Predicting Basketball Player's Player Efficiency Rating Using In-Game Stats Metrics

Lyndon Kelley
Tola Ouk
Jay Song
Martha Valdes

Introduction:

Basketball is a team sport created by James Naismith in December of 1891. The sport involves two opposing teams made up of 15 players each. Five players from each team will be on the court at any one time, and player substitutions are during stops in action. The court is rectangular and has two elevated hoops on opposite sides. When players from one team are on the court, they will advance a ball toward the opposing team's basket and try to get the ball into the basket. They will also try to prevent the opposing team from getting the ball into their basket. There are a multitude of basketball leagues, but the basketball organization held in the highest regard and used for this report is the National Basketball Association, abbreviated as the NBA.

Player efficiency rating, abbreviated as PER, is an advanced statistic created by John Hollinger that helps quantify the value that a player adds to his team. The formula is in the appendix. The PER aggregates individual stats and team stats to measure collected during a game to measure player productivity. In other words, it provides another way to see which player outputs the gaudiest stats. A key aspect of PER is that it can adjust for how much a player plays and team pace. Players that play more minutes have more possession to do something with the ball. In addition, teams with a higher pace have players with larger per-game outputs due to an increased number of possessions. PER takes both into account so that players are not penalized for these factors. Though PER cannot perfectly measure how good a player is due to its inability to capture parts of defense or offense not recorded as stats, it can still look at a player's statistical output by examining metrics that are measured.

The following question was examined for this study: which in-game statistics are the best predictors of Player Efficiency Rating (PER) of NBA players? Knowing which stats increase or decrease PER is a question that NBA front offices must answer. First, players that can produce large output for their teams are instrumental for teams winning. Most championship contenders have NBA players that can produce at a high level. In other words, having players with high PERs can indicate how likely a team is to win a championship for a given year. Knowing the circumstances in which players have higher statistical outputs can help determine whether a team can contend. In addition, PER can also help front offices get an idea of how much a player might make during negotiations. Generally, players with higher productivity get paid more. They can also win awards during a season based on a plethora of factors. Winning these awards and producing at a high level means that a player can make significantly more money during contract negotiations. By predicting a player's PER, front-offices can predict how much their players and team will get paid.

The dataset used is from Basketball-Reference. It is from December 22, 2020, to February 18, 2021 from the 2021 NBA season. Six explanatory variables were chosen from this data set: win shares, age, free-throw percentage, usage, true shooting, and three point frequency. Win shares are a statistic that estimates how many wins a player adds to his teams through a season. When a player on offense takes more contact than deemed necessary, players will get two shots without defenders guarding them. Free throw percentage is how accurate a player is at taking those shots. Age is how old a player is in years. Usage indicates how often a player will have the ball in his hands when on the court. True shooting measures how accurate a player is from the free-throw line and the field. Three point frequency measures how often a player shoots threes.

Methodology:

To determine if PER can be predicted using the various statistics and advanced metrics, we will go through a full process of selecting, building, and evaluating a linear model. To begin, we will examine our explanatory variables through box plots and correlation matrices. This will give us an initial idea of how our data looks, as well as if we should be wary of any potential outliers or signs of multicollinearity.

Once we have examined our variables, we will go through the model selection process to determine if we can eliminate any predictors to simplify our model. Having too many predictors that do not contribute to the model decreases its robustness, so we want to figure out what the most important explanatory variables to keep are. For our dataset, we will use the regression with all six explanatory variables as our full candidate model for testing. We will use a few different techniques for selecting which final model to use. First, we will conduct general linear hypothesis tests by using analysis of variance (ANOVA) and the F-statistics determined via the ratio of the mean squares of regression (MSR) and the mean squares of error (MSE). We will also use stepwise regression, as well as forward and backward elimination methods. The stepwise selection uses partial F-tests/t-tests to add or remove predictors from a model given a cutoff.

We will then diagnose our model to determine if they fit assumptions made for creating a linear regression. For a linear regression model, we have a set of four assumptions we need to meet. We need to assume that the response is a linear function of the predictors, the errors are independent, the errors are normally distributed, and the errors have equal variances. To analyze this, we will investigate the residuals from our linear model. We will take a look at a Residuals vs. Fitted values plot, a sequence plot, and a normal probability plot to see if our model meets those assumptions. We will also take a look at what points are potentially influential using the following methods: studentized deleted residuals, DFFITS, and Cook's distances. This will allow us to determine if any remedial measures should be used.

In this report, we will use k-fold cross validation to assess our models. k will be set equal to 5. For this, our data is split into 5 random subsets, then the linear model will be trained on 4/5 of the data five times, which each subset left out once. This allows us to use a "larger" test set to have tighter confidence intervals for our parameters, and allow us to test the accuracy of our linear model.

Data Analysis:

As previously stated in the introduction, the following question was examined: which in-game statistics are the best predictors of Player Efficiency Rating (PER) of NBA players? To do so, we need to perform various tests with our explanatory variables and compare multiple models.

To prepare the data for analysis, we filtered out the players who played less than 200 minutes to limit our model to players with significant statistics gathered. For our data set, we ended up with 334 players. Since we have only 6 predictors, our data set is more than sufficient to build a model.

Before building the model, it is necessary to understand how the variables are related to each other. The scatter plots in Figure 1.1 show how each predictor variable is related to the response variable. This analysis is important to making decisions regarding the variables that contribute to the robustness of a regression model.

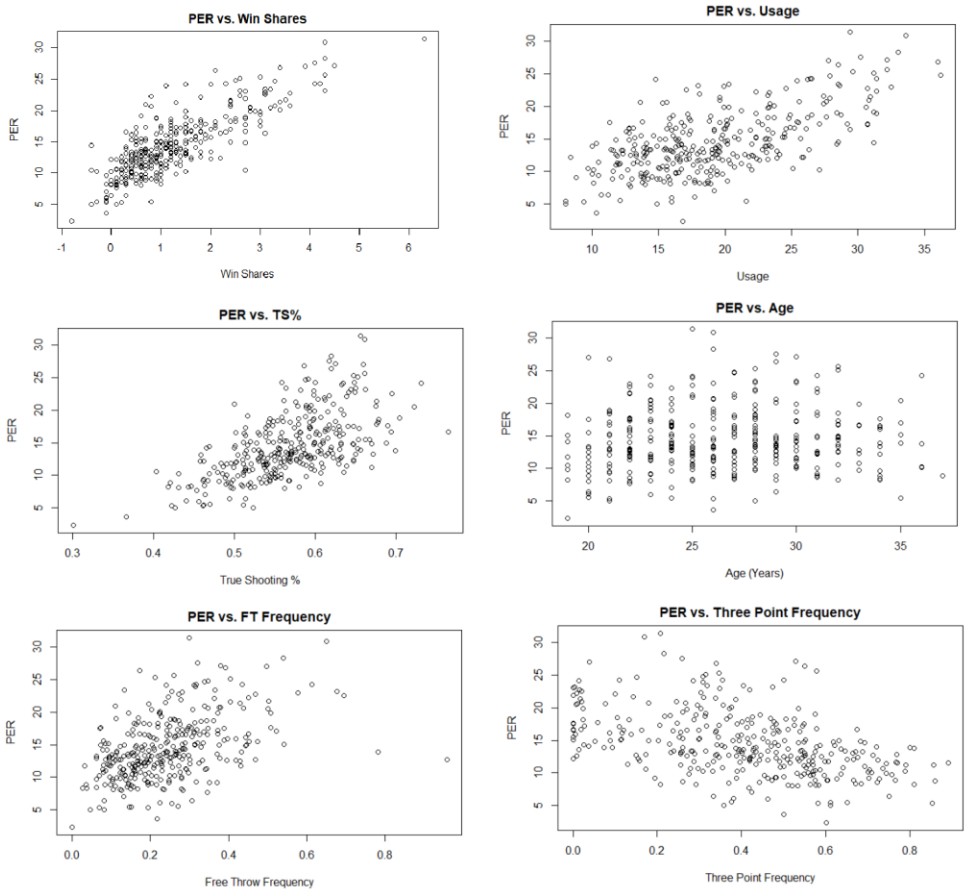

Figure 1.1. Scatter Plot of each predictor variable vs response variable PER.

In the PER vs Win Shares diagram, a strong and positive relationship can be detected. When this result is compared with the other scatterplots, it can be concluded that Win Shares is the variable with the strongest linear relationship with the response variable. This means that it must be one of the variables to use in the linear regression model.

