Multiple Linear Regression on 1976 NFL Data

Hunter Worssam
August 2nd, 2025
Johns Hopkins University

Abstract
This project applies multiple linear regression modeling to historical National Football League (NFL) 1976 team performance data to identify which team performance metrics are most predictive of win totals. The dataset includes offensive, defensive, and special teams statistics for all 28 teams, as well as opponent statistics and turnover differential. After standardizing predictors to enable comparability, a null model using all available predictors was constructed as a baseline. Stepwise regression with entry and removal p-value thresholds of 0.20 and 0.30, respectively, was used for feature selection, yielding a final model with three statistically significant predictors. Model diagnostics, including residual plots, Q-Q plots, leverage plots, Cook's Distance, and externally studentized residuals, confirmed that the assumptions of ordinary least squares regression were reasonably met. Bootstrap sampling demonstrated stability in coefficient estimates and model performance metrics, while the PRESS statistic suggested strong generalizability with minimal overfitting. The final model predicts team win totals within approximately 2 wins on new data, offering a robust and interpretable framework for understanding performance drivers in the 1976 NFL season.

Introduction
In team sports, quantifying the factors that drive winning outcomes has long been of interest to coaches, analysts, and fans. In professional football, where the margin between success and failure can be narrow, understanding which aspects of team performance most strongly influence win totals can inform both strategic and roster decisions. This project investigates the relationship between team-level performance metrics and season win totals using data from the 1976 National Football League (NFL) season. The primary objective is to develop a multiple linear regression model that identifies the most predictive variables for win totals while ensuring the model is statistically valid, interpretable, and generalizable.

Dataset
The chosen dataset for this project is the National Football League 1976 Team Performance table found on page 574 of the class text. It is also provided on **pages 12-13 of the code file.** In this dataset we are given data from all 28 teams in the league including their win total, offensive, defensive and special teams statistics, opponent statistics and turnover differential.

Methods
The following analysis **assumes** that the ordinary least squares regression model satisfies the following conditions. The relationship between each predictor and the response variable is linear in parameters. Observations are independent of one another, with no autocorrelation in residuals. The variance of residuals is constant across all levels of the fitted values. Model residuals are approximately normally distributed. The validity of these assumptions are routinely checked and confirmed throughout the model diagnostic and validation processes.

To begin, the dataset above was examined for any **missing values and obvious outliers**. It was noted that the range of values varies from thousands of yards to

percentages to differentials, however no categorical values were present that required encoding. However, due to the variance in scale among variables, particularly between the total yards variables x_1 & x_2 and the percentage variables x_4 and x_7, all predictors will be standardized prior to model building.

OLS can easily handle unscaled predictors, however not scaling these parameters could impact the interpretation and comparison of the magnitude of coefficients in the final model. A **standard scaler** is used here to remove the mean from each predictor and scale them to unit variance. This is an attractive option for our dataset given the uneven scales (percentages, differentials and large continuous variables) and the ability to observe how unit changes in a predictor's standard deviation impacts win totals. The response variable y will remain unscaled here to maintain our desired output of wins. The code and output of the standardization procedure can be found on **page 2** of the code file.

With the predictors now scaled, a **null model** is constructed incorporating all of our predictor variables. This model will serve as a baseline for which we can measure improvement on in terms of the adjusted R-squared and coefficient p-values. The code and results for the null model specification are found on **page 3** of the code file. Refer to the results section for more specific null model analysis.

Prior to feature selection, **variance inflation factors** were calculated for each of the predictor variables to gauge for multicollinearity. The code and results for this check are on **page 3** of the code file. Overall, variables x_1, x_7 and x_8 had moderate VIF around 4-5, so multicollinearity with respect to these variables will be kept in mind should they enter our final model.

The **feature selection** method of choice is stepwise regression. The selected variable list is left empty to begin with. Of the remaining available predictors, the one with the most statistically significant p value less than the defined threshold is added to the model. Next, the variables in the model are assessed for their p-values, and any that fall below the removal threshold are removed. The addition and removal steps are then repeated until no new variables qualify for addition into the model, and all selected variables do not qualify for removal from the model. Finally, a list of selected variables is returned. Code for this process is found on **page 4** of the code file. Thresholds of p = 0.20 and p = 0.30 were chosen for variable entry and exit respectively, and the final variables chosen were x_2, x_7 and x_8.

