NFL*

Yuchen Chen

April 4, 2024

This paper investigates the strategic components that influence victory in NFL games, focusing on the quantifiable impact of offensive yardage. Utilizing comprehensive play-by-play data from the 2023 NFL season, we examine the correlation between various offensive metrics and game outcomes. The study employs linear regression models to analyze how passing and rushing yards contribute to winning games, shedding light on the oft-debated significance of a balanced offense. Our results indicate that not only does offensive yardage have a predictive relationship with winning, but certain types of offensive yardage, like passing yards, are more strongly correlated with success on the field. This analysis, powered by the R programming language and the NFLverse data repository, provides a statistical backbone to the tactical discussions prevalent among football professionals and enthusiasts, potentially influencing future coaching strategies and sports analytics discourse.

Table of contents

*Code and data are available at: https://github.com/Victor1114/NFL.git

1 Introduction

Given the ongoing debate among coaches, commentators, and fans regarding the most effective strategies for securing victories in professional football, this research seeks to provide a data-driven analysis of the impact of offensive yardage on winning NFL games. While conventional wisdom has often underscored the importance of a balanced offense, combining both a robust passing game and a dynamic rushing attack, empirical evidence supporting this notion has largely been anecdotal. This study aims to explore the intricate relationship between various offensive yard measurements—encompassing both passing and rushing yards—and the likelihood of securing a win, thereby contributing a statistical perspective to the tactical discussions surrounding football games.

Utilizing the R programming language(R Core Team (2023)), along with analytical tools such as Tidyverse(Wickham et al. (2019)) and ggplot2 Wickham (2016), this paper analyzes play-by-play data from the 2023 NFL season, compiled from NFLVerse, to systematically examine how offensive performance influences game outcomes. Through statistical methods like logistic regression, this analysis quantifies the relative contributions of passing and rushing yards towards achieving victories, offering new insights into offensive strategy and enhancing our understanding of football analytics. By shedding light on these dynamics, the research not only enriches the ongoing conversation about sports analytics but also aids in the optimization of offensive strategies, providing valuable information for analysts, players, and fans alike.

2 Data

The data used in this paper was taken from nflverse (Carl et al. 2023), which consists of "a set of packages dedicated to data of the National Football League."

2.1 2.1 Dataset

The basis for the final dataset used in this article is play-by-play data from the 2023 NFL season. This includes an incredibly dense amount of information from every snapshot throughout the season, encompassing 372 variables. These include: seconds left in the half, touchdown probability, expected score, probability of a safety being called, kickoff location, play description and more. To filter this information, match-by-match data was cleaned and new variables created.

3 Model

For the "First Model":The intercept is -2.618, suggesting that when all other variables are held at zero, the outcome is negative."completions" has a positive coefficient (1.315), indicating a

	First Model	Second Model
(Intercept)	−2.618	−3.076
	(1.043)	(0.926)
completions	1.315	0.835
	(0.160)	(0.151)
attempts	−0.974	−1.315
	(0.105)	(0.090)
passing_tds	4.914	
	(0.459)	
passing_yards		0.124
		(0.008)
Num.Obs.	318	318
R2	0.489	0.599
R2 Adj.	0.484	0.595
AIC	2182.4	2105.5
BIC	2201.2	2124.3
RMSE	7.37	6.53

positive relationship with the dependent variable."attempts" has a negative coefficient (-0.974), suggesting that more attempts may lead to a decrease in the outcome variable.The model has an R2 of 0.489, meaning it explains approximately 48.9% of the variability in the dependent variable.

