Curtis Schrack & Mason Seftas

Dr. Borselli

MA 242 Statistics II

5/2/2022

Project Report

Today with the legalization of sports gambling, it has become one of the biggest new ways to gamble. Many sports fans want to try their luck and take home a big payday with oddsmakers creating enticing lines that pull in even the casual fan. One of the most enticing bets football fans like to get involved in is NFL MVP winners. Although choosing the right player could be one of the most challenging predictions with over a thousand candidates every year. In this report, we will show how a fan can accurately predict the next MVP of the league by using statistical modeling techniques.

We started the process of building our models by creating a dataset from statistics found on Pro Football Reference. We collected statistics over the last ten years, going back to 2011. We restricted are collection to players that received votes for MVP, Offensive Player of the Year, and Defensive Player of the Year awards. This came out to 103 observations used in producing our models. As we are looking to predict the number of votes a player will receive, we made our response variable number of votes. We included the common stats collected throughout the season, such as passing yards, rushing yards, and touchdowns, as the predictor variables.

From these variables, we created two models using multiple linear regression. The difference between these two models was how we determined the polynomial relationship between the predictor variables and the response variable. In the first model, we used K-Fold cross-validation. In the second, we examined scatterplots and box plots to try and visually determine which polynomial degree to assign to each predictor variable. Using backward elimination, we narrowed down our variables to only include significant predictors in our model. A significant predictor is anything that has a p-value of less than 0.05. As you can see in figure 1 and figure 2, we found significant predictors such as passing completions, passing touchdowns, and total touchdowns to help determine how many votes an NFL player will receive for MVP voting. Using these predictors, we built algebraic equations that can be see in figure 3 and figure 4 for each model. From these equations, we will be able to predict future winners of the award.

Fig 1: Model using K-fold validation

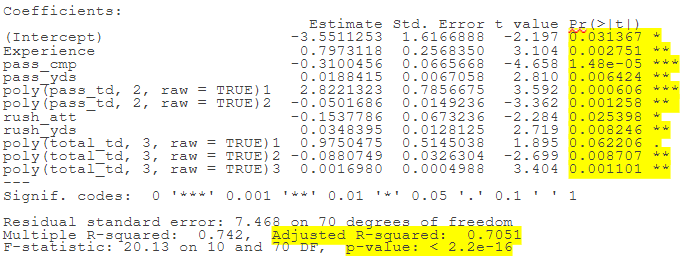
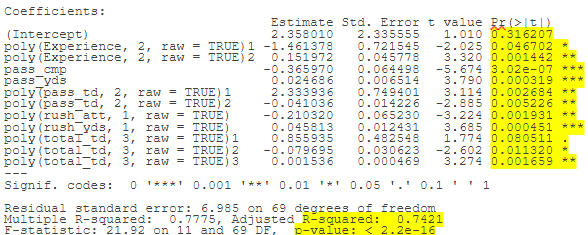


Fig 2: Model using visual observations

To determine the validity of our models based on unknown data, we ran testing statistics such as RMSE. RMSE stands for root mean square error, which finds the difference between the predicted values of the model and the actual values, finds the average, then takes the square root. A good test RMSE is as close to zero as possible. Our RMSE came to 12.25 on the 2021 MVP race results for the K-Fold model. For our created model, we obtained an RMSE of 9.89, which is slightly better in terms of the year 2021. Figure 5 represents the players of the 2021 race and how many votes they received. Their predicted number of votes is next to their actual votes based on the two models. Alarming results would be the different models predicting different winners in the race. The cross-validation model had the ending results giving Tom Brady the most votes with 32. Our created model gave the win to Aaron Rodgers with 14 votes. Another observation is that Aaron Rodgers received 39 votes in the actual race, which is much higher than the models predicted. From our understanding, this is due to the number of players who received zero votes in the race. Since many players received zero votes, the model does not escalate as quickly as the actual results. As this may seem troublesome on the surface, our initial question is not asking how many votes each player receives. Since we ask who wins the race, we are only concerned with who gets the most votes against their competition. This comparison is what matters in sports betting. Our model will be tested year-to-year on new player statistics to further the two models' accuracy.

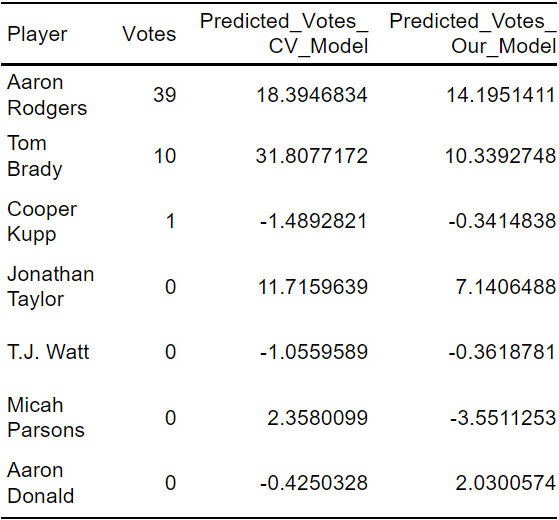


Fig 5: Results of 2021 MVP voting

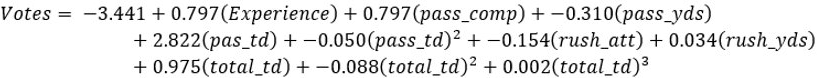


Fig 4: Equation from Visual Observations

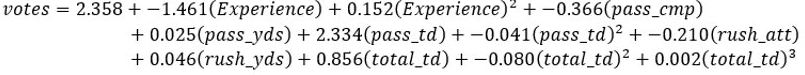


Fig 3: Equation from K-fold Cross Validation

When it comes to statistical modeling, using our multiple linear regression model will give all who are interested in the NFL MVP race an edge over the competition. From our results, we learned passing yards and total touchdowns positively impact the number of votes, whereas having too many completions can hinder your chances since it is negatively correlated. Updating the data set with more players and statistics will only help our model in the future. In addition, more years of modeling will better test the validity of our model each season the race is conducted. Nonetheless, our established model is an excellent start for years to come.