B. Brewer, D. Nguyen, A. Scheerer **February 16, 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. 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 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 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.

<u>OBJECTIVE 1:</u>

Problem:

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

Overall Approach:

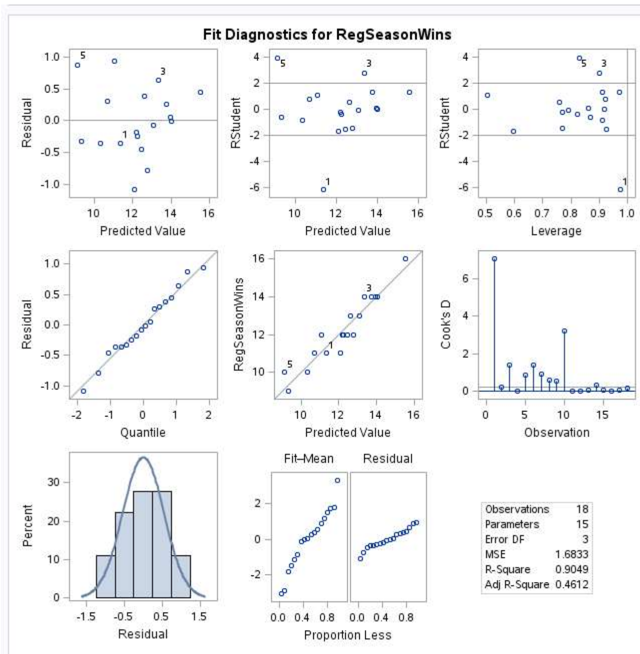
Using the aggregated regular season statistics, we will determine which variables correlate with the New England Patriots' regular season record and subsequently build a model using those parameters to interpret the relationship between those parameters and the New England Patriots' regular season record.

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:

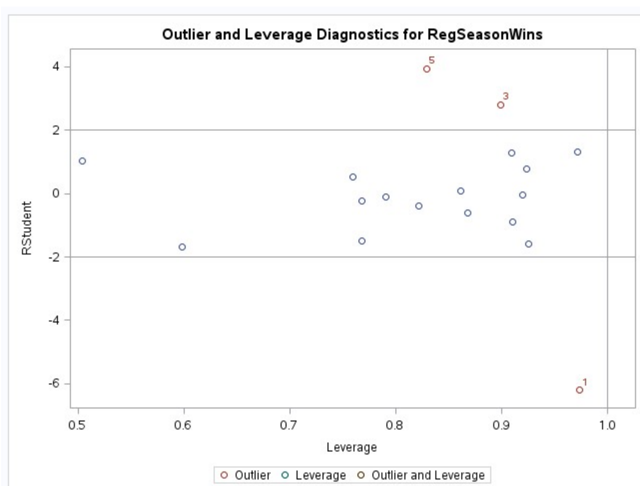
We then checked to see if the assumptions for multiple regression were met:[4]



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.

Correlation

Variable	BradyPasserRating	PointsFor	PointsAgainst	PointsDifferential	MarginOfVictory	SimpleRatingSystem	Turnovers	FirstDowns
BradyPasserRating	1.0000	0.7828	-0.0090	0.7817	0.7805	0.7001	-0.7944	0.4784
PointsFor	0.7828	1.0000	0.2641	0.8464	0.8451	0.7687	-0.6845	0.7847
PointsAgainst	-0.0090	0.2641	1.0000	-0.2901	-0.2924	-0.2898	-0.2463	0.3622
PointsDifferential	0.7817	0.8464	-0.2901	1.0000	1.0000	0.9228	-0.5431	0.5786
MarginOfVictory	0.7805	0.8451	-0.2924	1.0000	1.0000	0.9240	-0.5397	0.5764
SimpleRatingSystem	0.7001	0.7687	-0.2898	0.9228	0.9240	1.0000	-0.4185	0.4478
Turnovers	-0.7944	-0.6845	-0.2463	-0.5431	-0.5397	-0.4185	1.0000	-0.5160
FirstDowns	0.4784	0.7847	0.3622	0.5786	0.5764	0.4478	-0.5160	1.0000
PassTouchdowns	0.8830	0.8700	0.1631	0.7732	0.7735	0.7500	-0.6578	0.5344
PassInterceptions	-0.7255	-0.5281	0.0748	-0.5653	-0.5621	-0.4131	0.8863	-0.3344
NetYardsPerPass	0.8601	0.6933	0.0607	0.6544	0.6538	0.5769	-0.5717	0.6128
DriveScorePercent	0.7301	0.8944	0.2801	0.7328	0.7307	0.6282	-0.7452	0.8015
DriveTurnoverPercent	-0.7705	-0.6429	-0.2241	-0.5142	-0.5109	-0.3626	0.9775	-0.4503
AvgDriveTime	0.3992	0.3612	-0.2944	0.5209	0.5210	0.5386	-0.1569	0.2473
AvgDrivePoints	0.8296	0.9432	0.2185	0.8152	0.8138	0.7352	-0.7215	0.7718
RegSeasonWins	0.6313	0.5381	-0.4838	0.8010	0.8012	0.7225	-0.4218	0.2144

