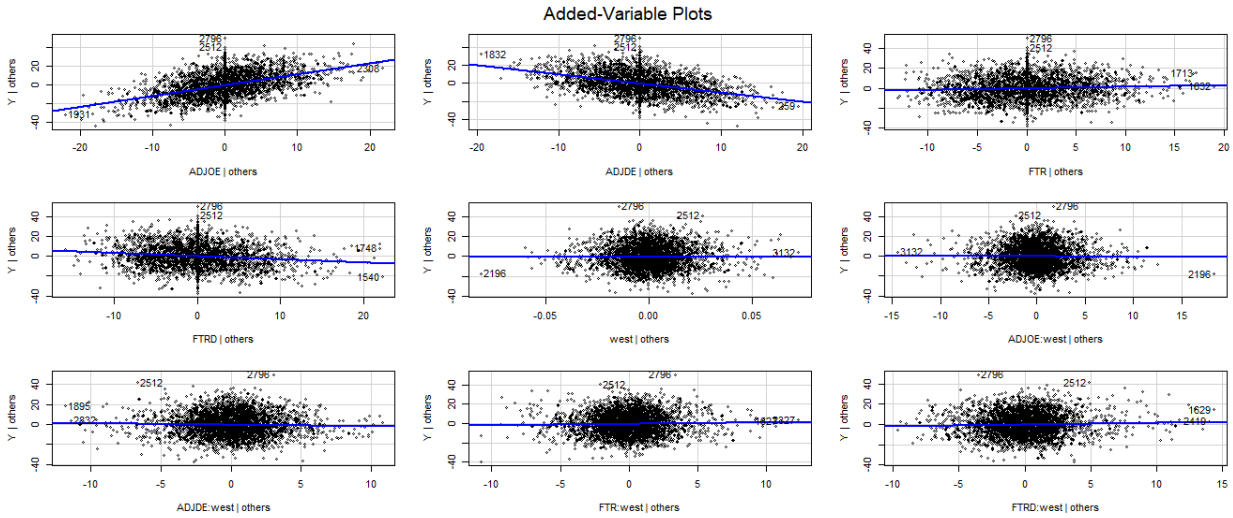
```
> vif(lm(Y~ADJOE+ADJDE+FTR+FTRD+west,cbb))
   ADJOE    ADJDE      FTR     FTRD     west
1.462299 1.329827 1.205103 1.320981 1.018714
```

This indicates no excessive multicollinearity.

```
library(car)
```

A.1.5

```
avPlots(initialmod)
```



Based on the added-variable plots, it looks like FTR, west, and west and the interaction it has with FTR, ADJOE, ADJDE, and FTRD are not needed given that all other predictors are considered in the model.

A.1.6

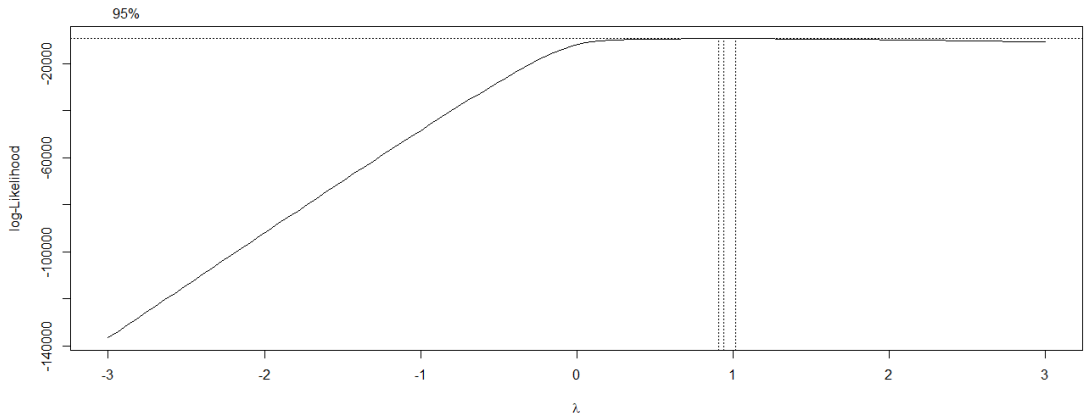
Now we try transforming Y:

```
library(MASS)

bcmle = boxcox(initialmod, lambda=seq(-3, 3, by = 0.1))

> bcmle$x[which.max(bcmle$y)]
[1] 0.9393939

cbb$transformY = cbb$Y^0.9393939
```

```
> bptest(lm(transformY~ADJOE+ADJDE+FTR+FTRD+west+west*ADJOE+west*ADJDE+west*FTR+west*FTRD, cbb))

        studentized Breusch-Pagan test

data:  lm(transformY ~ ADJOE + ADJDE + FTR + FTRD + west + west * ADJOE +     west * ADJDE + west * FTR + west * FTRD, cbb)
BP = 32.949, df = 9, p-value = 0.0001363
```

BP-test for model using transformed Y has a worse result because the p-value is smaller.

Lambda value = 1 is also within the best lambda value range so we choose to not transform



Diagnosed issues are: non-normal, non-constant variance, and existence of outlier terms.

A.1.7

Now we move to model selection. After selecting, we will diagnose again, and then apply remedial methods:

```
> library(ALSM)
> bs = BestSub(cbb[, c(5, 6, 14, 15, 26)], cbb$Y, num = 5)
> bs
   p 1 2 3 4 5      SSEp           r2        r2.adj          Cp      AICp     SBCp    PRESSp
1  2 1 0 0 0 0  610339.8 0.4750227364  0.4748736374 1056.457488 18164.02 18176.35  610978.5
1  2 0 1 0 0 0  696490.2 0.4009213739  0.4007512294 1702.290474 18629.19 18641.52  697208.1
1  2 0 0 0 1 0 1072871.4 0.0771811610  0.0769190710 4523.859489 20151.27 20163.60 1074083.8
1  2 0 0 1 0 0 1147251.6 0.0132037707  0.0129235104 5081.456647 20387.41 20399.75 1148540.9
1  2 0 0 0 0 1 1162485.0 0.0001009117 -0.0001830698 5195.655066 20433.89 20446.22 1163856.6
2  3 1 1 0 0 0  478591.1 0.5883449425  0.5881110476   70.793883 17309.33 17327.83  479318.3
2  3 1 0 0 1 0  604802.8 0.4797853004  0.4794897239 1016.949196 18133.91 18152.41  605798.3
2  3 1 0 1 0 0  607504.5 0.4774615307  0.4771646338 1037.202092 18149.61 18168.12  608476.2
2  3 1 0 0 0 1  610266.1 0.4750861688  0.4747879223 1057.904640 18165.59 18184.09  611276.5
2  3 0 1 0 1 0  663307.9 0.4294628232  0.4291386543 1455.536486 18459.21 18477.72  664393.5
3  4 1 1 0 1 0  474509.4 0.5918558531  0.5915079041   42.194420 17281.16 17305.83  475508.6
3  4 1 1 1 0 0  477124.2 0.5896066960  0.5892568296   61.797026 17300.52 17325.19  478112.0
3  4 1 1 0 0 1  478177.3 0.5887008705  0.5883502318   69.691780 17308.29 17332.95  479192.8
3  4 1 0 1 1 0  596699.5 0.4867553202  0.4863177715  958.201748 18088.39 18113.06  598026.8
3  4 1 0 0 1 1  604508.5 0.4800384716  0.4795951967 1016.742674 18134.20 18158.87  605872.8
4  5 1 1 1 1 0  469694.8 0.5959970717  0.5955377165    8.101486 17247.23 17278.06  470950.2
4  5 1 1 0 1 1  473740.5 0.5925171411  0.5920538291   38.430942 17277.45 17308.28  475023.8
4  5 1 1 1 0 1  476851.8 0.5898410387  0.5893746840   61.754605 17300.51 17331.34  478126.5
4  5 1 0 1 1 1  596563.7 0.4868721447  0.4862887133  959.183560 18089.59 18120.42  598254.4
4  5 0 1 1 1 1  641652.0 0.4480898799  0.4474623528 1297.191729 18346.28 18377.11  643447.5
5  6 1 1 1 1 1  469147.6 0.5964676666  0.5958939784    6.000000 17245.12 17282.13  470683.4
```

Judge by AIC and SBC because of assumption violations. Judge by PRESS because we are interested in best predictive power. For AIC and PRESS, the full model is the best and the second best is the model with everything except west, and the third best is the model with everything except FTR. For SBC, the best model is the model with everything except west, 2nd best is full model, 3rd best is model without FTR and west.

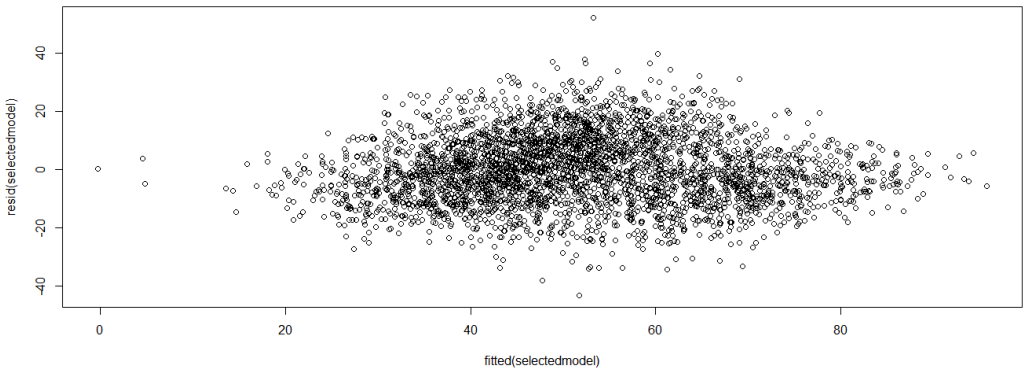
```
> step(initialmod, method="both", trace=1)
Start:  AIC=17234.46
Y ~ ADJOE + ADJDE + FTR + FTRD + west + west * ADJOE + west *
    ADJDE + west * FTR + west * FTRD

             Df Sum of Sq    RSS   AIC
- ADJOE:west  1     90.59 466759 17233
<none>                    466669 17235
- ADJDE:west  1    419.73 467088 17236
- FTR:west    1    527.56 467196 17236
- FTRD:west   1    800.02 467469 17239

Step:  AIC=17233.14
Y ~ ADJOE + ADJDE + FTR + FTRD + west + ADJDE:west + FTR:west +
    FTRD:west

             Df Sum of Sq    RSS   AIC
<none>                    466759 17233
- ADJDE:west  1       329 467088 17234
- FTR:west    1       467 467226 17235
- FTRD:west   1      1088 467847 17239
- ADJOE       1    172269 639028 18338

Call:
lm(formula = Y ~ ADJOE + ADJDE + FTR + FTRD + west + ADJDE:west +
    FTR:west + FTRD:west, data = cbb)

Coefficients:
(Intercept)        ADJOE        ADJDE          FTR         FTRD         west   ADJDE:west     FTR:west    FTRD:west
    43.1517       1.1646      -1.0381       0.1893      -0.3265      -1.1272      -0.1032       0.1554       0.2035
```

Stepwise says to remove interaction between west and ADJOE. Because of the research goals, west must be in the model, and the interaction between west and ADJOE, west and ADJDE, must also be in the model. Based on this and the best subsets algorithm, we decide that the selected model will be the

initial model with FTR dropped, because that model performed 3rd best for AIC, SBC, PRESS, and meets the research goals. Plus, the avPlot indicated that FTR and interaction between FTR and west were not needed given that the other predictors were considered, so that is another reason to remove it.

A.1.8

```
selectedmodel = lm(Y ~ ADJOE + ADJDE + FTRD + west + west*ADJOE + west*ADJDE + west*FTRD, cbb)
plot(fitted(selectedmodel), resid(selectedmodel))
```




```
> bptest(selectedmodel)

        studentized Breusch-Pagan test

data:  selectedmodel
BP = 29.278, df = 7, p-value = 0.0001287
```

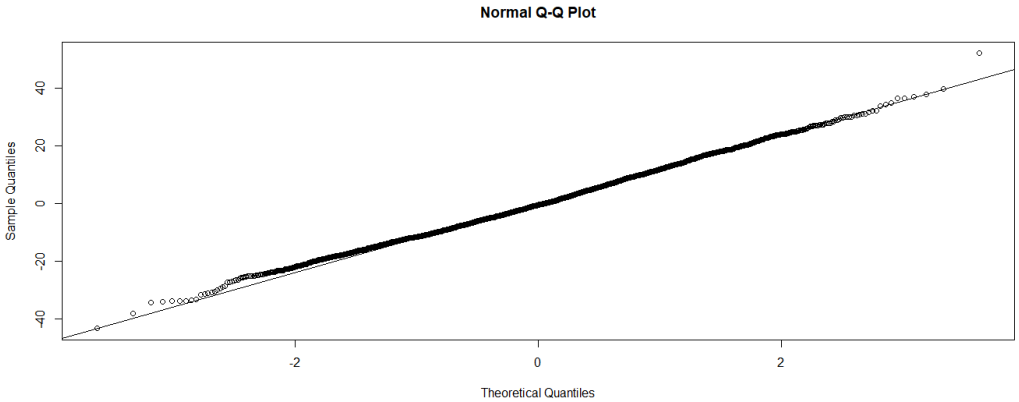
Violates non-constant variance.

```
> shapiro.test(resid(selectedmodel))

        Shapiro-Wilk normality test

data:  resid(selectedmodel)
W = 0.99767, p-value = 3.603e-05
```

Shapiro test indicates normality is violated.

qqnorm(resid(selectedmodel))

qqline(resid(selectedmodel))

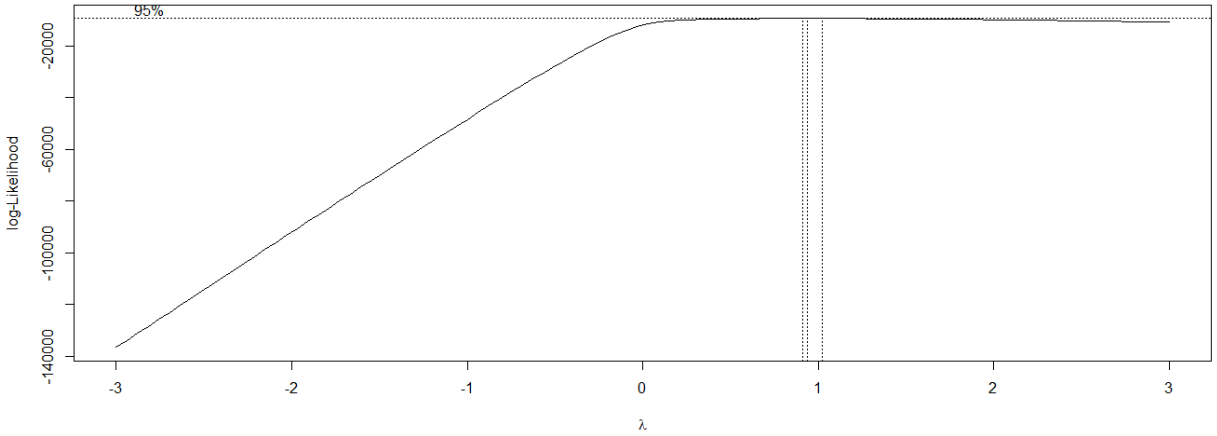


A.1.9

```
> bcmle = boxcox(selectedmodel, lambda=seq(-3, 3, by = 0.1))
> bcmle$x[which.max(bcmle$y)]
[1] 0.9393939
```

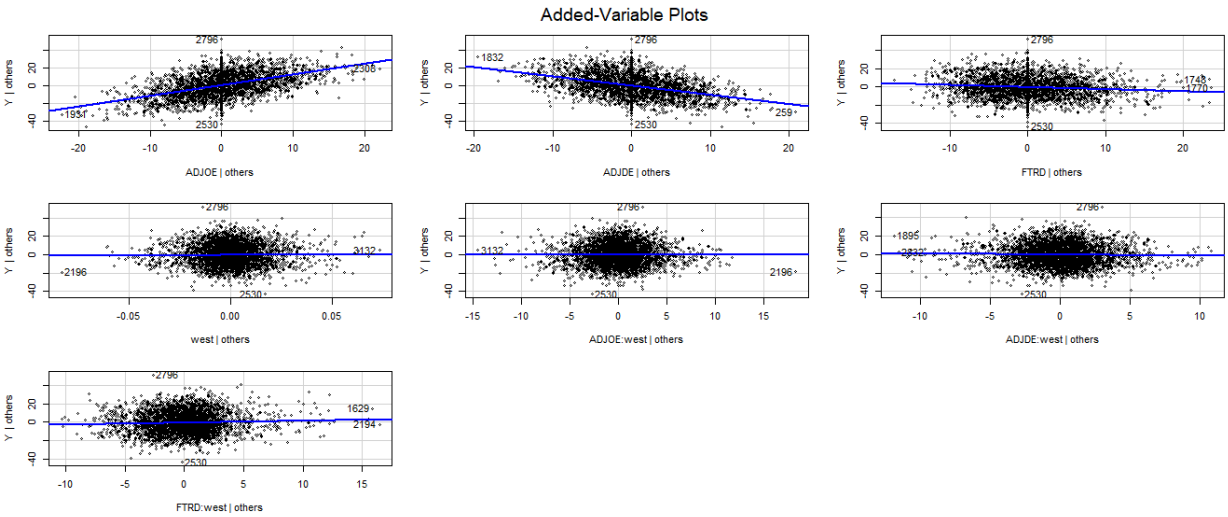
```
> bptest(lm(transformY ~ ADJOE + ADJDE + FTRD + west + west*ADJOE + west*ADJDE + west*FTRD, cbb))

        studentized Breusch-Pagan test

data:  lm(transformY ~ ADJOE + ADJDE + FTRD + west + west * ADJOE +     west * ADJDE + west * FTRD, cbb)
BP = 30.607, df = 7, p-value = 7.346e-05
```

Again, this is a worst result for regular Y and the plot shows that lambda = 1, is one of the top likelihoods:



A.1.10

avPlots(selectedmodel)



AvPlots indicate that west, and its interaction between ADJOE, ADJDE, and FTRD are not needed given all other predictors are included.

