A Regression Analysis of the Point Differential in NFL Games

Andrew Ivanov

Table of Contents

Section 1 Introduction to Data Set	**Pages 3-4**
Section 2 Simple Linear Regression	**Pages 5-8**
Section 3 Multiple Linear Regression	**Pages 9-14**
Section 4 Polynomial Regression	**Pages 15-17**
Section 5 Model Selection	**Pages 18- 25**
Section 6 Logistic Regression	**Pages 26-29**
Section 7 Cross-validation	**Pages 30-32**

Cover picture is via [ESPN Images]

Chapter 1: The Data Set

I used Yahoo! Finance as a source of 2019 price history for each asset I selected.

Every Sunday during the NFL season, play by play announcers often throw out random statistics which they believe have an impact on the final score of the game. For my second data set, I decided to look into which football statistics influence the final point differential between two teams in an NFL game. The NFL regular season has 256 each games each year. To increase the sample size of my data set, I gathered data from several recent NFL seasons, from the 2015 NFL season up to week 10 of the 2020 NFL season. This gave my dataset a sample size of 2856 observations. Each observation is the outcome of a single game and the variables in each observation are for a single NFL team.

To compile my dataset, I created an account on https://sportradar.us/sports-data/ which provides a tool to query www.pro-football-reference.com where I collected all of the single game football statistics for my dataset. I filtered through each NFL season I was interested in and compiled all of the variables I wanted to examine into an Excel file.

2015 to 2020 NFL Single Game Statistics

Obs	PtDiff	Location	DivGame	SackDiff	TODiff	PressuresAllowed	YPC	TotalPenalties	NetPassYdsPerAtt	NetYdsAllowed	ReturnYards	BigPlayDiff	SecondHalfRushingYds
1	4	0	1	1	-1	7	9	5.6000	366	50	3	65	15
2	-3	1	0	3	-3	3	6	7.6571	322	99	2	33	-10
3	20	0	0	2	-1	6	9	9.9730	285	39	7	11	28
4	3	1	1	2	1	5	5	7.5000	572	31	-2	5	-3
5	2	1	0	-3	0	5	6	7.3750	369	54	3	77	-7
6	-13	1	0	1	-2	11	5	8.0370	383	104	2	-25	-1
7	-4	1	0	0	1	13	7	5.9737	437	39	-1	30	-14
8	-7	1	1	-2	-1	10	3	5.5946	437	76	2	21	17
9	-1	1	0	0	-1	10	7	7.6667	386	3	-3	21	8
10	7	1	0	-1	0	14	4	7.7429	405	72	-1	-35	32
11	17	0	0	0	2	7	3	7.3750	304	41	8	170	-14
12	14	0	0	2	-1	5	5	8.9565	343	97	4	48	24
13	2	0	0	3	1	7	12	6.4074	364	130	3	-21	-4
14	14	0	0	-2	1	4	5	6.7826	339	44	-5	55	-6
15	10	1	1	0	0	10	7	6.6522	254	7	3	4	3
16	3	1	0	-2	0	12	5	8.3030	478	0	-3	-60	7
17	-26	0	0	-1	-3	3	10	5.9783	334	94	-2	-12	-9
18	8	0	1	4	1	9	10	6.8837	191	97	3	43	3
19	10	1	0	-2	4	10	6	10.1579	419	165	2	-3	-2
20	-4	1	0	-1	0	5	5	7.6176	372	43	-2	20	-14

In total, I compiled 13 variables that I would further examine. They are defined below.

PtDiff will be my Y variable. It is the number of points that a team won or lost by. PtDiff will be positive if a team won a game and negative if a team lost a game.

Location is a dummy variable that is 0 if the team is playing away and 1 if the team is playing at home

DivisionalGame is a dummy variable that is 1 if the team is a playing a team in the same division and 0 otherwise.

SackDiff is the sack differential between the team and its opponent. SackDiff will be positive if the team tackled the opposing quarterback more than the opposing team tackled the team's quarterback and negative otherwise.

TODiff is the turnover differential between the team and its opponent. TODiff will be positive if the team turned over the football less times than the opposing team in the game and negative otherwise.

PressuresAllowed is number of times the team allowed the quarterback to get rushed, knocked down or sacked.

YPC is the yards per rushing attempt that a team averaged in a game.

TotalPenalties is the total number of offensive and defensive penalties a team committed in a game.

NetPassYdsPerAtt is the number of passing yards a quarterback had minus the number of yards lost due to sacks and then divided by the number of pass attempts by the quarterback.

NetYdsAllowed is the total number of rushing, passing and return yards given up by a team's defense in a game.

ReturnYards is the total number of punt return and kick return yards a team had in a game

BigPlayDiff is the big play differential between two teams. A big play is counted as any rush attempt that was 10 yards or more or any passing play that was 20 yards or more.

SecondHalfRushingYds is the total number of rushing yards a team had in the second half of the game.

Chapter 2: A Simple Linear Regression Model

Using SAS, to create the regression scatterplot showing both the 95% confidence interval and the 95% prediction interval. Also, show the Analysis of Variance and Parameter Estimation sections output.

I selected point differential as a y-variable and sack differential as the x-variable.

Point Differential vs Sack Differential
with 95% confidence and prediction intervals

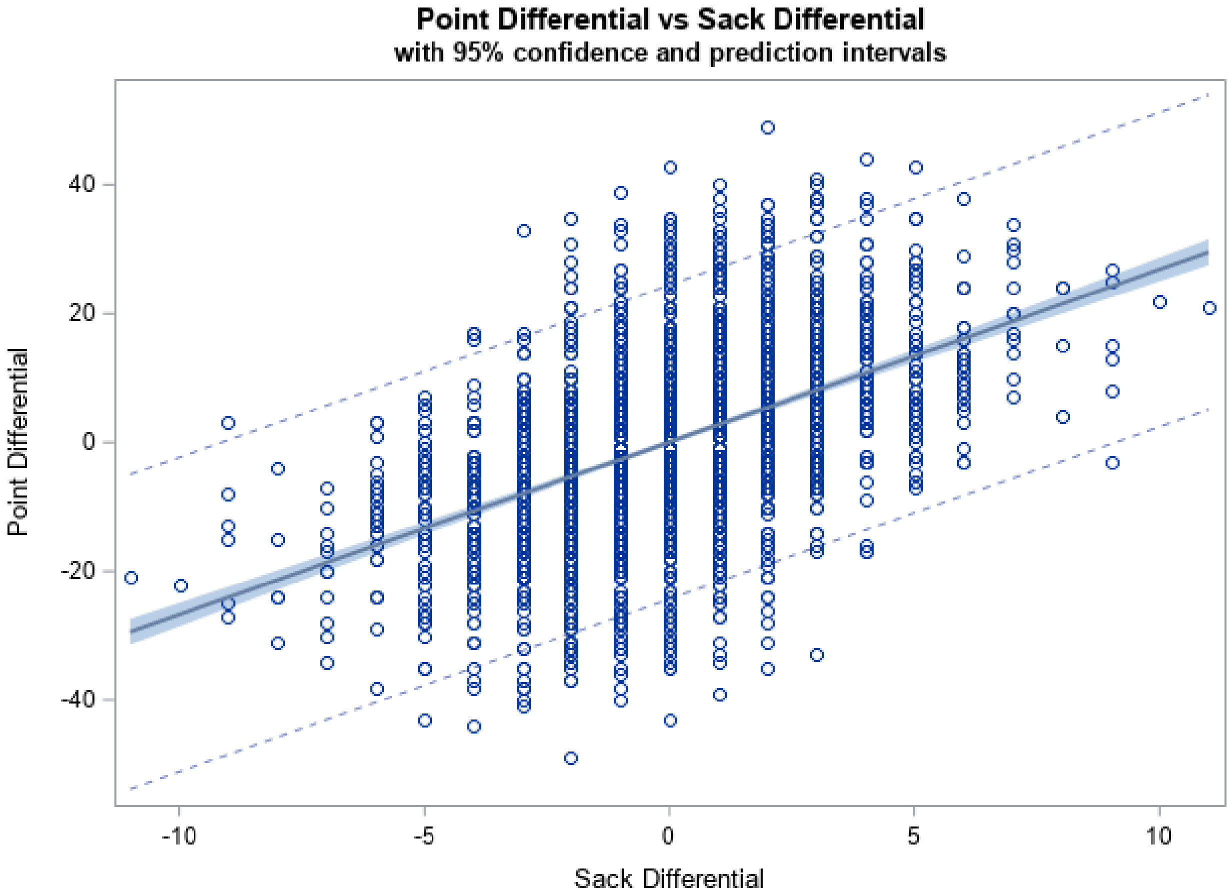
The REG Procedure
Model: MODEL1
Dependent Variable: PtDiff

Number of Observations Read	2856
Number of Observations Used	2856

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	134740	134740	872.68	<.0001
Error	2854	440654	154.39861		
Corrected Total	2855	575394			

Root MSE	12.42572	R-Square	0.2342
Dependent Mean	0	Adj R-Sq	0.2339
Coeff Var	.		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.04035	0.23251	0.17	0.8622
SackDiff	1	2.67996	0.09072	29.54	<.0001



By using the parameter estimates for the variables from the SAS output, we can see that the simple linear regression model for point differential using sack differential is:

