A LINEAR REGRESSION ANALYSIS ON NCAA DIVISION I FOOTBALL DATA

NATHAN AKERHIELM

Abstract. NCAA Division I college football games are highly watched and advertised, and they generate significant revenue for the respective schools and television networks. These games are also of high interest to sports gamblers. Bets are placed at sportsbooks on whether which team will cover the point spread. Using a regression analysis, I will build a model that predicts whether a team will cover the point spread in a given game using offensive and defensive game data. The data used to build the model is from the 2018 Atlantic Coast Conference (ACC) regular season, and then the model is validated using data from the 2019 ACC regular season.

Contents

Date: December 7, 2021.

This document is a senior thesis submitted to the Mathematics and Statistics Department of Haverford College in partial fulfillment of the requirements for a major in Mathematics.

1. Introduction

The Football Bowl Subdivision (FBS) is the top level of NCAA Division I football in the United States, and consists of 130 teams from 10 conferences. My thesis fouces on the point spread of the FBS games. A *point spread* (P.Spread) is used in sports betting to even the odds between two unevenly matched teams. Each team is given a point total (+/-) by a Las Vegas oddsmaker that is added to the final score, thus factoring into if the bet was won or lost. The stronger team is indicated by a minus (-) sign, and will have a certain number of points taken away from the final score of the game, while the weaker team is indicated by a plus (+) sign and will have the same number of points added to the final score [36].

For example, suppose Boston College is playing Wake Forest. I consider this example from Boston College's perspective because the data is arranged in ascending alphabetical order (A-Z), and 'B' comes before 'W'. Below is a portion of the dataset from the Boston College vs. Wake Forest game to help illustrate the point spread concept.

Team	Opponent	Team.Score	Opp.Score	P.Spread
Boston College	Wake Forest	41	34	-4.5

Table 1. Boston College vs. Wake Forest Data

As reported in Table 1, the point spread is "Boston College -4.5", which means that Boston College is favored to win the game by at least 4.5 points. It must be noted that this point spread is placed on the game *before* it occurs. The final score is Boston College: 41 and Wake Forest: 34. To determine if Boston College has "covered the spread", one simply computes:

$$\text{Cover.Spread} = \text{Team.Score}–\text{Opp.Score} + \text{P.Spread}. \tag{1.1}$$

If Cover.Spread > 0, then Boston College covers the spread, and those who bet on Boston College win while those that bet on Wake Forest lose. If Cover.Spread < 0, Boston College does not cover the spread, and those who bet on Boston College lose, while those that bet on Wake Forest win. If

Cover.Spread = 0, then this is called a push, and bettors receive their money back. Often times, an extra half point is added to or subtracted from the point spread to ensure a push does not happen. In this example, we see that Cover.Spread $= 41 - 34 + (-4.5) = 2.5 > 0$, so Boston College covers the spread, i.e. those that bet on Boston College win.

In my thesis, the response variable is Cover.Spread, where

$$\text{Cover.Spread} = \text{Team.Score} - \text{Opp.Score} + \text{P.Spread}. \tag{1.2}$$

The goal of my thesis is to answer the question: What factors are useful in predicting Cover.Spread?

2. Methods

In Section 2, I will explain the theory that I use to analyze my data, specifically the linear regression model. I list several important definitions and theorems that I will use later.

2.1. The Linear Regression Model.

The multiple regression model is commonly used and well understood. The linear regression model for $p-1$ predictors is given by ([3], p.221):

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_{p-1} x_{i,p-1} + \epsilon_i, \ i = 1, 2, \ldots, n \tag{2.1}$$

In Equation (2.1), the $\beta_0, \beta_1, \ldots \beta_{p-1}$ are unknown parameters. Y_i represents the response variable (Cover.Spread) for the ith data point. Furthermore, $x_{i1}, \ldots, x_{i,p-1}$ are **known** covariates. Denote X_j as the jth column in the matrix $\mathbf{X}$ defined below, where $X_j = [x_{1j}, x_{2j}, \ldots, x_{nj}]^T$. Additionally, ϵ_i are random errors. Note that $Y_i > 0$ means that those that bet on the ith game win, while $Y_i < 0$ indicates those that bet on the ith game lost.

In order to express the general multiple regression model (2.1) in matrix terms, define ([3], p.222-23):

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{n \times 1}$$

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1,p-1} \\ 1 & x_{21} & x_{22} & \ldots & x_{2,p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{n,p-1} \end{bmatrix}_{n \times p}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}_{p\times 1} \qquad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}_{n\times 1}$$

Therefore, the model (2.1) in matrix terms is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{2.2}$$

The assumptions of the linear regression are as follows ([24], p.101):

(1) *Linearity*: $E(\epsilon_i) = 0,\ i = 1, 2, \dots n.$
(2) *Constant variance*: $\mathrm{Var}(\epsilon_i) = \sigma^2$ for all i.
(3) *Independence*: ϵ_i, ϵ_j are independent for $i \neq j$.
(4) *Normality*: $\epsilon_i \sim N(0, \sigma^2)$.
(5) *$X_1, X_2, \dots, X_p$, the column vectors of the matrix $\boldsymbol{X}$ defined above, are not linearly correlated.*

Definition 2.1. ([3], p.181, 194) Given a random vector $\mathbf{X} = [X_1, X_2, \dots, X_n]^T$, the *expected value of a random vector* with *expectation*, $\boldsymbol{\mu}$, a vector of length n, is

$$E(\mathbf{X}) = \begin{bmatrix} E(X_1) \\ E(X_2) \\ \dots \\ E(X_n) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix} = \boldsymbol{\mu}.$$

Since $E(\epsilon_i) = 0$ by Assumption 1 and, by Definition 2.1, the random vector $\mathbf{Y}$ has expected value:

$$\begin{aligned} E(\mathbf{Y}) &= E(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) \\ &= \mathbf{X}\boldsymbol{\beta} + E(\boldsymbol{\epsilon}) \\ &= \mathbf{X}\boldsymbol{\beta}. \end{aligned}$$

Definition 2.2. ([3], p.194) The *variance-covariance matrix* of an $n \times 1$ random vector $\mathbf{Y} = [Y_1, Y_2, \dots, Y_n]^T$, denoted $\mathrm{Var}(\mathbf{Y})$, with $\boldsymbol{\mu} = E[\mathbf{Y}]$ is defined as

$$\mathrm{Var}(\mathbf{Y}) = E\left[(\mathbf{Y} - \boldsymbol{\mu})(\mathbf{Y} - \boldsymbol{\mu})^T\right].$$

In matrix form,

$$\mathrm{Var}(\mathbf{Y}) = \begin{bmatrix} \mathrm{Var}(Y_1) & \mathrm{Cov}(Y_1, Y_2) & \dots & \mathrm{Cov}(Y_1, Y_n) \\ \mathrm{Cov}(Y_2, Y_1) & \mathrm{Var}(Y_2) & \dots & \mathrm{Cov}(Y_2, Y_n) \\ \vdots & \vdots & & \vdots \\ \mathrm{Cov}(Y_n, Y_1) & \mathrm{Cov}(Y_n, Y_2) & \dots & \mathrm{Var}(Y_n) \end{bmatrix}_{n\times n}$$

Examining the variance-covariance matrix of $\boldsymbol{\epsilon}$, by Assumption 3, $\text{Cov}(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$, so all of the off-diagonal entries are zero. By Assumption 4, $\text{Var}(\epsilon_i) = \sigma^2$ for all i. The variance-covariance matrix of $\boldsymbol{\epsilon}$ ([3], p.223) is

$$\text{Var}(\boldsymbol{\epsilon}) = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix} = \sigma^2 \mathbf{I}.$$

where $\mathbf{I}$ is the $n \times n$ identity matrix.

Since $\mathbf{X}\boldsymbol{\beta}$ are not random, but rather are fixed constants, it follows that

$$\text{Var}(\mathbf{Y}) = \text{Var}\left(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}\right) = \text{Var}\left(\boldsymbol{\epsilon}\right) = \sigma^2 \mathbf{I}.$$

The following theorem will be helpful in subsequent work.

Theorem 2.3. *([3], p.196) Let $\boldsymbol{A} = [a_{ij}]$ be a $k \times n$ matrix of constants, and $\boldsymbol{Y}$ be an $n \times 1$ random vector such that $\boldsymbol{Y} = [Y_1, Y_2, \dots, Y_n]^T$. Then,*

(1) $E[\boldsymbol{A}] = \boldsymbol{A}$
(2) $E[\boldsymbol{A}\,\boldsymbol{Y}] = \boldsymbol{A} E[\boldsymbol{Y}]$
(3) $Var(\boldsymbol{A}\,\boldsymbol{Y}) = \boldsymbol{A}\, Var(\boldsymbol{Y}) \boldsymbol{A}^T$

Proof. The proof of (1) is trivial because the expected value of a constant is always a constant. For (2), define $\mathbf{X} = \mathbf{AY}$, a random vector of length k. In particular,

$$\mathbf{X} = \begin{bmatrix} \sum a_{1j} Y_j \\ \sum a_{2j} Y_j \\ \vdots \\ \sum a_{kj} Y_j \end{bmatrix}.$$

It follows that

$$E(\mathbf{X}) = \begin{bmatrix} E\left(\sum a_{1j} Y_j\right) \\ E\left(\sum a_{2j} Y_j\right) \\ \vdots \\ E\left(\sum a_{kj} Y_j]\right) \end{bmatrix} = \begin{bmatrix} \sum E\left(a_{1j} Y_j\right) \\ \sum E\left(a_{2j} Y_j\right) \\ \vdots \\ \sum E\left(a_{kj} Y_j\right) \end{bmatrix}$$

$$= \begin{bmatrix} \sum a_{1j} E(Y_j) \\ \sum a_{2j} E(Y_j) \\ \vdots \\ \sum a_{kj} E(Y_j) \end{bmatrix} = \mathbf{A} E\left(\mathbf{Y}\right).$$

Letting $\boldsymbol{\mu} = E(\mathbf{Y}) = [E(Y_1), E(Y_2), \ldots, E(Y_n)]^T$, I prove (3) as follows.

$$\begin{aligned}
\mathrm{Var}(\mathbf{X}) &= \mathrm{Var}(\mathbf{AY}) \\
&= E\left[(\mathbf{AY} - \mathbf{A}\mu)(\mathbf{AY} - \mathbf{A}\mu)^T\right] \\
&= E\left[\mathbf{A}(\mathbf{Y} - \mu)(\mathbf{A}(\mathbf{Y} - \mu))^T\right] \\
&= E\left[\mathbf{A}(\mathbf{Y} - \mu)(\mathbf{Y} - \mu)^T \mathbf{A}^T\right] \\
&= \mathbf{A}E\left[(\mathbf{Y} - \mu)(\mathbf{Y} - \mu)^T\right]\mathbf{A}^T \\
&= \mathbf{A}\mathrm{Var}(\mathbf{Y})\mathbf{A}^T.
\end{aligned}$$

□

2.2. **Method of Least Squares.** The method of least squares obtains the estimated parameters by minimizing the sum of the squared residuals as follows ([3], p.223):

$$Q = \sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1 X_{i1} - \ldots - \beta_{p-1} X_{i,p-1}\right)^2. \tag{2.3}$$

The least squares estimators seeks the values of $\beta_0, \beta_1, \ldots, \beta_{p-1}$, denoted by the vector $\hat{\beta}$, that minimize Q. In matrix form,

$$Q = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}).$$

Theorem 2.4. *([3], p.223) The least squares estimator, $\hat{\beta}$, that minimizes Q is $\hat{\beta} = \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{Y}$.*

Proof. To minimize Q, I take the partial derivative of Q with respect to $\boldsymbol{\beta}$ and set it equal to zero:

$$\begin{aligned}
\frac{\partial Q}{\partial \boldsymbol{\beta}} &= \frac{\partial}{\partial \boldsymbol{\beta}}\left[(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})\right] \\
&= \frac{\partial}{\partial \boldsymbol{\beta}}\left(\mathbf{Y}^T\mathbf{Y} - \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}\right) \\
&= \frac{\partial}{\partial \boldsymbol{\beta}}\left(\mathbf{Y}^T\mathbf{Y} - 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{Y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}\right) \\
&= -2\mathbf{X}^T\mathbf{Y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \\
&= 0
\end{aligned}$$

Note that the jump from the second to third line can be made because both $\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{Y}$ and $\mathbf{Y}^T\mathbf{X}\boldsymbol{\beta}$ are 1×1 matrices. Hence, scalars equal their own

transpose, which yields $2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{Y}$. Solving the above equation to find the desired $\hat{\beta}$:

$$\begin{aligned}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} &= \mathbf{X}^T\mathbf{Y}\\ \hat{\beta} &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}.\end{aligned}$$

Next, I show that $\hat{\beta}$ is a minimum, not a maximum. To do so, I take the second partial derivative of Q with respect to $\boldsymbol{\beta}$ in order to find the Hessian matrix.

$$\begin{aligned}\frac{\partial^2 Q}{\partial\boldsymbol{\beta}^2} &= \frac{\partial}{\partial\boldsymbol{\beta}}\left(-2\mathbf{X}^T\mathbf{Y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}\right)\\ &= 2\mathbf{X}^T\mathbf{X}\end{aligned}$$

I must show that $2\mathbf{X}^T\mathbf{X}$ is positive definite in order to prove that $\hat{\beta}$ is a minimum. In particular, for any non-zero vector, v, of length p, I want to prove that $v^T 2\mathbf{X}^T\mathbf{X}v > 0$. By Assumption 5, it follows that the columns of $\mathbf{X}$ are linearly independent, so the equation $\mathbf{X}v = 0$ if and only if $v = 0$. Thus,

$$\begin{aligned}v^T\left(2\mathbf{X}^T\mathbf{X}\right)v &= 2v^T\left(\mathbf{X}^T\mathbf{X}\right)v\\ &= 2\left(\mathbf{X}v\right)^T\left(\mathbf{X}v\right)\\ &> 0\end{aligned}$$

Therefore, $\hat{\beta}$ minimizes Q. □

Theorem 2.5. *$\hat{\beta}$ is unbiased, i.e. $E(\hat{\beta}) = \boldsymbol{\beta}$.*

Proof.

$$\begin{aligned}E\left(\hat{\beta}\right) &= E\left(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}\right)\\ &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T E(\mathbf{Y})\\ &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\left(\mathbf{X}\boldsymbol{\beta}\right)\\ &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\left(\mathbf{X}^T\mathbf{X}\right)\boldsymbol{\beta}\\ &= \boldsymbol{\beta}\end{aligned}$$

□

Theorem 2.6. *([3], p.207) $Var(\hat{\beta}) = \sigma^2\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$*

Proof.

$$\begin{aligned}
\mathrm{Var}(\hat{\beta}) &= \mathrm{Var}\left(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}\right) \\
&= \left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right]\mathrm{Var}(\mathbf{Y})\left[\left(\mathbf{X}^T\mathbf{X}^{-1}\right)\mathbf{X}^T\right]^T \\
&= \left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right]\left[\sigma^2(\mathbf{I})\right]\left[\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right] \\
&= \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\left(\mathbf{X}^T\mathbf{X}\right)\left(\mathbf{X}^T\mathbf{X}\right)^{-1} \\
&= \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}
\end{aligned}$$

□

3. Data

In Section 3, I will describe where I obtained my data and the steps I performed in order to merge and validate the data.

3.1. **Data Extraction.** There are 130 teams in the FBS, and roughly half of the teams belong to five major conferences, commonly referred to as the Power 5 Conferences. To narrow my focus, I analyze the games from one of the Power 5 Conferences, called the Atlantic Coast Conference (ACC), during the 2018 season. The 15 teams in the ACC are reported in Table 2. Notre Dame has an asterisk because, although it is a member of the ACC for

Team
Boston College
Clemson
Duke
Florida State
Georgia Tech
Louisville
Miami (FL)
North Carolina
North Carolina State
Notre Dame*
Pittsburgh
Syracuse
Virginia
Virginia Tech
Wake Forest

Table 2. 2018 ACC Teams

all other sports, it is not an official member of the ACC for football due to their exclusive TV contract with NBC. Notre Dame does, however, play five ACC teams each year, so I include them in the ACC for the purpose of my thesis.

Data from the 2018 regular season is used to build a model that uses college football game data in order to predict Cover.Spread. For a data source, I turn to Sports-Reference [1], which provides a comprehensive approach to data across many different sports, including professional baseball, basketball, football, and soccer, as well as college basketball and football.

All teams are guaranteed to play 12 regular season games each season, and the Sports-Reference dataset provides game-by-game data for each of the 15 teams in the ACC. I extracted the data based on ascending alphabetical order (A-Z), beginning with Boston College, then Clemson, etc.

