

Project Part 2: NBA Player Performance

Predictive Modeling & Testing

Austin Lacey, Nick Orecchia, David Woo

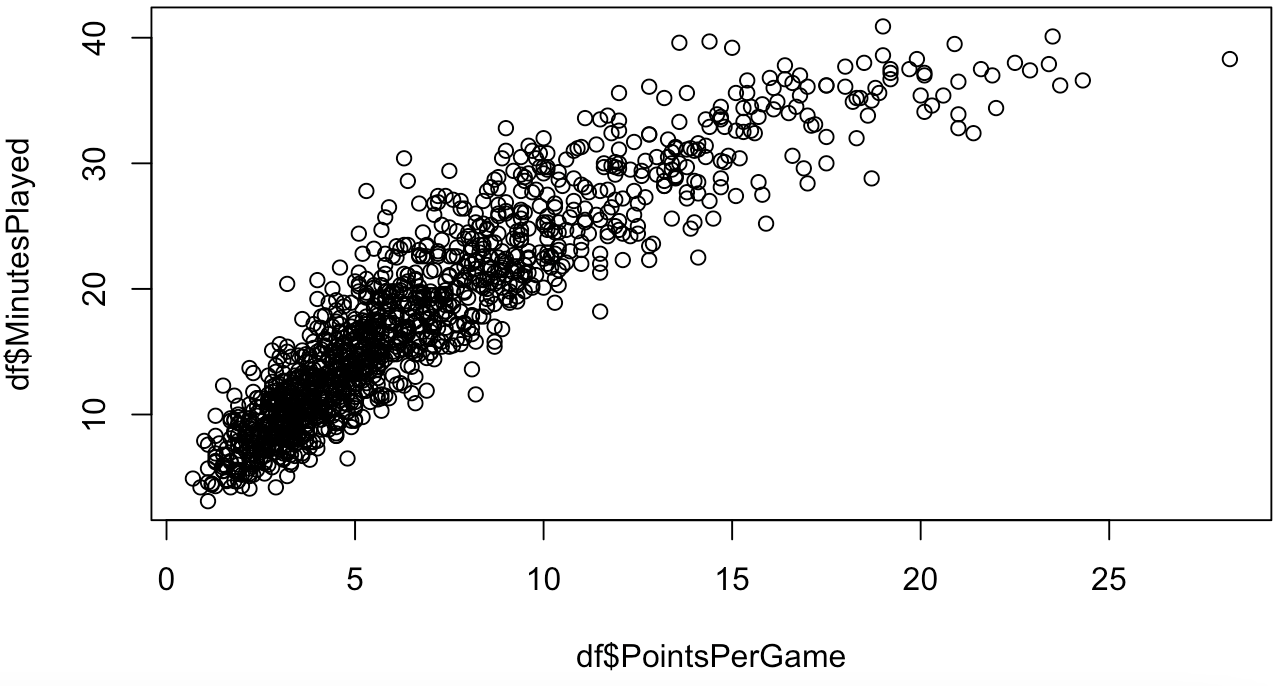
**Executive Summary**

The data that we used for both the first and second part of this project covers the average NBA statistics for a select number of players. We extracted our dataset from Kaggle (<https://www.kaggle.com/sachinsharma1123/performance-prediction>), with the hope of exploring the relationship between various statistics and player performance. The raw dataset included 1,340 unique values (which represented the individual number of players) and a total of 20 different statistics / variables per player. The data was structured in a cross-sectional format, as it consisted of various different players across all different era’s with no time factor included.

While going through the preprocessing portion of the first project, we were able to clean our dataset entirely and as a result have a TIDY dataset. Once we had accomplished this step in the analytics pipeline, we were able to run various descriptive analyses to gain a better understanding of the data that we had collected. We decided to run basic scatter plots using the plot function, plot(df$MinutesPlayed, df$PointsPerGame), and we did this to get a visual understanding of specific correlations or possible relationships that may be present between the variables that we have. After running these specific visualizations, and conducting our preliminary analysis in part one we were able to come up with specific variables of interest. The most important variable, and one that seemed to be related with almost all of the other variables was *Minutes Played*. The relationships that we were able to find through our tests further supported our initial hypothesis that this would be a strong variable of interest. We came to this hypothesis because no matter the player, no matter the style of play or position, the minutes that they play in a game is something that is comparable across all players and era’s. Our additional variables of interest; *Points per Game, Steals, Turnovers, Rebounds* resulted from the positive correlations we were able to see with the *Minutes Played* variable, our primary variable of interest. With this preliminary analysis completed, we were now tasked with looking to create linear regression models to help quantify the relationships that we have discovered through the first part of our analysis. To do this we had to transform specific variables of interest, separate the data into 70% training (which will be used to estimate the models) and 30% testing (which will then be used to assess the models created), and partition the data so that it was reproducible and constant for all of the models. After taking these steps, and building the different models, we can be able to compare them to find out which model best predicts our *Minutes Played* variable.