A.1.11

Remedial methods:

First we will look at the outliers. Our dataset is large (n = 3523) so a point is influential if the absolute value of DFFITS > $2\sqrt{\frac{p}{n}} = 2\sqrt{\frac{8}{3523}} = 0.095306$

```
thedffits = dffits(selectedmodel)
thedffits = abs(thedffits)
sort(thedffits, decreasing=TRUE)
```

All of these are influential points on single fitted values:

```
      2130       2292       2512       2196       2686       1954       3209       2796       3020       2773       2855       1494        889       1915
0.24029097 0.23703983 0.23650840 0.23076068 0.21794029 0.21732097 0.21370045 0.20474530 0.19694042 0.19281685 0.19020461 0.19007669 0.18888335 0.17621943
       313       1695       1895       1285       1938       2716       1523       3230       2873       2653       1442       1871       2530       2291
0.17443750 0.16706468 0.16660269 0.16293551 0.16249564 0.15911471 0.15751569 0.15671392 0.15571648 0.15062704 0.15003591 0.14972018 0.14596590 0.14288307
       309       2772       2594       2939       2618       2528        311       3001        898       1410       3352       2532       1460       2870
0.14267812 0.14170230 0.14133034 0.14116350 0.14109108 0.14059782 0.14056566 0.13806539 0.13724142 0.13695390 0.13422797 0.13338645 0.13315077 0.13201466
      1907       3073       1572       1672       2293       2516       3195       2297       1403        917       2526       2226       2678       1629
0.13004783 0.12936058 0.12922461 0.12834607 0.12790465 0.12771173 0.12753893 0.12697987 0.12613901 0.12545183 0.12533785 0.12382418 0.12359991 0.12358825
      2333       1474       2504       2513       3379       2961       1296       2708       3218       3512       3517       2639       2980        927
0.12278029 0.12205528 0.12190194 0.12188191 0.12180917 0.12155548 0.12138235 0.12127484 0.12094002 0.12076210 0.11883199 0.11864903 0.11793215 0.11778860
      2227       1579       2997       2912       1276        965       1397       2851       1747       2551       2326       2304       2630       3150
0.11715651 0.11687723 0.11601140 0.11504868 0.11452014 0.11334194 0.11331387 0.11196544 0.11175096 0.11161643 0.11129892 0.11110334 0.11038496 0.10952624
      1676       3226       1705       3089       3153       1540       1912       2642        849       2420       2790        375       1245       2305
0.10889663 0.10773610 0.10772826 0.10706227 0.10702891 0.10701693 0.10676995 0.10652572 0.10622033 0.10604678 0.10586993 0.10575975 0.10547025 0.10521694
      3256        291        455        316       2423       1371       1748       1352       1045       2829       3244       2782       1346       2658
0.10491932 0.10471031 0.10435014 0.10385399 0.10320980 0.10296135 0.10272280 0.10224348 0.10212639 0.10206725 0.10204877 0.10194741 0.10162717 0.10160418
      2877       2372       1425       3431       2544       1979       1560        366       2383       2864       1712       2793        478       1910
0.10144756 0.10123301 0.10112221 0.10083243 0.10064301 0.10042008 0.10041987 0.10029906 0.10020454 0.10005113 0.09988442 0.09954783 0.09952382 0.09947303
       378       2529       3358        310        306       2317       1479       2648       3381        328       1374        308       1412       3006
0.09946379 0.09942307 0.09928122 0.09898150 0.09881282 0.09881073 0.09819644 0.09795394 0.09782335 0.09774573 0.09759952 0.09743112 0.09740760 0.09687074
      2871       2795        320       1919        361       1377        957       2522       2713       1382       3043       2775       2858       1418
0.09678843 0.09677276 0.09669056 0.09665297 0.09659617 0.09641539 0.09636992 0.09636698 0.09634567 0.09620191 0.09617271 0.09575077 0.09560695 0.09509605
```

```
> qf(0.5, 8, 3523-8)
[1] 0.9181907
```

Any influential points will have a cooks distance > 0.9181907.

```
mycooks = cooks.distance(selectedmodel)
```

```
> sort(mycooks, decreasing=TRUE)
        2130         2292         2512
0.0072082690 0.0070112335 0.0069704502
```

The largest Cook's distance out of all of them is smaller than the threshold, so there are no influential points on all the fitted values.

For DFBETAS, an influential point will have an absolute value of DFBETAS > $\frac{2}{\sqrt{n}} = \frac{2}{\sqrt{3523}} = 0.0337$

```
mydfb = dfbetas(selectedmodel)
mydfb = abs(mydfb)
unique(which(mydfb[, 1] > 0.0337))
```

Doing from 1 to 8 shows that there are influential points on all the coefficients.

```
> vif(lm(Y~ADJOE+ADJDE+FTRD+west, cbb))
   ADJOE     ADJDE      FTRD      west
1.417951 1.325133 1.119859 1.014648
```

VIF indicates no excessive multicollinearity.

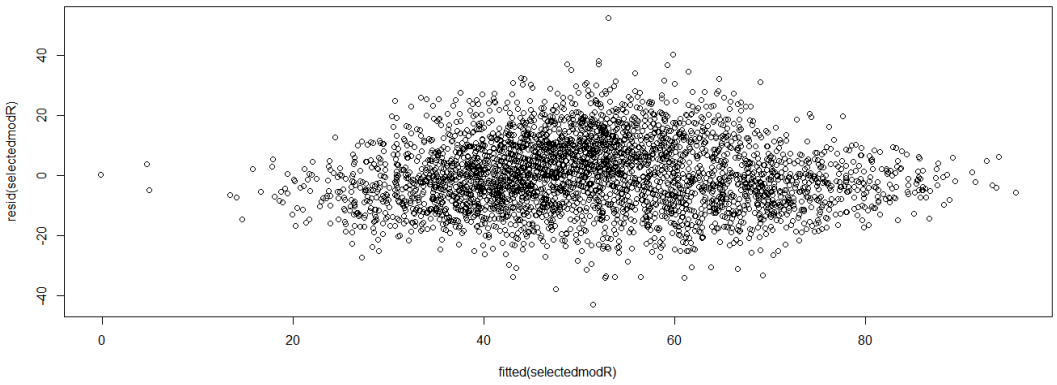
A.1.12

We do robust regression to dampen effect of outliers without removing them outright:

```
selectedmodR = rlm(Y~ADJOE+ADJDE+FTRD+west+west*ADJOE+west*ADJDE+west*FTRD, data=cbb, psi=psi.bisquare)
```

selectedmodR = rlm(Y~ADJOE+ADJDE+FTRD+west+west*ADJOE+west*ADJDE+west*FTRD, data=cbb,psi=psi.bisquare)

```
plot(fitted(selectedmodR), resid(selectedmodR))
```




```
> bptest(selectedmodR)

        studentized Breusch-Pagan test

data:  selectedmodR
BP = 29.278, df = 7, p-value = 0.0001287
```

This is a constant variance violation.

A.1.13

To address the constant variance violation, we do WLS:

```
wts1 = 1/fitted(lm(abs(residuals(selectedmodR))~Y, cbb))^2
```

```
> wts1 = 1/fitted(lm(abs(residuals(selectedmodR))~Y, cbb))^2
```

```
> selectedmodR2 = lm(Y~ADJOE+ADJDE+FTRD+west+west*ADJOE+west*ADJDE+west*FTRD, weight=wts1, data=cbb)
> bptest(selectedmodR2)

        studentized Breusch-Pagan test

data:  selectedmodR2
BP = 13.232, df = 7, p-value = 0.06665
```

selectedmodR2 = lm(Y~ADJOE+ADJDE+FTRD+west+west*ADJOE+west*ADJDE+west*FTRD, weight=wts1, data=cbb)

```
wts2 = 1/fitted(lm(abs(residuals(selectedmodR2))~Y, cbb))^2
```

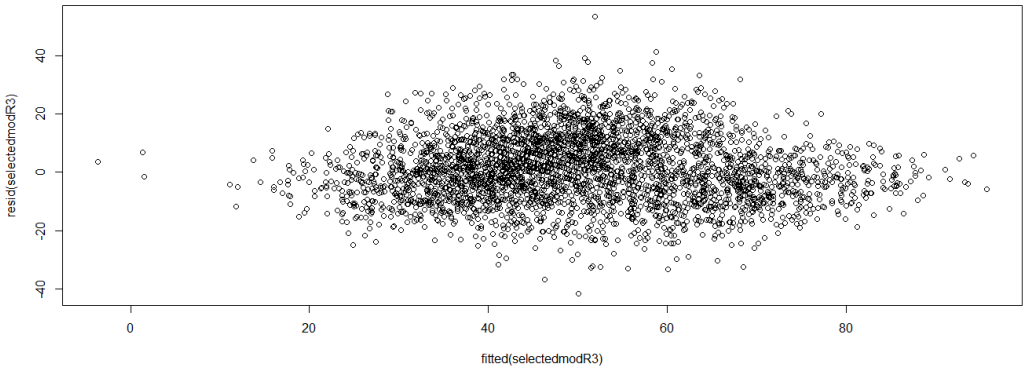
wts2 = 1/fitted(lm(abs(residuals(selectedmodR2))~Y, cbb))^2

```
> wts2 = 1/fitted(lm(abs(residuals(selectedmodR2))~Y, cbb))^2
> selectedmodR3 = lm(Y~ADJOE+ADJDE+FTRD+west+west*ADJOE+west*ADJDE+west*FTRD, weight=wts2, data=cbb)
> bptest(selectedmodR3)

        studentized Breusch-Pagan test

data:  selectedmodR3
BP = 12.892, df = 7, p-value = 0.07479
```

selectedmodR3 = lm(Y~ADJOE+ADJDE+FTRD+west+west*ADJOE+west*ADJDE+west*FTRD, weight=wts2, data=cbb)

A.1.14

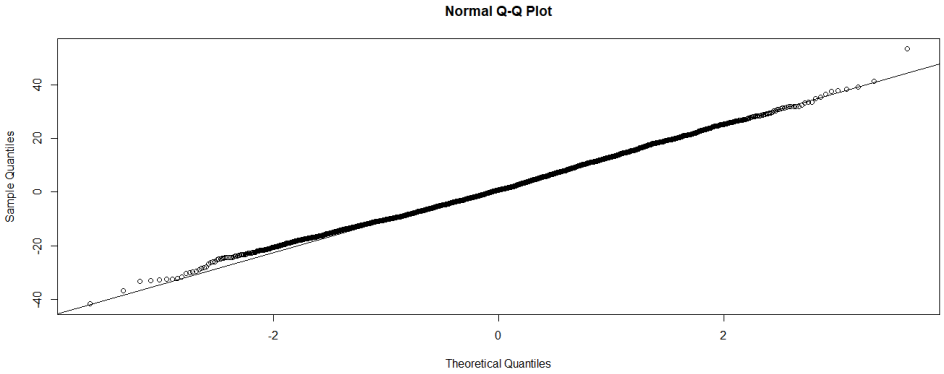
```
plot(fitted(selectedmodR3), resid(selectedmodR3))
```




```
> shapiro.test(resid(selectedmodR3))

        Shapiro-Wilk normality test

data:  resid(selectedmodR3)
W = 0.9976, p-value = 2.58e-05
```

The normality test failing error persists.

```
qqnorm(resid(selectedmodR3))
qqline(resid(selectedmodR3))
```




```
> selectedmodR3

Call:
lm(formula = Y ~ ADJOE + ADJDE + FTRD + west + west * ADJOE +
    west * ADJDE + west * FTRD, data = cbb, weights = wts2)

Coefficients:
(Intercept)        ADJOE        ADJDE         FTRD         west   ADJOE:west   ADJDE:west    FTRD:west
 39.0208906    1.2511877   -1.0618226   -0.2433003    2.6151519    0.0003846   -0.0970919    0.2311959
```

A.1.15

Finally, we test the hypotheses. We are interesting in testing the significance of interaction between west and ADJOE, and also testing interaction between west and ADJDE:

Let X1 = ADJOE, X2 = ADJDE, X3 = FTRD, X4 = west

This tests the significance of interaction between west and ADJOE

H_0: $\beta_{1,4} = 0$ given all other predictors are in the model

Reduced model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_{2,4} X_2 X_4 + \beta_{3,4} X_3 X_4$

$H_A: \beta_{1,4} \neq 0$ given all other predictors are in the model

Full model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_{1,4} X_1 X_4 + \beta_{2,4} X_2 X_4 + \beta_{3,4} X_3 X_4$

```
fullmod = selectedmodR3
```

```
> reducedmod = lm(Y~ADJOE+ADJDE+FTRD+west+west*ADJDE+west*FTRD, data = cbb, weights=wts2)
> reducedmod

Call:
lm(formula = Y ~ ADJOE + ADJDE + FTRD + west + west * ADJDE +
    west * FTRD, data = cbb, weights = wts2)

Coefficients:
(Intercept)        ADJOE        ADJDE         FTRD         west   ADJDE:west    FTRD:west
   39.00239      1.25130     -1.06177     -0.24327      2.67912     -0.09729      0.23109

> fullmod

Call:
lm(formula = Y ~ ADJOE + ADJDE + FTRD + west + west * ADJOE +
    west * ADJDE + west * FTRD, data = cbb, weights = wts2)

Coefficients:
(Intercept)        ADJOE        ADJDE         FTRD         west   ADJOE:west   ADJDE:west    FTRD:west
 39.0208906    1.2511877   -1.0618226   -0.2433003    2.6151519    0.0003846   -0.0970919    0.2311959

> anova(reducedmod, fullmod)
Analysis of Variance Table

Model 1: Y ~ ADJOE + ADJDE + FTRD + west + west * ADJDE + west * FTRD
Model 2: Y ~ ADJOE + ADJDE + FTRD + west + west * ADJOE + west * ADJDE +
    west * FTRD
  Res.Df    RSS Df  Sum of Sq  F Pr(>F)
1   3516 5508.2
2   3515 5508.2  1 4.8706e-05  0 0.9956
```

So, based on the p-value we conclude that there is no significant interaction effect between conference and ADJOE. Also, the fact that the p-value is extremely high shows that this interaction effect is very insignificant. In the context of the data, this means that when predicting a team's win rate, whether a team is in the western conference or not has no effect on the linear impact of ADJOE on win rate. In other words, when trying to predict a team's win rate, the interaction between conference and ADJOE doesn't matter and should not be considered.

This tests the significance of interaction between conference and ADJDE

$H_0: \beta_{2,4} = 0$ given all other predictors are in the model

Reduced model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_{1,4} X_1 X_4 + \beta_{3,4} X_3 X_4$

$H_A: \beta_{2,4} \neq 0$ given all other predictors are in the model

Full model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_{1,4} X_1 X_4 + \beta_{2,4} X_2 X_4 + \beta_{3,4} X_3 X_4$

```
> reducedmod = lm(Y~ADJOE+ADJDE+FTRD+west+west*ADJOE+west*FTRD, data = cbb, weights=wts2)
> anova(reducedmod, fullmod)
Analysis of Variance Table

Model 1: Y ~ ADJOE + ADJDE + FTRD + west + west * ADJOE + west * FTRD
Model 2: Y ~ ADJOE + ADJDE + FTRD + west + west * ADJOE + west * ADJDE +
    west * FTRD
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1   3516 5510.9
2   3515 5508.2  1    2.7168 1.7337  0.188
```

So, based on the p-value we conclude that there is no significant interaction effect between conference and ADJDE. In the context of the data, this means that when predicting a team's win rate, whether or not a team is in the western conference or not has no effect on the linear impact of ADJDE on win rate. In other words, when trying to predict a team's win rate, the interaction between conference and ADJDE

doesn't matter and should not be considered. So, overall, the bigger picture conclusion is that between western schools and eastern schools, the impact of ADJOE and ADJDE does not differ significantly between the two. Another interesting finding is that the largest estimated coefficient was for conference, which suggests that among the ADJOE, ADJDE, west, and FTRD, the one that has the biggest impact on win rate is whether a team is a western school or not. Furthermore, this means that holding all other variables constant, being in a western school results in an increase of 2.615 in win rate compared to the baseline of eastern schools.

A.1.16

```
library(MASS)
library(leaps)
library(caret)

set.seed(123)
train.control = trainControl(method="cv", number=10)
cvmodel = train(
    formula(selectedmodR3),
    data = cbb,
    method = "lm",
    trControl = train.control
)
```

```
> cvmodel
Linear Regression

3523 samples
   4 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 3170, 3170, 3171, 3172, 3170, 3171, ...
Resampling results:

  RMSE      Rsquared   MAE
  11.59519  0.5914609  9.304987
```

After crossvalidation we obtain: RMSE = 11.59519, Rsquared = 0.5914609, MAE = 9.304987

A.1.17

Bootstrap:

```
library(MASS)
boot.huber <- function(data, indices, maxit=3523) {
    data = data[indices,]
    mod = lm(Y~ADJOE+ADJDE+FTRD+west+west*ADJOE+west*ADJDE+west*FTRD, data=data, weights=wts2, maxit=maxit)
    coefficients(mod)
}
library(boot)
project.boot = boot(data = cbb, statistic = boot.huber, R=3523, maxit=3523)
```

Bootstrap Confidence interval for β_0: $(29.18,\ 57.21)$

```
boot.ci(boot.out = project.boot, type = "perc", index = 1)

Intervals :
Level     Percentile
95%   (29.18, 57.21 )
```

The Bootstrap Confidence interval for linear impact of ADJOE, is very precise. Also, all the values being positive makes sense, because if a team has a strong offense, they are likely to win more games β_1: $(1.137,\ 1.284)$

```
boot.ci(boot.out = project.boot, type = "perc", index = 2)

Intervals :
Level     Percentile
95%   ( 1.137,  1.284 )
```

The Bootstrap Confidence interval for linear impact of ADJDE, is also very precise. All the values in the interval are negative which makes sense, because the more points allowed, the more likely it is a team loses because they are letting their opponent score more points β_2: $(-\ 1.119,\ -\ 0.969)$:

```
boot.ci(boot.out = project.boot, type = "perc", index = 3)

Intervals :
Level     Percentile
95%   (-1.119, -0.969 )
```

The Bootstrap Confidence interval for β_3: $(-\ 0.3307,\ -\ 0.1733)$ All the values in this interval are negative which makes sense, because if a team allows a lot of free throws, that means they allow their opponents to score more points, which would result in more losses.

```
boot.ci(boot.out = project.boot, type = "perc", index = 4)

Intervals :
Level     Percentile
95%   (-0.3307, -0.1733 )
```

The Bootstrap Confidence interval for effect of conference alone, is not very precise β_4: $(-\ 18.123,\ 35.373)$:

```
boot.ci(boot.out = project.boot, type = "perc", index = 5)

Intervals :
Level     Percentile
95%   (-18.123,  35.373 )
```

The Bootstrap Confidence interval for interaction between ADJOE and conference suggests that the true value may be 0, which means no impact, and this is supported by the hypothesis testing earlier $\beta_{1,4}$: $(-\ 0.1694,\ 0.1045)$:

```
boot.ci(boot.out = project.boot, type = "perc", index = 6)

Intervals :
Level     Percentile
95%   (-0.1694,  0.1045 )
```

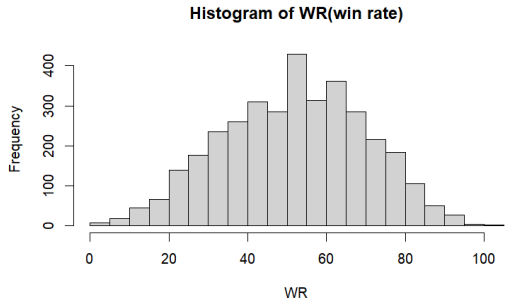
The Bootstrap Confidence interval for interaction between ADJDE and conference suggests that the true value may be 0, which means no impact, and this is supported by the hypothesis testing earlier $\beta_{2,4}$: $(-\ 0.259,\ 0.0277)$:

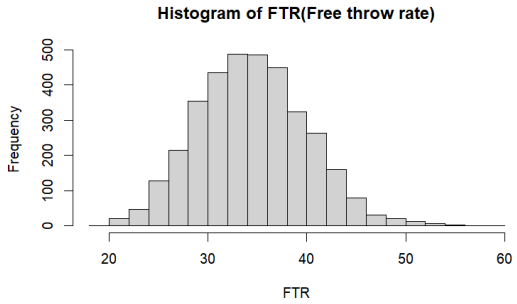
```
boot.ci(boot.out = project.boot, type = "perc", index = 7)

Intervals :
Level     Percentile
95%   (-0.2590,  0.0277 )
```

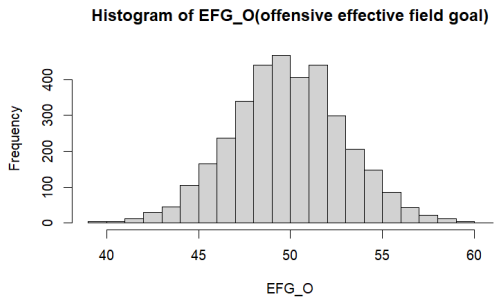
The Bootstrap Confidence interval for $\beta_{3,4}$: $(0.0725,\ 0.3593)$: This confidence suggests that being in the western conference reduces the negative impact of FTRD on winrate because all the values in the interval are positive, and the FTRD has all negative values. This suggests that giving up free throws matters less in the western conference.