Other predictor variables, such as usage and true shooting, also show a positive relationship with PER. However, in the case of these two variables, we can not say that it is a strong linear relationship. The pattern of the graph allows us to think that another type of relationship, such as the quadratic, might better describe the interaction between each of the variables and PER. Three point frequency is the other variable that shows an interesting relationship that should be taken into account. Unlike usage and true shooting, three point frequency and the response variable have a negative relationship.

The scatter plots showing the relationship between PER vs Age and PER and FT Frequency demonstrate a weak linear relationship between these two predictor variables and PER. As such, they may not have a significant contribution to the regression model. This result must be confirmed later with the construction and analysis of the regression model.

The correlation matrix in Figure 1.2 allows us to confirm the conclusions obtained with the analysis of the previous scatterplots. Win shares have a correlation coefficient of approximately 0.8 with PER. Usage and true shooting have a correlation of approximately 0.6 in regard to PER. Both free throw percentage and frequency of threes have a correlation of approximately 0.4 with respect to PER. Finally, the variable with smallest correlation to PER which has a correlation coefficient of less than 0.1. Consequently, there are 3 predictor variables that can be considered

significant for the model. In the case of the other variables with weaker relationships, the contribution to the regression model will be determined from the evaluation of the full model and the reduced models.

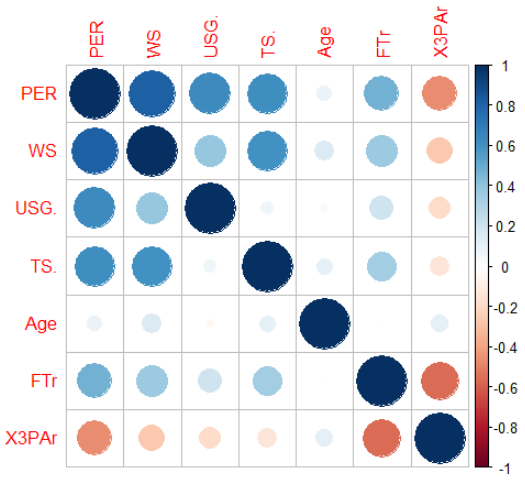

Figure 1.2. Correlation Matrix.

Through the boxplot diagrams it is possible to have a fairly complete vision of the descriptive measures of each variable, dispersion and symmetry of the study data. Additionally, we can clearly detect each observation that deviates in unusual ways from the rest of the data. Figure 1.3 shows the boxplot of the response variable. The boxplot of the predictor variables can be consulted in the appendix B figure B.1 of this document.

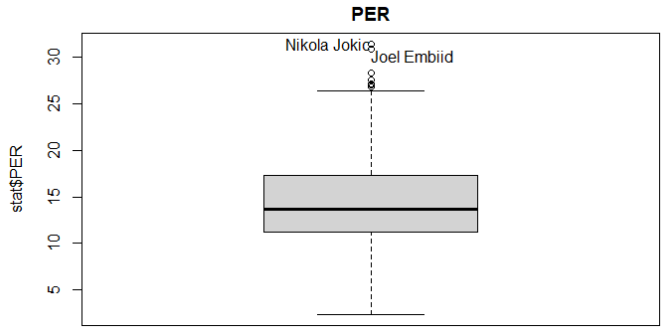

Figure 1.3. PER Boxplot.

The first element that must be analyzed is the dispersion of the data around its center. In the boxplot analysis we can detect that the variables have a symmetric distribution around the median. Age is the only variable that shows a slight asymmetry. Taking into account the context of the data, it is understandable. Players on professional teams tend to be under 30 years old. The age boxplot reflects this with 75% of the players less than 30 years old and the oldest player being above 35. The asymmetry in the distribution of the data indicates a normal distribution of these variables.

An important result obtained with the analysis of boxplot is the detection of possible outliers in the box plots corresponding to the predictor variables win shares, true shooting (TS) and three point frequency (X3PAr). Outliers are also detected in the values of the response variable PER. Therefore, we made an analysis of the atypical situation in the context. For example, Nikola Jokic and Joel Embiid have two outlier PER values. They were also the league's best-performing players in this period. It makes sense that their PER is well above the rest of the players'

observations. A similar analysis was performed for each variable regarding the players designated as outliers. Considering that players have particular roles in the game, it makes sense that other variables like win shares, usage percentage, true shooting and free-throw have outliers too. The outlier points detected correspond to players whose role in the game increase that particular statistic, so it makes sense that their results differ from the rest. Therefore, the outliers can be considered valid observations for the model. However, this conclusion should be complemented with the analysis of the influence that these observations have in terms of the prediction of the response variable.

To fit the data, a few models were built and tested. First is the Full model with all six variables. This results in this fitted model equation:

$$PER = -6.50389 + 1.89752\ WS + 0.361212\ USG. + 23.31425\ TS. + 0.03880\ Age - 0.45058\ FTr - 5.92438\ X3PAr$$

The adjusted R^2 of this model is 0.8956, suggesting that approximately 90% of the variation of PER in the data can be explained by this model. The full results are in Appendix B, and the hypothesis tests of all models can be found in Appendix C.

Hypothesis Testing:

$\alpha = 0.05$

Null Hypothesis Ho: $\beta_{WS} = \beta_{USG.} = \beta_{TS.} = \beta_{Age} = \beta_{FTr} = \beta_{X3PAr} = 0$

Alternative Hypothesis Ha: At least one of these parameters is not zero

F-Critical = F(0.05, 6, 327) = 0.2716

Reject Null Hypothesis if F-Statistics > F-Critical

Since F-Statistics = 477.1 > F-Critical, we reject the Null Hypothesis.

Therefore, at least one of the parameters is not zero.

The F-test of this model suggests that at least one of these six variables is significant. The T-Test for this full model suggests four variables to be significant at 95% level. They are win shares (WS), usage percentage (USG.), true shooting (TS.), and Three Point Frequency (X3PAr). The Reduced Model will use these four variables to fit the data.The regression function for the Reduced model is:

$$PER = -5.54312 + 1.93103\ WS + 0.35779\ USG. + 22.97922\ TS. - 5.65701\ X3PAr$$

The adjusted R^2 of this model is 0.8952, which is only 0.0006 less than that of the Full model. The hypothesis test using Anova F-test comparing this Reduced model with the Full model confirms that age and free throw rate (FTr) are not significant variables at 95% confidence level.

Next, the Stepwise Selection method was implemented. The Backward, Forward, and Both Directions methods were all used. All three result in the same model:

$$PER = -6.52887 + 1.89386\ WS + 0.36043\ USG. + 23.01520\ TS. - 5.78399\ X3PAr + 0.03847\ Age$$

The adjusted R^2 of this model is 0.8958. This model uses five variables; four variables that are in the reduced model plus the Age variable. The hypothesis test confirms that free throw rate (Ftr) is not a significant variable in estimating PER.

Figure 1.4 summarizes the results from the three models:

Model	Adjusted R^2
Full Model (6 Variables): PER = -6.50389 + 1.89752 WS + 0.361212 USG. + 23.31425 TS. + 0.03880 Age - 0.45058 FTr - 5.92438 X3PAr	0.8956
Reduced Model (4 Variables): PER = -5.54312 + 1.93103 WS + 0.35779 USG. + 22.97922 TS. - 5.65701 X3PAr	0.8952
Stepwise Regression Model (5 Variables): PER = -6.52887 + 1.89386 WS + 0.36043 USG. + 23.01520 TS. - 5.78399 X3PAr + 0.03847 Age	0.8958

Figure 1.4. Adjusted R^2 of Full and Reduced Models and Stepwise Regression Model

Overall, the model with the highest Adjusted R^2 is the one with 5 variables that is given by the Stepwise Regression method. However, the difference over the other two models' is not very significant; only 0.06% more than the Reduced Model for example. In addition, the full results of the Stepwise Regression method also show only a small change in AIC value when age is used in the model. Hence, the addition of variable Age does not add significant improvement to the model's ability to explain the variations in the data.

