UNITED STATES MILITARY ACADEMY

FINAL PROJECT: HOW MANY POINTS IS A WIN IN THE NFL?

MA388: SABERMETRICS

SECTION C1

LTC MICHAEL POWELL

By

CDT HALEY POTT '24, CO A2
CDT SAMANTHA SHEPPARD-MOORE '24, CO H1

WEST POINT, NEW YORK

8 MAY 2024

POTT.HALEY.RHODES.1591216399 Digitally signed By POTT.HALEY.RHODES.1591216399 Date: 2024.05.08 09:49:08 -04'00'

I CERTIFY THAT I HAVE COMPLETELY DOCUMENTED ALL SOURCES THAT I USED TO COMPLETE THIS ASSIGNMENT AND THAT I ACKNOWLEDGED ALL ASSISTANCE I RECEIVED IN THE COMPLETION OF THIS ASSIGNMENT.

_____ I CERTIFY THAT I DID NOT USE ANY SOURCES OR RECEIVE ANY ASSISTNCE REQUIRING DOCUMENTATION WHILE COMPLETING THIS ASSIGNMENT.

SIGNATURE: POTT.HALEY.RHODES.1591216399 Digitally signed by POTT.HALEY.RHODES.1591216399 Date: 2024.05.08 09:49:30 -04'00'

SHEPPARD-MOORE.SAMANTHA.MARIAN.1593582260 Digitally signed by SHEPPARD-MOORE.SAMANTHA.MARIAN.1593582260 Date: 2024.05.08 09:50:06 -04'00'

SCORE Project

How many points was a win in the 2022 season of the National Football League (NFL)?

CDT Haley Pott and CDT Samantha Sheppard

2024-05-10

```
library(tidyverse)
library(knitr)
library(ggrepel)
library(broom)
library(readxl)
```

Learning Goals

- Understand how many points was a win the the 2022 season of the National Football League.
- Understand linear regression to include the correct interpretation of coefficients.
- Create a linear regression model to determine how many points results in a win.
- Critique the model and discuss if a linear model is appropriate.

Introduction

At the end of the 2022 season the two teams with the worst record, the Chicago Bears and the Houston Texans, were concerned with how to secure more regular season wins. They could have analyzed whether it was more effective to invest in their defense or their offense, or which players were worth the high price tag. Sabermetrics suggests that a better statistic could be used to analyze, throughout a season, how many points a team needs to out score their opponent. When conducting this analysis on baseball, the general rule is that 10 runs will lead to an additional win. We will apply this analysis to football, and compare the score differential and the number of wins per team. We want to know: ***how many points was a win in the 2022 season of the National Football League (NFL)?***

Data

The dataset was found on Kaggle and includes play-by-play data from NFL games spanning from 1999-2022. The data was collected by a group of Carnegie Mellon University researchers Maksim Horowitz, Ron Yurko, and Sam Ventura who were interested in the statistical study of football. They were inspired by play-by-play datasets of other sports such as baseball (pitchF/x) and hockey (nhlscrapr). The team built this R package (nflscrapR) which includes data on the player, play, game, and season level (Horowitz, 2019). Our research explores how many points is a win in a typical NFL game. We sorted the dataset to only include seasonal data of each team during the 2022 NFL season. The data includes seasonal statistics on all 32 NFL teams such as pass completion percentage, average yards gained, total interceptions, total sacks, total wins and losses, score differential, etc.

Methods/Instructional Content

- Applied Linear Regression is a statistics textbook that explores the nuances of linear regression, non-linear regression, and binomial and Poisson regression. We will use this resource to understand how to format a learning module about linear regression.

Weisberg, S. 2014. Applied Linear Regression. 4th ed. Hoboken, NY: John Wiley & Sons Inc. Accessed from: https://ebookcentral.proquest.com/lib/usma/reader.action?docID=7103845.

- Mulholland and Jensen investigate tight ends' success in the NFL by using linear regression on pre-draft data to predict NFL draft order and career success. We will use this article as a model for how to conduct linear regression using NFL data and present linear regression results.

