

## In Class Exploration - LSN 24

20 points

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

```
fg <- read.csv("https://raw.githubusercontent.com/jkstarling/MA376/main/fg_2021-2.csv",
               header=T, stringsAsFactors = TRUE)

# setwd("C:/Users/user.name/...")
# field_goal <- read.csv("fg_2021-2.csv", header = TRUE, stringsAsFactors = FALSE)

fg$field_goal_result <- fct_rev(fg$field_goal_result)
```

For reference, we provide the table results for field goal results and the table results of surface type.

```
table(fg$field_goal_result)
```

```
##
## missed    made
##    158    918
```

```
table(fg$surface)
```

```
##
## grass    turf
##    637    439
```

- (1) Calculate the odds ratio comparing field goals attempted on turf and field goals attempted on grass.
- (2) How do we interpret the odds ratio?
- (3) What if we 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:

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

where log odds are linearly related to  $x$  (if we graph  $\pi$  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.**

- (4) We conduct inference on our coefficients as usual. In the context of the problem, what are the null and alternative hypotheses (in words and symbols) when using logistic regression? How will we know if the coefficient is statistically significant?
- (5) Fit the model with the binary explanatory variable that indicates whether or not the field goal was attempted on turf or grass. Use the `confint()` function to calculate the confidence intervals of the coefficients.  
(note: remember to check the levels of the **surface** variable with the function `contrasts()`)
- (6) How do we interpret the coefficients directly from the model? What conclusions can we draw based on the confidence interval?
- (7) What is the predicted probability of field goal attempted on grass? On turf?
- (8) Fit the model with the quantitative explanatory variable that captures kick distance.
- (9) How do we interpret the coefficients directly from the model? What conclusions can we draw based on the confidence interval?
- (10) 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?
- (11) 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?