**Bivariate Regression**

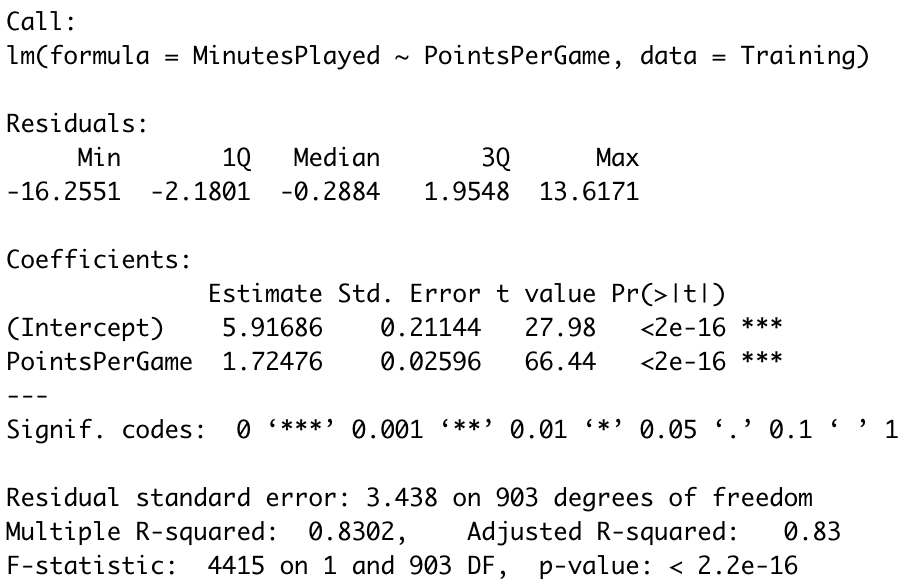
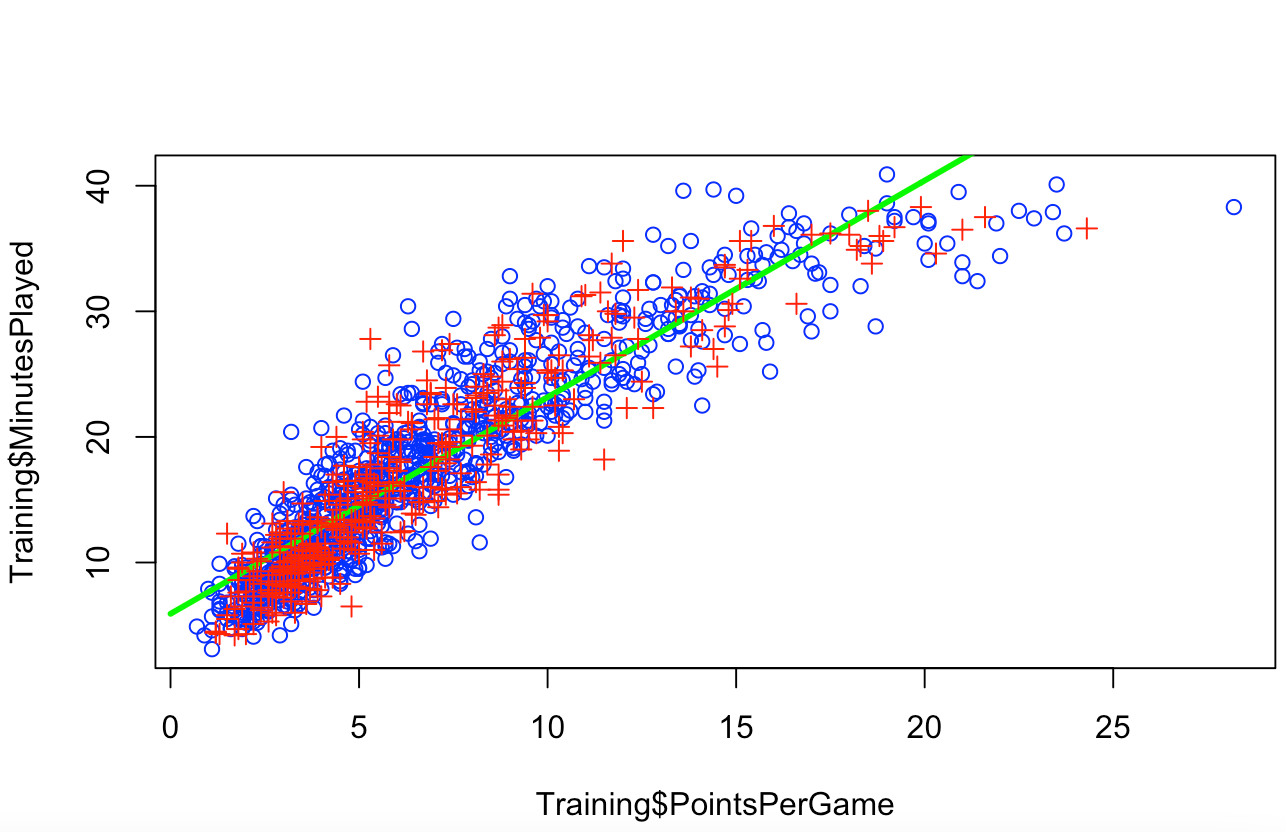
As stated in the executive summary, the output variable that we are looking to predict is *Minutes Played*. We came to this decision based on the strong relationships we found with other variables in our dataset but also because it is a variable that is comparable across all of the NBA players. The single independent variable that we chose to predict our output for our bivariate regression models was *Points per Game*. As the scatter plot below shows, there is a fairly strong positive correlation between the two variables, and because this relationship was the strongest with the *Minutes Played* we thought it would be the best fit to predict our output. While creating our models, we needed to transform the independent variable as well, and we transformed the *Points per Game* three separate times. We created a quadratic transformation, a cubic transformation and a 4th order transformation of the *Points per Game* variable. With these transformations, and our independent and dependent variables set, it was now up to building and assessing the models using the training and testing partitioned data.