3.2. **Data Merger.** The data on Sports-Reference contains two datasets, one for offensive data and one for defensive data. Both the offensive and defensive datasets have a total of 26 variables for a given team, with the *same* variable names. For both datasets, there are five variables that are repeating for both the offense and the defense. They are the game, date, opponent, location of the game, and result. The remaining 21 variables are variables that describe the defensive or offensive data. For example, Pass.Cmp represents the number of completed passes thrown by the quarterback. The following steps were taken to merge the data.

- I relabeled variables to differentiate between the offensive and defensive variable. For example, I denoted "Off.Pass.Cmp" to represent the number of completions thrown by the team's quarterback and "Def.Pass.Cmp" to represent the number of completions allowed to the opposing quarterback.

- I then added the 21 defensive variables into the accompanying columns in the offensive dataset, giving me $26 + 21 = 47$ variables in my dataset.

- Then, I deleted the location variable, for the location is already factored into the point spread created by the Las Vegas oddsmakers. For example, if Boston College is playing Wake Forest at Wake Forest's stadium, the oddsmakers will take this into account when creating the point spread. Therefore, it would be repetitive to include the location. This left me with $47 - 1 = 46$ variables.

- I added the variable "Team" the first column of the dataset in order to differentiate between the 15 ACC teams. This gave me $46+1 = 47$ variables.

- Originally, the Sports Reference data had the result for a team's game in a single column. For example, under "Result", it says "W(20-10)", indicating the team won the game by a score of 20 to 10. Consequently, I separated the "Result" variable into two variables, one named "Team.Score" and the other named "Opp.Score". For this example, I placed 20 under "Team.Score" and 10 under "Opp.Score". This left me with a total of $47 + 1 = 48$ variables.

- I also added the point spread to the dataset, labeling the variable "P.Spread". Sports-Reference did not provide the point spreads for the games, and I was not able to find the point spreads from all of the 2018 games in a single place. Rather, I was able to find the point spreads for games 1-4 on the SB Nation website ([20]-[23]), and the spreads for the remaining on Sports Illustrated website ([26]-[34]). This left me with $48 + 1 = 49$ variables.

- Furthermore, I deleted any observations from postseason games because I was only interested in a team's 12 regular season games.

- Finally, I removed repeat data values for teams in the ACC that played each other. For example, the data from the Clemson vs. Boston College game, both of whom are members of the ACC, appeared in both the Clemson and Boston College datasets as the mirror image of each other, so I deleted the data from Clemson's perspective since Boston College comes before Clemson in the alphabet.

After this stage, the dataset had a **total of 49 variables with 118 distinct games** played by teams in the ACC, and it must be noted that some of these games were played against opponents who are *not* members of the ACC.

3.3. **Data Validation.** Once the steps in Section 3.2 were complete, I performed the following data validation.

- Calculating the averages in Excel and compared these values with those provided by Sports-Reference for variables with an average (e.g. Off.Avg.Run.Yds). I concluded that all of the averages matched and then replaced the Sports-Reference value with my calculation because it provided more decimal places.

- Calculating any variable that involved a percentage (e.g. Off.Pass.Pct) then comparing to the Sports-Reference values. I concluded that all of the percentages matched and then replaced the Sports-Reference value with my calculation because it provided more decimal places.
- Calculating any variable that involved a total (e.g Off.Total.TO) and comparing to the Sports-Reference data. I concluded that all of the totals matched.

There were no discrepancies between my calculations and the Sports-Reference dataset.

3.4. **Choice of Covariates.** Recall that

$$\text{Cover.Spread} = \text{Team.Score} - \text{Opp.Score} + \text{P.Spread}.$$

Once Cover.Spread is calculated, the three variables, namely Team.Score, Opp.Score, and P.Spread, are no longer used in building the model. The following steps are taken to eliminate variables that are not relevant in predicting Cover.Spread.

(1) **Four variables are not included in the analysis**; Team, Game, Date, and Opponent were removed from consideration.

(2) Variables used to calculate an average are not included in the analysis. For example Off.Run.Att (total number of running attempts by the offense) and Off.Run.Yds (total number of running yards by the offense) are excluded because they were used to generate Off.Run.Avg.Yds (Total yards by team's offense/Total number of team's offensive plays). There are a total of four average variables in the dataset, each of which is generated by two variables, so **a total of 8 more variables are not included in my analysis**. For the remainder of Section 3.4, please reference the data dictionary in Table 12 for an explanation of the covariates.

(3) Variables used to calculate a percentage are not included. For example, Off.Cmp and Off.Att are excluded because they are used to generate Off.Pass.Pct. There are a total of two percentage variables in the dataset, both of which are generated by two variables, so **a total of 4 more variables are not included in my analysis**.

(4) Variables used to calculate a total are not included in my analysis. For example, Off.Fum and Off.Int are excluded because they are summed together to find Off.Total.TO. There are total of four variables that involved a total in the dataset, two of which are generated by two variables and two of which are generated by three variables. Thus, **a**

total of 10 more variables are not included in my analysis.

As a result, I am left with a total of $49 - 3 - 4 - 8 - 4 - 10 = 20$ covariates and one response variable, for a total of **21 variables**.

4. Preliminary Data Analysis

In Section 4, I will provide preliminary analysis for my response variable, Cover.Spread, as well as my covariates.

4.1. Preliminary Analysis for the Response Variable: Cover.Spread.

The preliminary data analysis begins with an analysis of the response variable, Cover.Spread. The summary statistics are reported in Table 3.

Min	Q1	Mean	Median	Q3	Max	sd	Outlier
-64.5	-10.875	-0.288	0.25	-0.288	13	17.34	Duke vs. Wake Forest

Table 3. Summary Statistics of Cover.Spread

Table 3 shows that the median for Cover.Spread is 0.25 and the mean is -0.288, both of which are close to zero. When Cover.Spread=0, this means that neither team covers the spread. Thus, the median and mean values indicate the the Las Vegas oddsmakers measure is relatively accurate, especially considering the fact that many of the point spreads are given an extra half point in order to ensure that one team has to cover the spread. The standard deviation is 17.34, and the interquartile range is 10.587. Cover.Spread ranges from -64.5 to 13.

Figure 1 clearly indicates that Cover.Spread is left-skewed with one outlier. The outlier comes from the 34th observation of the ACC dataset, which is the game Duke vs. Wake Forest. Duke was favored to win the game by 12.5 points, but lost 7 to 59. The value of the outlier is $7 - 59 + (-12.5) = -64.5$. This outlier can be seen in the boxplot in Figure 1 below.

4.2. Preliminary Analysis of Covariates.

All of the 20 covariates are numerical variables, 14 of which are continuous and 6 of which are discrete. Note that for discrete variables with 10 or more categories, I treat them as continuous. Off.Tot.First.Down, Off.No.Pen, Def.Tot.First.Down, and Def.No.Pen are the four variables I treat as continuous because they have more than 10 categories, denoted by a * in Table 4, which reports the summary statistics for the 14 continuous covariates.

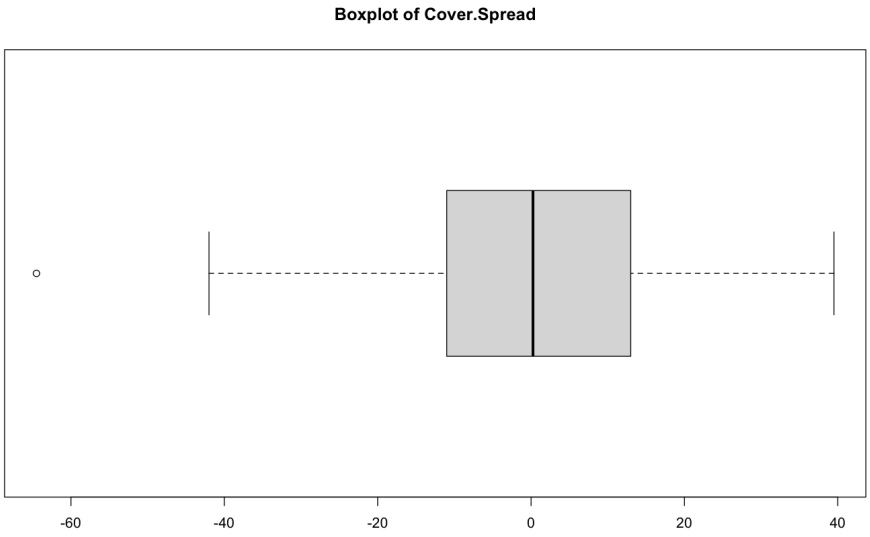

FIGURE 1. Boxplot of Cover.Spread

Variable	Mean	sd	Median	Min	Max	Shape	Outliers
Off.Pass.Pct	59.7	13.8	60.7	0	100	Roughly Symmetric	50,51,53,70
Off.Pass.Yds	232.3	99.9	240.5	0	473	Roughly Symmetric	None
Off.Run.Avg.Yds	4.9	2.2	4.8	0.3	13.7	Skewed to the right	18,21,102
Off.Avg.Yds.Play	6.2	1.7	6.1	2	13.9	Skewed to the right	10,21, 102
Off.Tot.First.Down *	21.5	5.8	21.0	8	38	Skewed to the right	23
Off.No.Pen *	5.9	3.2	6.0	0	17	Skewed to the right	42,60,61,62,71,98
Off.Pen.Yds	51.4	28.2	50.0	0	134	Skewed to the right	60,62,71,99
Def.Pass.Pct	56.4	11.1	57.0	25	80	Roughly Symmetric	18,56
Def.Pass.Yds	222.6	97.7	216.5	37	510	Skewed to the right	23,26,83,114
Def.Run.Avg.Yds	3.9	1.8	3.9	-0.6	9.3	Roughly Symmetric	20,31,33,59,104
Def.Avg.Yds.Play	5.3	1.4	5.3	1.3	9.7	Roughly Symmetric	31,64,104
Def.Tot.First.Down *	19.9	6.1	20.0	4	34	Skewed slightly to the left	None
Def.No.Pen *	5.5	2.7	5.0	0	16	Skewed to the right	20
Def.Pen.Yds	46.8	25.3	45.0	0	134	Skewed to the right	20,33

TABLE 4. Summary Statistics for Continuous Covariates

Examining the first row in Table 4, Off.Pass.Pct represents the number of completed passes thrown by the team divided by the total number of passes thrown by the team, multiplied by 100%. The distribution is relatively symmetric with outliers on both sides, where the outliers column refers to the observation number of the outlier. The typical success rate of the offensive passes is about 60% for a game, with a standard deviation of 13.8%. The

minimum value for Off.Pass.Pct is 0%, which means the team did not complete any of its passes. This outlier corresponds to observation 50, which is the game Georgia Tech vs. Virginia Tech. During the 2018 season, Georgia Tech used an offensive scheme that relied heavily on running the ball, and consequently, Georgia Tech had very few pass attempts each game. In the game against Virginia Tech, they only attempted one pass, which was incomplete, yielding Off.Pass.Pct= 0%.

The maximum value for Off.Pass.Pct is 100%, meaning the team's quarterback completes every throw in a game. This value corresponds to observation 51, which was the game between Georgia Tech vs. North Carolina. Georgia Tech attempted only two passes in the game, completing both of them, which yields the Off.Pass.Pct value of 100%. The remaining two outliers come from observations 53 (in which Georgia Tech completed only 1 out of 8 passes vs. Virginia), and 70 (in which Miami (FL) completed 6 out of 24 passes vs. Pittsburgh). The remaining 13 continuous covariates are analyzed using the same logic, and the shape of all 14 covariates can be visualized more clearly in Figures 8 and 9 of the Appendix.

Furthermore, Figures 2 and 3 show the scatter plots for each of the 14 continuous covariates plotted against Cover.Spread.

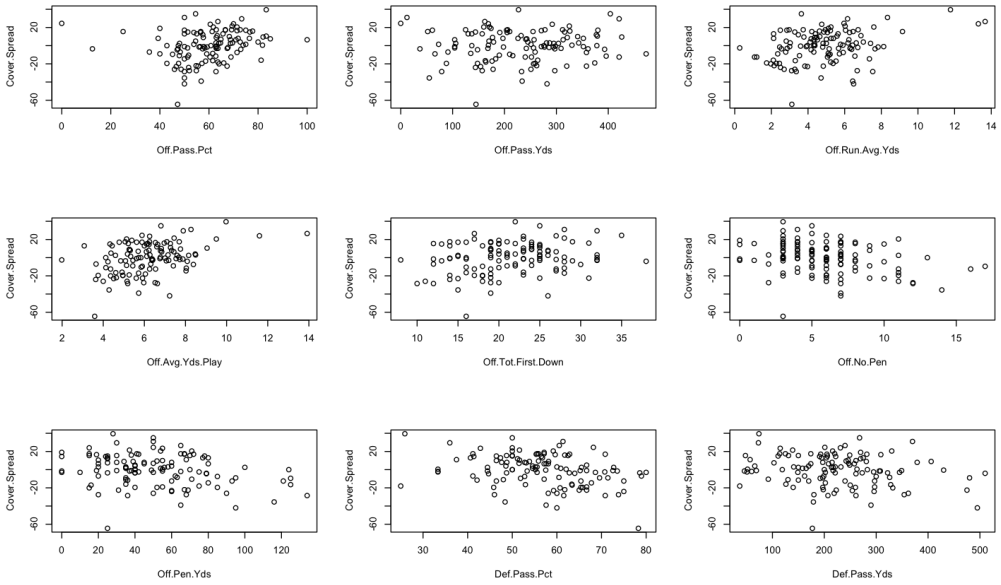

FIGURE 2. Scatter Plot of 9 Continuous Covariates vs. Cover.Spread

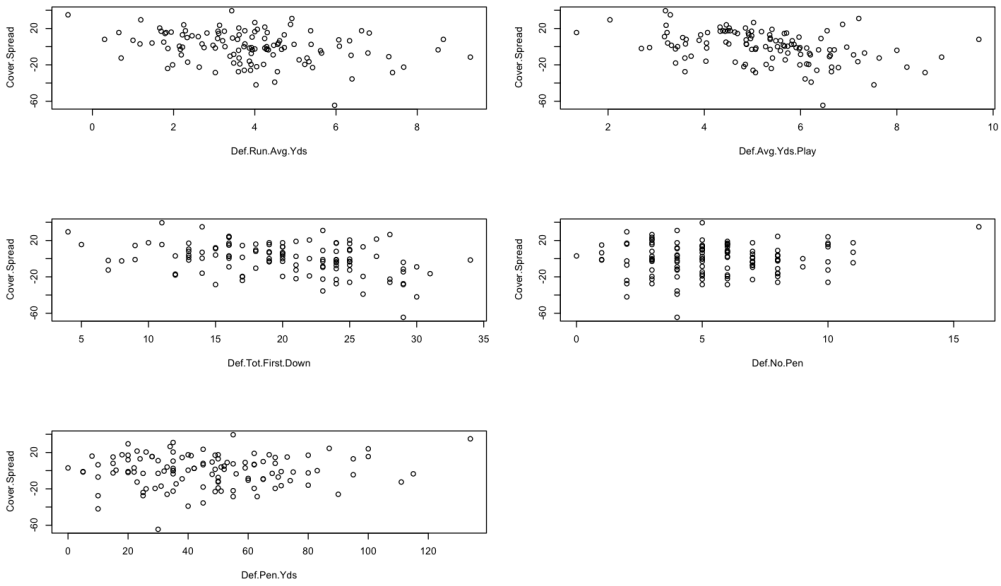

FIGURE 3. Scatter Plot of 5 Continuous Covariates vs. Cover.Spread

Observe that there is a positive relationship between Cover.Spread and the covariates Off.Pass.Pct, Off.Run.Avg.Yds, Off.Avg.Yds.Play, and Off.Tot.First.Down, and a negative relationship between Cover.Spread and Off.Pen.Yds, as well as Def.Avg.Yds.Play.

Moving to the discrete covariates, Table 5 reports the distribution for the six covariates.

Variable	0	1	2	3	4	5	6	7	8
Off.Pass.TD	7.8%	29.7%	24.6%	14.4%	9.3%	4.2%	0%	0%	0%
Off.Run.TD	12.7%	26.3%	22.9%	17.0%	10.2%	4.2%	2.5%	3.4%	0.9%
Off.Tot.TO	19.5%	38.1%	21.2%	15.3%	5.1%	0.9%	0%	0%	0%
Def.Pass.TD	27.2%	29.7%	22.0%	15.3%	5.1%	0.9%	0%	0%	0%
Def.Run.TD	25.4%	35.6%	17.8%	14.4%	3.4%	3.4%	0%	0%	0%
Def.Tot.TO	20.3%	27.2%	29.7%	19.5%	1.7%	0.9%	0.9%	0%	0%

TABLE 5. Distribution of Discrete Covariates

For an explanation of the covariates in Table 5, refer to the data dictionary in Table 12 in the Appendix. The ranges for Off.Pass.TD, Off.Tot.TO, Def.Pass.TD, and Def.Run.TD are from 0 to 5, the range for Off.Run.TD is from 0 to 8, and the range for Def.Tot.TO is from 0 to 6. Most games during

the 2018 ACC season have one or two offensive passing touchdowns as well as offensive running touchdowns. There were a total of zero, one, or two offensive turnovers in the majority of games. Furthermore, most games have zero, one, or two defensive passing touchdowns, defensive running touchdowns, and defensive total turnovers.

