Analyzing NFL Team Success Rates Using Regression Models and Panel Data

Sam Srivastava

ECON 4650-001 Principles of Econometrics

Professor Colby Young

December 13, 2024

Introduction

Understanding the cause for game success is a matter of great contention within American football. Particularly, understanding the impact multiple key metrics have on the results of a National Football Association game. Win-loss percentages are a primary indicator of success, while underlying metrics play a significant role as independent variables. Some econometric studies have ascertained NFL win probability using machine learning algorithms such as the random forest method by incorporating factors such as field position, distance, score, and remaining time (Lock & Nettleton, 2014). However, existing research fails to acknowledge broader key metrics and their correlation to the probability of a winning game. The purpose of this research project is to first identify broader performance metrics such as point differentials, yards gained, and turnovers. Then, we will examine performance metrics from a panel data set to quantify their effect on win-loss percentages and provide a rudimentary explanation of what independent factors have the greatest impact on win probability using panel regression. This paper will discuss relevant studies in regard to regression models in sports statistics, highlight key terms and the econometric model utilized in this study, assess the diagnostic tests and results of the model, and provide concluding insights. The primary goal is to use econometric analysis to determine which key performance metrics in the NFL have the greatest and least statistically significant impact on overall performance.

Literature Review

Various methodological approaches have been employed in studies to estimate win probabilities and team performances. The applicability of statistical models in predicting outcomes in the NFL has been a key focus of sports analytics. Researchers utilized random forest machine learning algorithms to estimate win probabilities before each gameplay in an NFL game, considering variables such as field position, current score, first down, yards remaining, and time remaining (Lock & Nettleton, 2014). This method was accurate in predicting late stage game outcomes and offered an outline for gauging specific plays and coaching strategies. A similar study assessed the calibration of win probability models, comparing modified Pro Football Reference (PRF) models with random forest models, which showed slightly better results for the PFR model in regard to overtime plays (Ruscio and Brady, 2021).

Similar machine learning techniques were applied in another study aiming to predict win-loss outcomes and point spreads by analyzing comprehensive play-by-play data from the nfldb database (Owen and Galle). This paper discussed Ordinary Least Squared regression as a baseline method for predicting point spreads in NFL games but questioned its effectiveness. The OLS regression from this study yielded an r-square value of 0.126, which was much lower than the other regression models used (0.157 for Ridge regression and 0.159 for Lasso regression). This indicated issues with its ability to properly mitigate multicollinearity issues. The study ultimately favored the usage of the Lasso regression, however, OLS regressions can have varying applicability in regard to the data or metric being analyzed, as in the case of win probabilities.

All three studies highlight the efficacy of regression models and machine learning algorithms in determining NFL team success. Incorporating both in-game and pre-game variables provides a more dynamic and real time estimation of win probabilities. These papers provide great insight into the importance and applicability of statistical modeling in understanding game outcomes.

Data and Econometric Model

This research utilized a panel dataset from Kaggle featuring 35 key metrics for 32 NFL teams spanning the years 2003 to 2023. This dataset included detailed performance metrics including penalties, turnovers, total points, passing statistics, point differentials, rushing statistics, and other key variables. The panel structure is arranged so that each row represents an NFL team's season performance over a period of several years, accounting for both cross-sectional and time-series factors.

Of the 35 variables, the win-loss percentage was chosen to be the dependent variable on which the independent variables would be contingent. After several regression model iterations and diagnostic testing issues, only 4 key metrics were chosen for comparison alongside an interaction term:

Turnover_PCT	Percentage of Drives Ending in Turnover.
Total_Yards	Offensive Yards Gained.
Points_Diff	Point Differential.
Pass_Net_Yds_Per_Att	Net Yards Gained Per Pass Attempt.

Certain variables were excluded to avoid multicollinearity such as wins and losses, which the win-loss percentage already denoted. Passing yards were omitted as total yards gained and yards per play were represented. Metrics such as the average margin of victory displayed unusually high correlation with other factors as well as omitted variables and were therefore excluded from the regression. Several other redundant and irrelevant variables were omitted from the final regression model to ensure robust reliability.

