David Risius
OA3103 Final Case Study
10 June 2013

How to Win Big by Betting on the NFL Playoffs

(Using Data Analysis to Model the Points Scored on an NFL Playoff Game Using Regular Season Data)

By: David Risius

INTRODUCTION

Professional sports gamblers and analysts have always longed for a reliable way to predict the outcome of football games based on statistics. There are entire sites devoted to analyzing sports statistics and their use in helping an individual pick their fantasy football team or win big money. Despite a plethora of data and statistics on the sport, the analysts seem to be wrong just as much as they are right in predicted both the scores and the overall outcomes of the game. Is this because they are just guessing or are they using some sort of prediction model that is not very good? The question that needs answering is whether there exists an accurate playoff score prediction model that one can use in triumphing over friends and winning a lot of money.

For the OA3103 Case Study, attempts were made to fit a model that reliably predicts the points scored by a team during an NFL Playoff game based on regular season offensive statistics for the team and the defensive statistics for the opponent. As will be shown in the analysis, there seems to be a lot of variability in what determines the overall points scored for a team which cannot be accurately modeled with just sports statistics variables. As such, the model produced is not quite as accurate as one would hope to achieve. The bottom line is that; if there were an accurate model out there that could accurately predict an NFL Playoff score, everyone would probably be using it.

DATA DESCRIPTION

All of the data used in this model was collected from nfl.com and espn.go.com/nfl. To get a sufficient number of data points for the model, data was initially collected from 18 different postseason playoff games in 2011 and 2012, giving 36 separate data points (two teams for each game for the response variable. The input variables used were statistics from the 2011/2012 regular season. After the initial model was fit, 59 more observations were collected to attempt to achieve a more accurate model. Table 1 below describes the response and input variables used in the model. There was initially one response variable (points scored) and nine input variables (X1-X9). For part II of the case study, six more variables were added to attempt a better fit (X10-X15).

Table 1. Description of Model Variables

Variable	Label	Variable Description
Points Scored	Y	Continuous-Numeric; Number of points scored in the playoff game by the NFL team of interest
Average Points per game	x1	Continuous-Numeric; The average points per game scored in the regular season.
Rush yards per game	x2	Continuous-Numeric; The average rushing yards per game achieved in the regular season.
Rushing touchdowns	x3	Continuous-Numeric; Total rushing touchdowns scored in the regular season.
Receiving yards per game	x4	Continuous-Numeric; The average receiving yards per game achieved in the regular season.

Receiving touchdowns	x5	Continuous-Numeric; Total receiving touchdowns scored in the regular season.
Quarterback Rating	x6	Continuous-Numeric; The overall calculated rating of the starting quarterback from the regular season.
Defensive Points per Game	x7	Continuous-Numeric; The average points allowed by opposing team in each game.
Defensive Yards per Play	X8	Continuous-Numeric; The average yards per play allowed by opposing team in each game.
Home or Away	X9	Categorical; Whether the game is played at the teams home field or away.
Total Fumbles	X10	Continuous-Numeric; Total times the offensive team fumbled the ball
Defensive Fumbles	X11	Continuous-Numeric; Total fumbles the opposing team defense caused in the regular season
Pass Yards	X12	Continuous-Numeric; Total pass yards the the offense
Defensive Penalty Yards	X13	Continuous-Numeric; Total penalty yards the opposing team defense received in the regular season.
Receiving Average Yards	X14	Continuous-Numeric; Average receiving yards in a paly
Total Yards Per Game	X15	Continuous-Numeric; Combination of Rush yards and Receiving yards per game

The response variable can vary from zero to any positive integer numbers, however the responses for this data varied from 2 to 51 points with a mean of 24.23 and a median of 24. Figure 1 and 2 below shows pairwise plots for each response and input variable using the original 36 observations.

