STA302 Final Project - Fall 2022

Word count: 1558

Sung-chi (William) Wu - 1006990446

2022/12/20

Introduction

This research is motivated by the increase in attention toward sports betting in recent years. The popularity of sports betting is evident in the recent 2022 FIFA World Cup, where the world beats a record of spending 136 billion Euros on sports betting this year [2]. As a participant in this trending global entertainment, predicting the best performing NBA player each season allows me accumulate wealth to pay for my basic monthly expenses. Hence the research question: Predicting the number of points an NBA player scores in the 2017 NBA regular season using their number of assists, steals, blocks, rebounds, and turnovers. Previous researches have built predictive models on point scoring based on player physical characteristics that do not vary by season. For example, [3] produced a predictive model for point scoring based on player's Body Mass Index and number of assists and steals. [4] produced one model based on university, birth place and date. However, none of the researches produced a linear regression model of point scoring on all major game play statistics, which vary greatly season-by-season. In addition to the common variables that were considered in the previous studies such as the number of assists and steals, in this research I will also consider other game play statistics including number of blocks, rebounds and turnovers to get a better picture of how these statistics can used to predict point scoring.

Methodology

The Dataset

Our research aims to find the 'best' model to predict an NBA player's points scored by their number of assists, steals, blocks, rebounds, and turnovers during the 2017 NBA regular season. Appendix 1 includes a description of all selected variables.

Model Violations and Diagnostics

The data originates from Kaggle [1]. We begin our investigation by randomly splitting the data into training and testing data, by a ratio of 50/50. We first build a full model by regressing PTS on the five predictors using training data. Then, we plot the response PTS against the fitted values to check condition 1. If the points are randomly scattered around the identity function, condition 1 is satisfied. We then use a scatterplot of all the pairs of predictors to identify any relationship between predictors. An absence of non-linear relationships will indicate condition 2 is satisfied. Given no severe violations of the two conditions, we will proceed to checking assumptions by using residual versus fitted plots, residual versus predictor plots, and normal Quantile-Quantile plots.

An absence of systematic patterns and large cluster of residuals in the residual plots will indicate the linearity assumption and independence assumption are satisfied. An absence of discernible patterns, especially a fanning pattern, in the residuals plots will indicate homoscedasticity assumption is satisfied. Finally, if

the points form a straight diagonal line in the QQ-plot with minimal deviations at the ends, the normality assumption is satisfied. If any of the assumptions are violated, we will implement Box-Cox power transformation to identify which variable(s) to be transformed and the most appropriate form of transformation to mitigate violations. After we select our preferred model, I will identify outliers, leverage points, and influential points, using Cook's Distance and DFFITS and remove them if given sufficient rationale to do so.

Variable and Model Selection

First, we will use two-way stepwise selection method based on AIC/BIC to choose a subset of predictors. By removing different variables, we will try to find balance between complexity and predictive power between different models by assessing their adjusted R^2 and AIC/BIC. Second, for each new model built, we will also conduct Partial-F tests to ensure that none of the predictors removed are significant. If p-value of F-test $>$ 0.05, we proceed with the reduced model, otherwise we stick with the original model. Third, we will also check multicollinearity for reduced model by assessing their VIFs and remove any predictor(s) with VIF$>$5 and that are insignificant to the model by checking with partial-F test. We will also very the conditions and assumptions and proceed to model validation only with the models that do not have severe assumption violations. If multiple models have violated assumptions, I will pick the one that has the minimal violation.