To determine if any of our cases were considered influential, we examined a few different factors. First, we plotted the deleted studentized residuals. Deleted residuals are a method where each point in the data set is fitted using a linear regression built without that point, and the residual is calculated. A deleted studentized residual transforms the residual by standardizing it using the standard errors, providing a sort of t-value. If the absolute value of a t-value is greater than the t-critical value for the family of t-residuals, it would be considered influential. The t-critical value for our case, with n=334 cases and 4 parameters, is $t_{(0.05/668,\ df\ =\ -329)} = 3.836$. The following plot shows the deleted studentized residuals with our cutoff:

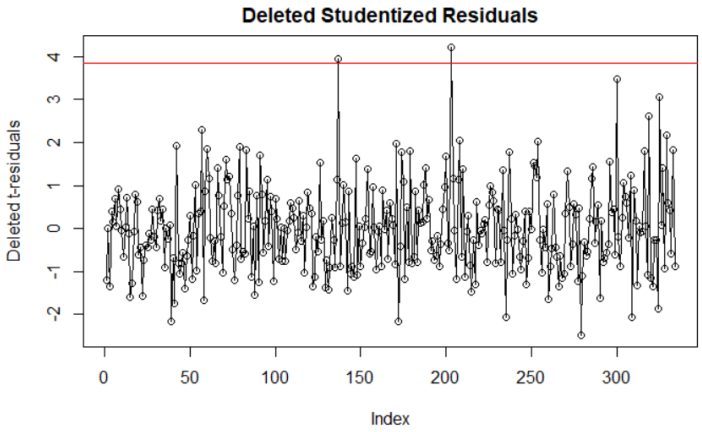

Figure 1.5. Deleted Studentized Residuals.

For our dataset, none of the cases reached the lower threshold, so it is not shown on the plot. We can see that there are two points which are above the threshold. The points belong to Willy Hernangomez and CJ McCollum, with t−values of 3.94 and 4.16 respectively. We also investigated the Cook's Distances, located in Appendix B13, which takes

into account the residuals and the effect the residual has on the predictions. All of the distances were much lower than 1, indicating that none were very influential. We also investigated the DFFITS, which is the number of standard deviations a fitted value changes when each case is omitted. We would consider an observation influential if a case changes the standard deviation in either direction greater than $2\sqrt{(p/n)}$. For this, we found that 11 players were considered influential to our model using DFFITS. The figures for the analysis are located in Appendix B14 and B15. We want to keep these data points though, since it is important to take into account exceptional players when creating the model. Based on the correlation matrix, it was unlikely that we had any issues with multicollinearity.

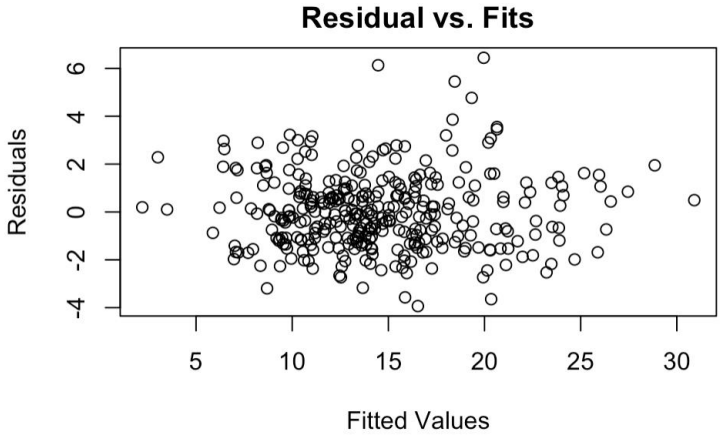

Figure 1.6. Residual vs Fitted Values Scatter Plot.

When creating and evaluating a linear model, an important aspect to keep is the acronym LINE. LINE is an acronym that lists 4 assumptions that must be met for linear regression models: linear relationship, independence among the error terms, normally distributed error terms, and equal variance error terms. As such, the constructed final linear model had to match these assumptions. Some type of linear relationship must exist between the explanatory and response variable. Based on the scatterplots in Figure 1.1, usage, win shares, three point frequency, all fit this assumption. To determine whether the error terms were independent, the sequence plot Figure B.16 was created to see the correlation between the error terms. There was no obvious correlation pictured, no fanning effect, and no funneling effect pictured in the plot, so the final model meets this assumption. To determine whether the final model had a normal distribution among the residuals, the Q-Q plot Figure B.17 was created. Most of the points line up along a straight line indicating normality, so the final model meets this assumption. Finally, the fitted and residual values were plotted against one another in Figure 1.6 to see whether the variances were equal. Though true shooting does exhibit some slight fanning, it's relatively minor meaning the final model meets this assumption as well. The final model is below:

PER = -5.54312 + 1.93103 WS + 0.35779 USG. + 22.97922 TS. - 5.65701 X3PAr

To determine the performance of the final model, the RMSE and R^2 were calculated using the k-fold cross validation method, with k=5. The output for the cross validation is the appendix. The RMSE is the root mean of square error which measures the difference between the values

predicted by the model and the observed values in a data set. Lower RMSEs indicate whether a linear regression model is performing well. For the final linear model for PER, the calculated RMSE was approximately 1.596. This indicates a mean difference of 1.596 units between the final model's predicted PER and the actual PER for each player. This relatively small RMSE indicates the final model is highly effective at predicting PER when using usage, true shooting, win shares, and three point frequency. Another way to measure a model's effectiveness is the R^2 value. R^2 is the percentage of the variation explained by a model and its explanatory variables. The R^2 for the final model was approximately 0.897 meaning about 89.7% of variation in PER value was explained by the model. When the RMSE and $R^{2.}$ are taken together, the final model appears to be highly effective at predicting PER based on its explanatory variables.

Conclusion:

The purpose of this study was to determine which in-game stats had the largest effects on a player's PER. Based on the results of the final regression model, win shares, usage, true shooting and three point frequency were the best explanatory variables at predicting PER. As previously stated in data analysis, these 4 variables explain almost 90% of the variation of PER when taken together. In contrast, age and free throw percentage had almost no impact on PER and did not need to be included in the final model. For our RMSE of 1.596, we can compare it to the range of PER, which is about 29 between the minimum and maximum of our dataset. An RMSE of ~6% is extremely good considering the data. Thus, we were able to fairly successfully determine which in-game stats were the most important for predicting a player's PER.

While the dataset was relatively easy to work with, there were some slight difficulties encountered while trying to analyze it. One of the issues was properly dealing with the outliers in the dataset. There were relatively few outliers on the low-end of PER, as many were removed out when the data was prepared and filtered. The larger issue occurred with players that had significantly higher PER that influenced the data set. Normally, it would be possible to simply remove these data points from the data set to get a more accurate picture of the mean. However, this was not possible for this dataset. Because many of the players that are outliers were the best in the NBA for that period of time, their statistics must be included. As a result, the data is slightly skewed by these players and their outlier PERs. Still, the data is relatively centralized in terms of PER. But most of PERs were grouped together meaning these high-end players probably did not skew the data too much. Overall, this dataset was relatively easy to prepare and analyze.

There are more ways by which the model's performance and accuracy could be improved. Something that would give a far better idea as to how well the final model performs is using a data set that includes player stats from multiple seasons. Increasing the size of the dataset by including a longer period of time would be useful in creating a model that is more representative of how each in-game metric directly affects PER. It might be less subject to outliers. In addition, it might further decrease the RMSE which would make for a robust model that is less subject to high and low-end outliers. In addition, considering the impact that higher order terms and interaction terms could also increase the performance of the model or increase the R^2 and explain more variance within the model. If the combination of two terms has a larger effect than considering the terms individually, it's important to use that to improve our model for the future.

Overall, this is a good starting point for analysis for this type of data. Sports analytics tend to use more complicated techniques to determine player value, but it is encouraging to see success with linear regression.

References:

Basketball Reference. (n.d.). *2020-21 NBA Player Stats: Per Game.* Retrieved November 4, 2021, from **https://www.basketball-reference.com/leagues/NBA_2021_per_game.html**

Basketball Reference. (n.d.). *2020-21 Calculating PER* Retrieved November 4, 2021 https://www.basketball-reference.com/about/per.html

National Basketball Association (NBA) Logo. (n.d.). *Symbols.com.* Retrieved November 28, 2021, from https://www.symbols.com/symbol/national-basketball-association-%28nba%29-logo.

Nba 2020-2021 Season Player Stats. (2021, February 18). Kaggle. Retrieved November 4, 2021, from **https://www.kaggle.com/umutalpaydn/nba-20202021-season-player-stats**

Appendix A (Response Variable Formula):

PER formula = (1 / MP) *[3P+ (2/3) * AST+ (2 - factor * (team_AST / team_FG)) * FG+ (FT *0.5 * (1 + (1 - (team_AST / team_FG)) + (2/3) * (team_AST / team_FG)))- VOP * TOV- VOP * DRB% * (FGA - FG)- VOP * 0.44 * (0.44 + (0.56 * DRB%)) * (FTA - FT)+ VOP * (1 - DRB%) * (TRB - ORB)+ VOP * DRB% * ORB+ VOP * STL+ VOP * DRB% * BLK- PF * ((lg_FT / lg_PF) - 0.44 * (lg_FTA / lg_PF) * VOP)]

Appendix B: Additional Tables and Figures.

