Applying MLR to the NBA 23/24 Season

Stats 101A, Professor Xu

John Tan

1 Introduction & Data Preparation

Context The data set used for analysis in this paper includes various match statistics from the NBA 23-24 season. Each row represents a single game, along with various other important team match statistics. The descriptions of the variables are as follows:

Table 1: Dataset Variables and Descriptions

Variable Name	Description
Team	Team that played the game
Match Up	Indicating the match against the opposing team
Game Date	The date of the game played
WinLose	Whether team won or lost the game
MIN	Total minutes played by team
PTS	Total points scored by the team
FGA	Field Goals Attempted (total shots taken)
FGM	Field Goals Made (total shots made)
FG%	Field Goal Percentage (FGM / FGA * 100), measure of shooting efficiency
3PA	Three-Point Field Goals Attempted
3PM	Three-Point Field Goals Made
3P%	Three-Point Percentage (3PM / 3PA * 100), measuring accuracy from three-point range
FTM	Free Throws Made
FTA	Free Throws Attempted
FT%	Free Throw Percentage (FTM / FTA * 100), measuring accuracy for free-throw
OREB	Offensive Rebounds
DREB	Defensive Rebounds
REB	Total Rebounds (OREB + DREB)
AST	Assists - Number of passes leading directly to a made basket
STL	Steals - Times the team took the ball from the opponent
BLK	Blocks - Shots blocked by the team
TOV	Turnovers - Number of times the team lost possession
PF	Personal Fouls - Fouls committed by the team
ScoreDiff	Team's overall point differential for the game

Using this dataset, we aim to fit a multiple linear regression (MLR) model to investigate which factors have more significance in determining the number of points (`PTS`) a team scores during an NBA game. The initial model will include all variables as predictors, with the exception of `3PM`, `3PA`, `FGM`, `FGA`, `FTM`, `FTA`, and `REB` due to issues with perfect collinearity. Using model selection, we aim to determine a model with a sufficient balance in complexity. For obvious reasons, variables such as `Team`, `GameDate`, and `MatchUp` will be omitted from the initial model as well. For more information on the descriptive statistics of the data set, refer to the Appendix at the end of this paper.

2 Initial MLR Model and its Significance

Fitting a Multiple Linear Regression Model

Table 2: Initial Regression Analysis Result

Predictor	β_i	**SE**	**t-value**	**Pr(> |t|)**
(Intercept)	-94.50245	4.80400	-19.672	< 2e-16 ***
WinLose	0.50839	0.37961	1.339	0.18062
MIN	0.32940	0.01853	17.776	< 2e-16 ***
FGPercent	1.57813	0.03536	44.636	< 2e-16 ***
ThreePtPercent	0.31278	0.01739	17.988	< 2e-16 ***
FTPercent	0.19236	0.01131	17.009	< 2e-16 ***
OREB	0.93596	0.03372	27.760	< 2e-16 ***
DREB	0.28547	0.03005	9.500	< 2e-16 ***
AST	0.32472	0.02808	11.565	< 2e-16 ***
STL	0.37876	0.04580	8.270	< 2e-16 ***
BLK	0.06720	0.04473	1.502	0.13315
TOV	-0.85883	0.03280	-26.181	< 2e-16 ***
PF	0.42882	0.02796	15.339	< 2e-16 ***
ScoreDiff	-0.04516	0.01619	-2.790	0.00531 **
$R^2 = 0.8138$, Adjusted $R^2 = 0.8129$				
F-statistic = 822.3 on 13 and 2445 DF, P-value: < 2.2e-16				

The initial MLR model gives us an estimated regression equation of:

$$\begin{aligned} \hat{PTS} = &-94.5 + 0.508(\text{WinLose}) + 0.329(\text{MIN}) + 1.58(\text{FG\%}) + 0.313(\text{3P\%}) \\ &+ 0.192(\text{FT\%}) + 0.936(\text{OREB}) + 0.285(\text{DREB}) + 0.325(\text{AST}) + 0.379(\text{STL}) \\ &+ 0.0672(\text{BLK}) - 0.859(\text{TOV}) + 0.429(\text{PF}) - 0.0452(\text{ScoreDiff}) \end{aligned} \tag{1}$$

Each coefficient represents the expected change in points when there is an increase in the respective predictor by one unit (assuming all other variables are held constant). The output shows that `FG%` has the largest effects on points, resulting in a 1.578 increase in points per unit increase of `FG%`. This makes logical sense, since majority of the points scored during a match come from field goals. All our predictors show a positive effect on points, except for `TOV` and `ScoreDiff`, which show a decrease in points. An increase in turnovers mean the team loses possession more often, which would likely lead to less points scored. However, it is more unclear why `ScoreDiff` may lead to a decrease.

Our $R^2 = 0.8138$ suggests that the model explains approximately 81.4% of the variance in points scored, indicating our combination of predictors are relatively strong in explaining the variation of points. The adjusted R^2 is also similarly high, at 0.8129, suggesting strong explanatory power. The summary statistics also show t-test scores with p-values that are less than 0.01 for most variables - except `WinLose`, `BLK`, and `ScoreDiff` - suggesting statistical significance. The variable `ScoreDiff` has a p-value that is less than 0.05, however both `WinLose` and `BLK` have p-values larger than 0.05 which indicates possible statistical insignificance and may require removal. The test of overall significance (F-statistic) has a small p-value that is less than 0.05, suggesting the linear regression model is statistically significant.

However, we still need to perform model diagnostics to test the assumptions of MLR and check if they hold. In addition, we want to avoid any issues of multicollinearity within our model as well as test various combinations of our variables selected to ensure the final model has strong explanatory power.