Mulholland, J. and Jensen, S. T. 2014. Predicting the draft and career success of tight ends in the National Football League. Journal of Quantitative Analysis in Sports, 10 (4): 381-396. doi: 10.1515/jqas-2013-0134.

- "Assumptions of Multiple Linear Regression Analysis." Statistics Solutions. Accessed from: https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/assumptions-of-linear-regression/.

Exercises/Activities

In this module, you will practice the following skills.

- Create a plot to visually analyze the score differential v. wins by team.
- Calculate the score differential that would typically result in an additional win.
- Justify if a linear model is appropriate for the data.
- Create a linear regression model and interpret each coefficient.
- Fit the linear regression model and summarize the results.
- Analyze the residuals and explain why a team might have a large residual.
- Determine how many wins are needed to make the playoffs.

Create a table that summarizes data. Include the variables: score differential, wins, playoffs, and team.

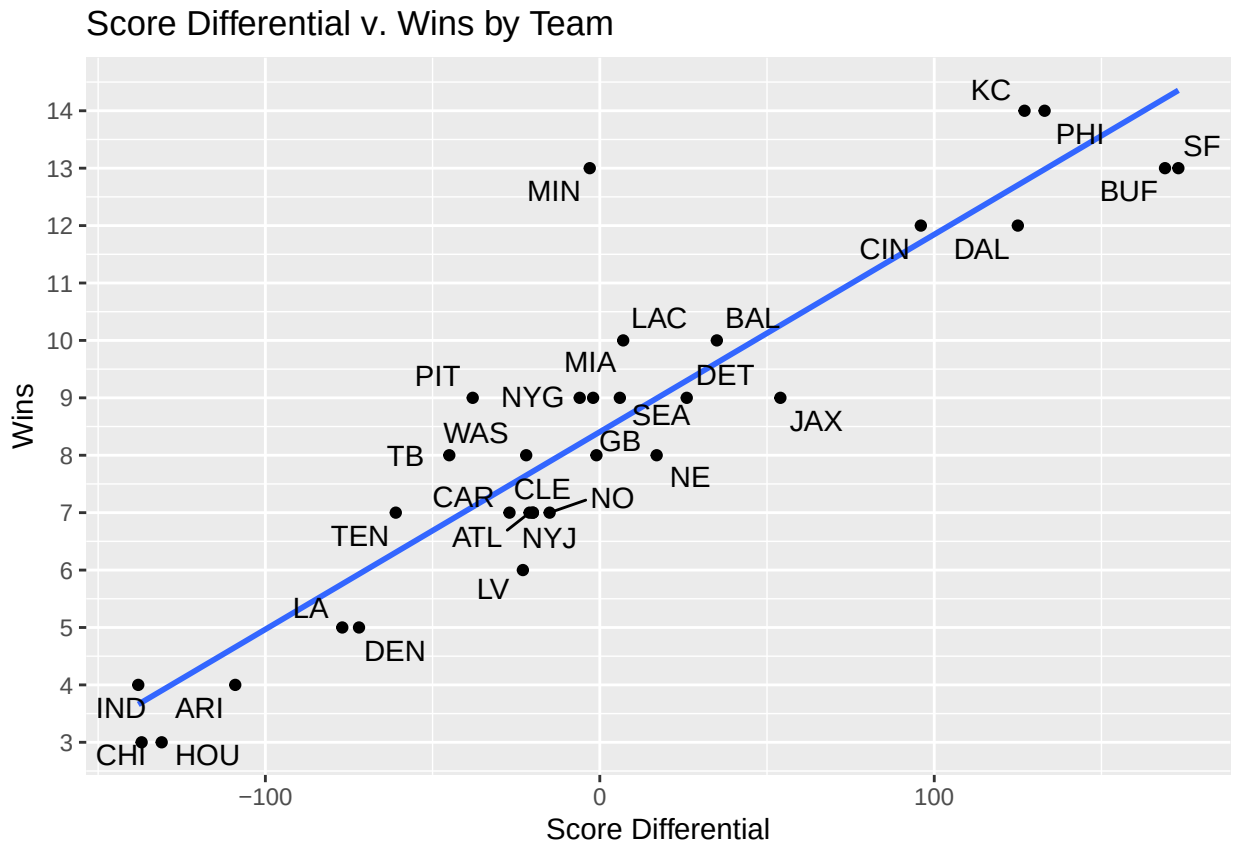
- Score Differential: The number of points scored - the number of points given up
- Wins: Number of games won in the 2022 season
- Team: Name of team
- Playoffs: Indicates if the team made it to the playoffs.

```
#Import the Excel file
nfl2022 <- read_excel("nfl-team-statistics.xlsx")

#Create a data frame
nfl2022summary <- nfl2022 %>%
  select(score_differential, wins, team, playoffs)

#Summarize the data in a table
nfl2022summary %>%
  kable(digits = 3)
```