5. Multiple Regression Analysis

In Section 5, I will examine possible transformations for the continuous and discrete covariates before performing the model selection.

5.1. **Transformations for Continuous Covariates.**

Since several of my covariates and response variable are skewed, I decided to use transformations. The main reason for a transformation is that the model assumptions are violated. A power transformation is used to suggest appropriate transformations for both covariates and response variables. Since the power transformation cannot deal with negative or zero values, seven covariates and the response variable were shifted so that all values are strictly greater than zero. Increasing a variable by a constant does not change the shape of the variable. In order to determine the amount by which to shift each of the eight variables, I examined their respective typical values.

For New.Off.Pass.Pct, I chose to added 0.2 since there is one observation where Off.Pass.Pct = 0, and the typical Off.Pass.Pct is 59.67% (Table 4). It should be noted that the values for Off.Pass.Pct are already multiplied by 100%, so I added 0.2 means I have added 0.2%. The remaining six covariates and response variable were shifted using the same logic.

$$\begin{aligned}
\text{New.Off.Pass.Pct} &= \text{Off.Pass.Pct} + 0.2 \\
\text{New.Off.Pass.Yds} &= \text{Off.Pass.Yds} + 1 \\
\text{New.Off.No.Pen} &= \text{Off.No.Pen} + 0.1 \\
\text{New.Off.Pen.Yds} &= \text{Off.Pen.Yds} + 0.1 \\
\text{New.Def.Run.Avg.Yds} &= \text{Def.Run.Avg.Yds} + 0.65 \\
\text{New.Def.No.Pen} &= \text{Def.No.Pen} + 0.1 \\
\text{New.Def.Pen.Yds} &= \text{Def.Pen.Yds} + 0.1 \\
\text{New.Cover.Spread} &= \text{Cover.Spread} + 64.6
\end{aligned}$$

The remaining 7 continuous covariates were left unchanged. In order to determine the appropriate transformation, the Box-Cox procedure was used to automatically identify a transformation from the family of power transformations on a random vector X. The family of power transformations

is of the form $X' = X^{\lambda}$ where λ is a parameter determined by the Box-Cox power transformation, X' is the transformed variable, and $\lambda = 0$ indicates $X' = \ln(X)$ by definition ([3], p.134-5). If $\lambda = 0.5$, for example, then $X' = \sqrt{X}$.

Box-Cox uses the method of maximum likelihood to estimate λ. The statistical software R performs these estimates, and the estimated powers for the continuous variables are reported in Table 6.

Variable	Estimated Power	Transformation
New.Off.Pass.Pct	1.4031016	None
New.Off.Pass.Yds	0.977626	None
Off.Run.Avg.Yds	0.6936627	Square-Root
Off.Avg.Yds.Play	0.4772026	None
Off.Tot.First.Down	0.592753	Square-Root
New.Off.No.Pen	0.51845	None
New.Off.Pen.Yds	0.5571395	Square-Root
Def.Pass.Pct	1.2057641	None
Def.Pass.Yds	0.6598441	None
New.Def.Run.Avg.Yds	0.9882215	None
Def.Avg.Yds.Play	0.7801285	None
Def.Tot.First.Down	1.0698116	None
New.Def.No.Pen	0.5595359	Square-Root
New.Def.Pen.Yds	0.5457683	None
New.Cover.Spread	1.3589012	None

TABLE 6. Estimated Powers for Continuous Covariates and Response Variable, and Used Transformations

The variables with estimated powers close to one were unaltered. This includes: New.Off.Pass.Pct, New.Off.Pass.Yds, Def.Pass.Pct, New.Def.Run.Avg.Yds, Def.Avg.Yds.Play, Def.Tot.First.Down, and New.Cover.Spread.

For the variable Off.Run.Avg.Yds, R reported an estimated power of 0.69. To determine the appropriate transformation for this variable. I considered the square-root transformation and the log transformation. I examined the scatter plots of Off.Run.Avg.Yds vs. Cover.Spread, sqrt(Off.Run.Avg.Yds) vs. Cover.Spread and log(Off.Run.Avg.Yds) vs Cover.Spread with all three regression lines on the plot, reported in Figure 10 of the Appendix. This plot shows that all three regression lines fit the data very similarly. I could not definitively say that one of the three transformations was the best, so I then compared the residual plots for simple regression of x versus y, reported in Figure 11 of the Appendix. From Figure 11, I noted that all three residual

plots are centered at zero and randomly spread. I did not feel comfortable drawing any conclusions at this stage.

As a final step, I examined the three marginal model plots for the three simple regressions. A marginal model plot is a scatterplot with the response variable on the vertical axis and the predictor on the horizontal axis. The plot contains both a nonparametric fit function for the variables (shown by a blue in R denoted as "Data"), and a function that shows the predicted values as a function of the variable on the horizontal axis (shown by a red line in R denoted as "Model") ([35]). When the curves are close to each other, there is evidence that the model fits well. The three marginal model plots are reported in Figures 12-14 in the Appendix. Figure 12 is the marginal model plot for the untransformed variable. There is deviation between the Data and Model curves from Off.Run.Avg.Yds = 6 to Off.Run.Avg.Yds = 10. Figure 14 is the marginal model plot for the log transformation. The data and model curves deviate significant from log(Off.Run.Avg.Yds) = -1.5 to log(Off.Run.Avg.Yds) = 0.75. Figure 13 is the marginal model plot for the square-root transformation. There are no major deviations between the Data and Model curves. Therefore, I chose to use the square-root transformation of the variable Off.Run.Avg.Yds.

I used the same steps and logic for the remaining seven variables. The results are reported in the third column of Table 6. Note that the only four transformed variables are Off.Run.Avg.Yds, Off.Tot.First.Down, New.Off.Pen.Yds, and New.Def.No.Pen. Both Off.Tot.First.Down and New.Def.No.Pen are count variables, and the square-root transformation is usually used on count data ([4]).

5.2. **Transformations for Discrete Covariates.** To determine the transformations for the discrete covariates, I examined the model with the four transformed continuous covariates, the remaining ten untransformed continuous covariates, and the six discrete covariates. The discrete variables were then added to the model using either the untransformed version or the square-root transformation because these discrete variables are counts, and the square-root transformation is a commonly used transformation used for count variables. I examined marginal model plots for the models.

For example, to determine if a transformation was needed for Off.Pass.TD, I first examined the marginal model plot for the non-transformed variable, found in Figure 15 of the Appendix. There are significant deviations between the Data and Model curves in Figure 15, so I decided to try the square root transformation of Off.Pass.TD. The corresponding marginal model plot is found in Figure 16. Although there are small deviations between the red and

blue curves, Figure 16 is a significant improvement from Figure 15. Therefore, I use the square-root transformation for Off.Pass.TD. I repeated this process for the other five discrete covariates. The marginal model plots for the remaining discrete covariates are located in Figures 17-25 of the Appendix. In the end, the square-root transformation was deemed appropriate for Off.Pass.TD, Off.Tot.TO, and Def.Tot.TO. The discrete covariates and their transformations are reported in Table 7. In total, among the 20 covariates, 7 predictors used the square-root transformation while the other 13 covariates were not transformed.

Variable	Transformation
Off.Pass.TD	Square Root
Off.Run.TD	None
Off.Tot.TO	Square Root
Def.Pass.TD	None
Def.Run.TD	None
Def.Tot.TO	Square Root

TABLE 7. Transformations for Discrete Covariates

5.3. **Model Selection.** Next, I turned my attention to determining which of the 20 covariates I use in the final model, using the backwards elimination method. In order to select the covariates, I started by running the model with the 20 covariates and identified the covariate with the highest p-value, which was Def.Pass.TD with a p-value of 0.825. The hypothesis testing is as follows.

$$H_0 : \beta_{\text{Def.Pass.TD}} = 0$$
$$H_a : \beta_{\text{Def.Pass.TD}} \neq 0$$

Using the typical significance value of 0.10, since $0.825 > 0.10$, I do not reject the null hypothesis. This means that when all the other variables are in the model, the coefficient for Def.Pass.TD=0, and thus Def.Pass.TD is removed from the model. Then, I rerun the model with the other 19 covariates, and repeat this process, until all of the covariates have a p-value less than 0.10. After this process was complete, I was left with the following eight covariates:

(1) sqrt(Off.Pass.TD)
(2) Off.Run.TD
(3) sqrt(Off.Tot.First.Down)
(4) sqrt(Off.Tot.TO)
(5) Def.Run.TD
(6) Def.Avg.Yds.Play

(7) Def.Pen.Yds
(8) sqrt(Def.Tot.TO)

The corresponding model is:

(Model M1)

$$\begin{aligned}\text{Cover.Spread} \approx 16.78 + 8.11 * \sqrt{\text{Off.Pass.TD}} + 4.65 * \text{Off.Run.TD} \\ -5.13 * \sqrt{\text{Off.Tot.First.Down}} - 7.23 * \sqrt{\text{Off.Tot.TO}} - 1.53 * \text{Def.Run.TD} \\ -3.14 * \text{Def.Avg.Yds.Play} + 0.10 * \text{Def.Pen.Yds} + 7.32 * \sqrt{\text{Def.Tot.TO}}\end{aligned}$$

6. Model Validation

In Section 6, I will validate Model M1 and then will compare Model M1 to the model *without* transformations, Model M2.

6.1. **Assumptions of Linear Regression.**

In order to determine if Model M1 is valid, I examine whether the assumptions of the linear regression model are satisfied. To do so, I analyze the plots in Figure 4.

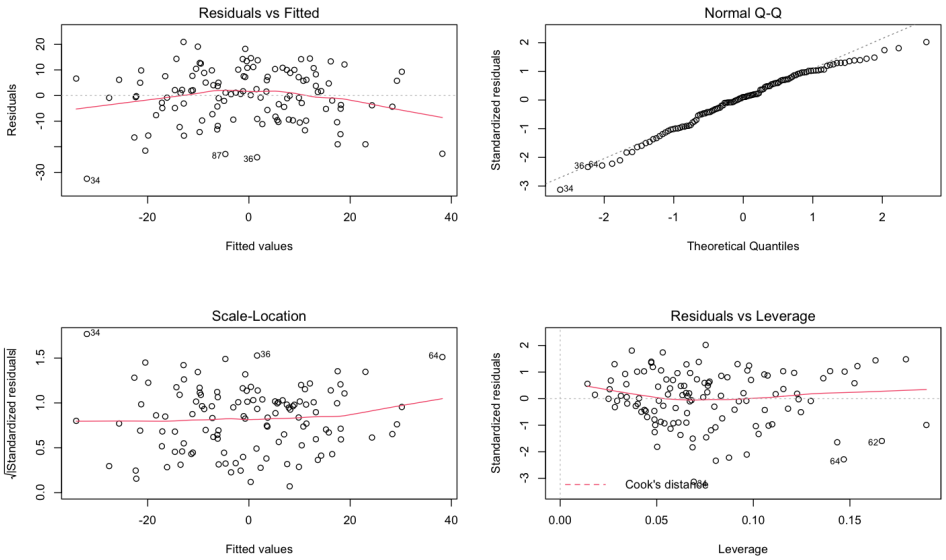

Figure 4. Residuals vs Fitted, Normal Q-Q, Scale-Location, and Residuals vs Leverage Plots of Model M1

The four plots allow me to check whether the assumptions of the linear regression model are satisfied for the ACC 2018 data. The Residuals vs.

Fitted plot provides information on whether linearity (Assumption 1) is violated. Linearity holds if the mean of the residual is zero. Although the red line and horizontal dashed line at $y = 0$ do not match perfectly, they are a rather close fit, so the model upholds the linearity assumption.

The Normal Q-Q plot shows whether the residuals are normally distributed (Assumption 4). A straight line of residuals along the dashed line is a strong indication that normality is upheld. Observe in Figure 4 that the tails on both ends deviate away from the line, so normality appears to be violated in Model M1. To confirm this observation, I use the Shapiro-Wilk test, which examines the coefficient of correlation between the ordered residuals and their expected values under normality. The hypothesis testing is as follows ([3], p.115):

$$H_0 : \text{The residuals are normally distributed}$$
$$H_a : \text{The residuals are not normally distributed}$$

Using R, I performed the Shapiro-Wilke test and obtained a p-value of 0.04999921. Thus, there is statistical evidence to reject the null hypothesis and accept the alternative hypothesis, which confirms that the residuals are not normal. Note that the sample size, $n = 118$, is not small, so the violation of normality is not that crucial.

The Scale-Location plot takes the square root of the standardized residuals and plots against the fitted values and shows if the residuals are spread equally along the ranges of predictors. A horizontal line with equally, randomly spread points indicates that constant variance (Assumption 2) is upheld ([5]). Observe that the red line is fairly horizontal, but does slope very slightly upwards in the right tail. Additionally, the spread of the residuals is randomly spread, so the assumption of constant variance is upheld.

The Residuals vs. Leverage plot is used to find influential cases. When cases are outside of the Cook's contour lines (meaning they have high Cook's distance scores), the cases are influential to the regression results ([5]). The regression results will be altered if we exclude those cases. Since we cannot even see the Cook's contour lines, at no point is Cook's distance greater than 0.5, so there are no influential cases.

In order to check the multicollinearity problem (Assumption 5) of Model M1, I use the variance-inflation factor (VIF). The VIF is used to detect the severity of multicollinearity in the least squares regression analysis. Specifically, the VIF is computed by

$$\text{VIF}_j = \frac{1}{1 - R_j^2}, \ j = 1, 2, \ldots, p - 1 \tag{6.1}$$

where R_j^2 is the coefficient of determination when X_j is regressed on the remaining $p-2$ other X variables in the model ([24], p.106). The VIF values for the eight covariates of Model (M1) are reported in Table 8.

Variable	VIF
sqrt(Off.Pass.TD)	1.38174
Off.Run.TD	1.880818
sqrt(Off.Tot.First.Down)	2.011721
sqrt(Off.Tot.TO)	1.145224
Def.Run.TD	1.491298
Def.Avg.Yds.Play	1.58078
Def.Pen.Yds	1.037897
sqrt(Def.Tot.TO)	1.102846

TABLE 8. VIF Values of Covariates in Model M1

I use the commonly accepted cutoff of 5 to determine whether the model satisfies the multicollinearity assumption ([3], p.409). Table 8 reports that all the VIF values are less than 2.1, indicating there is no multicollinearity problem. Another way in which to verify the multicollinearity assumption is to compare the VIF value for the entire model to those of the covariates. The R^2 of Model M1 is 64.05%, allowing me to calculate:

$$\text{VIF}_{\text{M1}} = \frac{1}{1-R^2} = \frac{1}{1-0.6405^2} = 1.70$$

Note that six out of the eight variables have a VIF value below $\text{VIF}_{\text{M1}} = 1.70$. Although two variables, Off.Run.TD and sqrt(Off.Tot.First.Down) have VIF values greater than 1.70, they are only slightly larger than the value for the whole Model M1, so the multicollinearity assumption is satisfied.

Furthermore, the independence assumption (Assumption 3) is violated because my dataset contains observations Team A vs. Team B and Team B vs. Team C. For example, I compare Clemson to Florida State and Florida State to Wake Forest. As a result, independence is not satisfied. Although a dependence relation does exist, I am confident that is not strong. In the Residuals vs. Fitted plot in Figure 4, there is no pattern amongst the residuals. Furthermore, I plotted the individual x's against the residuals and found no pattern. Thus, I am assured that the dependence relationship is not strong.

Next, I examine the marginal model plots for the Model M1 shown in Figure 5. Observe that for each of the eight covariates, there are no significant deviations between the Data and Model curves in the respective plots, a strong indication that the model is a good fit. Furthermore, the final plot, located in the third row, third column of Figure 5, shows the plot of with the fitted values on the horizontal axis. Although the red and blue lines are not a perfect match, the two curves match closely, which further emphasizes the adequacy of Model M1. Thus, Model M1 is verified.