3 MLR Diagnostics and Assumptions

Visually Checking Residual Plots

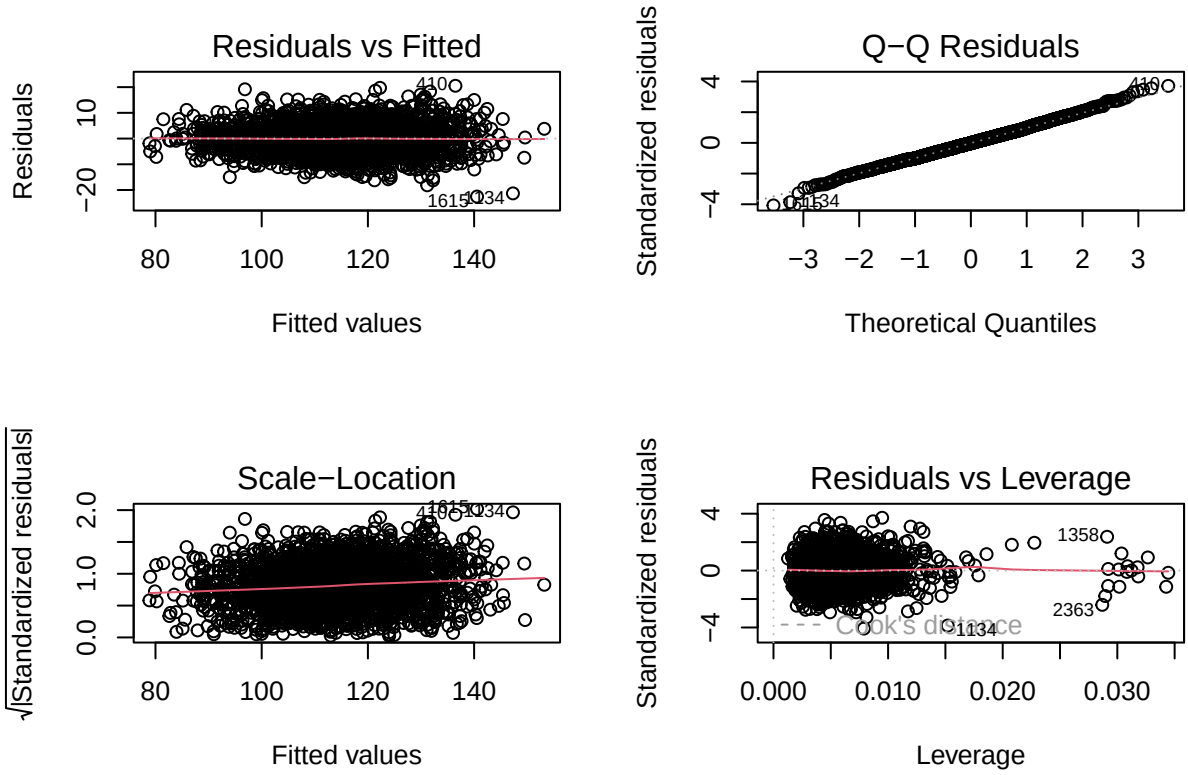

Figure 1: Diagnostic Plots

Residuals vs. Fitted: The residuals appear to be randomly scattered around zero (as indicated by the horizontal red line) with no discernible pattern. This suggests that the linearity assumption holds and there is no clear correlation among the residuals (indicating independence of errors).

Normal Q-Q Plot: Based on an initial visual analysis the standardized residuals follow closely to the straight dotted quantile line, with very minor deviations closer to the tail ends. This suggests that the residuals follow a normal distribution, and that our inference and hypothesis tests remain valid for this model.

Standardized Residuals vs. Fitted: The standardized residuals also appear to be randomly scattered around the horizontal line, and the width of the residuals stay generally invariant to the fitted values (standardized residuals do not "fan out" or "fan in" like a cone), suggesting homoscedasticity in the model and that the variance of residuals are constant across all levels of X.

Residuals vs. Leverage: At first glance we see various points with standardized residuals that are either greater than 4 or less than -4, which may indicate the presence of outliers. Most of the points have a small leverage, however there are a few points with high leverage which may indicate the presence of influential points.

Examining Potential Outliers and Influential Points

Outliers

Table 3: Bonferroni Test Results

Largest Residual Point	Studentized Residual	Benferroni P-value
1615	-4.08912	0.10992
No Studentized Residuals with Bonferroni P-value < 0.05		

We can use a Bonferroni Test as a formal test to identify possible outliers. The test focuses on the most extreme residuals in the dataset, while adjusting the p-value for multiple comparisons. The output states that no studentized residual reaches the adjusted threshold for significance (no residuals have a Bonferroni-adjusted p-value < 0.05), suggesting no outliers in the dataset. Therefore, even for the point with the largest residual (point 1615) it is not extreme enough to conclude it is an outlier.

