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(field_goal$field_goal_result)
```

```
##
## blocked    made  missed
##      14     918     144
```

```
table(field_goal$surface)
```

```
##
##  astroturf  fieldturf      grass matrixturf  sportturf
##        100        227        637         72         40
```

We notice that there are multiples types of turf and that there are 14 blocked field goals (which are also missed field goals)

(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=="fieldturf"~"turf",
                             surface=="matrixturf"~"turf",
                             surface=="sportturf"~"turf",
                             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=="blocked"~"missed",
                                       TRUE~field_goal_result)) %>%
  mutate(field_goal_result = fct_rev(as.factor(field_goal_result)))
```

(3) Create a contingency table using the `tabyl()` command from the janitor package (or the `table()` command).

```
# Contingency table
# table1(~field_goal_result|surface, data = fg)
# table(fg$field_goal_result, fg$surface)

fg %>%
  tabyl(field_goal_result, surface) %>%
  adorn_totals(c("row", "col")) %>% adorn_percentages() %>% adorn_pct_formatting() %>% adorn_ns(position="
```

```
##  field_goal_result       grass        turf          Total
##             missed 100 (63.3%)  58 (36.7%)   158 (100.0%)
##               made 537 (58.5%) 381 (41.5%)   918 (100.0%)
##              Total 637 (59.2%) 439 (40.8%) 1,076 (100.0%)
```

(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 <- 381/58
odds_grass <- 537/100
OR <- odds_turf/odds_grass
OR
```

```
## [1] 1.223271
```

```
(OR - 1) * 100
```

```
## [1] 22.3271
```

The odds of field goals attempted on turf being made are 1.22 times larger than the odds of field goals attempted on grass being made OR field goals attempted on turf are 22% more likely to be made compared to field goals attempted on grass

(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?

```
1/OR
```

```
## [1] 0.8174803
```

```
(1/OR - 1) * 100
```

```
## [1] -18.25197
```

Field goals attempted on grass are 18% less likely to be made compared to field goals attempted on turf.

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.

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

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

$$\hat{p}_i = \frac{exp\left[b_0 + b_1 x_i\right]}{1 + exp\left[b_0 + b_1 x_i\right]} = \frac{odds}{1 + odds}$$

(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?

Null: There is no linear relationship between the log odds and the explanatory variable

Alternative: There is a linear relationship between the log odds and the explanatory variable

$$H_0 : \beta_1 = 0 \text{ vs } H_A : \beta_1 \neq 0$$

Confidence interval containing the null value (0) indicates that the estimate is not 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(fg$surface)
```

```
##       turf
## grass    0
## turf     1
```

```
surface.glm <- fg %>% glm(field_goal_result ~ surface, data=., family='binomial')
summary(surface.glm)
```

```
##
## Call:
## glm(formula = field_goal_result ~ surface, family = "binomial",
##     data = .)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.6808     0.1089  15.433   <2e-16 ***
## surfaceturf   0.2015     0.1781   1.131    0.258
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 897.79  on 1075  degrees of freedom
## Residual deviance: 896.49  on 1074  degrees of freedom
## AIC: 900.49
##
## Number of Fisher Scoring iterations: 4
```

```r
coef(surface.glm)                # Predicted log odds of making the field goal
```

```
## (Intercept) surfaceturf
##   1.6808279   0.2015285
```

```r
confint(surface.glm)
```

```
##                 2.5 %    97.5 %
## (Intercept)  1.472349 1.8997523
## surfaceturf -0.144254 0.5551773
```

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

The intercept of 1.6808 is the predicted log odds for field goal attempted on grass

The $\hat{\beta}_1$ coefficient is the change in log odds going from **kicking on grass** to **kicking on turf**.

The confidence interval for $\hat{\beta}_1$ contains 0 indicating that there is not a statistically significant relationship between surface type and log odds.