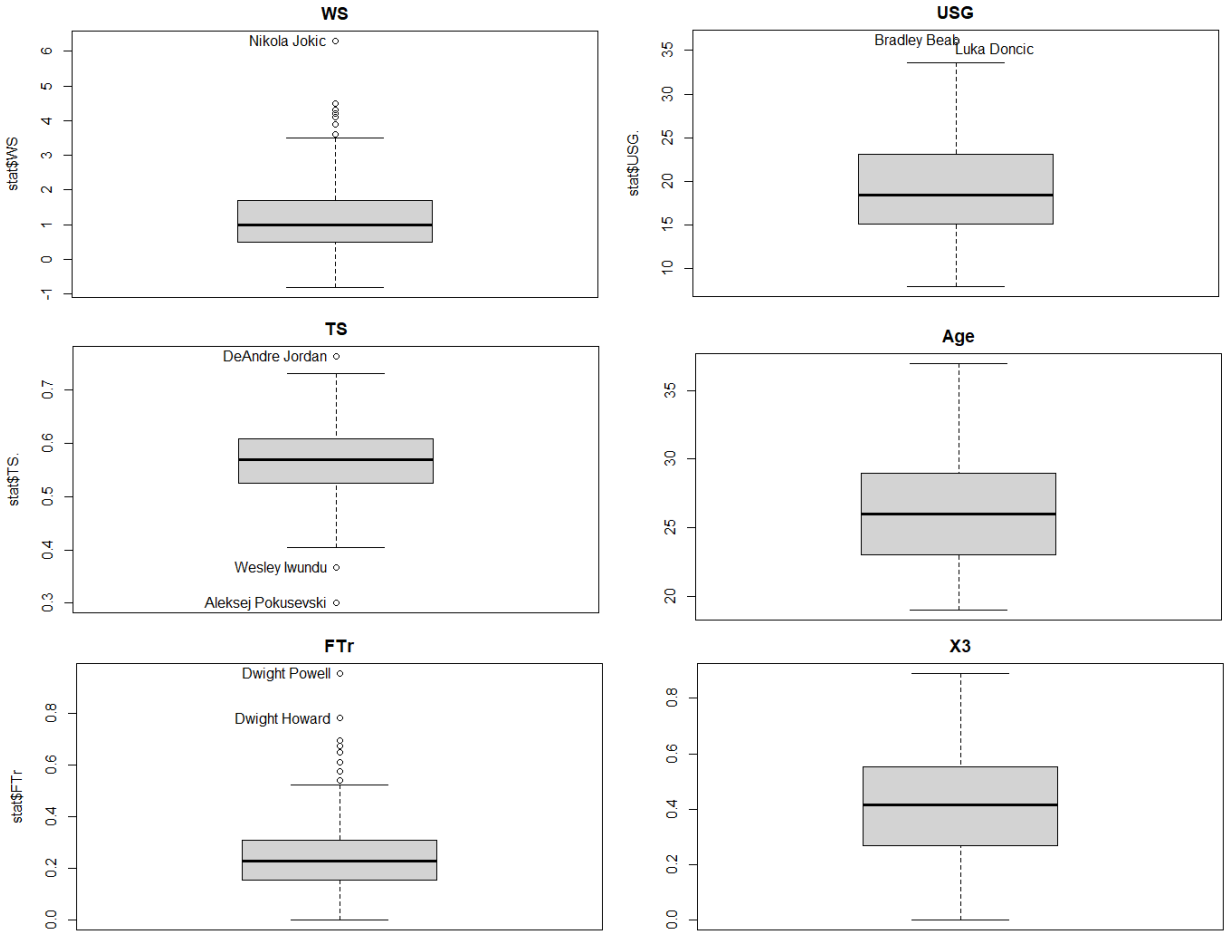

Figure B.1. Predictor variables boxplot.

	PER	WS	USG.	TS.	Age	FTr	X3PAr
PER	1.000	0.818	0.638	0.610	0.084	0.470	-0.467
WS	0.818	1.000	0.388	0.606	0.155	0.369	-0.267
USG.	0.638	0.388	1.000	0.066	-0.031	0.210	-0.183
TS.	0.610	0.606	0.066	1.000	0.104	0.340	-0.140
Age	0.084	0.155	-0.031	0.104	1.000	-0.006	0.120
FTr	0.470	0.369	0.210	0.340	-0.006	1.000	-0.567
X3PAr	-0.467	-0.267	-0.183	-0.140	0.120	-0.567	1.000

Figure B2. Correlation Matrix for Variables.

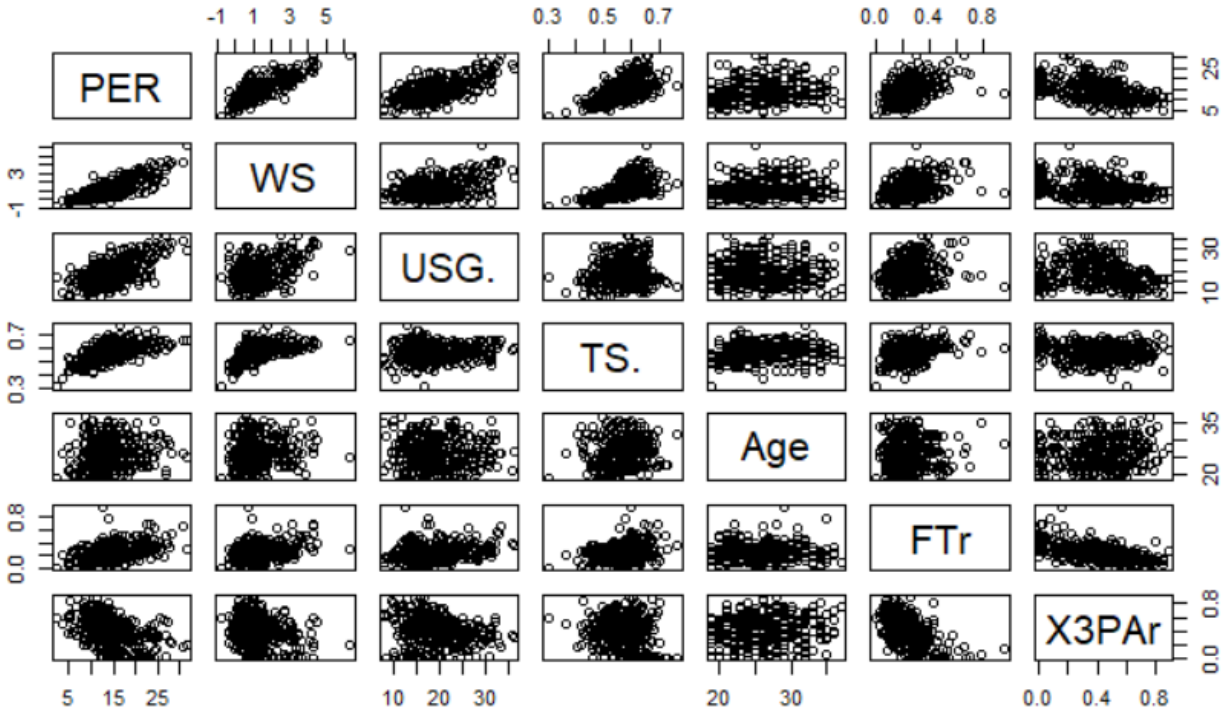

Figure B.3. Paired Correlation Scatter plots.