Model Validation

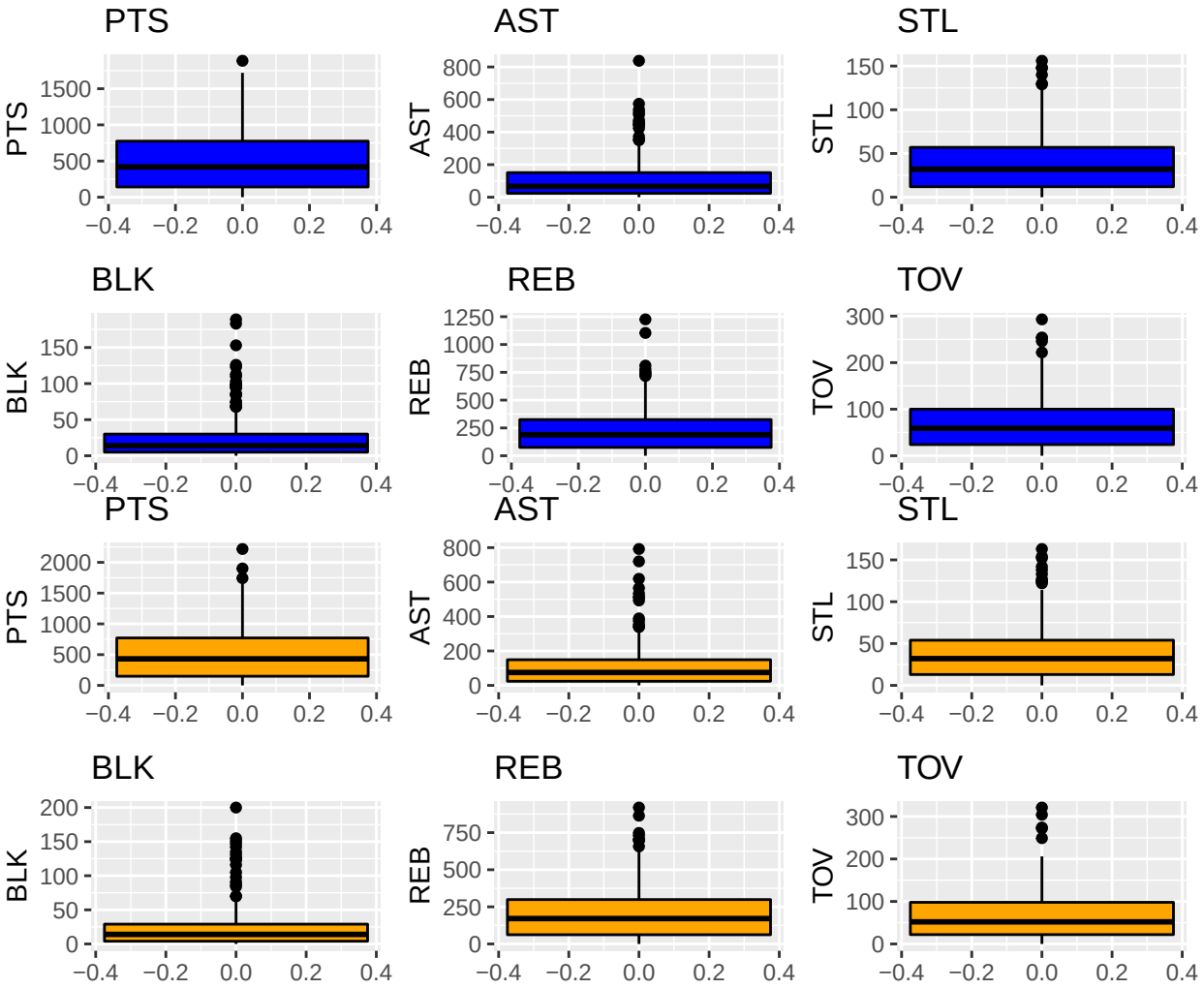
After checking for problematic points, we will proceed to use cross validation on testing data. We will build the same models and perform necessary transformations on the test data as we did with train data. We will then assess the predictor VIFs, number of influential points by Cooks and DFFITS, assumption violations, and summaries of the coefficients. If none of the assumptions are violated and the same predictors remain significant, then our model may be validated.

Results

Exploratory Data Analysis

Our train and test data each contains 245 observations and display. From Figure 1, we observe two similarities between training and testing data. First, the response and predictors between the two datasets share similar distribution that is right-skewed, indicating potential violations in Normality, Linearity and a poorly fitted model. Second, multiple outliers exist in all variables in both datasets, which might impact our ability to accurately measure means and spread.

