**B. Brewer, D. Nguyen, A. Scheerer February 13, 2019**

**MSDS 6372 Group Project I**

**INTRODUCTION:**

For our group project, we decided to look at the professional football team New England Patriots. Our goal is twofold:

1) To determine which, if any, variables explain the number of regular season wins for the Patriots since Tom Brady has been starting quarterback (last 18 seasons).

2) To determine if there is a statistical difference in Tom Brady’s performance over his wins and losses across his playoff career.

**DATA DESCRIPTION:**

We used two datasets, both obtained from [www.pro-football-reference.com](http://www.pro-football-reference.com). One dataset includes combined offensive and defensive stats, 70 variables total, for each of the patriots last 18 seasons (2001 through 2018). This data was compiled from multiple seasons on the [www.pro-football-reference.com](http://www.pro-football-reference.com) website, but the compiled dataset can be found at the “season stats” link below. The second dataset consists of metrics describing Tom Brady’s playoff performances during his tenure as starting quarterback for the New England Patriots. The data focuses on specifically passing metrics during both playoff and super bowl performances. The raw data can be found at the “raw data” link below and the cleaned data can be found at the project GitHub at the link below.

**Season Stats (all regular season statistics):**

<https://github.com/newtgunslinger/6372.404.AS.Project1/blob/master/PatriotsYearlyStats.csv>

**Tom Brady Playoff Stats:**

**Raw Data:** <https://www.pro-football-reference.com/players/B/BradTo00.htm#all_passing_playoffs>

**Cleaned Data:** <https://github.com/newtgunslinger/6372.404.AS.Project1/blob/master/BradyStats.csv>

Additional variables were created to support our analysis and some of the [www.pro-football-reference.com](http://www.pro-football-reference.com) require a little more explanation (see below).

**Season Stats (all regular, per season statistics):**

RegSeasonWins: Total wins per season.

BradyPasserRating: Tom Brady’s average passer rating.

PointsFor: Total points scored by the New England Patriots.

PointsAgainst: Total points allowed by the New England Patriots.

PointsDifferential: Total points scored minus total points allowed.

MarginOfVictory: Average point differential (points scored minus points allowed) per game.

StrengthOfSchedule: A metric describing average quality of opponent, measured by simple rating system.

SimpleRatingSystem: A metric describing team quality relative to average.

OffSimpleRatingSys: A metric describing offensive quality relative to average.

DefSimpleRatingSys: A metric describing defensive quality relative to average.

Yards: Total offensive yards made.

Plays: Total offensive plays.

YardsPerPlay: Total offensive yards per play.

Turnovers: Total turnovers.

FumblesLost: Total fumbles.

FirstDowns: Total number of first downs.

PassCompletions: Total number of pass completions.

PassAttempts: Total number of pass attempts.

CompletionPercentage: A ratio of PassCompletions to PassAttempts.

PassYards: Total number of passing yards.

PassTouchdowns: Total number of passing touchdowns.

PassInterceptions: Total number of interceptions.

NetYardsPerPass: Ratio of passing yards minus sack yards to passing attempts plus times sacked.

PassFirstDowns: Total number of passing first downs.

RushAttempts: Total number of rushing attempts.

RushYards: Total number of rushing yards.

RushTouchdowns: Total number of rushing touchdowns.

RushYardsPerAttempt: Average rushing yards per attempts.

RushFirstDowns: Total number of rushing first downs.

Penalties: Total number of penalties.

PenaltyYards: Total number of penalty yards.

PenaltyFirstDowns: Total number of New England Patriots penalties resulting in a first down.

NumberDrives: Total number of offensive drives.

DriveScorePercent: Percentage of an offensive drive resulting in a score.

DriveTurnoverPercent: Percentage of an offensive drive resulting in an offensive turnover.

AvgStartingPosition: Average yardage marker starting position for offense.

AvgDriveTime: Average amount of time run off the clock per drive.

AvgDrivePlays: Average number of plays per offensive drive.

AvgDriveYards: Average number of yards consumed per drive.

AvgDrivePoints: Average number of points scored per offensive drive.

**Tom Brady Playoff Stats (only three variables used):**

AgeBin: A categorical variable describing the point in Tom Brady’s career - early, mid, and late (1, 2, and 3 respectively).

Rate: Tom Brady’s passer rating for the individual playoff and Super Bowl games.

WonLost: Game won or lost by the New England Patriots.

**EXPLORATORY ANALYSIS:**

We used both SAS and R in our exploratory analysis and ultimately used SAS output for report quality diagrams. In both SAS and R, scatter plot matrix diagrams were produced to identify any variables that were colinear, identify non-linear trends in the data ideal for transformation, and highlight which variables had highest correlation with our response variable, RegSeasonWins. It was determined that opponent statistics were duplicative when considering strength of schedule. Additionally, it was found that many of the offensive variables were also duplicative, pass completions and passing attempts versus completion percentage, for instance. Using the scatter plot matrix, we were able to cherry-pick a handful of variables for our analysis.

**OBJECTIVE 1:**

**Problem:**

Which variables correlate to regular season wins for Tom Brady’s Patriots?

**Overall Approach:**

Using the aggregated regular season statistics, we will determine which variables correlate with the New England Patriots regular season record and subsequently build a model using those parameters to interpret the relationship between those parameters and the New England Patriots regular season record.