score_differential	wins	team	playoffs
-109	4	ARI	n
-21	7	ATL	n
35	10	BAL	y
169	13	BUF	y
-27	7	CAR	n
-137	3	CHI	n
96	12	CIN	y
-20	7	CLE	n
125	12	DAL	y
-72	5	DEN	n
26	9	DET	n
-1	8	GB	n
-131	3	HOU	n
-138	4	IND	n
54	9	JAX	y
127	14	KC	y
-77	5	LA	n
7	10	LAC	y
-23	6	LV	n
-2	9	MIA	y
-3	13	MIN	y
17	8	NE	n
-15	7	NO	n
-6	9	NYG	y
-20	7	NYJ	n
133	14	PHI	y
-38	9	PIT	n
6	9	SEA	y
173	13	SF	y
-45	8	TB	y
-61	7	TEN	n
-22	8	WAS	n

Create a plot of the wins vs. score differential for the 2022 NFL season.

```
#Create plot

nfl2022 %>%
  ggplot(aes(x = score_differential, y = wins, label = team)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  geom_text_repel(show.legend = FALSE) +
  labs(title = "Score Differential v. Wins by Team", x = "Score Differential", y = "Wins") +
  scale_y_continuous(breaks = seq(0, max(nfl2022$wins), by = 1))
```



Interpret your plot. How many points are needed for a team to win an extra game?

An extra 30 points in the score differential gives a team 1 extra win. This value was found by calculating the slope of the best fit line.

Linear Regression Models

Regression measures dependence. We will determine if explanatory variables influence a response variable. Linear regression is the most basic and commonly used method of regression. It summarizes the linear relationship between explanatory and response variables. To determine if the model meets the validity conditions for linear regression, we test for linearity, independence, normality, and equal variance.

- Linearity: There must be a linear relationship between the dependent variable and the independent variables. To determine if the model is linear, look at the scatter plot and decide if there is a straight-line relationship.
- Independence: The observations of the data set must be independent of each other. To determine if the data is independent, decide whether knowing the value of one variable in your data set could help you predict or give you any information about the value of another variable. For example, if you know the residual for the Detroit Lions, could you predict the residual of other teams?
- Normality: The residuals (the difference between the observed and predicted values) must be normally distributed. To determine if the model meets the normality condition, look at a histogram of the residuals and decide if it follows a bell-shaped, normal curve.
- Equal Variance: The variance of residuals must be "consistent across all levels of the independent variables" (Statistic Solutions, n.d.). To test for equal variance, look at a scatterplot of residuals vs

predicted values and determine if there is any noticeable pattern. Does the distribution funnel (look like a cone)?