Figure 2: Boxplots of Response and Predictors in Train Data and Testing Data



Model Violations and Diagnostics

The original full model shows violations in condition 1 and 2 (Appendix 2). In the response versus fitted values plot, a fanning pattern exists around the identity function, which implies a potential non-linear population relationship of the model, hence a violation in condition 1. We also observe that fanning patterns exist between every pairs of predictors, hence a violation in condition 2.

The original model shows violations in constant variance and normality (Appendix 3). The residual plots show fanning patterns in the residuals of each predictor. From the Q-Q plot, we observe there are severe

deviations from Normality at the two ends of the quantiles in all predictors, which may cause inaccurate p-values, confidence intervals and judgement on estimates. We use Box-Cox transformation by taking the square root of the response and all predictors in the transformed model, which shows significant improvement in the conditions and assumptions in all variables. We proceed to variable and model selection with the transformed full model, containing the response sqrt_PTS and predictors sqrt_AST, sqrt_STL, sqrt_BLK, sqrt_REB, and sqrt_TOV.

Variable Selection, Model Selection, and Checking for Multicollinearity

A test of multicollinearity on the transformed model shows that sqrt_TOV has a VIF > 5, implying multicollinearity. We try to mitigate this by building a reduced model without sqrt_TOV and use partial-F to see if this is valid. An F-value of 257.4 suggests sqrt_TOV is significant and should not be removed. Therefore, we will proceed to model selection with the full model.

The stepwise selection by AIC generates a reduced model without sqrt_BLK, whereas both sqrt_AST and sqrt_BLK are removed when stepwise selection by BIC is performed. As shown in Table 1, full model and the reduced models have similar adjusted R^2, AIC and BIC. Reduced Model 1 with sqrt_BLK removed has the lowest AIC, whereas Reduced Model 2 with two predictors removed have lowest BIC. To determine whether one or both predictors should be removed, I used partial F-test to compare the reduced models with the full model. When only sqrt_BLK is removed, the partial-F test generates an F-value of 1.53, suggesting sqrt_BLK can be removed. When both predictors are removed, F = 3.17, indicating we cannot remove both predictors.

The residual plots of Reduced Model 1 (Appendix 5) shows no severe deterioration in the conditions and assumptions. Therefore, we will proceed to model validation with Reduced Model 1, which contains four predictors that are all significant, as shown in Table 2.

Table 1: Summary of goodness measures for models fit to sqrt_PTS. The variables that were included in the full model were sqrt_AST, sqrt_STL, sqrt_BLK, sqrt_REB, sqrt_TOV. With a slight approximately equal adjusted-R^2, reduced model 1 has a slightly lower AIC value and the reduced model 2 has a slighly lower BIC value, indicating the model with sqrt_BLK removed is a better fitting model than the full model.

Model	Adjusted R^2	AIC	BIC
Full model	0.84	678.03	706.54
Reduced Model 1(-sqrt_BLK)	0.84	677.59	702.6
Reduced Model 2(-sqrt_BLK, -sqrt_AST)	0.84	680.43	701.94

Table 2: Parameter Estimates for the Transformed Full Model using Training Data

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	-0.4136	0.6498	-0.6366	0.5250
sqrt_AST	0.2062	0.0942	2.1887	0.0296
sqrt_STL	0.6575	0.1809	3.6355	0.0003
sqrt_REB	0.4315	0.0598	7.2186	0.0000
sqrt_TOV	1.2070	0.1882	6.4119	0.0000

Identifying Problematic Observations

As shown in Table 3, Cook's Distance identified 0 and DFFITS identified 18 influential points in both models, representing 7% of the training data. We do not have contextual reasons to remove them.

Model Validation

Adjusted R^2 is similar in both training and testing models with the testing model being slightly higher (0.8424 vs 0.9031). By comparing Table 2 and 4, we observe that there are minimal differences in the regression coefficients estimates and their standard errors. The predictors that are significant in the training model also significant in the testing model. Residual plots and QQ-plot in Appendix 5 also show that the assumptions are fulfilled in the testing data. With these criteria met, the final model is successfully validated.

