NFL analytics

Michael Lopez, Skidmore College

Overview

In this lab, we'll explore the `nflfastR` data set, and look at football metrics such as expected points and win probability.

Link: "https://raw.githubusercontent.com/guga31bb/nflfastR-data/master/data/play_by_play_2020.rds"

```r
library(tidyverse)
pbp_20 <- readRDS(url("https://raw.githubusercontent.com/guga31bb/nflfastR-data/master/data/play_by_pla
dim(pbp_20)

pbp_scrimmage <- pbp_20 %>%
  select(game_id, posteam, defteam, play_type, play_id, down, ydstogo, yardline_100, home_score, away_s
         passer_player_name, receiver_player_name, air_yards, complete_pass, desc)  %>%
  filter(play_type == "pass"|play_type == "run")


set.seed(10)
pbp_scrimmage %>%
  sample_n(6) %>%
  print.data.frame()
```

We'll be using `pbp_scrimmage` for the bulk of this lab. The variables above correspond to play-level data from the first few weeks of the 2020 season (the data updates after each game).

Exploratory data analysis

Let's start with some basic data analysis, which should be the first thing we think about when looking at a new data set.

1. Use a plot to examine the distribution of expected points added – the center, shape, and spread of this variable.
2. Same as the above, but facet by the play type (rush versus pass). Is either play type generally more successful? Any differences in the distributions? Why might that matter as far as coaching preferences?
3. Estimate the average epa within each team, both on run and pass plays. *Hint*: This should be three lines of code
4. Using only passing plays, find the quarterback (`passer_player_name`) with the highest average epa.
5. Same as above, except filter for quarterbacks with at least 30 attempts (in `summarize`, you can use `n_attempts = n()` to get a sample size for each QB)

Quarterback metrics

Expected points are nice because it allows us to compare how plays helped or hurt an offense put points on the board.

A similar metric can be used with players at each position. Let's look with quarterbacks.

```r
passes <- pbp_scrimmage %>%
  filter(play_type == "pass", !is.na(air_yards), air_yards >= -10)

passes %>%
  group_by(passer_player_name) %>%
  summarise(completion_pct = mean(complete_pass),
            n_passes = n()) %>%
  filter(n_passes >= 30) %>%
  arrange(-completion_pct)
```

The above code ranks quarterbacks with at least 30 attempts by their completion percentage.

However, not all completions are equal. It's much easier for quarterbacks to complete 5 yard passes than to complete 25 yard passes. Which quarterbacks tend to throw shorter and deeper passes?

We can look at the `air_yards` variable

```r
passes %>%
  group_by(passer_player_name) %>%
  summarise(completion_pct = mean(complete_pass),
            avg_pass_length = mean(air_yards),
            n_passes = n()) %>%
  filter(n_passes >= 30)
```

6. Which quarterbacks have tended to throw the shortest and longest passes?

QB performance versus expected

Not surprisingly, it's more difficult to complete passes that travel further in the air.

```r
fit_1 <- glm(complete_pass ~ air_yards, data = passes, family = "binomial")
library(broom)
tidy(fit_1)
```

We can use the model above to estimate probabilities for each pass being complete.

```r
passes <- passes %>%
  mutate(complete_hat = predict(fit_1, type = "response"))
```

The predict command provides a probability of being complete for each pass in the data set, at least using the model above.

Let's take a look at the first pass in the data set, from Jimmy Garoppolo to George Kittle.

```r
passes %>% head(1) %>% print.data.frame()
```

The pass to Kittle traveled four yards in the air, which comes with an estimated completion probability of about 72 percent.

7. Identify the average expected completion percentage for each quarterback – that is the average of their `complete_hat`. This represents how an average quarterback would do if they were given the same pass lengths as was given to each quarterback.

8. Using your average above, as well as the observed completion percentages (See earlier code), which quarterbacks have the highest completion percentage above their expectation? The lowest?

Visualizing QB performance

9. Use the `passes` data set, and make a plot of air-yards (x-axis) versus `complete_hat` on the y-axis. How do you feel about this association?

10. What's going on in the plot below? Identify the quarterbacks doing (i) the best and worst at the most difficult passes and (ii) the best and worst at the easiest passes.

```
passes %>%
  group_by(passer_player_name) %>%
  summarise(complete_rate_obs = mean(complete_pass),
            complete_rate_exp = mean(complete_hat),
            complete_above_exp = complete_rate_obs - complete_rate_exp,
            n_passes = n()) %>%
  filter(n_passes >= 30) %>%
  ggplot(aes(complete_rate_exp, complete_rate_obs)) +
  geom_text(aes(label = passer_player_name)) +
  geom_point()
```