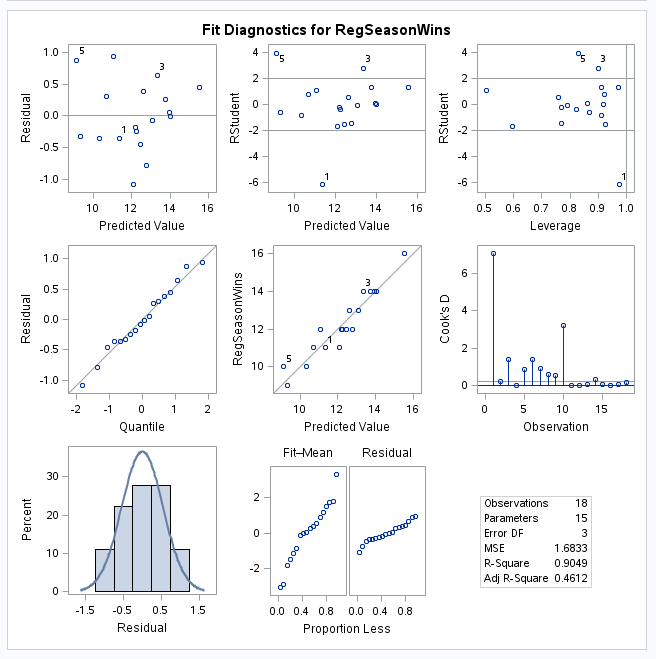
**Determining Predictors:**

First, we loaded the dataset into SAS.[1] Then, we performed exploratory data analysis on the data by creating multiple scatterplot matrices to identify which, if any, showed a correlation with regular season wins.[2] We identified 16 variables that showed a correlation with regular season wins and it was determined that none of the variables required a transformation. Our next step was to create a scatterplot matrix with these 16 variables.[3]



**Checking Assumptions:**

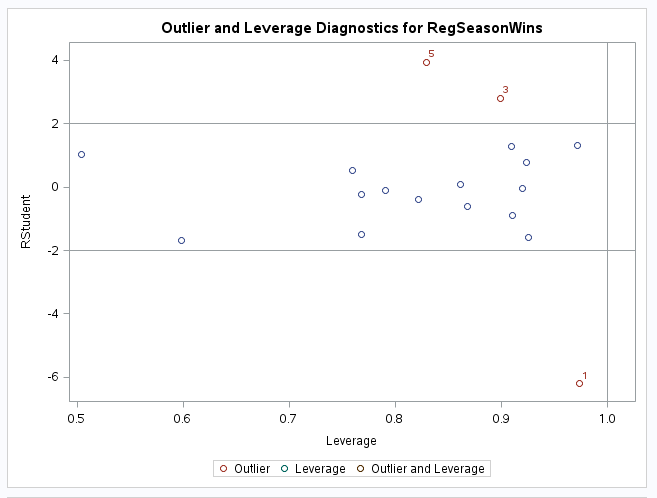
We then checked to see if the assumptions for multiple regression were met:[4]



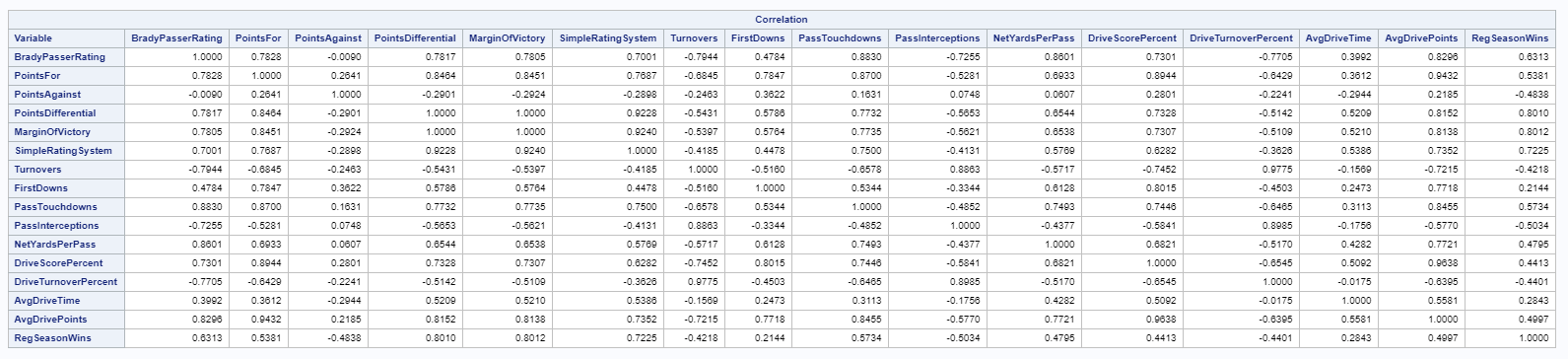
1) The residuals are normally distributed (predictors and response variables don’t have to be).

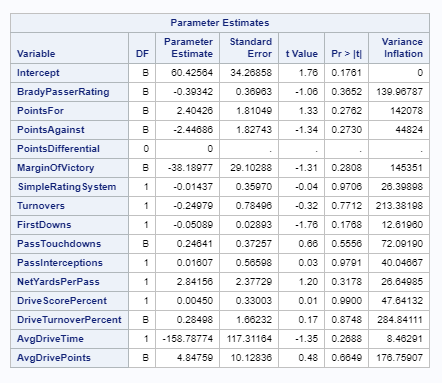
2) There is constant variance.

3) The observations (different seasons) are independent from and of one another.



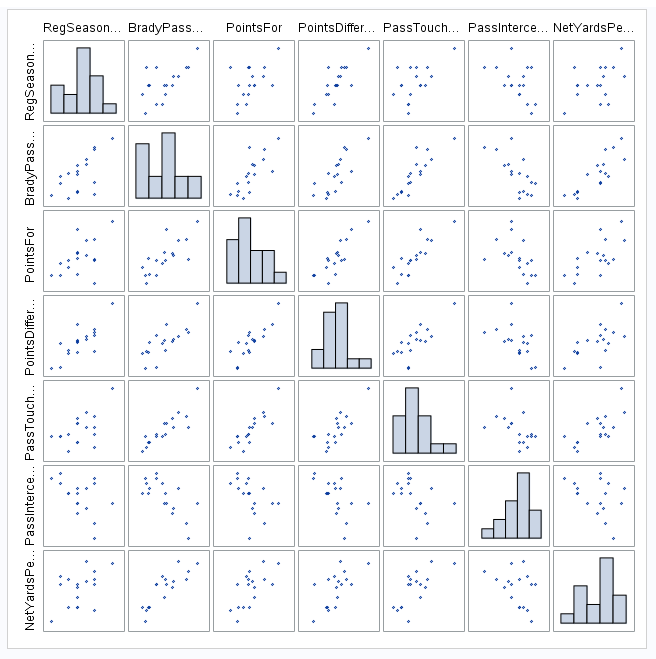
4) Checking through the residual diagnostics, there are three outliers and no leverage points. We will determine if any or all the three outliers need to be removed.





