**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 statistics explain the number of regular season wins for the Patriots since Tom Brady has been playing on the team (last 18 seasons).

2) \*\*\*whatever we are doing for our two-way ANOVA\*\*\*

**DATA DESCRIPTION:**

Our dataset includes all offensive and defensive stats (70 variables) for each of the Patriots’ past 18 seasons (2001 to 2018). Each season is one observation.

**EXPLORATORY ANALYSIS:**

We used SAS to complete our first objective and R to complete our second.

**OBJECTIVE 1:**

**Problem:**

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

**Overall Approach:**

We are going to determine which variables are predictors for regular season wins. Then, we are going to create a multiple regression model to predict regular season wins.

**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 see which if any variables showed a correlation with regular season wins.[2] The reason why we created multiple matrices is that there were 70 variables, and nothing would show if we created one single matrix. We determined that 16 variables showed a correlation with regular season wins. (None of the variables had to be transformed.) We created a scatterplot matrix with these 16 variables.[3]



**Checking Assumptions:**

multiple regression assumptions:

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

constant variance

observations are independent

multicollinearity (check thru VIFs)

outliers and leverage (check thru residual diagnostics)

**Model Selection:**

LARS

LASSO

stepwise

leave-one-out cross validation is K-fold cross validation taken to its logical extreme

**Parameter Interpretation:**

RegSeasonWins = 8.567177 + 0.023588(PointsDifferntial)

**Interpretation:**

If the New England Patriots finish the regular season with a 0 point differential, they will win 8.567177 games. 315.11 PD

**Confidence Intervals:**

95% confidence limits intercept 7.09350 10.04086

95% confidence limits pointsdifferential 0.01498 0.03220

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

Bill Belichick is the real goat, not Tom Brady.

**OBJECTIVE 2:**

**Problem:**

After exploring what regular season stats contributed to the Patriots success over the last 19 seasons, we elected to find out if there is any difference in Tom Brady’s passer rating in playoff games that they won, and as he has aged over time.

**Overall Approach:**

Since we are using a two way Anova, we decided to bin every game by its end results (win or loss) and his age into one of 3 bins (24-30,30-36, and 36-42). In these 3 different age bins, we are looking to see if Brady’s passer rating was higher in games that the Patriots won or 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, interceptions because the Passer rating stat encompasses all of those metrics. This will also keep the covariance between variables from being a problem in our analysis.

**Two Way Anova Interpretation**

When we run the model

Rate = AgeBin +WonLost,

The p-value for AgeBin is 0.77, which indicates that there is no difference in levels when it comes to Tom Brady’s age. The p-value for WonLost however is 0.0578, and although that is usually too high to be considered significant, since we have such a small sample size and are unlikely to see another quarterback accompany a franchise in succeeding as frequently as the Patriots have, we have decided to consider this variable significant, and run a one way Anova test with the lone variable as WonLost. We also decided to use the type III sum of squares value as opposed to the type I sum of squares value since interactions are not present, and we are treating one of the levels as significant. It really does not matter however, since the p-value for WonLost are the same across different types and we are dropping the AgeBin for further one way Anova analysis.

**One Way Anova Interpretation**

The model

Rate = WonLost

The p-value for the WonLost variable drops a little further to 0.0561, further indicating that there is some significant difference in Tom Brady’s passer rating and in turn his playoff performance in games that the Patriots win and games that the Patriots lose in the Postseason.

**Conclusion**

As we can see in the interaction plot in section 4 of the appendix, there was a slight increase in Brady’s passer rating as he progressed through each of the three age bins, we also see the average difference between games the Patriots won and lost. We also see in the QQplot and histogram evidence of normality in the dataset and our residual plot also looks good in terms of spread.