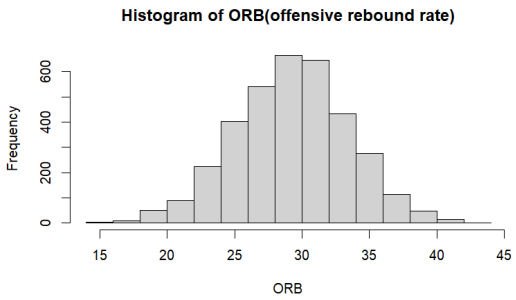
```
boot.ci(boot.out = project.boot, type = "perc", index = 8)

Intervals :
Level     Percentile
95%   ( 0.0725,  0.3593 )
```

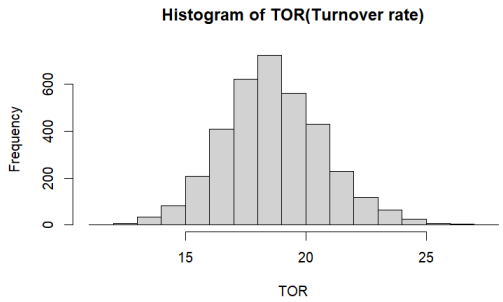
Jerry Huang's Appendix

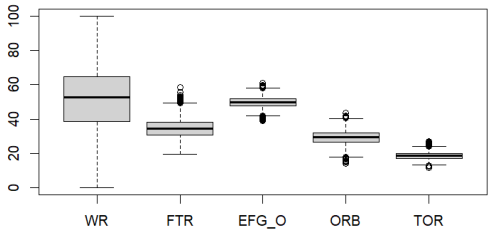








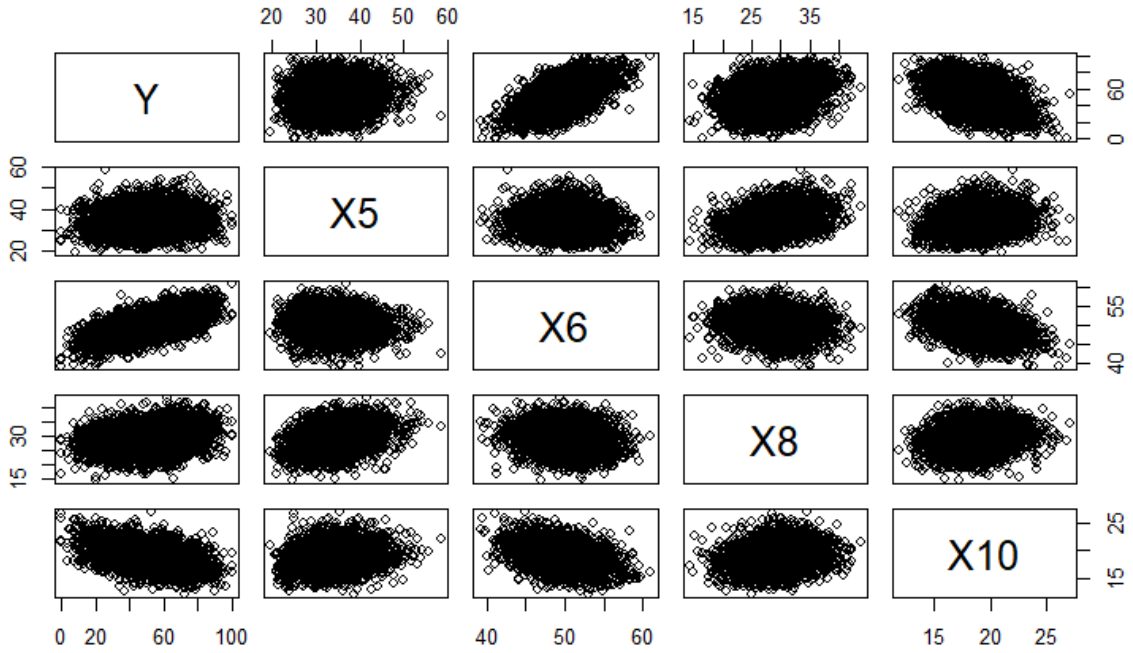




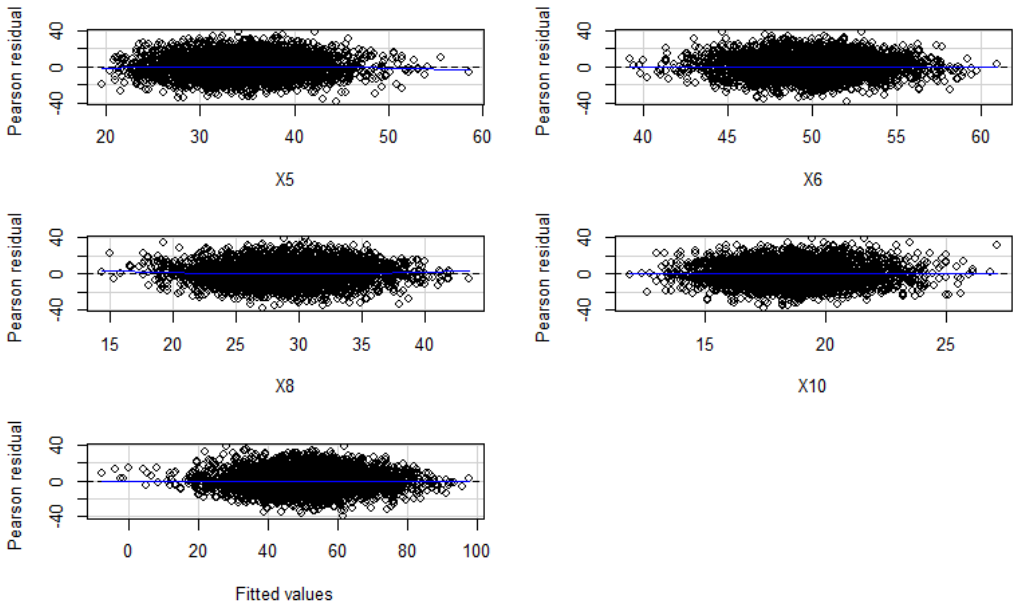
A.2.1 various histogram and boxplot

	Y	X5	X6	X8	X10
Y	1.0000000	0.11432167	0.61745443	0.2633390	-0.4311758
X5	0.1143217	1.00000000	-0.06681857	0.3344543	0.1618832
X6	0.6174544	-0.06681857	1.00000000	-0.1453579	-0.3710900
X8	0.2633390	0.33445432	-0.14535793	1.0000000	0.1727949
X10	-0.4311758	0.16188322	-0.37109003	0.1727949	1.0000000

A.2.2 Correlation matrix



A.2.3 Scatterplots of all pairs of variable



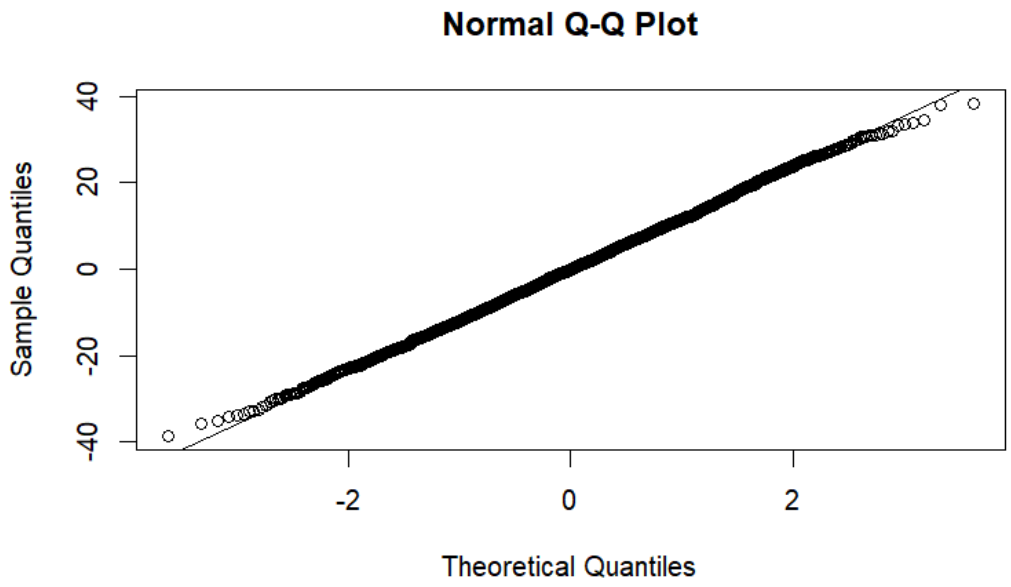
A.2.4 Residual plot against all variables and fitted values

```
        studentized Breusch-Pagan test

data:  initial_mod
BP = 8.3803, df = 4, p-value = 0.0786
```

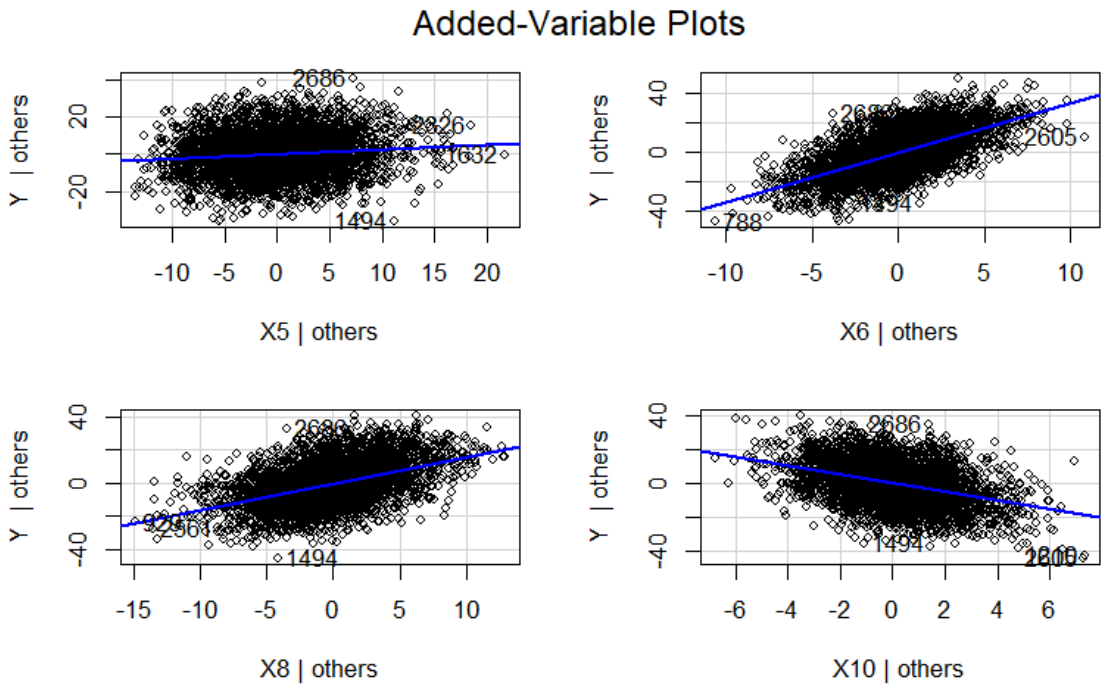
A.2.5 BP test on initial model

```
        Shapiro-Wilk normality test

data:  resid(initial_mod)
W = 0.99948, p-value = 0.4646
```

A.2.6 Shapiro-Wilk normality test



A.2.7 QQ plot for normality



A.2.8 Added-Variable Plots

```
[1] "Outliers based on studentized deleted residuals:"
named numeric(0)
[1] "Outliers based on leverage values:"
          1           2           6          10          12          16          19
33          34          48          50         111
0.004063283 0.003401161 0.004636025 0.003680670 0.003496191 0.002878477 0.003390723
0.002979180 0.003265000 0.003198386 0.003046280 0.003746991 0.003445881
        116         147         156         186         195         202         228
262         283         376         387         414         423
0.002967107 0.003135828 0.003310615 0.003631340 0.003698995 0.004317864 0.002849396
0.003205804 0.003377788 0.003661329 0.002976517 0.003142807 0.004127675
        447         452         470         493         522         551         607
647         683         693         696         785         786
0.003648459 0.003352751 0.002898908 0.003777334 0.003868042 0.003526562 0.002961951
0.004439753 0.003774376 0.004067713 0.003017854 0.003036835 0.003207560
        787         788         792         797         798         819         832
833         849         851         864         887         889
0.003555070 0.005396452 0.002936495 0.002850189 0.003167524 0.002990415 0.004410493
0.003111181 0.004539880 0.003164851 0.003469866 0.004001962 0.003685376
        901         920         929         954         986        1009        1159
1191        1203        1210        1212        1254        1260
0.005327599 0.003152017 0.004586621 0.003689841 0.003215011 0.003276100 0.004098975
0.003070650 0.003195850 0.004812502 0.003061085 0.003888421 0.003268309
```

A.2.9 Outliers detected by studentized deleted residuals and hat leverage values. For more details, refer to the Rmd file provided.

```
[1] "Threshold:"
[1] 0.06740084
         48          50          76         110         164         172         189
195         243         245         259         283         291
 0.10620450  0.07371442 -0.07143894 -0.07606051 -0.10162701 -0.08118552  0.06953181
0.11902549 -0.07023755 -0.07759126 -0.10747324 -0.10340499 -0.10404166
        344         350         355         362         365         368         375
423         455         493         497         505         509
-0.07749265 -0.08880308 -0.09354653 -0.07717881 -0.06866004 -0.07137635 -0.06839188
0.12110492 -0.09667030 -0.07194894 -0.06790519  0.08442525 -0.06883279
        529         556         631         644         664         670         676
686         710         711         741         788         804
-0.07313864 -0.07247146 -0.07777350  0.07380142 -0.08043737  0.06985964  0.07904704
-0.06908022 -0.10100225 -0.06756827 -0.07160991 -0.07314249  0.09133308
        815         864         868         901         917         920         923
957         988        1048        1070        1159        1171
-0.08137292  0.11276436  0.07060874 -0.07930120 -0.07180797  0.07055883 -0.06925566
-0.07857450 -0.08298816 -0.09853063 -0.06901655 -0.08896734 -0.07642100
       1197        1210        1254        1266        1287        1291        1313
1343        1353        1369        1403        1410        1412
-0.07711673 -0.14087030 -0.12044391 -0.07158226  0.07800111 -0.06990886 -0.07216974
0.07461703 -0.07949718 -0.08118178 -0.07619436  0.07042099  0.07361421
```

A.2.10 Outliers detected by DFFITS. For more details and full output, refer to the Rmd file provided.

```
Threshold for 20th percentile of F-distribution: 0.4121736
Threshold for 50th percentile of F-distribution: 0.8393353
Observations exceeding 20th percentile threshold:
Observations exceeding 50th percentile threshold:
```