5) Looking at the correlation, we removed PointsAgainst, Turnovers, FirstDowns, DriveTurnoverPercent, and AvgDriveTime for having low correlations with RegSeasonWins. Looking through the VIFs, we also removed MarginOfVictory, SimpleRatingSystem, and AvgDrivePoints for having correlation with other variables (multicollinearity assumption).

We created a scatterplot matrix with the remaining predictors:[5]

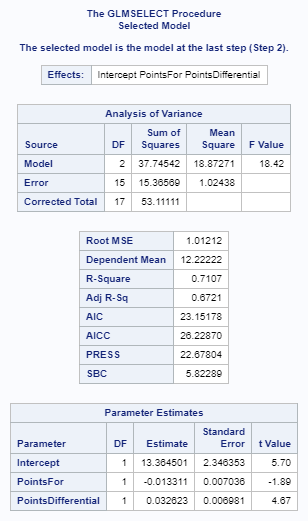


**Model Selection:**

We obtained our model through a series of feature selection tools: LARS, LASSO, stepwise, forward, and leave-one-out cross validation which was a k-fold cross validation taken to its extreme.[6]

**Final Model:**

Leave-One-Out Cross Validation:



**Model Selection Interpretation:**

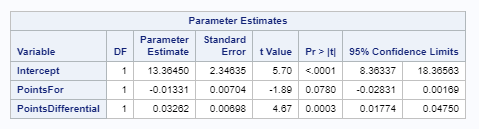
The LARS model gave us a R-square of 0.6281, the LASSO model gave us a R-square of 0.6281, the stepwise gave us an adjusted R-square of 0.6721, the partial with 2 parameters gave us an adjusted R-square of 0.5940, the partial with 1 parameter gave us a R-square of 0.6417, and the leave-one-out validation gave us an adjusted R-square of 0.6721.

The LARS model gave us an AIC of 25.67176, the LASSO model gave us an AIC of 25.67176, the stepwise gave us an AIC of 23.15178, and the leave-one-out validation gave us an AIC of 23.15178.

Out of all the models, we selected the leave-one-out validation as the best model because our dataset only contains 18 observations. The k-fold cross validation is best when there is a low number of observations because it makes use of all the observations through an iterative process where it removes one observation for testing validation. The R-Square and AIC results we obtained support this.

**Parameter Interpretation and Confidence Intervals:**

We ran a test to determine the confidence intervals for the variables.[7]



*RegSeasonWins = 13.3645 – 0.0133(PointsFor) + 0.0326(PointsDifferential)*

95% Confidence Limits for Intercept: (8.3634, 18.3656)

95% Confidence Limits for PointsFor: (-0.02831, 0.00169)

95% Confidence Limits for PointsDifferential: (0.01774, 0.04750)

If PointsFor and PointsDifferential are both zero, the Patriots will win between 8.3634 and 18.36560 with a 95% confidence.

Holding PointsDifferential constant, for every 1 increase in PointsFor, the Patriots will win between

-0.02831 and 0.00169 more games with a 95% confidence.

Holding PointsFor constant, for every 1 increase in PointsDifferential, the Patriots will win between 0.01774 and 0.04750 more games with a 95% confidence.

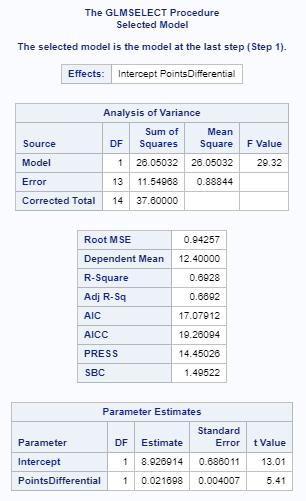
**Interpretation:**

If the New England Patriots finish the regular season with a 0-point differential, they will win 13.3645 games, keeping PointsFor constant. Logically, you would expect that if a team scores the same points as their opponents did you would see an 8-8 record. However, the tests we ran indicate that even when the Patriots score the same amount as their opponent, in the long run they will have a record above .500 (13-3). We’ve attributed this to parts of football that were not captured in our dataset. These attributes include coaching strategy, defensive and offensive schemes, and player management; all of which do not show up in the raw, end of season statistics we studied.

**Model with Outliers Removed Interpretation:**

We created a leave-one-out validation model removing the 3 outliers to see if it would result in a better model.[8]

Leave-One-Out Cross Validation w/ Outliers Removed:



Removing the 3 outliers resulted in a model with a R-square of 0.6692 which is in fact lower than the R-square for the model with the 3 outliers left in (0.6721). Therefore, we concluded that it would be better to include the 3 outliers in our model. Because of the low number of observations, it is best to have as many observations as possible.

**Final Conclusion from the Analyses of Objective 1:**

Against our expectation, it was determined that the regular season record was not directly affected by the individual performance of Tom Brady as Tom Brady’s individual statistics were not used to build the strongest model. Bill Belichick is the real goat, not Tom Brady.

**OBJECTIVE 2:**

**Problem:**

Is Tom Brady’s performance directly responsible for the New England Patriots’ success?

After exploring what regular season stats contributed to the Patriots’ success over the last 18 seasons, we sought to discover if there was any difference in Tom Brady’s passer rating between the New England Patriots wins and losses over three stages of Tom Brady’s career – early, mid, and late.

**Overall Approach:**

Since we are using a two-way ANOVA, we decided to bin every playoff and Super Bowl game by its end results (win or loss) and his career stage, identified by the quarterback’s age between 3 bins (24-30,30-36, and 36-42). In these 3 different stages, we are trying to determine if Tom Brady’s passer rating was higher in games that the Patriots won versus the games that the Patriots lost.

**Determining Predictors:**

We decided to use Brady’s passer rating as opposed to individual passing statistics such as touchdowns, yards, and interceptions because the passer rating statistic encompasses those metrics.