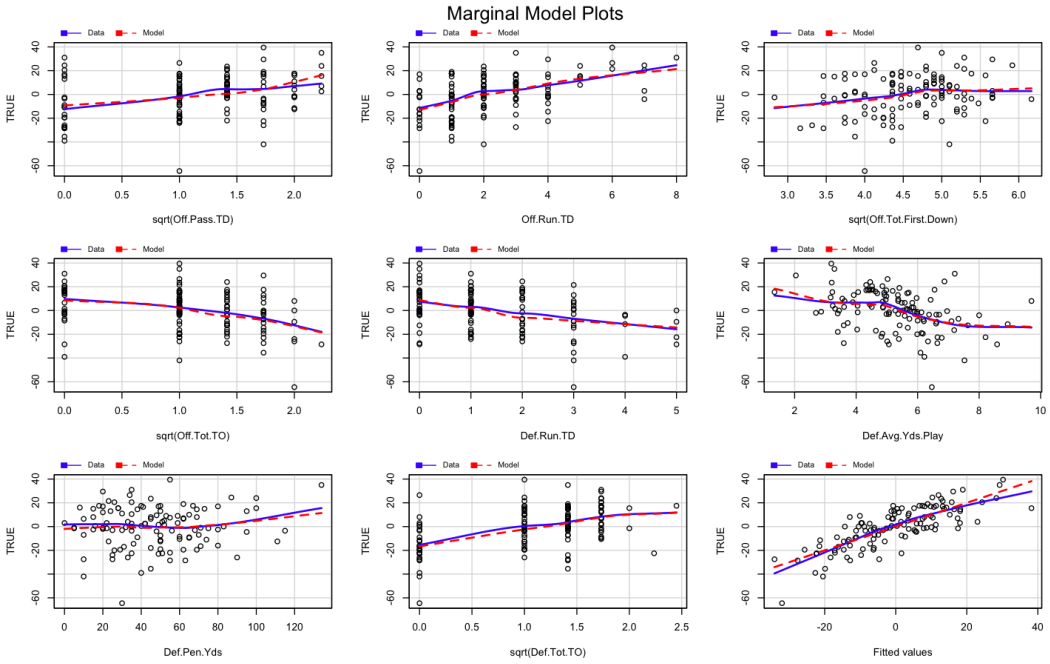

FIGURE 5. Marginal Model Plots of Model M1

6.2. **Comparison to Simple Model.** In addition to Model M1, I also examined the model with the 20 *untransformed* covariates and performed the model selection. The reason for doing so is because it is easier to interpret the untransformed covariates as opposed to the transformed covariates. Upon performing the model selection, I obtained the following model:

(Model M2)

$$\begin{aligned}\text{Cover.Spread} \approx 4.34 + 2.09 * \text{Off.Pass.TD} + 3.48 * \text{Off.Run.TD} \\ -4.10 * \text{Off.Tot.TO} - 1.85 * \text{Def.Run.TD} - 3.24 * \text{Def.Avg.Yds.Play} \\ +0.08 * \text{Def.Pen.Yds} + 3.51 * \text{Def.Tot.TO}\end{aligned}$$

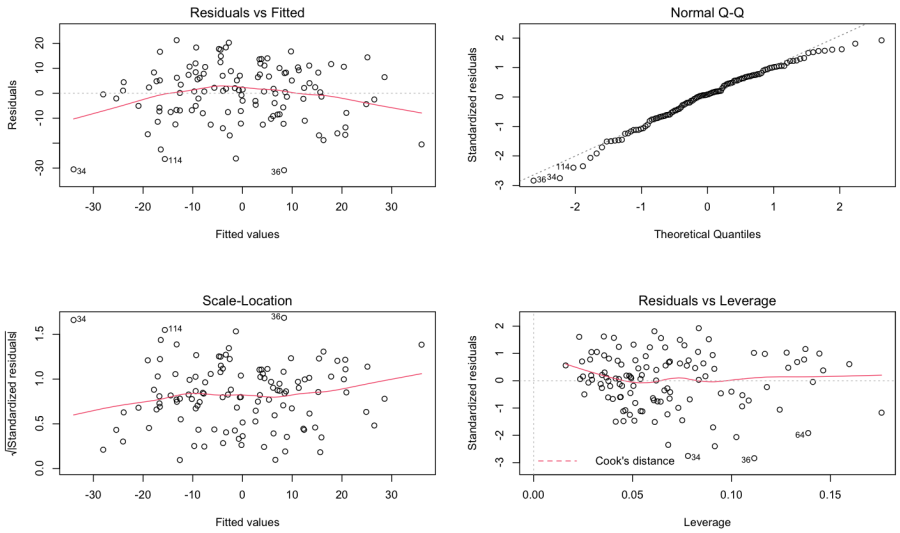

FIGURE 6. Residuals vs Fitted, Normal Q-Q, Scale-Location, and Residuals vs Leverage Plots of Model M2

Model M2 has $R^2 = 58.39\%$. To examine whether Model M2 upholds the assumptions, I examine Figure 6.

Analyzing the Residuals vs. Fitted Plot, although there is some deviation between the dashed line at $y = 0$ and red curve, the two are relatively close, so linearity (Assumption 1) holds. Based on the two plots, the two models are approximately the same in upholding the linearity assumption. The Normal Q-Q plot indicates that normality (Assumption 4) is violated, so both Model M1 and Model M2 violate the normality assumption. The positive slope of the red line in the Scale-Location plot indicates that the assumption of constant variance (Assumption 2) is violated. The corresponding plot for Model M1 is better in terms of the constant variance assumption. Finally, in the Residuals vs. Leverage plot, since we cannot even see the Cook's contour lines, at no point is Cook's distance greater than 0.5, so there are no influential cases.

To check the multicollinearity assumption (Assumption 5), the VIF values for the covariates in Model M2 are reported in Table 9.

Variable	VIF
Off.Pass.TD	1.125801
Off.Run.Avg.Yds	1.08369
Off.Tot.TO	1.15412
Def.Run.TD	1.477894
Def.Avg.Yds.Play	1.606903
Def.Pen.Yds	1.050349
Def.Tot.TO	1.07954

TABLE 9. VIF Values of Covariates in Model M2

Since all values are less than 5, Model M2 upholds the assumption of multicollinearity. The summary of the comparison of the two models is reported in Table 10.

Assumption	Satisfied by M1?	Satisfied by M2?
Linearity	Yes	Yes
Normality	No	No
Constant Variance	Yes	No
Independence	No	No
Multicollinearity	Yes	Yes

TABLE 10. Assumptions of the Linear Model: Model M1 vs M2

Thus, Model M1 is better than Model M2 in terms of upholding the assumptions of the linear model. Next, I examine the marginal model plots of Model M2 in Figure 7 to see whether the covariates are modeled correctly.

Note the discrepancies between the the Data and Model curves in the plots for Off.Pass.TD, Off.Tot.TO, and Def.Tot.TO. The square-root transformation was applied to these covariates in Model M1, which significantly improved their marginal model plots, as seen in Figure 5. Therefore, Model M1 is better than Model M2 based on their respective marginal model plots.

To assess the models further, I used data from the 2019 ACC regular season ([6]-[19]). Recall that data from the 2018 ACC regular season was used to build the models M1 and M2. I used Models M1 and M2 to predict Cover.Spread values with 2019 ACC regular season data and then found the mean squared error (MSE) to measure the average of the squared errors. The MSE values for the two models, as well as the R^2 values, are reported in Table 11.

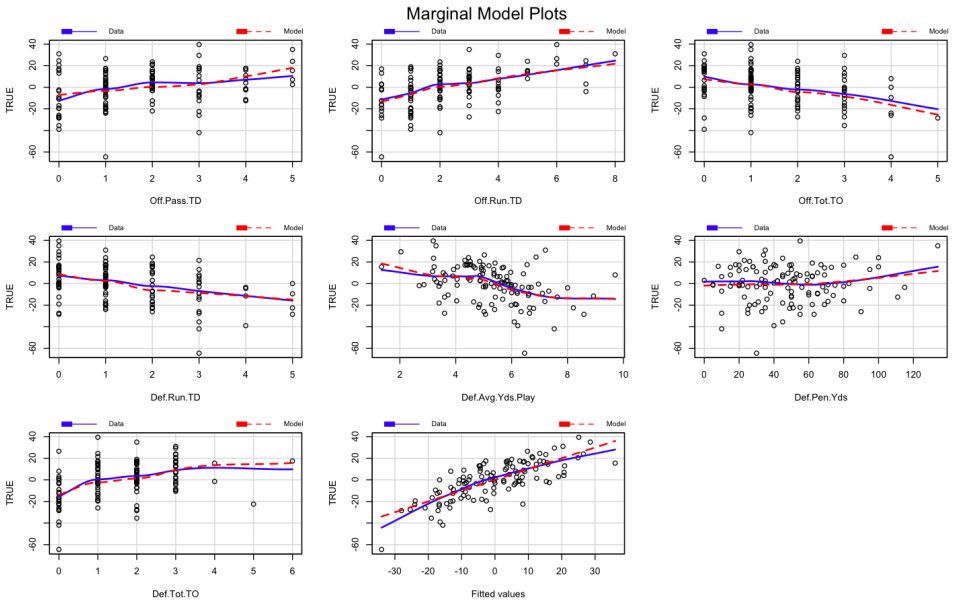

FIGURE 7. Marginal Model Plots of Model M2

Model	MSE	R^2
Model M1	6.955	0.6405
Model M2	29.84	0.5839

TABLE 11. Mean Square Errors and R^2 of Model M1 and Model M2

That the MSE for Model M1 is much smaller than that of Model M2 indicates Model M1 has better predictability. Additionally, Model M1 has a larger R^2 than model M2 by about 6%. Therefore, even though Model M2 is easier to interpret, I conclude to use Model M1, restated below. The coefficients will be interpreted in detail in Section 7.

(Model M1)

$$\text{Cover.Spread} \approx 16.78 + 8.11 * \sqrt{\text{Off.Pass.TD}} + 4.65 * \text{Off.Run.TD}$$
$$-5.13 * \sqrt{\text{Off.Tot.First.Down}} - 7.23 * \sqrt{\text{Off.Tot.TO}} - 1.53 * \text{Def.Run.TD}$$
$$-3.14 * \text{Def.Avg.Yds.Play} + 0.10 * \text{Def.Pen.Yds} + 7.32 * \sqrt{\text{Def.Tot.TO}}$$

7. Model Interpretation

Examining the coefficients in Model M1, note that a larger value of Cover.Spread corresponds to the team performing better than the oddsmakers in Las Vegas

anticipated. For the coefficient interpretation, I examine each coefficient one by one. Note that I do not examine the y-intercept because all the covariates would not equal zero in an actual college football game, so the y-intercept is hence not in the data range.

- $\sqrt{\text{Off.Pass.TD}}$: The coefficient is 8.11. The positive sign makes sense because the more running and passing touchdowns the team's offense earns, then the higher Cover.Spread will be. From Table 4, the typical value of Off.Pass.TD is 1. When Off.Pass.TD increases from 1 to 4, which corresponds to $\sqrt{\text{Off.Pass.TD}}$ increasing from 1 to 2, then Cover.Spread increases by 8.11 points, holding all other variables constant.

- Off.Run.TD: The coefficient is 4.65. The positive coefficient makes sense because the more running touchdowns the team's offense earns, then the higher Cover.Spread will be. The magnitude means that when Off.Run.TD increases by one touchdown, and the other variables remain fixed, then Cover.Spread increases by 4.65 points.

- $\sqrt{\text{Off.Tot.First.Down}}$: The coefficient is -5.13. The negative sign does *not* make sense. Intuitively, the more offensive first downs a team has, the more likely the offense will score more points, and hence lead to a higher Team.Score and a higher Cover.Spread.

- $\sqrt{\text{Off.Tot.TO}}$: The coefficient is -7.23. The negative coefficient is logical because a larger number of offensive total turnovers indicates a poorer performance by the team, leading to the potential for a lower Team.Score, a higher Opp.Score, and hence a lower Cover.Spread. From Table 4, the typical value of Off.Tot.TO is 1. When Off.Tot.TO increases from 1 to 4, which corresponds to $\sqrt{\text{Off.Tot.TO}}$ increasing from 1 to 2, then Cover.Spread decreases by 7.23 points, when all other variables remain fixed.

- Def.Run.TD: The coefficient is -1.53. The negative sign is sensible because the larger Def.Run.TD is, then the more running touchdowns the team's opponent scores, which creates the potential for a lower Cover.Spread. The magnitude indicates that when Def.Run.TD increases by one touchdown, and all other variables remain fixed, then Cover.Spread decreases by 1.53 points.

- Def.Avg.Yds.Play: The coefficient is -3.14. The negative sign makes sense because a higher value of Def.Avg.Yds.Play indicates the team's

defense is allowing a high average yards per play to the opponent's offense, leading to the potential for a lower Team.Score, a higher Opp.Score, and thus a lower Cover.Spread. The magnitude means that when Def.Avg.Yds.Play increases by one yard, Cover.Spread falls by 3.14 points, ceteris paribus.

- Def.Pen.Yds: The coefficient is 0.10. The positive sign is reasonable because the more yards awarded to the team due to the opponent's penalties, the greater the potential for the team performing better than the oddsmakers predicted. The magnitude means that when Def.Pen.Yds increases by one yard, then Cover.Spread increases by 0.10 points, ceteris paribus.

- $\sqrt{\text{Def.Tot.TO}}$: The coefficient is 7.32. The positive sign makes sense because a larger number of total turnovers by the opponents is a good indication of that the team is performing well.

In summary, the signs of the coefficients were as expected, with the exception of the negative sign for $\sqrt{\text{Off.Tot.First.Down}}$. I removed this variable from Model M1 in order to see the effect. The R^2 of the model fell from 64.05% to 62.29%, and the adjusted R^2 fell from to 0.6141 to 0.5988, so I decided to keep $\sqrt{\text{Off.Tot.First.Down}}$ in Model M1 despite the nonlogical negative sign. This irregularity could be explained by a recent trend in college football in which the top teams have more long scrimmage plays (plays of 10+ yards) rather than a large number of short-yardage plays ([2]). This trend has been adopted by many college football teams, and as a result, it is possible that the relationship between offensive total first downs and Team.Score can no longer be described as positive.

The magnitudes of the coefficients are logical. Offensive passing touchdowns are the most common form of scoring in college football, so it makes sense its magnitude is largest. The smallest magnitude, 0.10, for $\sqrt{\text{Def.Tot.TO}}$ is logical because typically, defensive penalty yards do not have a significant impact on the outcome of a college football game.

8. Discussion

Moving into the final discussion, Model M1 has been finalized and validated as a model that predicts whether a team will cover the point spread in a given football game. It contains 5 discrete variables and 3 continuous variables. There are many factors that determine the outcome of a college football game, but Model M1 indicates that roughly 64% of the variation of Cover.Spread can be explained by the variables in the model. Although

Model M1 does not satisfy the independence and normality assumptions, the other assumptions are upheld.

There are, undoubtedly, shortcomings of the model. Perhaps the most significant is that bettors must rely on previous season's data to make predictions about the current season. This is a disadvantage because a college football team can look significantly different from year to year, with seniors graduating, a handful of juniors leaving early and declaring for the NFL Draft, and a new class of freshman entering.

I analyzed the 2018 ACC postseason to assess Model M1 further. The ACC played a total of 12 postseason games in 2018 postseason, which includes the conference championship, bowl games, playoff games, and the national championship. I used Model M1 to predict the Cover.Spread values, obtaining a MSE of 589.4388. This MSE value is significantly larger than the MSE of Model M1 of 6.955 (Table 8). There are two possible contributing factors to the larger MSE value. The first is that the high value could be attributed to the relatively small sample size of only 12 games. The second, and more plausible explanation, is that, besides the conference championship game, the remaining 11 postseason games were played against opponents *not* in the ACC. Therefore, it may be unreasonable to expect the model performs well in games between ACC teams and non-ACC teams.

Additionally, the Model M1 may not work that well on other Power 5 Conference teams. In particular, I used Model M1 to predict Cover.Spread values with data from two other major conferences in the Power 5. Using Model M1 with 2018 regular season data from the Big Ten conference, I obtained an MSE of 183.94. Repeating this process with 2018 regular season data from the Big 12 conference yielded an MSE of 1099.787. A possible explanation could be that although each team plays a few out of conference games each season, the majority (typically 66-75%) of the competition a team faces is within conference. Therefore, Model M1 is inadequate when applied to teams outside of the ACC. With the benefit of more time, I would use data from the regular season games from the other four major conferences in order to have a larger sample size with which to build my model.

Although not a complete success, the influence of several predictors on covering the point spread was determined. This kind of analysis can help bettors have a better idea of the factors that determine whether a team will cover the point spread in a given game.