With three statistically significant variables selected, the **subset model** was constructed and examined in the context of our null model performance. The derivation and results of the subset model are found on **page 5** of the code file. More details of the subset model are found in the results section.

Regardless of how strong the subset model looks, it's important to consider the **model diagnostics** to ensure our underlying OLS assumptions are holding up. As a first check, the **residuals vs fitted values plot** was created to assess for any unusual patterns in variance. This code and plot is found on **page 5** of the code file. The red line that I included is the LOWESS smoother which is useful in condensing the residual trend into

something more digestible. Interpretation of the residuals plot is found in the results section.

The **normal Q-Q plot** found on **page 6** of the code file is a diagnostic plot for evaluating whether the model residuals are approximately normally distributed, which is one of our key underlying OLS assumptions. The x-axis provides the theoretical quantiles, in other words what our residuals should look like if they were perfectly normal. The y-axis provides what the model residuals actually look like. See the results section for interpretation of this plot.

A **standardized residuals vs leverage plot** was built using the code found on **page 6** of the code file. This plot is an important diagnostic for determining influential observations and overall model stability. The x-axis quantifies how impactful a datapoint is in terms of the overall model coefficients. The y-axis simply shows how far off the model prediction was in standard deviations. Again we leverage the LOWESS smoother to visualize the overall pattern, and analysis of this plot is provided in the results section.

The **Cook's Distance plot** found on **page 7** of the code file is another method of assessing influence of individual observations within the subset model. The x-axis indicates the observation index, or row number of the dataset, and the y-axis provides the calculated Cook's Distance. As a rule of thumb, the threshold for identifying influential points is a distance of 4/n, or in this case 0.143. See the results section for more details on the Cook's Distance plot.

As a final diagnostic check, the **externally studentized residuals plot** was created and studied, with the code and plot available on **page 7** of the code file. The externally studentized residual plot builds on our existing residuals plot by excluding the observation corresponding to the residual in the model fitting process. By doing this, we avoid artificially influencing our own error estimate, which makes externally studentized residuals more reliable for identifying genuine outliers. The two dashed lines represent 2 standard deviations, or observations that would be significant outliers at an alpha level of 0.95. The results section details the findings of this plot.

In the **model validation** stage, two methods were deployed to assess the stability and predictive ability of the subset model: bootstrap sampling and the PRESS statistic. **Bootstrap sampling** enables us to build a distribution of possible coefficient values for our model, expanding our understanding of the variability we can expect while establishing confidence in the model's stability. The code and results for bootstrap sampling are found on **pages 9 & 10** in the code file.

For the **bootstrap procedure**, we randomly resample our 28 observation dataset n times (1000 in this case) to simulate what would happen to the dataset if it had been drawn differently from the same population. This yields n groups of 28 samples. From there, we fit a linear model in the same form as our subset model for each of the n bootstrap samples, and we calculate the r-squared and coefficient estimates for each. We can then re-calculate 95% confidence intervals and view distributions of each model parameter across the n models to gauge the stability and generalizability of our model. Details from the bootstrap sampling distributions and confidence intervals can be found in the results section.

To further evaluate the model's predictive capability we can calculate and examine the **PRESS statistic**. This statistic is the prediction sum of squares, meaning it analytically simulates leave-one-out cross-validation for linear regression. This gives us a total squared error that we'd get if we had fit the model excluding each observation one at a timer and predicted it, effectively simulating how the model will behave on new, unseen data. Details from the PRESS statistic calculation can be found in the results section, and the code for the derivation can be found on **page 11** of the code file.

Results

Our **null model** yields an R-squared of 0.815 and an Adjusted R-squared of 0.723. We're explaining a large portion of the variance in our model, but the drop in adjusted R-squared suggests we may have some redundant or unimportant variables still in our model. The F-Statistic of 8.839 and corresponding p-value of 5.33e-05 suggest our null model is significant, so this will be our baseline as we develop our subset model.

Our **subset model** has an R-squared value of 0.786, slightly down from the null model's 0.815. However, the adjusted R-squared is 0.759, an increase from the 0.723 in the null model. We've increased the predictive power of our model while removing 6 variables, an important boost in model simplicity. Additionally, all three variables are statistically significant at the alpha = 0.05 level. Predictors x2 and x7 are positive effects, and both do not have zero in their confidence intervals meaning we can be confident in their signs. The same is true for x8, although it is clear Opponent rushing yards is a negative effects variable. AIC of our subset model is 113.1, a decrease from our null model's 121.0 value. Our F statistic and corresponding model p-value are 29.35 and 3.36e-08 respectively, indicating strong statistical significance.

