Logistic regression and NFL kickers

Michael Lopez, Skidmore College

Overview

In this lab, we'll gain experience implementing logistic regression to estimate the probability of successful NFL field goals given play and game specific conditions. Logistic regression will be a tool we use throughout the semester, and it is useful for a variety of sports/statistics questions. While field goal kickers are not the *most* exciting players in the game, their analysis provides an insight into our topic on expected points and decision-making, which we'll cover next week.

I uploaded a .csv file with lots of kicker data. You can view that data by clicking here.

```
## 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)
```

We load this as the `url` using the `getURL` command, and load it into R.

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.

```
ggplot(data = nfl_kick, aes(Year, Distance)) +
  geom_boxplot()


ggplot(data = nfl_kick, aes(as.factor(Year), Distance)) +
  geom_boxplot()
```

1. Identify the code change in the two lines above. What does it insinuate about how to deal with a time variable that's continuous, but that you want to look at it as a factor?
2. Describe the distributions of field goal distance by year. Do there appear to be any changes over time?

```
nfl_kick %>%
  group_by(Grass) %>%
  summarise(ave_success = mean(Success))

nfl_kick %>%
  mutate(is_late = GameMinute >= 57) %>%
  group_by(is_late) %>%
  summarise(ave_success = mean(Success))
```

3. Describe the link between success rates (`Success`) and `Grass` and the `is_late` variables.
4. Propose one alternative reason for field goals in the game's final minutes showing lower success rates.
5. Use exploratory data analysis tools to (i) identify if kickers on grass surfaces attempt shorter kicks than those on non-grass surfaces and (ii) identify if there is any link between the distance of the kick and the game minute of the kick.

Logistic regression

At this point, we are aware that, overall, there seems to be a link between surface, distance, game minute, and field goal success. However, there's also so correlation between a few of these predictors; kickers take longer field goals in the final few minutes of each half, for example. This makes it difficult to discern if the link between `GameMinute` and lower `Success` rates is real, or if its just simply because the kickers were kicking longer field goals.

We are going to walk through several logistic regression models of field goal kickers, with a focus on (i) identifying the most important parts of the output, (ii) interpreting the results, and, eventually (iii) linking back to the probability of a successful kick.

First, let's fit the model that we used in class on Tuesday:

$$log(\frac{P(Success=1)}{1-P(Success=1)}) = \beta_0 + \beta_1 * Distance$$

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

We can write the estimated regression equation as:

$$log(\frac{P(Success=1)}{1-P(Success=1)}) = 5.7246 + (-0.1026) * Distance$$

Estimated probabilities

Recall, probability estimates can be made by transforming the above equation (written on the log-odds scale) into probabilities.

$$P(Success = 1) = \frac{e^{5.7246+(-0.1026)*Distance}}{1+e^{5.7246+(-0.1026)*Distance}}$$

R will allow us to make such calculations quickly, and there are several ways to do this.

First, we can use the `exp()` function to compute the exponential function. Let's say we have a 50-yard field goal:

```
phat_50 <- exp(5.7246 + (-0.1026)*50) / (1 + exp(5.7246 + (-0.1026)*50))
phat_50
```

Our model suggests that the probability of a successful 50-yard field goal is 64.4%.

We can also use the `predict()` function in R to get the predicted probabilities for an assortment of distances. In the code below, the `type = "response"` ensures that the predictions are done on the probability scale (as opposed to, say, the log-odds scale).

```
df_predict <- data.frame(Distance = c(20, 30, 40, 50, 60))
pred <- predict(fit_1, df_predict, type = "response")
pred
```

Slope estimates from logistic regression

Recall that the estimated slopes from a logistic regression model are given on the log-odds scale. We can exponentiate to get estimated odds ratio.

```
exp(-0.1026)
exp(fit_1$coeff)
exp(confint(fit_1))
```

The first command calculated the estimated odds ratio by hand, while the second used the fact that R stores regression coefficients as `coeff`. In the third command, we use the `confint` command to get estimated confidence intervals for each parameter.

In the logistic regression output, we also see an AIC value (more on that later), as well as a z-test statistic for each parameter. In this case, the estimated β_1 is significantly different from 0. There is strong evidence that distance is linked to field goal success (shocker!).

Multiple logistic regression

Ultimately, we are interested in how several of our variables link to field goal success. Let's create one model, `fit_2`, below.

```
options(scipen=99)
fit_2 <- glm(Success ~ Distance + Grass + Year + GameMinute,
             data = nfl_kick, family = "binomial")
tidy(fit_2)
```

There's a bunch to get to in this model. Let's start by getting our estimated odds ratios:

```
odds_ratios <- exp(fit_2$coeff)
round(odds_ratios, 3)
odds_intervals <- confint(fit_2)
round(odds_intervals, 3)
```

First, notice that the estimated effecet of `Distance` did not change much from `fit_1`. This suggests that, when including the other predictors, the link between `Distance` and `Success` remains the same.

Next, `Grass` is a categorical variable. We interpret this as follows:

Given 'Year', 'GameMinute', and 'Distance', the odds of a successsful kick on grass surfaces are about 0.856 times that of one kicked on non-grass surfaces.

Given 'Year', 'GameMinute', and 'Distance', the odds of a successful kick on grass surfaces are about 14.4% less than that of one kicked on non-grass surfaces.

Either interpretation above is fine; I tend to think of the second one (relative percent odds) as more intuitive.

In this model, `Year` and `GameMinute` are treated as continuous, and their interpretations would be similar to those for `Distance`.

On your own

6. Interpret the coefficient on `Year` in `fit_2` (using the odds-ratio scale).
7. Using `fit_2`, estimate the probability of a successful 40-yard kick in 2010, made on a grass surface in the 10th minute of the game. Hint: make a data frame with the identical columns, and use the `predict()` command.

```
df_predict <- data.frame(Distance = c(40),
                         Grass = TRUE,
                         GameMinute = 10,
                         Year = 2010)
```

8. Do the same as in (7), only now using a non-grass surface.
9. Turn your probability estimates from (7) and (8) into odds, and then take their ratio. Where else do you see this number?
10. In the summary of `fit_2`, the coefficient for `Grass` is larger than the coefficient for `Distance`. Does this mean that surface is a more important predictor of field goal success than distance? If not, give a guess as to *why* the results are the way they are.
11. There are several variables in this data set. Using the AIC criteria, identify the model that is the best fit for our `Success` outcome.
12. One variable that is difficult to include in a regression model is the kicker Using the idea of `Expected points`, think of ways that we can estimate the net benefit or harm of each kicker, at least based on their field goal attempts.
13. 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?