****

Model One

For our first model we chose to create a linear regression equation by having solely the independent variable, no transformations. The regression equation for this model is; *Minutes Played* = 5.92 + 1.72*PointsperGame*. We chose to do this for our first model to create a baseline linear regression equation that we can gain insights on and then build on for our future models. After creating this linear model, we were able to gather the output that is provided in the image below, which allows us to look at the residuals, coefficients, significance levels, residual standard error and r squared values. With the adjusted r square value, we can see how well the model is fitting the training data and interpret that 83% of the variance found in the *Minutes Played* variable can be explained by *Points per Game.* While this variance will not be the highest out of all of the models we create, it is interesting to see that with only this variable, and without it being transformed at all, it can still explain 83% of variation in our dependent variable. This further explains the relationship we are able to see while running our scatter plot but also shows just how much points per game has an effect on the minutes players play in the game. Expanding on this insight, we can look at the specific effects the coefficients have on the output. As the visual below shows, for every additional point scored per game the minutes played in the game increases by 1.72. This allows us to conclude that the more points you score in a game, the more minutes a player will play which is something we would expect to see. However, we are now able to quantify this conclusion and approximate with the testing data the effects that additional points scored can have on time spent in the game. In terms of significance, we can see that the overall model, the intercept and *Points per Game* coefficient are statistically significant at the highest confidence level.

