score problems IPR

2024-04-04

Figure 1: mvp_pic

1 - Learning Goals

1. Understand the mechanics of statistical analysis through sports
2. Create and Understand a Generalized Linear Model
3. Be able to explain this project to someone unfamiliar with statistics
4. Understand how important touchdowns are to building an MVP case

2 - Introduction

What makes an NFL MVP? Is it strictly a QB award? Through this SCORE module we want to explore the main factors that determine how NFL MVP awards are won. By looking at individual stats such touchdowns, total yards, and turnovers, we will develop a model to identify the stats most important to winning the most prestigious award in the sport. This model may not be implemented by the MVP voting committee in future seasons, but it will give football fans an understanding of how MVPs are determined.

3 - Data

We found a data set from NFLverse that has play-by-play data dating back to 1999. It was found at this link: https://nflverse.nflverse.com/. We are going to group together by player, by season, and attach a variable about MVP winners. After the compilation of the data set we have the variables passing yards, passing touchdowns, rush yards, rushing touchdowns, and a turnover variable that combines interceptions and fumbles.

We also created a data frame with each season's MVP winner and attached it to the NFL verse data set through a left join.

Problem 1: Find individual QB statistics for 2018. Create a table of the top 15 arranged in descending order of Passing Yards.

4 - Methods / Instructional Content

These two links are both integral to understanding our project. The first one is an explanation of logistic regression: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3936971/. The second link is to a journal article from UC Berkeley that analyzes the difference in what NFL MVP awardees win the award for and what should really matter to voters. https://sportsanalytics.berkeley.edu/articles/nfl-mvp.html.

5 - Exercises and Activities

The following code is a problem set-esque step-by-step breakdown of how we got our results and how we would expect someone attempting follow on research to start.

Problem 2: Create a function that will do Problem 1 for any year.

```
QB_stats <- function(season) {
nflyear <- nflfastR::load_player_stats(season)
nflyear %>%
  select(player_id, player_display_name, position, passing_yards, interceptions, passing_tds,
         rushing_yards, rushing_tds, season) %>%
  filter(position == "QB" | position == "RB") %>%
  group_by(player_id) %>%
  mutate(total_yards = sum(passing_yards),
         total_picks = sum(interceptions),
         total_rush_yards = sum(rushing_yards),
         total_tds = sum(passing_tds) + sum(rushing_tds)) %>%
  select(-passing_yards,
         -interceptions,
         -passing_tds,
         -rushing_tds,
         -rushing_yards) %>%
  arrange(-total_yards) %>%
  rename(Name = player_display_name) %>%
  unique() %>%
  ungroup() %>%
  head(15)
```

```
}
```

Problem 3: Create a DataFrame of all the seasons since 2000. Do not include 2000, 2006, and 2012. Running Backs won MVP these seasons, so including them will influence the QB MVP prediction process.

```
seasons <- c(2001, 2002, 2003, 2004, 2005, 2007, 2008, 2009, 2010, 2011, 2013, 2014,
             2015, 2016, 2017, 2018, 2019, 2020, 2021)




results_list <- list()


for (season in seasons) {
  season_stats <- QB_stats(season)
  results_list[[as.character(season)]] <- season_stats
}

# Combine results from all seasons into a single dataframe
combined_df <- bind_rows(results_list, .id = "Season")
```

Problem 4: We have created a list of all MVP winners for the years within the dataframe. Use a left join to add an MVP column with binary identification: a 1 for MVPs and a 0 for everyone else.

Problem 5: Create a model that predicts if a player will be MVP based on the stats of previous MVP winners.

```
model.mvp <- glm(MVP == "1" ~ total_yards + total_picks + total_tds,
                 family = "binomial",
                 data = MVP_seasons)
MVP_seasons %>%
  mutate(mvp_pred = predict(model.mvp,
                            type = "response",
                            newdata = .)) -> MVP_seasons

summary(model.mvp)
```

```
##
```

```
## Call:
## glm(formula = MVP == "1" ~ total_yards + total_picks + total_tds,
##     family = "binomial", data = MVP_seasons)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.6707241  2.3168234  -3.743 0.000182 ***
## total_yards  0.0004735  0.0007047   0.672 0.501668
## total_picks -0.1767555  0.0815260  -2.168 0.030152 *
## total_tds    0.1565802  0.0505934   3.095 0.001969 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 128.821  on 284  degrees of freedom
## Residual deviance:  78.595  on 281  degrees of freedom
## AIC: 86.595
##
## Number of Fisher Scoring iterations: 7
```

The reason we are using a Generalized Linear Model (glm) in this scenario is because regular linear regression is restricted to only data with a normal distribution.