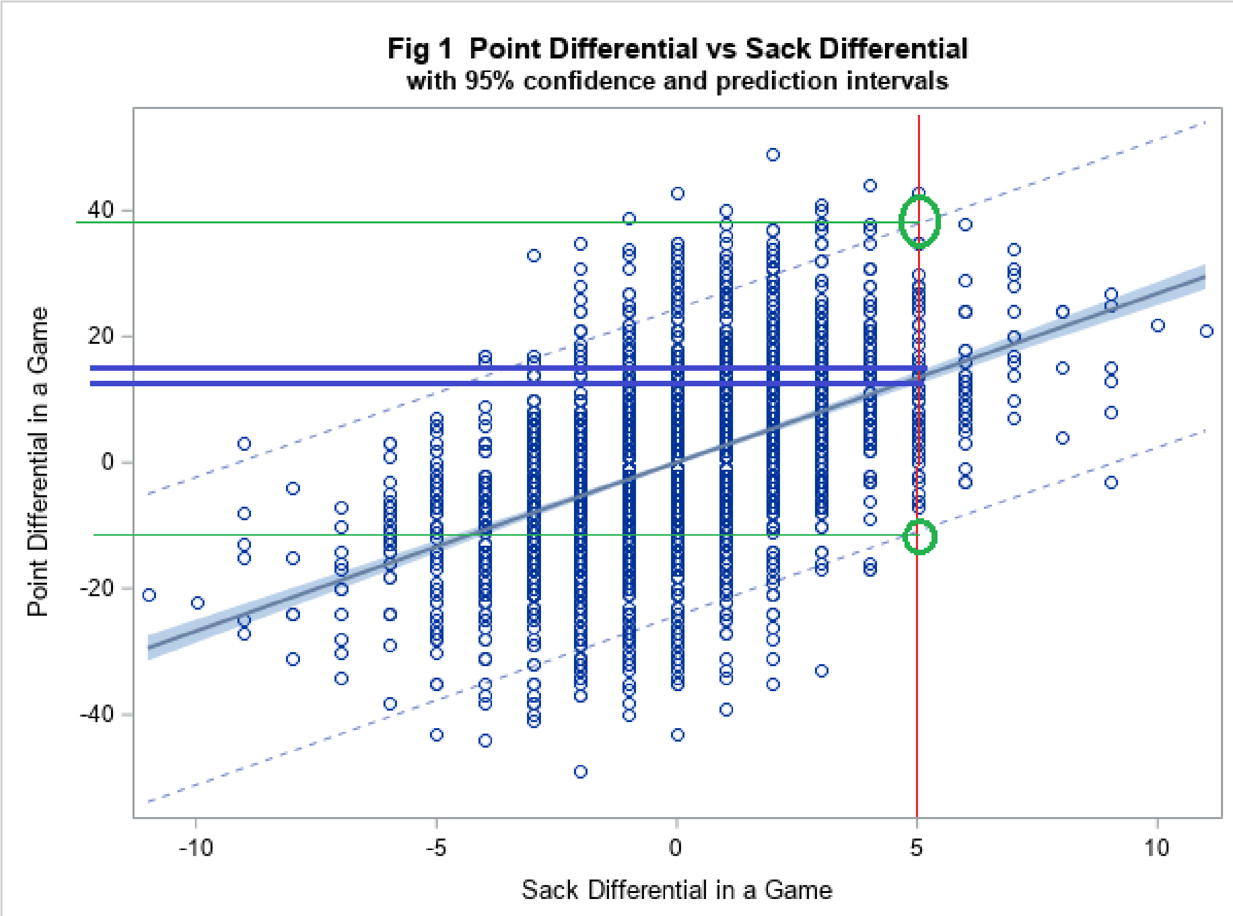
$$\hat{y} = 0.04035 + 2.67996x$$

Since sack differential has a positive sample of 2.6799, the linear regression model is estimating that for each additional sack differential, the predicted value of point differential will increase by 2.6799.

Y-hat stands for the predicted value of Y, and it can be obtained by plugging an individual value of x into the equation and calculating y-hat.

Story of Many Possible Samples

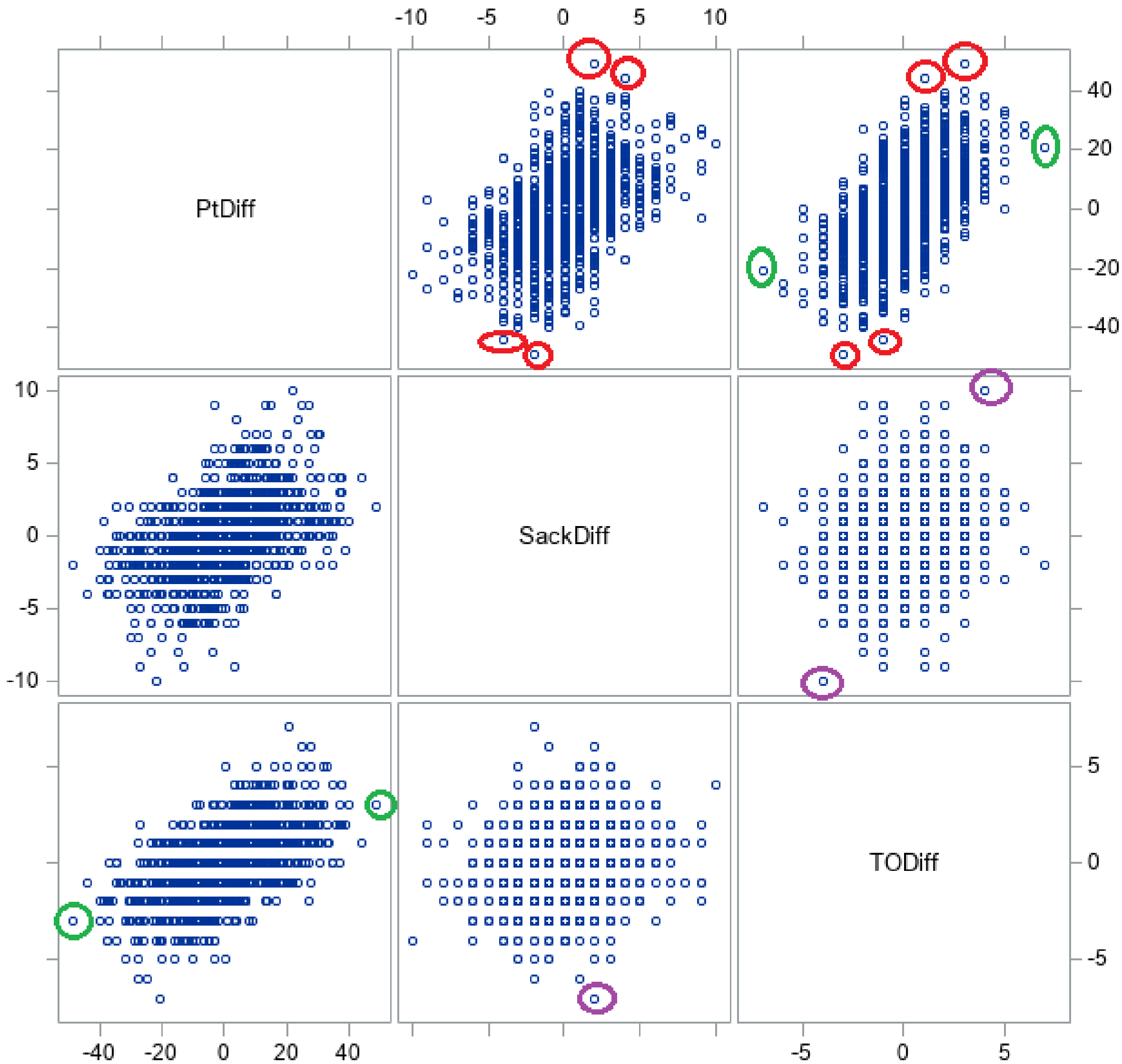
Imagine that we were going to take a sample of outcomes of games in a given NFL week where each team plays from our dataset of 2856 outcomes. For any given week we would have 32 possible outcomes which would make our n = 32 out of the possible N = 2856. Using our sample of 32 outcomes, we can calculate the average point differential and the average sack differential in the given week. Now, if we were to take another a sample of outcomes in another given week, once again n would equal 32 of the possible N = 2856. However, it is almost certain that the sample average in our new sample would not be equal to the sample averages of the previous dataset. This is because there are many possible samples of n = 32 that we can choose and each possible samples its own sample average. The total number of samples N = 2856 has its own population average and each possible sample of n =32 has its own sample average.



In the plot of point differential versus sack differential, the 95% confidence interval is the blue area along the regression line and the 95% prediction interval is the dotted line. For an arbitrary x value such as 5, we can say that the 95% confidence interval is the range between the two blue lines. The 95% confidence interval is a range of values that you can be 95% certain contains the true mean of the population. The 95% in a 95% confidence interval means that we are 95% confident that the population mean will be between the lower and upper values of the confidence interval. In this example the 95% confidence interval is 13 to 16. We can say with 95% confidence that the point differential of a team that has a sack differential of 5 is between 13 and 16. A prediction interval predicts in what range a future observation will fall and tries to estimate the location of the parent population distribution, while a confidence interval is a range of values that tries to capture the population average of a parent population. The 95% prediction interval is between -11 and 38. Therefore, we can say that the 95% prediction interval for the point differential in a game when a team has a sack differential of 5 is -11 to 38. The reason the confidence interval is small is because the mean values are around 13 and 16 meanwhile the prediction interval has to account for all possible point differentials when sacks are equal to 5.