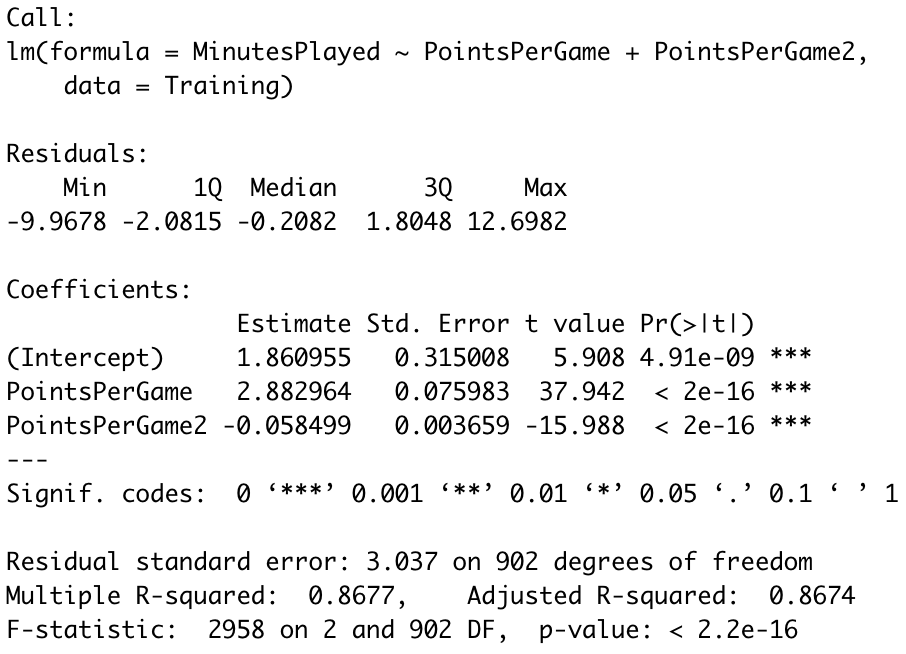
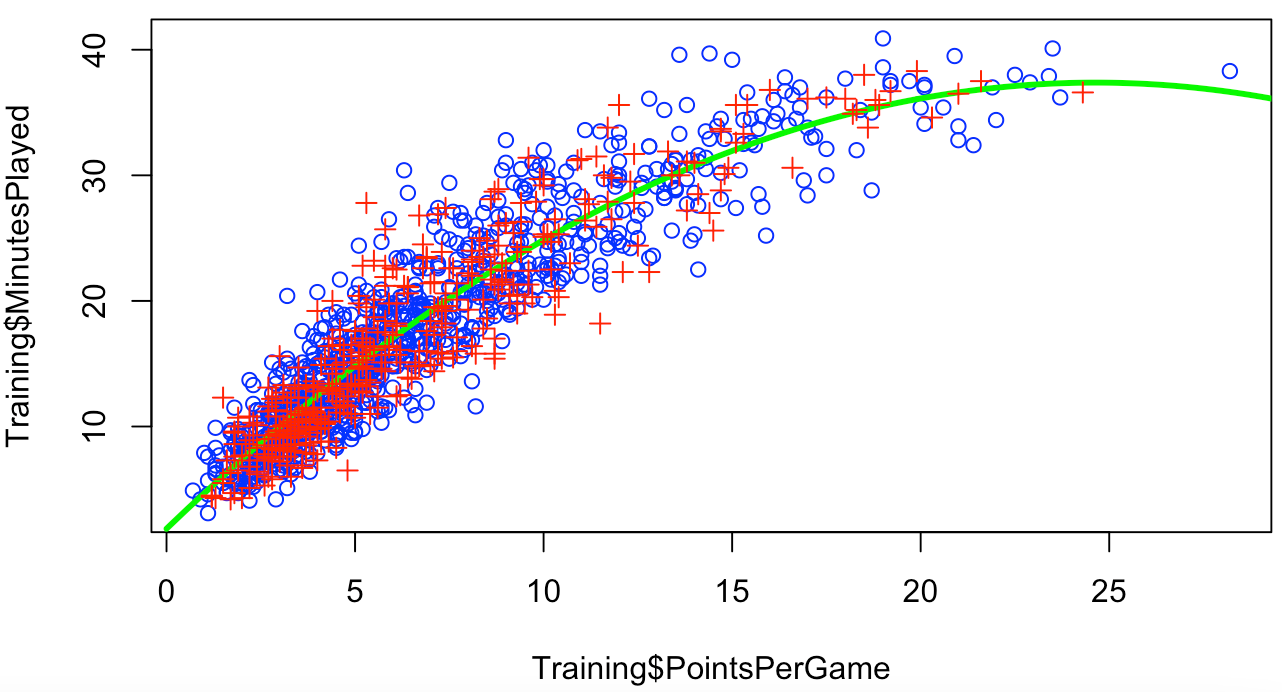
While assessing the model we looked to the root mean squared error for in-sample and out-of-sample data points to gain further insight. Since this is our first model we will use these two values to compare how well the model fits the in-sample data and how accurate it can predict the out-of-sample data with our next five models. As the output of R code below shows, the in-sample error for model one is 3.434382 while the out-of-sample error is 3.409132.