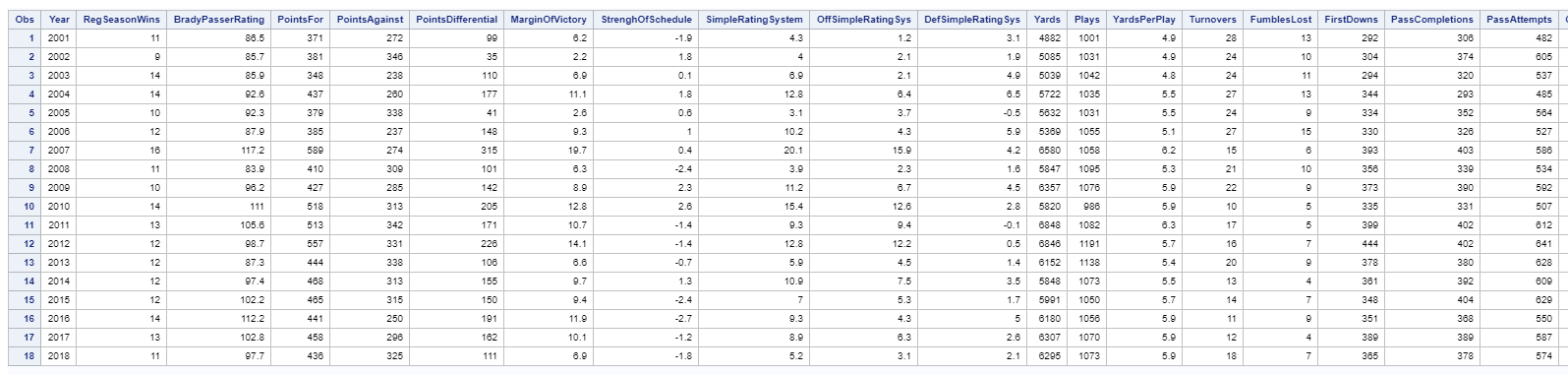
Section 5 of the appendix has some graphics regarding the one way Anova performed after age is taken out of the dataset. The box and whisker plot shows the evidence that Tom Brady’s performance does have a significant effect on their results in playoff games. This kind of strays from the evidence we saw in part one of our analysis which focused on the Patriots and Tom Brady’s regular season success. We were unable to find evidence that Tom Brady really had an effect in the regular season, and thus we attribute that success to Bill Belichick’s coaching. However, when the Patriots get to the post season, we see that for the Patriots to make the run deep in the playoffs and get to the Super Bowl, Tom Brady has to be at his best. The last graphic in our appendix has a series of confidence intervals, for age 24-30 we see his rating from 79.4 to 104.4 in games they won and 55.3 to 94.3 in games they lost. In age group 30-36, his rating in wins is 80.8 to 114.7 whereas when they lose it drops to 66.6 to 97.6. In the final age group 36-42, in wins the rating interval is 85.4 to 109.2 and in losses it again drops to 61.8 to 98.6. In every age bin there is a jump between games the Patriots won and lost depending on Brady’s passer rating. In conclusion, although Bill Belichick is the greatest of all time when it comes to putting his team in a playoff spot and in position to win a title, the Patriots need and 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;

proc print data=pats;  
run;



[2]

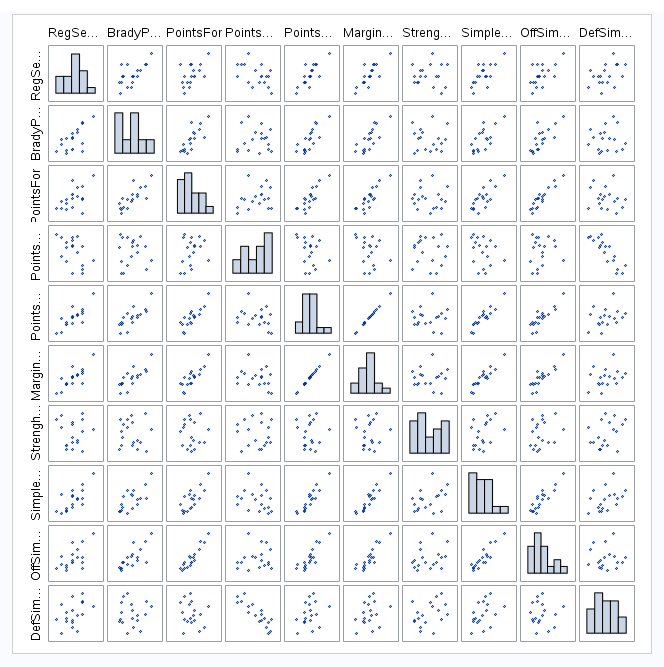
proc sgscatter data=pats;

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;



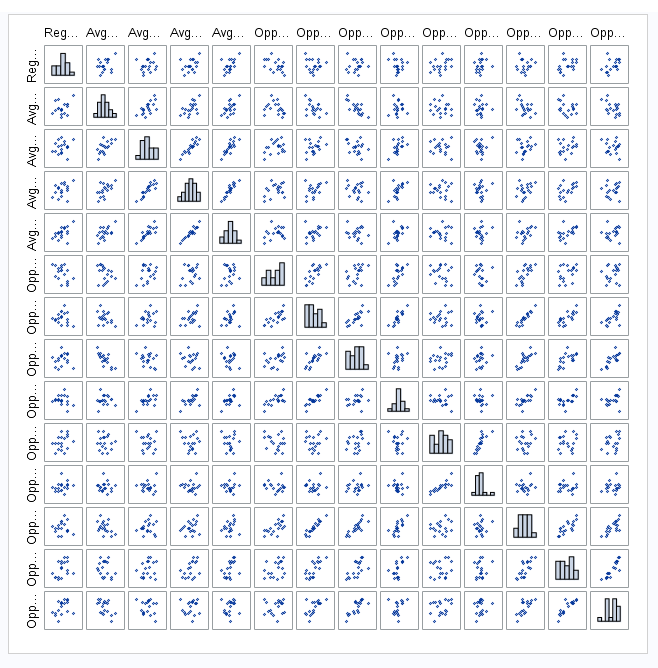
proc sgscatter data=pats;