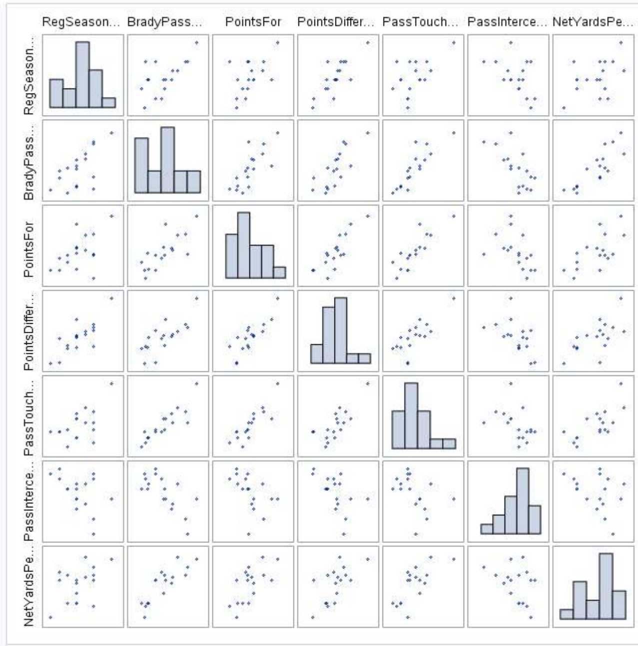
Variable	PassTouchdowns	PassInterceptions	NetYardsPerPass	DriveScorePercent	DriveTurnoverPercent	AvgDriveTime	AvgDrivePoints	RegSeasonWins
BradyPasserRating	0.8830	-0.7255	0.8601	0.7301	-0.7705	0.3992	0.8296	0.6313
PointsFor	0.8700	-0.5281	0.6933	0.8944	-0.6429	0.3612	0.9432	0.5381
PointsAgainst	0.1631	0.0748	0.0607	0.2801	-0.2241	-0.2944	0.2185	-0.4838
PointsDifferential	0.7732	-0.5653	0.6544	0.7328	-0.5142	0.5209	0.8152	0.8010
MarginOfVictory	0.7735	-0.5621	0.6538	0.7307	-0.5109	0.5210	0.8138	0.8012
SimpleRatingSystem	0.7500	-0.4131	0.5769	0.6282	-0.3626	0.5386	0.7352	0.7225
Turnovers	-0.6578	0.8863	-0.5717	-0.7452	0.9775	-0.1569	-0.7215	-0.4218
FirstDowns	0.5344	-0.3344	0.6128	0.8015	-0.4503	0.2473	0.7718	0.2144
PassTouchdowns	1.0000	-0.4852	0.7493	0.7446	-0.6465	0.3113	0.8455	0.5734
PassInterceptions	-0.4852	1.0000	-0.4377	-0.5841	0.8985	-0.1756	-0.5770	-0.5034
NetYardsPerPass	0.7493	-0.4377	1.0000	0.6821	-0.5170	0.4282	0.7721	0.4795
DriveScorePercent	0.7446	-0.5841	0.6821	1.0000	-0.6545	0.5092	0.9638	0.4413
DriveTurnoverPercent	-0.6465	0.8985	-0.5170	-0.6545	1.0000	-0.0175	-0.6395	-0.4401
AvgDriveTime	0.3113	-0.1756	0.4282	0.5092	-0.0175	1.0000	0.5581	0.2843
AvgDrivePoints	0.8455	-0.5770	0.7721	0.9638	-0.6395	0.5581	1.0000	0.4997
RegSeasonWins	0.5734	-0.5034	0.4795	0.4413	-0.4401	0.2843	0.4997	1.0000

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	B	60.42564	34.26858	1.76	0.1761	0
BradyPasserRating	B	-0.39342	0.36963	-1.06	0.3652	139.96787
PointsFor	B	2.40426	1.81049	1.33	0.2762	142078
PointsAgainst	B	-2.44686	1.82743	-1.34	0.2730	44824
PointsDifferential	0	0	.	.	.	.
MarginOfVictory	B	-38.18977	29.10288	-1.31	0.2808	145351
SimpleRatingSystem	1	-0.01437	0.35970	-0.04	0.9706	26.39898
Turnovers	1	-0.24979	0.78496	-0.32	0.7712	213.38198
FirstDowns	1	-0.05089	0.02893	-1.76	0.1768	12.61960
PassTouchdowns	B	0.24641	0.37257	0.66	0.5556	72.09190
PassInterceptions	1	0.01607	0.56598	0.03	0.9791	40.04667
NetYardsPerPass	1	2.84156	2.37729	1.20	0.3178	26.64985
DriveScorePercent	1	0.00450	0.33003	0.01	0.9900	47.64132
DriveTurnoverPercent	B	0.28498	1.66232	0.17	0.8748	284.84111
AvgDriveTime	1	-158.78774	117.31164	-1.35	0.2688	8.46291
AvgDrivePoints	B	4.84759	10.12836	0.48	0.6649	176.75907

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:

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 2).

Effects:	Intercept PointsFor PointsDifferential

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	2	37.74542	18.87271	18.42
Error	15	15.36569	1.02438	
Corrected Total	17	53.11111		

Root MSE	1.01212
Dependent Mean	12.22222
R-Square	0.7107
Adj R-Sq	0.6721
AIC	23.15178
AICC	26.22870
PRESS	22.67804
SBC	5.82289

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	13.364501	2.346353	5.70
PointsFor	1	-0.013311	0.007036	-1.89
PointsDifferential	1	0.032623	0.006981	4.67

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]

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	13.36450	2.34635	5.70	<.0001	8.36337	18.36563
PointsFor	1	-0.01331	0.00704	-1.89	0.0780	-0.02831	0.00169
PointsDifferential	1	0.03262	0.00698	4.67	0.0003	0.01774	0.04750

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:

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 1).