Model Two

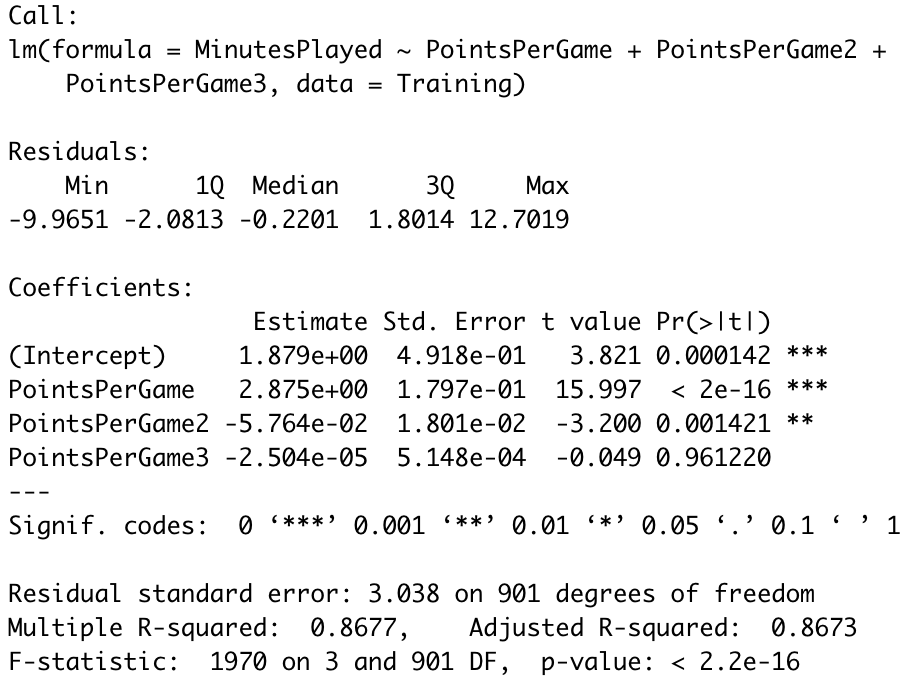
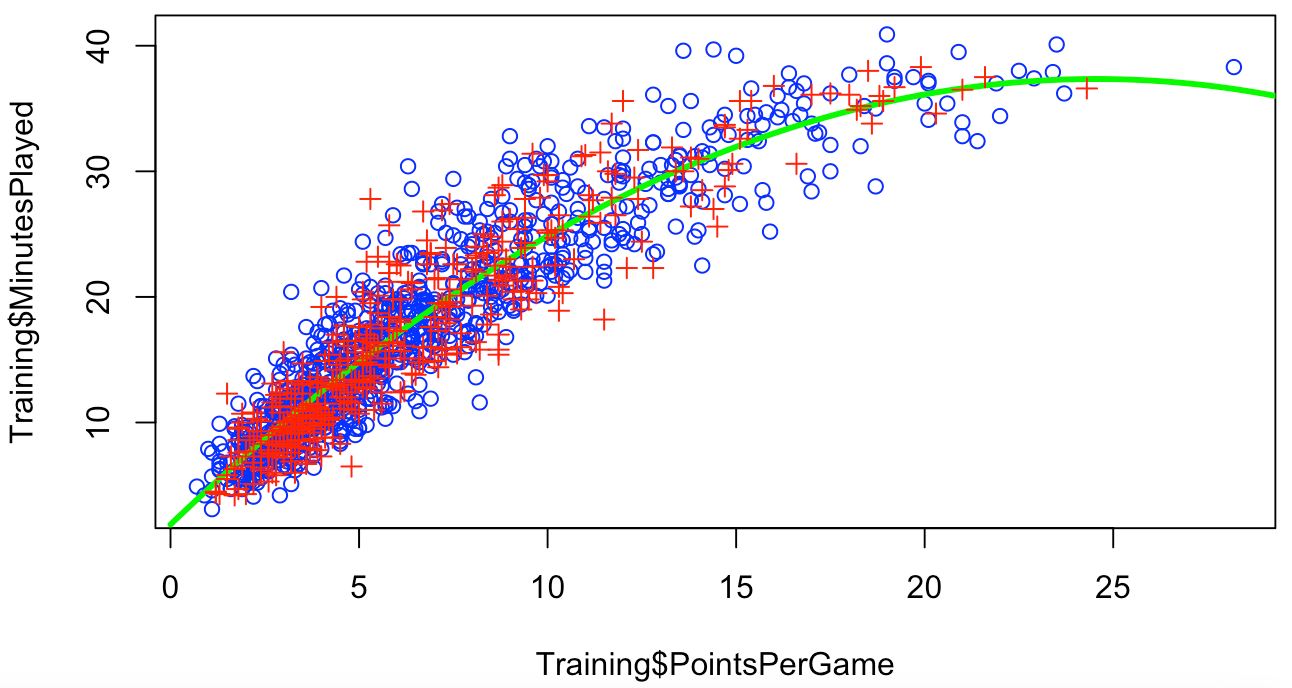
For our second model, we chose to transform our independent variable by raising it to the second power. This will allow our line of best fit within the regression model to better fit the data by giving the line some type of curvature. The green line which represents this line of best fit is no longer linear, but quadratic. However, the model entirely still remains linear because there aren’t any nonlinear transformations to the parameters. It can be expressed by the following equation; *Minutes Played* = 1.86 + 2.88*PointsperGame* – 0.06*PointsperGame*^2. Now, 86.74% of the variance found in *Minutes Played* can be explained by *Points per Game* so this model explains an additional 3% of variance found in the data meaning that it is a better fitting model than model 1*.* The two key factors to look at in the output of the model are the root mean squared error for in-sample and out-of-sample data points. As seen in the output of our R code, this model has a lower RMSE for in-sample and out-of-sample than the previous model. This means that this model fits the in-sample data better than the previous model, and also can more accurately predict the out-of-sample data than the previous model as well.