matrix RegSeasonWins AvgDriveTime AvgDrivePlays AvgDriveYards AvgDrivePoints OppPointsFor OppYards OppPlays OppYardsPerPlay

OppTurnovers OppFumblesLost OppFirstDowns OppPassCompletions OppPassAttempts

/ diagonal=(histogram);

run;



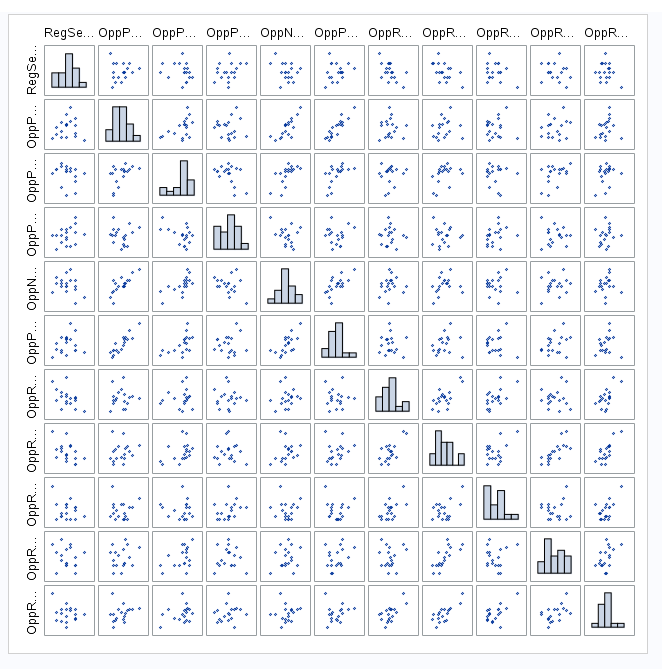
proc sgscatter data=pats;

matrix RegSeasonWins OppPassYards OppPassTouchdowns OppPassInterceptions OppNetYardsPerPass OppPassFirstDowns OppRushAttempts

OppRushYards OppRushTouchdowns OppRushYardsPerAttempt OppRushFirstDowns

/ diagonal=(histogram);

run;



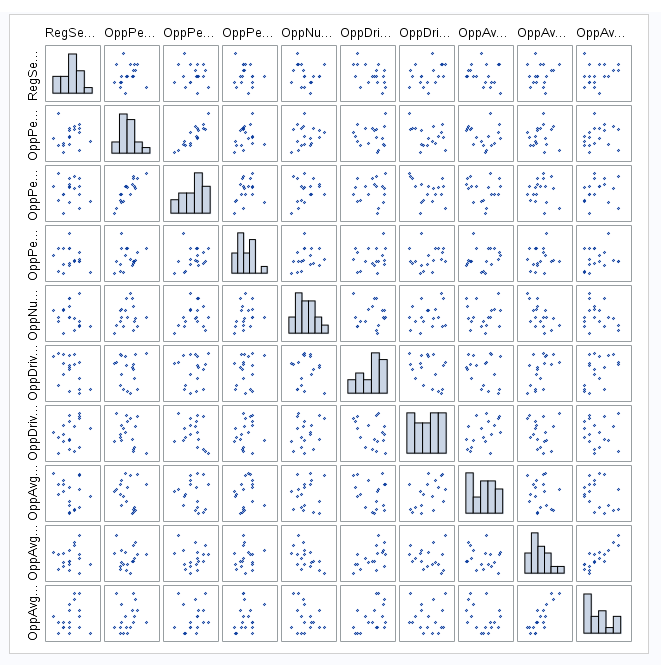
proc sgscatter data=pats;

matrix RegSeasonWins OppPenalties OppPenaltyYards OppPenaltyFirstDowns OppNumberDrives OppDriveScorePercent OppDriveTurnoverPerent

OppAvgStartingPosition OppAvgDriveTime OppAvgDrivePlays

/ diagonal=(histogram);

