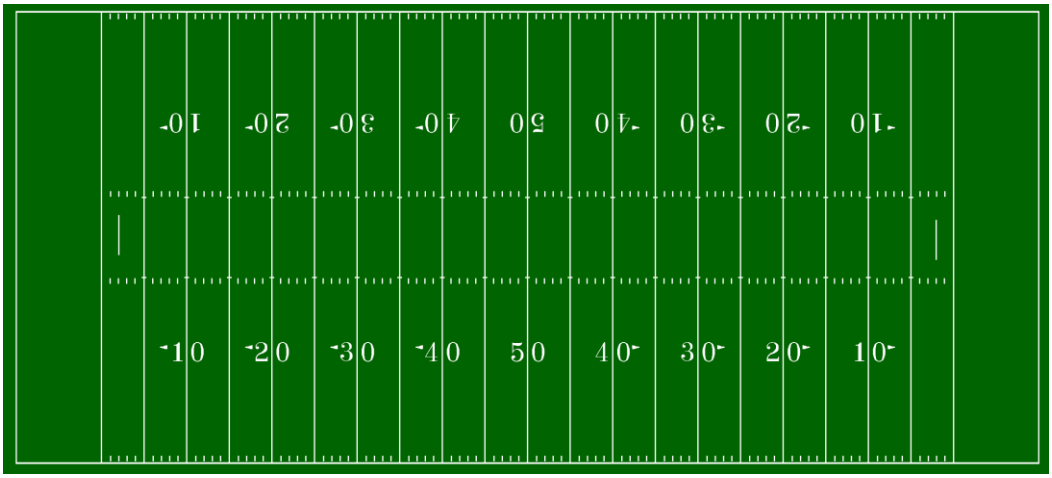
Jacob Anderson
MATH 4753
Linear Regression Project

Introduction

American football is a relatively complex sport with a large pool of statistics available for evaluating what makes a "good" team. One important facet of the game, often called the "field position battle", involves both minimizing the distance your team's offense must move the ball to score and maximizing the distance your opponent's offense must move the ball to score.

To understand the data used in this project, a "drive" must be defined and a description of the measurements of an NFL playing field is necessary. A "drive" is simply a series of plays in which a team's offense has possession of the ball. From the offensive perspective, a drive may consist of any non-zero number of plays and may end in a score, an intentional change of possession, a turnover (unintentional change of possession), or the end of either half (set amount of time to be played). An NFL playing field is 100 yards long, split into two sections of 50 counting down from the middle. So there is one "50 yard line" which is the center of the field and two "x yard lines" where x is a whole number between 1 and 49.



The relevant data used in this project consists of the average number of points scored per drive and the average drive start position (in yards from their own end zone) for each of the 32 teams in the NFL for the 2015 season. The average number of points scored per drive at first seems like a direct measure of offensive efficiency, but this project will attempt to show a linear relationship between the drive start position (strongly influenced by both defense and special teams play) and the average points scored per drive. This will also serve to show how a good/bad defense and/or special teams can positively/negatively influence a team's scoring. The results may also give a model for predicting the change in scoring that will occur due to a fairly significant 2016 rule change involving field position.

Research Question

This project will aim to determine how the average number of points scored per drive by NFL teams is affected by the average starting field position of their drives.

Analysis Used

The research question will be answered using simple linear regression analysis. This method is appropriate for the given data because knowledge of the average drive start of an NFL team allows for more accurate estimation of that team's average points per drive than simply using the mean points per drive for all NFL teams.

Simple Linear Regression Analysis in this case will involve examination of the expected value of Avg_Pts given Avg_Start. The following assumptions will be made and shown to be true in the next section:

