Running the Numbers: A Data-Driven Exploration of Utilizing Linear Regression to Predict Forty Yard Dash Times in the NFL Combine

Hannah Connell

STATS 401: Applied Statistical Methods II

December 12, 2023

Background

Each year, hundreds of college football players declare their intent to enter the NFL draft, a process where professional teams select individual players based on their college career performances, as well as team needs. Prior to this though, a subset of those players are invited to participate in the NFL Scouting Combine, which allows them to show off their individual skills and be directly evaluated by scouts on physical, mental, and medical criteria. The data set this analysis is performed on contains data from the physical portion of the NFL Scouting Combine (hereby referred to as "the Combine") on select defensive positions from the 2015 season.

The variables included in this analysis were forty-yard dash times ("fortyyd"), player height in inches ("heightinchestotal"), player weights in pounds ("weight"), vertical jump heights in inches ("vertical"), twenty yard shuttle times in seconds ("twentyss"), and position, which was either CB, DE, DT, or OLB ("position"). Going forward, each of these variables will be referred to by their name as it appears in R. The overall goal of this analysis was to leverage each of these predictors in order to create a linear model that would predict a player's forty yard dash time.

Analysis

Initial Model

The initial model utilized fortyyd as the response, with heightinchestotal, weight, position, and vertical as the predictors. This model yielded a multiple R-squared value of 0.8011, and an adjusted R-squared value of 0.7854 (Figure 1). As evidenced by the diagnostic plots, the assumptions of linearity and constant variance appeared to hold (Figure 2), as did the assumption of normality (Figure 3), implying that the data set did not appear to have any OLS violations.

Model Change #1: Including an Interaction Term

When exploring the data, it was found that weight seemed to vary largely by position (Figure 4), with CB's appearing to have, on average, the lowest weights, and DT's having the highest weights. From there, a decision was made to attempt to include an interaction between weight and position in an updated version of the model (Figure 5). As evidenced by the model however, there appeared to be no statistically significant difference in slopes between the weights of the reference category (CB) and the weights of the other three positions. The slopes themselves had very little difference from one another as well, which indicated very little practical significance. The overall impact on the model from including this interaction variable was relatively small, but negative, with the adjusted R-squared dropping to 0.7791, a decrease from the initial model. While I was hopeful that the differences in weight by position evidenced by Figure 4 would provide a statistically significant impact on the model, it appeared that would not be the case, and thus I removed the interaction term.

Model Change #2: Addition of a New Predictor

At this point in the analysis, it was decided that the exploration of additional predictors could be beneficial to the strength of the model. From there, twentyss was added to the model. The scatterplot matrix indicated a strong, positive linear relationship between this new addition and fortyyd (Figure 6). When the linear model was updated to include twentyss, the adjusted R-squared value increased to 0.7928 (Figure 7), implying a better fit than both the initial model and the previous model that included an interaction term.

Model Change #3: Revisiting an Interaction Term

With the addition of twentyss, I decided to revisit the idea of including position as an interaction term. When comparing twentyss and fortyyd by position (Figure 8), there was less clear distinction between positions than when comparing weight and fortyyd, but still enough

that I decided to test the interaction in the model (Figure 9). To my surprise, the adjusted R-squared value increased to 0.7944, implying that this model may be stronger than the previously tested ones. The differences in slope were considerable as well, with each position having a very practically significant slope. However, only twentyss:positionDT appeared to have a statistically significant difference in slope between itself and the CB position. Despite this, I kept this interaction in the model due to its impact on the adjusted R-squared value, and the practical significance of each of the new slopes.

Final Model

The final model contained fortyyd as the response variable, with heightinchestotal, weight, vertical, twentyss, position, and an interaction term between twentyss and position as the response variables (Figure 10).

Final Assessments of OLS Assumptions

While the assumption of linearity was met, the assumption of constant variance appeared to be (debatably) violated when compared to the initial model (Figure 11). The assumption of normality continued to be met (Figure 12). While subject to interpretation, I assessed that every assumption of OLS was reasonably met, except for constant variance. There was also potential for a violation of independence if a player were to participate in the Combine one year, go undrafted and unsigned for a whole year, then re-participate in the Combine the following year, as their data in year one would not be independent of their data in year 2. However, due to the fact that this model only accounted for data in the 2015 Combine, dependence in the form of the previously mentioned scenario was very unlikely to be an issue unless the data set was expanded to include multiple years.

Assessing Model Fit

The final R-squared value for the model was 0.8195, meaning that 81.95% of the variability in forty yard dash times could be explained by the linear relationship with predictors that were used. The RMSE for this model was 0.0975 seconds, meaning that the model had an average error of approximately that much. In the context of the forty yard dash times however, an error of 0.0975 seconds could be considerable, given that the average time for the data set was 4.698 seconds, whereas the max was 5.250 seconds and the minimum time was 4.310 seconds (Figure 13).

Interpretation of Significance

Given the F-Test statistic of 32.69 and the reported p-value of < 2.2e-16 for the model, this very strongly suggested that the overall model explained a significant amount of variation in forty yard dash times (Figure 9). The most statistically significant, and thus strong, linear predictor of forty yard dash times was vertical jump heights (Pr(>|t| = 0.00532). For the interaction terms, there was a statistically significant interaction effect between CB and DT (Pr(>|t| = 0.09868). While statistically significant predictors were scarce, there were several practically significant ones. For example, positionDT yielded a slope of -1.2745688, implying that each DT would have a decreased forty yard dash time of approximately 1.27 seconds. Which given the spread of these times (Figure 13), is an incredibly significant drop. Other notably practically significant slopes were twentyss:positionDT (0.3477345) and twentyss:positionOLB (0.2158058). These slopes implied that for every 1-second increase in twenty yard shuttle times, the forty yard dash time would increase by approximately 0.35 and 0.22 seconds respectively, holding all other variables constant.

Other Considerations: Multicollinearity

When testing for Variable Inflation Factors (VIFs) on each individual predictor, besides the categorical variable position, no evidence of multicollinearity was found (Figure 14). While the initial scatterplot matrix had me concerned about potential multicollinearity between weight and height, as well as weight and twenty yard shuttle time (Figure 6), the calculated VIF for weight was 4.030159, thereby less than the threshold of a VIF of 5 for suspecting multicollinearity.

Conclusion

Ultimately, the final model was able to account for approximately 81.95% of variability in NFL Combine forty yard dash times, with vertical jump heights appearing to be the strongest linear predictor. From the initial model, the adjusted R-squared value increased from 0.7854 to 0.7944, giving confidence that the final model is a better predictor of variation in forty yard dash times. While a large amount of the variability was captured by this model, the RMSE was rather high given the context of how much small increments can matter to forty yard dash times. Future iterations of this model would likely include the addition of offensive positions, the addition of multiple years worth of Combine data, and further exploration of other potential predictors for forty yard dash times.