Effects:	Intercept PointsDifferential

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	1	26.05032	26.05032	29.32
Error	13	11.54968	0.88844	
Corrected Total	14	37.60000		

Root MSE	0.94257
Dependent Mean	12.40000
R-Square	0.6928
Adj R-Sq	0.6692
AIC	17.07912
AICC	19.26094
PRESS	14.45026
SBC	1.49522

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	8.926914	0.686011	13.01
PointsDifferential	1	0.021698	0.004007	5.41

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;
```

Obs	Year	RegSeasonWins	BradyPasserRating	PointsFor	PointsAgainst	PointsDifferential	MarginOfVictory	StrenghOfSchedule	SimpleRatingSystem
1	2001	11	86.5	371	272	99	6.2	-1.9	4.3
2	2002	9	85.7	381	346	35	2.2	1.8	4
3	2003	14	85.9	348	238	110	6.9	0.1	6.9
4	2004	14	92.6	437	260	177	11.1	1.8	12.8
5	2005	10	92.3	379	338	41	2.6	0.6	3.1
6	2006	12	87.9	385	237	148	9.3	1	10.2
7	2007	16	117.2	589	274	315	19.7	0.4	20.1
8	2008	11	83.9	410	309	101	6.3	-2.4	3.9
9	2009	10	96.2	427	285	142	8.9	2.3	11.2
10	2010	14	111	518	313	205	12.8	2.6	15.4
11	2011	13	105.6	513	342	171	10.7	-1.4	9.3
12	2012	12	98.7	557	331	226	14.1	-1.4	12.8
13	2013	12	87.3	444	338	106	6.6	-0.7	5.9
14	2014	12	97.4	468	313	155	9.7	1.3	10.9
15	2015	12	102.2	465	315	150	9.4	-2.4	7
16	2016	14	112.2	441	250	191	11.9	-2.7	9.3
17	2017	13	102.8	458	296	162	10.1	-1.2	8.9
18	2018	11	97.7	436	325	111	6.9	-1.8	5.2

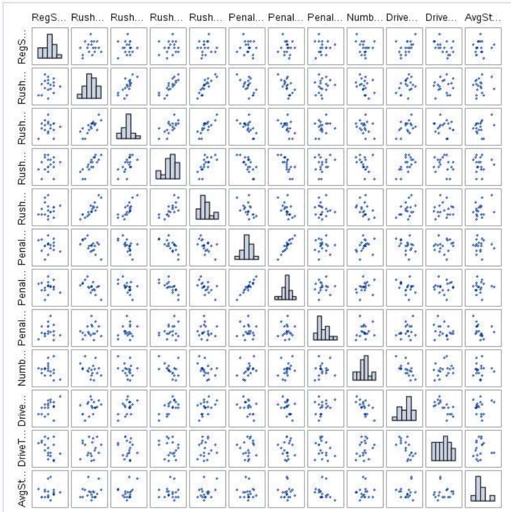
Obs	OffSimpleRatingSys	DefSimpleRatingSys	Yards	Plays	YardsPerPlay	Turnovers	FumblesLost	FirstDowns	PassCompletions	PassAttempts
1	1.2	3.1	4882	1001	4.9	28	13	292	306	482
2	2.1	1.9	5085	1031	4.9	24	10	304	374	605
3	2.1	4.9	5039	1042	4.8	24	11	294	320	537
4	6.4	6.5	5722	1035	5.5	27	13	344	293	485
5	3.7	-0.5	5632	1031	5.5	24	9	334	352	564
6	4.3	5.9	5369	1055	5.1	27	15	330	326	527
7	15.9	4.2	6580	1058	6.2	15	6	393	403	586
8	2.3	1.6	5847	1095	5.3	21	10	356	339	534
9	6.7	4.5	6357	1076	5.9	22	9	373	390	592
10	12.6	2.8	5820	986	5.9	10	5	335	331	507
11	9.4	-0.1	6848	1082	6.3	17	5	399	402	612
12	12.2	0.5	6846	1191	5.7	16	7	444	402	641
13	4.5	1.4	6152	1138	5.4	20	9	378	380	628
14	7.5	3.5	5848	1073	5.5	13	4	361	392	609
15	5.3	1.7	5991	1050	5.7	14	7	348	404	629
16	4.3	5	6180	1056	5.9	11	9	351	368	550
17	6.3	2.6	6307	1070	5.9	12	4	389	389	587
18	3.1	2.1	6295	1073	5.9	18	7	365	378	574

[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;
```

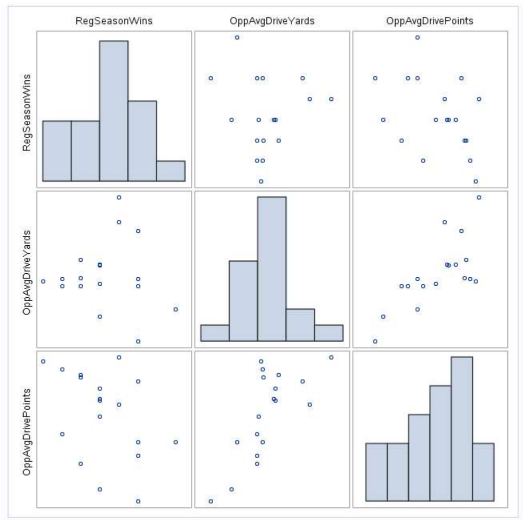


```
proc sgscatter data=pats;
matrix RegSeasonWins AvgDriveTime AvgDrivePlays AvgDriveYards AvgDrivePoints OppPointsFor
OppYards OppPlays OppYardsPerPlay
OppTurnovers OppFumblesLost OppFirstDowns OppPassCompletions OppPassAttempts
/ diagonal=(histogram);
run;
```

```
proc sgscatter data=pats;
matrix RegSeasonWins OppPassYards OppPassTouchdowns OppPassInterceptions OppNetYardsPerPass
OppPassFirstDowns OppRushAttempts
OppRushYards OppRushTouchdowns OppRushYardsPerAttempt OppRushFirstDowns
/ diagonal=(histogram);
run;
```