The linear regression equation of our model would be:

$$Wins_i = \beta_0 + \beta_1 SD_i + \epsilon_i \quad \epsilon_i \sim \text{Normal}(0, \sigma^2)$$

where $Wins_i$ and SD_i are the wins and score differential for team i. β_0 is the intercept (the value when Wins = 0). β_1 is the slope (the rate of change in $Wins$ for a unit change in PD). ϵ_i applies the normal distribution structure to the error.

```
# Fit a linear regression model where wins is a function of score differential
wins_lm = lm(wins ~ score_differential, data = nfl2022summary)

# Summarize the linear model
wins_lm %>%
  summary()
```

```
##
## Call:
## lm(formula = wins ~ score_differential, data = nfl2022summary)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6153 -0.7912 -0.4248  0.6698  4.6969
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       8.406250   0.224668   37.42  < 2e-16 ***
## score_differential 0.034390   0.002768   12.42 2.34e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.271 on 30 degrees of freedom
## Multiple R-squared:  0.8372, Adjusted R-squared:  0.8318
## F-statistic: 154.3 on 1 and 30 DF,  p-value: 2.338e-13
```

```
# Print a table of results.
wins_lm %>%
  tidy() %>%
  kable(digits = 2)
```

term	estimate	std.error	statistic	p.value	
(Intercept)		8.41	0.22	37.42	0
score_differential		0.03	0.00	12.42	0

If a team were to have a score differential equal to 0, how many games would they win? Approximately what proportion of all games in a season is this value? Identify which coefficient is being interpreted.

If a team had a score differential equal to 0, they would win 8.41 games which is approximately half of the games in a season. This represents β_0.

How many points result in an extra win for a team? Identify which coefficient is being interpreted.

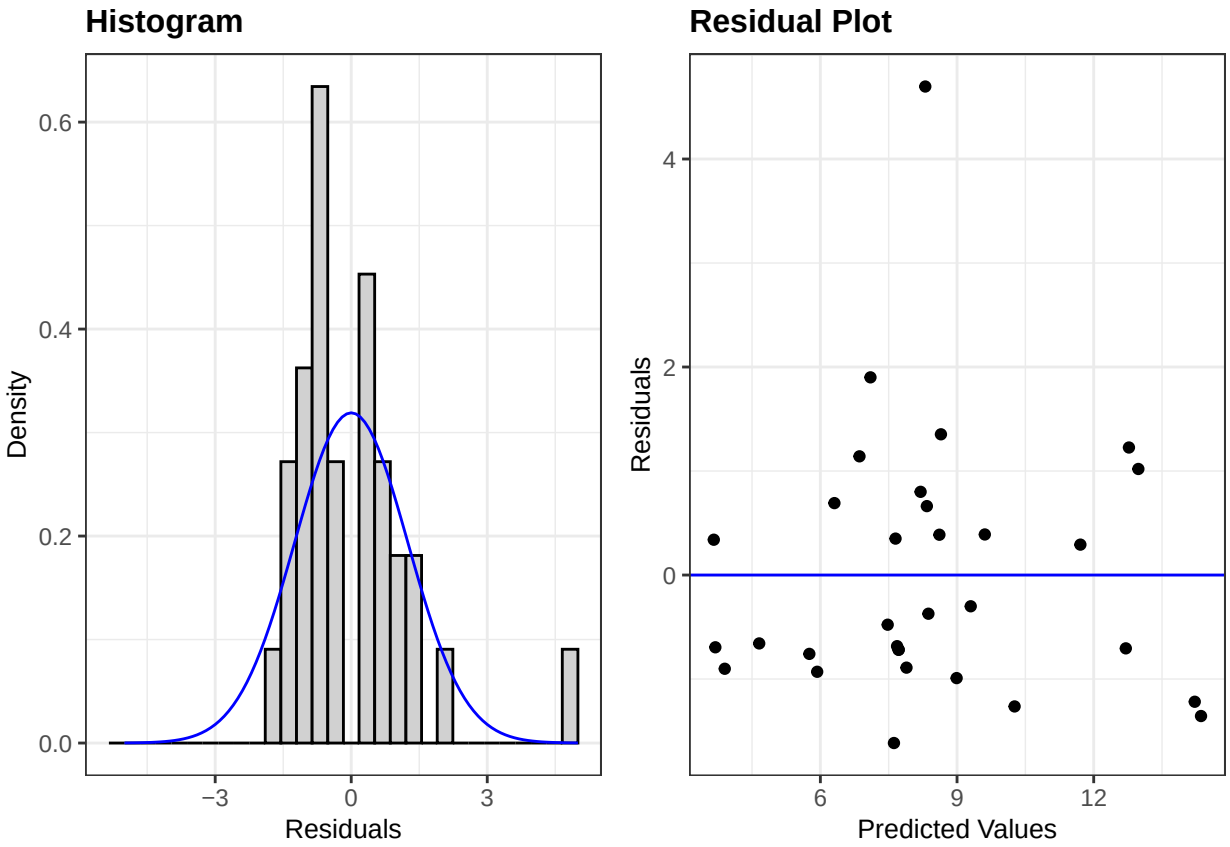
An extra 30 points in the score differential gives a team 1 extra win. This represents β_1.