Figure 1. Pairwise Plot 1 for Response and Original Regressor Variables

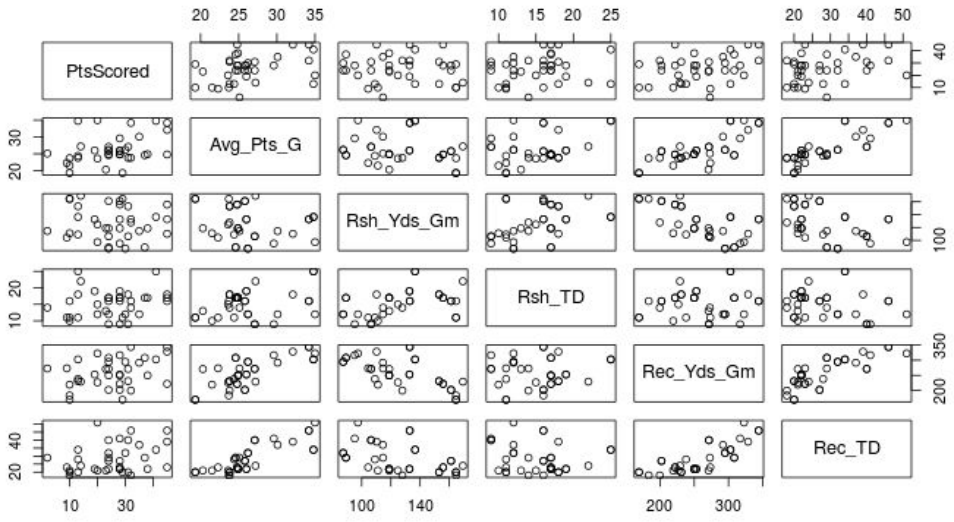
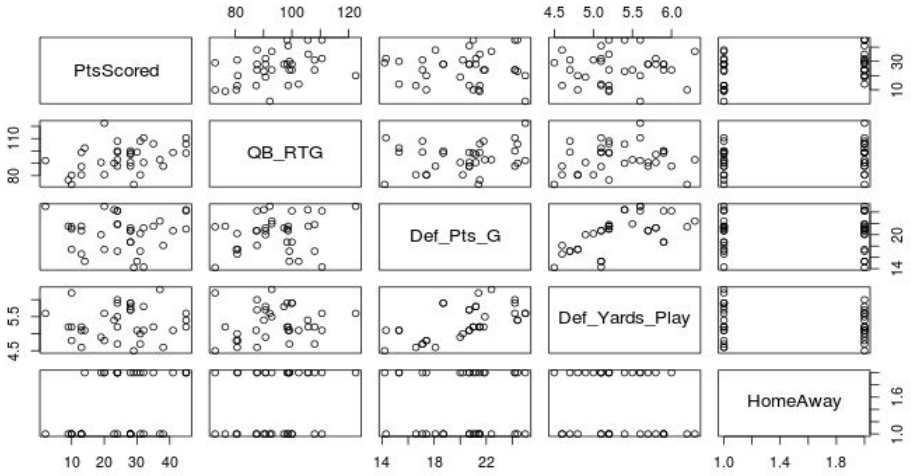


Figure 2. Pairwise Plot 2 for Response and Original Regressor Variables



In both of these plots, the response variable is plotted as a function of the input variables across the top row. At first glance, the strongest linear relationships to total points scored appear Average Points per Game, Quarterback rating, and possibly Receiving Yards per game. The rest of the variables may not have a strong linear relationship to the response variable, however they were tested as part of the model anyway.

ANALYSIS

In this section, two types of analysis are presented. First a one way ANOVA was used to study the relationship between Points Scored and whether the team enjoyed a home field advantage. We specifically want to see whether playing at home will actually improve the teams mean score in the playoffs. The ANOVA will be followed by multiple linear regression analysis. In my analysis, I attempted to fit the following model:

ANOVA:

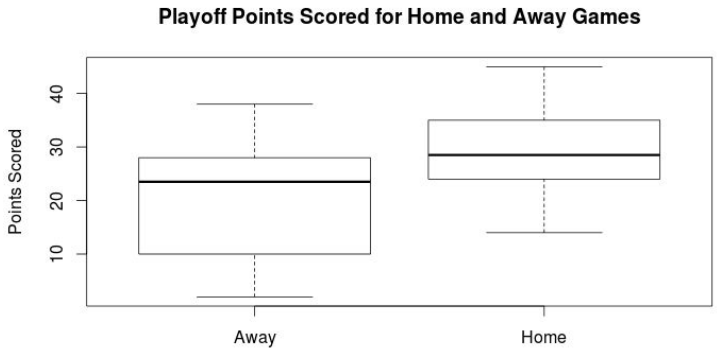
For this portion of the analysis, we wanted to see determine whether playing at home actually has an effect on the predicted points scored. For our test, the following hypothesis was used where μ_i is the average points scored in a game for i = Home or Away:

$$Ho: \ \mu_H = \mu_A$$

$$Ha: \ \mu_H \neq \mu_A$$

Figure 3 below shows a boxplot of the points scored versus Home or Away games. It appears from the figure that there could possibly be a difference in the mean points scored however further analysis is necessary.