The following regression model was used:

$$\text{win\_loss\_percit} = \beta 1\text{turnover\_pctit} + \beta 2\text{centered\_total\_yardsit} + \beta 3\text{centered\_points\_diffit} + \beta 4\text{interaction\_termit} + \beta 5\text{pass\_net\_yds\_per\_attit} + \mu i + \text{uit}$$

μi = Team-specific fixed effects.

uit = Idiosyncratic error term.

Results and Discussion

Specification tests:

The primary objective of this analysis was to determine the relationship between win-loss percentage with turnover percentage, offensive gained, point differentials, and net yards gained using the panel data model. Initially, a Breusch-Pagan LM test indicated the appropriateness of other regression models, however, conducting a Hausman test revealed a result of $X^2 = 1.033$ and a p-value of 0.960, exceeding the 0.05 level of significance, suggesting a rejection of the null hypothesis and confirming the applicability of the random effects model over a fixed effects model. Various diagnostic tests were performed to gauge the validity of the model assumptions. These tests included linearity, homoskedasticity, autocorrelation, multicollinearity, normality, model specification, and the appropriateness of using either a fixed or random model.

A Ramsey RESET test was conducted to evaluate the assumption of linearity and specification errors. The results were as follows: RESET = 2.295, p = 0.102. As the p-value is greater than significance level, we reject the null hypothesis and infer that the model does not have any misspecifications. The model was therefore confirmed to be in the correct linear functional form.

The Breusch-Pagan test was utilized to test for potential heteroskedasticity or inconsistent variance across observable points. The results were as follows: X^2 = 0.55, p = 0.815. We fail to reject the null as the p-value exceeds the significance level and we conclude that the error terms have constant variance. Thus the assumption for homoskedasticity is satisfied.

To test for serial correlation of residual data, a Durbin-Watson test was employed, yielding the following values: DW=2.0636, p = 0.7903. As the p-value is greater than the significance level, we fail to reject the null hypothesis and conclude that there is no autocorrelation in the model. Omitting certain independent variables did result in autocorrelation in previous model iterations. This demonstrates the importance of model specification which can greatly affect the model's validity.

The Variance Inflation Factor (VIF) test was used to test for multicollinearity in the model. The following were the results: turnover_pct = 1.343475, centered_total_yards = 3.631908, centered_points_diff = 2.149837, interaction_term = 1.003552, pass_net_yds_per_att = 4.134693. Since all values were below the baseline of 5, we conclude that there were no issues with multicollinearity in the regression model.

As a test of normality of residuals, the Shapiro-Wilk test was employed to ensure that the residuals were normally distributed. The results were as follows: W = 0.99819, p = 0.7099. As the p-value is bigger than the set significance level, the null hypothesis is rejected and we assume the assumption of normality in the residuals is satisfied.

Interpretations:

The main regression results provided a profound understanding of the relationship between the win-loss percentage and other performance metrics. The random effects model was chosen as the primary framework for analysis after conducting the Hausman test. In order to ensure reliable results and combat multicollinearity, the total-yards variable and the points_diff variable were centered for better interpretation. Centering allowed the variables to be evaluated at the mean level of other variables, thus avoiding potential misinterpretation. Furthermore, an interaction term was created by multiplying centered_points_diff and centered_total_yards to account for their joint influence on overall team performance.

The model's R-squared term of 0.831 and adjusted R-squared term of 0.830 suggests that 83.1% of the variability in a teams' win-loss percentage is taken into consideration by the various independent factors. The F-statistic of 3291.707 with a p-value of 0.001 implies that the model is statistically significant, confirming that one of the predictors are substantially related with the dependent variable. The five independent variables together explained a large portion of variance within the dependent variable, win-loss percentage.