Leverage Plots

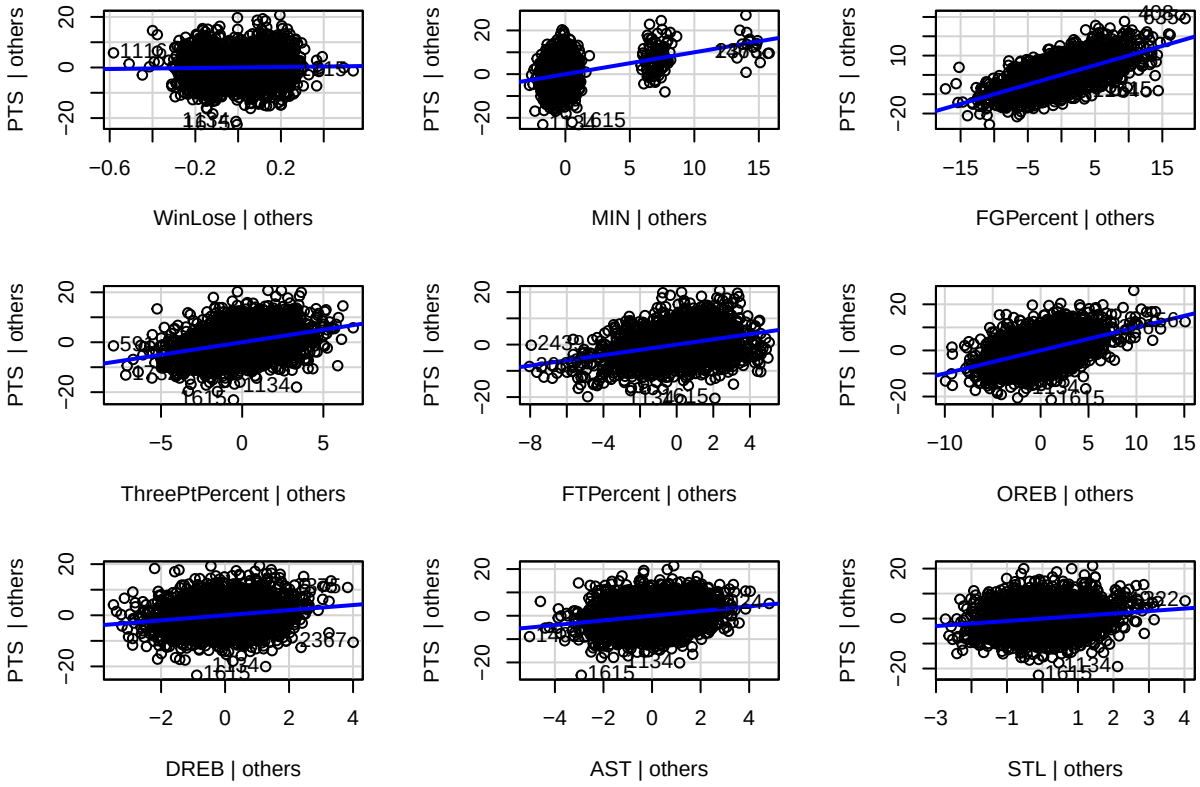

Figure 2: Leverage Plots for all Predictor Variables

The leverage plots show the residual response variable (y-axis) against a specific predictor variable. Through this plot we can observe how far an observation's combination of predictor values is for the rest of the dataset. For values of high leverage, we would see these points distinctly separate from the bulk of the data (horizontally far to the left or right of the rest of the data). For most of the variables, there are no obvious high-leverage points that stand out from the rest of the data. However, we can see three distinct clusters of points in `MIN`, which is likely a result of how overtime works in NBA. We will see if removal is required during model selection.

