HW 5: NFL kicker eval

Stats and sports class

Fall 2019

Preliminary notes for doing HW

1. All files should be knit and compiled using R Markdown. Knit early and often! I do not recommend waiting until the end of the HW to knit.
2. All questions should be answered completely, and, wherever applicable, code should be included.
3. If you work with a partner or group, please write the names of your teammates.
4. Copying and pasting of code is a violation of the Skidmore honor code

Homework questions

Part I: Multiple regression and player metrics

For this homework, we will be using data provided in the `nfl.kick` data set, as was done during class. Our goals will be to confirm our knowledge of logistic regression, our interpretations of slopes, as well as kicker-specific analysis.

```
## Note: if using your personal computer, run `install.packages(RCurl)`
library(RCurl); library(tidyverse)
url <- getURL("https://raw.githubusercontent.com/statsbylopez/StatsSports/master/Data/nfl_fg.csv")
nfl_kick <- read.csv(text = url)
head(nfl_kick)
```

Exploratory data analysis

Question 1

Use R to find the **kicker** with the best percentage of successful field goals. Why might one argue that this specific kicker may not be the most accurate, even though he has the highest percentage? Return to your lab – where we use `group_by()` and `summarize()` – for suggested code.

Question 2

Use R to find the **team** with the best percentage of successful field goals. Why might one argue that this team may not have had the best kickers even though they've posted the highest overall percentage?

Question 3

Identify the teams that have kicked the highest percentage of their field goals on grass (recall: the `Grass` variable is a TRUE/FALSE indicator for whether or not each kick was kicked on a grass surface.).

Logistic regression

Question 4

Use the following code for the next several questions.

```
library(broom)
fit_1 <- glm(Success ~ Distance + Grass + Year,
             data = nfl_kick, family = "binomial")
tidy(fit_1)
```

Using the model above, interpret the coefficient for Grass on the odds scale

Question 5

Using your model from (4), interpret the coefficient for `Distance` on the odds scale.

Question 6

Odds ratios are multiplicative. That is, if the odds of a successful outcome are e^{β_1} given a one-unit increase in x_1, the odds of a successful outcome are $e^{c*\beta_1}$ given a c-unit increase in x_1. Given your model in (4) what are the odds of making a field goal that is 10 yards longer?

Question 7

Estimate the probability of a successful 40-yard field goal, kicked on a non-grass surface in 2015.

Expected points

Question 8

Use your answer to Question (7) to estimate the expected points of a 40-yard field goal, kicked on a non-grass surface in 2015.

Question 9

Kicker A hits the field goal in Question (8) while Kicker B misses it. How many expected points has Kicker A added to his team given this single kick? How about Kicker B?

Question 10

It is straightforward to estimate the value of kickers using expected points.

First, we generate predicted probabilities for each field goal using `fit_1`. Next, we use that to estimate the expected points for each field goal (`predict_points`). Finally, we use the result of the field goal (`Success` = 0 or a 1) and the value of the kick (3 points) to get an expected points added (`EPA`) for each kicker on each kick.

```
nfl_kick <- nfl_kick %>%
  mutate(predict_Success = predict(fit_1, nfl_kick, type = "response"),
                   predict_points = 3*predict_Success,
                   EPA = Success*3 - predict_points)
nfl_kick %>% head()
```

The first row corresponds to a David Akers kick in 2005. What was the predicted success rate for Akers on this kick? What relative worth (in terms of `EPA`) did Akers provide on this kick?

Question 11

One metric we may be interested in is the relative worth, in terms of total `EPA`, among all kickers in our data set. The `dplyr` function makes it simple.

```
options(dplyr.print_max = 1e9)
kick_summary <- nfl_kick %>%
  group_by(Kicker) %>%
  summarize(percent_success = mean(Success),
            total_kicks = n(),
            total_EPA = sum(EPA)) %>%
  arrange(total_EPA)
kick_summary
```

The above function calculates kicker-specific percentages, each kicker's total number of kicks, and each kickers total EPA.

Since 2005, who has been worth the most (and least) total `EPA` to their teams?

Question 12

Interpret the R-squared calculated below. What does it suggest about the fraction of unexplained variability when it comes to kicker EPA?

```
kick_summary <- kick_summary %>% filter(total_kicks >= 50)
ggplot(kick_summary, aes(percent_success, total_EPA)) +
  geom_point()

kick_summary %>%
  summarise(r2_epa_pct = cor(total_EPA, percent_success)^2)
```

Question 13

Given your readings, are there any other variables that you would want to account for when measuring field goal success that aren't in the current data set? How may it effect the ranking of kickers in Question Question (12)?

Question 14

Make a better plot than the one shown above. Be creative!