9. Acknowledgements

I would like to thank Weiwen Miao for being an incredible mentor and guiding me through this process. I would also like to thank the Haverford College Mathematics Department for providing me the opportunity and support, not just during the thesis, but over the course of my four years here.

References

[1] *2018 College Football Statistics*. 2019. URL: www.sports-reference.com/cfb.

[2] *2018 National Leaders 2018 Long Scrimmage Plays -All Games*. 2019. URL: http://www.cfbstats.com/2018/leader/national/team/offense/split01/category30/sort01.html.

[3] Michael Kutner et al. *Applied Linear Regression Models*. Fourth. New York: McGraw-Hill Irwin, 2004.

[4] M.J. Crawley et al. *An Introduction to Data Analysis using S-Plus*. London, UK: John Wiley and Sons, 2003.

[5] Kim Bommae. *Understanding Diagnostic Plots for Linear Regression Analysis*. University of Virginia Library. 2015.

[6] Jimmy Boyd. *Week 1 NCAAF Football Betting Lines: Point Spreads and Game Totals*. 2019. URL: https://www.boydsbets.com/week-1-college-football-odds/.

[7] Jimmy Boyd. *Week 10 NCAAF Football Betting Lines: Point Spreads and Game Totals*. 2019. URL: https://www.boydsbets.com/week-10-college-football-odds/.

[8] Jimmy Boyd. *Week 11 NCAAF Football Betting Lines: Point Spreads and Game Totals*. 2019. URL: https://www.boydsbets.com/week-11-college-football-odds/.

[9] Jimmy Boyd. *Week 12 NCAAF Football Betting Lines: Point Spreads and Game Totals*. 2019. URL: https://www.boydsbets.com/week-12-college-football-odds/.

[10] Jimmy Boyd. *Week 13 NCAAF Football Betting Lines: Point Spreads and Game Totals*. 2019. URL: https://www.boydsbets.com/week-13-college-football-odds/.

[11] Jimmy Boyd. *Week 14 NCAAF Football Betting Lines: Point Spreads and Game Totals*. 2019. URL: https://www.boydsbets.com/week-14-college-football-odds/.

[12] Jimmy Boyd. *Week 2NCAAF Football Betting Lines: Point Spreads and Game Totals*. 2019. URL: https://www.boydsbets.com/week-2-college-football-odds/.

[13] Jimmy Boyd. *Week 3 NCAAF Football Betting Lines: Point Spreads and Game Totals*. 2019. URL: https://www.boydsbets.com/week-3-college-football-odds/.

[14] Jimmy Boyd. *Week 4 NCAAF Football Betting Lines: Point Spreads and Game Totals*. 2019. URL: https://www.boydsbets.com/week-4-college-football-odds/.

[15] Jimmy Boyd. *Week 5 NCAAF Football Betting Lines: Point Spreads and Game Totals*. 2019. URL: https://www.boydsbets.com/week-5-college-football-odds/.

[16] Jimmy Boyd. *Week 6 NCAAF Football Betting Lines: Point Spreads and Game Totals*. 2019. URL: https://www.boydsbets.com/week-6-college-football-odds/.

[17] Jimmy Boyd. *Week 7 NCAAF Football Betting Lines: Point Spreads and Game Totals*. 2019. URL: https://www.boydsbets.com/week-7-college-football-odds/.

[18] Jimmy Boyd. *Week 8 NCAAF Football Betting Lines: Point Spreads and Game Totals*. 2019. URL: https://www.boydsbets.com/week-8-college-football-odds/.

[19] Jimmy Boyd. *Week 9 NCAAF Football Betting Lines: Point Spreads and Game Totals*. 2019. URL: https://www.boydsbets.com/week-9-college-football-odds/.

[20] Bill Connelly. *Spread for every college football game in 2018's Week 1*. 2018. URL: https://www.sbnation.com/college-football/2018/8/30/17800050/college-football-picks-week-1-2018-predictions-odds-spreads.

[21] Bill Connelly. *Spread for every college football game in 2018's Week 2*. 2018. URL: https://www.sbnation.com/college-football/2018/9/6/17826820/college-football-picks-week-2-2018-predictions-odds-spreads.

[22] Bill Connelly. *Spread picks for every college football game in 2018's Week 3*. 2018. URL: https://www.sbnation.com/college-football/2018/9/13/17854190/college-football-picks-week-3-2018-predictions-odds-spreads.

[23] Bill Connelly. *Spread picks for every college football game in 2018's Week 4*. 2018. URL: https://www.sbnation.com/college-football/2018/9/20/17878870/college-football-picks-week-4-2018-predictions-odds-spreads.

[24] John Fox. *Applied Regression Analysis and Generalized Linear Model*. Ed. by 2nd. Thousand Oaks, California: Sage, 1997.

[25] Corporate Finance Institute. *Variance Inflation Factor (VIF)*. URL: https://corporatefinanceinstitute.com/resources/knowledge/other/variance-inflation-factor-vif/.

[26] Kaelen Jones. *Opening Lines for Every Week 10 College Football Game*. 2018. URL: https://www.si.com/college/2018/10/29/week-10-college-football-game-spreads-betting-lines-odds-openers.

[27] Kaelen Jones. *Opening Lines for Every Week 12 College Football Game*. 2018. URL: https://www.si.com/college/2018/11/11/week-12-college-football-game-spreads-betting-lines-odds.

[28] Kaelen Jones. *Opening Lines for Every Week 13 College Football Game*. 2018. URL: https://www.si.com/college/2018/11/19/week-13-college-football-game-spreads-betting-lines-odds.

[29] Kaelen Jones. *Opening Lines for Every Week 8 College Football Game*. 2018. URL: https://www.si.com/college/2018/10/14/week-8-college-football-game-spreads-betting-lines-odds.

[30] Kaelen Jones. *Opening Lines for Every Week 9 College Football Game*. 2018. URL: https://www.si.com/college/2018/10/22/week-9-college-football-game-spreads-betting-lines-odds.

[31] Kaelen Jones. *Opening Spreads for Every Week 5 College Football Game*. 2018. URL: https://www.si.com/college/2018/09/24/week-5-college-football-game-spreads-betting-lines-odds.

[32] Kaelen Jones. *Opening Spreads for Every Week 6 College Football Game*. 2018. URL: https://www.si.com/college/2018/10/01/week-6-college-football-game-spreads-betting-lines-odds.

[33] Kaelen Jones. *Opening Spreads for Every Week 7 College Football Game*. 2018. URL: https://www.si.com/college/2018/10/08/week-7-college-football-game-spreads-betting-lines-odds.

[34] Kalen Jones. *Opening Lines for Every Week 11 College Football Game*. 2018. URL: https://www.si.com/college/2018/11/04/week-11-college-football-game-spreads-betting-lines-odds.

[35] Warren Kuhfeld. "Marginal model plots". In: (2018). URL: https://blogs.sas.com/content/graphicallyspeaking/2018/08/23/marginal-model-plots/.

[36] Frank Schwab. "Sports Betting 101: What Does the Point Spread Mean, and Why Do People Bet It?" In: *MSN* (2020).

10. Appendix

Variable	Description
Off.Pass.Pct	(Number of times the offense completes a pass)/(Total number of offensive pass attempts) *100
Off.Pass.Yards	Total yards the team gains by passing the ball
Off.Pass.TD	Touchdowns the team scored from passing
Off.Run.Avg.Yds	Average yards the team gained per running attempt
Off.Run.TD	Touchdowns the team scored from running the ball
Off.Avg.Yds.Play	(Total yards gained by team's offense)/(Total number of team's offensive Plays)
Off.Tot.First.Down	Total number of first downs earned by the team (sum of the first down the team earns via passing, running, and penalities)
Off.No.Pen	Number of penalties commited by the team
Off.Pen.Yds	Total yards awarded to opponent due to the team's penalties in that game
Off.Tot.TO	Number of the team's interceptions + number of team's fumbles
Def.Pass.Pact	(Number of completions allowed to opponent)/(Total number of opponent's passing attempts)*100
Def.Pass.Yards	Total number of passing yards allowed to the opponent
Def.Pass.TD	Number of passing touchdowns allowed to the opponent
Def.Run.Avg.Yds	Average yards allowed to the opponent per running attempt
Def.Run.TD	Running touchdowns allowed to the opponent
Def.Avg.Yds.Play	(Total yards the team's defense allowed to the opposing offense)/(Total number of plays performed by the team's defense)
Def.Tot.First.Down	Total number of first downs allowed to the opponent in that game
Def.No.Pen	Number of penalties commited by the opponet
Def.Pen.Yds	Total yards awarded to team due to the opponent's penalties in that game
Def.Tot.TO	Number of opponent's fumbles + number of opponent's Interceptions

Table 12. Data Dictionary of the 20 Covariates

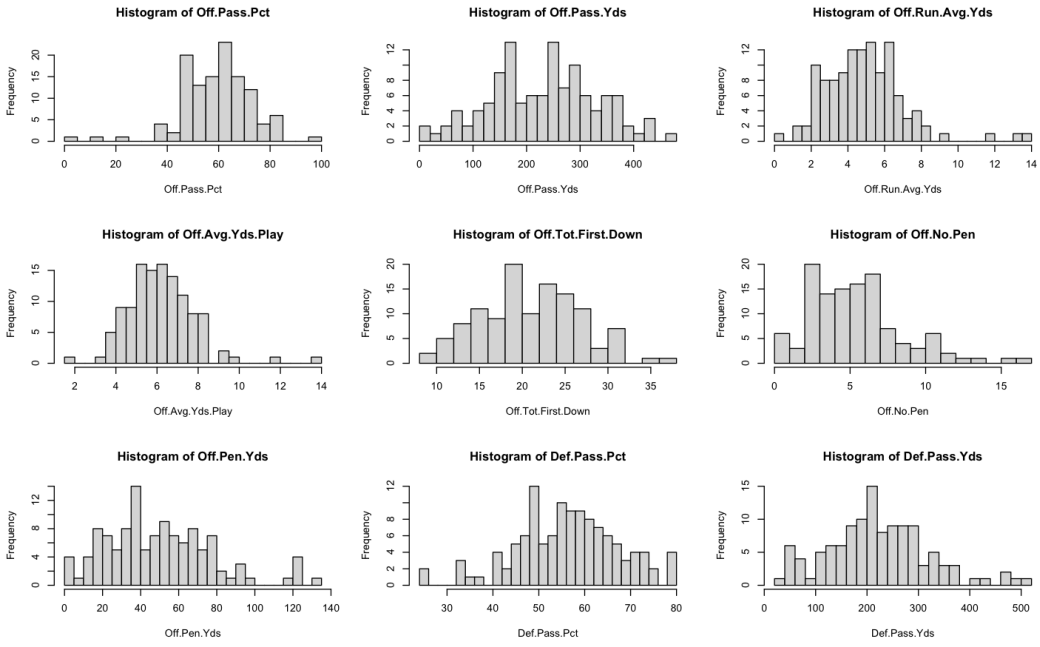

Figure 8. Histograms of 9 Continuous Covariates

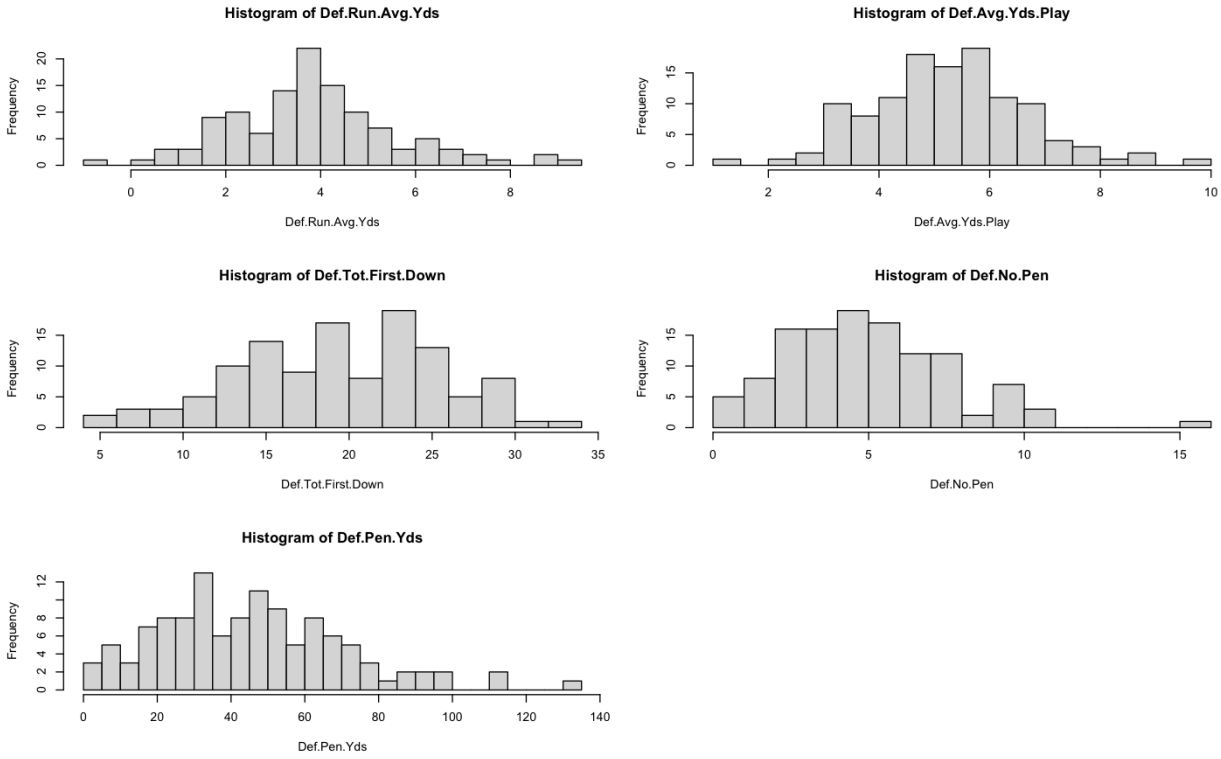

FIGURE 9. Histograms of the Remaining 5 Continuous Covariates

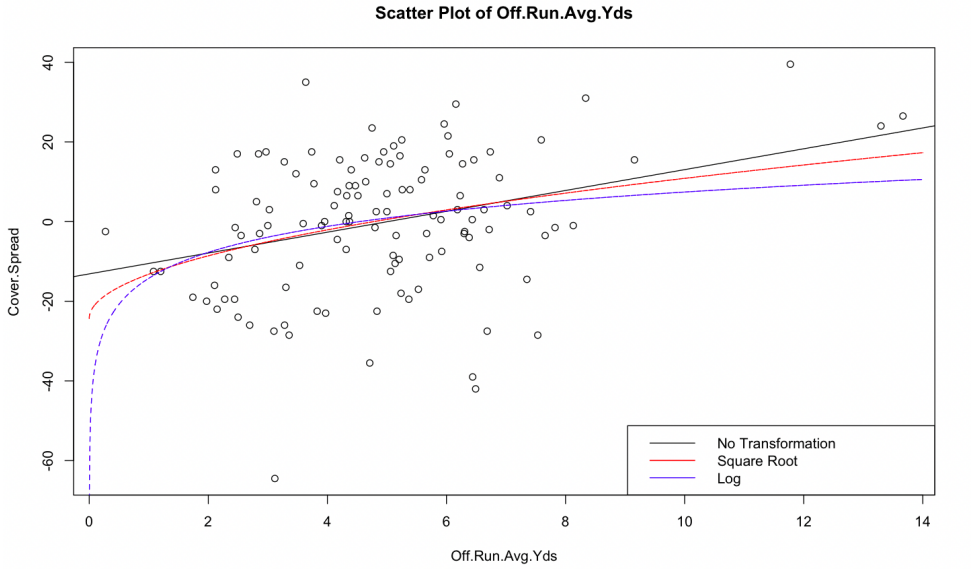

FIGURE 10. Scatter Plot of Off.Run.Avg.Yds with Linear, Square Root, and Log Regression Curves