PER	WS	USG.	TS.	Age	FTr	X3PAr
Min. : 2.40	Min. :-0.800	Min. : 8.00	Min. :0.3010	Min. :19.00	Min. :0.0000	Min. :0.0000
1st Qu.:11.20	1st Qu.: 0.500	1st Qu.:15.10	1st Qu.:0.5258	1st Qu.:23.00	1st Qu.:0.1542	1st Qu.:0.2677
Median :13.65	Median : 1.000	Median :18.45	Median :0.5690	Median :26.00	Median :0.2270	Median :0.4160
Mean :14.50	Mean : 1.258	Mean :19.18	Mean :0.5670	Mean :26.33	Mean :0.2432	Mean :0.4034
3rd Qu.:17.30	3rd Qu.: 1.700	3rd Qu.:23.07	3rd Qu.:0.6090	3rd Qu.:29.00	3rd Qu.:0.3068	3rd Qu.:0.5515
Max. :31.40	Max. : 6.300	Max. :36.20	Max. :0.7640	Max. :37.00	Max. :0.9570	Max. :0.8910

Figure B.4. Five Number Summary for Variables

WS	USG.	TS.	Age	FTr	X3PAr
2.088855	1.272026	1.751187	1.063785	1.674808	1.544824

Figure B.5. Variance Inflation Factors

Table. Adjusted R^2 of Full and Reduced Models

Model	**Adjusted R^2**
Full Model: PER = -6.50389 + 1.89752 WS + 0.361212 USG. + 23.31425 TS. + 0.03880 Age - 0.45058 FTr - 5.92438 X3PAr	0.8956
Reduced Model: PER = -5.54312 + 1.93103 WS + 0.35779 USG. + 22.97922 TS. - 5.65701 X3PAr	0.8952

Stepwise Regression:

Stepwise Regression: (Full results)
start.model = lm(PER ~ 1, data = stat)
step.backward = step(per.lm, direction = "backward")
step.forward = step(start.model, direction = "forward", scope = formula(per.lm))
step.both = step(per.lm, direction = "both")

```
> step.backward = step(per.lm, direction = "backward")
Start:  AIC=321.25
PER ~ WS + USG. + TS. + Age + FTr + X3PAr

        Df Sum of Sq     RSS    AIC
- FTr    1      0.67  838.71 319.52
<none>                838.04 321.25
- Age    1      8.22  846.26 322.51
- X3PAr  1    333.53 1171.57 431.16
- TS.    1    428.35 1266.39 457.15
- WS     1    640.41 1478.45 508.86
- USG.   1   1125.15 1963.19 603.58

Step:  AIC=319.52
PER ~ WS + USG. + TS. + Age + X3PAr

        Df Sum of Sq     RSS    AIC
<none>                838.71 319.52
- Age    1      8.09  846.80 320.73
- X3PAr  1    440.54 1279.25 458.53
- TS.    1    441.71 1280.43 458.83
- WS     1    640.21 1478.92 506.97
- USG.   1   1129.11 1967.82 602.36
```

Figure B.6. Backward results.

```
> step.forward = step(start.model, direction = "forward", scope = formula(per.lm))
Start:  AIC=1070.03
PER ~ 1

        Df Sum of Sq    RSS     AIC
+ WS     1    5474.5 2700.6  702.09
+ USG.   1    3328.7 4846.4  897.40
+ TS.    1    3045.7 5129.4  916.35
+ FTr    1    1809.3 6365.8  988.48
+ X3PAr  1    1786.3 6388.7  989.68
+ Age    1      57.7 8117.3 1069.67
<none>               8175.1 1070.03

Step:  AIC=702.09
PER ~ WS

        Df Sum of Sq    RSS    AIC
+ USG.   1    987.73 1712.8 552.01
+ X3PAr  1    545.45 2155.1 628.73
+ FTr    1    267.55 2433.0 669.24
+ TS.    1    169.36 2531.2 682.46
<none>               2700.6 702.09
+ Age    1     15.22 2685.4 702.20

Step:  AIC=552.01
PER ~ WS + USG.

        Df Sum of Sq    RSS    AIC
+ TS.    1    433.28 1279.6 456.61
+ X3PAr  1    425.66 1287.2 458.59
+ FTr    1    195.29 1517.6 513.58
<none>               1712.8 552.01
+ Age    1      0.60 1712.2 553.90

Step:  AIC=456.61
PER ~ WS + USG. + TS.

        Df Sum of Sq    RSS    AIC
+ X3PAr  1    432.77  846.8 320.73
+ FTr    1    108.00 1171.6 429.16
<none>               1279.6 456.61
+ Age    1      0.32 1279.2 458.53

Step:  AIC=320.73
PER ~ WS + USG. + TS. + X3PAr

       Df Sum of Sq    RSS    AIC
+ Age   1    8.0858 838.71 319.52
<none>              846.80 320.73
+ FTr   1    0.5394 846.26 322.51

Step:  AIC=319.52
PER ~ WS + USG. + TS. + X3PAr + Age

       Df Sum of Sq    RSS    AIC
<none>              838.71 319.52
+ FTr   1   0.67283 838.04 321.25
```

Figure B.7. Forward results.

```
> step.both = step(per.lm, direction = "both")
Start:  AIC=321.25
PER ~ WS + USG. + TS. + Age + FTr + X3PAr

        Df Sum of Sq     RSS    AIC
- FTr    1      0.67  838.71 319.52
<none>                838.04 321.25
- Age    1      8.22  846.26 322.51
- X3PAr  1    333.53 1171.57 431.16
- TS.    1    428.35 1266.39 457.15
- WS     1    640.41 1478.45 508.86
- USG.   1   1125.15 1963.19 603.58

Step:  AIC=319.52
PER ~ WS + USG. + TS. + Age + X3PAr

        Df Sum of Sq     RSS    AIC
<none>                838.71 319.52
- Age    1      8.09  846.80 320.73
+ FTr    1      0.67  838.04 321.25
- X3PAr  1    440.54 1279.25 458.53
- TS.    1    441.71 1280.43 458.83
- WS     1    640.21 1478.92 506.97
- USG.   1   1129.11 1967.82 602.36
```

Figure B.8. Both directions results.

```
> step.backward

Call:
lm(formula = PER ~ WS + USG. + TS. + Age + X3PAr, data = stat)

Coefficients:
(Intercept)           WS         USG.          TS.          Age        X3PAr
   -6.52887      1.89386      0.36043     23.01520      0.03847     -5.78399

> step.forward

Call:
lm(formula = PER ~ WS + USG. + TS. + X3PAr + Age, data = stat)

Coefficients:
(Intercept)           WS         USG.          TS.        X3PAr          Age
   -6.52887      1.89386      0.36043     23.01520     -5.78399      0.03847

> step.both

Call:
lm(formula = PER ~ WS + USG. + TS. + Age + X3PAr, data = stat)

Coefficients:
(Intercept)           WS         USG.          TS.          Age        X3PAr
   -6.52887      1.89386      0.36043     23.01520      0.03847     -5.78399
```

Figure B.9. Stepwise Model

Backward:

```
Step:  AIC=319.52
PER ~ WS + USG + TS + Age + X3

       Df Sum of Sq     RSS    AIC
<none>              838.71 319.52
- Age   1      8.09  846.80 320.73
- X3    1    440.54 1279.25 458.53
- TS    1    441.71 1280.43 458.83
- WS    1    640.21 1478.92 506.97
- USG   1   1129.11 1967.82 602.36
```

Forward:

```
Step:  AIC=319.52
PER ~ WS + USG + TS + X3 + Age

       Df Sum of Sq     RSS    AIC
<none>              838.71 319.52
+ FTr   1   0.67283 838.04 321.25
```

```
Step:  AIC=319.52
PER ~ WS + USG + TS + Age + X3

       Df Sum of Sq     RSS    AIC
<none>              838.71 319.52
- Age   1      8.09  846.80 320.73
+ FTr   1      0.67  838.04 321.25
- X3    1    440.54 1279.25 458.53
- TS    1    441.71 1280.43 458.83
- WS    1    640.21 1478.92 506.97
- USG   1   1129.11 1967.82 602.36
```

Both directions:

Figure B.10. Final steps of Backward, Forward and Both directions.

```
> summary(step.lm)

Call:
lm(formula = PER ~ WS + USG. + TS. + X3PAr + Age, data = stat)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9575 -1.1247 -0.1144  0.8993  6.4758

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.52887    1.18408  -5.514 7.12e-08 ***
WS           1.89386    0.11969  15.823  < 2e-16 ***
USG.         0.36043    0.01715  21.013  < 2e-16 ***
TS.         23.01520    1.75111  13.143  < 2e-16 ***
X3PAr       -5.78399    0.44066 -13.126  < 2e-16 ***
Age          0.03847    0.02163   1.778   0.0763 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.599 on 328 degrees of freedom
Multiple R-squared:  0.8974,    Adjusted R-squared:  0.8958
F-statistic: 573.8 on 5 and 328 DF,  p-value: < 2.2e-16
```

