NFL 2020 Weeks 8-10:

Best Models to Forecast Spread, Total Points, and Result of Games

Authors: Abby Chen, Aman Depani,
Ellis Fitzhugh, Hayden Dewey, Kush Patel

I. Data Information

a. Data Cleaning and Joining

Given that team aggregate season statistics were on different datasets, separated by season and side (offense or defensive), we wanted to combine all of this so we had one dataset that had all team data for all years. The data had extra rows at the top indicating which category each particular set of variables fell under (passing, rushing, penalties) along with rows at the bottom for averages and totals across the league, which we got rid of to make it easy to concatenate all the rows across the years. Once this was done, we added a year variable in order to be able to join the offense and defense statistics together, finally obtaining one large dataset, each row representing a team's aggregate statistics on a particular season. Two missing values were found, both of which were in 2020; upon research, the two values were for the offensive turnovers for Green Bay and defensive turnovers for Houson, both of which were zero during the first four weeks of the season, so these missing values were replaced with zero.

The next step was joining the other datasets with information about the stadium, teams, and game results to then be joined with the team's aggregate statistics created earlier. Cleaning had to be done in the results dataset to ensure that the stadium names there matched the ones listed on the stadium dataset. Minor discrepancies such as "Mercedes-Benz Stadium" versus "Mercedes-Benz Superdome", and stadium name changes like Jacksonville's EverBank Field to TIAA Bank Field had to all be accounted for to ensure all the data could be joined properly. Some 2019 were played in the new Tottenham Hotspur Stadium, which was not listed in the stadium dataset, so research about the stadium had to be done to add this row manually. From there, the teams dataset, containing team ID information useful to account for any organization name changes that occurred over the past 20 years, as the ID stayed the same, for example, for the St. Louis Rams and Los Angeles Rams, which helps to compare data between seasons. In order for an observation to contain the necessary team information for both the home and away team, the team data had to be joined twice.

The final step was to join the aggregate team statistics from earlier to the newly joined game results data. We decided it would make the most sense to join aggregate statistics from the season before to join with a given season's game results. First of all, there is not enough from the 2020 season to make valuable predictions on that season's game results, when only three games had been played. In addition, when running models, we have all of the data for the other seasons to use, meaning if we were to join by the same year as the individual games, the models would be representing how a team's overall season performance reflected on all the games they played that season. However, we will never know the 2020 overall season performance; we would be essentially projecting the games they played so far onto future games, which is not consistent with what the models truly represent, even if we normalized statistics by the number of games played. The data was then joined twice by the home team ID and the away team ID, and we finally ended up with a starting dataset containing 5047 observations and 140 columns.

b. Engineered Variables

One variable that we created that was not in the original dataset is the proportion of 1st downs to the total number of offensive plays. We believed that this would be a good measure of the offensive skills of a team since more first downs means that that team was able to move the ball down the field more. Dividing the number of first downs by the total number of plays would standardize the measurement and allow a fair comparison between different teams. This proportion was made for both home and away offenses and both measurements were statistically significant predictors of their respective scores when modeled in a linear regression.

c. Outside Data

From the Harvard Business Review article, "Who's the Most Important Member of an NFL Franchise," we drew inspiration by how large the quarterback's impact on the team is, especially now that it has been growing in recent years (Groysberg et. al). Therefore, we decided to import season-level passing data from Pro Football Reference (an example dataset can be found at https://www.pro-football-reference.com/years/2019/passing.htm) to join quarterback statistics to see which statistics impacted the results the most. In this case, it made sense to join on the current year since who the starting quarterback is for a team stays more consistent within a season than across seasons.