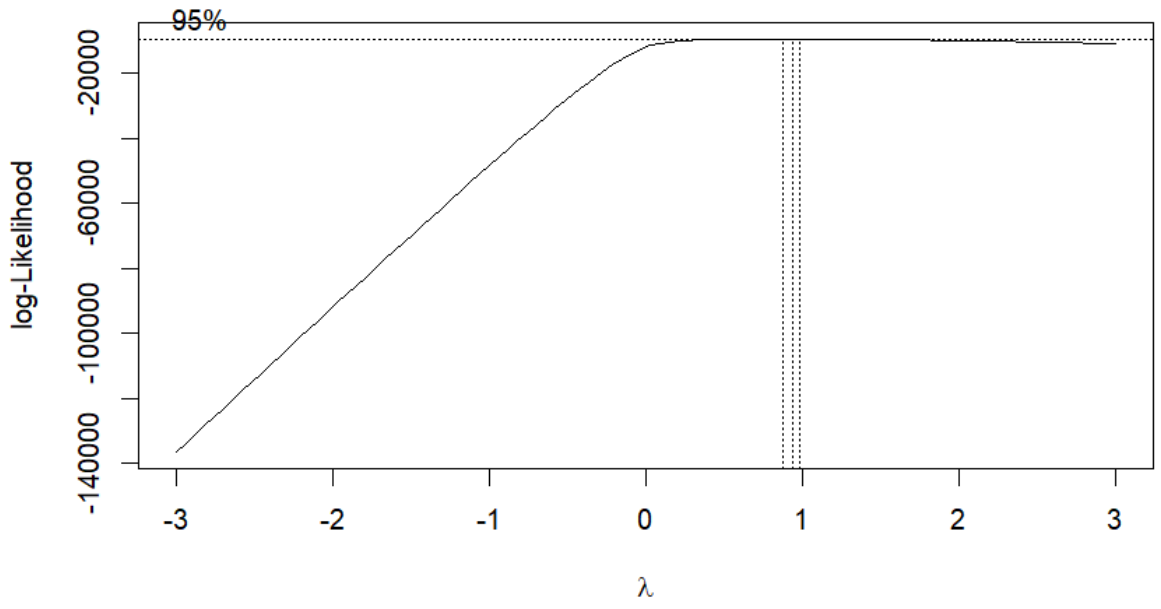
A.2.11 Outliers detected by Cook's Distance. For more details, refer to the Rmd file provided.

```
Threshold for dfbeta: 0.03370042
Points with at least one influence exceeding the threshold: 14 19 41 48 50 55 56 63 76 90 96 110 125 127 129 154 155 160 163 164 172 185 189 195 197 212 214 223 227 243
244 245 259 263 264 267 276 283 291 336 344 348 350 353 355 359 361 362 365 368 375 387 403 409 415 423 455 456 465 475 479 493 497 498 505 509 513 518 521 522 529 539
550 552 556 559 569 573 590 600 607 611 612 613 631 633 640 644 659 664 670 672 674 676 686 687 690 691 700 710 711 718 728 741 748 758 780 786 788 795 799 800 804 813
815 820 832 833 842 843 850 851 860 864 868 875 884 894 895 897 901 911 917 920 923 940 947 954 957 965 985 986 988 989 992 993 997 998 1001 1003 1017 1022 1042 1045
1048 1058 1060 1066 1070 1072 1082 1095 1106 1118 1131 1159 1168 1171 1178 1179 1197 1204 1210 1221 1230 1237 1242 1248 1254 1255 1262 1263 1266 1269 1271 1275 1287 1291
1303 1306 1308 1310 1313 1335 1341 1343 1347 1353 1369 1378 1384 1389 1391 1394 1403 1408 1412 1413 1430 1458 1460 1474 1477 1481 1488 1489 1491 1494 1502 1505 1507 1511
1512 1518 1520 1523 1532 1536 1539 1547 1560 1562 1564 1567 1569 1570 1572 1577 1579 1582 1584 1587 1588 1590 1591 1594 1597 1601 1604 1606 1618 1625 1628 1632 1636 1638
1639 1644 1649 1651 1657 1663 1665 1669 1677 1679 1682 1690 1694 1695 1703 1704 1705 1710 1714 1717 1721 1727 1737 1742 1747 1756 1768 1784 1788 1791 1807 1813 1825 1829
1832 1833 1841 1851 1867 1871 1895 1907 1911 1914 1917 1919 1920 1923 1926 1928 1929 1931 1935 1938 1939 1948 1957 1990 1996 2018 2029 2032 2036 2037 2041 2052 2056 2076
2083 2084 2085 2093 2094 2095 2097 2104 2116 2119 2120 2121 2128 2133 2136 2151 2162 2163 2172 2178 2179 2191 2196 2198 2206 2223 2224 2228 2230 2237 2250 2251 2260 2281
2282 2292 2297 2301 2304 2305 2309 2315 2317 2319 2326 2327 2330 2331 2337 2339 2341 2342 2350 2354 2356 2358 2360 2364 2369 2370 2388 2392 2395 2396 2398 2402 2409 2410
2417 2419 2426 2443 2445 2447 2449 2453 2462 2463 2471 2477 2482 2487 2488 2495 2497 2500 2501 2505 2508 2509 2512 2513 2514 2516 2522 2523 2524 2526 2527 2528 2529 2530
2531 2533 2538 2539 2541 2559 2564 2573 2602 2605 2608 2618 2621 2624 2625 2642 2644 2645 2646 2648 2651 2653 2664 2672 2683 2686 2689 2690 2695 2696 2698 2702 2708 2710
2714 2715 2716 2722 2723 2728 2736 2751 2754 2755 2759 2763 2772 2773 2775 2777 2780 2782 2783 2788 2789 2792 2793 2795 2796 2799 2800 2801 2809 2813 2816 2820 2825 2828
2836 2840 2845 2850 2853 2857 2859 2864 2868 2879 2891 2892 2893 2920 2930 2937 2938 2940 2944 2946 2960 2963 2967 2969 2971 2972 2973 2980 2985 2990 2993 2995 3000 3001
3005 3008 3009 3013 3019 3024 3051 3055 3086 3088 3089 3105 3111 3116 3117 3122 3123 3124 3127 3131 3146 3149 3152 3155 3158 3163 3167 3174 3184 3189 3192 3205 3207 3208
3209 3213 3215 3220 3224 3231 3234 3238 3242 3243 3245 3248 3252 3256 3270 3272 3278 3281 3285 3289 3312 3336 3340 3354 3355 3357 3358 3377 3400 3427 3428 3430 3438 3443
3463 3466 3476 3496 3500 3506 3508 3511 3516 3522
```

A.2.12 Outliers detected DFBETAS. For more details, refer to the Rmd file provided.

```
      X5       X6       X8      X10
1.140832 1.169479 1.154169 1.194574
```

A.2.13 Variation Inflation Factors of each variable



A.2.14 Box-Cox transformation graph

```
[1] "Lambda value that gives the greatest log likelihood:"
[1] 0.9393939

Call:
lm(formula = Y ~ X5 + X6 + X8 + X10, data = transformed_cbb)

Residuals:
    Min      1Q  Median      3Q     Max
-38.756  -8.117   0.021   7.915  38.504

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -122.36318    4.76693 -25.669  < 2e-16 ***
X5             0.25025    0.03821   6.549 6.63e-11 ***
X6             3.33722    0.06930  48.158  < 2e-16 ***
X8             1.60022    0.05029  31.822  < 2e-16 ***
X10           -2.56732    0.10248 -25.051  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.71 on 3517 degrees of freedom
Multiple R-squared:  0.5844,    Adjusted R-squared:  0.584
F-statistic:  1237 on 4 and 3517 DF,  p-value: < 2.2e-16
```

A.2.15 Best lambda found and its associated transformed regression model summary

```
        studentized Breusch-Pagan test

data:  trans_mod
BP = 14.247, df = 4, p-value = 0.006547
```

A.2.16 BP test on the transformed model

	p	1	2	3	4	SSEp	r2	r2.adj	Cp	AICp	SBCp	PRESSp
1	2	0	1	0	0	717582.9	0.38124997	0.38107419	1718.52517	18729.98	18742.31	718335.9
1	2	0	0	0	1	944121.5	0.18591257	0.18568130	3371.67933	19696.30	19708.63	945187.5
1	2	0	0	1	0	1079305.6	0.06934741	0.06908302	4358.17852	20167.61	20179.94	1080501.8
2	3	0	1	1	0	569874.8	0.50861419	0.50833491	642.63278	17920.26	17938.76	570773.6
2	3	0	1	0	1	662679.8	0.42859126	0.42826651	1319.87245	18451.64	18470.14	663778.1
2	3	1	1	0	0	689385.9	0.40556338	0.40522553	1514.75908	18590.79	18609.29	690489.5
3	4	0	1	1	1	487826.6	0.57936188	0.57900318	45.89007	17374.74	17399.41	488885.4
3	4	1	1	1	0	567944.5	0.51027863	0.50986101	630.54651	17910.31	17934.97	569162.6
3	4	1	1	0	1	620719.1	0.46477272	0.46431630	1015.66619	18223.25	18247.92	622078.9
4	5	1	1	1	1	481949.2	0.58442979	0.58395715	5.00000	17334.05	17364.88	483265.5

A.2.17 Result of best subset algorithm, for all variables, choosing the top 3 best models for each number of p.

	nvmax <dbl>	RMSE <dbl>	Rsquared <dbl>	MAE <dbl>	RMSESD <dbl>	RsquaredSD <dbl>	MAESD <dbl>
1	4	11.71798	0.5836182	9.37313	0.294409	0.03497883	0.3456121

A.2.18 Result from 10-fold cross validation

```
Call:
lm(formula = Y ~ X5 + X6 + X8 + X10, data = cbb, weights = wts1)

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-4.0297 -0.8640  0.0007  0.8449  4.1344

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -122.48115    4.75826 -25.741  < 2e-16 ***
X5             0.24385    0.03827   6.371 2.12e-10 ***
X6             3.33948    0.06907  48.348  < 2e-16 ***
X8             1.60760    0.05008  32.099  < 2e-16 ***
X10           -2.56682    0.10279 -24.972  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.251 on 3517 degrees of freedom
Multiple R-squared:  0.5854,    Adjusted R-squared:  0.585
F-statistic:  1242 on 4 and 3517 DF,  p-value: < 2.2e-16
```

A.2.19 Weight Least Square model summary

```
Call: rlm(formula = Y ~ X5 + X6 + X8 + X10, data = cbb, psi = psi.bisquare)
Residuals:
      Min        1Q    Median        3Q       Max
-38.80169  -8.05479   0.07088   7.95963  38.69835

Coefficients:
            Value     Std. Error t value
(Intercept) -122.2031    4.9134   -24.8716
X5             0.2533    0.0394     6.4322
X6             3.3427    0.0714    46.7996
X8             1.6135    0.0518    31.1308
X10           -2.6190    0.1056   -24.7931

Residual standard error: 11.88 on 3517 degrees of freedom
```

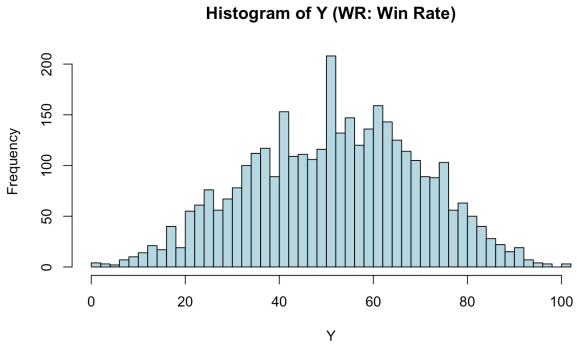
A.2.20 Robust regression model summary

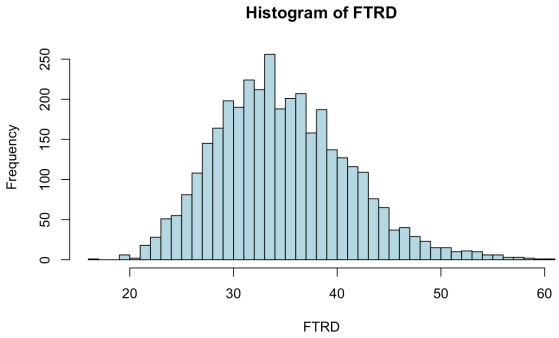
```
Bootstrap Statistics :
        original        bias    std. error
t1* -122.3631776  0.2038761101  4.82783753
t2*    0.2502474 -0.0009249626  0.03814611
t3*    3.3372214 -0.0017898252  0.06735271
t4*    1.6002196  0.0015936228  0.04823721
t5*   -2.5673214 -0.0072561465  0.10853202
```

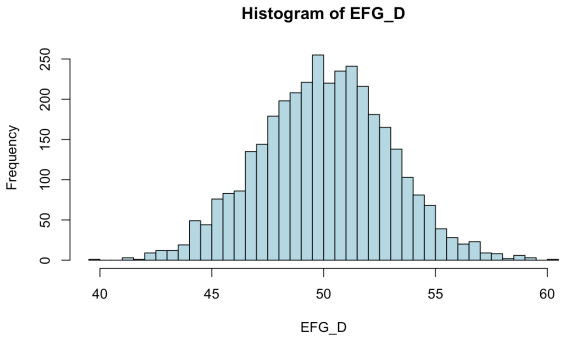
```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot.ci(boot.out = cbb.boot, type = "perc", index = 1)

Intervals :
Level     Percentile
95%   (-131.3, -112.0 )
Calculations and Intervals on Original Scale
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot.ci(boot.out = cbb.boot, type = "perc", index = 2)

Intervals :
Level     Percentile
95%   ( 0.1720,  0.3244 )
Calculations and Intervals on Original Scale
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot.ci(boot.out = cbb.boot, type = "perc", index = 3)

Intervals :
Level     Percentile
95%   ( 3.198,  3.465 )
Calculations and Intervals on Original Scale
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot.ci(boot.out = cbb.boot, type = "perc", index = 4)

Intervals :
Level     Percentile
95%   ( 1.507,  1.698 )
Calculations and Intervals on Original Scale

ORDINARY NONPARAMETRIC BOOTSTRAP
```

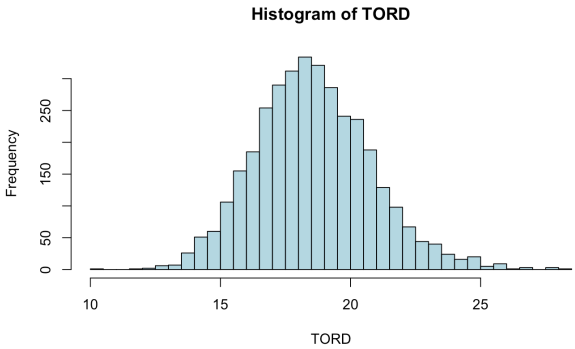
A.2.21 Results from Bootstrapping

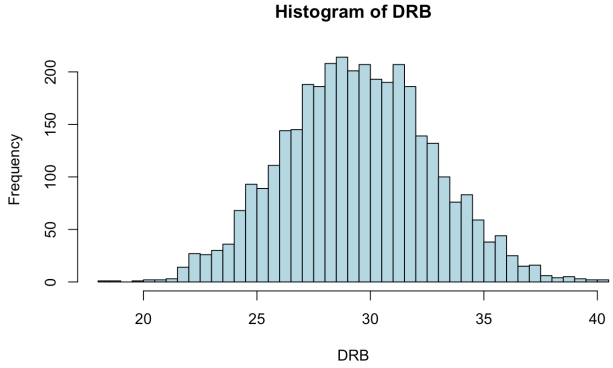
Nikhil Venkatachalam's Appendix

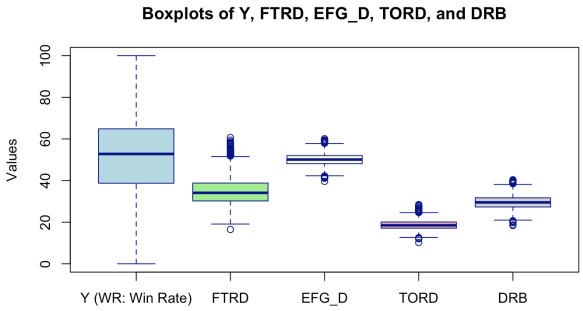








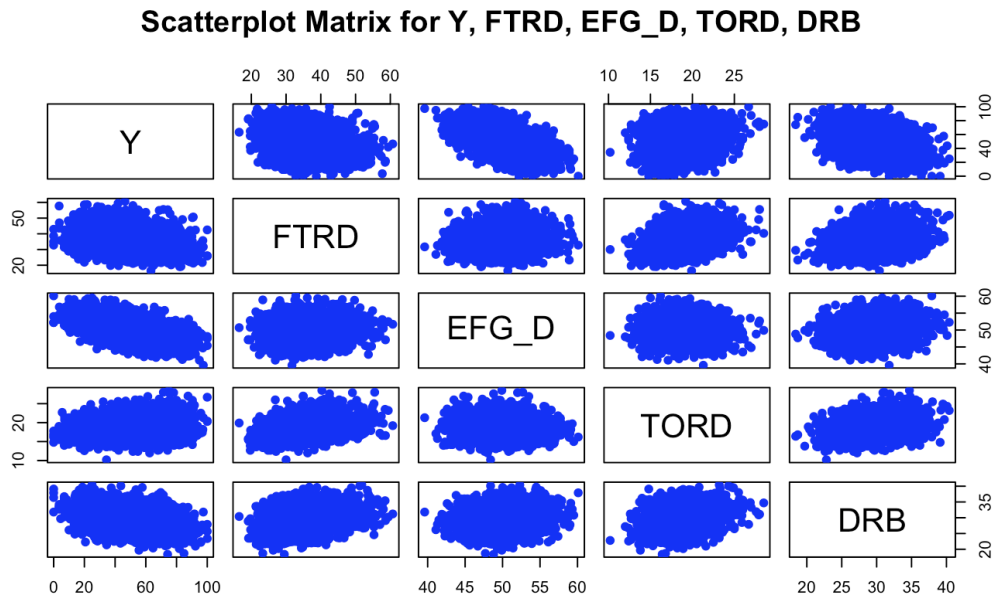




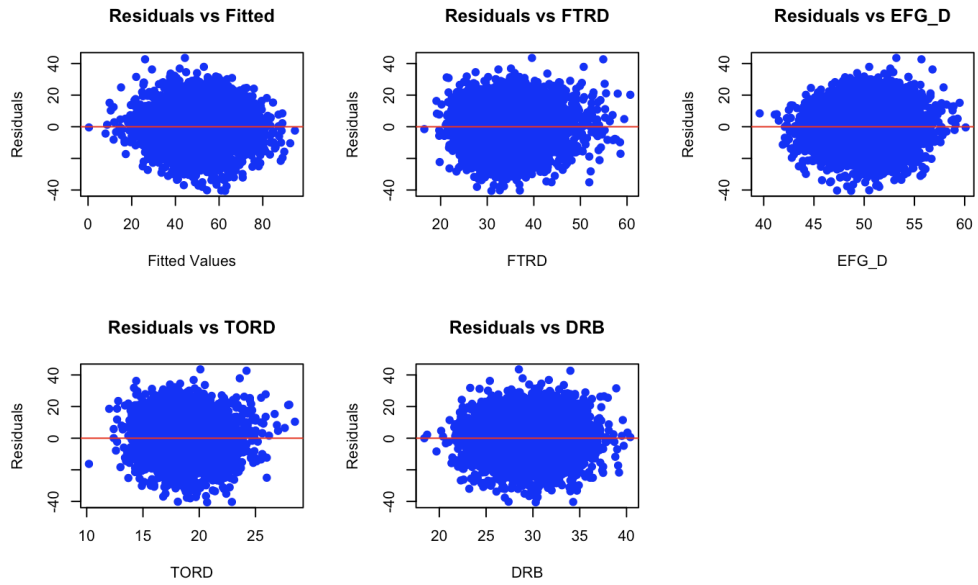
A.3.1 and A.3.2: Histograms and boxplots of variables examined

	Y	FTRD	EFG_D	TORD	DRB
Y	1.0000000	-0.2776990	-0.5753987	0.1612089	-0.3837487
FTRD	-0.2776990	1.0000000	0.0680408	0.3582435	0.3143469
EFG_D	-0.5753987	0.0680408	1.0000000	-0.0587971	0.1458278
TORD	0.1612089	0.3582435	-0.0587971	1.0000000	0.3040078
DRB	-0.3837487	0.3143469	0.1458278	0.3040078	1.0000000