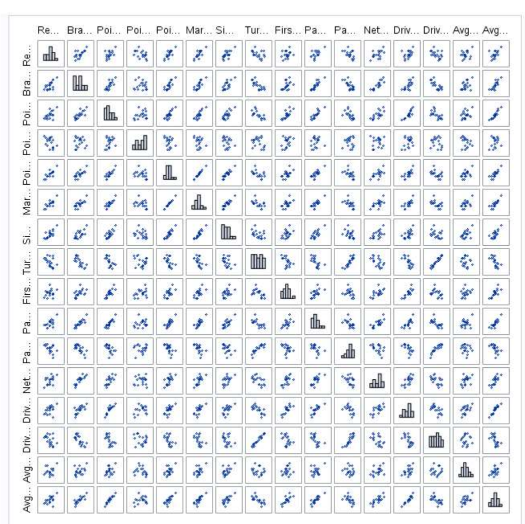
```
proc sgscatter data=pats;
matrix RegSeasonWins OppPenalties OppPenaltyYards OppPenaltyFirstDowns OppNumberDrives
OppDriveScorePercent OppDriveTurnoverPerent
OppAvgStartingPosition OppAvgDriveTime OppAvgDrivePlays
/ diagonal=(histogram);
run;
```

```
proc sgscatter data=pats;
matrix RegSeasonWins OppAvgDriveYards OppAvgDrivePoints/ diagonal=(histogram);
run;
```



[3]

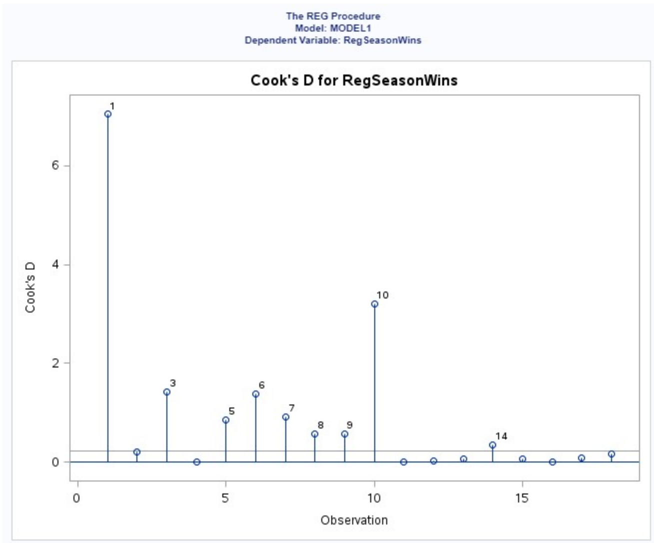
```
/* Scatterplot matrix of the predictors */
proc sgscatter data=pats;
matrix RegSeasonWins BradyPasserRating PointsFor PointsAgainst PointsDifferential MarginOfVictory
SimpleRatingSystem Turnovers FirstDowns
PassTouchdowns PassInterceptions NetYardsPerPass DriveScorePercent DriveTurnoverPercent
AvgDriveTime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

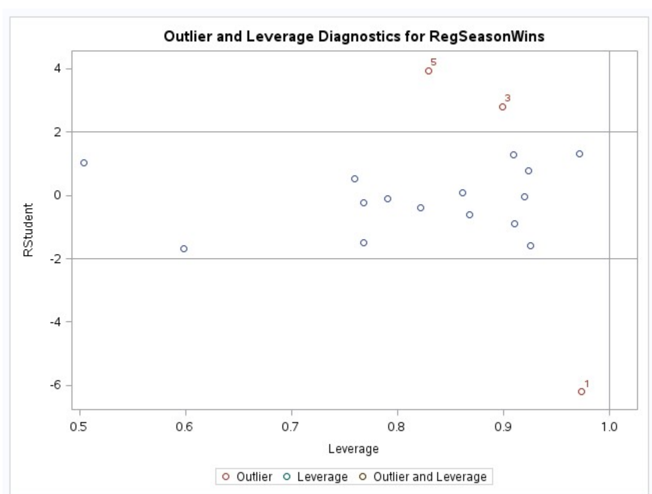


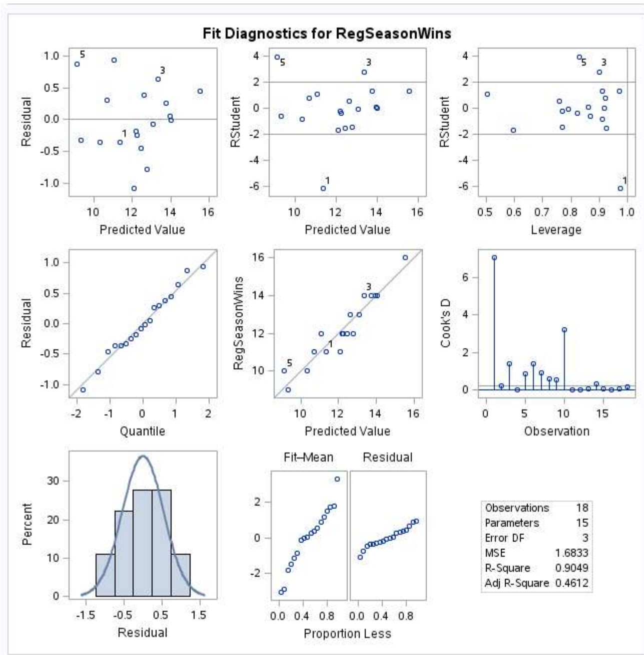
[4]