Table 3: Summary of characteristics of two candidate models in the training and test datasets. Full Model uses sqrt_AST, sqrt_STL, sqrt_BLK, sqrt_REB, sqrt_TOV as predictors, while Reduced Model uses sqrt_ASL, sqrt_STL, sqrt_REB, sqrt_TOV as predictors. Response is sqrt_PTS (Points scored) in both models. Coefficients are presented as estimate ± SE (* = significant t-test at $\alpha = 0.05$)

Characteristic	Full (Train)	Full (Test)	Reduced (Train)	Reduced (Test)
Largest VIF value	7.0202951	14.231884	7.0120678	14.1287309
# Cook's D	0	0	0	0
# DFFITS	18	18	18	18
Violations	none	none	none	none

Table 4: Table 3: Parameter Estimates for the Transformed Full Model Using Testing Data

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	-0.4513	0.4919	-0.9176	0.3598
sqrt_AST	0.2780	0.1178	2.3593	0.0191
sqrt_STL	0.6800	0.1761	3.8602	0.0001
sqrt_REB	0.4126	0.0623	6.6262	0.0000
sqrt_TOV	1.1871	0.2098	5.6577	0.0000

Discussion

Our final model explains 84.24% of the total variation in the response square root PTS. The model suggests that there is are positive relationship between the square root of the response and all the predictors. Specifically, one unit of increase in the square root of AST, STL, REB, TOV will lead to an expected increase in the square root of PTS by 0.2062, 0.6575, 0.4315, and 1.2070, respectively, keeping other predictors fixed. Therefore, to predict the top-scoring candidate in the 2017 NBA season, we will want to look for players that achieve high performance in these four key metrics.

There are two limitations to our study. First, due to our decision to keep sqrt_TOV in our model, multicollinearity remains a problem in the final model as shown in Table 3, where the training model and testing model both contain predictors with VIF > 5. This may potentially reduce the precision of our coefficient estimates and weakens the statistical power of our model. Second, we identify a slight violation in constant variance among the predictors even after Box-cox transformation, which may deteriorate the accuracy of our p-value in t-tests and misinterpret the significance of some predictors. For instance, with a p-value close to 0.05, AST may potentially be insignificant and hence be removed from the model during stepwise selection.

Reference

[1] Goldstein, O. (2017, May 3). NBA players stats - 2014-2015. Kaggle. Retrieved December 20, 2022, from https://www.kaggle.com/datasets/drgilermo/nba-players-stats-20142015

[2]Li, X. (2021, December 1). National Basketball Association Most Valuable player prediction based on machine learning methods. SPIE Digital Library. Retrieved October 17, 2022, from https://www.spiedigitallibrary.org/conference-proceedings-of-spie/12079/120791Q/National-Basketball-Association-Most-Valuable-Player-prediction-based-on-machine/10.1117/12.2623094.full?SSO=1

[3] Yakowicz, W. (2022, November 18). Gamblers expected to wager more than $160 billion on the World Cup-here's where the smart money is going. Forbes. Retrieved December 20, 2022, from https://www.forbes.com/sites/willyakowicz/2022/11/17/gamblers-expected-to-wager-more-than-160-billion-on-the-world-cup-heres-where-the-smart-money-is-going/?sh=728e2f3e7e17