The data provided by Pro Football Reference indicates in the position column which quarterback was the primary starter for his respective team, and therefore it was simple to filter the 32 necessary quarterbacks during each season for our analysis. The assumption here is that the starting quarterback gives an overall representation of the team's performance, but there are discrepancies here due to the fact that it is likely that during a season, a team may have to start different quarterbacks in some games due to injury or lack of performance. Therefore, the quarterback data we use to regress for previous team results are not completely matched. In the future, it would be beneficial to aggregate and weigh the quarterback performance by how many games each quarterback within a team started. Due to simplicity, we used whoever was indicated as a starter and we took quarterbacks starting from the 2012 season, as it became more apparent as went back more seasons that there were more and more discrepancies with who could have been indicated as the starting quarterback.

Once we had this data, we had to decide which variables from the data would be the most interesting and likely to have an impact on the game results. We got rid of any statistics that were not normalized (total completions, total yards, etc.), and chose typical variables like quarterback rating, completion percentage, touchdown rate, interception rate, and yards per gain. Some interesting ones that we added that we wished to explore were Adjusted Net Yards per Pass Attempt, defined by the following equation:

$$Adjusted\ Net\ Yards\ Per\ Attempt = \frac{Passing\ Yards - Sack\ Yards + 20*(Passing\ TD) - 45*(Interceptions)}{Pass\ Attempts + Times\ Sacked}$$

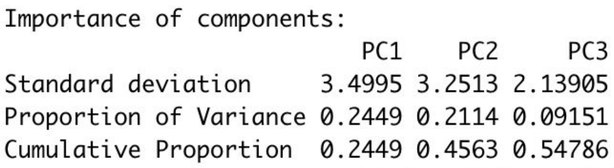
Other interesting variables added were comebacks led by the quarterback and game-winning-drives led by the quarterback. The ultimate goal was to join the past 2012-2019 quarterback data with our dataset created before by home and away team ID's (which had to be checked and cleaned in the quarterback data as well). From this, we could see which metrics had the most impact on the results, and we could potentially use significant predictors within our models to predict 2020 games. We keep this data set with the joined quarterback data separate from our original to be able to include 2000-2011 seasons in our analysis.

II. Methodology for Spread

The very first thing we tried was using a linear regression model on all the team aggregate statistic variables, home and away. This was a 104 variable model that ended up having 14 statistically significant predictors, with an adjusted R-squared of .08. Some of the noteworthy predictors (that would come up in later models as well) included home offensive points, home and away offensive passing first downs, and home defensive rushing touchdowns allowed. From this however, there are clearly issues in the fact that we use 104 variables; we want our final model to be parsimonious, meaning containing less variables while having higher predictive power.

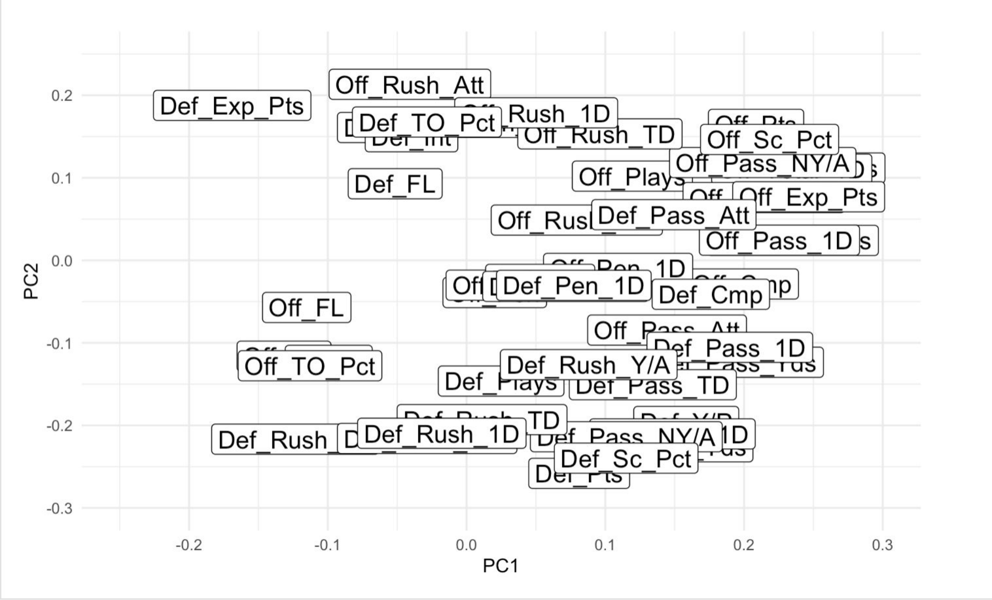
The next step to take was to perform backward selection to be able to pick out a few very significant variables, and using regsubsets in R, we ended up with a 5 variable model that included a couple of the previous significant predictors like home offensive points and home defensive rushing touchdowns allowed. However, the R-squared dropped to .064. This could mean that the existing collinearity amongst the variables led to a lack of additional predictive value, despite having a p-value of less than .001.