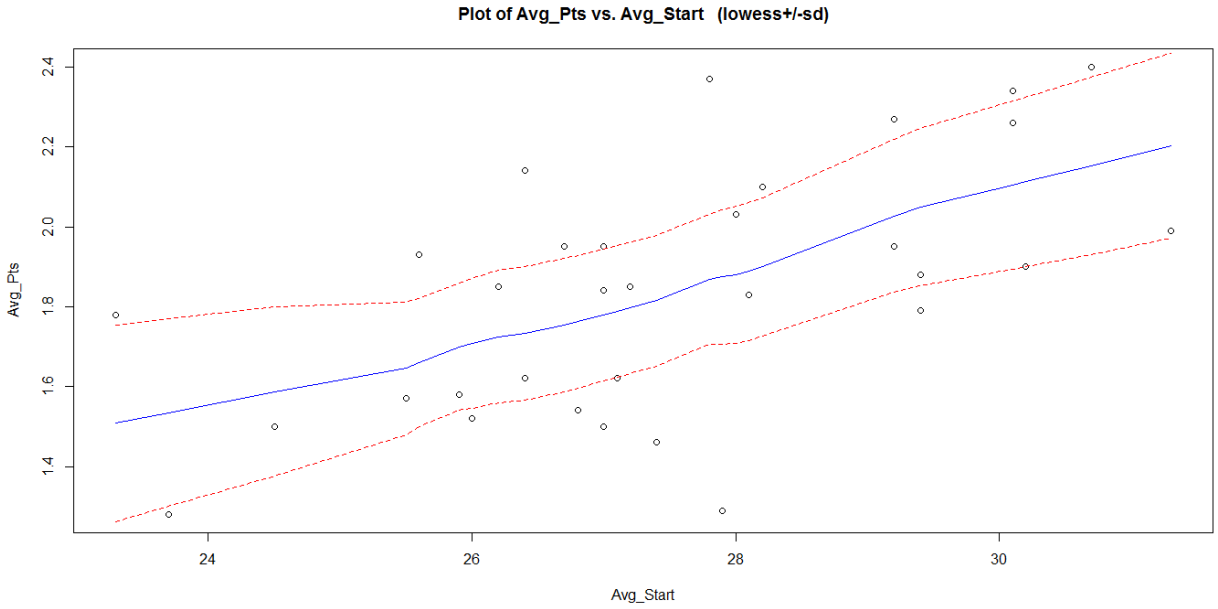
- The trend is linear.
- $\varepsilon_i \sim N(0, \sigma^2)$
- This assumes that errors are independent.

Using this analysis method, a linear model will be created using R and the s20x library will be used for various plotting. The primary goal will be to determine values of $\hat{\beta}_0$ and $\hat{\beta}_1$ such that $\overline{Y} = \hat{\beta}_0 + \hat{\beta}_1 * \overline{X}$ where Y and X are the Avg_Pts and Avg_Start data, respectively. These estimates will be found in the following ways:

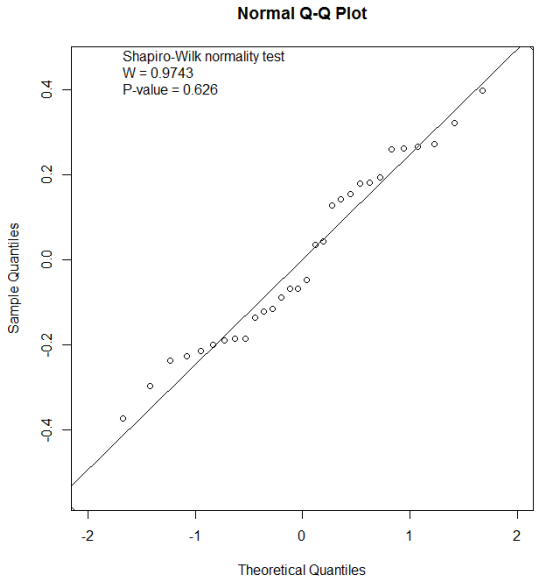
- $\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}$
- $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 * \overline{X}$

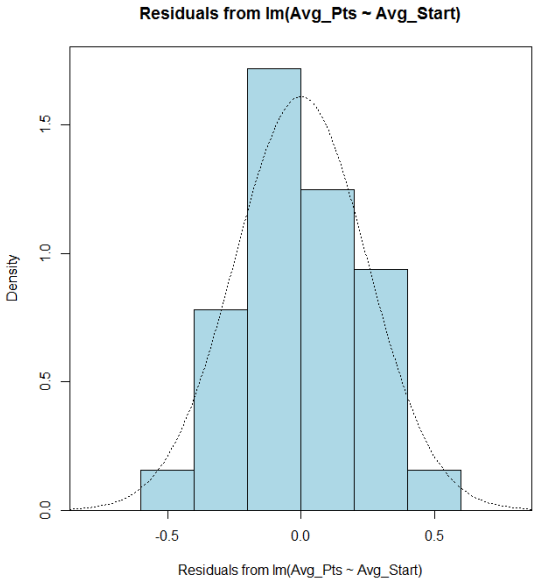
Validity

First, to show that there is a linear trend in the data to justify creating a linear model, we plot a lowess smoother with a smoothing constant of 2/3 over a scatter plot of Avg_Pts vs. Avg_Start. The smoothing constant was chosen by trial and error. There is an immediately identifiable positive linear relationship among the data.