```
/* 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;
```

Correlation

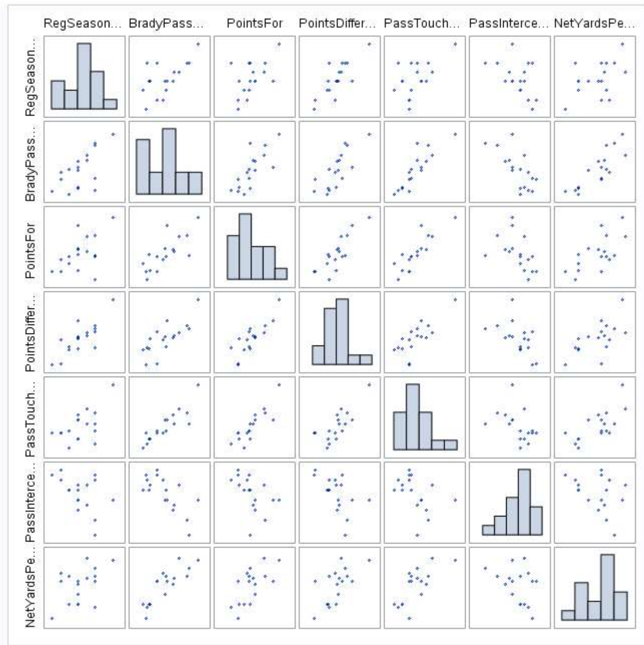
Variable	BradyPasserRating	PointsFor	PointsAgainst	PointsDifferential	MarginOfVictory	SimpleRatingSystem	Turnovers	FirstDowns
BradyPasserRating	1.0000	0.7828	-0.0090	0.7817	0.7805	0.7001	-0.7944	0.4784
PointsFor	0.7828	1.0000	0.2641	0.8464	0.8451	0.7687	-0.6845	0.7847
PointsAgainst	-0.0090	0.2641	1.0000	-0.2901	-0.2924	-0.2898	-0.2463	0.3622
PointsDifferential	0.7817	0.8464	-0.2901	1.0000	1.0000	0.9228	-0.5431	0.5786
MarginOfVictory	0.7805	0.8451	-0.2924	1.0000	1.0000	0.9240	-0.5397	0.5764
SimpleRatingSystem	0.7001	0.7687	-0.2898	0.9228	0.9240	1.0000	-0.4185	0.4478
Turnovers	-0.7944	-0.6845	-0.2463	-0.5431	-0.5397	-0.4185	1.0000	-0.5160
FirstDowns	0.4784	0.7847	0.3622	0.5786	0.5764	0.4478	-0.5160	1.0000
PassTouchdowns	0.8830	0.8700	0.1631	0.7732	0.7735	0.7500	-0.6578	0.5344
PassInterceptions	-0.7255	-0.5281	0.0748	-0.5653	-0.5621	-0.4131	0.8863	-0.3344
NetYardsPerPass	0.8601	0.6933	0.0607	0.6544	0.6538	0.5769	-0.5717	0.6128
DriveScorePercent	0.7301	0.8944	0.2801	0.7328	0.7307	0.6282	-0.7452	0.8015
DriveTurnoverPercent	-0.7705	-0.6429	-0.2241	-0.5142	-0.5109	-0.3626	0.9775	-0.4503
AvgDriveTime	0.3992	0.3612	-0.2944	0.5209	0.5210	0.5386	-0.1569	0.2473
AvgDrivePoints	0.8296	0.9432	0.2185	0.8152	0.8138	0.7352	-0.7215	0.7718
RegSeasonWins	0.6313	0.5381	-0.4838	0.8010	0.8012	0.7225	-0.4218	0.2144

Variable	PassTouchdowns	PassInterceptions	NetYardsPerPass	DriveScorePercent	DriveTurnoverPercent	AvgDriveTime	AvgDrivePoints	RegSeasonWins
BradyPasserRating	0.8830	-0.7255	0.8601	0.7301	-0.7705	0.3992	0.8296	0.6313
PointsFor	0.8700	-0.5281	0.6933	0.8944	-0.6429	0.3612	0.9432	0.5381
PointsAgainst	0.1631	0.0748	0.0607	0.2801	-0.2241	-0.2944	0.2185	-0.4838
PointsDifferential	0.7732	-0.5653	0.6544	0.7328	-0.5142	0.5209	0.8152	0.8010
MarginOfVictory	0.7735	-0.5621	0.6538	0.7307	-0.5109	0.5210	0.8138	0.8012
SimpleRatingSystem	0.7500	-0.4131	0.5769	0.6282	-0.3626	0.5386	0.7352	0.7225
Turnovers	-0.6578	0.8863	-0.5717	-0.7452	0.9775	-0.1569	-0.7215	-0.4218
FirstDowns	0.5344	-0.3344	0.6128	0.8015	-0.4503	0.2473	0.7718	0.2144
PassTouchdowns	1.0000	-0.4852	0.7493	0.7446	-0.6465	0.3113	0.8455	0.5734
PassInterceptions	-0.4852	1.0000	-0.4377	-0.5841	0.8985	-0.1756	-0.5770	-0.5034
NetYardsPerPass	0.7493	-0.4377	1.0000	0.6821	-0.5170	0.4282	0.7721	0.4795
DriveScorePercent	0.7446	-0.5841	0.6821	1.0000	-0.6545	0.5092	0.9638	0.4413
DriveTurnoverPercent	-0.6465	0.8985	-0.5170	-0.6545	1.0000	-0.0175	-0.6395	-0.4401
AvgDriveTime	0.3113	-0.1756	0.4282	0.5092	-0.0175	1.0000	0.5581	0.2843
AvgDrivePoints	0.8455	-0.5770	0.7721	0.9638	-0.6395	0.5581	1.0000	0.4997
RegSeasonWins	0.5734	-0.5034	0.4795	0.4413	-0.4401	0.2843	0.4997	1.0000

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	B	60.42564	34.26858	1.76	0.1761	0
BradyPasserRating	B	-0.39342	0.36963	-1.06	0.3652	139.96787
PointsFor	B	2.40426	1.81049	1.33	0.2762	142078
PointsAgainst	B	-2.44686	1.82743	-1.34	0.2730	44824
PointsDifferential	0	0	.	.	.	.
MarginOfVictory	B	-38.18977	29.10288	-1.31	0.2808	145351
SimpleRatingSystem	1	-0.01437	0.35970	-0.04	0.9706	26.39898
Turnovers	1	-0.24979	0.78496	-0.32	0.7712	213.38198
FirstDowns	1	-0.05089	0.02893	-1.76	0.1768	12.61960
PassTouchdowns	B	0.24641	0.37257	0.66	0.5556	72.09190
PassInterceptions	1	0.01607	0.56598	0.03	0.9791	40.04667
NetYardsPerPass	1	2.84156	2.37729	1.20	0.3178	26.64985
DriveScorePercent	1	0.00450	0.33003	0.01	0.9900	47.64132
DriveTurnoverPercent	B	0.28498	1.66232	0.17	0.8748	284.84111
AvgDriveTime	1	-158.78774	117.31164	-1.35	0.2688	8.46291
AvgDrivePoints	B	4.84759	10.12836	0.48	0.6649	176.75907

[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;
```

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 1).