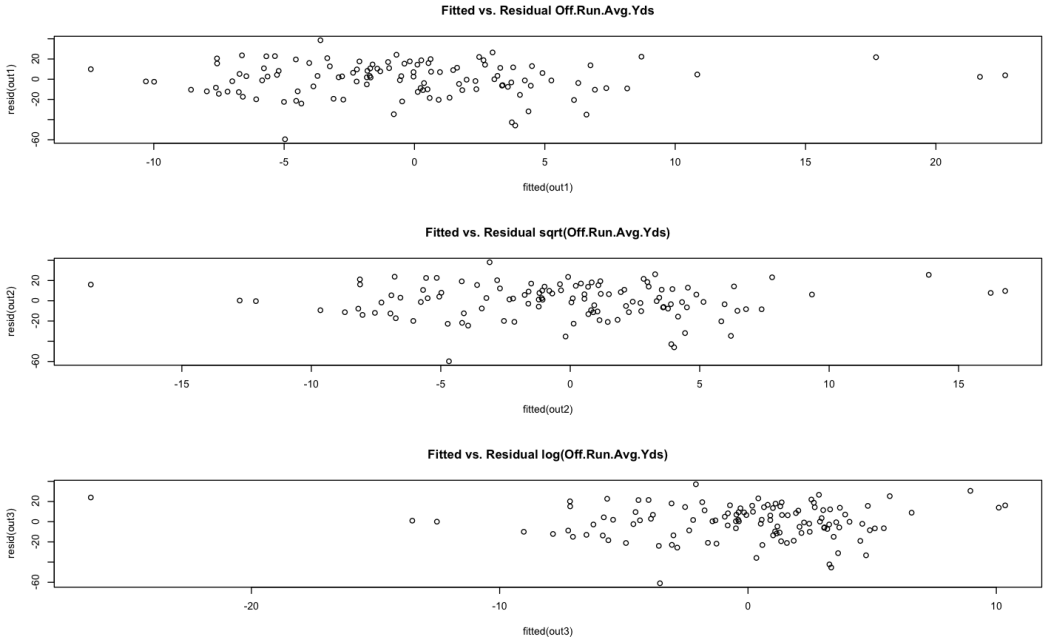

FIGURE 11. Fitted vs. Residual Plots for Off.Run.Avg.Yds vs. Cover.Spread, sqrt(Off.Run.Avg.Yds) vs. Cover.Spread, and log(Off.Run.Avg.Yds) vs. Cover.Spread

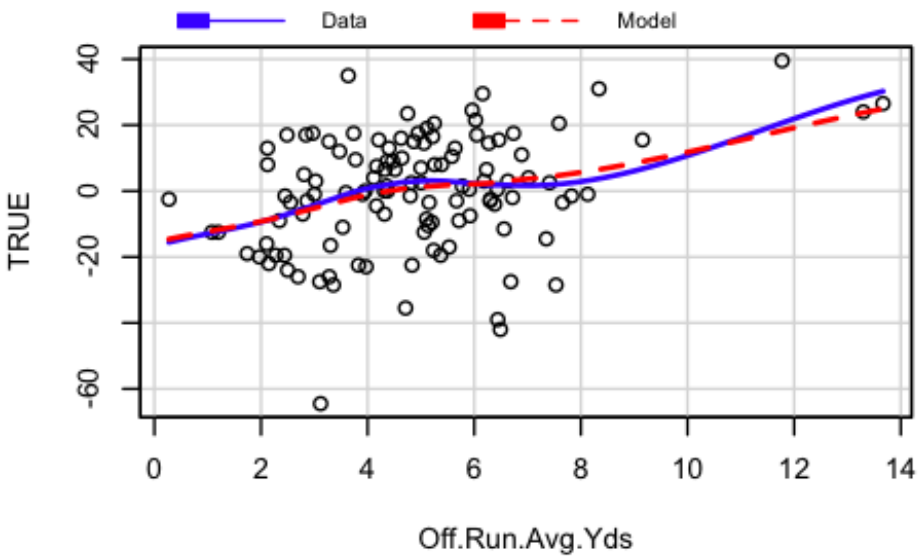

FIGURE 12. Marginal Model Plot of Off.Run.Avg.Yds

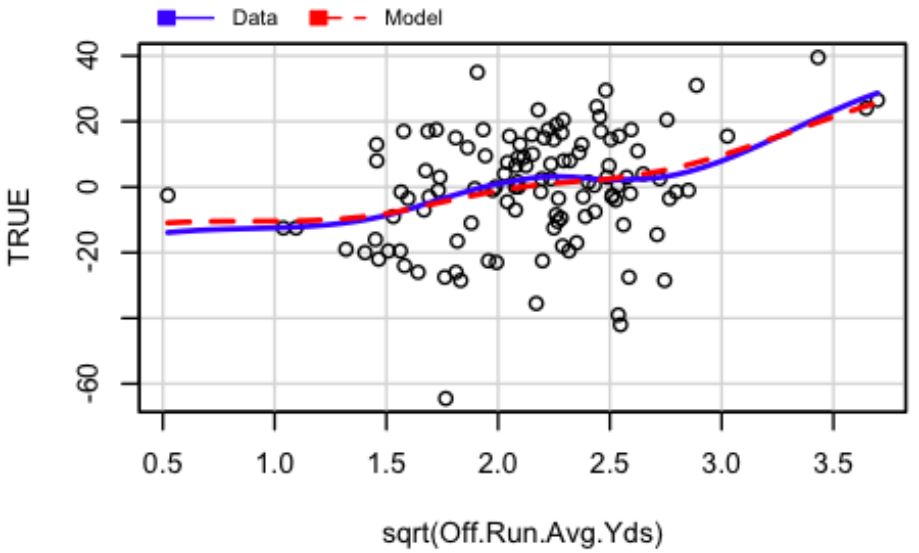

FIGURE 13. Marginal Model Plot of sqrt(Off.Run.Avg.Yds)

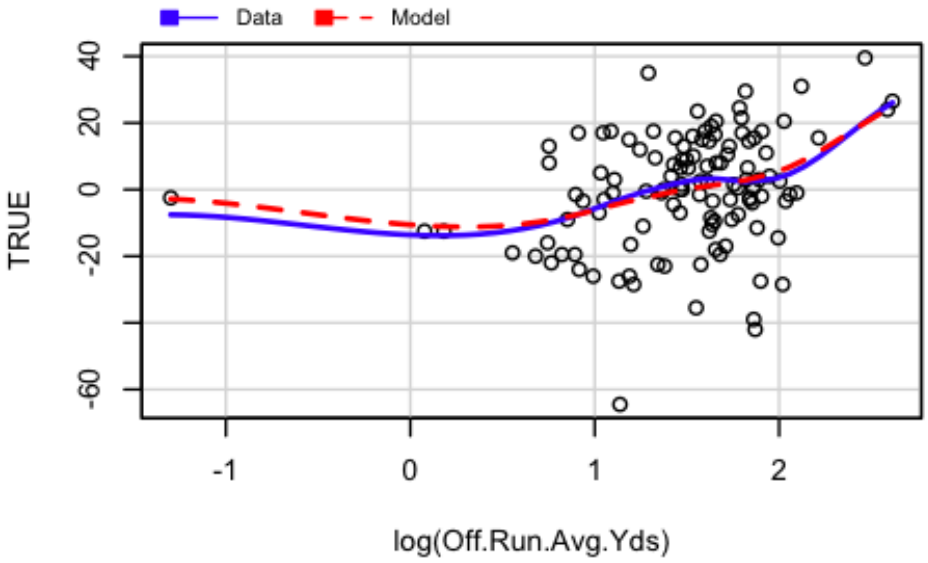

FIGURE 14. Marginal Model Plot of log(Off.Run.Avg.Yds)

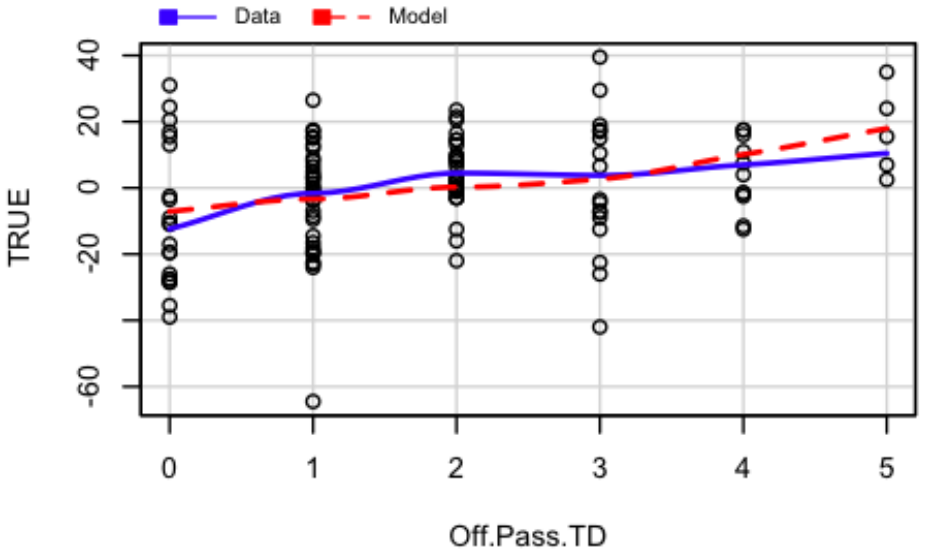

FIGURE 15. Marginal Model Plot of Off.Pass.TD

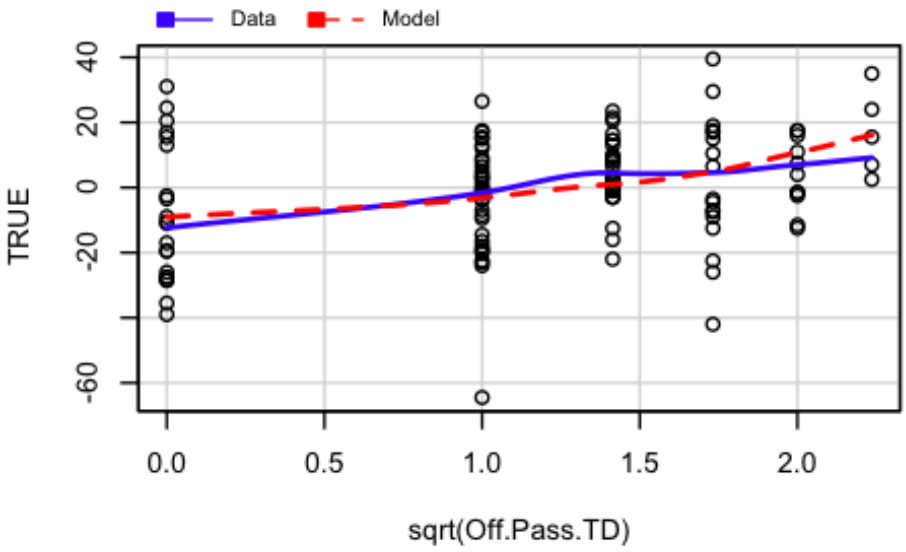

FIGURE 16. Marginal Model Plot of sqrt(Off.Pass.TD)

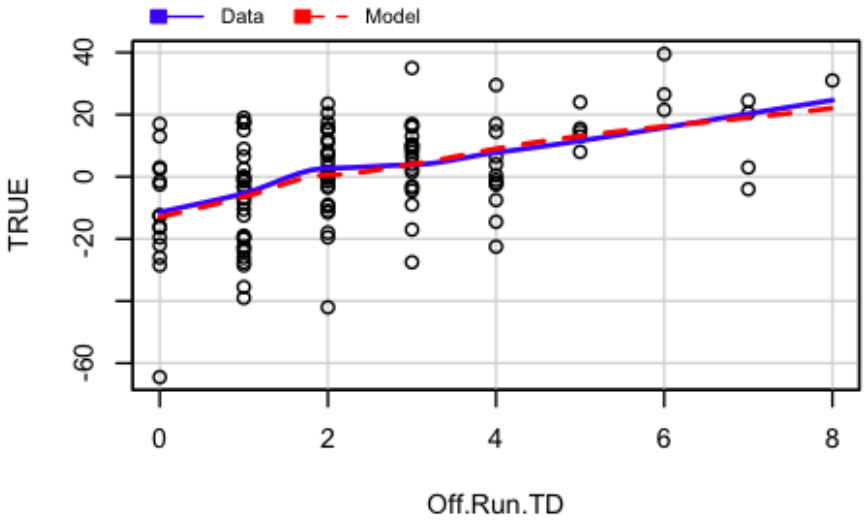

FIGURE 17. Marginal Model Plot of Off.Run.TD

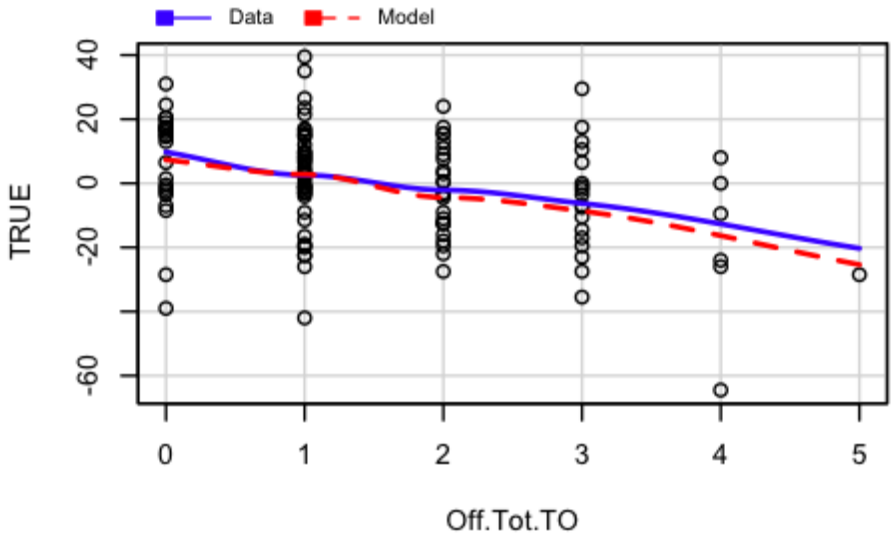

FIGURE 18. Marginal Model Plot of Off.Tot.TO

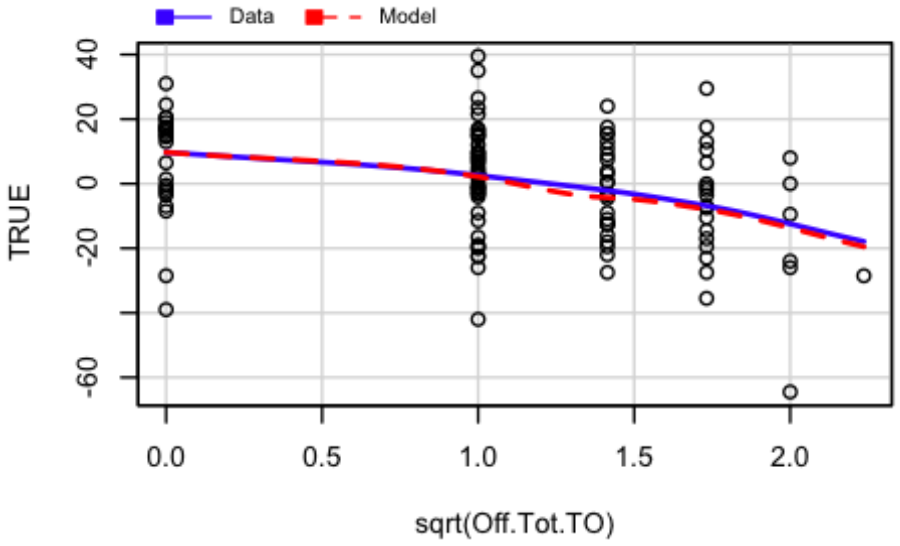

FIGURE 19. Marginal Model Plot of sqrt(Off.Tot.TO)