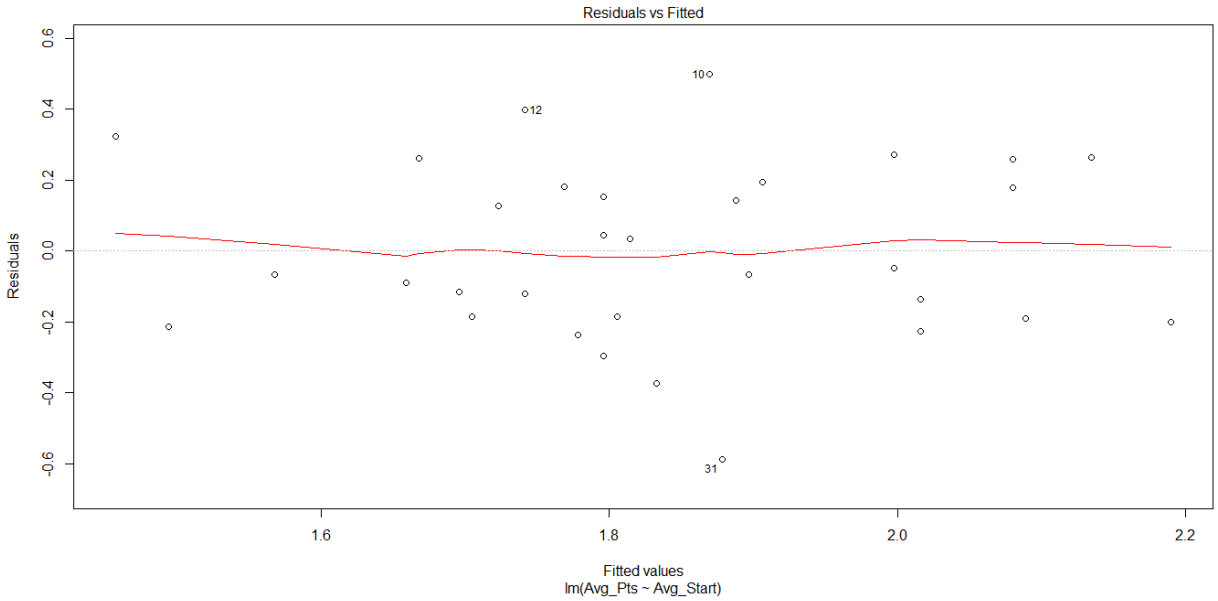


Additionally, the normality assumption of the data must be satisfied. This is done using a Q-Q Plot and a Shapiro-Wilk normality test. This test involves a null hypothesis H_0: $\varepsilon_i \sim N$, and the resulting p-value is 0.626. This shows that there is not nearly enough evidence to support the rejection of the null hypothesis. Additionally, because n = 32 (> 30), the CLT could also be applied to validate the normality assumption.





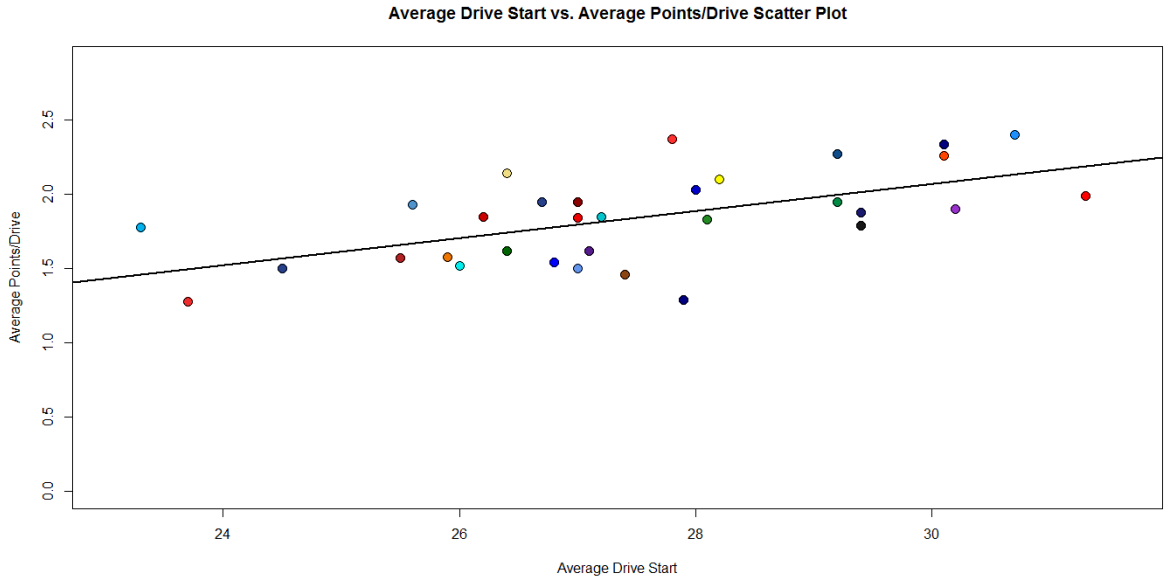
Next, we show constant variance by plotting the residuals vs. fitted values and observe the trend. The line is not strictly constant, but the variance is small enough to ignore, satisfying the constant variance assumption.