Figure B.11. R output for step.lm

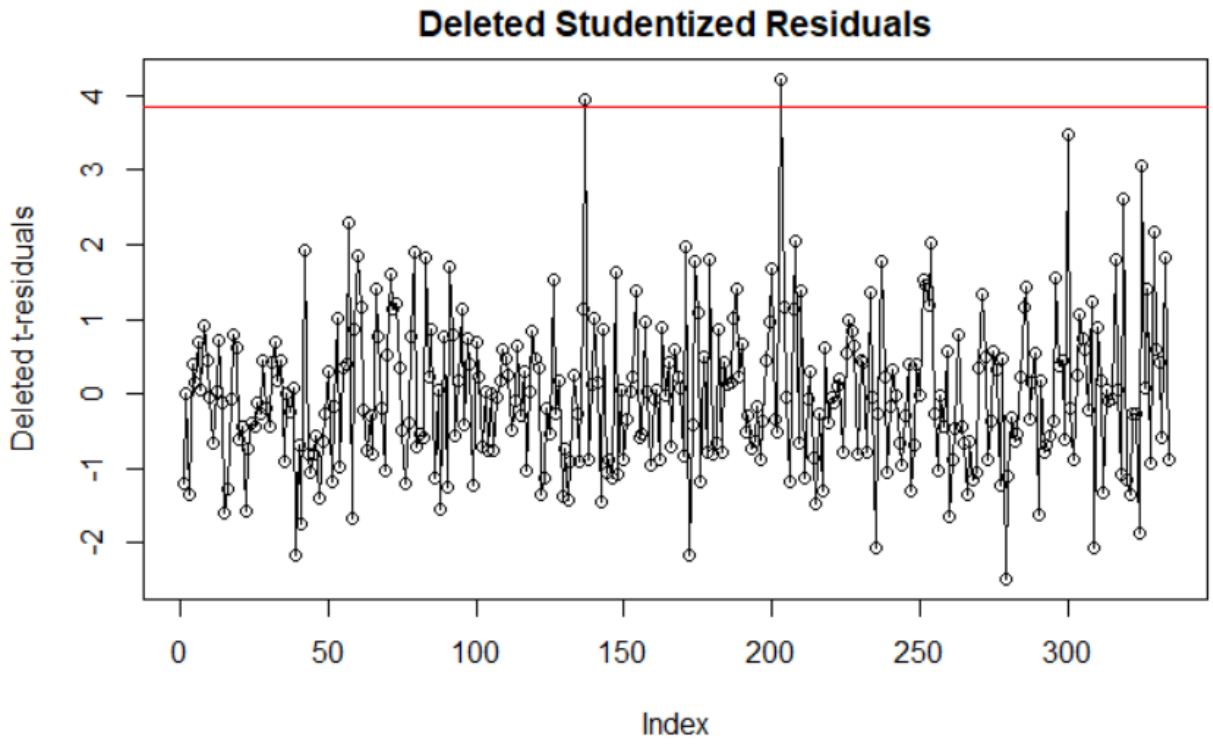

Figure B.12. Deleted Studentized Residuals vs. T-critical Value

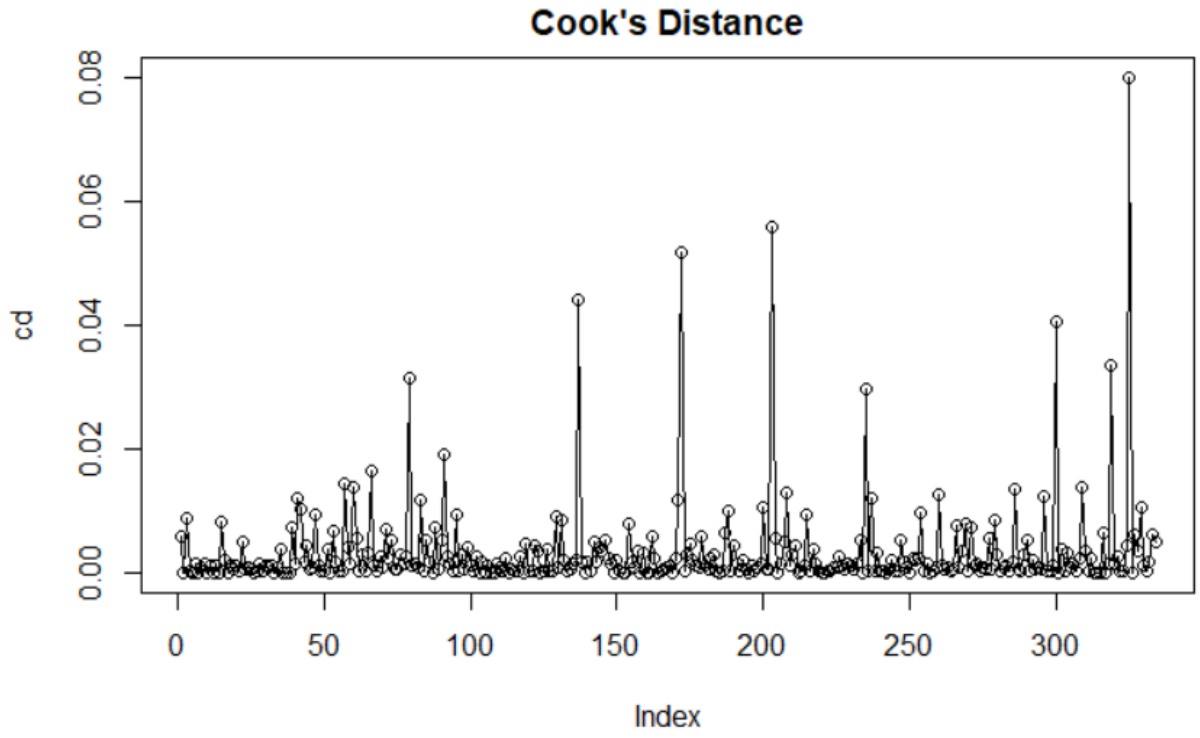

Figure B.13. Cooks Distance vs. Index of players

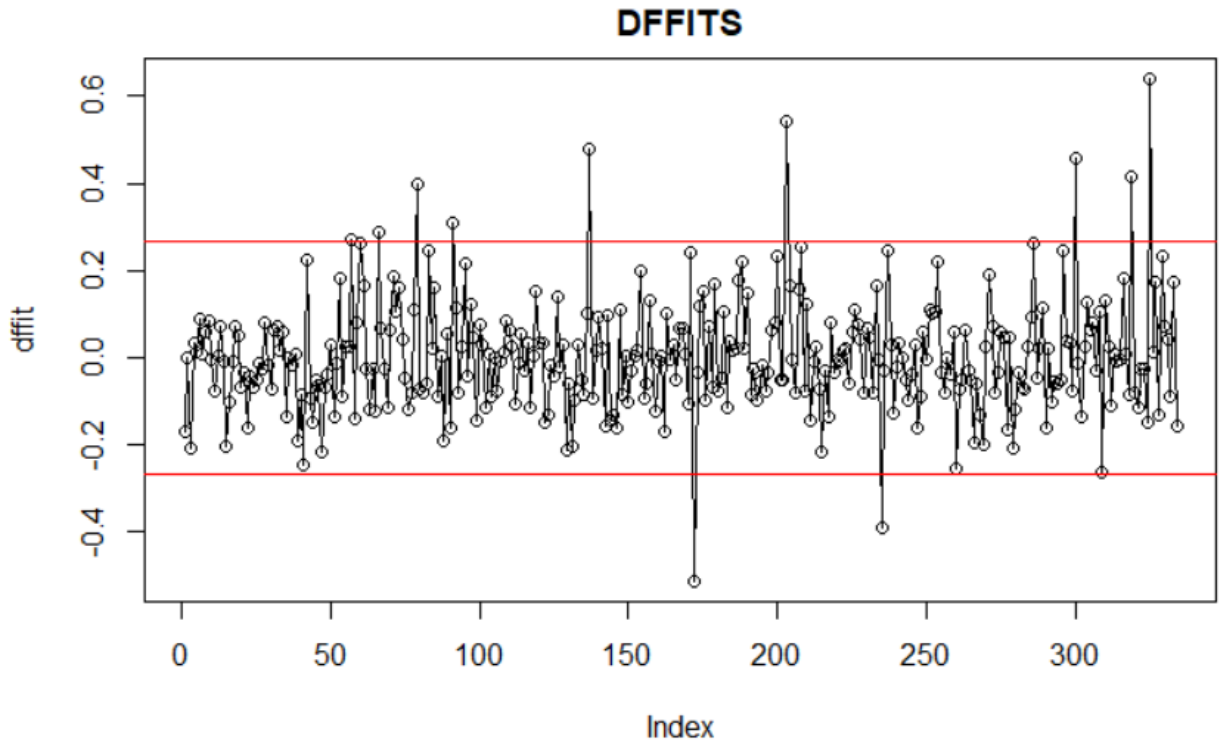

Figure B.14. DFFITS

```
     Jimmy Butler         Gary Clark          Ed Davis      Andre Drummond  Willy Hernangomez
        0.2704196          0.2862941          0.3973267          0.3108872          0.4804631
   DeAndre Jordan       CJ McCollum  Royce O'Neale Karl-Anthony Towns   Hassan Whiteside
       -0.5118730          0.5427606         -0.3882464          0.4589757          0.4134815
Robert Williams
        0.6412931
```