Chapter 3: A Multiple Regression Model with Two Regressors

Scatterplot Matrix for Point Differential Sack Differential and Turnover Differential



There does not appear to be any significant collinearity, curvature, or heteroscedasticity when we plot point difference versus sack differential and turnover differential. The potential outliers for point differential are circled in red. They are the points with the most extreme y values in the data set. The potential high leverage points for point differential are circled in green. These are the points with the most extreme x values. The points circled in purple are potential outliers for sack differential and turnover differential, they are the points with the most extreme y values. There are no obvious influential points, which are both outliers and leverage points.

The two-regressor multiple regression model is:

$\hat{y}$ = B0 + B1x + B2x + ε_i.

The model is proposing that the two variables, B1,B2 along with the intercept B0 can be used to predict the value of Y.

Two-regressor model output

The REG Procedure
Model: MODEL1
Dependent Variable: PtDiff

Number of Observations Read	2856
Number of Observations Used	2856

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	263661	131831	1206.52	<.0001
Error	2853	311733	109.26489		
Corrected Total	2855	575394			

Root MSE	10.45298	R-Square	0.4582
Dependent Mean	0	Adj R-Sq	0.4578
Coeff Var	.		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.03043	0.19560	0.16	0.8764
SackDiff	1	2.02086	0.07869	25.68	<.0001
TODiff	1	3.73115	0.10862	34.35	<.0001

"Sweeping out" operation can be used to find the partial sample slope coefficient of x2.

SUMMARY OUTPUT				
PtDiff | SackDiff TODiff				
Regression Statistics				
Multiple R	0.676924923			
R Square	0.458227351			
Adjusted R Square	0.457847559			
Standard Error	10.45298475			
Observations	2856			
ANOVA				
	df	*SS*	*MS*	*F*
Regression	2	263661.2683	131830.6342	1206.523284
Residual	2853	311732.7317	109.2648902	
Total	2855	575394		
	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	0.030426057	0.19560015	0.15555232	0.876396915
SackDiff	2.020856251	0.07869199	25.68058388	5.0631E-131
TODiff	3.731147916	0.108622829	34.34957402	1.0387E-216

Regressing sack differential and turnover differential produces an output identical to the Proc Reg in SAS for my two regressor model.

Next, we will regress Y, PtDiff on SackDiff and PtDiff on TODiff. We will then regress X1, SackDiff on TODiff. Afterwards we will regress the residuals of PtDiff on TODiff on the residuals of SackDiff on TODiff which will produce an x variable coefficient identical to SackDiff in the original multiple regression.

SUMMARY OUTPUT				
PtDiff | Sack Diff				
Regression Statistics				
Multiple R	0.483911778			
R Square	0.234170609			
Adjusted R Square	0.233902274			
Standard Error	12.42572377			
Observations	2856			
ANOVA				
	df	*SS*	*MS*	*F*
Regression	1	134740.3636	134740.3636	872.678598
Residual	2854	440653.6364	154.3986112	
Total	2855	575394		
	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	0.040349542	0.232514524	0.173535577	0.862242772
Sack Diff	2.679960292	0.090719615	29.541134	1.3974E-167

SUMMARY OUTPUT				
Pt | TO Diff				
Regression Statistics				
Multiple R	0.577055051			
R Square	0.332992532			
Adjusted R Square	0.332758822			
Standard Error	11.59634235			
Observations	2856			
ANOVA				
	df	*SS*	*MS*	*F*
Regression	1	191601.9049	191601.9049	1424.812662
Residual	2854	383792.0951	134.4751559	
Total	2855	575394		
	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	0	0.216991103	0	1
TO Diff	4.411334552	0.116866789	37.74669074	2.8E-253

SUMMARY OUTPUT					
SackDiff |TODiff					
Regression Statistics					
Multiple R	0.243838394				
R Square	0.059457163				
Adjusted R Square	0.05912761				
Standard Error	2.486467597				
Observations	2856				
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1115.437335	1115.437	180.4179	
Residual	2854	17644.91526	6.182521		
Total	2855	18760.35259			
	Coefficients	*Standard Error*	*t Stat*	*P-value*	
Intercept	-0.015056022	0.046526856	-0.3236	0.746266	
X Variable 1	0.336583384	0.025058374	13.43197	6.25E-40	

SUMMARY OUTPUT				
Regression Statistics				
Multiple R	0.433308489			
R Square	0.187756247			
Adjusted R Square	0.187471649			
Standard Error	10.4511533			
Observations	2856			
ANOVA				
	df	*SS*	*MS*	*F*
Regression	1	72059.36339	72059.36	659.7235
Residual	2854	311732.7317	109.2266	
Total	2855	383792.0951		
	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	4.13894E-16	0.195562291	2.12E-15	1
X Variable 1	2.020856251	0.078678203	25.68508	4.6E-131

Chapter 4: Polynomial Models

4.1. [6] Using the same y and the same two x-variables from above, Problem 3.3, create the variables needed to run a quadratic polynomial model, including squares and the interaction term. Run the model in SAS. In terms of variance inflation and the BIC criteria, how well does this quadratic polynomial perform compare to the original model, above?

For the quadratic polynomial model, I created 3 new variables which are used to create a polynomial model. SqSack was the sack differential squared, SqTO was the turnover differential squared, and Sack_TO was sack differential times turnover differential.

The Polynomial Model

Point Differential vs SackDiff TODiff SackDiff Squared TODiff Squared and SackDiff * TODiff

The REG Procedure
Model: MODEL1
Dependent Variable: PtDiff

Number of Observations Read	2856
Number of Observations Used	2856

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	263661	52732	482.10	<.0001
Error	2850	311733	109.37984		
Corrected Total	2855	575394			

Root MSE	10.45848	R-Square	0.4582
Dependent Mean	0	Adj R-Sq	0.4573
Coeff Var	.		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	0.03127	0.26511	0.12	0.9061	0
SackDiff	1	2.02084	0.07874	25.67	<.0001	1.06326
TODiff	1	3.73118	0.10868	34.33	<.0001	1.06327
SqSack	1	-0.00021985	0.01862	-0.01	0.9906	1.09912
SqTO	1	-0.00045527	0.04043	-0.01	0.9910	1.08017
Sack_TO	1	0.00187	0.04592	0.04	0.9675	1.17971

Fit Statistics	
-2 Res Log Likelihood	21529.9
AIC (Smaller is Better)	21531.9
AICC (Smaller is Better)	21531.9
BIC (Smaller is Better)	21537.9

The previous quadratic model had the following values:

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	0.10248	0.27714	0.37	0.7116	0
SackDiff	1	2.01756	0.10882	18.54	<.0001	1.04790
TODiff	1	3.60894	0.15144	23.83	<.0001	1.04790

Fit Statistics	
-2 Res Log Likelihood	21514.7
AIC (Smaller is Better)	21516.7
AICC (Smaller is Better)	21516.7
BIC (Smaller is Better)	21522.7

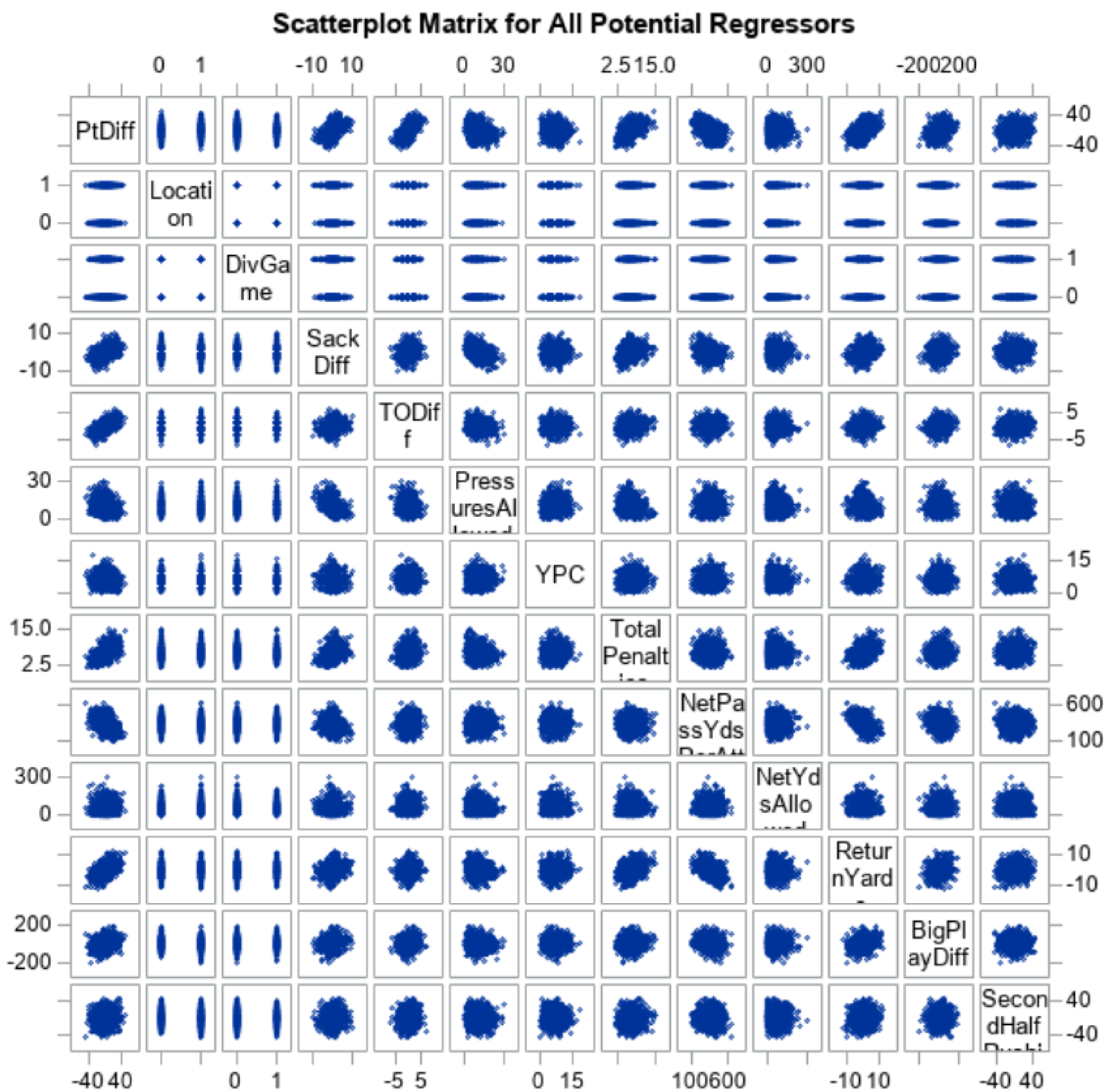
As we can see because none of the terms added to the polynomial model was statistically significant many elements of the two models are similar. The original model has a slightly larger R squared value and a slightly lower BIC value. The variance inflation is also higher in the polynomial model than the original model. The original model is superior because it uses less terms that are all statistically significant.