Model Three

To conclude with our third model, we choose to take our independent variable to the cubic root. This allowed us to continue to explore the relationship between these variables. As our statistical output continued to increase in accuracy, this was a necessary step. We hypothesized that the higher power we took our variable to the more accurate it would be. But as we will see below this is not the case. As stated above we have not included any non-linear transformations to this model. The equation representing the data below is *Minutes Played* = 1.88 + 2.875*PointsperGame* – 0.056*PointsperGame*^2 – 0.000025*PointsperGame*^3. At this point we start observing inconsistencies with our data. Below you can see that *PointsperGame*^3 loses its significance, moving well out of the 5% range that we desire. The adjusted r-squared dropped very slightly by .01% to 86.73%, which for the added complexity, it is in no way significantly improving this model up to this point. Next, we will analyze the RMSE output. When using the memorized data to test for the in-sample error it increased its accuracy by 4.0 x 10^-6 to 3.031579. The out-of-sample error also improved slightly by 3.0 x 10^-6 to 3.076112. This model proves to be more accurate in learning the trained data and also predicting the out-of-sample data. Finally, we added a fourth order to our equations to see if this would continue to increase the accuracy of the RMSE. The added complexity ended up negatively impacting our model. This proves that our polynomial equation is most accurate when *PointsperGame* has been taken to its cubic root.



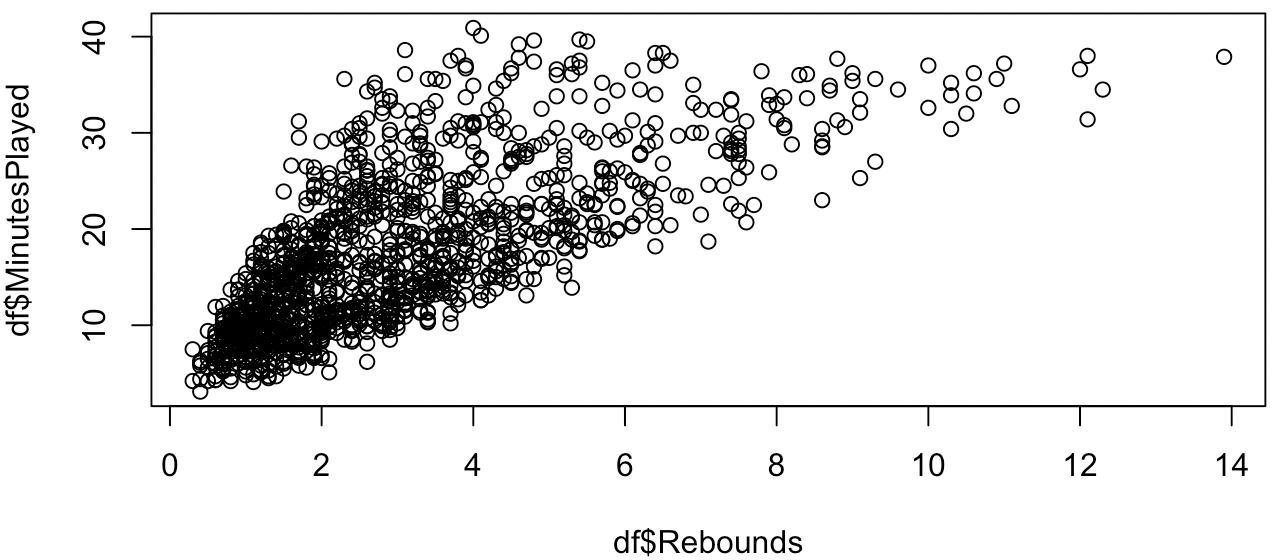
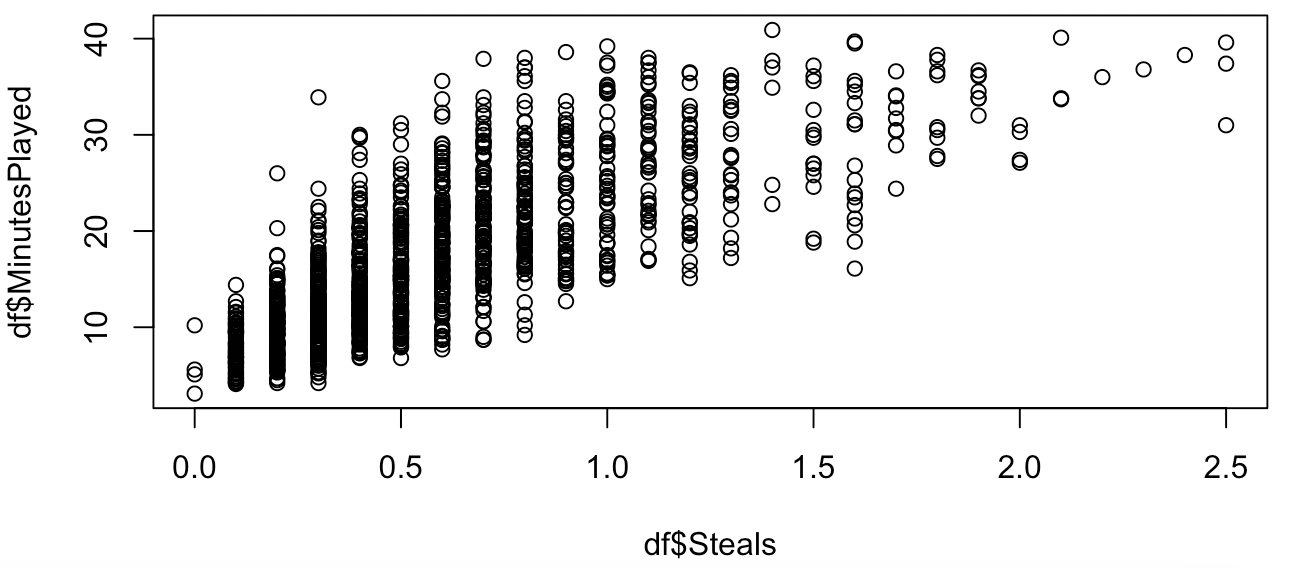