Figure B.15. Players exceeding DFFITS Threshold

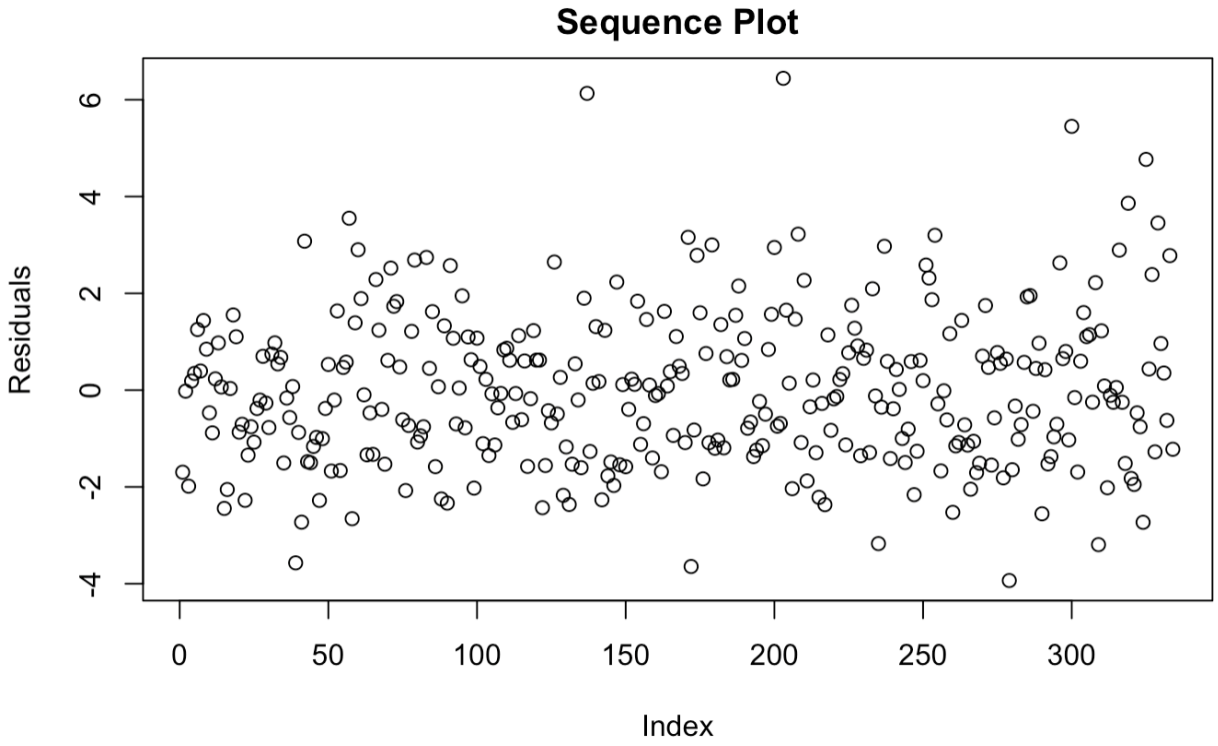

Figure B.16. Sequence Plot for Residual Terms.

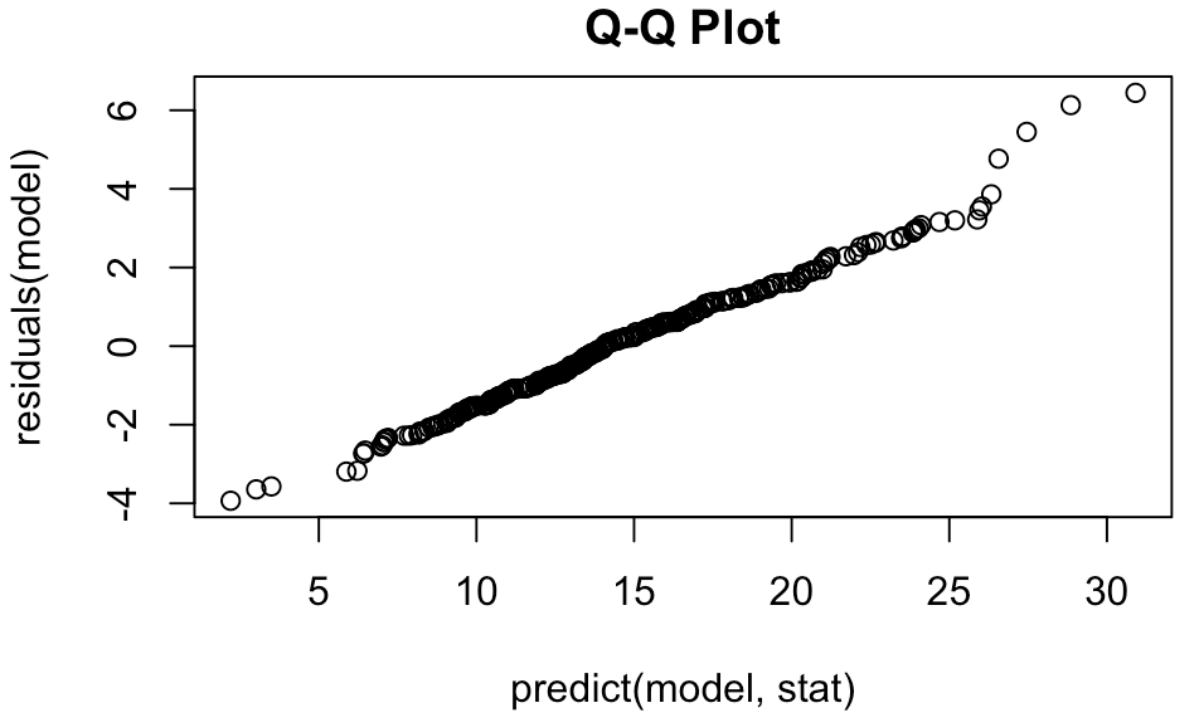

Figure B.17. Q-Q Plot of Final Model.

Appendix C. Hypothesis Tests.

Full Model, Reduced Model, and Stepwise Regression Model:

Full Model

```
> summary(per.lm)

Call:
lm(formula = PER ~ ., data = stat)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9346 -1.1361 -0.1097  0.9011  6.4429

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.50389    1.18642  -5.482 8.42e-08 ***
WS           1.89752    0.12004  15.808  < 2e-16 ***
USG.         0.36121    0.01724  20.953  < 2e-16 ***
TS.         23.21425    1.79561  12.928  < 2e-16 ***
Age          0.03880    0.02167   1.791   0.0742 .
FTr         -0.45058    0.87938  -0.512   0.6087
X3PAr       -5.92438    0.51932 -11.408  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.601 on 327 degrees of freedom
Multiple R-squared:  0.8975,    Adjusted R-squared:  0.8956
F-statistic: 477.1 on 6 and 327 DF,  p-value: < 2.2e-16
```

Figure C.1. R output for Full Model.

Model Equation:
PER = -6.50389 + 1.89752 WS + 0.361212 USG. + 23.31425 TS. + 0.03880 Age - 0.45058 FTr - 5.92438 X3PAr
Adjusted R^2: 0.8956

Hypothesis Testing:
$\alpha = 0.05$
Null Hypothesis Ho: $\beta_{WS} = \beta_{USG.} = \beta_{TS.} = \beta_{Age} = \beta_{FTr} = \beta_{X3PAr} = 0$
Alternative Hypothesis Ha: At least one of these parameters is not zero
F-Critical = F(0.05, 6, 327) = 0.2716
Reject Null Hypothesis if F-Statistics > F-Critical
Since F-Statistics = 477.1 < F-Critical, we reject the Null Hypothesis.
Therefore, at least one of the parameters is not zero.

T-Test for this full model suggests four variables to be significant at 95% level. They are WS USG., TS., and X3PAr. These four variables will be used in the Reduced Model below.

Reduced Model

```
> summary(reduced.lm)

Call:
lm(formula = PER ~ WS + USG. + TS. + X3PAr, data = stat)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9297 -1.1518 -0.0778  0.9210  6.5550

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.54312    1.04974   -5.28 2.35e-07 ***
WS           1.93103    0.11824   16.33  < 2e-16 ***
USG.         0.35779    0.01714   20.87  < 2e-16 ***
TS.         22.97922    1.75674   13.08  < 2e-16 ***
X3PAr       -5.65701    0.43626  -12.97  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.604 on 329 degrees of freedom
Multiple R-squared:  0.8964,    Adjusted R-squared:  0.8952
F-statistic: 711.8 on 4 and 329 DF,  p-value: < 2.2e-16
```

Figure C.2. R output for Reduced Model.