**Two-Way ANOVA Interpretation:**

In our model Rate = AgeBin + WonLost, the p-value for AgeBin is 0.77, which indicates that there is no difference in means between levels of Tom Brady’s age. The p-value for WonLost is 0.0578, although that is on the cusp of our 95% confidence level, we have decided to consider this variable significant and run a one-way ANOVA test with the lone variable as WonLost.[9]

**One-Way ANOVA Interpretation:**

In our model Rate = WonLost, the p-value for the WonLost variable drops a little further to 0.0561, further indicating that there is a significant difference between the mean of Tom Brady’s passer rating between the Patriots’ wins and losses in playoff and Super Bowl matches.[10]

**Conclusion:**

As we can see from the interaction plot[9], there was a slight increase in Brady’s passer rating if we compare his late and mid-career passer rating to his early career passer rating. The comparison between his wins versus losses passer rating is more evident. Additionally, the parallel quality of the lines in the interaction plot indicate that there is no interaction between AgeBin and WonLost. The Q-Q plot indicates a normal distribution and our residual plot implies normality and equal variance among observations.

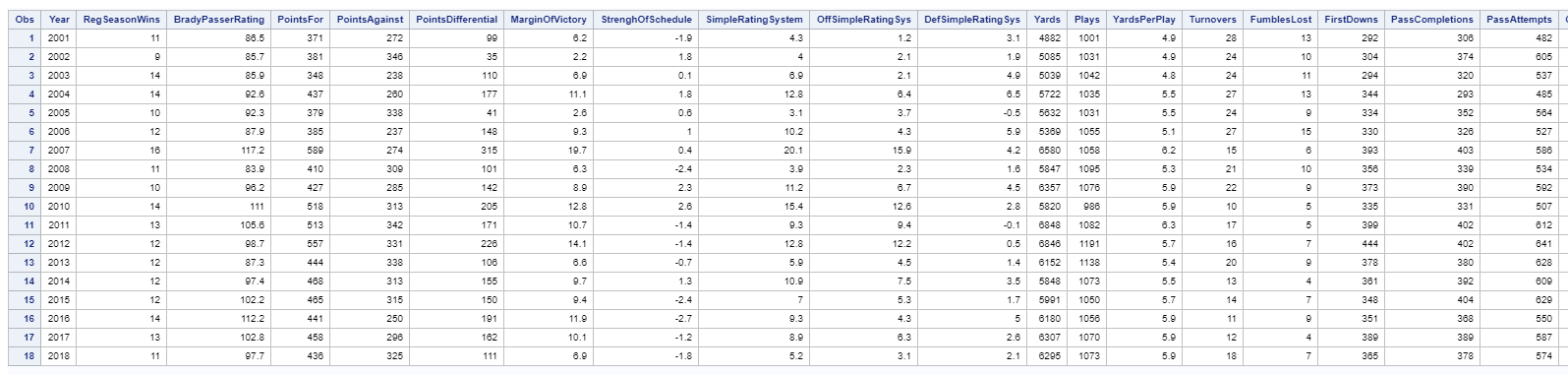
Section 10 of the appendix has some graphics regarding the one-way ANOVA performed after age is taken out of the dataset. The box and whisker plot show the evidence that Tom Brady’s performance does have a significant effect on their results in playoff games. This kind of strays from the evidence we saw in part one of our analysis which focused on the Patriots and Tom Brady’s regular season performance. We were unable to find evidence that Tom Brady’s individual performance had an effect in the regular season, but this seems to shift when the Patriots get to the post season. In conclusion, the Patriots have needed Tom Brady to play at his best in playoff games for them to come out on top in February.

**APPENDIX:**

[1]

PROC IMPORT OUT= WORK.pats  
 DATAFILE= "/home/daveknockwin0/PatriotsYearlyStats.csv"  
 DBMS=CSV REPLACE;  
 GETNAMES=YES;  
 DATAROW=2;  
RUN;

/\* Print dataset \*/  
proc print data=pats;  
run;



[2]

/\* Scatterplot matrices to determine predictors \*/

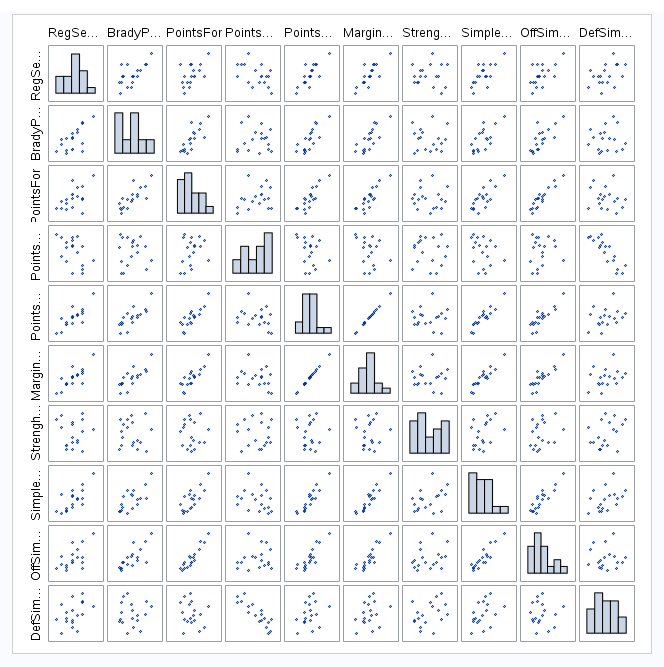
proc sgscatter data=pats;

matrix RegSeasonWins BradyPasserRating PointsFor PointsAgainst PointsDifferential MarginOfVictory StrenghOfSchedule

SimpleRatingSystem OffSimpleRatingSys DefSimpleRatingSys

/ diagonal=(histogram);

run;



proc sgscatter data=pats;

matrix RegSeasonWins Yards Plays YardsPerPlay Turnovers FumblesLost FirstDowns PassCompletions PassAttempts PassYards

PassTouchdowns PassInterceptions NetYardsPerPass PassFirstDowns RushAttempts

/ diagonal=(histogram);

run;



proc sgscatter data=pats;

matrix RegSeasonWins RushYards RushTouchdowns RushYardsPerAttempt RushFirstDowns Penalties PenaltyYards PenaltyFirstDowns

NumberDrives DriveScorePercent DriveTurnoverPercent AvgStartingPosition

/ diagonal=(histogram);

run;