One method to reduce collinearity amongst variables, as well selecting variables that lead to the most variation amongst observations, is principal component analysis. This was done on all the team aggregate statistics, which originally had 52 total variables (before combining home and away). Once we ran the algorithm, we ended up with the following importance of components:



```
Importance of components:
                          PC1    PC2     PC3
Standard deviation     3.4995 3.2513 2.13905
Proportion of Variance 0.2449 0.2114 0.09151
Cumulative Proportion  0.2449 0.4563 0.54786
```

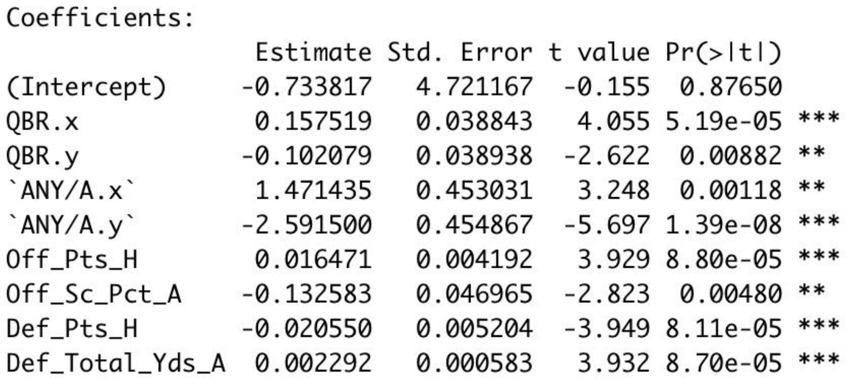
From this, we can see that the first and second principal components gave a cumulative proportion of about half of the variance the data, so we used these two components to graph the variable weights on each of these components, and the ones with the most weight (whichever was located along the edges) would be the ones to include in our reduced dataset.



We chose variables that had a PC1 weight greater than .2 or variables that had a PC2 weight of less than -.2 since those seemed to be the furthest away from the graph's origin. We also added Def_Exp_Pts and Off_Rush_Att. This reduced the amount of variables from 52 to 20, which we would then rejoin with the rest of the data to perform another round of regression models. After performing backward selection with these variables, we ended up with a model containing home offensive points, home defensive points allowed, offensive scoring drive percentage, and defensive total yards allowed, with an adjusted R-squared of .063. All variables were significant given a .001 alpha.

We next wanted to see if there were any predictors from the quarterback data. Using the 2012-2019 quarterback metrics, we ran another linear regression and the most significant predictors were QBR for home and away, and the Adjusted Net Yards per Attempts (ANY/A) for home and away, and overall the model gave an R-squared of .178. Seeing how significant QBR was compared to the rest of the variables inspired us to try something interesting - how does the QBR of the home team interact with the QBR of the away team with respect to spread? Unfortunately, when running this model with the interaction variable, the p-value was .36, and therefore we could not use this. When testing the interaction variable for ANY/A for home and away, we get a p-value of .819.