Effects:	Intercept PointsDifferential

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	1	33.35926	33.35926	27.02
Error	16	19.75185	1.23449	
Corrected Total	17	53.11111		

Root MSE	1.11108
Dependent Mean	12.22222
R-Square	0.6281
Adj R-Sq	0.6049
AIC	25.67176
AICC	27.38605
SBC	7.45250

Parameter Estimates		
Parameter	DF	Estimate
Intercept	1	9.529105
PointsDifferential	1	0.018327

```
/* LASSO model */
proc GLMSELECT data=pats;
model RegSeasonWins = BradyPasserRating PointsFor PointsDifferential PassTouchdowns
PassInterceptions NetYardsPerPass / selection = LASSO;
run; quit;
```

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 1).

Effects:	Intercept PointsDifferential

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	1	33.35926	33.35926	27.02
Error	16	19.75185	1.23449	
Corrected Total	17	53.11111		

Root MSE	1.11108
Dependent Mean	12.22222
R-Square	0.6281
Adj R-Sq	0.6049
AIC	25.67176
AICC	27.38605
SBC	7.45250

Parameter Estimates

Parameter	DF	Estimate
Intercept	1	9.529105
PointsDifferential	1	0.018327

```
/* Stepwise model */
proc GLMSELECT data=pats;
model RegSeasonWins = BradyPasserRating PointsFor PointsDifferential PassTouchdowns
PassInterceptions NetYardsPerPass / selection = stepwise;
run; quit;
```

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 2).

Effects:	Intercept PointsFor PointsDifferential

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	2	37.74542	18.87271	18.42
Error	15	15.36569	1.02438	
Corrected Total	17	53.11111		

Root MSE	1.01212
Dependent Mean	12.22222
R-Square	0.7107
Adj R-Sq	0.6721
AIC	23.15178
AICC	26.22870
SBC	5.82289

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	13.364501	2.346353	5.70
PointsFor	1	-0.013311	0.007036	-1.89
PointsDifferential	1	0.032623	0.006981	4.67

```
/* Partial model with PointsFor and PointsDifferential */
proc reg data=pats;
model RegSeasonWins = PointsFor PointsDifferential /partial;
run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: RegSeasonWins

Number of Observations Read	18
Number of Observations Used	18

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	37.74542	18.87271	18.42	<.0001
Error	15	15.36569	1.02438		
Corrected Total	17	53.11111			

Root MSE	1.01212	R-Square	0.7107
Dependent Mean	12.22222	Adj R-Sq	0.6721
Coeff Var	8.28095		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	13.36450	2.34635	5.70	<.0001
PointsFor	1	-0.01331	0.00704	-1.89	0.0780
PointsDifferential	1	0.03262	0.00698	4.67	0.0003

```
/* Partial model with PointsDifferential */
proc reg data=pats;
model RegSeasonWins = PointsDifferential /partial;
run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: RegSeasonWins

Number of Observations Read	18
Number of Observations Used	18

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	34.07901	34.07901	28.65	<.0001
Error	16	19.03211	1.18951		
Corrected Total	17	53.11111			

Root MSE	1.09065	R-Square	0.6417
Dependent Mean	12.22222	Adj R-Sq	0.6193
Coeff Var	8.92346		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	9.07117	0.64238	14.12	<.0001
PointsDifferential	1	0.02144	0.00401	5.35	<.0001

```
/* Leave-one-out validation model */
proc GLMSELECT data=pats;
model RegSeasonWins = BradyPasserRating PointsFor PointsDifferential PassTouchdowns
PassInterceptions NetYardsPerPass / selection=forward(STOP=Press);
run;
```

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 2).

Effects:	Intercept PointsFor PointsDifferential

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	2	37.74542	18.87271	18.42
Error	15	15.36569	1.02438	
Corrected Total	17	53.11111		

Root MSE	1.01212
Dependent Mean	12.22222
R-Square	0.7107
Adj R-Sq	0.6721
AIC	23.15178
AICC	26.22870
PRESS	22.67804
SBC	5.82289

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	13.364501	2.346353	5.70
PointsFor	1	-0.013311	0.007036	-1.89
PointsDifferential	1	0.032623	0.006981	4.67

[7]

```
/* Confidence intervals for the final model */
proc reg data=pats;
model RegSeasonWins = PointsFor PointsDifferential / clb;
run;
```

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	13.36450	2.34635	5.70	<.0001	8.36337	18.36563
PointsFor	1	-0.01331	0.00704	-1.89	0.0780	-0.02831	0.00169
PointsDifferential	1	0.03262	0.00698	4.67	0.0003	0.01774	0.04750