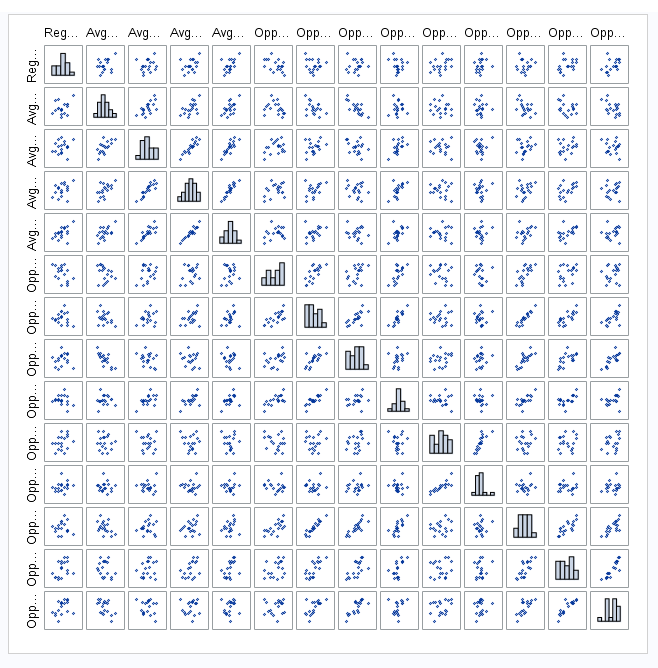
proc sgscatter data=pats;

matrix RegSeasonWins AvgDriveTime AvgDrivePlays AvgDriveYards AvgDrivePoints OppPointsFor OppYards OppPlays OppYardsPerPlay

OppTurnovers OppFumblesLost OppFirstDowns OppPassCompletions OppPassAttempts

/ diagonal=(histogram);

run;



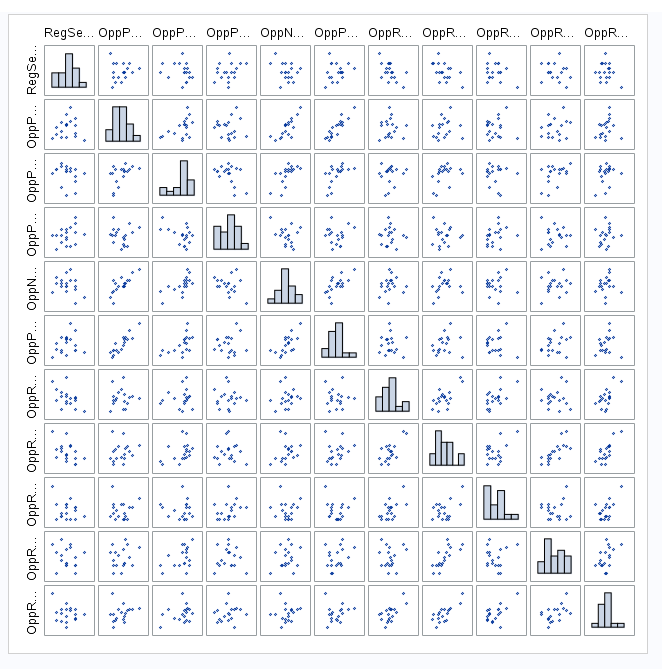
proc sgscatter data=pats;

matrix RegSeasonWins OppPassYards OppPassTouchdowns OppPassInterceptions OppNetYardsPerPass OppPassFirstDowns OppRushAttempts

OppRushYards OppRushTouchdowns OppRushYardsPerAttempt OppRushFirstDowns

/ diagonal=(histogram);

run;



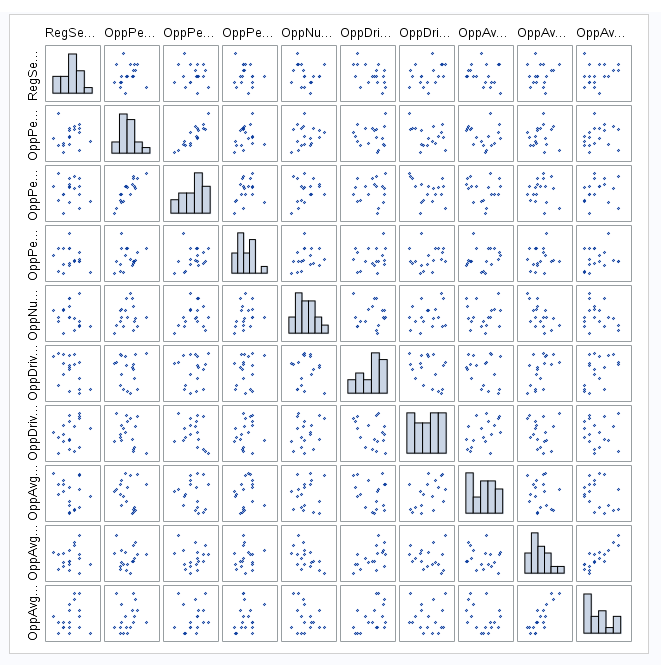
proc sgscatter data=pats;

matrix RegSeasonWins OppPenalties OppPenaltyYards OppPenaltyFirstDowns OppNumberDrives OppDriveScorePercent OppDriveTurnoverPerent

OppAvgStartingPosition OppAvgDriveTime OppAvgDrivePlays

/ diagonal=(histogram);

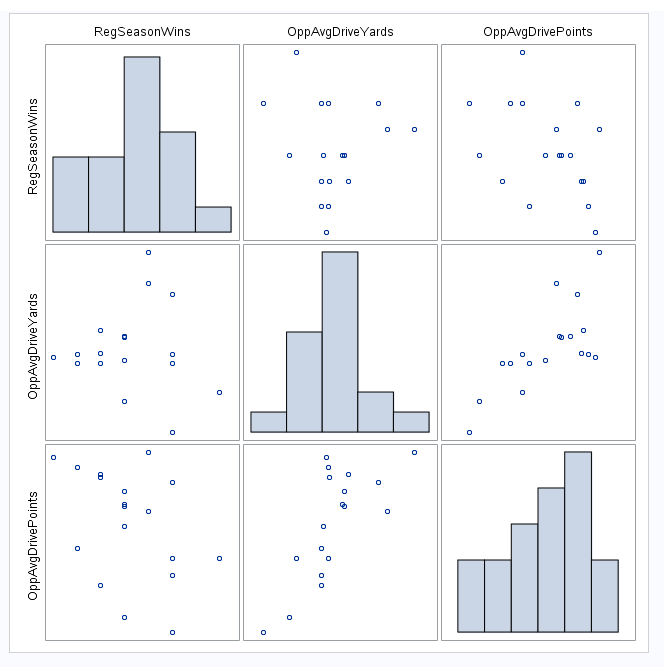
run;



proc sgscatter data=pats;

matrix RegSeasonWins OppAvgDriveYards OppAvgDrivePoints/ diagonal=(histogram);

run;