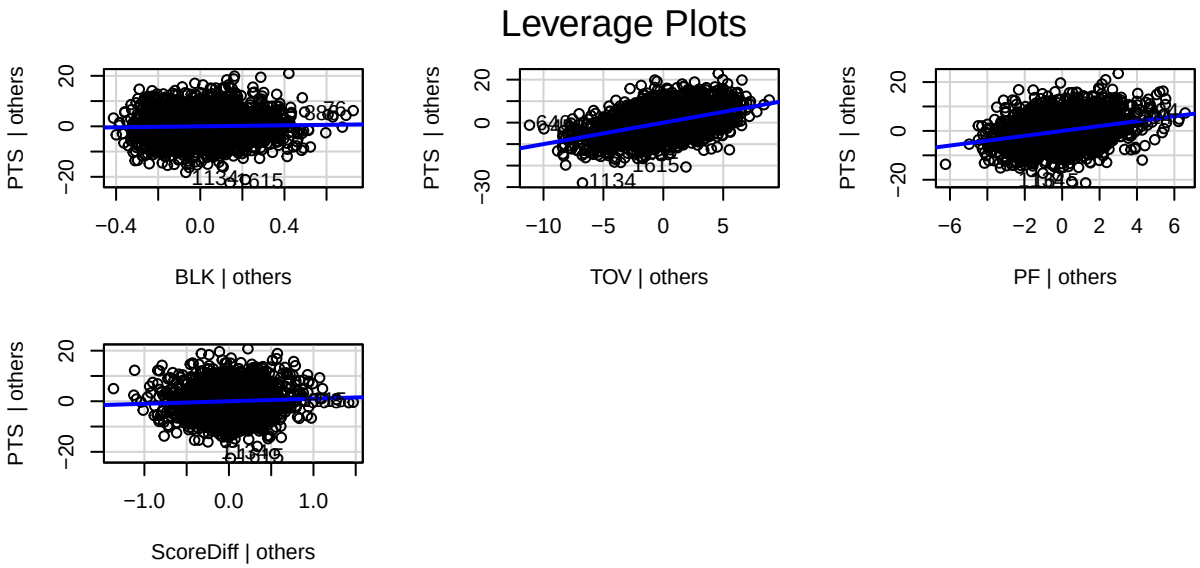

Figure 3: Leverage Plots for all Predictor Variables

Cook's Distance

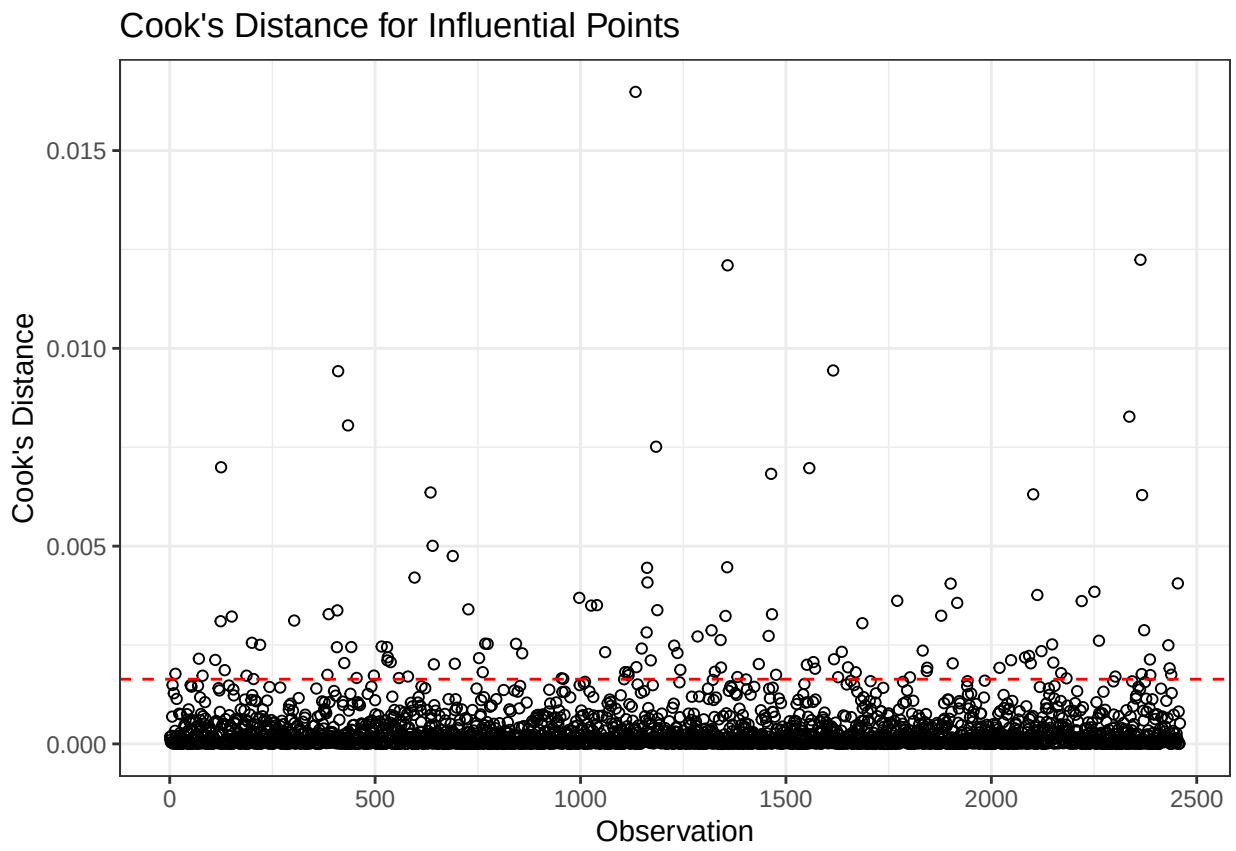

Figure 4: Cook's Distance for Observations

This plot shows the Cook's Distance for the various observations in our dataset. Points above the dotted red line are above the threshold for what is considered an influential point. We can see that most of the dataset lie below the red line, with only a few points having a Cook's Distance far above the threshold. However, the overall distribution is low enough that no single point appears to drastically influence our overall model.

Testing the Assumptions of Multiple Linear Regression

Linearity of the Data

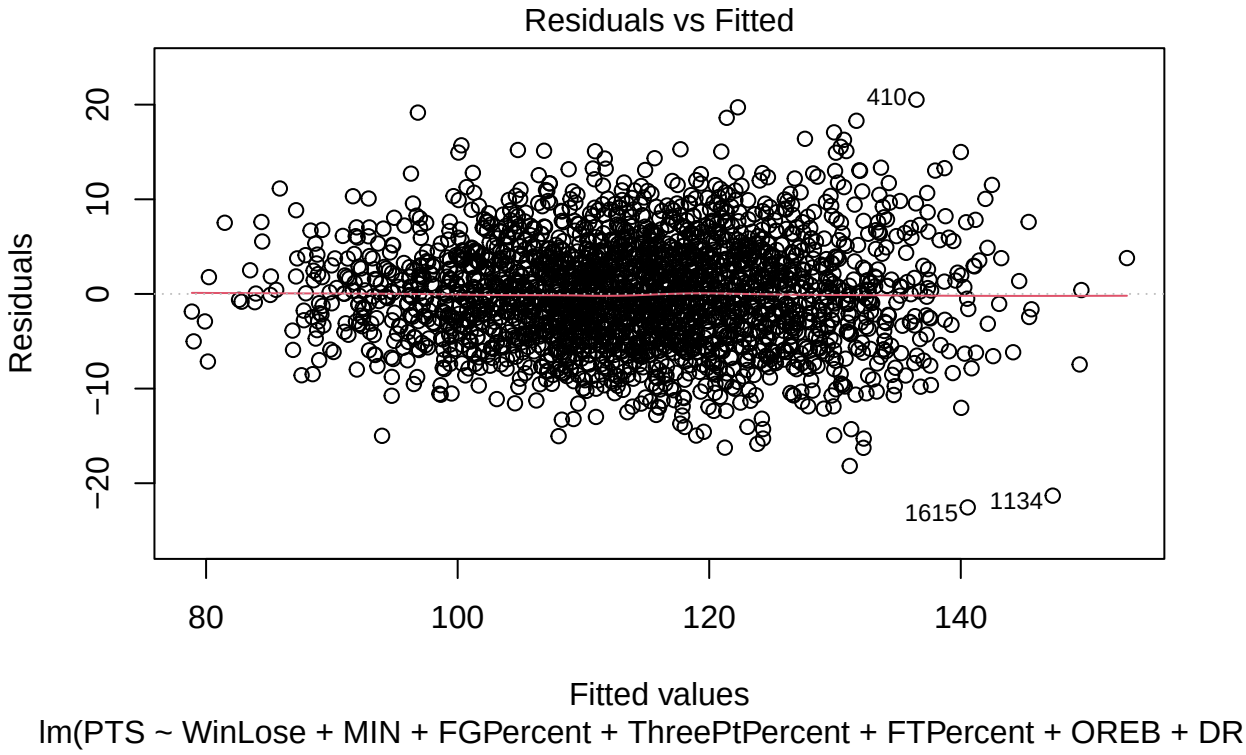

Figure 5: Residuals against Fitted Values

As stated in our initial visual analysis, the residuals are scattered randomly around 0 with a lack of a discernable pattern. This suggests that linearity holds for the model, with our specified response variable and predictors.