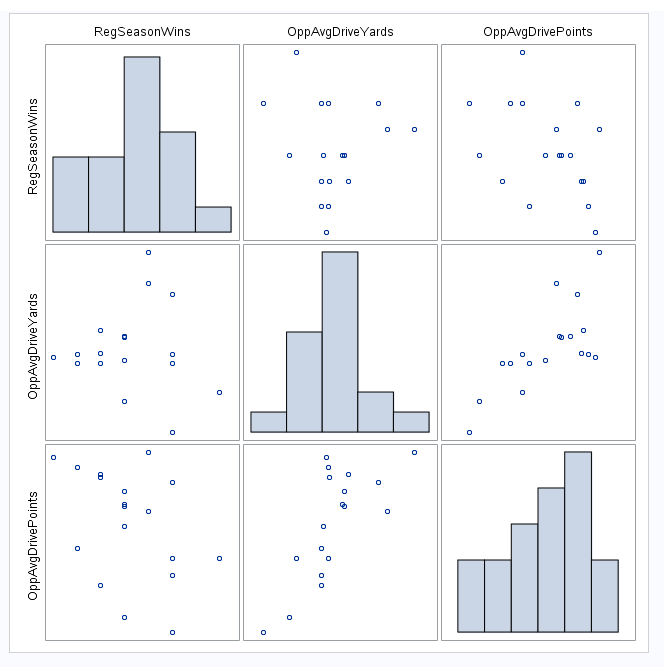
run;



proc sgscatter data=pats;

matrix RegSeasonWins OppAvgDriveYards OppAvgDrivePoints/ diagonal=(histogram);

run;



[3]

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]

proc anova data=work.import;

class AgeBin WonLost;

model Rate = AgeBin WonLost;

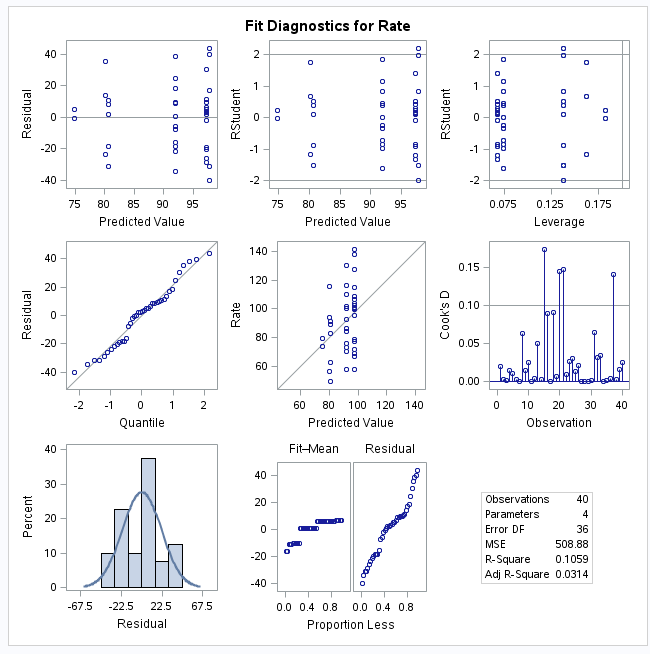
run;

proc glm data=work.import plots=all;

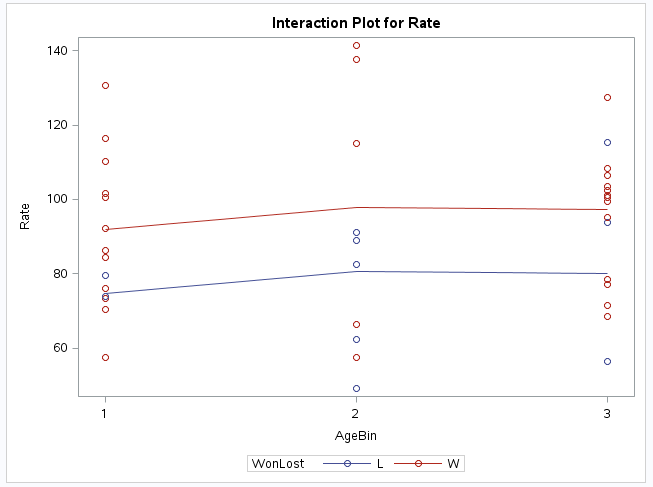
class AgeBin WonLost;

model Rate = AgeBin WonLost / clm;

run;







[5]

proc anova data = work.import;

class WonLost;

model Rate = WonLost;

run;

proc glm data = work.import;

class WonLost;

model Rate = WonLost;

run;