The percentage of drives resulting in turnovers in a game, denoted as turnover_pct, was negatively associated with win-loss percentage. A one unit increase in the turnover percentage is linked to a 0.4% decrease in win-loss percentage, ceteris paribus. The coefficient of -0.00405 coupled with a p-value less than 0.001 confirms that the relationship is in fact statistically significant. Similar to the results for turnover, the total offensive yards gained also displayed a negative correlation with the dependent variable, with a coefficient of -0.0000254 and a p-value of 0.01. As the total yards increase by a single unit, the dependent win-loss percentage variable decreases by 0.00254%, ceteris paribus. Though this decrease is marginally small, this relationship suggests that a team accumulating significant yardage without points may have a detrimental effect on their overall performance.

The point differential metric demonstrated a very strong and significant correlation with win-loss percentage. The coefficient for centered_points_diff of 0.00165 and a p-value of 0.001 indicates that for every additional point a team scores relative to the other team, the team's win-loss percentage increases by 0.165%, holding all other variables constant. The net yards gained per pass attempt intriguingly also suggested a positive and statistically significant correlation. As the coefficient for pass_net_yds_per_att was 0.021 and the p-value was less than 0.05, we conclude that all else being equal, a single unit increase in total yards gained on a completion of a football pass is related to a 2.1% increase in win-loss percentage. This emphasized the crucial role of passing efficiency in predicting game success.

Conclusion

This study analyzed a comprehensive panel data and utilized a random effects panel regression model to quantify the effects of important performance statistics in determining an NFL team's success rate by observing the win-loss percentage. The analysis considered data from 20 NFL teams over a 20 year duration, primarily relying on five metrics: turnover percentage, total offensive yards, point differentials, net yards per pass attempt, and an interaction term for avoiding multicollinearity. The interaction term aimed to explain the collective impact of offensive yards gained and point differentials on success rate. Point differentials and total yards were centered to ensure a better interpretation of the main effects. The Hausman test was conducted to confirm the random effects approach to the model. The regression model satisfied all main CLRM assumptions and passed all key diagnostic tests.

When examining sports analytics, often times our intuitive understanding of what factors influence success can be challenged. In the case of the five predictors observed in this study, point differential had the most profound and statistically significant correlation to the dependent variable in relation to the other variables. Turnover percentage conversely had a negative overall impact on win-loss percentage, as a higher turnover rate resulted in a lower win-loss percentage. Total yards gained displayed a smaller but significant negative correlation to the dependent variable, whereas net yards gained per pass attempt had a positive relationship with the dependent variable. This analysis provides an insightful understanding of the significance of factors in relation to overall team success rate in the NFL. The research further corroborates the importance of econometric analysis in sports analytics in evaluating key statistics and overall performance.

Appendix

Regression Output

```
===============================================
                  Dependent variable:
               ---------------------------
                        win
-----------------------------------------------
turnover_pct             -0.004***
                         (0.001)
centered_total_yards     -0.00003***
                         (0.00001)
centered_points_diff     0.002***
                         (0.00004)
interaction_term         -0.00000
                         (0.00000)
pass_net_yds_per_att     0.021***
                         (0.008)
Constant                 0.424***
                         (0.050)
-----------------------------------------------
Observations             672
R2                       0.831
Adjusted R2              0.830
F Statistic              3,291.707***
===============================================
Note:           *p<0.1; **p<0.05; ***p<0.01
```

References

Lock, D., & Nettleton, D. (2014). Using random forests to estimate win probability before each play of an NFL game. *Journal of Quantitative Analysis in Sports, 10*(2), 197–205.

Owen, Z., & Galle, V. (2014). **Predicting the NFL**. Retrieved from https://vgalle.github.io/files/NFLReport.pdf

Ruscio, J., & Brady, K. (2021). **Estimating win probability for NFL games**. The College of New Jersey. Draft available from https://ruscio.pages.tcnj.edu/files/2021/01/NFL-Win-Probability.pdf