A.3.3: Correlation matrix of variables examined



A.3.4: Scatterplot matrix of variables examined



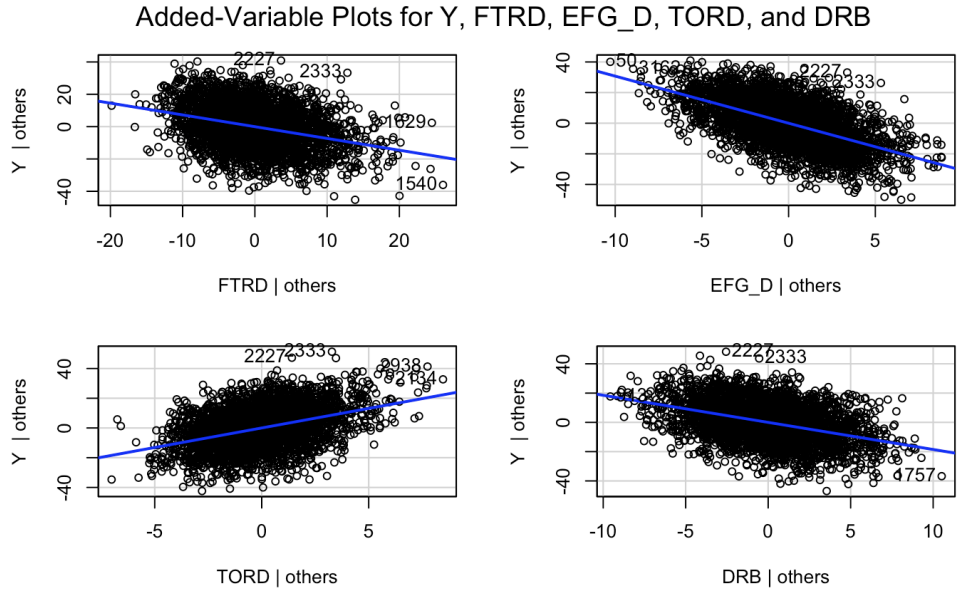
A.3.5: Residual plots against fitted values and all variables examined

```
        studentized Breusch-Pagan test

data:  model
BP = 18.508, df = 4, p-value = 0.0009817
```

A.3.6: Breush-Pagan (BP) test on model

```
        Shapiro-Wilk normality test

data:  residuals
W = 0.99928, p-value = 0.1774
```

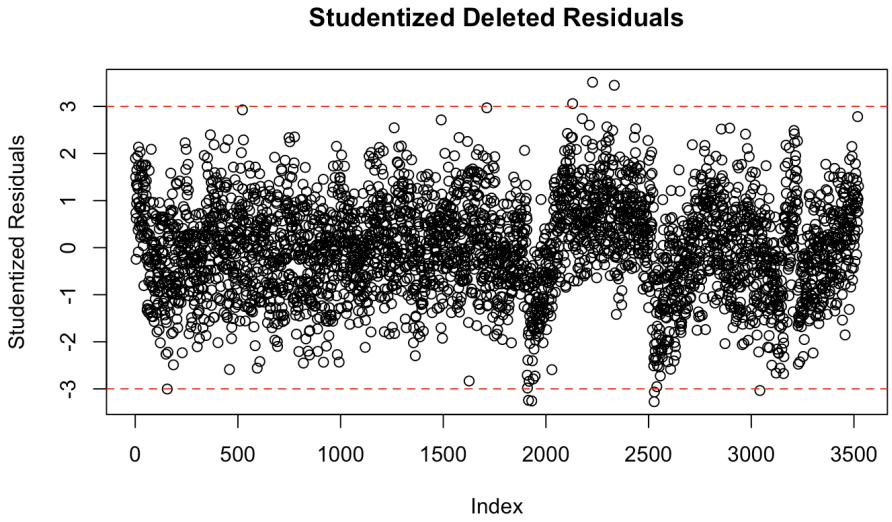
A.3.7: Shapiro-Wilk test on model residuals



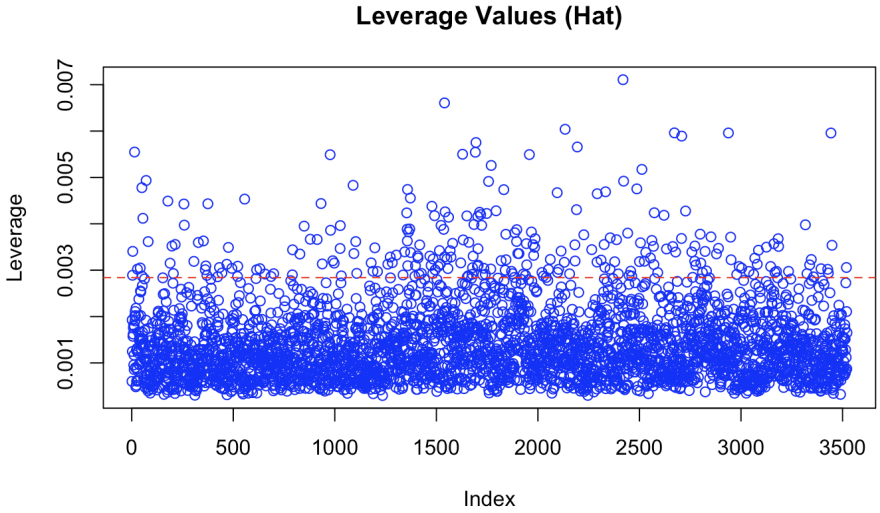
A.3.8: Added-Variable plots for all variables examined

```
[1] "Leverage threshold: 0.00283929585462805"
[1] "DFFITS threshold: 0.0753564311074782"
[1] "Cook's Distance threshold: 0.00113571834185122"
[1] "DFBETAS threshold: 0.0337004204996202"
[1] "Mahalanobis Distance threshold: 11.0704976935164"
```

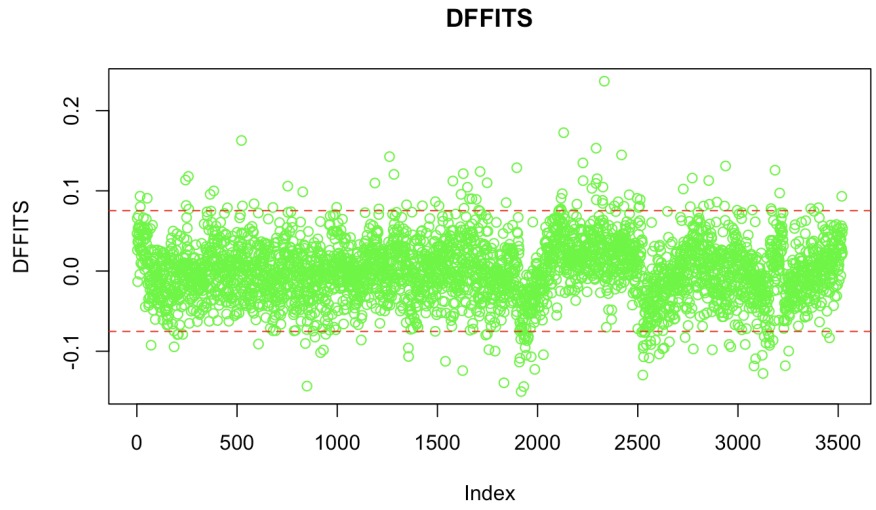
A.3.9: Thresholds calculated for outlier determination methods (see attached Rmd file for precise formulas)



A.3.10: Outlier graph for studentized deleted residuals

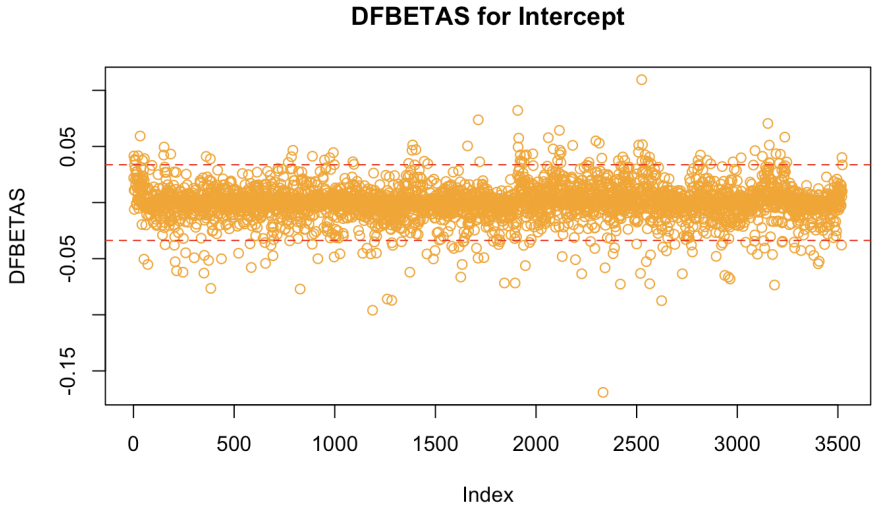


A.3.11: Outlier graph for hat leverage values

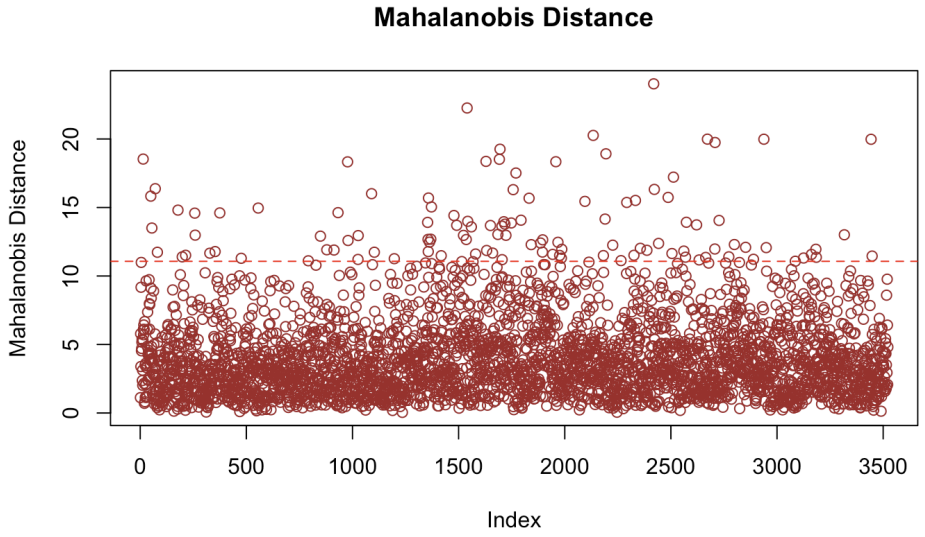


A.3.12: Outlier graph for DFFITS

A.3.13: Outlier graph for Cook's distance



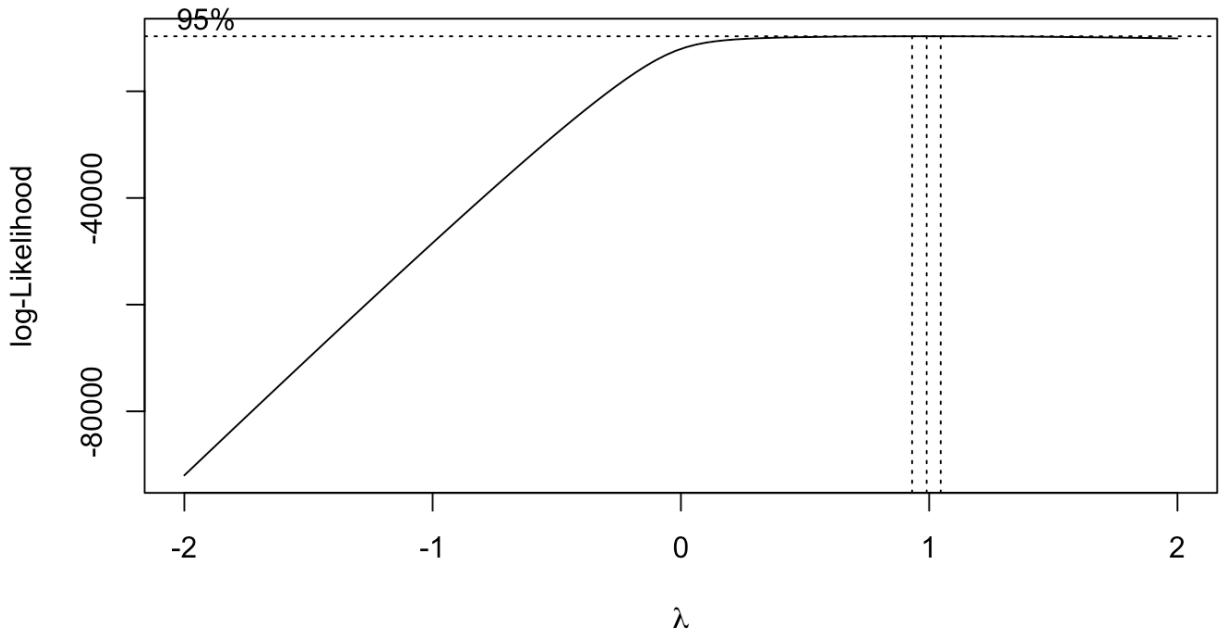
A.3.14: Outlier graph for DFBETAS



A.3.15: Outlier graph for Mahalanobis distance

```
    FTRD     EFG_D      TORD       DRB
1.216016 1.037633 1.221400 1.192553
```

A.3.16: Variation Inflation Factors (VIFs) for each variable examined




```
[1] "Optimal lambda: 1"
```

A.3.17: Box-Cox transformation graph and optimal lambda

```
  p 1 2 3 4      SSEp         r2     r2.adj        Cp     AICp     SBCp    PRESSp
1 2 0 1 0 0  775762.3 0.33108362 0.33089359 1516.6756 19004.55 19016.88  776592.8
1 2 0 0 0 1  988944.5 0.14726306 0.14702081 2900.2221 19859.66 19871.99  990060.6
1 2 1 0 0 0 1070295.3 0.07711672 0.07685454 3428.1865 20138.08 20150.41 1071506.1
1 2 0 0 1 0 1129590.5 0.02598830 0.02571159 3813.0104 20327.99 20340.32 1130864.7
2 3 0 1 0 1  669232.8 0.42294082 0.42261285  827.3019 18486.30 18504.80  670332.0
2 3 1 1 0 0  709460.5 0.38825363 0.38790595 1088.3790 18691.89 18710.39  710672.0
2 3 0 1 1 0  756880.5 0.34736483 0.34699391 1396.1332 18919.76 18938.26  758144.4
2 3 0 0 1 1  890280.1 0.23233834 0.23190205 2261.8932 19491.49 19509.99  891756.7
3 4 0 1 1 1  604838.0 0.47846650 0.47802176  411.3815 18131.97 18156.64  606184.3
3 4 1 1 0 1  642053.6 0.44637658 0.44590447  652.9100 18342.27 18366.94  643527.6
3 4 1 1 1 0  647930.0 0.44130957 0.44083314  691.0474 18374.36 18399.03  649405.1
3 4 1 0 1 1  811008.1 0.30069222 0.30009588 1749.4200 19165.04 19189.70  812805.3
4 5 1 1 1 1  541913.0 0.53272483 0.53219338    5.0000 17747.06 17777.90  543447.0
```

A.3.18: Result of the best subset algorithm for all variables examined

```
Linear Regression

3522 samples
   4 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 3168, 3170, 3170, 3171, 3170, 3170, ...
Resampling results:

  RMSE      Rsquared   MAE
  12.41292  0.5332026  9.955891

Tuning parameter 'intercept' was held constant at a value of TRUE
```

A.3.19: Results from 10-fold cross-validation of the model

```
Call:
lm(formula = Y ~ FTRD + EFG_D + TORD + DRB, data = cbb, weights = weights)

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-6.3735 -2.9104  0.4452  2.9089  6.5967

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 237.54563    1.35396  175.45   <2e-16 ***
FTRD         -0.73113    0.01328  -55.08   <2e-16 ***
EFG_D        -3.08207    0.02320 -132.86   <2e-16 ***
TORD          2.63062    0.02944   89.35   <2e-16 ***
DRB          -1.86471    0.02404  -77.58   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.156 on 3517 degrees of freedom
Multiple R-squared:  0.9264,    Adjusted R-squared:  0.9263
F-statistic: 1.107e+04 on 4 and 3517 DF,  p-value: < 2.2e-16
```

A.3.20: Weighted Least Squares regression model summary

```
        studentized Breusch-Pagan test

data:  wls_model
BP = 2.1351, df = 4, p-value = 0.7109
```

A.3.21: Breusch-Pagan test on the WLS model

```
Call: rlm(formula = Y ~ FTRD + EFG_D + TORD + DRB, data = cbb)
Residuals:
     Min       1Q   Median       3Q      Max
-40.7625  -8.5254   0.1549   8.4392  43.5552

Coefficients:
            Value    Std. Error t value
(Intercept) 238.5494   4.3408    54.9551
FTRD         -0.7438   0.0373   -19.9560
EFG_D        -3.1018   0.0761   -40.7701
TORD          2.6650   0.1066    24.9922
DRB          -1.8708   0.0730   -25.6198

Residual standard error: 12.56 on 3517 degrees of freedom
```

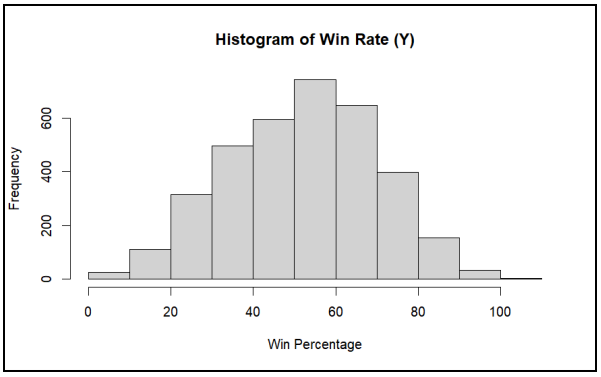
A.3.22: Robust regression model summary

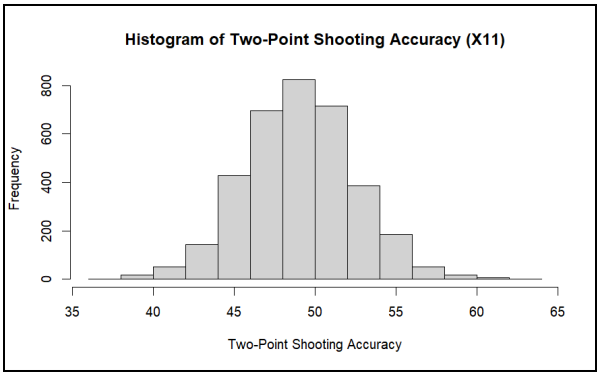
Parth Gandhi's Appendix

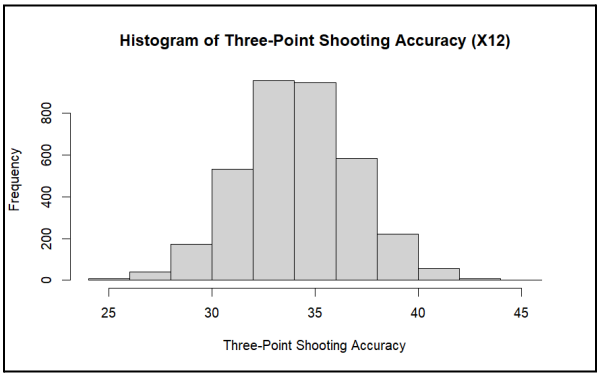
```
#Map all the variables
y <- bbdata$WinPercent
x11 <- bbdata$X2P_O
x12 <- bbdata$X3P_O
x8 <- bbdata$ORB
x9 <- bbdata$TORD
x5 <- bbdata$FTR
```