What is the likelihood that the results for score differential is from chance alone? Make a conclusion about the significance using the standardized statistic and the p-value.

With a very low p-value of 2.338e-13 and a very high standardized staistic of 37.42 for the intercept and 12.42 for the score differential, we can conclude that our results are statistically significant.

Let's look at the normality and equal variance of our data.

```
# Test your validity conditions.

sd_table = nfl2022summary %>%
group_by(team) %>%
summarize(n = n(),
group_sd = sd(wins))
#Validity Condition 3
library(ggResidpanel)
resid_panel(wins_lm, plots = c('hist', 'resid'))
```

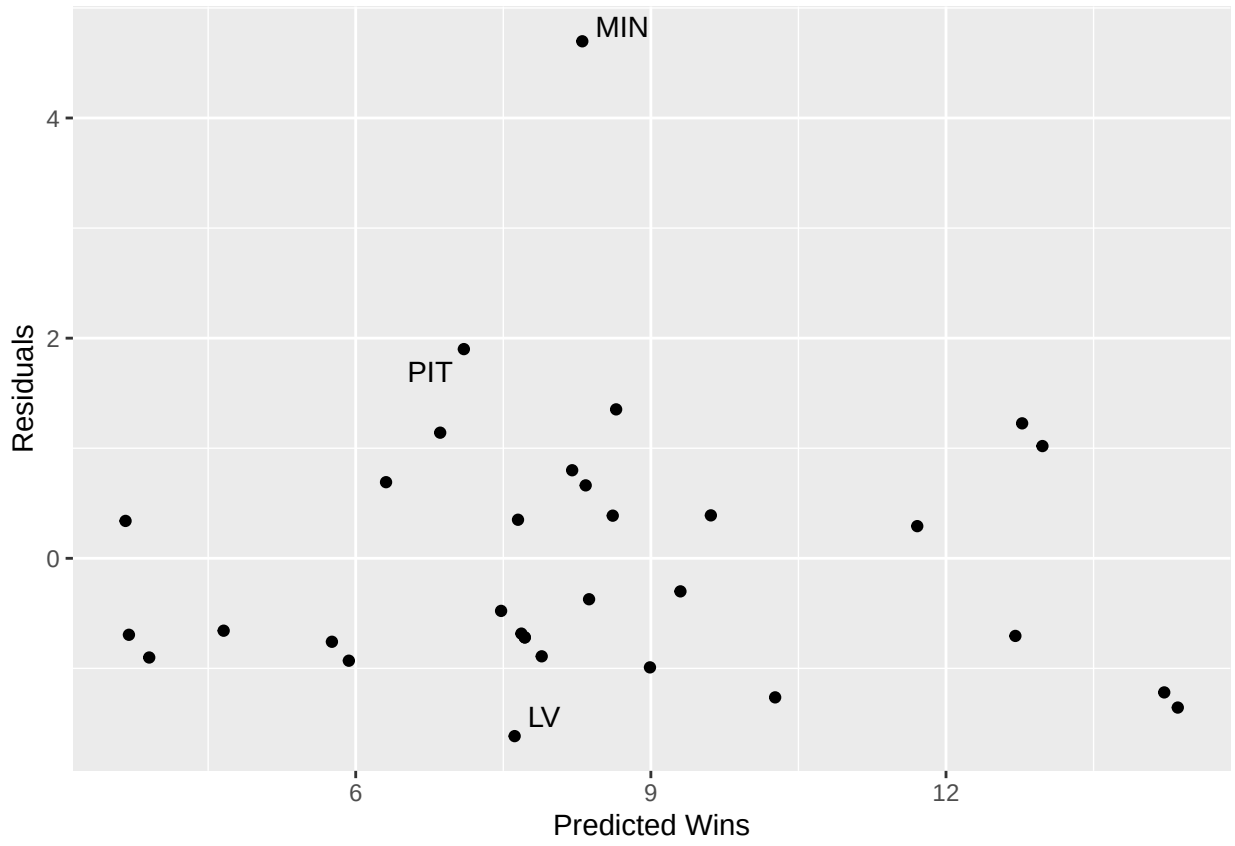


Does it appear our validity conditions are met? Justify your answer.

Based on the plot that displays the wins v. score differential, there appears to be a linear relationship. We conclude the data is independent as we would not be able to predict one team's residual based on another team's residual. Based on the histogram