We had four significant variables from the quarterback data and four variables from the previous regression without quarterback data, so we now have 8 significant variable candidates for our model to predict spread. Running this model, we get the following results:



```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -0.733817   4.721167  -0.155  0.87650
QBR.x           0.157519   0.038843   4.055 5.19e-05 ***
QBR.y          -0.102079   0.038938  -2.622  0.00882 **
`ANY/A.x`       1.471435   0.453031   3.248  0.00118 **
`ANY/A.y`      -2.591500   0.454867  -5.697 1.39e-08 ***
Off_Pts_H       0.016471   0.004192   3.929 8.80e-05 ***
Off_Sc_Pct_A   -0.132583   0.046965  -2.823  0.00480 **
Def_Pts_H      -0.020550   0.005204  -3.949 8.11e-05 ***
Def_Total_Yds_A  0.002292   0.000583   3.932 8.70e-05 ***
```

The p-values here are all significant at the .01 level, and our adjusted R-squared increased from our model without quarterback data to .2. We decided that this would be the best model to go with due to the fact that it is parsimonious, only using 8 variables, and adjusted R-squared here was the best of any model tested.

We tried using other models that implemented K-clustering to cluster the teams by their aggregate statistics, using the 20 variables obtained from principal component analysis. This way, we would only have to represent each team by a cluster they belonged to, and see if those clusters had any predictive value on spread. Two of the five clusters had significant value for both home and away, but this would only mean that our model would only be useful for teams within those clusters. When trying to test the interaction of the clusters together, there was only significance between the interaction of two clusters. Because of this, and the fact the adjusted R-squared was only .03, we chose the 8 variable model pictured above as the final model for spread.

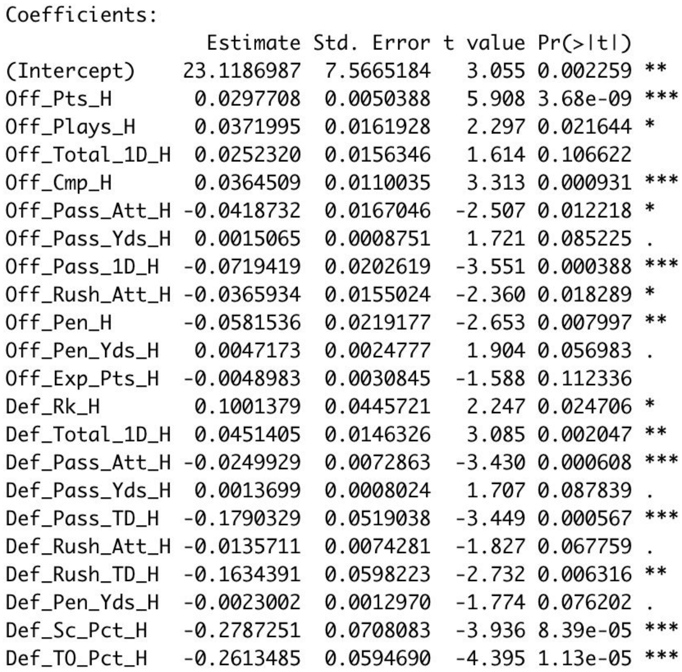
For the cross-validation process, the data was split into training and test sets at 85% and 15%, respectively. We then applied the model to both and examined the residuals and calculated the RMSE. The training set had an RMSE of 12.91 and the test set had an RMSE of 13.55. These values were very similar, indicating that the model performed fairly consistently across both sets, with the test predictions having more error, marginally.

To predict the values for the Week 8-10 games, we created an R function to take in the home team ID and the away team ID as parameters, and using those, obtain the necessary values from the 2019 team aggregate data and the starting quarterback data for this season (we made sure to use the data from the quarterback's that are most likely to start, i.e. Andy Dalton instead of the injured Dak Prescott). We ran these values with the final prediction model, and obtained the spread values (after rounding to the nearest whole number). Any zero values that we got were assigned 1 or -1 by looking at which side of 0 the actual spread value lied, along with consideration of who was injured. For example, the San Francisco 49ers no longer have their best defensive players due to injury, so in a tie setting between them and Green Bay, we chose Green Bay to win by 1 point.

