Lab 6: NFL play by play

Michael Lopez, Skidmore College

Important note 1

First, we'll open RStudio by going to http://r.skidmore.edu/.

Important note 2

Open a new R Markdown file (File / New File / R Markdown...). You can create a basic name – Lab0, for example – and that'll set you up with a new file ready to go.

Overview

After looking at field goal kickers and vague NFL play-by-play data from a few years ago, we are ready to look at the most modern ways of approaching play success in the NFL: expected points added and win probability added (EPA and WPA).

Expected points added reflects the difference in a teams' expected points after a play to their expected points prior to the play.

For example, moving from 1.5 to 2.7 expected points is worth an expected points added of 1.2.

Win probability added is sort of the same idea, but instead of using expected points (which are tied to down, distance, and field position), win probability also factors in time on the clock and score.

One thing that is interesting to note: both EPA and WPA can be negative, and often are.

Summary of EPA and WPA

First, we load in the data, which contains information on all pass plays run so far in the 2018 season.

```
library(RCurl)
library(tidyverse)
url.18 <- getURL("https://raw.githubusercontent.com/ryurko/nflscrapR-data/
        master/play_by_play_data/regular_season/reg_pbp_2018.csv")
pbp <- read.csv(text = url.18)
head(pbp)
passes <- pbp %>% filter(play_type == "pass")
passes <- passes %>%
  select(game_id, home_team, away_team, posteam, yardline_100, game_seconds_remaining,
         qtr, down, ydstogo, yards_gained, shotgun, no_huddle, pass_length,
         pass_location, epa, wpa, passer_player_name, receiver_player_name) %>% na.omit()
```

1. Make a histogram of both EPA and WPA. Identify the shape, center, and spread of each.
2. Make side by side boxplots of EPA for each possession team (`posteam`). Which team has the highest median EPA?
3. Now, use the summarize command to double check your answer in No. 2.

Density plots, groupings, facets

An analyst is interested in comparing `wpa` between passes thrown short and those thrown deep.

Our traditional way of doing this is side-by-side boxplots

```
ggplot(data = passes, aes(x = pass_length, y = epa)) + geom_boxplot()
```

An alternative to boxplots is density plots. You can think of density plots like pieces of spaghetti that are put on top of histograms.

```
ggplot(data = passes, aes(x = epa, colour = pass_length)) + geom_density()
```

In the code above, each colour reflects a different pass length – and the density curves are overlaid.

Key takeaway: Density curves are best analyzed not by the y-axis, but based on the area that corresponds to different portions of the x-axis. For those familiar with calculus, the curves are designed so that the total "area under the curve" is equal to 1.

4. Summarize what you see in the density curves of `epa` based on `pass_length`.
5. What features are apparent in the density plots that are not apparent in the boxplots? What features are apparent in the boxplots that are not apparent in the histograms?
6. Make overlapping density curves of `wpa`, with a different colour for each `no_huddle` (0 or 1, for whether a team was in no huddle). Focus on the middle part of the curves by adding the code `+ xlim(c(-0.1, 0.1))`. Passes to what location tend to add the most win probability?
7. Facet the chart above by pass distance. You should end up with 2 graphs, each showing 3 curves. Where is it most obvious that passing to the middle tends to increase `wpa`? Deep throws or short throws?

On-your-own

8. Make one scatter plot of `game_seconds_remaining` and EPA. Make another of `game_seconds_remaining` and WPA. How are EPA and WPA linked to the amount of time left in the game? Note: think carefully about what it means to have a small number of seconds remaining in a game.
9. Using all receivers who have been targeted at least 20 times, identify the ones with the five highest expected points added. Note: each row in this data set reflects a target.
10. Using all QBs who have thrown at least 50 passes, identify those who threw deep (`pass_length == "deep"`) most often. For those of you who know football well, what do you notice about two of these players?