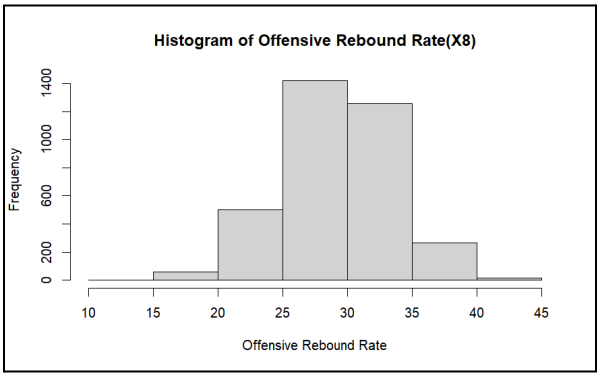
```
#Make a new dataframe with all the variables
clgdata <- data.frame(y, x11, x12, x8, x9, x5)
```

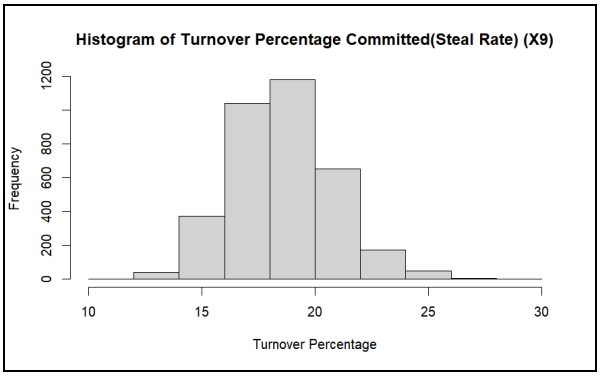
A.4.1 - Mapping of variables to ensure uniformity with conventions used in the report and creation of new dataframe

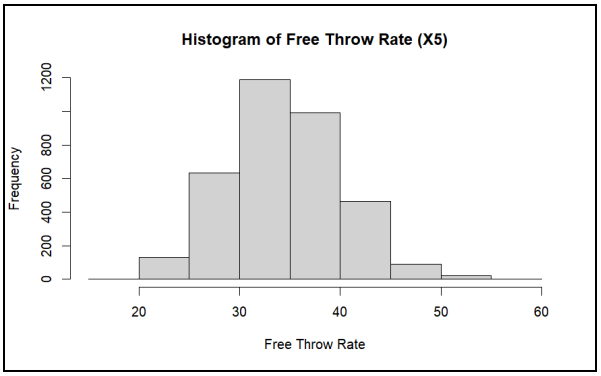


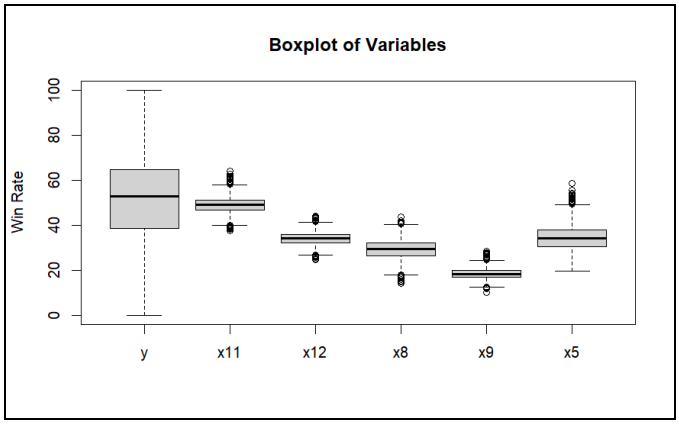








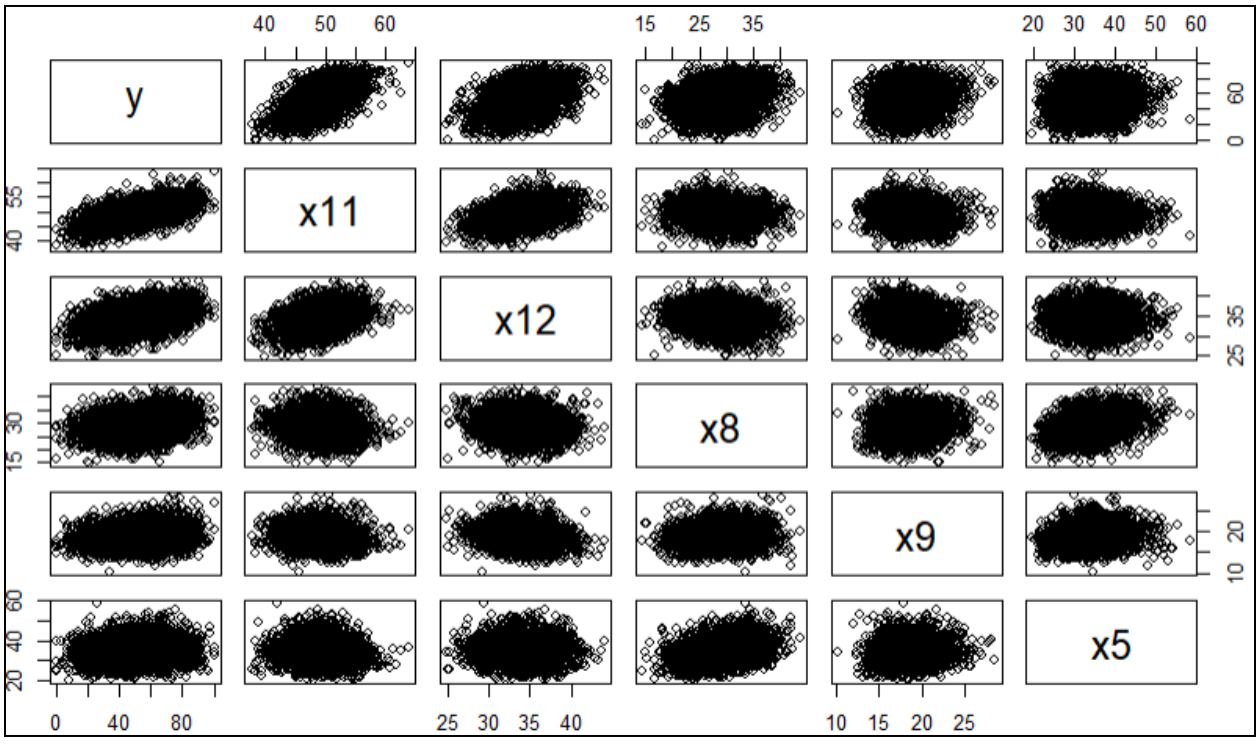




A.4.2 - Histogram and Boxplot of dependent and independent variables

	y	x11	x12	x8	x9	x5
y	1.0000000	0.58772174	0.42021536	0.2633390	0.16120886	0.11432167
x11	0.5877217	1.00000000	0.38231605	-0.1083672	-0.09964824	-0.07142672
x12	0.4202154	0.38231605	1.00000000	-0.1182776	-0.17535905	-0.01458056
x8	0.2633390	-0.10836717	-0.11827765	1.0000000	0.16173902	0.33445432
x9	0.1612089	-0.09964824	-0.17535905	0.1617390	1.00000000	0.11224962
x5	0.1143217	-0.07142672	-0.01458056	0.3344543	0.11224962	1.00000000

A.4.3 - Correlation matrix



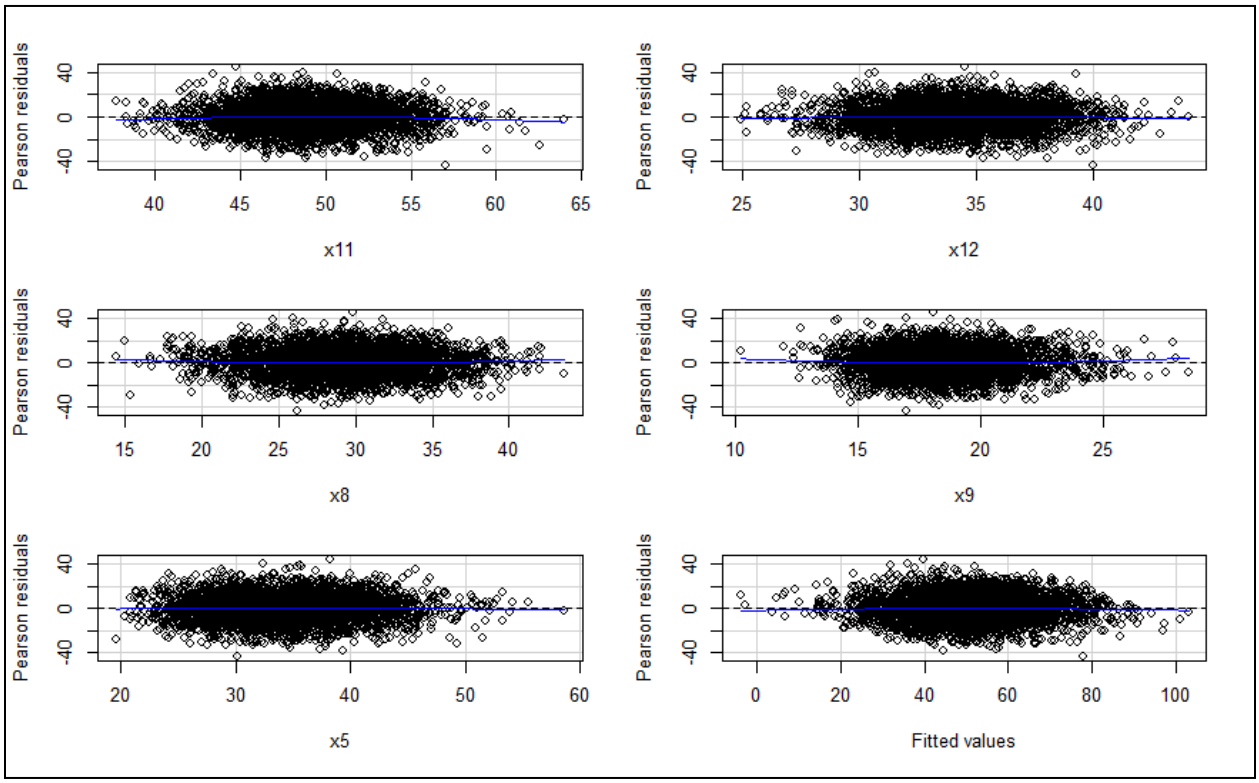
A.4.4 - Scatterplots of all pairs of variable

```
Anova Table (Type II tests)

Response: y
           Sum Sq   Df  F value  Pr(>F)
x11        279643    1 1905.988 < 2e-16 ***
x12         81243    1  553.736 < 2e-16 ***
x8          97146    1  662.131 < 2e-16 ***
x9          49082    1  334.537 < 2e-16 ***
x5            839    1    5.718 0.01684 *
Residuals 515860 3516
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

A.4.5 - Type II Anova Table for full model

```
     x11      x12      x8       x9       x5
1.177992 1.213564 1.146589 1.069382 1.120616
```

A.4.6 - Variance Inflation Factor for Dataset



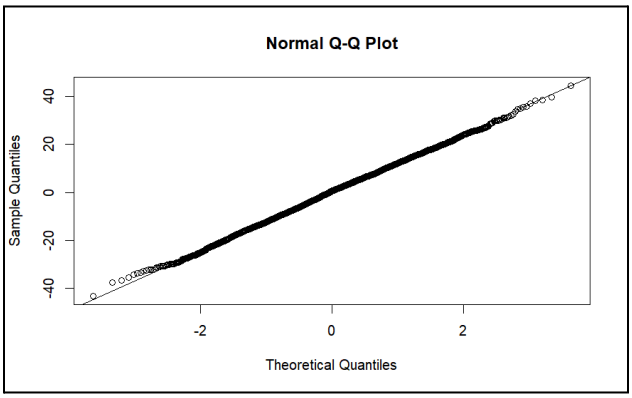
A.4.7 - Residual plot for all variables

```
        studentized Breusch-Pagan test

data:  full_model
BP = 9.0963, df = 5, p-value = 0.1053
```

```
        Shapiro-Wilk normality test

data:  residuals(full_model)
W = 0.99925, p-value = 0.1559
```

A.4.8 - BP Test for non constant variance

A.4.9 - Shapiro Test for normality



A.4.10 - Normal Q-Q Plot

```
       90       162       163       164       170       172       185       187       195       197       245       283
-2.466006  2.117020 -2.206132 -2.272990  2.144193 -2.466097 -2.233559  2.078937  3.180353  2.034929 -2.171302 -2.047965
      291       345       350       362       365       415       505       615       676       691       710       718
-2.375405 -2.411886 -2.175902 -2.196332 -2.061305 -2.459675  2.341919 -2.218118  2.470572  2.062178 -2.424726 -2.046863
      728       815       820       860       884       895       897       904       917       949       957       986
-2.145175 -2.355553 -2.193000  2.069276 -2.397625  2.222393  2.131648  2.163774 -2.068710 -2.100524 -2.154012 -2.112848
      988      1048      1095      1111      1159      1171      1172      1197      1210      1254      1263      1266
-2.487527 -2.408918 -2.537863 -2.179654 -2.627199 -2.100917 -2.073053 -2.452822 -2.918577 -3.110732 -2.647320 -2.160860
     1273      1275      1303      1313      1394      1412      1460      1488      1489      1494      1509      1511
-2.435846 -2.066503 -2.584230 -2.194335 -2.683589  2.140158 -2.071301 -2.573029 -2.548888 -2.480056 -2.279038 -2.083543
     1518      1523      1557      1570      1584      1587      1590      1597      1628      1669      1682      1704
-2.033037 -2.278472 -2.022224 -2.274658 -2.345369 -2.394212 -2.531232 -2.260028 -2.230810 -2.168967 -2.659166 -2.098896
     1705      1737      1907      1911      1914      1919      1926      1927      1928      1929      1935      1938
-2.265996 -2.211004  2.059104 -2.269674 -2.771310 -2.229344 -2.509536 -2.464881 -2.836615 -3.015985 -2.025429 -2.285929
     1957      2020      2037      2041      2056      2085      2104      2116      2162      2172      2228      2230
-2.037872 -2.280699  2.092826  2.268295  2.229161  2.172377  2.128636  2.212270  2.390081  2.488556  2.014642  2.689261
     2234      2293      2294      2297      2305      2341      2356      2417      2443      2497      2504      2509
 2.157852  2.115198  2.187023  3.151397  2.036244  2.933896  3.061416  2.451412  2.608122  3.674042  2.005974  2.052466
     2523      2525      2526      2529      2530      2541      2542      2561      2564      2567      2605      2618
 2.389706 -2.178476 -2.412815 -2.661003 -2.298349 -2.578886 -2.167805 -2.474807 -2.547944 -2.113379 -3.578017  2.048596
     2644      2657      2686      2689      2695      2715      2716      2723      2759      2775      2792      2793
 2.245287 -2.001068  2.453728  2.566349  2.104196  2.125947  2.089265  2.164401  2.132832  2.126696  2.625957  2.357857
     2797      2799      2813      2816      2817      2828      2857      2864      2868      2960      2963      2967
 2.167299  2.256452  3.273137  2.037986  2.784834  2.083654  2.119566  2.481454  2.361026  2.860956  2.029990  2.429676
     2972      2979      3000      3019      3111      3176      3206      3208      3234      3235      3238      3270
 2.867891  2.567168 -2.133956 -2.709521 -2.100321  2.576049  2.101731  2.550742 -2.764611 -2.085241 -2.477386 -2.154713
     3312      3358      3466      3489      3495      3508      3515      3516      3520
-2.318199  2.519856  2.471362  2.138013  2.947927  2.094234  2.189678  2.018466  2.231056
```

A.4.11 Outliers detected by studentized deleted residuals