[3]

/\* Scatterplot matrix of the predictors \*/

proc sgscatter data=pats;  
matrix RegSeasonWins BradyPasserRating PointsFor PointsAgainst PointsDifferential MarginOfVictory SimpleRatingSystem Turnovers FirstDowns   
PassTouchdowns PassInterceptions NetYardsPerPass DriveScorePercent DriveTurnoverPercent AvgDriveTime AvgDrivePoints / diagonal=(histogram);  
run;



[4]

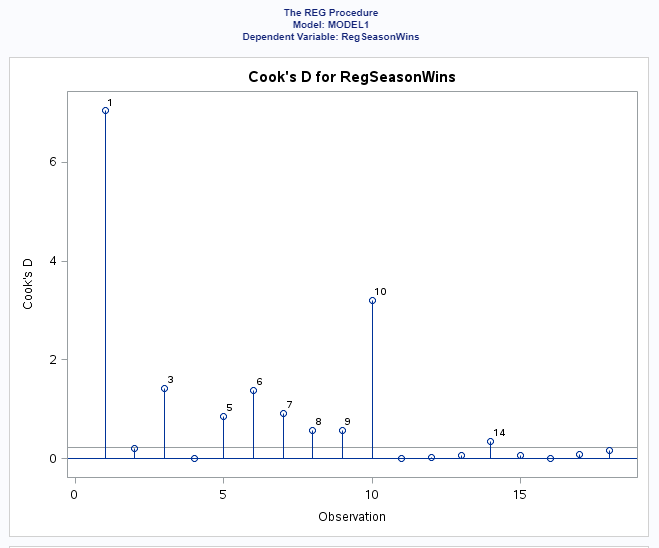
/\* Checking assumptions including outliers and leverage points \*/

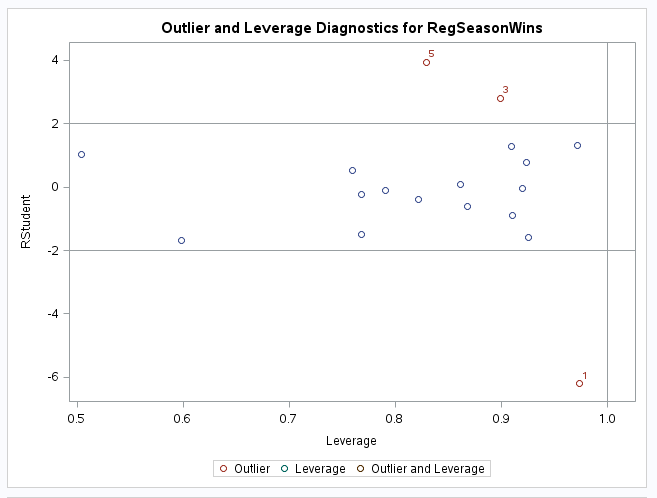
proc reg data=pats plots(labels) = (rstudentleverage cooksd);

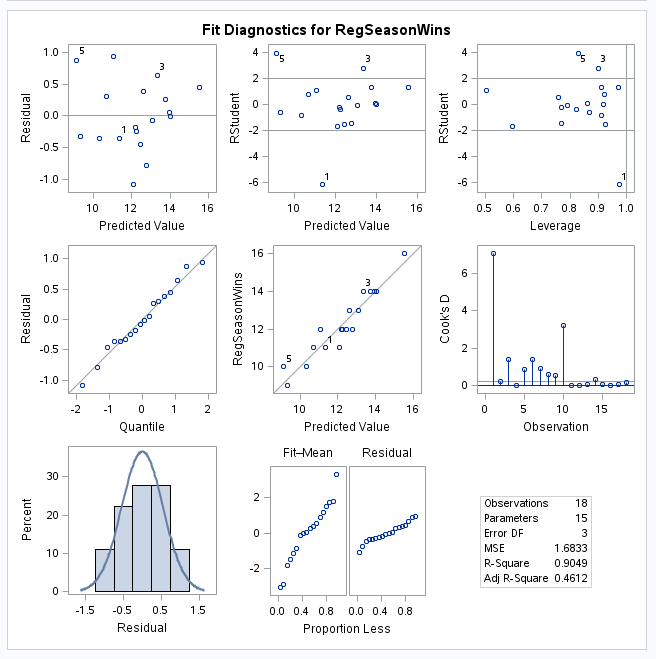
model RegSeasonWins = BradyPasserRating PointsFor PointsAgainst PointsDifferential MarginOfVictory SimpleRatingSystem Turnovers FirstDowns

PassTouchdowns PassInterceptions NetYardsPerPass DriveScorePercent DriveTurnoverPercent AvgDriveTime AvgDrivePoints;

run; quit;