Figure 3. BoxPlot of Points Scored in Home and Away Games



To determine whether there is actually a difference between the points scored in home and away games, a one way ANOVA table was used. The output below shows the ANOVA table generated from R.

Figure 4. ANOVA table generated for playoff HomeAway regressor

```
> anova(HomeAway.lm)
Analysis of Variance Table

Response: Playoffs$PtsScored
          Df Sum Sq Mean Sq F value  Pr(>F)
HomeAway   1  650.3  650.25  6.1732 0.01805 *
Residuals 34 3581.4  105.33
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The p-value generated in R is .01805, which causes us to reject the null hypothesis and conclude that the two means are not equal and whether the game is played at home or away does affect the average points scored. This is to be expected since in the NFL playoffs, the team with the better record in the regular season usually enjoys the home field advantage.

Regression Analysis:

The initial model for predicting the total points scored was assumed to have the following form:

$$y = \beta_0 + \beta_1 x1 + \beta_2 x2 + \beta_3 x3 + \beta_4 x4 + \beta_5 x5 + \beta_6 x6 + \beta_7 x7 + \beta_8 x8 + \beta_9 x9 + \varepsilon$$

Figure 5. Original Playoff Linear Model Fit

```
> summary(Playoff.lm1)

Call:
lm(formula = PlayoffReorder$PtsScored ~ Avg_Pts_G + Rsh_Yds_Gm +
    Rsh_TD + Rec_Yds_Gm + Rec_TD + QB_RTG + Def_Pts_G + Def_Yards_Play +
    HomeAway, data = PlayoffReorder)

Residuals:
    Min      1Q  Median      3Q     Max
-21.626  -4.801  -0.363   5.015  19.998

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    -57.7836    45.7504  -1.263   0.2182
Avg_Pts_G        0.0875     1.6670   0.052   0.9586
Rsh_Yds_Gm       0.1532     0.1565   0.979   0.3370
Rsh_TD          -0.4371     1.2675  -0.345   0.7331
Rec_Yds_Gm       0.1442     0.1252   1.152   0.2602
Rec_TD          -0.5303     1.0222  -0.519   0.6085
QB_RTG           0.3533     0.3359   1.052   0.3030
Def_Pts_G       -0.5430     0.7931  -0.685   0.4998
Def_Yards_Play   3.7677     5.1487   0.732   0.4711
HomeAwayHome     7.3223     3.8445   1.905   0.0684 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.49 on 25 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared: 0.3117, Adjusted R-squared: 0.06387
F-statistic: 1.258 on 9 and 25 DF,  p-value: 0.307
```

Based on the R output, we failed to reject the null hypothesis and state that we cannot conclude that any of the coefficients has any significant effect on the predicted points scored in a game. Additionally, we can see that the R-squared value is much closer to zero than one which means that our model is not a very good fit for predicting points scored. Below is the current fitted model for predicting points scored:

$\hat{y} = -57.78 + .0875x1 + .153x2 - .437x3 + .144x4 - .53x5 + 353x6 - .543x7 + 3.768x8 + 7.322x9$, *where x9 is equal to one when the HomeAway category is home, zero otherwise.*

If we go back to our original scatter plot in which hypothesized that Average Points per Game, Quarterback rating, and Receiving Yards per game appeared to have a linear relationship with total points scored we get the following model:

Figure 5. Improved Model Fit Using Pairwise Plots as Guide

```
> summary(Playoff.lm2)

Call:
lm(formula = PlayoffReorder$PtsScored ~ Avg_Pts_G + Rec_Yds_Gm +
    QB_RTG + HomeAway, data = PlayoffReorder)

Residuals:
    Min      1Q  Median      3Q     Max
-19.836  -6.737   1.325   5.912  17.614