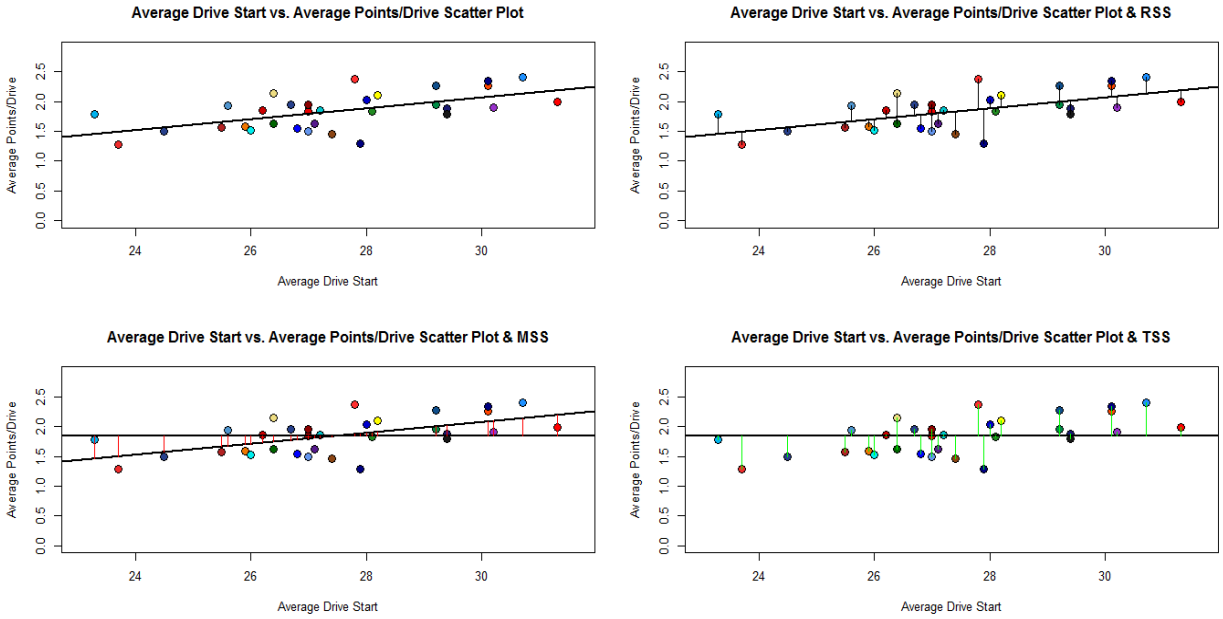
Lastly, the observed data must be shown to be independent. Because of the nature of the sport, one team's average points scored is trivially independent of its opponent's. In an individual game, one team's average drive start position may affect the average drive start position of the other team, but this influence is fairly small among various more strongly influencing factors. Additionally, this data is averaged from an entire season where teams face 13 distinct opponents, basically eliminating this minor influence. Thus, the data independence assumption is satisfied.

Analysis

After creating the linear model, we can plot the trend line over the data to get a quick judge of its accuracy. At first glance it seems to be a fairly good predictor, although the values are not as tightly grouped around the trend line as we'd like.



To judge the accuracy of the model further, we can compute and plot the residual sum of squares, model sum of squares, and total sum of squares over the data. The resulting MSS/TSS value (Multiple R^2) is 0.3457813. This means that the created model explains ≈34.6% of the data variance. This is not a particularly high value, and indicates that there are other factors at work here. These other factors are outside the scope of this model, however, and the model is still useful.



A summary of the model is shown below. From this we can see that the calculated Multiple R^2 value was correct and that the Adjusted R^2 value is similar. We also get point estimates for $\hat{\beta}_0$ and $\hat{\beta}_1$ of -0.67607 and 0.09157, respectively.

```
Call:
lm(formula = Avg_Pts ~ Avg_Start)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5886 -0.1863 -0.0573  0.1844  0.5005

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.67607    0.63342  -1.067 0.294340
Avg_Start    0.09157    0.02300   3.982 0.000401 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2515 on 30 degrees of freedom
Multiple R-squared:  0.3458, Adjusted R-squared:  0.324
F-statistic: 15.86 on 1 and 30 DF,  p-value: 0.0004012
```