Shifting gears to our model diagnostics, the **raw residuals plot** indicates the residuals are reasonably centered around zero with no strong or obvious skew in distribution. There is also no evidence of fanning, suggesting that we are not in the presence of heteroskedasticity. The LOWESS line is slightly wavy which suggests there may be some nonlinearity present around fitted values 5 - 8. There also appear to be some large residuals around +/- 3 in magnitude, however these points do not flag in our subsequent Cook’s Distance plot.

In the **normal Q-Q plot** can see that the vast majority of the residuals do fall slightly below or slightly above this ideal red line down the center, which supports the validity of our statistical inference capabilities. There is a slight upward bend in the right tail of the residual plot, corresponding to a few large residual values that we already noted in our residuals graph, but overall the normality assumption does appear to hold.

For the **standardized residuals vs leverage plot** the most dangerous combination to look for are points that have both high residual values and high leverage, which would indicate a single datapoint is controlling a large portion of the model definition. We do not observe that here, however there are a couple observations with high leverage > 0.30 that should be watched closely. Overall, most standardized residuals fall between +/- 2, and the majority of leverage points are below 0.25. While there is not cause for concern, the top three leverage points do appear to be pulling the LOWESS smoother downward. On **page 8** of the code file, the leverage points > 0.30 were identified as the

Washington and New York Giants observations. These points did not end up flagging in the subsequent Cook's Distance plot, so they are retained in the model.

In the **Cook's Distance plot,** none of the observations exceed the threshold of 0.143, so we can conclude that no single observation is disproportionately influencing the regression model. There are three or so points that contribute significantly, but in the context of our residual and leverage plots, we can conclude that our model is likely robust to individual data points.

The only observation lying outside of the +/- 2 SD range in the **externally studentized residuals plot** is the Washington observation, which was also flagged in the standard residual and leverage plots. Because this point did not also arise Cook's Distance plot, it will simply be flagged and left in the model.

Compared to our subset model performance, the **bootstrap estimates** for our three coefficients and r-squared value are incredibly similar, suggesting our model is stable in the presence of new data subsets from the same population. The bootstrap r-squared mean is 0.797, very similar to the subset model r-squared of 0.786. The subset model coefficients of 6.96, 1.77, 1.06 and -1.71 (for the intercept, x2, x7 and x8 respectively), are all very similar to our bootstrap coefficient estimates of 6.97, 1.78, 1.05 and -1.69 respectively. We can also see that the bootstrap sampling has very similar 95% confidence intervals for all coefficients, albeit the bootstrap sample ranges are slightly narrowed due to the confidence gained from having more N in the distribution. Overall, the bootstrap sampling process has given us the added insight of what coefficient estimates we could expect across the full 95% confidence interval, and it also has established confidence that our model will be able to withstand different data samples from our population of NFL teams.

Regarding the **PRESS statistic**, our training data RMSE is 1.58, meaning on average the model misses actual win totals by 1.58 wins on the data it trained on. The PRESS RMSE of 1.77 suggests that on average, if we apply the model to a new NFL team not in the training set, the prediction would likely miss by ~1.77 wins. The small increase between both the training RMSE and PRESS RMSE suggests that our model is not overfitting, and than it is stable and generalizable. Since wins are whole numbers, both models will likely have the same ~2 win error, so it's predictability on new data is robust.

Conclusion

Through careful feature selection, diagnostic evaluation, and validation, this analysis identified three key performance metrics from the 1976 NFL season that reliably predict team win totals. These variables were x_2: Passing yards (season), x_7: Percent rushing (rushing plays / total plays) and x_8: Opponents' rushing yards (season). All three variables were individually statistically significant, as was the overall model. The final model, $y = 6.9643 + 1.7716x_2 + 1.0569x_7 - 1.706x_8$, was also stable under resampling, and generalizes well to new data, with an expected prediction error of roughly two wins. While specific to the 1976 season, the approach demonstrates a transferable process for building interpretable and predictive sports performance models.