III. Methodology for Total

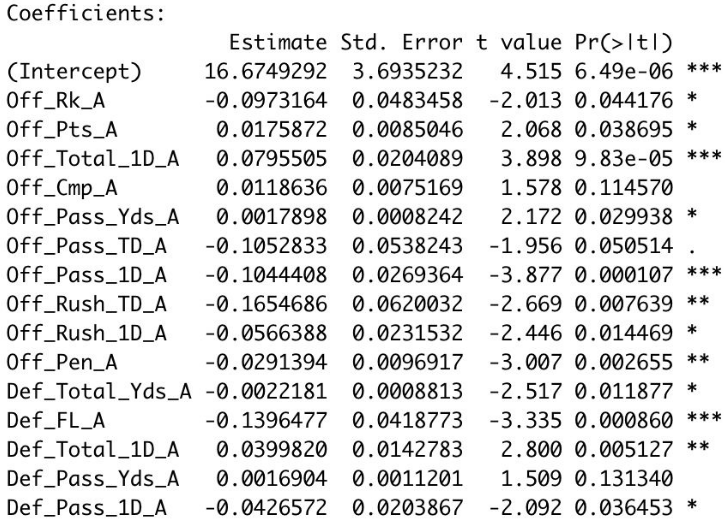
The first thing we tried for the total measurement was to use the 20 variables previously obtained from the principal component analysis to see if different variables had predictive value. After using backward selection, the four most significant variables to predict the total score turned out to be home offensive total yards, home offensive rushing attempts, away offensive passing yards, and away defensive rushing yards allowed. All were very significant predictors with p-values less than .001. However, we wanted to try seeing which quarterback metrics gave predictive value. When a regression was run with quarterback metrics, we ended up with two significant predictors: the home and away yards per gain. Using these two variables along with the four we obtained earlier, we created another model which only had an adjusted R-squared of .07.

The predictive model we ended up with for the total measurement consists of two individual linear models, one predicting the number of points scored by the home team and the other predicting the number of points scored by the away team. For the home team points model, we created a full model that included all the numerical offensive and defensive statistics that corresponded with the home team and then created a model through the stepwise model selection method. This home team points model included twenty one predictors, eleven were offensive statistics and ten were defensive statistics. All but two of the predictors were statistically significant in the model. Reducing the model to exclude these non-significant predictors led to a lower R^2 value, so we decided to include the non-significant predictors in order to have a more holistic model to predict the home score. The predictors and their p-values for the home model are pictured below.



```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)    23.1186987  7.5665184   3.055 0.002259 ** 
Off_Pts_H       0.0297708  0.0050388   5.908 3.68e-09 ***
Off_Plays_H     0.0371995  0.0161928   2.297 0.021644 *  
Off_Total_1D_H  0.0252320  0.0156346   1.614 0.106622    
Off_Cmp_H       0.0364509  0.0110035   3.313 0.000931 ***
Off_Pass_Att_H -0.0418732  0.0167046  -2.507 0.012218 *  
Off_Pass_Yds_H  0.0015065  0.0008751   1.721 0.085225 .  
Off_Pass_1D_H  -0.0719419  0.0202619  -3.551 0.000388 ***
Off_Rush_Att_H -0.0365934  0.0155024  -2.360 0.018289 *  
Off_Pen_H      -0.0581536  0.0219177  -2.653 0.007997 ** 
Off_Pen_Yds_H   0.0047173  0.0024777   1.904 0.056983 .  
Off_Exp_Pts_H  -0.0048983  0.0030845  -1.588 0.112336    
Def_Rk_H        0.1001379  0.0445721   2.247 0.024706 *  
Def_Total_1D_H  0.0451405  0.0146326   3.085 0.002047 ** 
Def_Pass_Att_H -0.0249929  0.0072863  -3.430 0.000608 ***
Def_Pass_Yds_H  0.0013699  0.0008024   1.707 0.087839 .  
Def_Pass_TD_H  -0.1790329  0.0519038  -3.449 0.000567 ***
Def_Rush_Att_H -0.0135711  0.0074281  -1.827 0.067759 .  
Def_Rush_TD_H  -0.1634391  0.0598223  -2.732 0.006316 ** 
Def_Pen_Yds_H  -0.0023002  0.0012970  -1.774 0.076202 .  
Def_Sc_Pct_H   -0.2787251  0.0708083  -3.936 8.39e-05 ***
Def_TO_Pct_H   -0.2613485  0.0594690  -4.395 1.13e-05 ***
```