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -7.70690   14.43646  -0.534   0.5974
Avg_Pts_G     0.13873    0.78017   0.178   0.8601
Rec_Yds_Gm    0.01560    0.05536   0.282   0.7800
QB_RTG        0.23650    0.25159   0.940   0.3547
HomeAwayHome  6.44766    3.45618   1.866   0.0719 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.999 on 30 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared: 0.2491, Adjusted R-squared: 0.149
F-statistic: 2.488 on 4 and 30 DF,  p-value: 0.06444
```

This is by no mean a "good" model as the R-Squared and Adjusted R-Squared are closer to zero than one. However, the p-value is .064 is better than our first model. Using ANOVA to compare the two models, we get the following result for the null hypothesis.

$$Ho: \beta_i = 0 \; for \; i = \{RushYards, RushTDs, RecTD, DefPoints, DefYards\}$$

$$Ha: At \; least \; one \; \beta_i \; is \; not \; equal \; to \; zero \; given \; the \; other \; terms \; are \; in \; the \; model$$

Figure 6. ANOVA Comparison of Two Initial Models

```
> anova(Playoff.lm2, Playoff.lm1)
Analysis of Variance Table

Model 1: PlayoffReorder$PtsScored ~ Avg_Pts_G + Rec_Yds_Gm + QB_RTG +
    HomeAway
Model 2: PlayoffReorder$PtsScored ~ Avg_Pts_G + Rsh_Yds_Gm + Rsh_TD +
    Rec_Yds_Gm + Rec_TD + QB_RTG + Def_Pts_G + Def_Yards_Play +
    HomeAway
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     30 2999.2
2     25 2749.3  5    249.91 0.4545  0.806
```

Based on this, we must fail to reject the null hypothesis and conclude that we cannot determine whether the coefficients listed in the null hypothesis are zero.

Initial Conclusions:

The model developed in part I was not reliable enough for professional sports gambling in Las Vegas or even wowing friends with a perceived sixth sense, however with more data points and possibly modifying the input variables, we hoped to be able to more accurately predict the points scored in an NFL playoff game. The next step in the analysis was to gather more data, scope and add additional input variables and check the adequacy of the model via the residuals. Further analysis was completed in the second part of the case study and is detailed in the remainder of the report.

David Risius
OA3103 Final Case Study
10 June 2013

PART II. MODEL ADEQUECY and VALIDATION

Below is the fitted regression equations for the final model (FinalPlayoff.3) used in the case study. FinalPlayoff.3 was fitted by first running a model including some of the new regressor variables (X10-X15) in Table 1. Once a base model was fit with all two factor interactions, backward stepwise regression was to get a more accurate model. Next, 24 iterations of the drop1 command were performed to delete unnecessary and insignificant terms until we had a manageable model in which all terms were significant with p-values under .05.

Fitted Final Regression Model

$$\hat{PtsScored} = 55.395 + 5.727x_{QB\_RTG} + 11.884x_{Pass\_Int} + 3.461x_{HomeAway} - 1.026x_{Def\_Plys} + 14.084x_{Def\_FUM} + 134.16x_{Home}$$

; where $x_{HomeAway}$ equals 1 when HomeAway category is Home, 0 otherwise.

Figure 7 shows the summary statistics for FinalPlayoff.3. Notice that in this model, each of the regressor variables appears to be significant in the presence of the others given a significance level of .05.

Figure 7. Summary Statistics for FinalPlayoff.3

```
> summary(FinalPlayoff.3)

Call:
lm(formula = Playoffs$PtsScored ~ QB_RTG + Pass_Int + Def_Plys +
    Def_FUM + HomeAway + QB_RTG:Rec_Avg + Pass_Int:Rec_Avg +
    Rec_Avg:Def_Plys + Rec_Avg:Def_FUM + Def_Plys:HomeAway, data = Playoffs)

Residuals:
     Min       1Q   Median       3Q      Max
-20.1953  -6.4494   0.2314   5.9568  20.1421