The regression table above shows the results of our model. The intercept represents the odds that a quarterback with 0 statistics would win MVP. The value of -8.67 has no meaning within the context of our model, so it is equitable to zero. Intuitively, a QB with no stats has no chance at winning MVP. Each coefficient within the column titled Estimate below the intercept represents the change in odds a player wins MVP with a 1 unit increase of the variable name. Therefore, the log odds of winning MVP increase 0.15 per touchdown. Per the table, which statistic matters the most to MVP voters?

Problem 6: Create a table of the top 10 most likely MVP seasons of players who did NOT win the MVP award.

```
MVP_seasons %>%
  arrange(desc(mvp_pred)) %>%
  select(Name, Season, mvp_pred, total_tds, MVP) %>%
  filter(MVP == "0") %>%
  head(15) %>%
  kable(caption = "Top-10 Highest MVP Probabilities for non-MVP's",
        col.names = c("Name", "Season", "P(MVP)", "TDs", "MVP"), digits = 3)
```

Table 1: Top-10 Highest MVP Probabilities for non-MVP's

Name	Season	P(MVP)	TDs	MVP
Aaron Rodgers	2016	0.648	53	0
Drew Brees	2011	0.497	54	0
Josh Allen	2020	0.477	51	0
Tom Brady	2020	0.457	54	0
Patrick Mahomes	2020	0.403	45	0

Name	Season	P(MVP)	TDs	MVP
Tom Brady	2011	0.350	51	0
Tom Brady	2021	0.343	48	0
Patrick Mahomes	2021	0.327	51	0
Josh Allen	2021	0.279	51	0
Tom Brady	2015	0.273	43	0
Daunte Culpepper	2004	0.251	47	0
Matthew Stafford	2021	0.234	52	0
Drew Brees	2009	0.214	44	0
Patrick Mahomes	2019	0.213	40	0
Drew Brees	2018	0.183	40	0

```
MVP_seasons %>%
  arrange(desc(mvp_pred)) %>%
  select(Name, Season, mvp_pred, total_tds, MVP) %>%
  filter(MVP == "1") %>%
  head(15) %>%
  kable(caption = "Top-10 Highest MVP Probabilities for MVP's",
        col.names = c("Name", "Season", "P(MVP)", "TDs", "MVP"), digits = 3)
```

Table 2: Top-10 Highest MVP Probabilities for MVP's

Name	Season	P(MVP)	TDs	MVP
Peyton Manning	2013	0.833	61	1
Aaron Rodgers	2020	0.822	57	1
Tom Brady	2007	0.749	58	1
Patrick Mahomes	2018	0.660	56	1
Matt Ryan	2016	0.606	48	1
Aaron Rodgers	2011	0.561	50	1
Peyton Manning	2004	0.538	54	1
Aaron Rodgers	2014	0.330	44	1
Cam Newton	2015	0.313	50	1
Aaron Rodgers	2021	0.257	40	1
Tom Brady	2017	0.246	40	1
Tom Brady	2010	0.188	39	1
Peyton Manning	2003	0.061	38	1
Rich Gannon	2002	0.044	37	1
Peyton Manning	2009	0.041	39	1

6 - Wrap Up / Conclusions

While exploring the prediction of NFL MVP probabilities using a generalized linear model (GLM) with predictors such as total yards, touchdowns, and turnovers, we learned the versatility of a GLM. It allows us to model and predict outcomes that have non-normal distributions, like MVP awards. This model also shows the significance of stats such as yards, turnovers, and touchdowns and how that determines a player's impact on the field. There have been some examples where our model's predicted winner did not win the award. We considered almost all of the QB statistics, but an important variable we did not include was wins. So, do you think that the MVP is the best QB of the season with the best stats or the QB on the most successful team of that season? Here are some examples of Predicted winners vs. the guy who won MVP:

```
winscomp %>%
  kable()
```

Year	Winner	Winner.Odds	Wins	Predicted	Predicted.Odds	Wins.1
2016	Matt Ryan	0.592	11	Aaron Rodgers	0.638	10
2009	Peyton Manning	0.035	14	Drew Brees	0.199	13
2021	Aaron Rodgers	0.244	13	Tom Brady	0.324	13

A future skill would be to delve into machine learning techniques such as deep learning. Having a better understanding of this skill could allow more intricate patterns to be found in player performance. Being more skilled would improve the accuracy and sureness of the patterns found.