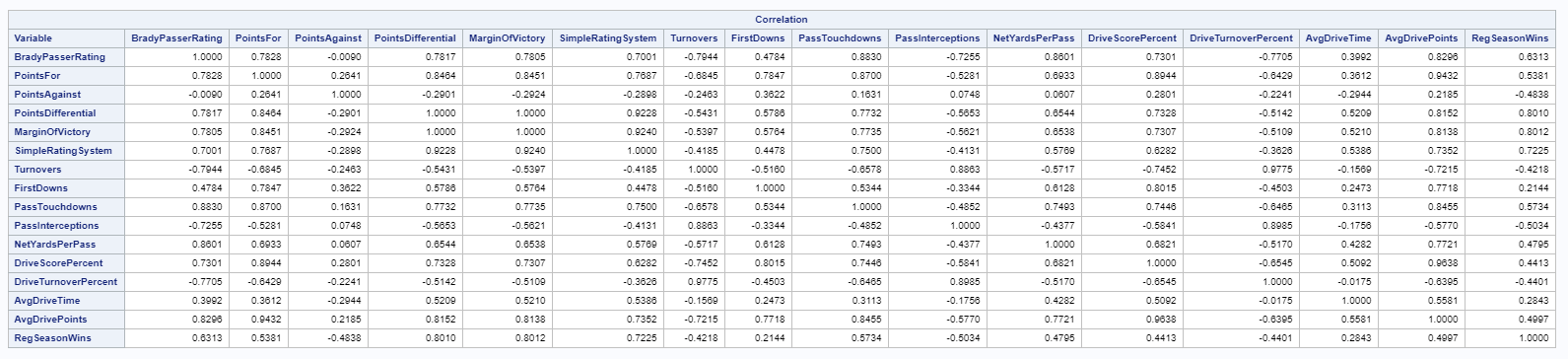
/\* Checking multicollinearity thru VIFs \*/

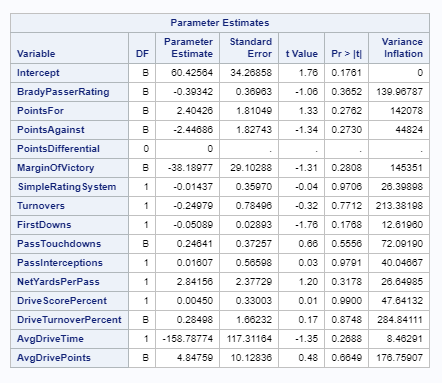
proc reg data=pats corr plots(label)=(rstudentleverage cooksd);

model RegSeasonWins = BradyPasserRating PointsFor PointsAgainst PointsDifferential MarginOfVictory SimpleRatingSystem Turnovers FirstDowns

PassTouchdowns PassInterceptions NetYardsPerPass DriveScorePercent DriveTurnoverPercent AvgDriveTime AvgDrivePoints / VIF;

run; quit;





[5]

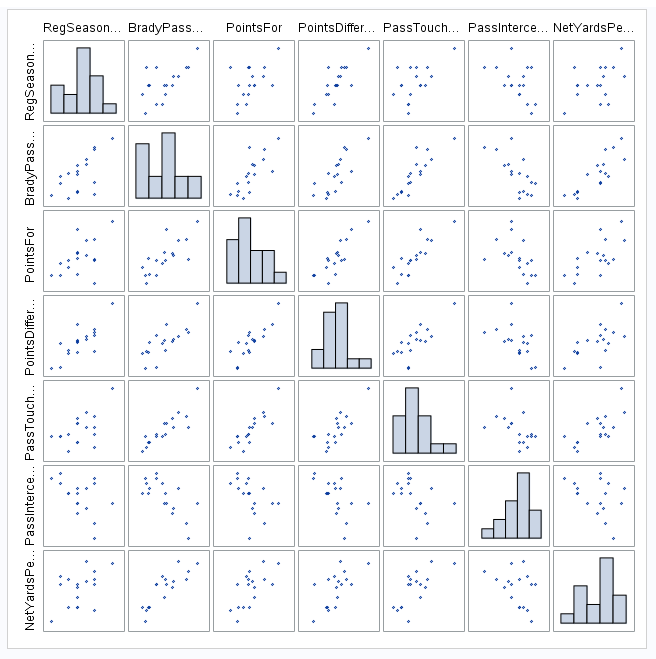
/\* Scatterplot matrix of the predictors minus the low correlation and multicollinear ones \*/

proc sgscatter data=pats;

matrix RegSeasonWins BradyPasserRating PointsFor PointsDifferential

PassTouchdowns PassInterceptions NetYardsPerPass / diagonal=(histogram);

run;



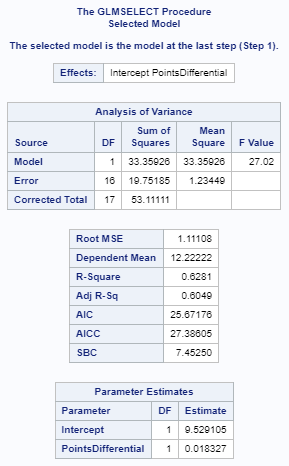
[6]

/\* LARS model \*/

proc GLMSELECT data=pats;

model RegSeasonWins = BradyPasserRating PointsFor PointsDifferential PassTouchdowns PassInterceptions NetYardsPerPass / selection = LARS;

run; quit;

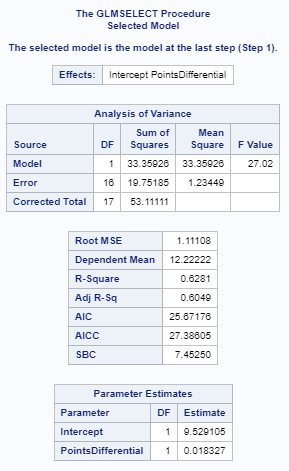


/\* LASSO model \*/

proc GLMSELECT data=pats;

model RegSeasonWins = BradyPasserRating PointsFor PointsDifferential PassTouchdowns PassInterceptions NetYardsPerPass / selection = LASSO;

run; quit;

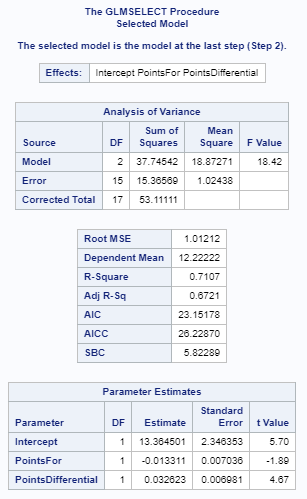


/\* Stepwise model \*/

proc GLMSELECT data=pats;

model RegSeasonWins = BradyPasserRating PointsFor PointsDifferential PassTouchdowns PassInterceptions NetYardsPerPass / selection = stepwise;

run; quit;