**Multivariate Regressions**

Throughout this project we have found that there is a strong correlation with several of the variables and *Minutes Played*. For this reason, we will continue to focus on *Minutes Played* as our output variable. Originally, we started with 20 unique variables which gave us plenty of options to combine and find the most relevant model to use out-of-sample information for the benchmarking of our predictive analysis. We each were trying to create the right combination of variables to see who could make the most accurate model. This took us down many different avenues that created many nice models with fair accuracy. The more that we tweaked the more that we understood there must be a more direct route.

First, we decided to go over the analysis that we had completed above in the bivariate regression models. After all the transformations that were performed, we concluded that the highest degree of accuracy was taking the x-variable *Points per Game* to the cubic transformation (3rd order polynomial). So, we will focus on the quadratic and polynomial transformation completing this section. Next, was the need to find the best variables to use with these multivariate models. When originally starting this process, we were focusing on logical combinations that seemed to make sense for us. Examples are, we started out with a strategy of including all offensive statistics. This was done to see if offensive production was a driving factor to see if you stayed off the bench. This proved to be not the most fruitful information. We also did the same thing with defense but received similar results. After multiple unsuccessful attempts we decided to leverage the data that we produced from the first part of this project. For this we turned our focus to our scatterplots (displayed below) and correlations. The final 3 variables that after testing that made the cut were *Points Per Game*, *Steals*, and *Rebounds*. Producing dramatically more accurate models.

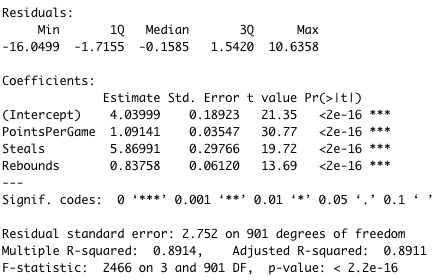


Model 5

For our first multivariate regression model and fifth total model we chose to create a linear regression equation by having three independent variables (*Points per Game*, *Steals*, *Rebounds*) but did not include any transformations. The regression equation for this model is; *Minutes Played* = 4.04 + 1.09*PointsperGame* + 5.87*Steals* + 0.84*Rebounds*. The reason we decided to not include any transformations on our first multivariate regression model was because we wanted to see how the additional variables we included in the model interacted with the *Minutes Played* variable and the overall model. With the adjusted r square value, we can see how well the model fits the training data. We interpreted that 89.1% of the variance found in the *Minutes Played* variable can be explained byour three independent variables*.* This adjusted r square value is the highest out of the three models we have already made, which shows us that with the addition of these variables the variation in the output variable is better explained. While looking at the coefficients we can see that *Steals* has the highest effect on the *Minutes Played* variable. With everything else constant, for every additional steal the *Minutes Played* increases by 5.87. This is not entirely what we would have expected, because initially we thought *Points per Game* would have had the biggest effect on increasing minutes played. However, with everything else constant for every additional point per game the minutes a player plays only increases by 1.09. This was in line with our expectations, and in line with what we saw in our previous three models, but it is still interesting to see that every additional steal has a larger impact on the minutes played for this specific model. In terms of rebounds, for every additional rebound, with everything else held constant, minutes played increases by 0.837. Even with the addition of the new variables, all of them are statistically significant at the highest confidence level.

Next, we looked to the root mean squared error for in-sample and out of sample data points. As the output of R code below shows, the in-sample error for model one is 2.746093 while the out-of-sample error is 2.738222. In this model, both the in-sample and out-of-sample data better fit and more accurately predict the output. This shows us that these specific independent variables are not only better able to fit the training data that was used to model the data but we are better able to predict the output compared to the previous three models we have taken. This makes sense to us because by adding more statistical variables to the equation, and accounting for more styles of play the model should be better able to predict the *Minutes Played* for all styles of NBA players.