All of the additional terms in the polynomial model were statistically insignificant. I believe that this is because in football sacks and turnovers both lead to negative consequences and less points scored by a team. Therefore, when we square turnovers and sacks, teams that have a negative differential in either category have a similar squared value as teams that have a positive differential. Similarly, teams with both a negative differential in sacks and turnovers have an interaction term that is similar to teams with a positive differential in sacks and turnovers. SAS recognizes this and identifies the terms as statistically insignificant.

I believe that a polynomial model is not effective with my dataset because all the regressors I selected have a linear relationship with point differential. Squaring the regressors or creating an interaction term with two regressors will create a new term that will likely be statistically insignificant. This will be explored in the next section in the matrix scatterplot of all variables.

Chapter 5: Model Selection Methods

Below is a matrix scatterplot for Point Differential and all possible regressors in the dataset. As we can see many of the regressors appear to have a relationship with Point Differential in the shape of a slanted oval. However, there are some variables where there appears to be no relationship.

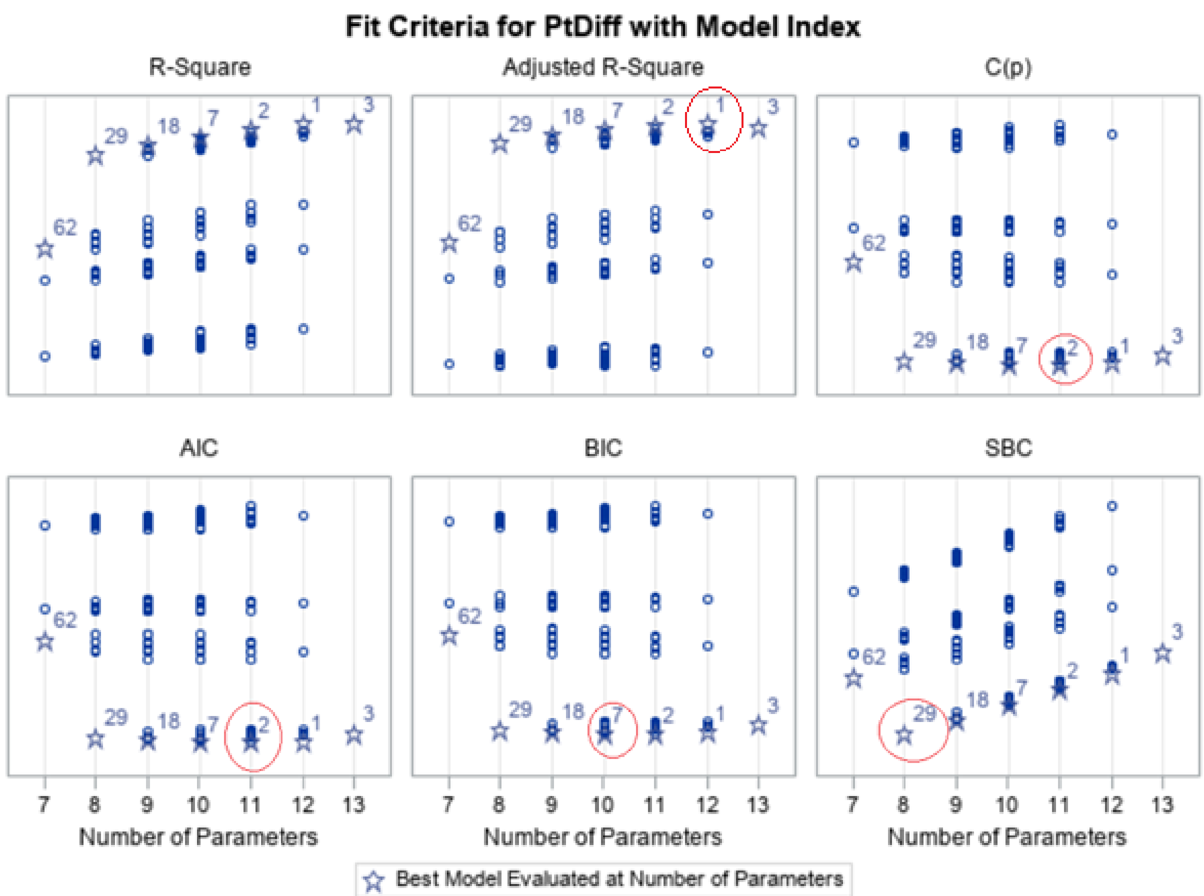


"Model selection," its goals, and the number of possible subset models.

In regression model selection involves selecting the best possible model out of the possible regressors. The best model varies depending on what measure we are looking at. In model selection, the goal is to minimize bias and variance as much as possible. The number of regressors in a model influences both bias and variance. Using a low number of regressors decreases the estimation variance, but increases the bias, while using a larger number of regressors will increase the variance but decrease the bias. The goal is to select a number of regressors where we are comfortable with the bias variance trade off.

The formula to calculate the number of possible subset models is 2^k where K is the number of possible regressors. In my dataset we are regressing Point Differential on 12 possible regressors therefore the number of possible subset models is $2^{12} = 4096$.

<u>The best subset model selection routine from SAS for Adjusted Rsq, Cp, AIC, BIC, and SBC:</u>