[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;
```

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 1).

Effects:	Intercept PointsDifferential

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	1	26.05032	26.05032	29.32
Error	13	11.54968	0.88844	
Corrected Total	14	37.60000		

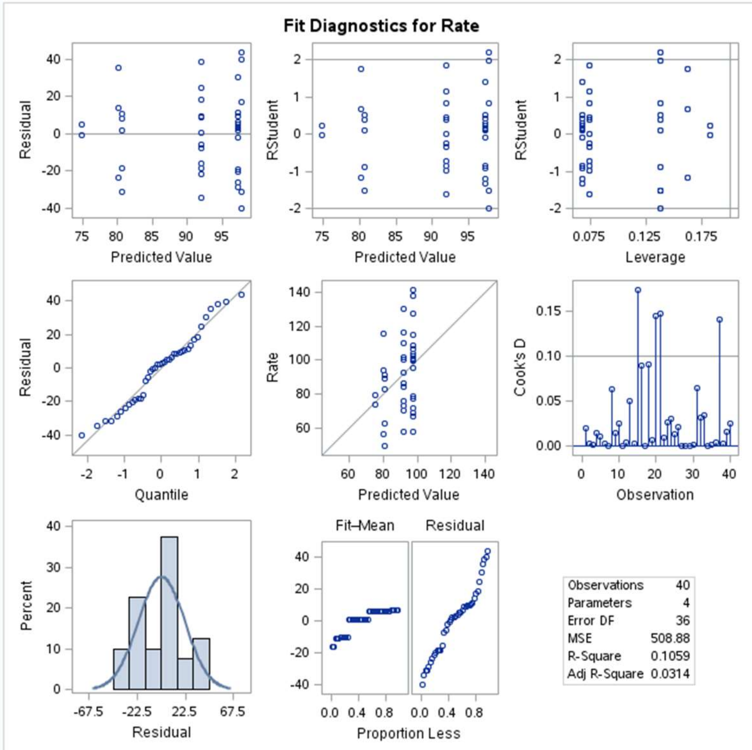
Root MSE	0.94257
Dependent Mean	12.40000
R-Square	0.6928
Adj R-Sq	0.6692
AIC	17.07912
AICC	19.26094
PRESS	14.45026
SBC	1.49522

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	8.926914	0.686011	13.01
PointsDifferential	1	0.021698	0.004007	5.41

[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;
```



The GLM Procedure

Class Level Information		
Class	Levels	Values
AgeBin	3	1 2 3
WonLost	2	L W

Number of Observations Read	40
Number of Observations Used	40

The GLM Procedure

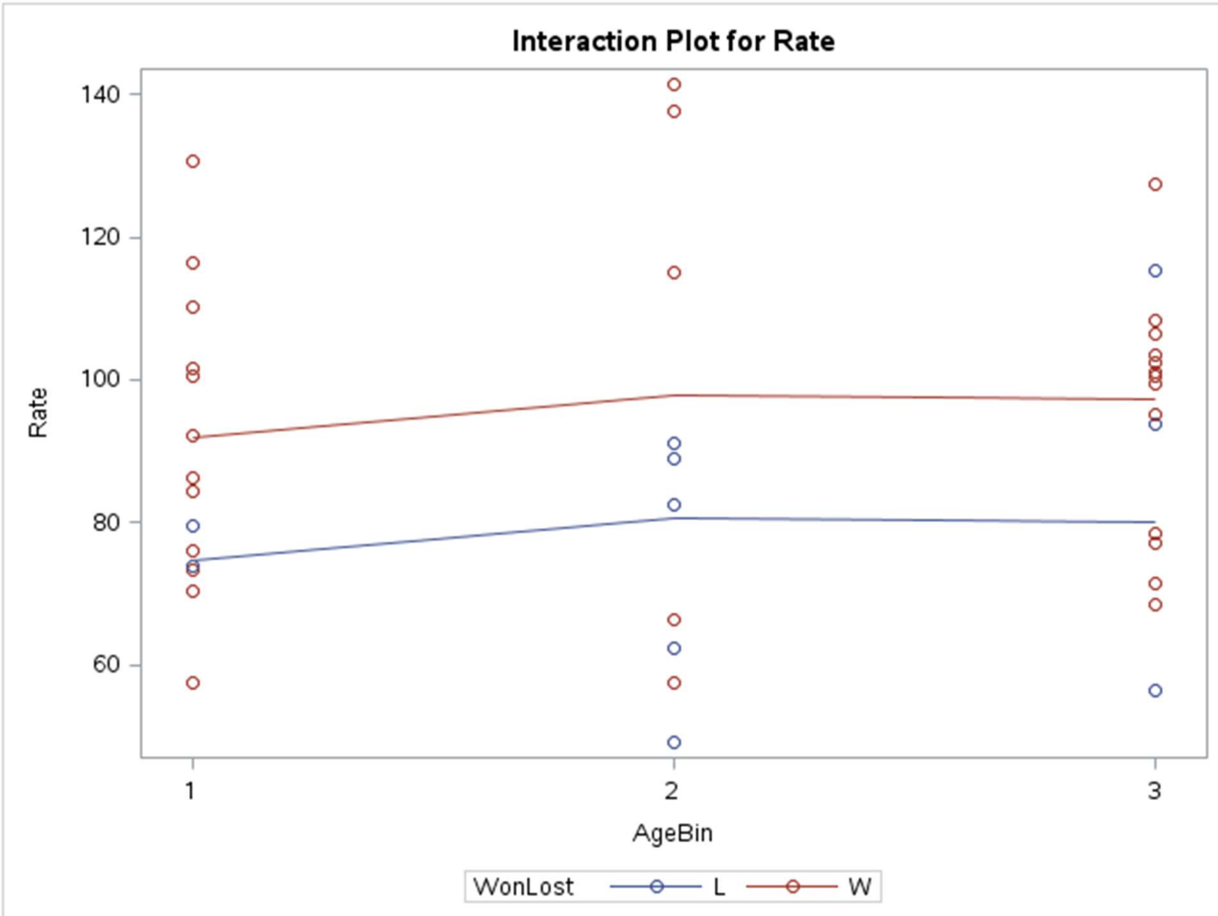
Dependent Variable: Rate Rate

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	2170.46725	723.48908	1.42	0.2524
Error	36	18319.55050	508.87640		
Corrected Total	39	20490.01775			

R-Square	Coeff Var	Root MSE	Rate Mean
0.105928	24.71938	22.55829	91.25750

Source	DF	Type I SS	Mean Square	F Value	Pr > F
AgeBin	2	215.971607	107.985804	0.21	0.8098
WonLost	1	1954.495640	1954.495640	3.84	0.0578