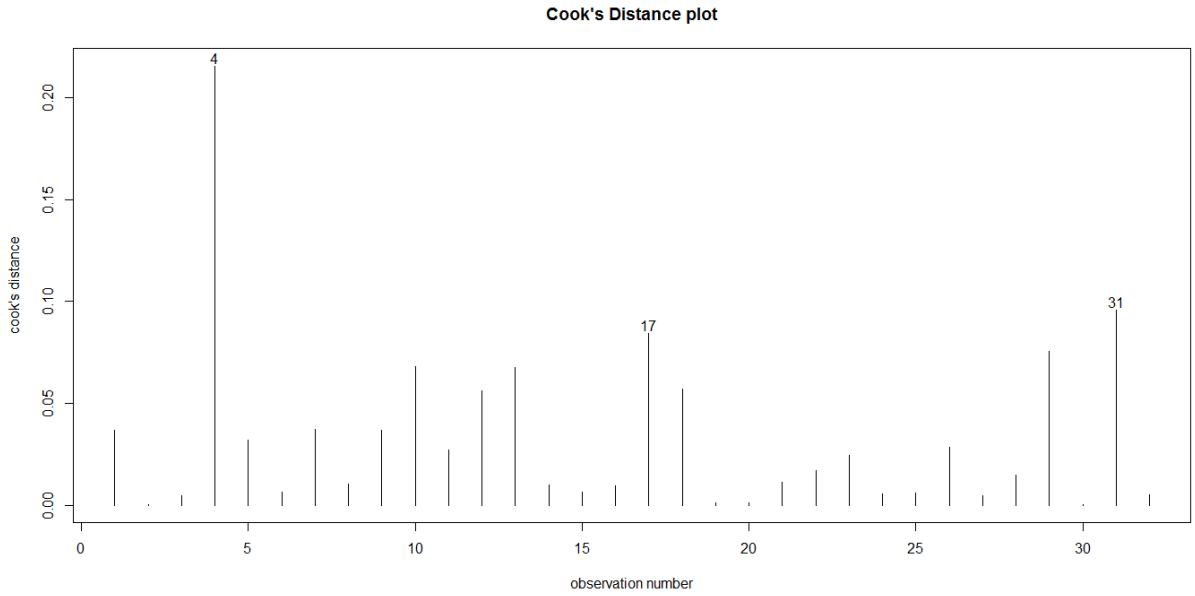
Additionally, we can obtain 95% confidence intervals for $\hat{\beta}_0$ and $\hat{\beta}_1$:

```
            95 % C.I.lower    95 % C.I.upper
(Intercept)       -1.96969           0.61756
Avg_Start          0.04460           0.13853
```

Using the model, we can predict values of Avg_Pts from given values of Avg_Start. Some examples are given:

- E(Avg_Pts | 20) = 1.155256
- E(Avg_Pts | 25) = 1.613088
- E(Avg_Pts | 30) = 2.070919

Using the below Cook's Distance plot, we can identify potential outliers in the data. Because each piece of data is a 16 game average, the possibility of erroneous data can be ignored. This leads to the conclusion that these outliers are good indicators of the additional influencers not accounted for by the model. A quick look at the additional data shows that the largest outlier, the Dallas Cowboys (4 below), had the highest turnover percentage among all NFL teams. The other two labeled values in the plot are the Baltimore Ravens and the Jacksonville Jaguars who each had worse-than-average turnover percentages. This is fairly convincing evidence that an improved version of this model would include turnover percentage as an additional predicting variable.



Conclusion

The goal of this project was to determine how the average number of points scored per drive by NFL teams is affected by the average starting field position of their drives. This goal was achieved with a simple linear regression model. The model satisfies all necessary assumptions, and fits the data relatively well.

Therefore, the research question can be answered with the estimates given by the model. With 95% confidence, we can say that for every yard of average drive starting field position, the average points per drive scored by a team will change by between 0.0446 and 0.1385 points. Assuming an average of 11 drives per game (a rough estimate) this translates to between 0.4906 and 1.5235 points per game.

As expected, evidence of other contributing factors arose during the analysis of this model. An improved version of this model utilizing the concepts of multiple regression would likely incorporate turnover percentage data along with other suspected factors.

One immediate application of the model involves a rule change for the upcoming 2016 NFL season. This rule involves placing the ball at the 25 yard line (instead of the 20) after a touchback. Estimating (very roughly) that this will increase the average drive starting position by about 2 yards, the model (+ my assumptions) would roughly predict that this rule change will increase team per game scoring by between 1 and 3 points.

Data Source: http://www.pro-football-reference.com/years/2015/