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           55.39520   36.07799   1.535  0.12844
QB_RTG                 5.72664    1.82969   3.130  0.00241 **
Pass_Int              11.88377    4.40375   2.699  0.00842 **
Def_Plys              -1.02551    0.23585  -4.348 3.83e-05 ***
Def_FUM               14.08380    3.26775   4.310 4.42e-05 ***
HomeAwayHome        -134.15762   47.37823  -2.832  0.00579 **
QB_RTG:Rec_Avg         -0.45345    0.15093  -3.004  0.00351 **
Pass_Int:Rec_Avg       -0.99188    0.37386  -2.653  0.00954 **
Def_Plys:Rec_Avg        0.08119    0.01925   4.219 6.17e-05 ***
Def_FUM:Rec_Avg        -1.19971    0.27544  -4.356 3.73e-05 ***
Def_Plys:HomeAwayHome   0.13536    0.04729   2.862  0.00531 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.119 on 84 degrees of freedom
Multiple R-squared: 0.3549, Adjusted R-squared: 0.2781
F-statistic: 4.622 on 10 and 84 DF,  p-value: 3.11e-05
```

Assumptions: Next, the model was checked to see how it meets each of the six assumptions.

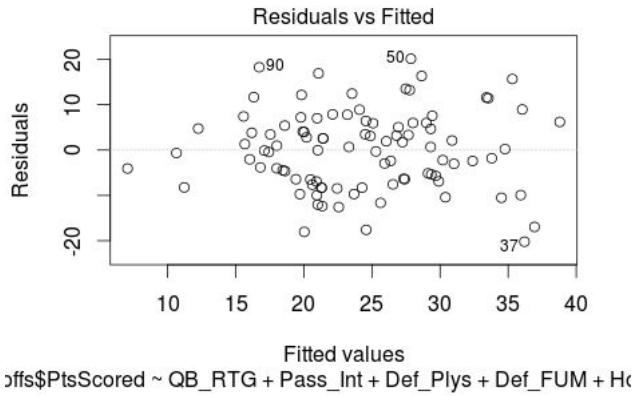
1. The error term ε has mean = zero.

> mean(FinalPlayoff.3$residuals) #check
[1] 6.141834e-17

2. The error term ε has constant variance.

Figure 7a. Residual versus Fitted Plot for PlayoffFinal.3 model

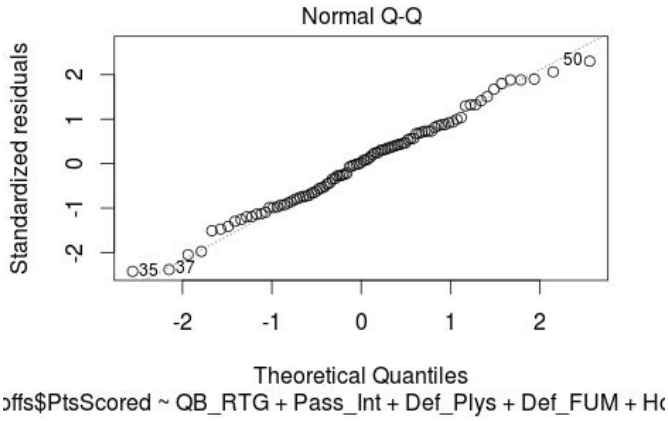
Looking at the residual versus fitted plot in Figure 7a, it appears that the error terms are evenly distributed throughout the plot.



3. The errors are normally distributed. **By the Shapiro-Wilk test and the normal Q-Q plot in figure 8, it appears that the data is normally distributed.**

Figure 8. Normal Plot for FinalPlayoff.3 Model

>shapiro.test(residuals(FinalPlayoff.2))
Shapiro-Wilk normality test
data: residuals(FinalPlayoff.2)
W = 0.9935, p-value = 0.9294



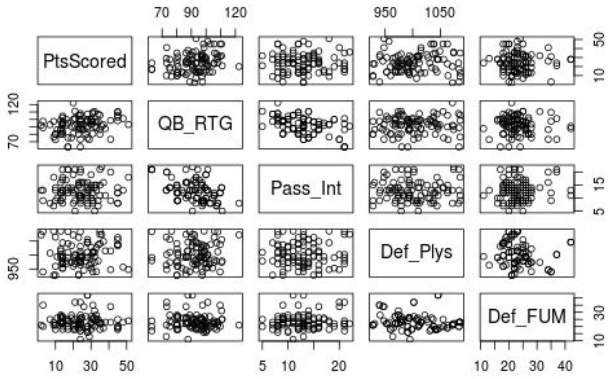
4. The errors are uncorrelated. **Using the residual versus fitted plot in Figure 7, it appears that there is no evident sequence of points. Also utilizing the Durbin-Watson test on the model gives us a p-value of .394 which causes us to Fail to Reject the null hypothesis that the error terms are uncorrelated.**

> durbinWatsonTest(FinalPlayoff.3)
lag Autocorrelation D-W Statistic p-value
1 -0.08698335 2.158074 0.394
Alternative hypothesis: rho != 0

5. The relationship between the response and the regression variables is correct.

Figure 9. Pair Plot of Main Effect Terms in FinalPlayoff.3

Looking at Figure 9 (pairs plot) for the main effects terms in FinalPlayoff.3, there does not appear to be any non linear relationships between the response and regressors.



6. The regression variables are independent. **Based on the variance inflation factors shown in Figure 8, we do have strong evidence of multi-collinearity among the regressor variables. This was to be expected as there are several two factor interactions in the model. In an attempt to correct this and lower the vif's, several variables were removed from the model starting with the highest vif's and using the drop1 commands in R. Although removing several of the terms does lower the variance inflation factors, it also significantly increased the RSE and lowered the R squared and adjusted R squared and in general lowered the performance of the model. See Figure 10b. Therefore, the decision was made to keep the model in its current form and live with the multi-collinearity.**

Figure 10. Variance Inflation Factors for FinalPlayoff.3