```
[1] "DFFITs Threshold: 0.0825488343412996"
        90        162        164        172        185        195        245        283        291        344
-0.1493267  0.0845595 -0.0962611 -0.0835515 -0.0895276  0.1157817 -0.0903293 -0.1083026 -0.1052673 -0.1006938
       423        505        552        569        676        709        710        711        718        780
 0.1198105  0.0950202 -0.0894634  0.0924565  0.0983838 -0.0860584 -0.1128470 -0.0923558 -0.0928149  0.0843106
       787        815        842        850        864        884        895        917        977        986
-0.0838828 -0.1031050 -0.0972486 -0.0841276  0.1160672 -0.0834648  0.1073388 -0.0829615 -0.0842590 -0.1516811
       988       1048       1095       1118       1159       1171       1178       1197       1210       1254
-0.1217870 -0.0971221 -0.0871100  0.0927595 -0.1356209 -0.1129351 -0.0835064 -0.0836594 -0.1090193 -0.0942730
      1259       1263       1266       1268       1303       1313       1321       1343       1372       1394
-0.0900947 -0.1016619 -0.0933063  0.0826537 -0.0907714 -0.1264535 -0.0956662  0.0965362 -0.1040291 -0.0870246
      1474       1488       1489       1491       1494       1518       1523       1570       1587       1590
-0.1150890 -0.1500250 -0.1112319 -0.0901480 -0.1261211 -0.0977085 -0.0931762 -0.0956075 -0.0928949 -0.1336892
      1628       1649       1660       1685       1705       1713       1737       1770       1825       1833
-0.1199109 -0.1007218  0.1023939 -0.0908946 -0.1125744  0.0970207 -0.1295800 -0.1075042 -0.0874336  0.0832665
      1886       1907       1911       1914       1919       1926       1927       1929       1938       1996
-0.0870546  0.1049788 -0.0842896 -0.1039288 -0.0969759 -0.1461083 -0.0846435 -0.0991701 -0.1063382 -0.0867254
      2020       2056       2085       2104       2116       2162       2172       2196       2230       2237
-0.1143861  0.1008623  0.0869949  0.0887198  0.1055283  0.1205590  0.1050419 -0.0990190  0.1139769  0.0861991
      2293       2294       2297       2314       2341       2356       2398       2410       2443       2462
 0.1063523  0.1002438  0.1528088  0.1088450  0.1287062  0.1228932 -0.0832253 -0.1091312  0.1185626  0.0827788
      2497       2508       2509       2512       2522       2525       2526       2529       2530       2531
 0.1170214  0.0829031  0.1075549  0.1261106  0.0838147 -0.1154691 -0.0893093 -0.0961211 -0.1177630  0.0840382
      2541       2543       2554       2561       2564       2567       2605       2618       2648       2686
-0.0976420 -0.0965596  0.0891118 -0.1821481 -0.0922010 -0.0915530 -0.1778487  0.1246017  0.1457563  0.1308818
      2689       2695       2723       2759       2783       2792       2799       2800       2813       2817
 0.1378483  0.0869236  0.1202367  0.0895725  0.0910868  0.0937184  0.1078028  0.0970334  0.1151961  0.0851354
      2828       2830       2849       2857       2864       2938       2960       2967       2973       2979
 0.0874029  0.0874264  0.0831079  0.0885607  0.1212300  0.1123083  0.0935226  0.0916632  0.0861776  0.1274107
      3000       3019       3051       3122       3131       3208       3215       3234       3238       3270
-0.0940882 -0.0905642  0.1067802  0.0960496  0.1007028  0.1229268  0.0889104 -0.1163507 -0.1111674 -0.0887376
      3332       3358       3377       3476       3495       3516
-0.0901958  0.1144015  0.0929853  0.0840176  0.1001254  0.1057192
```

A.4.12 - Influential Point in DFFITs for Dataset

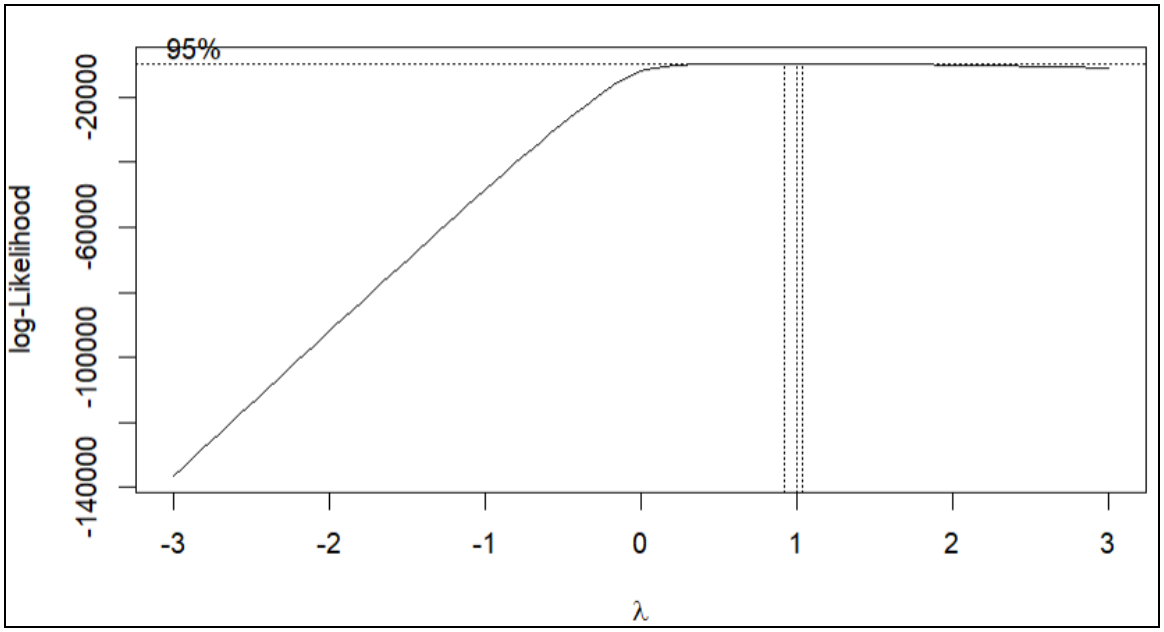
```
  [1] 0.04461646 0.05193015 0.03449363 0.05559339 0.03497359 0.03921733 0.05679746 0.07683326 0.04361743 0.04764849
 [11] 0.03966758 0.05643788 0.04389361 0.06432038 0.06354470 0.04417671 0.07149253 0.03626580 0.05218729 0.06361223
 [21] 0.04050417 0.04525009 0.04855585 0.04863238 0.03690450 0.07128246 0.08910218 0.03410703 0.05371465 0.05546901
 [31] 0.03781812 0.04576485 0.04357737 0.05585284 0.03926817 0.05906471 0.05212112 0.03469536 0.05233421 0.05621902
 [41] 0.04498298 0.03880083 0.03961037 0.04767510 0.06829849 0.04611628 0.04889407 0.03938675 0.06207998 0.06161000
 [51] 0.03581045 0.04521977 0.06829417 0.06489173 0.05414052 0.03785935 0.03572387 0.06896423 0.04786284 0.05835590
 [61] 0.04060674 0.05782420 0.08440433 0.03887459 0.03909630 0.04906806 0.04285193 0.04525225 0.06934480 0.07826073
 [71] 0.04277695 0.09761251 0.03916369 0.06396972 0.03909662 0.03406488 0.04100729 0.06274650 0.03412463 0.03626052
 [81] 0.05009081 0.03658539 0.07617806 0.03864565 0.03462688 0.03816099 0.03695553 0.04709264 0.05974389 0.03755811
 [91] 0.05840275 0.03834415 0.04537841 0.07694744 0.07910035 0.03819738 0.03717255 0.05349926 0.04072555 0.03854292
[101] 0.09192849 0.04382324 0.03948375 0.03424555 0.05346030 0.04664332 0.05708759 0.03872987 0.03692141 0.03934486
[111] 0.03410661 0.03746029 0.03566089 0.03718907 0.04273572 0.05086934 0.03596301 0.03382836 0.04222486 0.03475704
[121] 0.03959578 0.04139832 0.04759631 0.03968828 0.04495867 0.03886344 0.05480725 0.04698590 0.05320835 0.05683244
[131] 0.03873136 0.05099566 0.03389435 0.03599480 0.03695999 0.03583167 0.06444251 0.03928729 0.04081645 0.04982605
[141] 0.03465692 0.04869182 0.08239239 0.07916397 0.04524443 0.04047242 0.03597947 0.04066363 0.04043298 0.04508866
[151] 0.03734428 0.03695199 0.05282317 0.04207471 0.04038891 0.04324027 0.04301391 0.04258874 0.04114836 0.03946389
[161] 0.04462456 0.03646426 0.04337250 0.03449874 0.04143382 0.04929804 0.03737387 0.04023114 0.03524676 0.04364143
[171] 0.04406470 0.04060438 0.03529828 0.04348027 0.04393776 0.05545949 0.05972751 0.04351128 0.06145401 0.03774213
[181] 0.03696216 0.03580082 0.04288203 0.05161551 0.04264904 0.05112166 0.04664229 0.03841605 0.05373827 0.04828515
[191] 0.04340006 0.03740849 0.05200616 0.04782954 0.03407702 0.04686347 0.03854091 0.04173815 0.03729257 0.07140954
[201] 0.04392351 0.03463579 0.04322469 0.04010776 0.04175919 0.05950896 0.04304746 0.03974004 0.04441799 0.05222816
[211] 0.05248139 0.03917463 0.06041632 0.06137155 0.03524160 0.04967932 0.04218697 0.03849903 0.03481315 0.04116479
[221] 0.03373906 0.04985820 0.03695418 0.04308282 0.03405261 0.04854962 0.05051740 0.04026443 0.03524296 0.03719251
[231] 0.05326528 0.03623873 0.05756133 0.04484075 0.04228555 0.03609699 0.03803109 0.03750885 0.03523683 0.03381919
[241] 0.03466581 0.03573584 0.07109191 0.05510054 0.04220855 0.04044945 0.03905638 0.03570630 0.05277063 0.03714441
[251] 0.03835681 0.04066387 0.04266886 0.03748256 0.03829820 0.04320301 0.05799012 0.03785387 0.04460294 0.07683806
[261] 0.03813023 0.04526099 0.05073009 0.09217366 0.04748014 0.03373084 0.05317358 0.04896352 0.04323184 0.03547317
[271] 0.03741903 0.04025694 0.06371509 0.03813134 0.03825311 0.07034573 0.03912805 0.04450040 0.04420004 0.04097490
[281] 0.04403673 0.04937137 0.05827134 0.06744174 0.07881755 0.03783725 0.03807895 0.03616883 0.03865114 0.05253694
[291] 0.04021602 0.05705806 0.06381669 0.04570556 0.04916087 0.06681597 0.03441913 0.05927787 0.03710734 0.05198120
[301] 0.03989615 0.03801660 0.03898574 0.05308996 0.07129009 0.03444956 0.05043140 0.04859111 0.05191196 0.03834558
[311] 0.06016821 0.03375280 0.04078359 0.03941202 0.09097787 0.03967621 0.05257235 0.03425303 0.06867142 0.03955695
[321] 0.05211261 0.03492408 0.05410371 0.03735510 0.05362482 0.04512349 0.06356108 0.03570361 0.03905147 0.03371265
[331] 0.04667040 0.06133669 0.03473278 0.04459325 0.04573358 0.03500340 0.03943200 0.03501745 0.03560702 0.03533102
[341] 0.05862088 0.03461849 0.06206305 0.03530714 0.03717036 0.03444599 0.03676207 0.03512808 0.03997098 0.05056927
[351] 0.05273648 0.04930186 0.04074670 0.07487337 0.06965355 0.08277847 0.06146766 0.04964891 0.03765772 0.03894027
[361] 0.03600801 0.04800843 0.05277622 0.04229634 0.04007405 0.04031419 0.03496801 0.04402453 0.03986291 0.03904750
[371] 0.04741341 0.05084039 0.03488429 0.06047313 0.04132147 0.04768046 0.04566168 0.03536467 0.04135389 0.04623168
[381] 0.03403544 0.05155018 0.04503090 0.03747126 0.03774617 0.05036643 0.04044812 0.06658880 0.03959857 0.03406339
[391] 0.04116708 0.06142528 0.08297300 0.04396057 0.14568477 0.06314230 0.03502086 0.05761661 0.04621760 0.03884911
[401] 0.03504344 0.04949021 0.03458429 0.05918267 0.03796959 0.04070191 0.03956362 0.04534838 0.03903761 0.03825925
[411] 0.03466600 0.04841719 0.05934803 0.03841248 0.03619672 0.03707418 0.03974447 0.04764713 0.03842551 0.05534580
[421] 0.04217118 0.05862150 0.04454245 0.05470619 0.03723345 0.04842297 0.05847675 0.04880707 0.05130349 0.06177437
[431] 0.03656370 0.08622594 0.04612307 0.03981370 0.05475798 0.04323776 0.03644009 0.04373110 0.07912064 0.04539925
```

```
[431] 0.03656370 0.08622594 0.04612307 0.03981370 0.05475798 0.04323776 0.03644009 0.04373110 0.07912064 0.04539925
[441] 0.04703024 0.03797401 0.03378721 0.03713016 0.03845921 0.05209030 0.05115840 0.04410109 0.04965619 0.06365445
[451] 0.06938469 0.05752539 0.03604728 0.04772409 0.03795673 0.07639209 0.03803927 0.04298999 0.03583640 0.05320558
[461] 0.04480794 0.04222702 0.08498801 0.03422070 0.06490356 0.03822598 0.03666600 0.05140226 0.05719490 0.04339335
[471] 0.05974450 0.05043269 0.03814349 0.03764803 0.03991456 0.08647669 0.04988558 0.05209406 0.03682604 0.03667656
[481] 0.03676759 0.11542224 0.03546505 0.04937392 0.03884558 0.06897615 0.08602820 0.03541267 0.05379996 0.05798405
[491] 0.05024154 0.04392917 0.03578465 0.04808309 0.08294370 0.07982172 0.03681207 0.04261472 0.05887315 0.04161838
[501] 0.10386364 0.03713565 0.04598067 0.03412141 0.05047927 0.03746260 0.03707797 0.04326784 0.03396476 0.04515723
[511] 0.05373607 0.04498444 0.04833161 0.06890293 0.04933256 0.07632548 0.05102372 0.05194148 0.04017403 0.03747156
[521] 0.03472463 0.05045221 0.03976372 0.03598656 0.07589174 0.03678380 0.05505923 0.03517729 0.03596972 0.06022907
[531] 0.06509604 0.04353819 0.03634685 0.04638988 0.03867677 0.06242913 0.04209527 0.03673979 0.03937216 0.03837006
[541] 0.04334588 0.03839927 0.04602295 0.04218315 0.03610996 0.04766131 0.03618769 0.05199177 0.03767674 0.05106280
[551] 0.04304842 0.03881056 0.03663828 0.04239385 0.06903958 0.03478279 0.03579443 0.04380437 0.03494619 0.05261105
[561] 0.06743798 0.03833474 0.05659952 0.07194772 0.04303933 0.07348217 0.03885068 0.03948050 0.04712148 0.04026317
[571] 0.04327999 0.03888068 0.03514673 0.05070844 0.03801966 0.07349416 0.04555672 0.04403244 0.05286948 0.10219990
[581] 0.03860120 0.03390187 0.05984767 0.03768909 0.04372642 0.09469588 0.04791152 0.05859480 0.07428720 0.03431217
[591] 0.04986306 0.04827649 0.04290447 0.03688051 0.03419303 0.04122629 0.05399230 0.07354482 0.03632630 0.06087744
[601] 0.03463241 0.03520350 0.04518580 0.05333455 0.03841568 0.04298413 0.03755868 0.03583356 0.04497571 0.04789773
[611] 0.04564602 0.03589700 0.04418131 0.06164742 0.04492508 0.06809097 0.04002414 0.03578997 0.04082507 0.03763177
```

A.4.13 -Influential points in DFBETAS for dataset

```
[1] "Cook's Distance Threshold for major influence points: 0.891524342607 37"
named numeric(0)
```

A.4.14 - Influential point in Cook's Distance for Dataset



A.4.15 Box-cox procedure on dataset

```
[1] "Lambda with greatest log likelihood: 1"
```

A.4.16 - Lambda with the greatest log likelihood

	p	1	2	3	4	5	SSEp	r2	r2.adj	Cp	AICp	SBCp	PRESSp
1	2	1	0	0	0	0	759139.6	0.3454168	0.3452309	1656.146041	18928.26	18940.59	759955.4
2	3	1	0	1	0	0	633635.3	0.4536355	0.4533250	802.732960	18293.79	18312.29	634652.5
3	4	1	1	1	0	0	566859.3	0.5112144	0.5107976	349.600942	17903.57	17928.24	568075.2
4	5	1	1	1	1	0	516698.9	0.5544661	0.5539594	9.718023	17579.26	17610.09	518109.2
5	6	1	1	1	1	1	515860.0	0.5551895	0.5545570	6.000000	17575.53	17612.53	517555.9

A.4.17 - Results from Best Subset Algorithm for Model Selection

```
Call: rlm(formula = y ~ x11 + x12 + x8 + x9 + x5, data = clgdata, psi = psi.bisquare)
Residuals:
     Min       1Q   Median       3Q      Max
-43.7089  -8.4578   0.4985   8.0465  44.5089

Coefficients:
            Value     Std. Error t value
(Intercept) -231.9298    4.5271   -51.2311
x11            2.8931    0.0672    43.0410
x12            1.9514    0.0846    23.0596
x8             1.3476    0.0537    25.0734
x9             1.7161    0.0966    17.7684
x5             0.0986    0.0406     2.4267

Residual standard error: 12.28 on 3516 degrees of freedom
```

A.4.18 - Robust Regression Summary for our dataset

```
Linear Regression with Backwards Selection

3522 samples
   5 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 3168, 3170, 3170, 3171, 3170, 3170, ...
Resampling results:

  RMSE      Rsquared   MAE
  12.11805  0.5539729  9.746416

Tuning parameter 'nvmax' was held constant at a value of 5
```

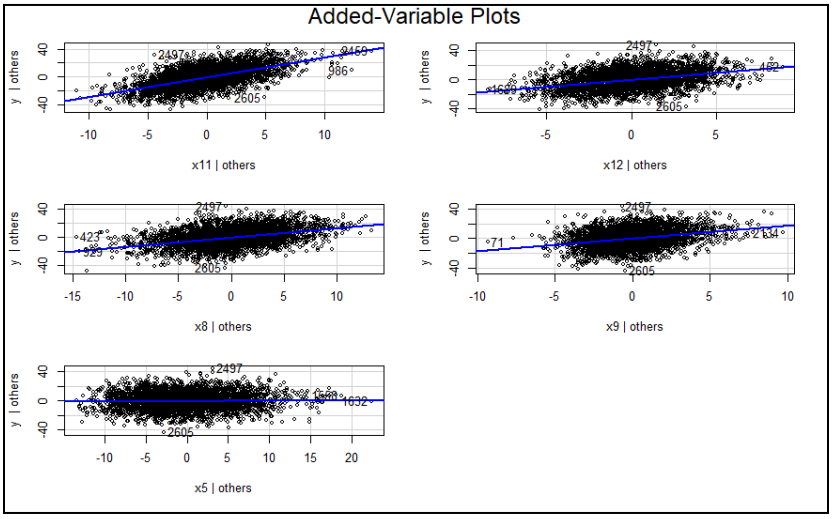
A.4.19 - Results from 10-Fold Cross Validation

```
Call:
lm(formula = y ~ x11 + x12, data = clgdata)

Residuals:
    Min      1Q  Median      3Q     Max
-46.793  -9.747   0.376   9.805  44.322

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -131.51439    3.90436  -33.68   <2e-16 ***
x11            2.66996    0.07605   35.11   <2e-16 ***
x12            1.52255    0.09472   16.07   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’