Normality of Errors

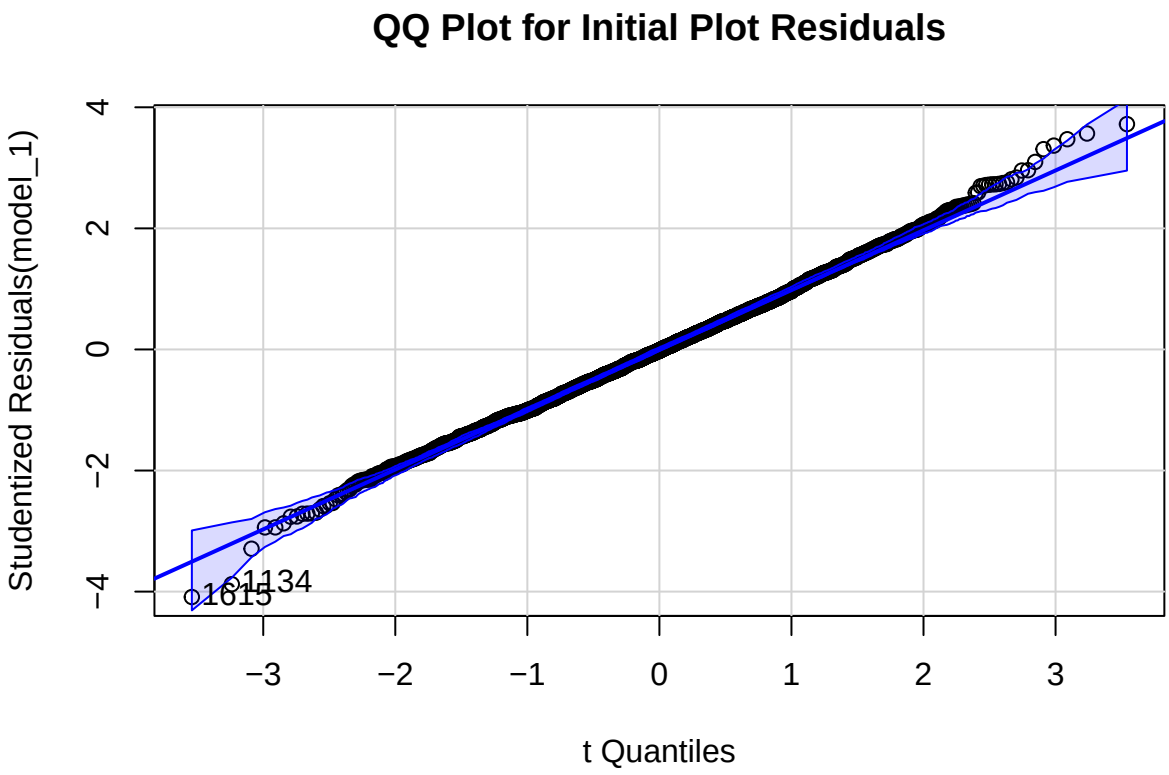

Figure 6: QQ Plot for Initial Plot Residuals

```
## [1] 1134 1615
```

In the QQ plot, majority of the points appear to follow the linear blue line, with only slight deviations in the tail ends. This suggests that our studentized residuals follow a normal distribution, which we can also visualize using a histogram.

Aside from minor deviations, we can see that the overall histogram follows a normal distribution, suggesting that our assumption of normality of errors holds.

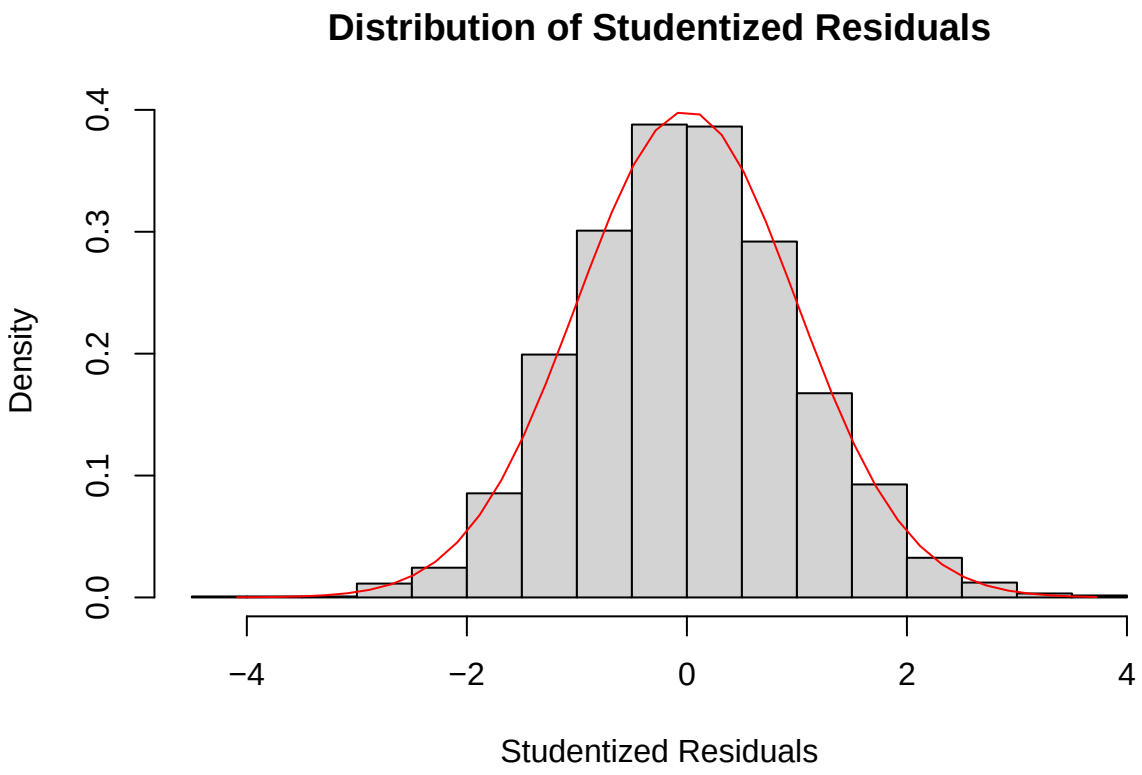

Figure 7: Distribution of Studentized Residuals

Homoscedasticity (Constant Variance)

