MA256 Lesson 20 Logistic Regression

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?

```
# table(XXX)
# table(XXX)
```

(2) Using what you found in the last question, use the `case_when()` function to change *blocked* field goals to *missed* and change various types of turf to say *turf*. Hint: mimic the example for the surface below.

```
# #changing the variations of turf to just turf
# fg <- field_goal %>%
#   mutate(surface = case_when(surface=="astroturf"~"turf",
#                              surface==XXX....
#                              TRUE~surface)) %>%
#   mutate(surface = as.factor(surface))

# # Selecting only field goals and changing blocked field goals to be missed.
# fg <- fg %>%
#   mutate(field_goal_result = case_when(field_goal_result==XXXX,
#                                        TRUE~field_goal_result)) %>%
#   mutate(field_goal_result = fct_rev(as.factor(field_goal_result)))
```

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

```
# # Contingency table
# table1(XXX)
# table(XXX)
```

(4) Calculate the odds ratio comparing field goals attempted on turf and field goals attempted on grass? How do we interpret the odds ratio?

```
# odds_turf <- XXX
# odds_grass <- XXX
# OR <- XXX
# OR
```

(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?

```
# XXX
```

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.)

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:

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

(6) With this model we can estimate the probability of success π (estimated proportion of success). Show how you can do this.

(7) 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?

Categorical Explanatory Variable

(8) Fit the model with the binary explanatory variable that indicates whether or not the the field goal was attempted on turf or grass.
Calculate the confidence interval of the model's coefficients using the `confint()` function. (Check the levels of the **surface** variable with the function `contrasts()`)

```
# contrasts(XXX)
# surface.glm <- fg %>% glm(XXX)
# summary(XXX)
# coef(XXX)
# confint(XXX)
```

(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 ratio? What conclusions can we draw based on the confidence interval? (Hint: Check out p. 28 in the handout)

```
# XXX
# XXX
```

(11) What is the predicted probability of field goal attempted on grass? On turf? BONUS: What do you notice about the OR between the two odds?

```
# #On Grass
# odds_grass <- XXX
# prob_grass <- XXX
# #On Turf
# odds_turf <- XXX
# prob_turf <- XXX
#
# c(prob_grass, prob_turf)
# # OR
# XXX ???
```

Quantitative Explanatory Variable

(12) Fit the model with the quantitative explanatory variable that captures kick distance. Calculate the 95% CI for the coefficient estimates.

```
# distance.glm <- XXXX
# # summary(distance.glm)
# coef(distance.glm)
# 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?

```
# XXX
# XXX
```

(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(XXXXXX))
#
# predict.glm(distance.glm, newdata=newobs.data, 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=XXX, y=XXX) ) +
#   geom_point() + geom_line() + theme_classic() +
#   labs(x="Kick distance", y="kick probability", title = "Probability of successful kick vs. distance (in
```