A similar method was used to find a model to best predict the score of the away team. We again created a full model that included all the numerical offensive and defensive statistics, but corresponding to the away team and used the stepwise model selection method to select which variables were the best for predicting away team score. The away team points model included fifteen predictors, nine of which were offensive statistics and six were defensive statistics. Like the home scoring model, the away scoring model also had two non-significant predictors, but we decided to include them for similar reasons we included similar predictors in the home scoring model. The predictors and their p-values for the away model are shown below.



```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)     16.6749292  3.6935232   4.515 6.49e-06 ***
Off_Rk_A        -0.0973164  0.0483458  -2.013 0.044176 *  
Off_Pts_A        0.0175872  0.0085046   2.068 0.038695 *  
Off_Total_1D_A   0.0795505  0.0204089   3.898 9.83e-05 ***
Off_Cmp_A        0.0118636  0.0075169   1.578 0.114570    
Off_Pass_Yds_A   0.0017898  0.0008242   2.172 0.029938 *  
Off_Pass_TD_A   -0.1052833  0.0538243  -1.956 0.050514 .  
Off_Pass_1D_A   -0.1044408  0.0269364  -3.877 0.000107 ***
Off_Rush_TD_A   -0.1654686  0.0620032  -2.669 0.007639 ** 
Off_Rush_1D_A   -0.0566388  0.0231532  -2.446 0.014469 *  
Off_Pen_A       -0.0291394  0.0096917  -3.007 0.002655 ** 
Def_Total_Yds_A -0.0022181  0.0008813  -2.517 0.011877 *  
Def_FL_A        -0.1396477  0.0418773  -3.335 0.000860 ***
Def_Total_1D_A   0.0399820  0.0142783   2.800 0.005127 ** 
Def_Pass_Yds_A   0.0016904  0.0011201   1.509 0.131340    
Def_Pass_1D_A   -0.0426572  0.0203867  -2.092 0.036453 *  
```

For cross-validation, the data was once again split into training and test sets, with the training set being composed of 85% of the original data set and the test set with 15%. The model was applied and we extracted the residuals and calculated RMSE, which resulted in a value of 10.99. While this value is “high”, it does make sense because of the range of scores possible in a given game. Mean Absolute Error (MAE) was also calculated and returned a value of 8.32. This value indicates that whatever score is predicted using our model, on average we would expect an actual score +/- 8.32 of the prediction. This also tells us that our predictive accuracy is around one touchdown off.

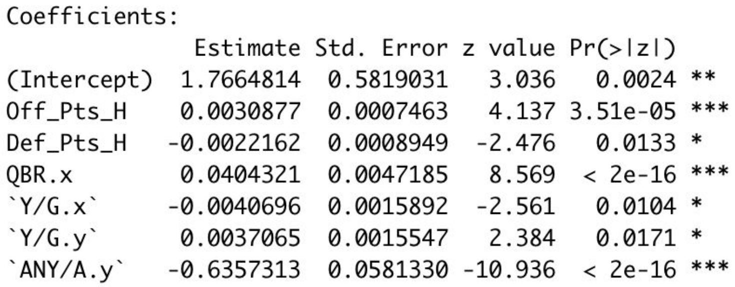
To predict the total scores for the upcoming games, we created new datasets based on the teams playing and designated whether they were the home or away team. We then applied the respective models to each dataset to predict what the individual home and away scores would be, and then added those values together to predict the total score of the game. We also explored adding squared and cubic terms to the stepwise models to see if the addition would help increase the model’s predictive power. However, adding non-linear terms to the model ended up decreasing the fit, so we decided to continue with the linear model since it provided simplicity as well as accuracy.