```
> vif(FinalPlayoff.3)
          QB_RTG          Pass_Int          Def_Plys          Def_FUM          HomeAway
        508.3328          355.2492          115.1529          358.5909          641.0145
   QB_RTG:Rec_Avg  Pass_Int:Rec_Avg  Def_Plys:Rec_Avg  Def_FUM:Rec_Avg Def_Plys:HomeAway
        695.8060          341.8564          394.6209          363.3181          633.2190
```

Figure 10b. Final Model After Colinearity is Reduced (Notice Decreased Performance)

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -2.800e+01  2.865e+01  -0.977  0.33100
QB_RTG               2.872e-01  1.088e-01   2.639  0.00984 **
Pass_Int             4.367e-01  3.081e-01   1.417  0.15989
Def_Plys             2.327e-02  3.351e-02   0.694  0.48938
Def_FUM             -1.368e-02  2.023e-01  -0.068  0.94624
Def_Plys:Rec_Avg    -3.328e-04  1.473e-03  -0.226  0.82184
Def_Plys:HomeAwayHome  2.105e-03  2.347e-03   0.897  0.37224
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.47 on 88 degrees of freedom
Multiple R-squared: 0.1088, Adjusted R-squared: 0.04803
F-statistic: 1.791 on 6 and 88 DF,  p-value: 0.1102

> vif(FinalPlayoff.3v)
          QB_RTG          Pass_Int          Def_Plys          Def_FUM  Def_Plys:Rec_Avg Def_Plys:HomeAway
        1.364176          1.318479          1.763246          1.042335          1.753772          1.182849
```

Model Validation

The first step in validating the model was performing manual cross validation. To perform the cross validation, the Playoff data set was split into a Playoff.main and Playoff.test data set. Since the total number of observations was 95, the test data set consisted of a random sample of 15 observations (approximately 16%) Each of the fitted models was recalculated using the main data and used to predict the PtsScored on the test data. The following R code was used to split the test data.

```
test.rows <- sample(1:95, 15,replace = F)
Playoffs.main <-
data.frame(Playoffs[-test.rows,]
Playoffs.test <-
data.frame(Playoffs[test.rows,])
```

The FinalPlayoff.3 model was refit using the Playoff.main data set and used to predict the Playoff.test data set. Figure 11 shows the summary statistics for this. Notice that the RSPE for this model was 10.5 compared to 8.988 RSE on the new fitted model. This was a difference of about 16.9 percent.