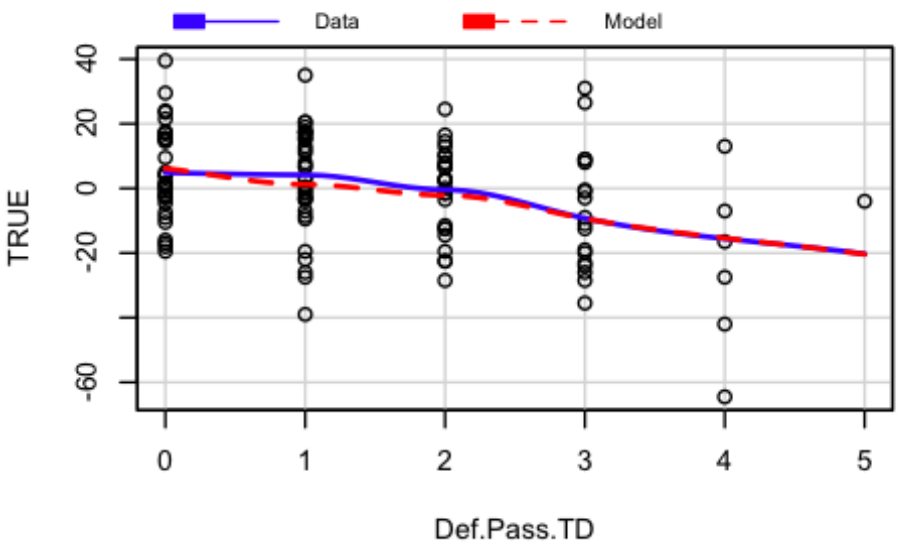

FIGURE 20. Marginal Model Plot of Def.Pass.TD

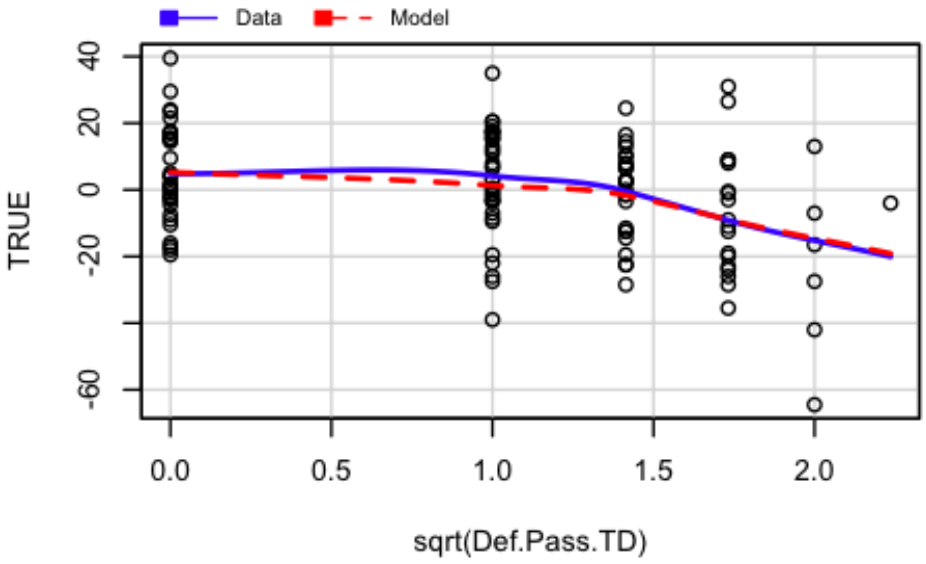

FIGURE 21. Marginal Model Plot of sqrt(Def.Pass.TD)

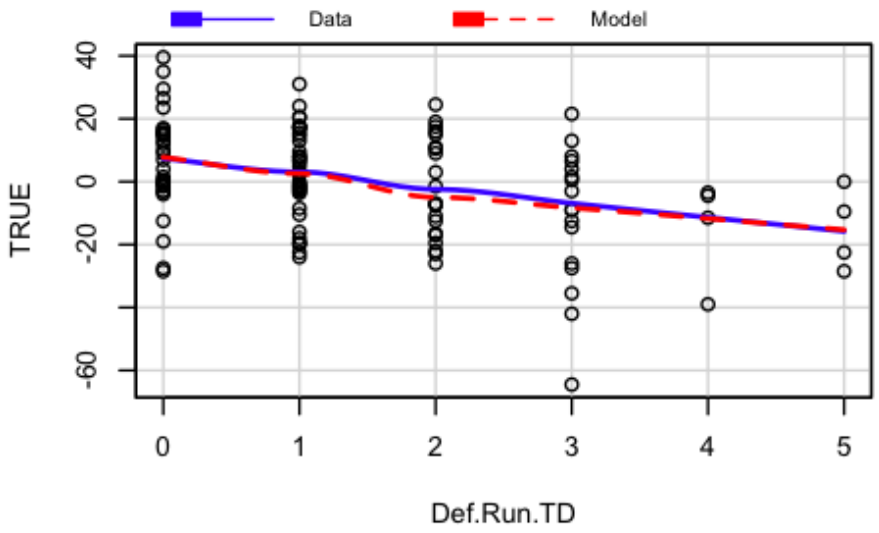

FIGURE 22. Marginal Model Plot of Def.Run.TD

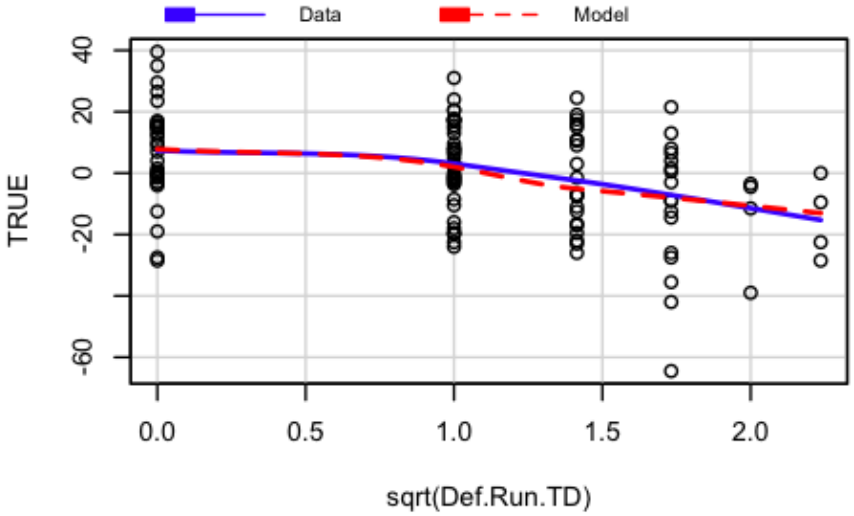

FIGURE 23. Marginal Model Plot of sqrt(Def.Run.TD)