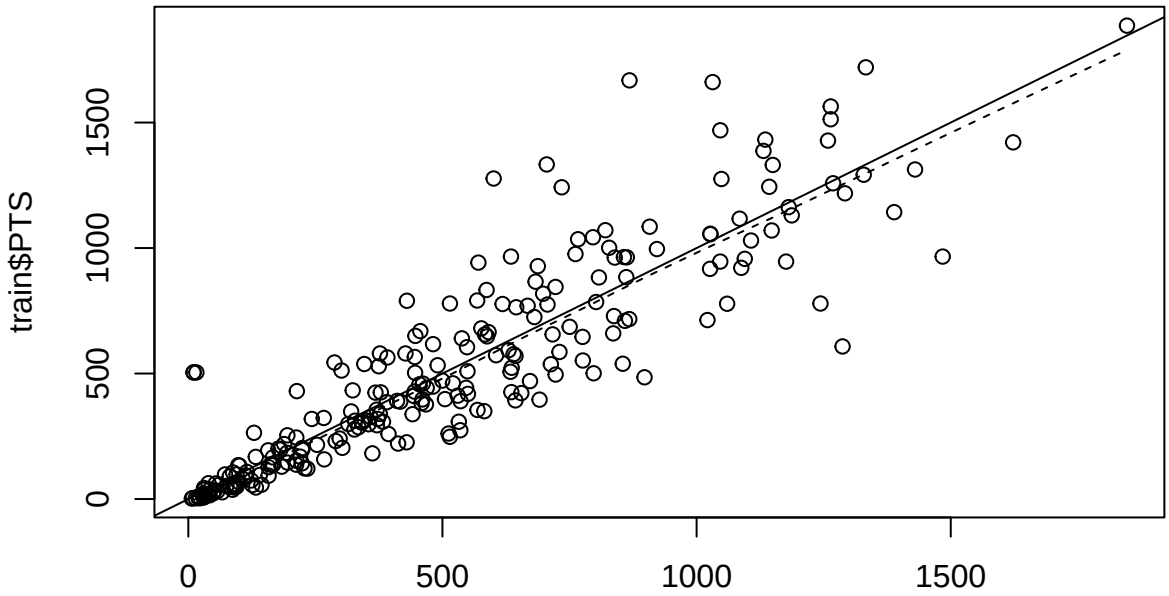
[4] Zhao, C. (2017). Predictive model for NBA teams, player metrics, and optimal strategies for team lineups (Order No. 10681438). Available from ProQuest Dissertations & Theses Global. (2023809112). Retrieved from http://myaccess.library.utoronto.ca/login?qurl=https%3A%2F%2Fwww.proquest.com%2Fdissertations-theses%2Fpredictive-model-nba-teams-player-metrics-optimal%2Fdocview%2F2023809112%2Fse-2%3Faccountid%3D14771