Model Equation:
PER = -5.54312 + 1.93103 WS + 0.35779 USG. + 22.97922 TS. - 5.65701 X3PAr
Adjusted R-squared = 0.8952

Hypothesis Testing:
$\alpha = 0.05$
Null Hypothesis Ho: $\beta_{Age} = \beta_{FTr} = 0$
Alternative Hypothesis Ha: At least one of these two parameters is not zero
F-Critical = F(0.05, 2, 327) = 3.023345
Reject Null Hypothesis if F-Statistics > F-Critical

```
> anova(reduced.lm, per.lm)
Analysis of Variance Table

Model 1: PER ~ WS + USG. + TS. + X3PAr
Model 2: PER ~ WS + USG. + TS. + Age + FTr + X3PAr
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    329 846.80
2    327 838.04  2    8.7586 1.7088 0.1827
```

Figure C.3. Anova table.

Since F-Statistics = 1.7088 < F-Critical, we fail to reject the Null Hypothesis.
Therefore, there is no evidence that β_{Age} and β_{FTr} are significant.

Stepwise Regression Model:

Hypothesis Testing:
$\alpha = 0.05$
Null Hypothesis Ho: $\beta_{FTr} = 0$

Alternative Hypothesis Ha: $\beta_{FTr} \neq 0$
F-Critical = F(0.05, 1, 327) = 3.87
Reject Null Hypothesis if F-Statistics > F-Critical

```
> anova(step.lm, per.lm)
Analysis of Variance Table

Model 1: PER ~ WS + USG. + TS. + X3PAr + Age
Model 2: PER ~ WS + USG. + TS. + Age + FTr + X3PAr
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    328 838.71
2    327 838.04  1   0.67283 0.2625 0.6087
```

Figure C.4. Anova table

Since F=0.2625 < F-critical = 3.87, we fail to reject the null hypothesis. We are 95% confident that the reduced model is appropriate.

Appendix D. R Code

```
## ----setup, include=FALSE-------------------------------------------------
-------------------------------------
knitr::opts_chunk$set(echo = TRUE)


## -------------------------------------------------------------------------
-------------------------------------
library(mltools)
library(data.table)
library(corrplot)
library(here)
library(dplyr)
library(ALSM)

main_df = read.csv(here("nba2021_advanced.csv"), sep=',')
#write.csv(main_df,"C:\\Users\\JS Home\\Documents\\MSDS\\MATH
6357\\M6357Project\\marketing_campaign_fixed.csv",sep=',',row.names = FALSE)


data = main_df[!(duplicated(main_df$Player)),] #delete entries for separate
teams
summary(data$MP)
summary(data$G)
boxplot(data$MP)

# filter data to players with >200 mins played
bymp = data[(data$MP>=200),]
bymp$Player

#obtain the stats and rename rows for analysis
stat = bymp[,c("PER","WS", "USG.","TS.", "Age", "FTr", "X3PAr")]
row.names(stat)=bymp$Player

#correlation matrix & correlation plots
round(cor(stat[]),3)
pairs(stat)
summary(stat$PER)
summary(stat)

corrplot(cor(stat[]))
model = lm(PER~., data=stat)

vif(model)
# since all VIF are <4, no need to adjust for multicollinearity

#scatterplots for all variables against PER
plot(stat[,2],stat[,1],ylab="PER",xlab="Win Shares",main="PER vs. Win
Shares")
plot(stat[,3],stat[,1],ylab="PER",xlab="Usage",main="PER vs. Usage")
plot(stat[,4],stat[,1],ylab="PER",xlab="True Shooting %",main="PER vs. TS%")
plot(stat[,5],stat[,1],ylab="PER",xlab="Age (Years)",main="PER vs. Age")
plot(stat[,6],stat[,1],ylab="PER",xlab="Free Throw Frequency",main="PER vs.
FT Frequency")
```

```
plot(stat[,7],stat[,1],ylab="PER",xlab="Three Point Frequency",main="PER vs.
Three Point Frequency")


## ------------------------------------------------------------------------
-----------------------------------
# library to show names on boxplots
library(car)
# full model
per.lm = lm(PER~., data=stat)
summary(per.lm)
anova(per.lm)

#show boxplots for each variable
Boxplot(stat$PER, main="PER",id=list(labels=rownames(stat),n=2, location =
"u"))
Boxplot(stat$WS, main="WS",id=list(labels=rownames(stat),n=1))
Boxplot(stat$USG., main="USG", id=list(labels=rownames(stat),n=2,
location="u"))
Boxplot(stat$TS.,main="TS",id=list(labels=rownames(stat),avoid=T))
boxplot(stat$Age,main="Age")
Boxplot(stat$FTr,main="FTr",id=list(labels=rownames(stat),n=2))
boxplot(stat$X3PAr,main="X3")

## ------------------------------------------------------------------------
-----------------------------------
# Full Model:
per.lm = lm(PER~., data=stat)
summary(per.lm)
## F-critical
qf(0.95, 6, 327)



# Reduced Model:
reduced.lm = lm(PER~WS+USG.+TS.+X3PAr, data=stat)
summary(reduced.lm)



anova(reduced.lm, per.lm)
# F-Critical
qf(0.95, 2, 327)



# Stepwise Regression:

start.model = lm(PER ~ 1, data = stat)

step.backward = step(per.lm, direction = "backward")
step.forward = step(start.model, direction = "forward", scope =
formula(per.lm))
step.both = step(per.lm, direction = "both")

step.backward
step.forward
```

```
step.both

step.lm = lm(PER~ WS + USG. + TS. + X3PAr + Age, data=stat)
summary(step.lm)
anova(step.lm, per.lm)
# F-Critical
qf(0.95, 1, 327)


## ------------------------------------------------------------------------
-------------------------------------


library(caret)
set.seed(12345)

#set up k-fold cross validation
train.control = trainControl(method="cv", number = 5)
# models with all, 5, and 4 predictors
modelall = train(PER~., data=stat, method="lm", trControl = train.control)
model4 = train(PER~WS+USG.+TS.+X3PAr, data=stat, method="lm", trControl =
train.control)
model5 = train(PER~WS+USG.+TS.+X3PAr+FTr, data=stat, method="lm", trControl =
train.control)
# show R-squared and RMSE for cross-validated models
print(modelall)
print(model4)
print(model5)

c(modelall$results$RMSE, model4$results$RMSE, model5$results$RMSE)

# anova for 4 predictors vs full model
model4pred = lm(PER~WS+USG.+TS.+X3PAr+FTr, data=stat)
anova(model4pred, per.lm)


## ------------------------------------------------------------------------
-------------------------------------
# redisplay cross validated 4 predictor model results
print(model4)

# rebuild model with 4 predictors
model = lm(PER~WS+USG.+TS.+X3PAr, data=stat)

#deleted studentized residuals
dtres = rstudent(model)

#t critical value
qt((0.05/(2*334)), (334-4-1))

plot(rstudent(model), type="o", xlab = "Index", main="Deleted Studentized
Residuals", ylab="Deleted t-residuals")
# t* = 3.83
abline(h=3.836, col="red")
abline(h=-3.836, col="red")

# Show players exceeding t-critical value
```

```
dtres[dtres>=3.836]
# Willy Hernangomez        CJ McCollum
#          3.943443           4.162823

# Cook's distance plot
cd = cooks.distance(model)
plot(cd, type="o", xlab = "Index", main = "Cook's Distance")
#text(cd, labels=rownames(stat), cex=0.9, font=2)
# all cook's distances are very small, <<1, so likely not influential

# calculate DFFITS
dffit = dffits(model)
cutoff = 2*sqrt(6/334)
# cutoff ~ 0.26726
plot(dffit, type="o", xlab="Index", main = "DFFITS")
# show players that exceed DFFITS
dffit[dffit>.267|dffit<(-.267)]
#text(dffit, labels=rownames(stat), cex=0.9, font=2)
abline(h=cutoff, col="red")
abline(h=-cutoff, col="red")
# lots of influential points


## ------------------------------------------------------------------------
-------------------------------------
# Fitted vs residuals plot
plot(predict(model, stat),residuals(model), main="Residual vs. Fits", xlab =
"Fitted Values", ylab = "Residuals")
# sequence plot
plot(1:334,residuals(model), main="Sequence Plot", xlab = "Index", ylab =
"Residuals",type="o")
abline(h=0, col="red")

# boxplot of residuals
boxplot(residuals(model))

# QQ-plot for normality
qqplot(predict(model, stat),residuals(model))
qqline(predict(model, stat),residuals(model))
```