(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?

```r
exp(coef(surface.glm))
```

```
## (Intercept) surfaceturf
##    5.370000    1.223271
```

```r
exp(confint(surface.glm))
```

```
## Waiting for profiling to be done...
```

```
##                 2.5 %   97.5 %
## (Intercept) 4.3594623 6.684238
## surfaceturf 0.8656678 1.742250
```

The odds of making a field goal on grass is 5.37:1 or in other words we would expect to make 5.37 field goals for every miss.

The odds of making a field goal on grass is (5.370+1.223):1 = 6.593:1 or we can say that we would expect to make 6.593 field goals for every miss.

We are 95% confident that field goals kicked on turf have a 0.86 times smaller odds to 1.74 times larger odds of being made than field goals kicked on grass. This CI contains 1 indicating that there is not a significant difference in the odds of making a kick on turf vs grass

(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?

```r
#On Grass
#log odds = 1.6808
odds_grass <- exp(1.6808)
prob_grass <- odds_grass/(1+odds_grass)
#On Turf
#log odds = 1.6808+0.2015 = 1.8823
odds_turf <- exp(1.8823)
prob_turf <- odds_turf/(1+odds_turf)

c(prob_grass, prob_turf)
```

```
## [1] 0.8430104 0.8678751
```

```
odds_turf/odds_grass
```

```
## [1] 1.223236
```

Our predicted success probability of a field goal attempted on grass is 0.8430

Our predicted success probability of a field goal attempted on turf is 0.8679

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 <- glm(field_goal_result ~  kick_distance, data=fg, family='binomial')
# summary(distance.glm)
coef(distance.glm)                # Predicted log odds of survival
```

```
##   (Intercept) kick_distance
##     6.5465796    -0.1127346
```

```
confint(distance.glm)
```

```
##                    2.5 %      97.5 %
## (Intercept)    5.5771665  7.59618940
## kick_distance -0.1351956 -0.09162642
```

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

The intercept of 6.55 is the predicted log odds for field goal attempted with a distance of 0 yards

The $\hat{\beta}_1$ coefficient of -0.11 is the change in log odds for an increase in 1 yard of kick distance.

The confidence interval for $\hat{\beta}_1$ does not contain 0 indicating that there is a statistically significant relationship between kick distance and log odds.

(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?

```
exp(coef(distance.glm))
```

```
##   (Intercept) kick_distance
##   696.8565647     0.8933877
```

```
exp(confint(distance.glm))
```

```
## Waiting for profiling to be done...
```

```
##                     2.5 %      97.5 %
## (Intercept)   264.3215934 1990.596069
## kick_distance   0.8735451    0.912446
```

The slope coefficient indicates a multiplicative change (x 0.8934) for each increase of 1 yard in distance.

The confidence interval does not contain 1 indicating a significant relationship between kick distance and odds.

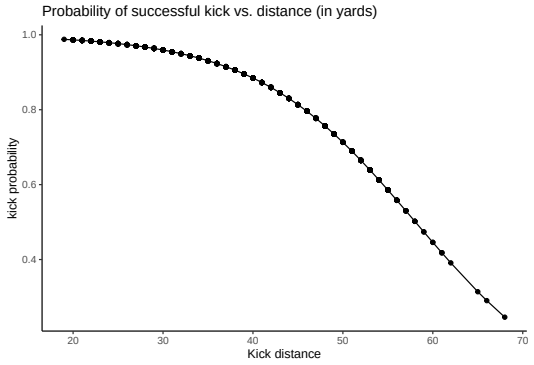
(16) What is the predicted probability of making a field goal from 20 yards? 44 yards? 60 yards? What is the relationship between predicted probability and kick distance?

```
newobs.data <- data.frame(kick_distance = c(20, 44, 60))

predict.glm(distance.glm, newdata=newobs.data, type="response")
```

```
##         1         2         3
## 0.9865055 0.8300888 0.4458390
```

```
# 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() +
  labs(x="Kick distance", y="kick probability", title = "Probability of successful kick vs. distance (in y
```



The predicted probability of a 20/44/60 yard field goal is 0.9865/0.8301/0.4458. There is a negative relationship between predicted probability and kick distance.