Lesson 22 NFL Field Goals

LTC J.K. Starling

Last compiled on 18 March, 2022

```
# Load packages
library(tidyverse)
library(table1)

### Data management
#loading play by play for the 2021 season
# library(nflfastR)
# data <- load_pbp(2021)
# field_goal <- data %>% filter(is.na(field_goal_result)!=T)


# field_goal <- read.csv("https://raw.githubusercontent.com/jkstarling/MA376/main/fg_2021.csv",header=T

# setwd("C:/Users/james.starling/OneDrive - West Point/Teaching/MA376/JimsLessons/Block III/LSN22/")
# field_goal <- read.csv("fg_2021.csv", header = TRUE, stringsAsFactors = FALSE)
```

Background The 2021 NFL Season had seen some historically bad weeks for the field goal kickers. In this lesson we will look at the relationship between kick distance and and field goal result, and the relationship between field surface and field goal result as part of the model building process.

(1) Explore the variables `field_goal_result` and `surface` using the `table()` command. What do you notice? Can we simplify the data? (ask for the email...)

(2) Create a contingency table using the `table1()` command.

```
# Contingency table
```

(3) Calculate the odds ratio comparing field goals attempted on turf and field goals attempted on grass?

(4) How do we interpret the odds ratio?

(5) What if I wanted the odds ratio between field goals attempted on grass and field goals attempted on turf, how do I calculate and interpret it?

Logistic Regression Model To be able to adjust for the field surface we need some type of model. Because we have a binary response we will use the logistic regression model. The logistic regression model is a linear equation that takes the form:

$$logit(\pi_i) = ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_i$$

where log odds are linearly related to x (if we graph π and x they are related in the S-shaped curve.) **Notice we do not have an error term in the logit model. This is because we are not modelling individual values of the response variable.**

I am assuming that you are familiar with logs with base e, where e represents Euler's constant of approx 2.718. R uses base log base e by default when you use the *log* function.

The model estimate is:

$$\text{Predicted log odds} = logit(\hat{p}_i) = ln\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = b_0 + b_1 x_i$$

With this model we can estimate the probability of success π (estimated proportion of success) by following the rules of logs and exponents.

$$exp\left[ln\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right)\right] = exp\left[b_0 + b_1 x_i\right]\frac{\hat{p}_i}{1-\hat{p}_i} = exp\left[b_0 + b_1 x_i\right]\hat{p}_i = \frac{exp\left[b_0 + b_1 x_i\right]}{1 + exp\left[b_0 + b_1 x_i\right]}$$

(6) We conduct inference on our coefficients as usual. What are the null and alternative hypotheses? How will we know if the coefficient is statistically significant?

(7) How do we get the odds ratios given the log odds?

(8) Our hypotheses are:

$$H_0 : OR = 1 \text{ vs } H_A : OR \neq 1$$

- Confidence interval containing the null value indicates that the estimate is not statistically significant. So if we are dealing with OR, when do we have a statistically significant result?

Categorical Explanatory Variable

(9) Fit the model with the binary explanatory variable that indicates whether or not the the field goal was attempted on turf or grass. (remember to check the levels of the **surface** variable with the function `contrasts()`)

```
# contrasts(...)
# surface.glm <- ...
# summary(surface.glm)
# coef(surface.glm)                # Predicted log odds of making the field goal
# confint(surface.glm)
```

(10) How do we interpret the coefficients directly from the model? What conclusions can we draw based on the confidence interval?

(11) Typically these interpretations don't make much sense to the consumer; how do we write them and interpret them in terms of odds ratio? What conclusions can we draw based on the confidence interval?

(12) What is the predicted probability of field goal attempted on grass? On turf?

```
#On Grass

#On Turf
```

Quantitative Explanatory Variable

(13) Fit the model with the quantitative explanatory variable that captures kick distance.

```
# distance.glm <- glm(...)
# coef(distance.glm)                # Predicted log odds of survival
# confint(distance.glm)
```

(14) How do we interpret the coefficients directly from the model? What conclusions can we draw based on the confidence interval?

(15) Typically these interpretations don't make much sense to the consumer; how do we write them and interpret them in terms of odds ratios? What conclusions can we draw based on the confidence interval?

(16) What is the predicted probability of making a field goal from 20 yards? 44 yards? 60 yards? What is the relationship between predicted probability and kick distance?

```
# newobs.data <- data.frame(kick_distance = c(...))
#
# predict.glm(...., type="response")
#
# # Predicted probability (inverse logit)
# distance.dat.sub <- fg %>% mutate(distance.pp = distance.glm$fitted.values)
# #view(smoke.dat.sub)
# distance.dat.sub %>% ggplot(aes(x=kick_distance, y=distance.pp) ) +
#   geom_point() +
```

```
#   geom_line()+
#   theme_classic()
```

In this lesson we looked at the relationship between field surface and making a field goal, and the relationship between kick distance and making a field goal as part of the model building process. Next week we will put it all together to answer a research question with a multivariate logistic regression model

References Nathan Tintle et al.(2019). Intermediate Statistical Investigations for U.S. Military Academy at West Point.