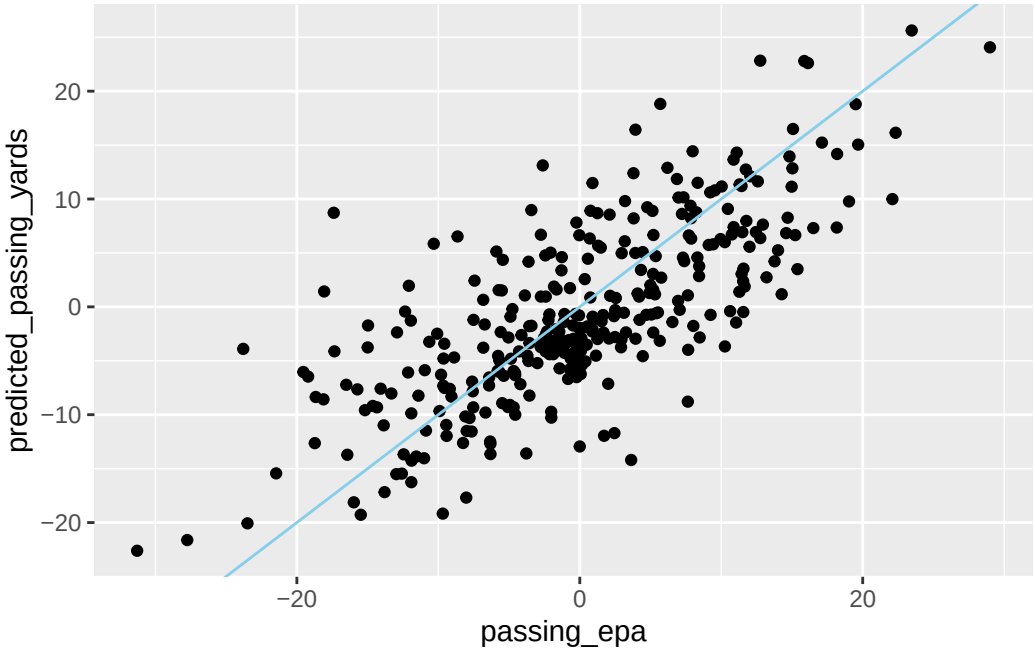
For the "Second Model":The coefficients for "completions" and "attempts" are both smaller in absolute value compared to the "First Model", indicating a less steep slope for these variables.The "passing_tds" has a smaller positive effect compared to the "First Model", but it still maintains a positive relationship."passing_yards" shows a positive coefficient, similar to the "First Model".This model has a higher R2 value of 0.599, indicating it explains approximately 59.9% of the variability in the outcome variable, making it a better fit than the "First Model".

$$y_i|\mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma) \tag{1}$$

$$\mu_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} \tag{2}$$

In the pursuit of understanding the dynamics between offensive plays and their outcomes in NFL games, linear regression emerges as an instrumental statistical tool. The utilization of linear regression is underpinned by its ability to model the relationship between an independent variable, such as passing EPA, and a dependent variable, like predicted passing yards. This technique is particularly favored for its simplicity, interpretability, and efficiency in handling continuous data. It allows us to distill complex relationships into comprehensible models that highlight the strength and direction of associations. In our analysis, linear regression enables us to draw a direct line of inference from the expected points added by passing plays to the actual yardage achieved, which is invaluable for crafting strategies that leverage efficient offensive maneuvers. By offering a quantifiable measure of the impact that passing EPA has on game performance, linear regression not only aids in tactical planning but also provides a foundation for further predictive modeling that could anticipate game outcomes and inform coaching decisions.

4 Results



The regression analysis conducted to examine the predictive power of passing expected points added (EPA) on passing yards generated significant insights. As depicted in the scatter plot, a clear positive trend emerges, indicating a strong correlation between the EPA of passing plays and the predicted passing yards. Each point on the plot represents an observation from the dataset, with the x-axis denoting the passing EPA and the y-axis representing the predicted passing yards from our model.

A closer look at the scatter plot reveals that higher values of passing EPA are generally associated with an increase in predicted passing yards, as indicated by the upward trajectory of the fitted line. This linear relationship is highlighted by the sky blue line, suggesting a unit increase in passing EPA tends to yield a proportional increase in predicted passing yards. This pattern is consistent with the premise that efficient passing plays, which contribute positively to a team's expected points, are likely to result in more substantial yardage gains.

References

Carl, Sebastian, Ben Baldwin, Lee Sharpe, Tan Ho, and John Edwards. 2023. *Nflverse: Easily Install and Load the 'Nflverse'*. https://nflverse.nflverse.com/.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.