Source	DF	Type III SS	Mean Square	F Value	Pr > F
AgeBin	2	270.415163	135.207582	0.27	0.7682
WonLost	1	1954.495640	1954.495640	3.84	0.0578



[10]

```
proc anova data = work.import;
class WonLost;
model Rate = WonLost;
run;

proc glm data = work.import;
class WonLost;
model Rate = WonLost;
run;
```

The GLM Procedure

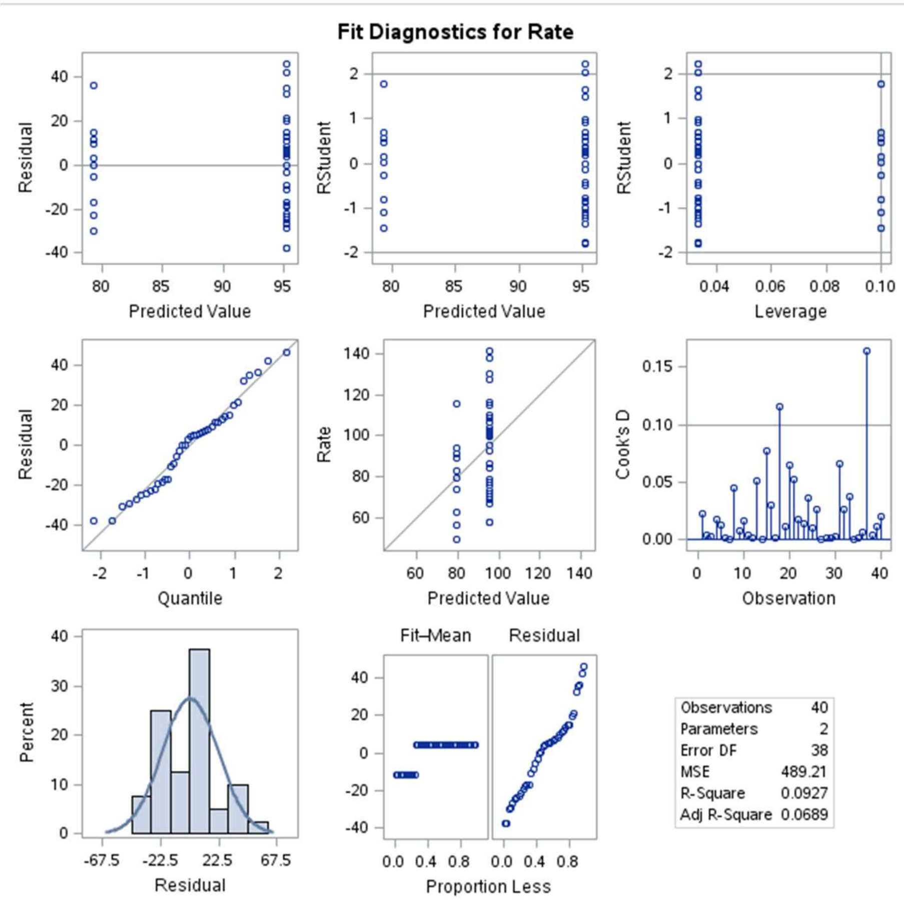
Dependent Variable: Rate Rate

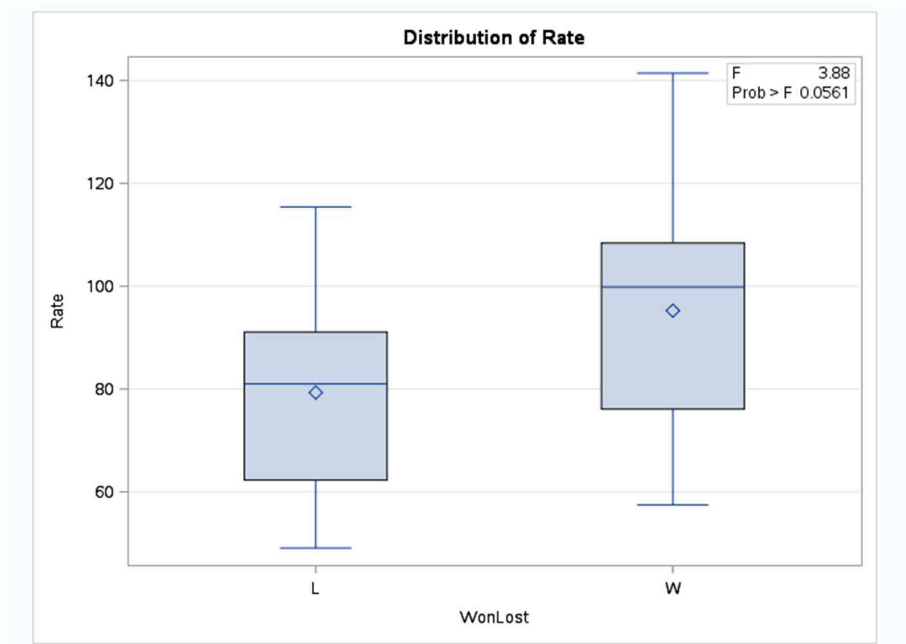
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1900.05208	1900.05208	3.88	0.0561
Error	38	18589.96567	489.20962		
Corrected Total	39	20490.01775			

R-Square	Coeff Var	Root MSE	Rate Mean
0.092731	24.23700	22.11808	91.25750

Source	DF	Type I SS	Mean Square	F Value	Pr > F
WonLost	1	1900.052083	1900.052083	3.88	0.0561

Source	DF	Type III SS	Mean Square	F Value	Pr > F
WonLost	1	1900.052083	1900.052083	3.88	0.0561





Distribution of Rate
F 3.88
Prob > F 0.0561
140
120
100
80
60
Rate
L
W
WonLost