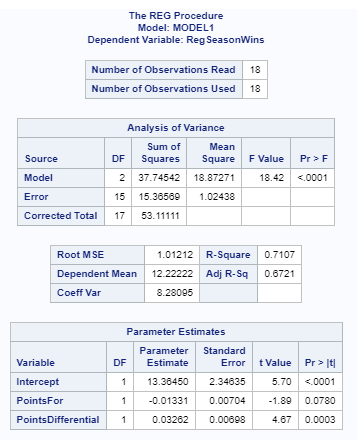


/\* Partial model with PointsFor and PointsDifferential \*/

proc reg data=pats;

model RegSeasonWins = PointsFor PointsDifferential /partial;

run;

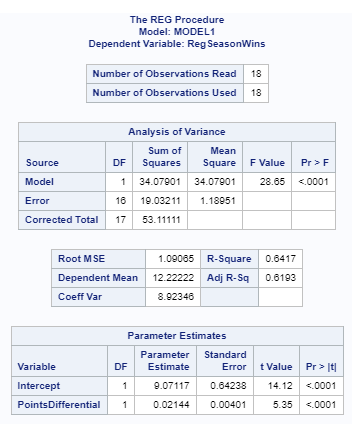


/\* Partial model with PointsDifferential \*/

proc reg data=pats;

model RegSeasonWins = PointsDifferential /partial;

run;

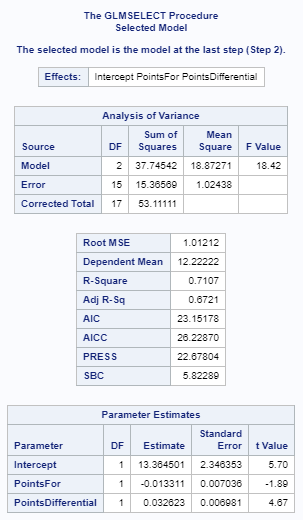


/\* Leave-one-out validation model \*/

proc GLMSELECT data=pats;

model RegSeasonWins = BradyPasserRating PointsFor PointsDifferential PassTouchdowns PassInterceptions NetYardsPerPass / selection=forward(STOP=Press);

run;



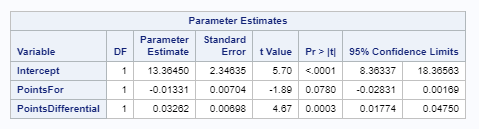
[7]

/\* Confidence intervals for the final model \*/

proc reg data=pats;

model RegSeasonWins = PointsFor PointsDifferential / clb;

run;



[8]

/\* New dataset without outliers \*/

data pats2;

set pats;

if \_n\_=1 then delete;

if \_n\_=2 then delete;

if \_n\_=3 then delete;

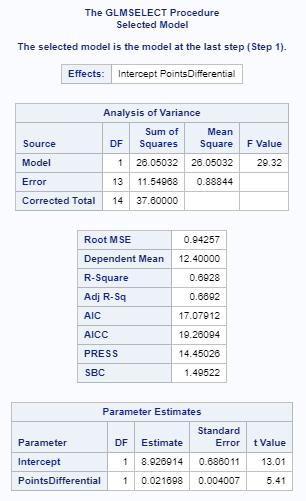
run;

/\* Model for K-fold cross validation (leave-one-out) on the new dataset\*/

proc GLMSELECT data=pats2;

model RegSeasonWins = BradyPasserRating PointsFor PointsDifferential PassTouchdowns PassInterceptions NetYardsPerPass / selection=forward(STOP=Press);

run;



[9]

proc anova data=work.import;

class AgeBin WonLost;

model Rate = AgeBin WonLost;

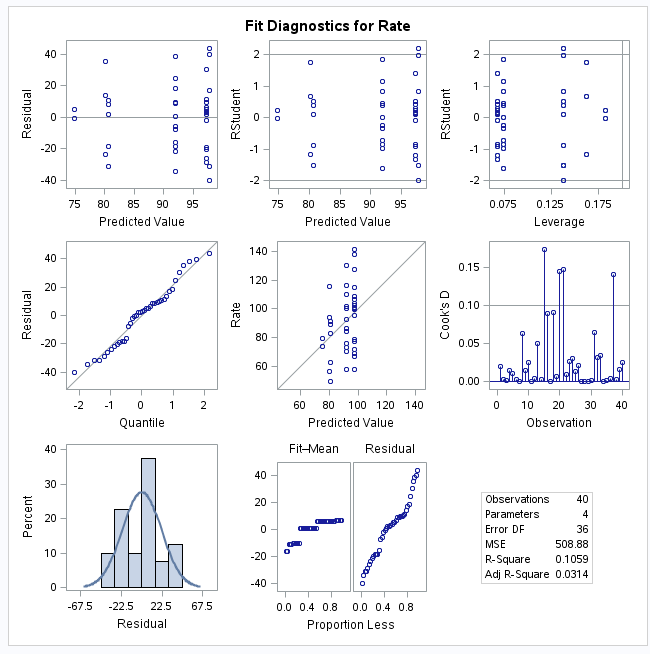
run;

proc glm data=work.import plots=all;

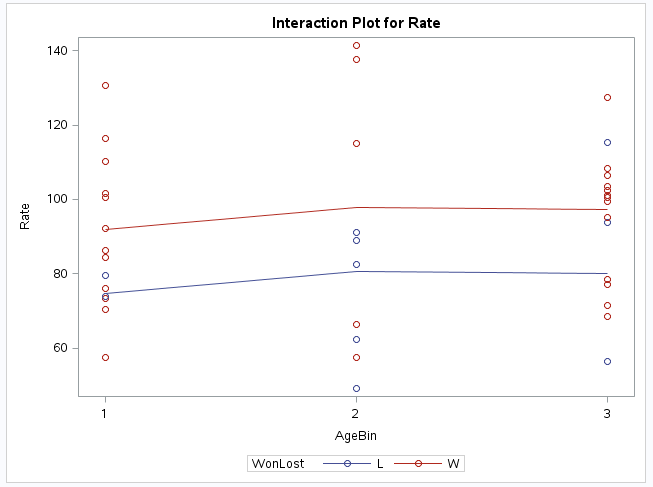
class AgeBin WonLost;

model Rate = AgeBin WonLost / clm;

run;







[10]

proc anova data = work.import;

class WonLost;

model Rate = WonLost;

run;

proc glm data = work.import;

class WonLost;

model Rate = WonLost;

run;