The REG Procedure
Model: MODEL1
Dependent Variable: PtDiff

Adjusted R-Square Selection Method

Number of Observations Read	1428
Number of Observations Used	1428

Model Index	Number in Model	Adjusted R-Square	R-Square	C(p)	AIC	BIC	SBC	Variables in Model
1	11	0.7090	0.7113	11.1619	5801.2361	5803.4536	5864.40444	Location SackDiff TODiff PressuresAllowed YPC TotalPenalties NetPassYdsPerAtt NetYdsAllowed ReturnYards BigPlayDiff SecondHalfRushingYds
2	10	0.7089	0.7109	10.7873	5800.8753	5803.0493	5858.77960	Location SackDiff TODiff PressuresAllowed YPC TotalPenalties NetPassYdsPerAtt ReturnYards BigPlayDiff SecondHalfRushingYds
3	12	0.7088	0.7113	13.0000	5803.0727	5805.3114	5871.50506	Location DivGame SackDiff TODiff PressuresAllowed YPC TotalPenalties NetPassYdsPerAtt NetYdsAllowed ReturnYards BigPlayDiff SecondHalfRushingYds
4	10	0.7088	0.7108	11.2575	5801.3491	5803.5157	5859.25341	Location SackDiff TODiff PressuresAllowed YPC TotalPenalties NetPassYdsPerAtt NetYdsAllowed ReturnYards BigPlayDiff
5	10	0.7088	0.7108	11.3939	5801.4866	5803.6511	5859.39089	Location SackDiff TODiff YPC TotalPenalties NetPassYdsPerAtt NetYdsAllowed ReturnYards BigPlayDiff SecondHalfRushingYds
6	11	0.7087	0.7110	12.6515	5802.7383	5804.9305	5865.90669	Location DivGame SackDiff TODiff PressuresAllowed YPC TotalPenalties NetPassYdsPerAtt ReturnYards BigPlayDiff SecondHalfRushingYds
7	9	0.7087	0.7105	10.7885	5800.8909	5803.0206	5853.53117	Location SackDiff TODiff PressuresAllowed YPC TotalPenalties NetPassYdsPerAtt ReturnYards BigPlayDiff
8	11	0.7086	0.7109	13.1104	5803.2008	5805.3851	5866.36919	Location DivGame SackDiff TODiff PressuresAllowed YPC TotalPenalties NetPassYdsPerAtt NetYdsAllowed ReturnYards BigPlayDiff
9	9	0.7086	0.7104	11.1855	5801.2904	5803.4145	5853.93070	Location SackDiff TODiff YPC TotalPenalties NetPassYdsPerAtt ReturnYards BigPlayDiff SecondHalfRushingYds
10	11	0.7086	0.7108	13.2473	5803.3388	5805.5207	5866.50713	Location DivGame SackDiff TODiff YPC TotalPenalties NetPassYdsPerAtt NetYdsAllowed ReturnYards BigPlayDiff SecondHalfRushingYds
11	10	0.7086	0.7106	12.2511	5802.3499	5804.5010	5860.25423	SackDiff TODiff PressuresAllowed YPC TotalPenalties NetPassYdsPerAtt NetYdsAllowed ReturnYards BigPlayDiff SecondHalfRushingYds
12	9	0.7085	0.7104	11.5863	5801.6936	5803.8121	5854.33389	Location SackDiff TODiff YPC TotalPenalties NetPassYdsPerAtt NetYdsAllowed ReturnYards BigPlayDiff
13	10	0.7085	0.7105	12.6652	5802.7667	5804.9114	5860.67107	Location DivGame SackDiff TODiff PressuresAllowed YPC TotalPenalties NetPassYdsPerAtt ReturnYards BigPlayDiff
14	9	0.7084	0.7103	11.9534	5802.0628	5804.1761	5854.70312	SackDiff TODiff PressuresAllowed YPC TotalPenalties NetPassYdsPerAtt ReturnYards BigPlayDiff SecondHalfRushingYds
15	9	0.7084	0.7103	12.0259	5802.1358	5804.2480	5854.77606	SackDiff TODiff PressuresAllowed YPC TotalPenalties NetPassYdsPerAtt NetYdsAllowed ReturnYards BigPlayDiff
24	9	0.7082	0.7100	13.1719	5803.2878	5805.3838	5855.92807	Location DivGame SackDiff TODiff YPC TotalPenalties NetPassYdsPerAtt ReturnYards BigPlayDiff
25	8	0.7082	0.7098	12.2484	5802.3690	5804.4418	5849.74530	SackDiff TODiff YPC TotalPenalties NetPassYdsPerAtt ReturnYards BigPlayDiff SecondHalfRushingYds
26	8	0.7082	0.7098	12.2489	5802.3694	5804.4422	5849.74571	SackDiff TODiff YPC TotalPenalties NetPassYdsPerAtt NetYdsAllowed ReturnYards BigPlayDiff
27	10	0.7082	0.7102	14.2494	5804.3605	5806.4805	5862.26482	DivGame SackDiff TODiff YPC TotalPenalties NetPassYdsPerAtt NetYdsAllowed ReturnYards BigPlayDiff SecondHalfRushingYds
28	9	0.7081	0.7100	13.5250	5803.6424	5805.7335	5856.28273	DivGame SackDiff TODiff PressuresAllowed YPC TotalPenalties NetPassYdsPerAtt ReturnYards BigPlayDiff
29	7	0.7080	0.7095	12.0194	5802.1461	5804.1908	5844.25833	SackDiff TODiff YPC TotalPenalties NetPassYdsPerAtt ReturnYards BigPlayDiff
30	9	0.7080	0.7098	14.1290	5804.2491	5806.3317	5856.88942	DivGame SackDiff TODiff YPC TotalPenalties NetPassYdsPerAtt NetYdsAllowed ReturnYards BigPlayDiff
31	9	0.7080	0.7098	14.1421	5804.2623	5806.3446	5856.90257	DivGame SackDiff TODiff YPC TotalPenalties NetPassYdsPerAtt ReturnYards BigPlayDiff SecondHalfRushingYds
32	8	0.7078	0.7095	13.9231	5804.0495	5806.1010	5851.42578	DivGame SackDiff TODiff YPC TotalPenalties NetPassYdsPerAtt ReturnYards BigPlayDiff
33	10	0.7046	0.7067	31.6219	5821.7219	5823.5745	5879.62619	Location SackDiff TODiff PressuresAllowed TotalPenalties NetPassYdsPerAtt NetYdsAllowed ReturnYards BigPlayDiff SecondHalfRushingYds
34	9	0.7044	0.7063	31.4845	5821.5708	5823.4111	5874.21112	Location SackDiff TODiff PressuresAllowed TotalPenalties NetPassYdsPerAtt ReturnYards BigPlayDiff SecondHalfRushingYds
35	11	0.7044	0.7067	33.5224	5823.6230	5825.4639	5886.79134	Location DivGame SackDiff TODiff PressuresAllowed TotalPenalties NetPassYdsPerAtt NetYdsAllowed ReturnYards BigPlayDiff SecondHalfRushingYds
36	9	0.7044	0.7062	31.8466	5821.9299	5823.7652	5874.57022	Location SackDiff TODiff PressuresAllowed TotalPenalties NetPassYdsPerAtt NetYdsAllowed ReturnYards BigPlayDiff
37	10	0.7042	0.7063	33.4068	5823.4937	5825.3192	5881.39800	Location DivGame SackDiff TODiff PressuresAllowed TotalPenalties NetPassYdsPerAtt ReturnYards BigPlayDiff SecondHalfRushingYds
38	8	0.7042	0.7059	31.6058	5821.6735	5823.5035	5869.04982	Location SackDiff TODiff PressuresAllowed TotalPenalties NetPassYdsPerAtt ReturnYards BigPlayDiff
39	10	0.7042	0.7062	33.7591	5823.8431	5825.6633	5881.74744	Location DivGame SackDiff TODiff PressuresAllowed TotalPenalties NetPassYdsPerAtt NetYdsAllowed ReturnYards BigPlayDiff
40	9	0.7040	0.7059	33.5378	5823.6062	5825.4182	5876.24653	Location DivGame SackDiff TODiff PressuresAllowed TotalPenalties NetPassYdsPerAtt ReturnYards BigPlayDiff
41	9	0.7040	0.7059	33.5734	5823.6415	5825.4530	5876.28179	Location SackDiff TODiff TotalPenalties NetPassYdsPerAtt NetYdsAllowed ReturnYards BigPlayDiff SecondHalfRushingYds
42	9	0.7039	0.7058	34.1377	5824.2004	5826.0042	5876.84072	SackDiff TODiff PressuresAllowed TotalPenalties NetPassYdsPerAtt NetYdsAllowed ReturnYards BigPlayDiff SecondHalfRushingYds
43	10	0.7038	0.7059	35.4917	5825.5606	5827.3545	5883.46491	Location DivGame SackDiff TODiff TotalPenalties NetPassYdsPerAtt NetYdsAllowed ReturnYards BigPlayDiff SecondHalfRushingYds
44	8	0.7038	0.7054	33.6862	5823.7329	5825.5371	5871.10914	Location SackDiff TODiff TotalPenalties NetPassYdsPerAtt ReturnYards BigPlayDiff SecondHalfRushingYds
45	8	0.7037	0.7054	33.9375	5823.9814	5825.7825	5871.35766	Location SackDiff TODiff TotalPenalties NetPassYdsPerAtt NetYdsAllowed ReturnYards BigPlayDiff
46	8	0.7037	0.7054	33.9669	5824.0105	5825.8112	5871.38675	SackDiff TODiff PressuresAllowed TotalPenalties NetPassYdsPerAtt NetYdsAllowed ReturnYards BigPlayDiff
47	10	0.7037	0.7058	36.0549	5826.1184	5827.9038	5884.02276	DivGame SackDiff TODiff PressuresAllowed TotalPenalties NetPassYdsPerAtt NetYdsAllowed ReturnYards BigPlayDiff SecondHalfRushingYds
48	8	0.7037	0.7054	34.1090	5824.1510	5825.9499	5871.52725	SackDiff TODiff PressuresAllowed TotalPenalties NetPassYdsPerAtt ReturnYards BigPlayDiff SecondHalfRushingYds
49	9	0.7036	0.7055	35.6258	5825.6731	5827.4564	5878.31340	Location DivGame SackDiff TODiff TotalPenalties NetPassYdsPerAtt ReturnYards BigPlayDiff SecondHalfRushingYds

Selection Method	Selection Value	Model Number	Number of Variables	Variables
R Squared	.7113	1	11	Location, SackDiff, TODiff, PressuresAllowed, YPC, Total Penalties, NetPassYdsPerAtt, NetYdsAllowed, ReturnYards, BigPlayDiff, SecondHalfRushingYds
C(P)	10.7873	2	10	Location, SackDiff, TODiff, PressuresAllowed, YPC, Total Penalties, NetPassYdsPerAtt, ReturnYards, BigPlayDiff, SecondHalfRushingYds
AIC	5800.8753	2	10	Location, SackDiff, TODiff, PressuresAllowed, YPC, Total Penalties, NetPassYdsPerAtt, ReturnYards, BigPlayDiff, SecondHalfRushingYds
BIC	5803.0206	7	9	Location, SackDiff, TODiff, PressuresAllowed, YPC, Total Penalties, NetPassYdsPerAtt, ReturnYards, BigPlayDiff
SBC	5844.25833	29	7	SackDiff, TODiff, YPC, Total Penalties, NetPassYdsPerAtt, ReturnYards, BigPlayDiff

To find the optimal select method for each of R square, C(p), AIC, BIC, and SBC, we have to look at the Fit Criteria plots. For R square we select the model with the largest R square value. That is the model with the index 1, circle in red. For the other selection method, we are looking for the lowest value. For C(p) and AIC that is the model with index 2, circled in red. For SBC the lowest value is the model with index 10, and for SBC the lowest value is the model with index 29. Afterwards, we can look at the summary table for all possible models and see the model index, model selection values, and variables in the model.