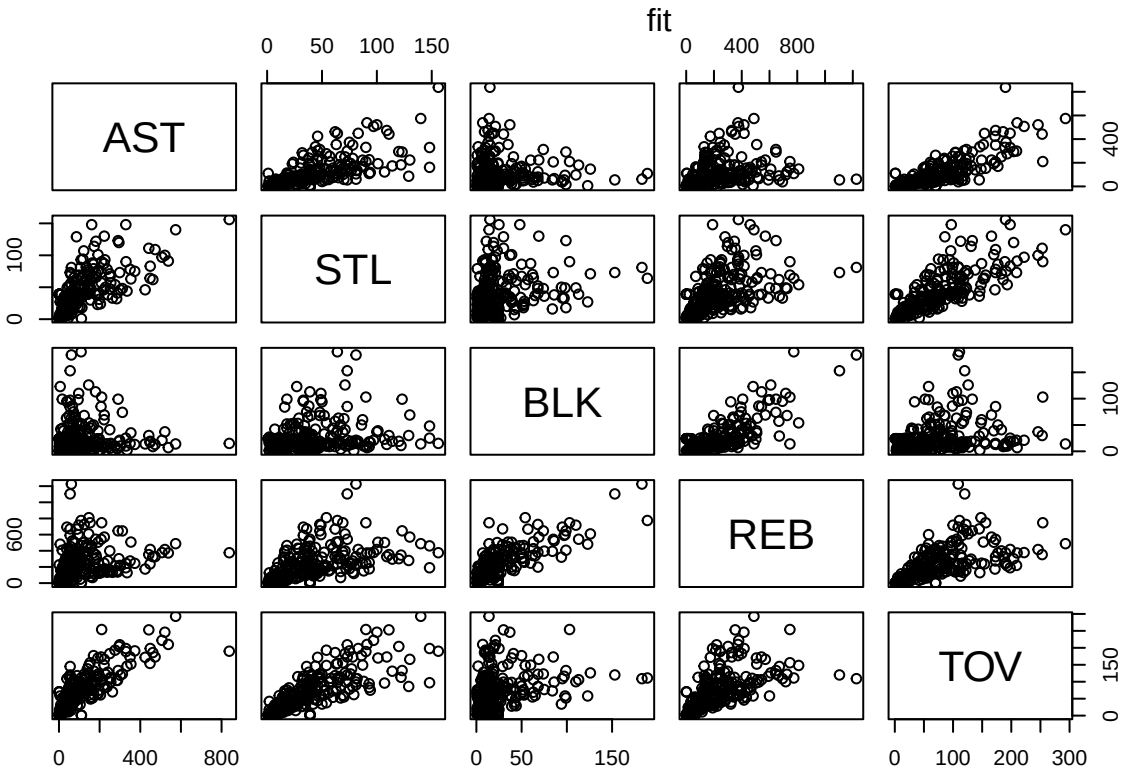
Appendices

Appendix 1: Description of Selected Variables

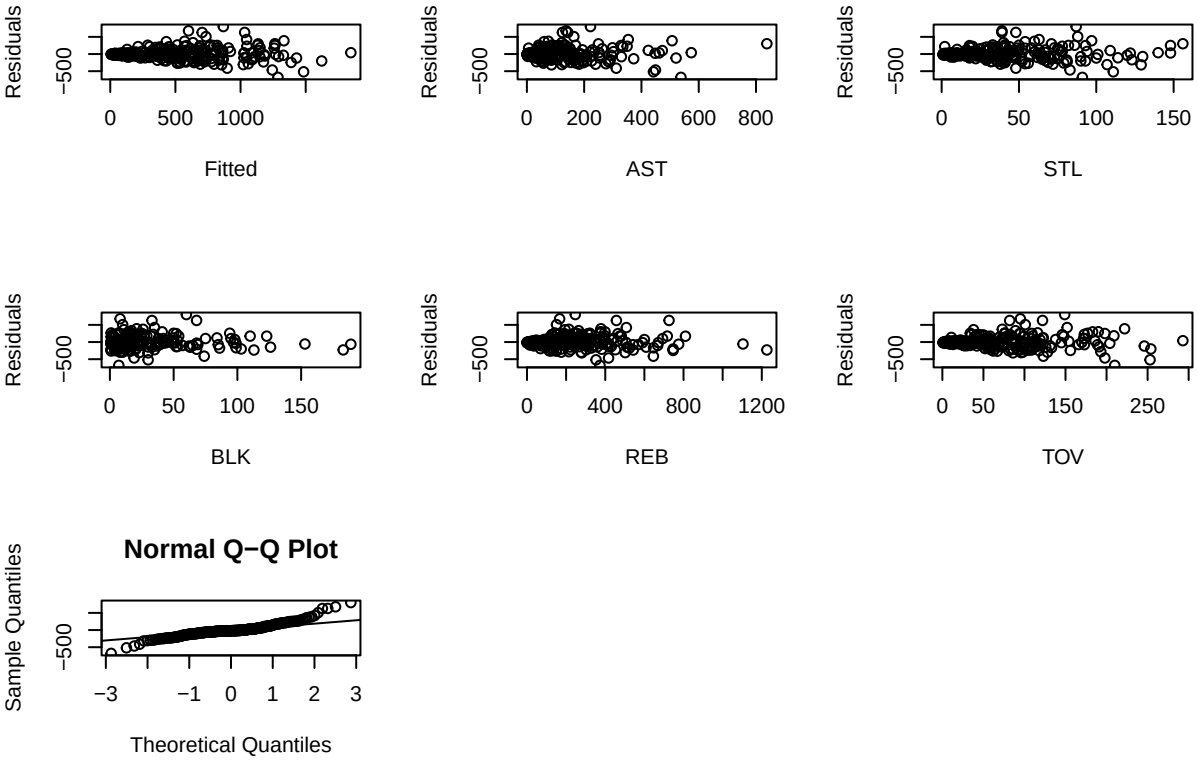
Variable	Description
PTS	Total Number of points the player scored in the 2017 NBA regular season
AST	Total Number of assists the player achieved in the 2017 NBA regular season
STL	Total Number of steals the player achieved in the 2017 NBA regular season
BLK	Total Number of blocks the player achieved in the 2017 NBA regular season
REB	Total Number of rebounds the player achieved in the 2017 NBA regular season
TOV	Total Number of turovers the player achieved in the 2017 NBA regular season