above, our data is fairly normal. We can also apply the normal distribution structure to our linear regression model. The residuals appear to have an equal variance across y=0.

Let's look at the residual for this regression.

```
nfl2022summary = augment(wins_lm, data = nfl2022summary)       #Makes a data frame that adds residuals

nfl2022summary %>%
  ggplot(aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_text_repel(data = filter(nfl2022summary, abs(.resid) > abs(1.5)),
                  aes(label = paste(team))) +
  labs(x = "Predicted Wins",
       y = "Residuals")
```



We define a residual to be the difference between the expected value and the observed value. What units are the residuals in?

The residuals are representative of wins.

The Minnesota Vikings appear to have the largest residual. Explain why they might have a large residual?

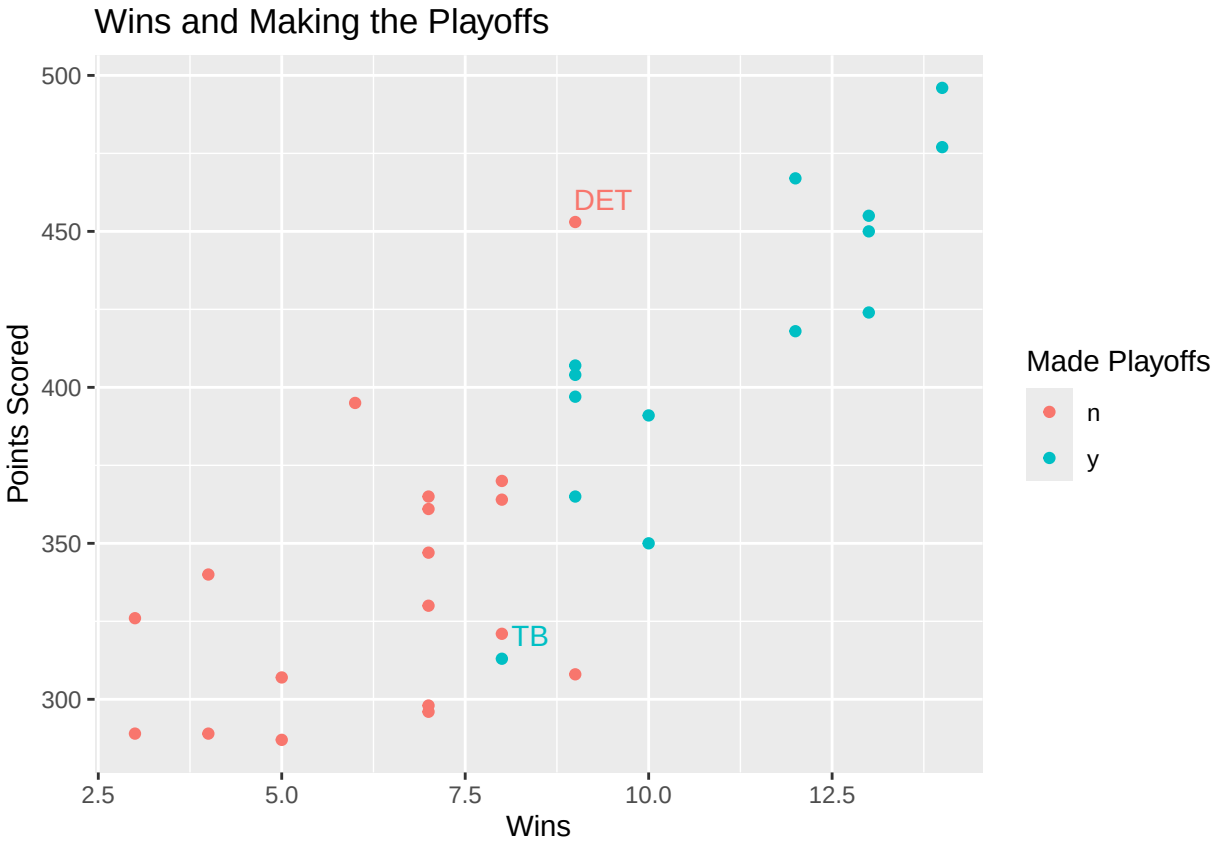
The Minnesota Vikings won more games than expected based on the score differential. There may have been a large number of close games.