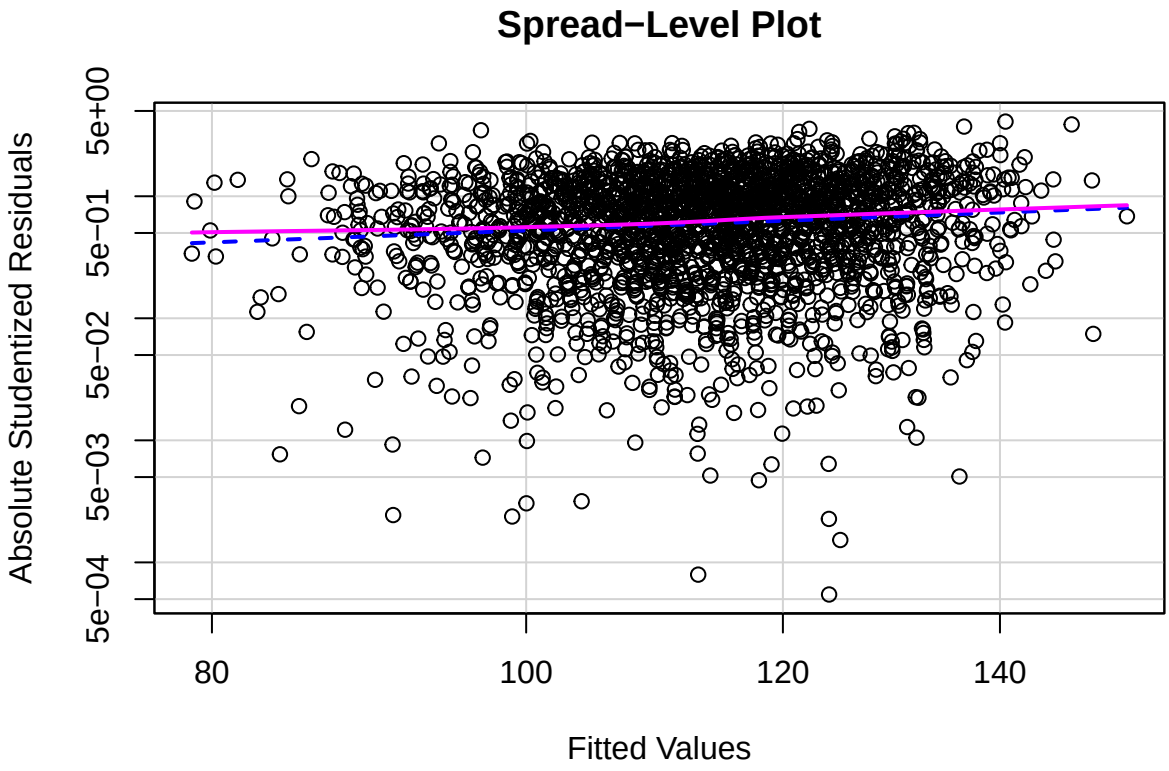

Figure 8: Spread-Level Plot to check for Constant Variance

```
##
## Suggested power transformation:  -0.006436933
```

Observing the Spread-Level Plot, our studentized residuals appear to form a slope that is relatively flat. However, it does seem to have an uneven width, which may suggest heteroscedasticity.

Table 4: Non-constant Varaince Score Test Results (BP Test)

Chisquare Statistic	**Df**	**P-value**
63.74801	1	1.414e-15
P-value < 0.05, suggests heteroscedasticity present in model		

We confirm this using the Breusch-Pagan test. The significant p-value (that is less than 0.05) indicates that there is heteroscedasticity present in the model. We thus will have to correct this, to ensure that our residuals have constant variance.

Independence of Errors

Table 5: Durbin-Watson Test Results

Lag	D-W Statistic	P-value
1	1.944005	0.188
P-value > 0.05, suggests no autocorrelation		

The Durbin-Watson test, which detects autocorrelation in residuals, give a p-value that is greater than 0.05, which suggests that our errors have independence. This means knowing the residual will not help in predicting the residual from another point.

Multicollinearity

The output for the Variance Inflation Factor (VIF) suggests that the variable `ScoreDiff`, potentially has multicollinearity with other predictors, due to the VIF value being greater than 5.

Table 6: VIF for Predictor Variables

Predictor	Variance Inflation Factor
WinLose	2.871282
MIN	1.103977
FGPercent	3.009054
ThreePtPercent	1.676404
FTPercent	1.050858
OREB	1.320049
DREB	2.104619
AST	1.634964
STL	1.330313
BLK	1.076318
TOV	1.245858
PF	1.069189
ScoreDiff	5.202396

To verify this we can check the Added Variable Plots. Most of the added-variable plots show a positive or negative linear relationship with the response variable, which suggests that those variables have a direct effect on response variable. However, the AVP for `WinLose` and `BLK` has a flat slope, which suggests they may be correlated with other predictor variables in the model. Using stepwise selection will allow us to remove the multicollinearity in the model.