Appendix 2: Checking Conditions of the Original Full Model (Training)

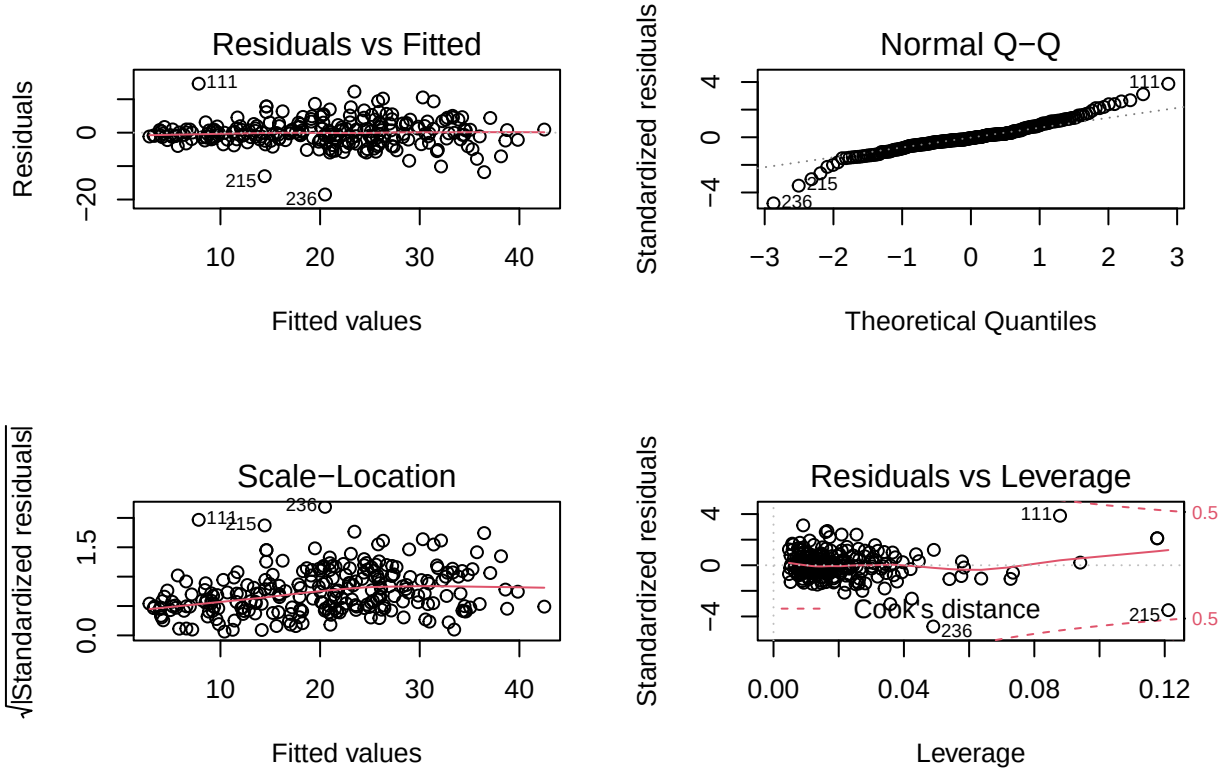




Appendix 3: Checking Assumptions of the Original Full Model (Training)



Appendix 4: Checking Conditions and Assumptions of the Reduced Model (Training)



Appendix 5: Checking Checking Conditions and Assumptions of the Final Model (Testing)