I will opt to use model 29. Model 29 has the smallest SBC value of all possible models, and out of the optimal models selected it has the smallest number of variables with 7. Model 29 also has R squared, C(p), AIC, and BIC values which are close to the optimal values selected for each method.

Model 29

The REG Procedure
Model: MODEL1
Dependent Variable: PtDiff

Number of Observations Read	1428
Number of Observations Used	1428

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	200526	28647	495.34	<.0001
Error	1420	82121	57.83193		
Corrected Total	1427	282647			

Root MSE	7.60473	R-Square	0.7095
Dependent Mean	-0.04482	Adj R-Sq	0.7080
Coeff Var	-16968		

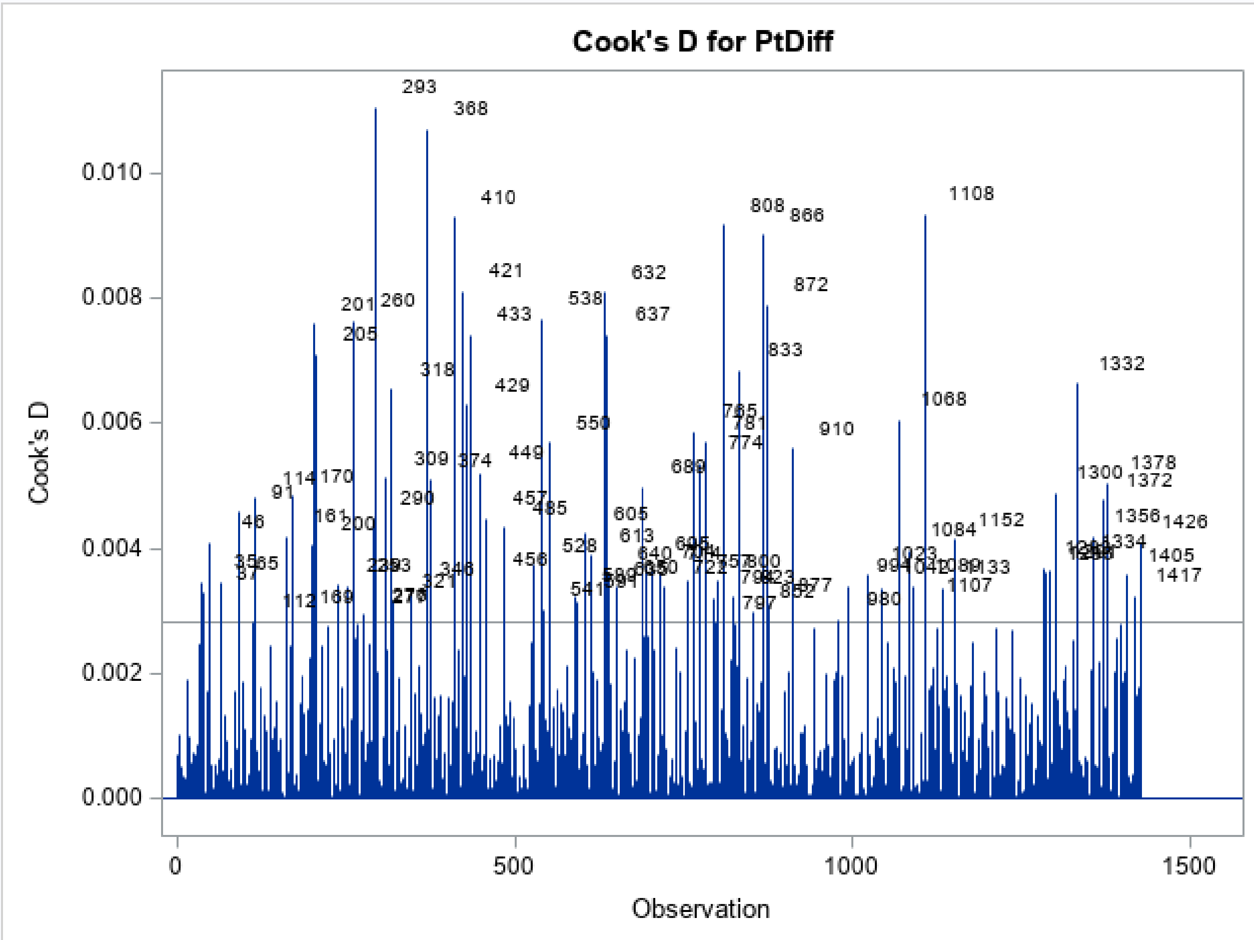
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3.44157	1.20892	2.85	0.0045
SackDiff	1	0.86745	0.08578	10.11	<.0001
TODiff	1	3.46551	0.11194	30.96	<.0001
YPC	1	-0.37847	0.07427	-5.10	<.0001
TotalPenalties	1	2.31496	0.11864	19.51	<.0001
NetPassYdsPerAtt	1	-0.04858	0.00293	-16.60	<.0001
ReturnYards	1	0.53808	0.07026	7.66	<.0001
BigPlayDiff	1	0.02445	0.00416	5.88	<.0001

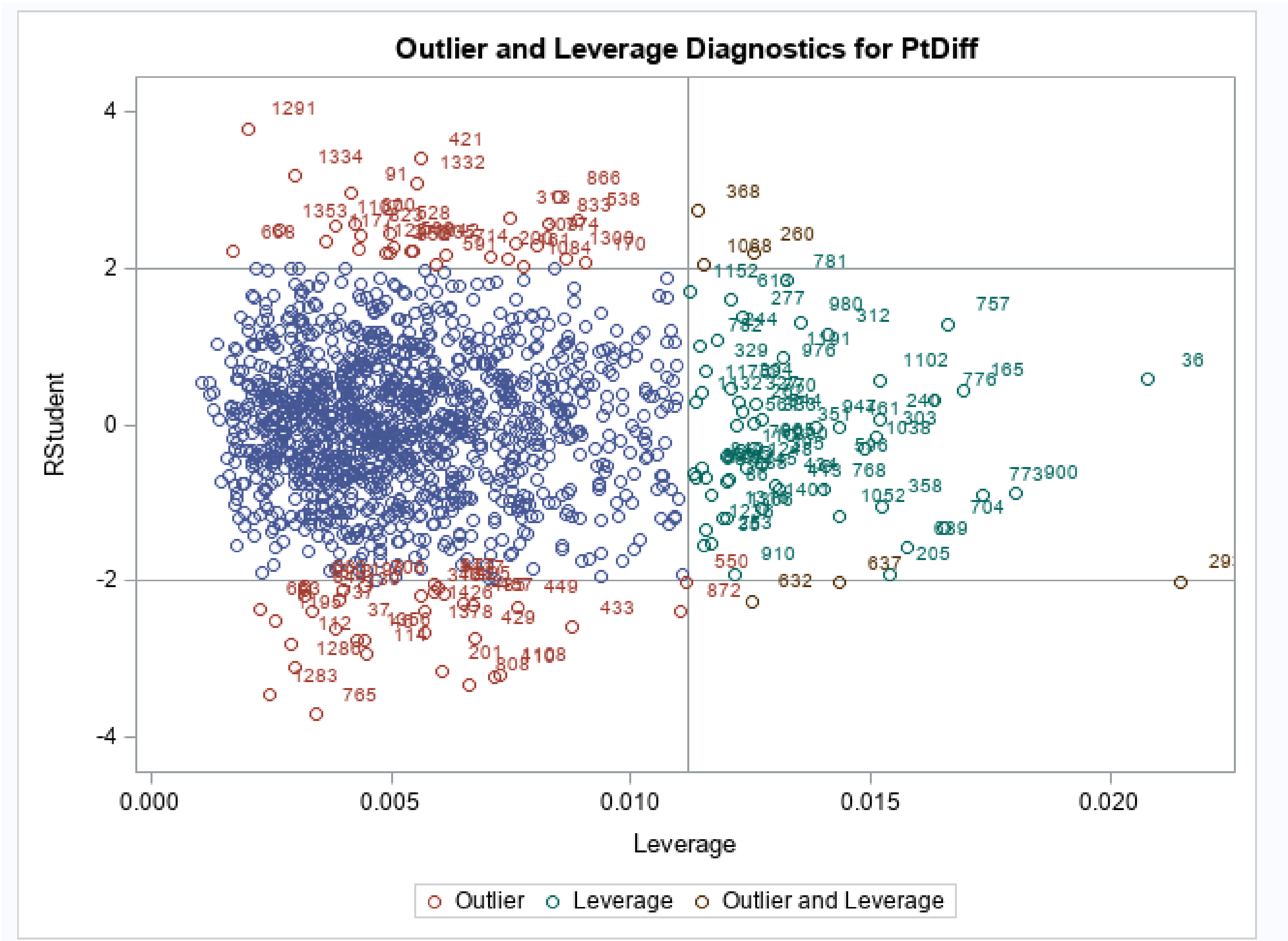
The fitted model using model 29 is:

$\hat{y}$ = 3.44157 + 0.86745(X1) + 3.46551(X2) -0.37847(X3) + 2.31496(X4) – 0.04858(X5) + 0.53808(X6) + 0.02445(X7)

Model 29

The REG Procedure
Model: MODEL1
Dependent Variable: PtDiff

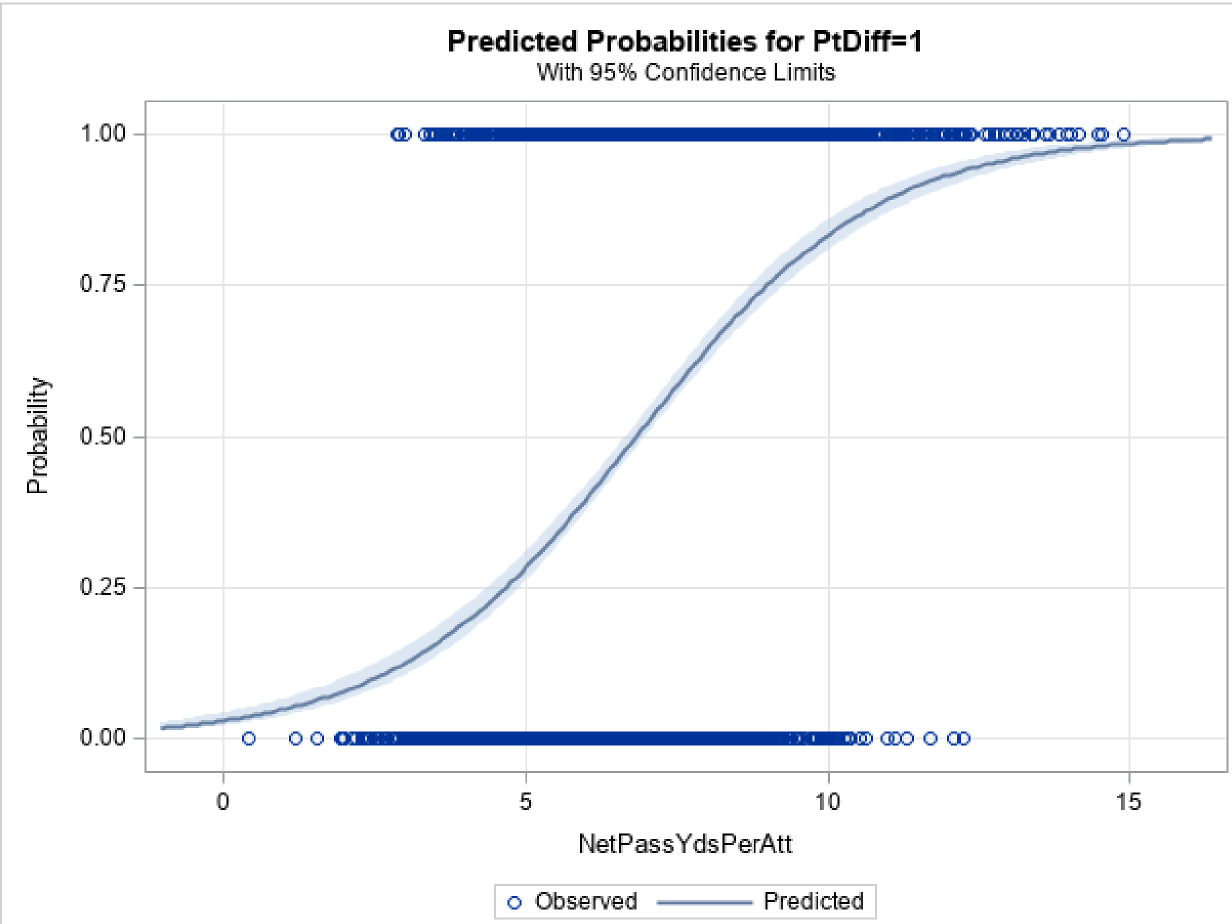



Outlier and Leverage Diagnostics for PtDiff
RStudent
Leverage
Outlier
Leverage
Outlier and Leverage

<u>Chapter 6</u>: Logistic Regression

I transformed PointDiff into 0 or 1 by splitting the variable into 0 for "low" and 1 "high" relative to the median of y. The median point differential in the population is 0. For the logistic regression is a point differential is greater than 0 it will be 1 and 0 otherwise.

First, I regressed the transformed y variable onto Net Pass Yards Per Attempt:

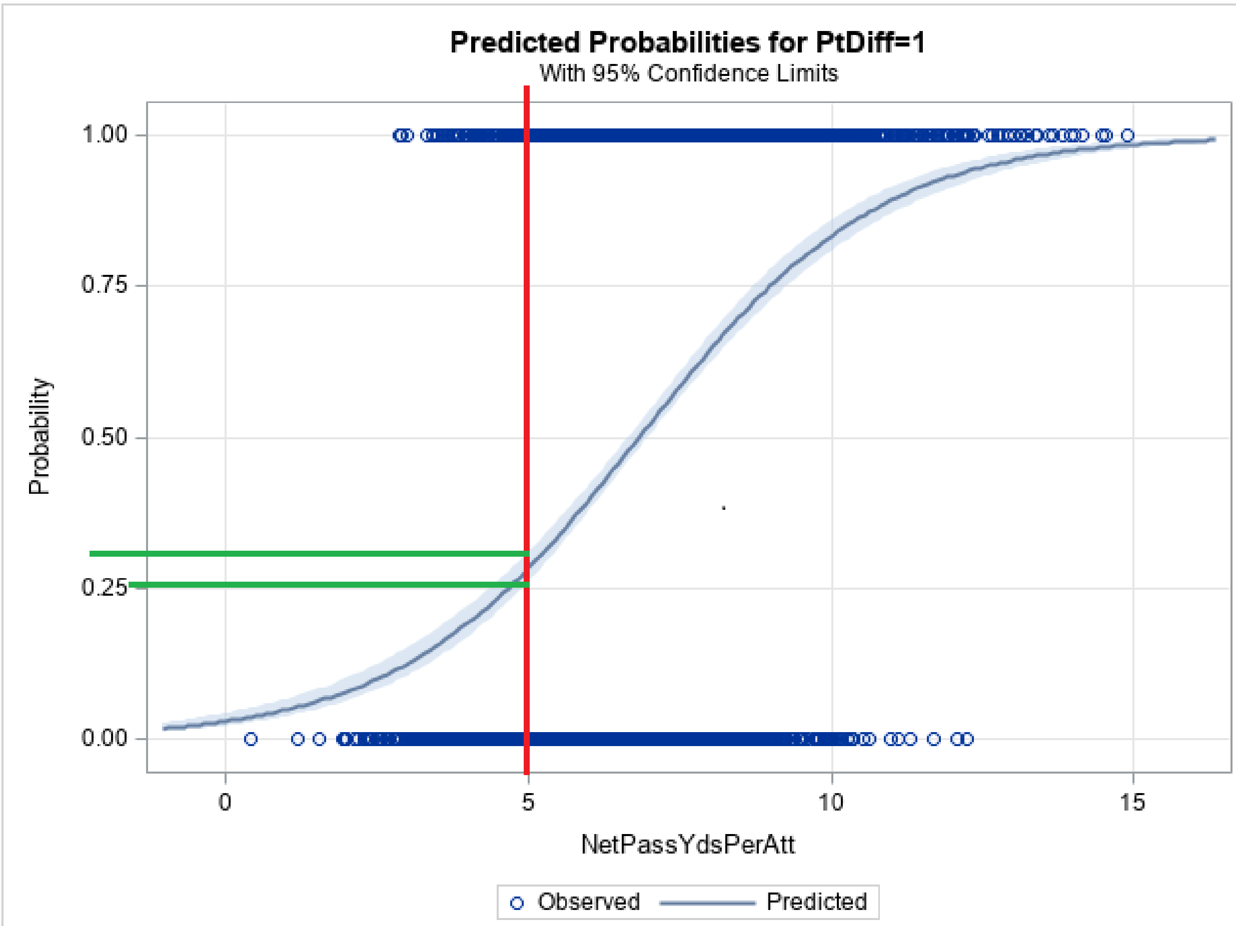


Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-3.4621	0.1801	369.6128	<.0001
NetPassYdsPerAtt	1	0.5074	0.0260	382.0616	<.0001

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	3961.206	3463.295
SC	3967.163	3475.210
-2 Log L	3959.206	3459.295

R-Square	0.1606	Max-rescaled R-Square	0.2141

On the above regression plot, select an x-value (on the horizontal axis), and add a vertical line through that point. On the graph, locate and label the endpoints for the corresponding 95% confidence interval. Explain what it is that you hope to capture with that interval.



Logistic regression estimates the expected value of y hat given x. In my dataset when point differential is greater than 0, its value is 1 and 0 otherwise. In a game when a team's point differential is positive that results in that team winning the game. In the above regression plot, when a team has 5 net passing yards per attempt in a game, we are 95% that the team's probability of winning the team will fall into the interval of the two green lines.

Using the AIC criterion, compare the performances of:

(a) the logistic regression with the one x-variable

(b) the logistic regression now using all the x-variables in the data set

(c) the logistic regression model chosen by using forward stepwise procedure starting with all x-variables in the data set.