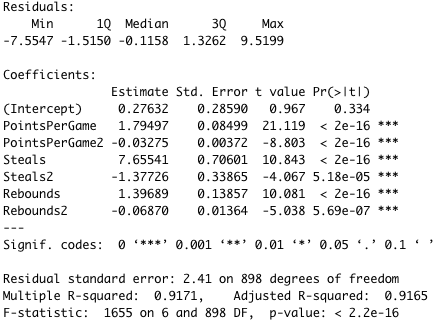


Model 6

For our second multivariate regression model, we took the same variables in model 5 but transformed them by raising them to the second power. The regression equation is represented by the following equation; *Minutes Played* = 0.28 + 1.79*PointsPerGame* – 0.03*PointsPerGame*^2 + 7.66*Steals* – 1.38*Steals*^2 + 1.40*Rebounds* – 0.07*Rebounds*^2. The adjusted r-squared value came out to 91.65%, meaning this model with the squared variables explains 91.65% of variance within the *Minutes Played* variable. This is the highest r-squared value out of all of our previously created models, meaning this model fits the data the best. Again, looking at the coefficients within the regression equation, it still seems that *Steals* has the largest impact on the *Minutes Played* variable.

After looking at the root mean squared error for the in-sample and out-of-sample data, we see a fairly large reduction in the values of both from model 5. The in-sample RMSE is 2.400526 and the out-of-sample RMSE is 2.465767. This means that this model with the additional transformation of the variables is able to more accurately model the in-sample data and more accurately predict the out-of-sample data points. This was to be expected, because in the Bivariate Regression portion of this project, we found that squaring the variable from model 1 to model 2 allowed for a better fit for both in-sample and out-of-sample data. Shown below is the output of model 6, the variables included, and the RMSE output.





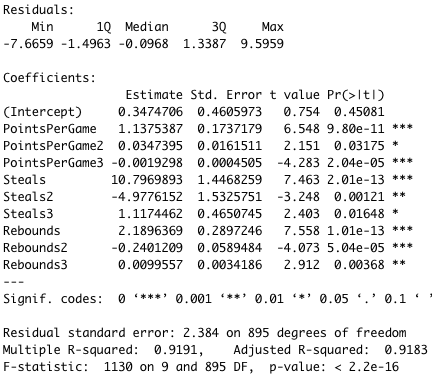


Model 7

For our final regression model, we reproduced Model 6 and then raised them to the cubic root. With the improvements shown in Model 6 this step was necessary to see if the trend would continue. The equation representing the data below is *Minutes Played* = 0.34 + 1.14*PointsPerGame* + 0.03*PointsPerGame*^2 – 0.002*PointsPerGame*^3 + 10.8*Steals* – 4.98*Steals*^2 + 1.12*Steals*^3 + 2.19*Rebounds* – 0.24*Rebounds*^2 + 0.001*Rebounds*^3. There is a very high statistical significance of all the variables in our two previous models, and with this we can see that with each root increase it is diminishing the impact of that variable. For example, *Rebounds*^2 goes from 0.24 to 0.001 for *Rebounds*^3. Showing that we are adding more information but not necessarily adding more value. *Steals* still has the most significant impact followed by *Rebounds* than *PointsPerGame*. Another interesting piece of data to note is that *PointPerGame^2* which was at the highest confidence level and is now in the 1% range which is quite the jump from < 2e – 16. Next, we will observe the results from our adjusted R-squared. It came in at an impressive 91.83%. Showing a high level of accuracy explaining the output variable due to our use of variables and their transformations.

Lastly, we will analyze how the use of our training data will help us predict the accuracy of our testing data. We see that there is improvement in the RMSE for our in-sample error. It has been reduced by 0.02987 to 2.370386. Although this is a plus it did reflect in the out-of-sample error. This has increased by 0.00423 bringing this to 2.469997. Once again, the added complexity has decreased the predictive accuracy of our model. Proving that Model 6 is superior when compared to this model and the previous simpler one.







**Conclusion**

After creating seven different models using various independent variables and several transformations of these variables, we found that model six had the lowest out-of-sample error. Although model seven had a lower in-sample error than model six, one of the problems with adding complexity to a model is that it may increase the out-of-sample error which is exactly what happened in this case. This is why model six is the best out of our seven models, because it has the lowest out-of-sample error. For forecasting purposes, a model’s ability to accurately predict out-of-sample data points is the most important characteristic of that model. If a player’s points per game, steals, and rebound statistics are available, this model can help predict the expected amount of time they will spend on the floor each game. After completing both projects, we were successfully able to identify the key variables that were statistically correlated through the use of visualizations, and then used these variables to create a model that can successfully predict (on average) how much playing time per game an athlete can expect to play.