Residual standard error: 14.18 on 3519 degrees of freedom
Multiple R-squared:  0.3902,    Adjusted R-squared:  0.3898
F-statistic:  1126 on 2 and 3519 DF,  p-value: < 2.2e-16
```

A.4.20 - Reduced Model Summary

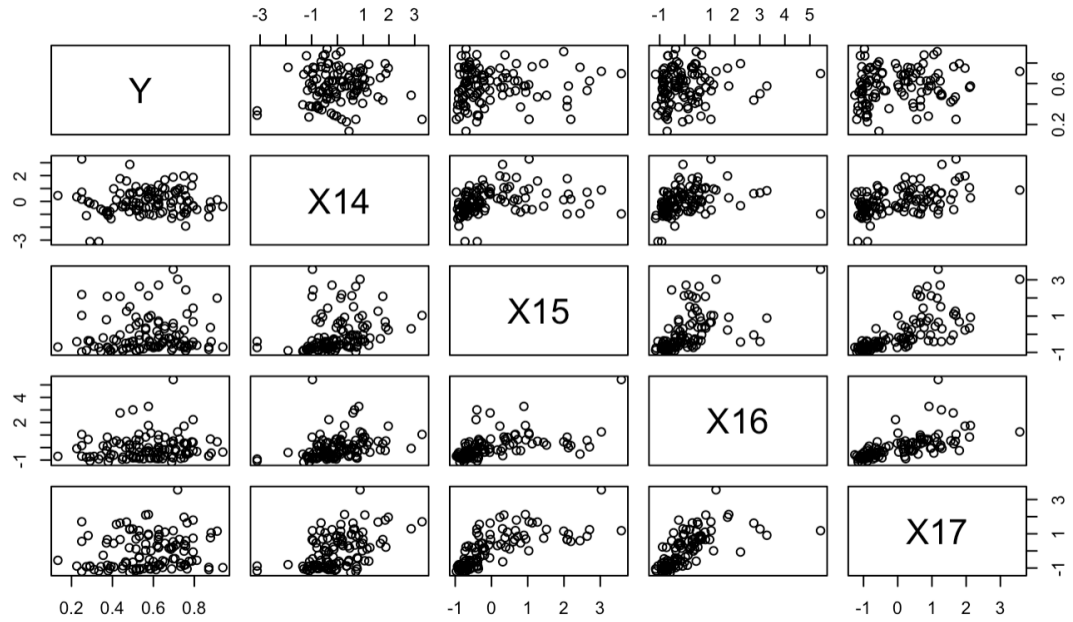
```
Analysis of Variance Table

Response: y
           Df Sum Sq Mean Sq F value    Pr(>F)
x11         1 400590  400590 1993.28 < 2.2e-16 ***
x12         1  51924   51924  258.36 < 2.2e-16 ***
Residuals 3519 707216     201
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
```

A.4.21 - Reduced Model ANOVA Table

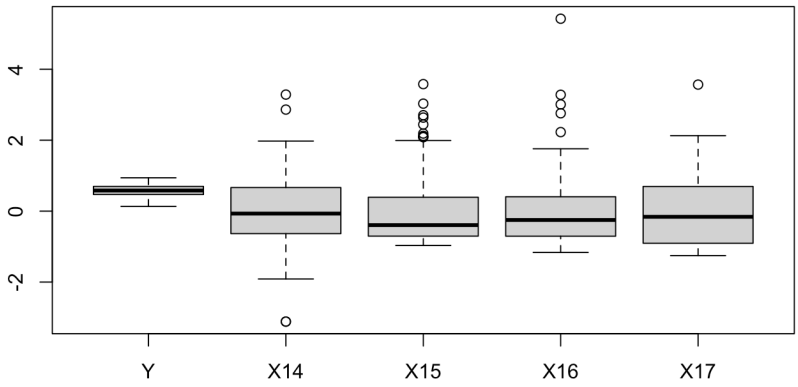


A.4.22 - Added-Variable Plot

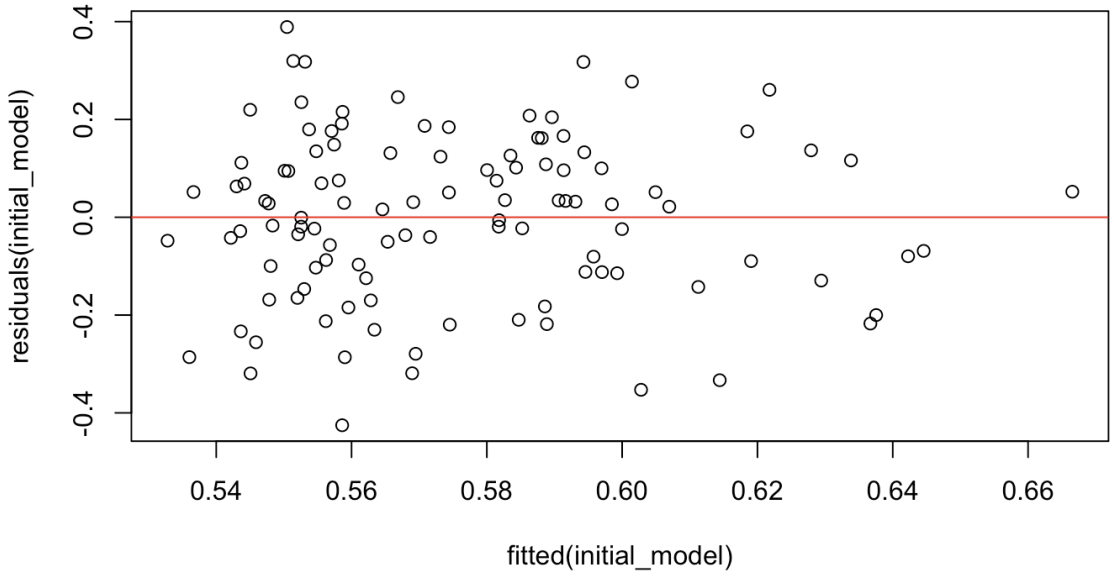
Le Rui Tay's Appendix



A.5.1. Scatterplot between the Y variable and the individual independent variables



A.5.2 Boxplot of variables



A.5.3 Residual plots

X14	X15	X16	X17
1.339664	2.133037	1.810791	3.073013

A.5.4 VIF results of first model

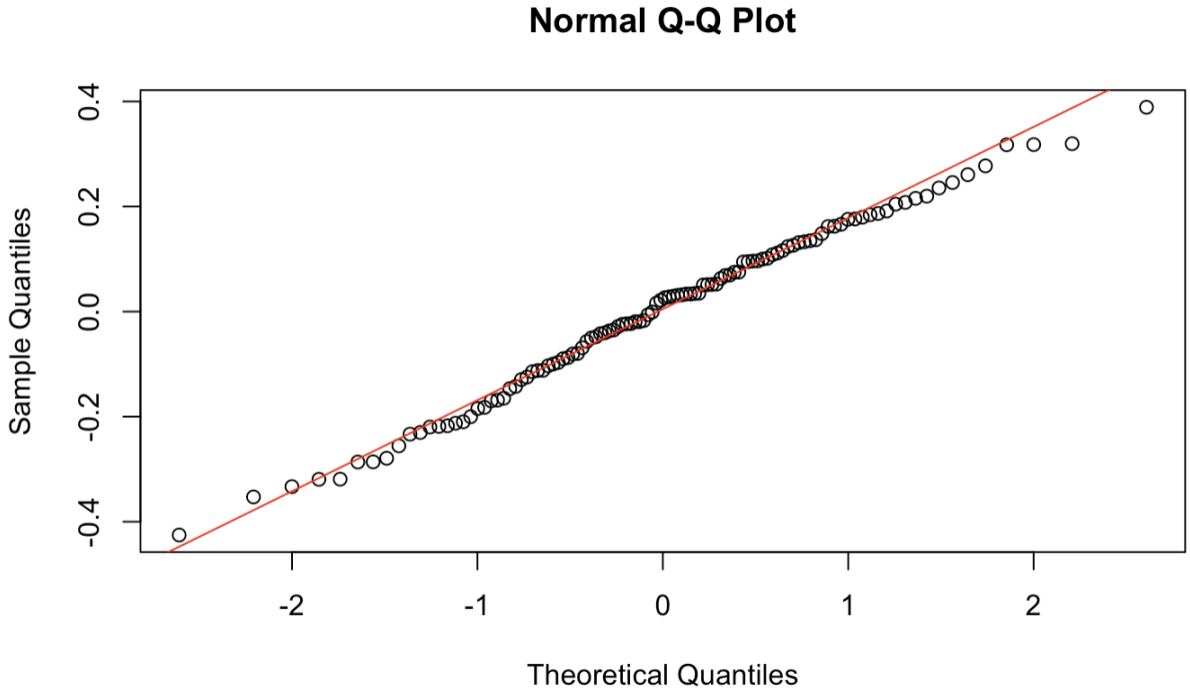
	Y	X14	X15	X16	X17
Y	1.00000000	0.04544029	0.07938113	0.1031885	0.1519371
X14	0.04544029	1.00000000	0.27434362	0.3154843	0.4911635
X15	0.07938113	0.27434362	1.00000000	0.5506193	0.7160233
X16	0.10318849	0.31548426	0.55061931	1.0000000	0.6596657
X17	0.15193709	0.49116353	0.71602331	0.6596657	1.0000000

A.5.5 Correlation table of variables

```
        studentized Breusch-Pagan test

data:  initial_model
BP = 1.1706, df = 4, p-value = 0.8829
```

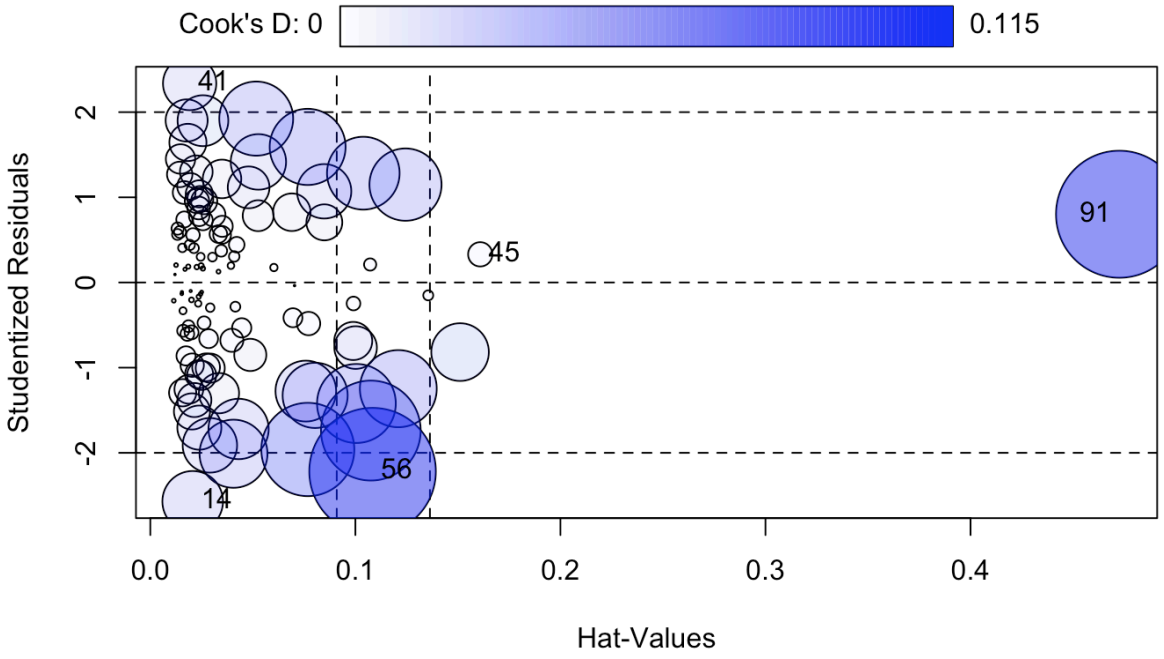
A.5.6 Bp test of residuals from first model

```
        Shapiro-Wilk normality test

data:  residuals(initial_model)
W = 0.99101, p-value = 0.6871
```



A.5.7 Shapiro test of residuals from first model and qqnorm diagram

	StudRes <dbl>	Hat <dbl>	CookD <dbl>
14	-2.5705340	0.02065694	0.026461351
41	2.3368188	0.01917177	0.020477664
45	0.3307513	0.16088993	0.004230994
56	-2.2176925	0.10837994	0.115263213
91	0.8012304	0.47256988	0.115432838



A.5.8 InfluencePlot results

	Model	Predictors	Adj_R2	Cp	AIC	BIC	PRESS
1	Y ~ 1	(Intercept)	0.000000000	-0.1409884	-73.78063	-68.37966	3.234168
2	Y ~ X14	X14	-0.007175321	1.6363022	-72.00799	-63.90655	3.315993
3	Y ~ X15	X15	-0.002899550	1.1793528	-72.47597	-64.37453	3.273911
4	Y ~ X16	X16	0.001487197	0.7105435	-72.95817	-64.85673	3.236840
5	Y ~ X17	X17	0.014039369	-0.6309006	-74.34973	-66.24829	3.215566
6	Y ~ X14 + X15	X14 + X15	-0.011655636	3.1140453	-70.54302	-59.74109	3.363032
7	Y ~ X14 + X16	X14 + X16	-0.007656851	2.6906542	-70.97868	-60.17675	3.319780
8	Y ~ X14 + X17	X14 + X17	0.005968395	1.2480138	-72.47621	-61.67429	3.300337
9	Y ~ X15 + X16	X15 + X16	-0.007100415	2.6317388	-71.03943	-60.23751	3.301937
10	Y ~ X15 + X17	X15 + X17	0.006632825	1.1776641	-72.54976	-61.74784	3.275770
11	Y ~ X16 + X17	X16 + X17	0.004840594	1.3674254	-72.35148	-61.54956	3.259071
12	Y ~ X14 + X15 + X16	X14 + X15 + X16	-0.016493782	4.6204546	-69.05108	-55.54867	3.387779
13	Y ~ X14 + X15 + X17	X14 + X15 + X17	-0.001159861	3.0120733	-70.72308	-57.22068	3.359299
14	Y ~ X14 + X16 + X17	X14 + X16 + X17	-0.003396637	3.2466897	-70.47760	-56.97519	3.343263
15	Y ~ X15 + X16 + X17	X15 + X16 + X17	-0.002628591	3.1661290	-70.56183	-57.05942	3.337693
16	Y ~ X14 + X15 + X16 + X17	X14 + X15 + X16 + X17	-0.010578517	5.0000000	-68.73573	-52.53285	3.419587

```
Start:  AIC=-382.9
Y ~ X14 + X15 + X16 + X17

       Df Sum of Sq    RSS     AIC
- X16   1  0.000355 3.0918 -384.89
- X14   1  0.004891 3.0964 -384.73
- X15   1  0.007263 3.0987 -384.64
- X17   1  0.047710 3.1392 -383.22
<none>              3.0915 -382.90

Step:  AIC=-384.89
Y ~ X14 + X15 + X17

       Df Sum of Sq    RSS     AIC
- X14   1  0.004875 3.0967 -386.72
- X15   1  0.006947 3.0988 -386.64
<none>              3.0918 -384.89
- X17   1  0.061887 3.1537 -384.71

Step:  AIC=-386.72
Y ~ X15 + X17

       Df Sum of Sq    RSS     AIC
- X15   1  0.005636 3.1023 -388.52
<none>              3.0967 -386.72
- X17   1  0.058935 3.1556 -386.64

Step:  AIC=-388.52
Y ~ X17

       Df Sum of Sq    RSS     AIC
<none>              3.1023 -388.52
- X17   1  0.073309 3.1756 -387.95

Call:
lm(formula = Y ~ X17, data = final_data)

Coefficients:
(Intercept)          X17
    0.57562      0.02593
```

A.5.9 Subset selection table and stepwise regression

```
Linear Regression

110 samples
  2 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 99, 99, 98, 100, 98, 100, ...
Resampling results:

  RMSE       Rsquared   MAE
  0.1703072  0.1146159  0.1403273

Tuning parameter 'intercept' was held constant at a value of TRUE
```

A.5.10 K-fold cross validation results

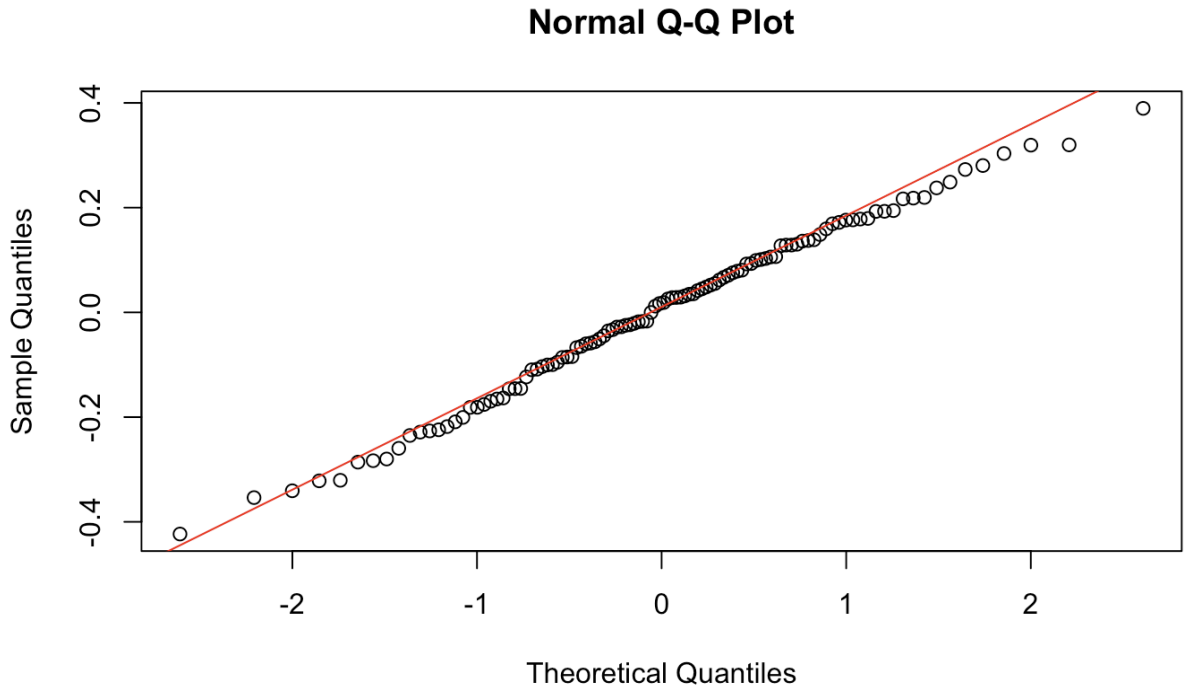
```
     X14      X17
1.317943 1.317943
```

A.5.11 VIF for second model

```
        studentized Breusch-Pagan test

data:  new_model
BP = 1.1012, df = 2, p-value = 0.5766
```

A.5.12 Bp test of residuals from second model

```
        Shapiro-Wilk normality test

data:  residuals(new_model)
W = 0.99184, p-value = 0.7588
```



A.5.13 Shapiro test of residuals from second model and qqnorm diagram

```
numeric(0)
         1          9         37         43         45         52         56         87
0.10316431 0.09967849 0.05868357 0.05767438 0.13503884 0.08444081 0.10823055 0.05687176
```

A.5.14 Outliers from hat leverage