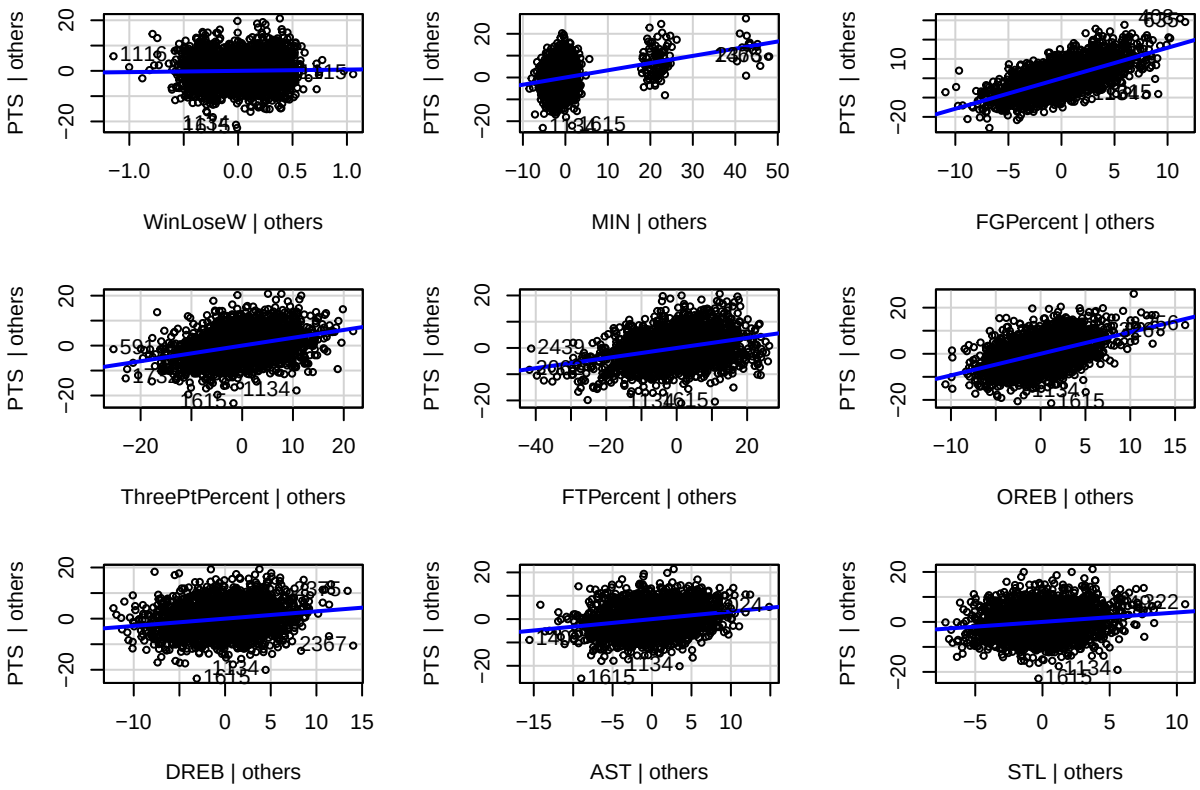

Figure 9: Added Variable Plots for Predictors

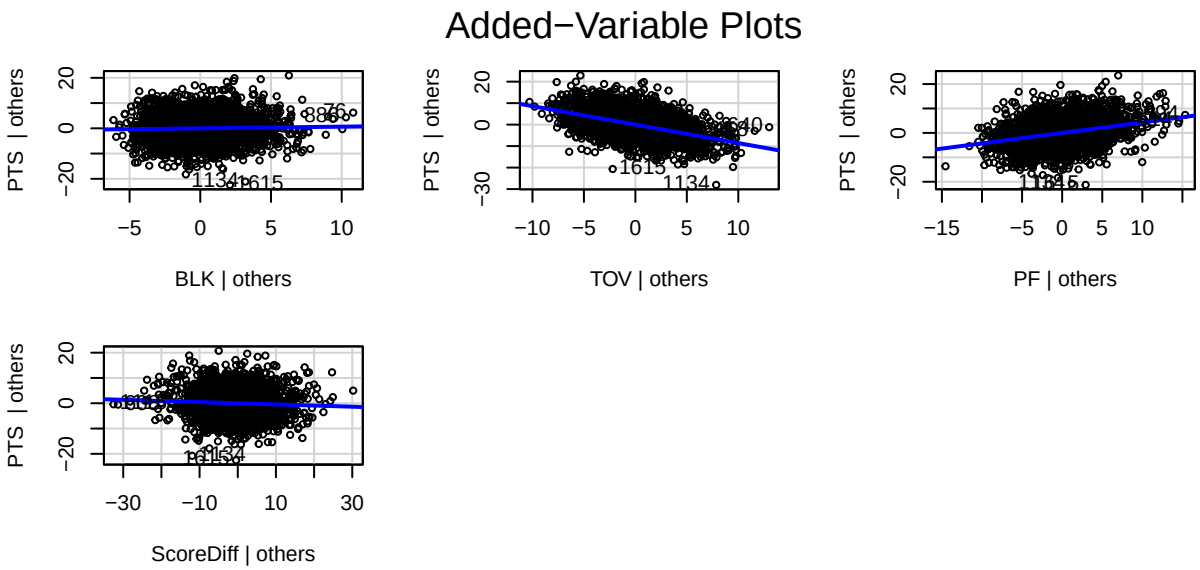

Figure 10: Added Variable Plots for Predictors

4 Model Selection

Firstly, before using a variable selection method, we will attempt using a log transformation on the response variable to correct for the heteroscedasticity in the model.

Table 7: Transformed Regression Analysis Result

Predictor	β_i	**SE**	**t-value**	**Pr(> |t|)**
(Intercept)	2.907e+00	4.224e-02	68.821	< 2e-16 ***
WinLose	4.349e-03	3.338e-03	1.303	0.19265
MIN	2.817e-03	1.629e-04	17.289	< 2e-16 ***
FGPercent	1.407e-02	3.109e-04	45.273	< 2e-16 ***
ThreePtPercent	2.762e-03	1.529e-04	18.066	< 2e-16 ***
FTPercent	1.691e-03	9.944e-05	17.002	< 2e-16 ***
OREB	8.340e-03	2.964e-04	28.133	< 2e-16 ***
DREB	2.533e-03	2.642e-04	9.587	< 2e-16 ***
AST	2.749e-03	2.469e-04	11.134	< 2e-16 ***
STL	3.395e-03	4.027e-04	8.431	< 2e-16 ***
BLK	6.502e-04	3.933e-04	1.653	0.09840
TOV	-7.684e-03	2.884e-04	-26.641	< 2e-16 ***
PF	3.823e-03	2.458e-04	15.554	< 2e-16 ***
ScoreDiff	-3.971e-04	1.423e-04	-2.790	0.00531 **
$R^2 = 0.8158$, Adjusted $R^2 = 0.8148$				
F-statistic: 832.9 on 13 and 2445 DF, P-value: < 2.2e-16				

The transformed MLR model is in the form:

$$\begin{aligned}\log(\hat{P}_{TS}) = \beta_0 &+ \beta_1(\text{WinLose}) + \beta_2(\text{MIN}) + \beta_3(\text{FG\%}) + \beta_4(\text{3P\%}) \\ &+ \beta_5(\text{FT\%}) + \beta_6(\text{OREB}) + \beta_7(\text{DREB}) + \beta_8(\text{AST}) + \beta_9(\text{STL}) \\ &+ \beta_{10}(\text{BLK}) + \beta_{11}(\text{TOV}) + \beta_{12}(\text{PF}) + \beta_{13}(\text{ScoreDiff}) \end{aligned} \tag{3}$$

Table 8: Non-constant Varaince Score Test Results (BP Test) for Transformed Regression

Chisquare Statistic	**Df**	**P-value**
0.08732345	1	0.76761
P-value > 0.05, suggests no heteroscedasticity present in model		

When we conduct the Breusch-Pagan test, the p-value is now greater than 0.05 and thus we fail to reject the null of homoscedasticity. We have successfully removed the heteroscedasticity from the model. We can now proceed with the stepwise selection of our independent variables.