IV. Methodology for Result

One method that our group utilized to predict NFL game results was using the linear models that predicted the home and away scores for each game. Since the linear models had significant predictors and were good fits of the data, we concluded that we could rely on the predicted scores to determine whether the home or away team won. So, comparing the predicted scores from the linear model, if the away team scored more than the home team, then the result measurement would be zero. On the other hand, if the predicted home score was more than the predicted away score, then the result measurement would be one. After applying this process to the remainder of the games, we went through the results to make sure they made logical sense with the current rankings of the NFL. There were instances when this method seemed to incorrectly predict the result. For example, this method predicted that the Patriots would win against the Bills in the game on November 1st. Given that the Bills are top ranked in the AFC East with a winning record while the Patriots have only won two games all season, it seems more likely that the Bills would win this game. Inconsistencies like this indicated that our group needed to continue with finding other methods to predict the results of the games.

Another approach we tried was to just use a binomial logistic regression that could give us either a 1 or 0 for whether or not the home team won that game, based on the predictors. The first generalized model we tried was with the 20 variables obtained from the principal component analysis from earlier (40 total with home and away). Ten of the predictors were statistically significant given a .05 alpha, and two, home offensive points and home defensive points allowed were significant given a .001 alpha. The AIC value was 6,667, which was really high, probably due to the lack of parsimony, and once we performed 10-fold cross validation, we obtained a prediction error of .235. Once this model was done, we performed the same analysis with just the quarterback metrics. This time, of the 18 variables in the model, 15 were statistically significant given a .05 alpha, and the AIC was 2,606, much better than what we got earlier. The predictive error in this model was .211, which is another improvement.

The best model we went with to predict our final results for the games was a combined version of what we obtained from these two binomial models. We used significant predictors from both to try to find the model that resulted in the lowest AIC and most significant predictors.



```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept)  1.7664814  0.5819031   3.036   0.0024 ** 
Off_Pts_H    0.0030877  0.0007463   4.137 3.51e-05 ***
Def_Pts_H   -0.0022162  0.0008949  -2.476   0.0133 *  
QBR.x        0.0404321  0.0047185   8.569  < 2e-16 ***
`Y/G.x`     -0.0040696  0.0015892  -2.561   0.0104 *  
`Y/G.y`      0.0037065  0.0015547   2.384   0.0171 *  
`ANY/A.y`   -0.6357313  0.0581330 -10.936  < 2e-16 ***
```

This model had an AIC of 2,616, which is only slightly higher than the model only using quarterback data. The prediction error here was also .211, which means on average, 80% percent of the observations in the testing sets yielded the correct result. One of the main reasons we decided that this was the best was because of its low prediction error and AIC. The other model with just quarterback data also gave similar values; however, we wanted to combine team statistics with quarterback statistics since we know the quarterback is not the only person on the team. We also tried other more advanced models such as classification trees to see if it had any predictive power. The root node errors across the trees were consistently around .43, with a calculated predictive error of .35 (once multiplied with the X-error), which was much worse than the predictive errors we obtained earlier.

We created an R function to run our final model based on the home and away team ID's passed in as parameters, and based on the log probabilities, we assigned 1 and 0 to each game. The values we obtained mostly matched with the sign of the spread values we predicted. The ones that did not match were games that were within 5 points. We decided not to alter any values however, since that would ruin the point of the models we created for each section. The only alterations we made in this case were on the spread values if there was a zero, since ties are extremely rare.