Figure 11. Manual Cross-Validation Results for FinalPlayoff.3 Model

```
> summary(FinalPlayoff.3Main)

Call:
lm(formula = Playoffs.main$PtsScored ~ QB_RTG + Pass_Int + Def_Plys +
    Def_FUM + HomeAway + QB_RTG:Rec_Avg + Pass_Int:Rec_Avg +
    Rec_Avg:Def_Plys + Rec_Avg:Def_FUM + Def_Plys:HomeAway, data = Playoffs.main)

Residuals:
     Min       1Q   Median       3Q      Max
-18.5647  -6.5271   0.6835   5.2420  19.3050

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         43.72973   37.60350   1.163 0.248869
QB_RTG               7.21865    2.00970   3.592 0.000611 ***
Pass_Int            14.58014    4.77129   3.056 0.003190 **
Def_Plys            -1.21247    0.25538  -4.748 1.08e-05 ***
Def_FUM             15.26279    3.32320   4.593 1.91e-05 ***
HomeAwayHome      -116.61874   48.56343  -2.401 0.019032 *
QB_RTG:Rec_Avg      -0.57169    0.16574  -3.449 0.000962 ***
Pass_Int:Rec_Avg    -1.19650    0.40409  -2.961 0.004202 **
Def_Plys:Rec_Avg     0.09700    0.02094   4.631 1.66e-05 ***
Def_FUM:Rec_Avg     -1.30001    0.28021  -4.639 1.61e-05 ***
Def_Plys:HomeAwayHome  0.11818  0.04839   2.442 0.017163 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.988 on 69 degrees of freedom
Multiple R-squared: 0.4302, Adjusted R-squared: 0.3477
F-statistic:  5.21 on 10 and 69 DF,  p-value: 1.162e-05

> predict.test3 <- predict(FinalPlayoff.3Main, newdata = Playoffs.test) #making pr
> RSPE3 = sqrt(sum((Playoffs.test$PtsScored-predict.test3)^2) / 15)
> RSPE3
[1] 10.50856
> #Calculate how far the predicted value is off from the model
> abs(8.988-RSPE3)/8.988*100 #about 2.8 percent off
[1] 16.91767
```

The RSE was also predicted using the cross validation function in R. Figure 12 shows the results.

Figure 12. Cross Validation in R Results

```
> cv.glm(data=Playoffs, FinalPlayoff.3a, K=6)$delta
[1] 98.28165 95.71153
> sqrt(98.28)
[1] 9.913627
> sqrt(95.71)
[1] 9.783149
```

Notice that this gives similar results for RSPE compared to the manual version in figure 11.

The PRESS statistic was then used to predict the R^2 value using the following equations.

$R^2=1-PRESS/SS_{total}$
$R^2= 1 - (SS_{res} / SS_{total})$
$SS_{total} = ((RSE^2)*(n-p))/(1- R^2)$

For the FinalPlayoff.3 model, we get the following results for the PRESS statistic and predicted R squared value.

Figure 13. Predicted R^2 using PRESS Statistic

```
Residual standard error: 9.119 on 84 degrees of freedom
Multiple R-squared: 0.3549, Adjusted R-squared: 0.2781
F-statistic: 4.622 on 10 and 84 DF,  p-value: 3.11e-05

> SSTotal.P3 <- ((9.119^2)*84)/(1-.3549)
> R2.P3.Pred <- 1- (PRESS.final.P3)/(SSTotal.P3)
> R2.P3.Pred
[1] 0.1260054
```

Notice the predicted R squared value of .126. This is significantly below the adjusted R squared value for the fitted model of .279. This shows that the model isn't explaining as much of the variability in the data as we originally predicted in the presence of new data.

Table 2 lists the overall summary of statistics for the FinalPlayoff.3 model.

Table 2. Overall FinalPlayoff.3 Model Validation Results

Model	RSE	R^2	R^2_{adj}	RSPE	PRESS	Pred R^2
FinalPlayoff.3	**9.119**	**.3549**	**.2781**	**10.51**	**9464**	**.126**

CONCLUSION

The final model produced is not exactly the most reliable predictor of NFL playoff score, however, it seemed to be the best out of over 40 models produced using techniques learned in the Data Analysis class. It appears that our initial hypothesis is correct. If there was a model that could accurately predict the score of an NFL Playoff game, it would probably be well known and widely used by sports analysts and gamblers. The model was initially fit using 36 observations and 9 regressor variables considered. It

was then expanded to 95 observations with 15 regressor variables considered with the hope that we could find a more accurate model with more observations and a larger set of possible regressor variables. The model developed in this case study is marginally better than a random guess at the score. For example, say the Minnesota Vikings were playing against the Seattle Seahawks in a playoff game and our model predicted the Vikings score was 20 points. Given an RSE of 9.119, and using one standard deviation, we would have an interval of approximately (11,29) points which is not exactly an accurate prediction. It may be possible that there is a model which accurately predicts an NFL Playoff score, however it was not found using the data that was collected in this case study. It is probable that there are just too many human factors involved in the sport of football to use just regular season statistics in a prediction model.