Predicting Expected Points Added in Football: A Linear Regression Approach*

Ben Li

April 2, 2024

Table of contents

1 Introduction

Football is an established industry in the United States with teams, fans, franchise merchandise, etc., and a complete economic system. Understanding the dynamics of a team can help a team succeed on the field. One of the important metrics when evaluating individual plays' effectiveness is Expected Points Added (EPA). This score effectively quantifies the impact of individual plays on scoring and can capture information beyond traditional methods. In this study, we predicted passing EPA through a linear regression model utilizing a variety of game-related variables. By focusing on quarterback performance, we analyzed the key factors that influence a team's ability to score.

*Code and data supporting this analysis is available at: https://github.com/UofT-DailinLi/NFL-model.git

2 Data

Our dataset was obtained from Opentoronto and includes 15449 observations and 53 variables (Gelfand 2022). We are only curious about the regional session QB position in the first nine weeks of 2023. Therefore, the number of observations dropped to 335. so the Individual data from games inside the NFL are included, with a focus on quarterbacks (QB). The selected predictors include pass attempts, completions, passing yards, passing touchdowns, interceptions, and sacks. These variables were chosen because they are closely related to passing performance and may have a potential impact on EPA. The dataset was divided into a training set and a test set, randomly assigned in an 80-20 ratio, to develop the model and test its predictive ability.

Data was cleaned and analyzed using the open source statistical programming language R (R Core Team 2022), and additional packages from `tidyverse` (Wickham et al. 2019), `ggplot2` (Wickham 2016), `nflverse` (Carl et al. 2023), `here`(Müller 2020), `nflreadr`(Ho and Carl 2023), `caret`(Kuhn and Max 2008), `lmtest`(Zeileis and Hothorn 2002) and `knitr` (Xie 2014). The cleaned dataset contains 318 observations, and four variables will be shown below.

3 Model

$$\begin{aligned}\text{passing\_epa} = {} & \beta_0 + \beta_1 \times \text{attempts} + \beta_2 \times \text{completions} + \\ & \beta_3 \times \text{passing\_yards} + \beta_4 \times \text{passing\_tds} + \beta_5 \times \text{interceptions} + \\ & \beta_6 \times \text{sacks} + \epsilon\end{aligned}$$

This is the linear model we used. In this equation β_0 = -0.79499 , β_1 = -0.82374, β_2 = 0.45150, β_3 = 0.09936, β_4 = 1.92008, β_5 = -3.37566 , β_6 = -1.91623 .

The scatterplot in Figure 1 illustrates the relationship between the predicted and actual values of epa predicted by the regression model. Each point represents a value in a test dataset. The y-axis represents the predicted value from the model, the x-axis represents the actual value, and this red line represents the perfect prediction line, which means that the predicted and actual values are equal. The distance from these surrounding points to the line is the error. Overall, our model has relatively predictive solid power. However, the error may be likely due to the fact that the selected variables do not fully capture all the variations in the passing-EPA.

4 Discussion

The resulting root mean square error is 4.2637 points, which indicates that, on average, the predicted values of our model do not differ much from the actual values of EPA. This error also

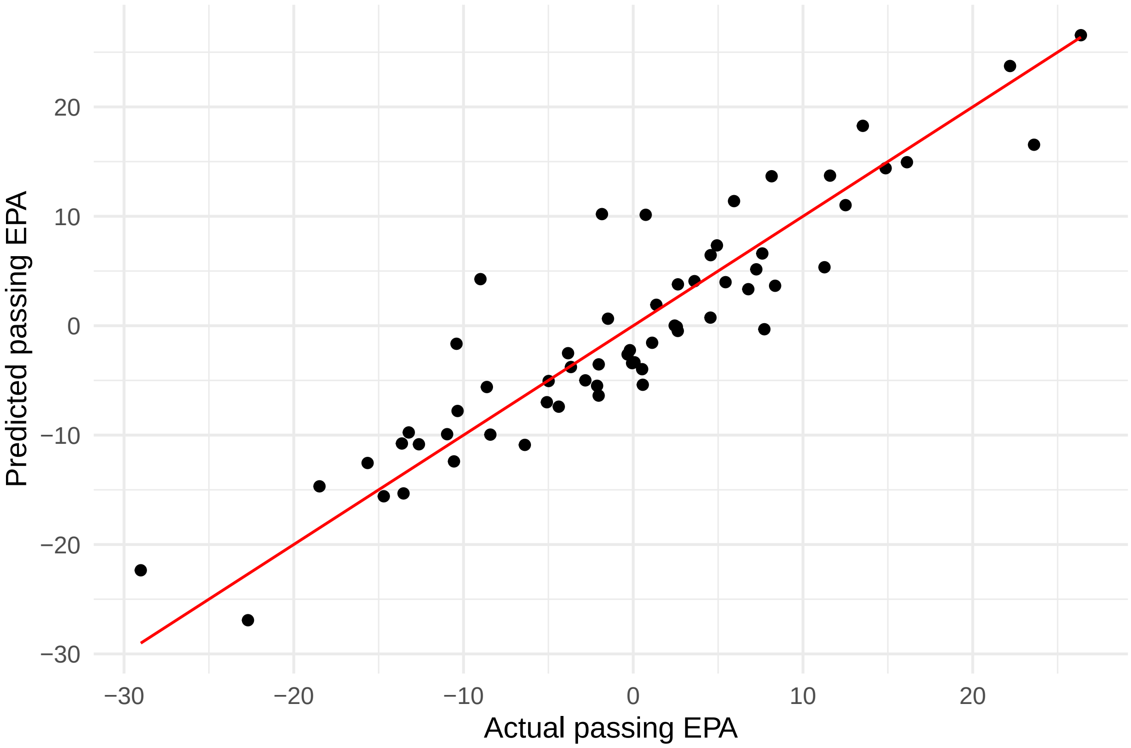

Figure 1: Actual vs. Predicted passing EPA

suggests that the range of EPA should be analyzed using a case-by-case approach in soccer analysis. Moreover, key factors such as pass attempts and touchdowns are expected to be positively correlated with EPA, reflecting successful offensive plays. Conversely, interceptions and sacks may show a negative correlation.

Reference

Carl, Sebastian, Ben Baldwin, Lee Sharpe, Tan Ho, and John Edwards. 2023. *Nflverse: Easily Install and Load the 'Nflverse'*. https://CRAN.R-project.org/package=nflverse.

Gelfand, Sharla. 2022. *Opendatatoronto: Access the City of Toronto Open Data Portal*. https://CRAN.R-project.org/package=opendatatoronto.

Ho, Tan, and Sebastian Carl. 2023. *Nflreadr: Download 'Nflverse' Data*. https://CRAN.R-project.org/package=nflreadr.

Kuhn, and Max. 2008. "Building Predictive Models in r Using the Caret Package." *Journal of Statistical Software* 28 (5): 1–26. https://doi.org/10.18637/jss.v028.i05.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://here.r-lib.org/.

R Core Team. 2022. *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.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.

Zeileis, Achim, and Torsten Hothorn. 2002. "Diagnostic Checking in Regression Relationships." *R News* 2 (3): 7–10. https://CRAN.R-project.org/doc/Rnews/.