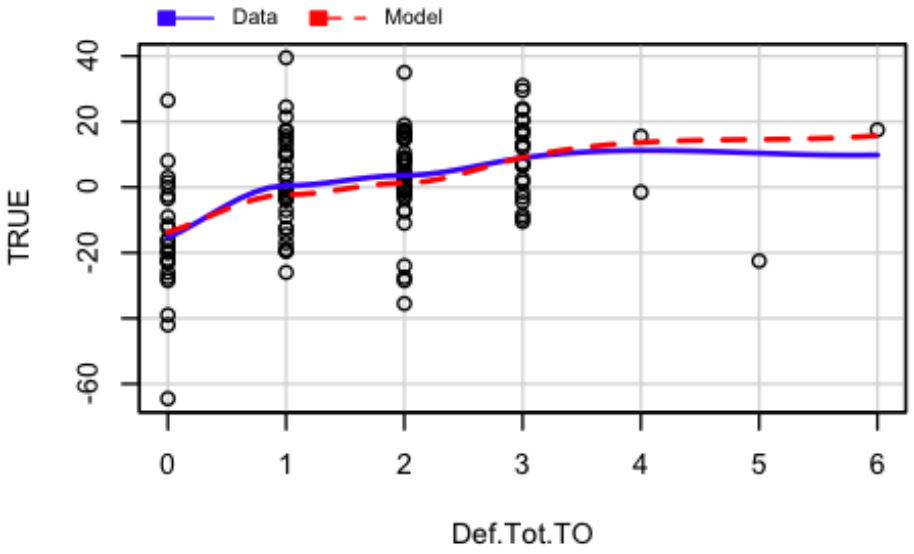

FIGURE 24. Marginal Model Plot of Def.Tot.TO

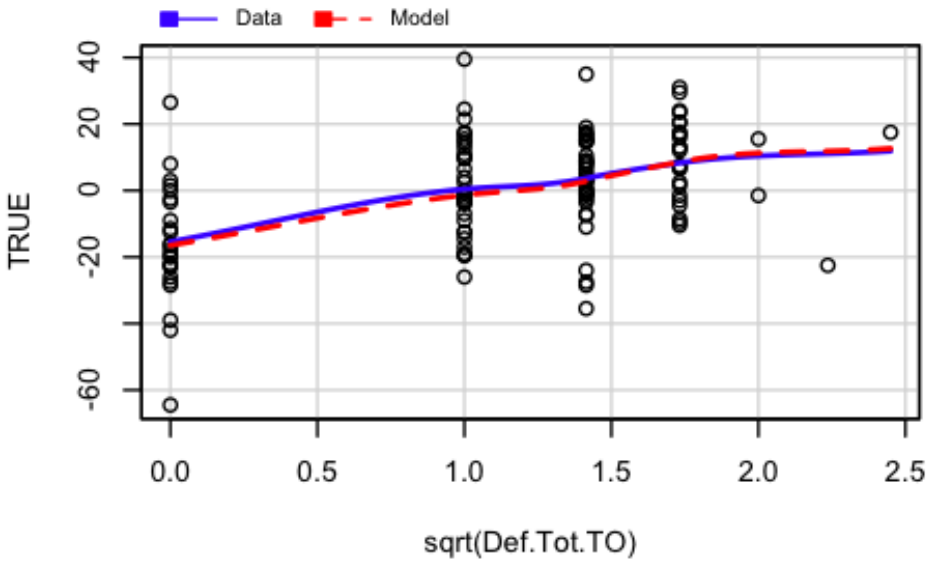

FIGURE 25. Marginal Model Plot of sqrt(Def.Tot.TO)

```
####################################
#R Code for Thesis
############################################
setwd("/Users/Nathan1/Documents")
 a= read.csv("RACC2018Data.csv")
 attach(a)
 #Define our y variable
y=Team.Score-Opp.Score+Spread
a$y=Team.Score-Opp.Score+Spread
#Get an overview of y
summary(y)
#Debug
is.na(y)
#Perform least square regression
output1=lm(y~Off.Pass.Pct+Off.Pass.Yds+Off.Pass.TD+Off.Run.Avg.Yds+Off.Run.TD+Off.Avg.Yds.Play+Off.Tot.First.Down+Off.No.Pen+O
 ff.Pen.Yds+Off.Tot.TO+Def.Pass.Pct+Def.Pass.Yds+Def.Pass.TD+Def.Run.Avg.Yds+Def.Run.TD+Def.Avg.Yds.Play+Def.Tot.First.Down+De
 f.No.Pen+Def.Pen.Yds+Def.Tot.TO)
#Get summary
summary(output1)
#Compare fitted to residual plot
plot(fitted(output1), resid(output1), main="Fitted vs. Residual of Output1", xlab="Fitted", ylab="Residual")

#Histogram of y
hist(y, main="Histogram of Cover.Spread")

#Determine the plots for output1
plot(output1)

#Determine which variables are extraneous for our model and which ones we should keep
step(output1, direction="both")
step(output1, direction="backward")

##lm(formula = y ~ Off.Pass.TD + Off.Run.TD + Off.Tot.First.Down +
  #  Off.Pen.Yds + Off.Tot.TO + Def.Pass.Pct + Def.Run.TD + #Def.Avg.Yds.Play +
    #Def.Pen.Yds + Def.Tot.TO)

#Determine Multiple R-squared for the variables we want to keep
output2=lm(y~Off.Pass.TD+Off.Run.TD+Off.Tot.First.Down+Off.Tot.TO+Def.Run.TD+ Def.Avg.Yds.Play+Def.Pen.Yds+Def.Tot.TO)
summary(output2)
#Fitted vs Residual plot of Output 2
plot(fitted(output2), resid(output2), main="Fitted vs. Residual of Output2", xlab="Fitted", ylab="Residual")
par(mfrow=c(2,2))
plot(output2)

#Correlation of the variables to y

#Marginal Model Plots
install.packages("car")
library(car)
mmps(output1)

b=subset(a, select=c(y,Off.Tot.First.Down,Off.Run.TD, Off.Pass.TD))
plot(b)

#Plot with 9 plots on per page
par(mfrow=c(3,3))
plot(Off.Pass.Pct, y)
plot(Off.Pass.Yds,y)
plot(Off.Run.Avg.Yds,y)
plot(Off.Avg.Yds.Play,y)
plot(Off.Tot.First.Down,y)
plot(Off.No.Pen, y)
```

```
plot(Off.Pen.Yds, y)
plot(Def.Pass.Pct,y)
plot(Def.Pass.Yds, y)

par(mfrow=c(3,2))
plot(Def.Run.Avg.Yds, y)
plot(Def.Avg.Yds.Play,y)
plot(Def.Tot.First.Down,y)
plot(Def.No.Pen, y)
plot(Def.Pen.Yds, y)
#Determine which values are outliers
3
a[50,] #Shows 50th row in dataset

#VIF of Output2
install.packages("car")
library(car)


#Histograms of 14 Continuous Covariates
par(mfrow=c(3,3))
hist(Off.Pass.Pct, breaks=20, main="Histogram of Off.Pass.Pct")
hist(Off.Pass.Yds, breaks=20, main="Histogram of Off.Pass.Yds")
hist(Off.Run.Avg.Yds, breaks=20, main="Histogram of Off.Run.Avg.Yds")
hist(Off.Avg.Yds.Play, breaks=20, main="Histogram of Off.Avg.Yds.Play")
hist(Off.Tot.First.Down, breaks=20, main="Histogram of Off.Tot.First.Down")
hist(Off.No.Pen, breaks=20, main="Histogram of Off.No.Pen")
hist(Off.Pen.Yds, breaks=20, main="Histogram of Off.Pen.Yds")
hist(Def.Pass.Pct, breaks=20, main="Histogram of Def.Pass.Pct")
hist(Def.Pass.Yds, breaks=20, main="Histogram of Def.Pass.Yds")

par(mfrow=c(3,2))
hist(Def.Run.Avg.Yds, breaks=20, main="Histogram of Def.Run.Avg.Yds")
hist(Def.Avg.Yds.Play, breaks=20, main="Histogram of Def.Avg.Yds.Play")
hist(Def.Tot.First.Down, breaks=20, main="Histogram of Def.Tot.First.Down")
hist(Def.No.Pen, breaks=20, main="Histogram of Def.No.Pen")
hist(Def.Pen.Yds, breaks=20, main="Histogram of Def.Pen.Yds")



#Power Transformation of 14 Continuous Covariates
#First, we redefine variables that have 0/negative values
New.Off.Pass.Pct=Off.Pass.Pct+0.2
New.Off.Pass.Yds=Off.Pass.Yds+1
New.Off.No.Pen=Off.No.Pen+0.1
New.Off.Pen.Yds=Off.Pen.Yds+0.1
New.Def.Run.Avg.Yds=Def.Run.Avg.Yds+0.65
New.Def.No.Pen=Def.No.Pen+0.1
New.Def.Pen.Yds=Def.Pen.Yds+0.1


#Now, we perform the power transformation
library(car)
powerTransform(cbind(New.Off.Pass.Pct,New.Off.Pass.Yds, Off.Run.Avg.Yds,Off.Avg.Yds.Play,Off.Tot.First.Down, New.Off.No.Pen,
 New.Off.Pen.Yds, Def.Pass.Pct, Def.Pass.Yds, New.Def.Run.Avg.Yds, Def.Avg.Yds.Play, Def.Tot.First.Down, New.Def.No.Pen,
 New.Def.Pen.Yds,y+64.6))



#Transformation Analysis of Off.Run.Avg.Yds
out1=lm(y~Off.Run.Avg.Yds)
out2=lm(y~sqrt(Off.Run.Avg.Yds))
```

```
out3=lm(y~log(Off.Run.Avg.Yds))


#Obtain Multiple R^2
summary(out1)
summary(out2)
summary(out3)

summary(Off.Run.Avg.Yds)
#Scatter Plot with the regression lines
plot(Off.Run.Avg.Yds, y, main="Scatter Plot of Off.Run.Avg.Yds", ylab="Cover.Spread",)
#Add linear regression model line
abline(out1)
f=seq(0,14,0.001)
yf=-24.327+11.125*sqrt(f)
lines(f,yf,col='red', lty=2)
yf2=-14.244+9.408*log(f)
lines(f,yf2, col='blue', lty=2)
legend(
  "bottomright",
  lty=c(1,1,1),
  col=c("black","red", "blue"),
  legend = c("No Transformation", "Square Root", "Log")
)

#Residual Plot

par(mfrow=c(3,1))
plot(fitted(out1), resid(out1), main="Fitted vs. Residual Off.Run.Avg.Yds")
plot(fitted(out2), resid(out2), main="Fitted vs. Residual sqrt(Off.Run.Avg.Yds)")
plot(fitted(out3), resid(out3), main="Fitted vs. Residual log(Off.Run.Avg.Yds)")



#Transformation Analysis of Off.Avg.Yds.Play
out1=lm(y~Off.Avg.Yds.Play)
out2=lm(y~sqrt(Off.Avg.Yds.Play))

#Obtain Multiple R^2
summary(out1)
summary(out2)

#Scatter Plot with the regression lines
plot(Off.Avg.Yds.Play, y, main="Scatter Plot of Off.Avg.Yds.Play", ylab="Cover.Spread")
#Add linear regression model line
abline(out1)
f=seq(1.6,14,0.001)
yf=-52.55+21.17*sqrt(f)
lines(f,yf,col='red', lty=2)
#Residual Plot

par(mfrow=c(2,1))
plot(fitted(out1), resid(out1), main="Fitted vs. Residual Off.Avg.Yds.Play")
plot(fitted(out2), resid(out2), main="Fitted vs. Residual sqrt(Off.Avg.Yds.Play)")



#Transformation Analysis of Off.Tot.First.Down
out1=lm(y~Off.Tot.First.Down)
out2=lm(y~sqrt(Off.Tot.First.Down))

#Obtain Multiple R^2
```

```
summary(out1)
summary(out2)

summary(Off.Tot.First.Down)
#Scatter Plot with the regression lines
plot(Off.Tot.First.Down, y, main="Scatter Plot of Off.Tot.First.Down", ylab="Cover.Spread")
#Add linear regression model line
abline(out1)
f=seq(7.5,38.5,0.001)

yf=-29.865+6.443*sqrt(f)
lines(f,yf,col='red', lty=2)

#Residual Plot

par(mfrow=c(2,1))
plot(fitted(out1), resid(out1), main="Fitted vs. Residual Off.Tot.First.Down")
plot(fitted(out2), resid(out2), main="Fitted vs. Residual sqrt(Off.Tot.First.Down)")

#Transformation Analysis of New.Off.No.Pen
out1=lm(y~New.Off.No.Pen)
out2=lm(y~sqrt(New.Off.No.Pen))

#Obtain Multiple R^2
summary(out1)
summary(out2)

summary(New.Off.No.Pen)
#Scatter Plot with the regression lines
plot(New.Off.No.Pen, y, main="Scatter Plot of New.Off.No.Pen", ylab="Cover.Spread")
#Add linear regression model line
abline(out1)
f=seq(0,17.5,0.001)

yf=15.736+-6.836*sqrt(f)
lines(f,yf,col='red', lty=2)

#Residual Plot

par(mfrow=c(2,1))
plot(fitted(out1), resid(out1), main="Fitted vs. Residual New.Off.No.Pen")
plot(fitted(out2), resid(out2), main="Fitted vs. Residual sqrt(New.Off.No.Pen)")

#Transformation Analysis of New.Off.Pen.Yds
out1=lm(y~New.Off.Pen.Yds)
out2=lm(y~sqrt(New.Off.Pen.Yds))

#Obtain Multiple R^2
summary(out1)
summary(out2)

summary(New.Off.Pen.Yds)
#Scatter Plot with the regression lines
plot(New.Off.Pen.Yds, y, main="Scatter Plot of New.Off.Pen.Yds", ylab="Cover.Spread")
#Add linear regression model line
abline(out1)
f=seq(0,135,0.001)

yf=12.6767+-1.8956*sqrt(f)
lines(f,yf,col='red', lty=2)

#Residual Plot
```

```
par(mfrow=c(2,1))
plot(fitted(out1), resid(out1), main="Fitted vs. Residual New.Off.Pen.Yds")
plot(fitted(out2), resid(out2), main="Fitted vs. Residual sqrt(New.Off.Pen.Yds)")

#Transformation Analysis of Def.Pass.Yds
out1=lm(y~ Def.Pass.Yds)
out2=lm(y~sqrt(Def.Pass.Yds))
out3=lm(y~log(Def.Pass.Yds))


#Obtain Multiple R^2
summary(out1)
summary(out2)
summary(out3)

summary(Def.Pass.Yds)
#Scatter Plot with the regression lines
plot(Def.Pass.Yds, y, main="Scatter Plot of Def.Pass.Yds", ylab="Cover.Spread")
#Add linear regression model line
abline(out1)
f=seq(36,511,0.001)
yf=13.4897+-0.9487*sqrt(f)
lines(f,yf,col='red', lty=2)
yf2=28.910+-5.522*log(f)
lines(f,yf2, col='blue', lty=2)
#Residual Plot

par(mfrow=c(3,1))
plot(fitted(out1), resid(out1), main="Fitted vs. Residual Def.Pass.Yds")
plot(fitted(out2), resid(out2), main="Fitted vs. Residual sqrt(Def.Pass.Yds)")
plot(fitted(out3), resid(out3), main="Fitted vs. Residual log(Def.Pass.Yds)")


##Transformation Analysis of New.Def.No.Pen
#####
out1=lm(y~ New.Def.No.Pen)
out2=lm(y~sqrt(New.Def.No.Pen))

#Obtain Multiple R^2
summary(out1)
summary(out2)

summary(New.Def.No.Pen)
#Scatter Plot with the regression lines
plot(New.Def.No.Pen, y, main="Scatter Plot of New.Def.No.Pen", ylab="Cover.Spread")
#Add linear regression model line
abline(out1)
f=seq(0,16.5,0.001)

yf=-4.590+1.880*sqrt(f)
lines(f,yf,col='red', lty=2)

#Residual Plot

par(mfrow=c(2,1))
plot(fitted(out1), resid(out1), main="Fitted vs. Residual New.Def.No.Pen")
plot(fitted(out2), resid(out2), main="Fitted vs. Residual sqrt(New.Def.No.Pen)")

#Transformation Analysis of New.Def.Pen.Yds
##
out1=lm(y~New.Def.Pen.Yds)
```

```
out2=lm(y~sqrt(New.Def.Pen.Yds))

#Obtain Multiple R^2
summary(out1)
summary(out2)

summary(New.Def.Pen.Yds)
#Scatter Plot with the regression lines
plot(New.Def.Pen.Yds, y, main="Scatter Plot of New.Def.Pen.Yds", ylab="Cover.Spread")
#Add linear regression model line
abline(out1)
f=seq(0,134.5,0.001)

yf=-3.7318+0.5241*sqrt(f)
lines(f,yf,col='red', lty=2)

#Residual Plot

par(mfrow=c(2,1))
plot(fitted(out1), resid(out1), main="Fitted vs. Residual New.Def.Pen.Yds")
plot(fitted(out2), resid(out2), main="Fitted vs. Residual sqrt(New.Def.Pen.Yds)")

#Do discrete transformations with unaltered continuous covariates
output_original=lm(y~Off.Pass.Pct+Off.Pass.Yds+Off.Pass.TD+Off.Run.Avg.Yds+Off.Run.TD+Off.Avg.Yds.Play+Off.Tot.First.Down+Off.No
 .Pen+New.Off.Pen.Yds+Off.Tot.TO+Def.Pass.Pct+Def.Pass.Yds+Def.Pass.TD+Def.Run.Avg.Yds+Def.Run.TD+Def.Avg.Yds.Play+Def.Tot.Fir
 st.Down+New.Def.No.Pen+Def.Pen.Yds+Def.Tot.TO)
library(car)
mmps(output_original)

#Discrete Transformation
output_transform=lm(y~Off.Pass.Pct+Off.Pass.Yds+Off.Pass.TD+Off.Run.Avg.Yds+Off.Run.TD+Off.Avg.Yds.Play+sqrt(Off.Tot.First.Down)
 +Off.No.Pen+sqrt(New.Off.Pen.Yds)+Off.Tot.TO+Def.Pass.Pct+Def.Pass.Yds+Def.Pass.TD+Def.Run.Avg.Yds+Def.Run.TD+Def.Avg.Yds.Pl
 ay+Def.Tot.First.Down+sqrt(New.Def.No.Pen)+Def.Pen.Yds+Def.Tot.TO)
summary(output_transform)

##Compare Marginal Model Plot
library(car)
mmps(output_transform)

#Stat wise selection
output_stat_wise=lm(y~Off.Pass.Pct+Off.Pass.Yds+sqrt(Off.Pass.TD)+sqrt(Off.Run.Avg.Yds)+Off.Run.TD+Off.Avg.Yds.Play+sqrt(Off.Tot.F
 irst.Down)+Off.No.Pen+sqrt(New.Off.Pen.Yds)+sqrt(Off.Tot.TO)+Def.Pass.Pct+Def.Pass.Yds+Def.Pass.TD+Def.Run.Avg.Yds+Def.Run.TD
 +Def.Avg.Yds.Play+Def.Tot.First.Down+sqrt(New.Def.No.Pen)+Def.Pen.Yds+sqrt(Def.Tot.TO))
step(output_stat_wise, direction='backward')



#Remove covariates one at a time with p-values>0.1, starting with all 20 variables in output_remove
output_remove=lm(y~sqrt(Off.Pass.TD)+Off.Run.TD+sqrt(Off.Tot.First.Down)+sqrt(Off.Tot.TO)+Def.Run.TD+Def.Avg.Yds.Play+Def.Pen.Y
 ds+sqrt(Def.Tot.TO))

summary(output_remove)
plot(output_remove)
plot(fitted(output_remove),resid(output_remove))

##Compare the manual process to the stat wise selection
step(output_remove, direction='backward')


##Find VIF of the model with the 8 covariates
model_1=lm(y~Off.Pass.TD+Off.Run.TD+sqrt(Off.Tot.First.Down)+Off.Tot.TO+Def.Run.TD+Def.Avg.Yds.Play+Def.Pen.Yds+Def.Tot.TO)
library(car)
```

```
vif(model_1)

#Rename the variables
sqrt.Off.Pass.TD=sqrt(Off.Pass.TD)
sqrt.Off.Tot.First.Down=sqrt(Off.Tot.First.Down)
sqrt.Off.Tot.TO=sqrt(Off.Tot.TO)
sqrt.Def.Tot.TO=sqrt(Def.Tot.TO)

#Final Model
model_final=lm(y~sqrt.Off.Pass.TD+Off.Run.TD+sqrt.Off.Tot.First.Down+sqrt.Off.Tot.TO+Def.Run.TD+Def.Avg.Yds.Play+Def.Pen.Yds+sq
 rt.Def.Tot.TO)
summary(model_final)

#VIF of Final Model
library(car)
vif(model_final)

#Plots of Final Model
par(mfrow=c(2,2))
plot(model_final)

library(car)
mmps(model_final)

#Shapiro-Wilk Test
t1=shapiro.test(resid(model_final))
names(t1)

#Gives exact value
t1$p.value



##Compare with model with no transformations
output1=lm(y~Off.Pass.Pct+Off.Pass.Yds+Off.Pass.TD+Off.Run.Avg.Yds+Off.Run.TD+Off.Avg.Yds.Play+Off.Tot.First.Down+Off.No.Pen+O
 ff.Pen.Yds+Off.Tot.TO+Def.Pass.Pct+Def.Pass.Yds+Def.Pass.TD+Def.Run.Avg.Yds+Def.Run.TD+Def.Avg.Yds.Play+Def.Tot.First.Down+De
 f.No.Pen+Def.Pen.Yds+Def.Tot.TO)

#Perform model selection, removing one at a time

model_m2=lm(y~Off.Pass.TD+Off.Run.TD+Off.Tot.TO+Def.Run.TD+Def.Avg.Yds.Play+Def.Pen.Yds+Def.Tot.TO)

summary(output1_remove)
#Plots
par(mfrow=c(2,2))
plot(output1_remove)

library(car)
mmps(output1_remove)


##Cross-Validate with 2019 Data
b=read.csv("RACCData2019.csv")

#Define variables appropriately
b$y=b$Team.Score-b$Opp.Score+b$Spread
b$sqrt.Off.Pass.TD=sqrt(b$Off.Pass.TD)
b$sqrt.Off.Tot.First.Down=sqrt(b$Off.Tot.First.Down)
b$sqrt.Off.Tot.TO=sqrt(b$Off.Tot.TO)
b$sqrt.Def.Tot.TO=sqrt(b$Def.Tot.TO)
```

```
#Get predicted values based on 2019 data with model_final
try=predict(model_final, b)

#Make sure the predict command was accurate
try2=summary(model_final)$coef[1]+summary(model_final)$coef[2]*b$sqrt.Off.Pass.TD+summary(model_final)$coef[3]*b$Off.Run.T
 D+summary(model_final)$coef[4]*b$sqrt.Off.Tot.First.Down+summary(model_final)$coef[5]*b$sqrt.Off.Tot.TO+summary(model_fina
 l)$coef[6]*b$Def.Run.TD+summary(model_final)$coef[7]*b$Def.Avg.Yds.Play+summary(model_final)$coef[8]*b$Def.Pen.Yds+summa
 ry(model_final)$coef[9]*b$sqrt.Def.Tot.TO

test=try-try2

#Get mean square error
mse_final=(sum(b$y-try2)^2)/118

#Repeat using Model M2
try=predict(model_m2, b)
try2=summary(model_m2)$coef[1]+summary(model_m2)$coef[2]*b$Off.Pass.TD+summary(model_m2)$coef[3]*b$Off.Run.TD+summ
 ary(model_m2)$coef[4]*b$Off.Tot.TO+summary(model_m2)$coef[5]*b$Def.Run.TD+summary(model_m2)$coef[6]*b$Def.Avg.Yds.Pl
 ay+summary(model_m2)$coef[7]*b$Def.Pen.Yds+summary(model_m2)$coef[8]*b$Def.Tot.TO
test=try-try2

mse_m2=(sum(b$y-try2)^2)/118

#Load in Big Ten 2018 Data
c=read.csv("BigTen2018.csv")

#Define variables
c$y=c$Team.Score-c$Opp.Score+c$Spread
c$sqrt.Off.Pass.TD=sqrt(c$Off.Pass.TD)
c$sqrt.Off.Tot.First.Down=sqrt(c$Off.Tot.First.Down)
c$sqrt.Off.Tot.TO=sqrt(c$Off.Tot.TO)
c$sqrt.Def.Tot.TO=sqrt(c$Def.Tot.TO)

#Get predicted values based on 2019 data with model_final
try=predict(model_final, c)
try2=summary(model_final)$coef[1]+summary(model_final)$coef[2]*c$sqrt.Off.Pass.TD+summary(model_final)$coef[3]*c$Off.Run.TD
 +summary(model_final)$coef[4]*c$sqrt.Off.Tot.First.Down+summary(model_final)$coef[5]*c$sqrt.Off.Tot.TO+summary(model_final)
 $coef[6]*c$Def.Run.TD+summary(model_final)$coef[7]*c$Def.Avg.Yds.Play+summary(model_final)$coef[8]*c$Def.Pen.Yds+summary
 (model_final)$coef[9]*c$sqrt.Def.Tot.TO

#Want to get all zeros
test=try-try2

mse_big_ten=(sum(c$y-try2)^2)/98
mse_big_ten

#Assess model_final with Big 12 2018 regular season data
d=read.csv("Big122018.csv")
d$y=d$Team.Score-d$Opp.Score+d$Spread
d$sqrt.Off.Pass.TD=sqrt(d$Off.Pass.TD)
d$sqrt.Off.Tot.First.Down=sqrt(d$Off.Tot.First.Down)
d$sqrt.Off.Tot.TO=sqrt(d$Off.Tot.TO)
d$sqrt.Def.Tot.TO=sqrt(d$Def.Tot.TO)

#Get predicted values based on 2019 data with model_final
try=predict(model_final, d)
try2=summary(model_final)$coef[1]+summary(model_final)$coef[2]*d$sqrt.Off.Pass.TD+summary(model_final)$coef[3]*d$Off.Run.T
 D+summary(model_final)$coef[4]*d$sqrt.Off.Tot.First.Down+summary(model_final)$coef[5]*d$sqrt.Off.Tot.TO+summary(model_fina
 l)$coef[6]*d$Def.Run.TD+summary(model_final)$coef[7]*d$Def.Avg.Yds.Play+summary(model_final)$coef[8]*d$Def.Pen.Yds+summa
 ry(model_final)$coef[9]*d$sqrt.Def.Tot.TO

#Want to get all zeros
```

```
test=try-try2
test

mse_big_12=(sum(d$y-try2)^2)/70
mse_big_12

#Load 2018 ACC Postseason Data
p=read.csv("RACC2018Postseason.csv")
p$y=p$Team.Score-p$Opp.Score+p$Spread
p$sqrt.Off.Pass.TD=sqrt(p$Off.Pass.TD)
p$sqrt.Off.Tot.First.Down=sqrt(p$Off.Tot.First.Down)
p$sqrt.Off.Tot.TO=sqrt(p$Off.Tot.TO)
p$sqrt.Def.Tot.TO=sqrt(p$Def.Tot.TO)

#Get predicted values based on 2019 data with model_final
try=predict(model_final, p)
try2=summary(model_final)$coef[1]+summary(model_final)$coef[2]*p$sqrt.Off.Pass.TD+summary(model_final)$coef[3]*p$Off.Run.T
 D+summary(model_final)$coef[4]*p$sqrt.Off.Tot.First.Down+summary(model_final)$coef[5]*p$sqrt.Off.Tot.TO+summary(model_fina
 l)$coef[6]*p$Def.Run.TD+summary(model_final)$coef[7]*p$Def.Avg.Yds.Play+summary(model_final)$coef[8]*p$Def.Pen.Yds+summa
 ry(model_final)$coef[9]*p$sqrt.Def.Tot.TO

#Want to get all zeros
test=try-try2
test

mse_acc_post=(sum(p$y-try2)^2)/12
mse_acc_post
```