(A) Logistic Regression with only Net Pass Yards Per Attempt:

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	3961.206	3463.295
SC	3967.163	3475.210
-2 Log L	3959.206	3459.295

R-Square	0.1606	Max-rescaled R-Square	0.2141

(B) Logistic Regression with all 12 x variables in the data set:

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	3961.206	1889.983
SC	3967.163	1967.426
-2 Log L	3959.206	1863.983

(C) Logistic regression model chosen using forward stepwise procedure starting with all x-variables in that data set. SAS found 9 x variables that met the 0.05 significance level for placement in the model.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	3961.206	1886.459
SC	3967.163	1946.031
-2 Log L	3959.206	1866.459

When we compare AIC values for the intercept and covariates, we can see that the logistic regression with only net pass yards per attempt has an AIC criterion of 3463.295. When we include all 12 possible regressors into the logistic model we have an AIC criterion of 1889.983 for the intercept and covariates. Lastly when we use forward stepwise selection we have an AIC criterion of 1886.459 for only 9 possible regressors.

Summary of Forward Selection					
Step	Effect Entered	DF	Number In	Score Chi-Square	Pr > ChiSq
1	TODiff	1	1	765.8591	<.0001
2	BigPlayDiff	1	2	457.5526	<.0001
3	SackDiff	1	3	225.2812	<.0001
4	NetPassYdsPerAtt	1	4	153.5884	<.0001
5	NetYdsAllowed	1	5	163.4600	<.0001
6	SecondHalfRushingYds	1	6	40.1441	<.0001
7	TotalPenalties	1	7	35.8350	<.0001
8	PressuresAllowed	1	8	13.1733	0.0003
9	Location	1	9	12.6364	0.0004

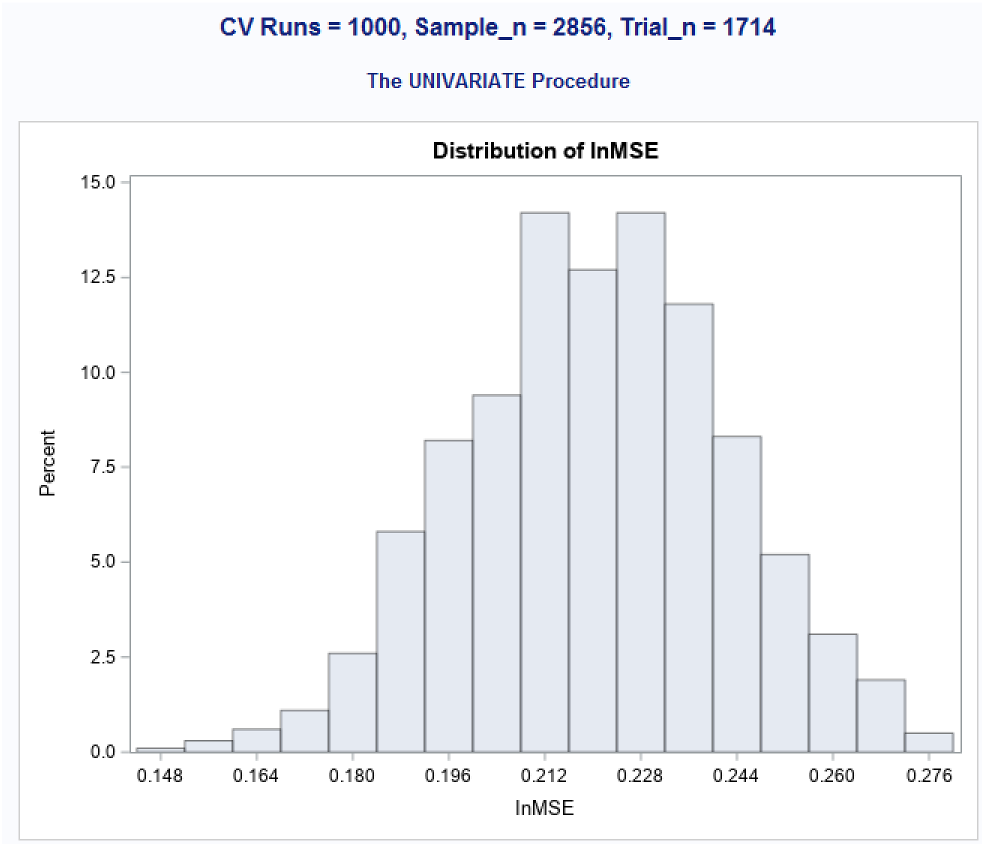
Selected Topic – Cross Validation

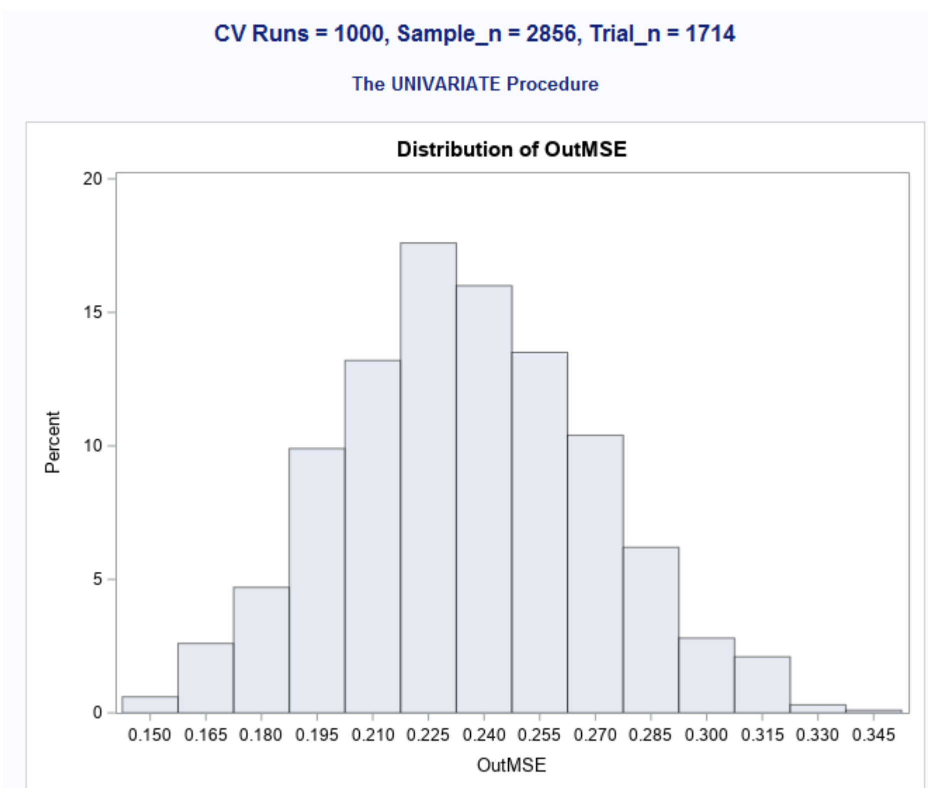
I chose to perform cross validation on my data set using the optimal model I selected in section 5. The optimal model consists of 7 variables: sack differential, turnover differential, yards per carry, total penalties, net passing yards per attempt, return yards, and big play differential.

Cross validation is a regression methodology where a dataset is divided into two separate datasets, a training dataset and a validation dataset. We randomly select a portion of the dataset to be our training data and the rest of the data is the validation data on which we test our model. When we create a model, we would like to test the quality of the model with new data and cross validation makes this possible. A model is formed from the training dataset and then we test the quality of the model using the validation dataset. We test the validation dataset by inputting the x values of the validation data into our regression model. To see whether or not our model is effective we look at the training MSE and the validation MSE.

The MSE of the validation data is obtaining by finding the y hat values of the validation data by using the model we created using the training dataset. We input the x values of the validation data into the model obtaining y hat values for each observation. Afterwards we subtract the y hat values from the known y values of the validation data to obtain the residuals. Then by taking the square and adding the residuals we have the MSE of the validation data. When comparing the MSE of the training data and the MSE of the validation data, if the values are similar then the model performed as effectively on the training data as it did on the validation data. However, we must note that the MSE of the training will often be less than the MSE of the validation data since the coefficients of the variables in the model are computed in order to minimize MSE.

For the cross validation I elected to use 60% of the original dataset as the training data and the other 40% as the validation dataset. Since I had 2856 observations in the original dataset, the training data consisted of 1714 observations and the validation dataset consisted of 1142. The distribution of the MSE of both the training and validation are below.


CV Runs = 1000, Sample_n = 2856, Trial_n = 1714
The UNIVARIATE Procedure
Distribution of InMSE
Percent
15.0
12.5
10.0
7.5
5.0
2.5
0.0
0.148
0.164
0.180
0.196
0.212
0.228
0.244
0.260
0.276
InMSE


CV Runs = 1000, Sample_n = 2856, Trial_n = 1714
The UNIVARIATE Procedure
Distribution of OutMSE
Percent
20
15
10
5
0
0.150
0.165
0.180
0.195
0.210
0.225
0.240
0.255
0.270
0.285
0.300
0.315
0.330
0.345
OutMSE

The MEANS Procedure

Variable	Mean	Std Dev	N
InMSE	0.2206331	0.0219979	1000
OutMSE	0.2351087	0.0347529	1000

As we can see the mean MSE of the training data is lower than the validation data, however the values are very similar to each other. Therefore, we can say model 29 from section 5 performed as effectively on the training dataset as it did on the validation dataset.