Stepwise Selection

Due to the large number of predictor variables in the initial model, we will use stepwise selection since it is less computationally expensive than performing a best subset selection.

We perform a backward stepwise selection, to compare the AIC and BIC of the different models when removing or adding each variable to an initial model. Based on the output of the ANOVA, using the stepwise selection with AIC results in the removal of the `WinLose` variable, which we had previously suspected of causing multicollinearity in the model. The other variables are kept in the model. However, the decrease in AIC is minimal, decreasing the AIC from -14834.7 to -14835.0.

We have also conducted the same stepwise procedure, but instead using BIC as the information criteria, which punishes additional variables added to the model more than AIC.

Table 9: Backward Stepwise Selection

Steps	**Variable Removed (AIC)**	**AIC**	**Variable Removed (BIC)**	**BIC**
(1)	-	-14834.7	-	-14753.39
(2)	WinLose	-14835.0	WinLose	-14759.49
(3)	-	-	BLK	-14764.21
(4)	-	-	ScoreDiff	-14766.74

When stepwise selection is performed with BIC, the suggested final model removes `WinLose`, `BLK`, and `ScoreDiff` to reduce the BIC from -14753.39 to -14766.74. This stepwise selection also removes the other variable (`BLK`) that we suspected caused multicollinearity, based on the AVP. Since the initial model started with a large number of variables, we will use the suggested model determined using BIC to more heavily punish added variables. BIC also removed the two variables we suspected of causing multicollinearity. Thus we have our final model:

Table 10: Final Regression Analysis Result

Predictor	β_i	**SE**	**t-value**	**Pr(> |t|)**
(Intercept)	2.938e+00	4.008e-02	73.294	< 2e-16 ***
MIN	2.885e-03	1.616e-04	17.855	< 2e-16 ***
FGPercent	1.369e-02	2.581e-04	53.057	< 2e-16 ***
ThreePtPercent	2.645e-03	1.455e-04	18.172	< 2e-16 ***
FTPercent	1.645e-03	9.739e-05	16.886	< 2e-16 ***
OREB	8.091e-03	2.789e-04	29.012	< 2e-16 ***
DREB	2.154e-03	1.929e-04	11.169	< 2e-16 ***
AST	2.753e-03	2.465e-04	11.169	< 2e-16 ***
STL	2.948e-03	3.573e-04	8.251	2.53e-16 ***
TOV	-7.386e-03	2.659e-04	-27.775	< 2e-16 ***
PF	3.743e-03	2.425e-04	15.435	< 2e-16 ***
$R^2 = 0.815$, Adjusted $R^2 = 0.8143$				
F-statistic: 1079 on 10 and 2448 DF, P-value: < 2.2e-16				

5 Interpretation and Discussion

Thus, our final model is given by:

$$\begin{aligned}\log(\hat{PTS}) = &-2.94 + 0.00289(\text{MIN}) + 0.0137(\text{FG\%}) + 0.00265(\text{3P\%}) \\ &+ 0.00165(\text{FT\%}) + 0.00809(\text{OREB}) + 0.00215(\text{DREB}) \\ &+ 0.00275(\text{AST}) + 0.00295(\text{STL}) - 0.00739(\text{TOV}) + 0.00374(\text{PF}) \end{aligned} \tag{2}$$

Due to the transformation, our coefficients are less intuitively interpretable. The coefficients represent the increase or decrease in the log points scored by an NBA team on average, when there is a 1 unit change in a given predictor (assuming all other predictors are held constant). Based on the model, the `FG%` still has the largest impact on the points scored by a team. All the other variables cause an increase in the number of points except for the `TOV` (number of times the team lost possession), which makes intuitive sense. An interesting observation is that the `PF` (fouls committed by the team) actually increase the points scored by a team, and actually have a greater impact than some other variables that one would expect, such as steals (`STL`) or assists (`AST`).

The methodology used to arrive at the final model has several limitations. The use of backwards stepwise selection prevents us from ensuring the best model; some combinations of predictor variables are not tested during this selection method. Moreover, using AIC instead of BIC during the stepwise selection would have resulted in a different final model. It was a judgement call that led to the BIC stepwise model being chosen. This model also does not take into account team or individual player specific factors, as some teams may play better against other specific teams or have better strategies. Moreover, since the data only covers one specific NBA season (23/24), this model may not necessarily be optimal in predicting the scores of future seasons and we must be careful not to extend this model to predict the number of points scored in any basketball game. We cannot necessarily make this generalization.

Despite these limitations however, the overall approach is sound, as all the assumptions of MLR are tested. The resulting model shows that all predictors are statistically significant in determining the number of points scored by a team in the NBA.

6 Appendix

Table 11: Descriptive Statistics Table of NBA 23/24 Data

Variable Name	**N**	**Mean**	**Median**	**Min**	**Max**	**SD**	**Skew**
MIN	2460	241.4	240	240	290	6.351	0.1119
PTS	2460	114.2	114	73	157	12.85	5.050
FG%	2460	47.52	47.5	27.7	67.1	5.498	0.06087
3P%	2460	36.49	36.55	6.9	64.5	8.341	0.07457
FT%	2459	78.33	78.9	33	100	10.15	-0.4311
OREB	2460	10.55	10	0	28	3.817	0.5381
DREB	2460	32.99	33	16	55	5.409	0.1752
AST	2460	26.67	27	11	50	5.101	0.2667
STL	2460	7.474	7	0	20	2.822	0.3619
BLK	2460	5.142	5	0	17	2.598	0.5720
TOV	2460	13.6	14	3	29	3.812	0.2404
PF	2460	18.73	19	4	34	4.149	0.2778
ScoreDiff	2460	0	0	-62	62	15.79	0