The Las Vegas Raiders appear to have the smallest residual. Explain why they might have a small residual?

The Las Vegas Raiders lost more games than expected based on the score differential.

How Many Wins to Make the Playoffs

```
nfl2022 %>%
 select(wins, points_scored, team, playoffs) %>%
 ggplot(aes(x = wins, y = points_scored, color = factor(playoffs))) +
 geom_point() +
 labs(x = "Wins",
      y = "Points Scored",
      color = "Made Playoffs",
      title = "Wins and Making the Playoffs") +
 geom_text_repel(data = filter(nfl2022, team == c("DET","TB")),
                 aes(x = wins, y = points_scored, label = paste(team)),
                 show.legend = FALSE)
```



How many wins should an NFL team strive for to have a good chance of making the playoffs? Discuss why you think Detroit did not make the playoffs.

Based on the 2022 season, NFL teams should expect to make the playoffs with 10 wins, however teams may make the playoffs with 8 or 9 wins. For a team to win an extra game, they would need to score 30 more points throughout the season, and for a team to win two extra games, they would need to score 60 more points throughout the season, and so on. Detriot may not have made the playoffs due to the structure of the NFL. For example, playoff teams are selected based on divisions and conferences. If your division is more competitive, you may have to fight for a wildcard spot rather than being able to easily win a division.

Conclusion

Congrats! You are now equipped with the skills to determine how many points is a win. Your newly obtained linear regression skills will allow you to determine how explanatory variables influence response variable and if a linear regression model is the appropriate model to use. This strategy can be applied many other sports including soccer and lacrosse. Future skills can include predicting the expected wins for a team and simulating a season, or determining what players contribute the most points!

References

"Assumptions of Multiple Linear Regression Analysis." Statistics Solutions. Accessed from: https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/assumptions-of-linear-regression/.

ChatGPT. Assistance given to the author, AI. We used ChatGPT to change the y-axis on the first plot to have continuous integers. This removed the half values to make the plot more clear. OpenAI, (https://chat.openai.com/chat). West Point, NY, 06MAY2024.

Horowitz, M. 2019. Detailed NFL Play-by-Play Data 2009-2018. Kaggle. Accessed from: https://www.kaggle.com/datasets/maxhorowitz/nflplaybyplay2009to2016/data.

Mulholland, J. and Jensen, S. T. 2014. Predicting the draft and career success of tight ends in the National Football League. Journal of Quantitative Analysis in Sports, 10 (4): 381-396. doi:10.1515/jqas-2013-0134.

Pott, Haley CDT A-2 ’24. Prior work used. We used code from CDT Pott’s MA376 Lab 1 to create a histogram and scatterplot of the residuals. West Point, NY, 06MAY2024.

Weisberg, S. 2014. *Applied Linear Regression*. 4th ed. Hoboken, NY: John Wiley & Sons Inc. Accessed from: https://ebookcentral.proquest.com/lib/usma/reader.action?docID=7103845.