Project 3

David Nguyen, Huy Ngo

March 9, 2021

```r
# packages used
library(dplyr)
library(car)
library(mcprofile)
```

Flagstick in or out? Edoardo Molinari putting experiment.

1) (51 total points) This problem is a continuation of the setting from Project #2. Start with the original set1 data frame read in from the file flagstick.csv. When you read in the data, use the stringsAsFactors = TRUE argument.

```r
# read in full golf data set
set1 <- read.csv("data/flagstick.csv", stringsAsFactors = TRUE)
set1 %>% knitr::kable()
```

Flagstick	BallSpeed	EntryLine	Success	Trials
Out	1	1	100	100
Out	2	1	100	100
Out	3	1	81	100
Out	1	2	100	100
Out	2	2	73	100
Out	3	2	0	100
Out	1	3	100	100
Out	2	3	38	100
Out	3	3	0	100
In	1	1	100	100
In	2	1	100	100
In	3	1	100	100
In	1	2	100	100
In	2	2	45	100
In	3	2	7	100
In	1	3	100	100
In	2	3	14	100
In	3	3	0	100

a) (3 points) There are three categorical explanatory variables in the data frame: Flagstick, BallSpeed, and EntryLine. Convert the BallSpeed and EntryLine variables to a factor type within the set1 data frame. For all three variables, show verification that these variables are factors.

```
set1$BallSpeed <- factor(set1$BallSpeed)
set1$EntryLine <- factor(set1$EntryLine)

class(set1$Flagstick)
```

```
## [1] "factor"
```

```
class(set1$BallSpeed)
```

```
## [1] "factor"
```

```
class(set1$EntryLine)
```

```
## [1] "factor"
```

b) (7 points) Bilder (2020) focused on a logistic regression model that included terms for flagstick, entry line, ball speed, and the interaction between flagstick and ball speed. Attempt to estimate this model using the observed counts as given in set1. Answer the following:

```
mod.fit.b1 <- glm(Success / Trials ~ Flagstick*BallSpeed + EntryLine, weights = Trials, data = set1,
                  family = binomial(link = "logit"))
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

i) What aspects of the output suggest there is a convergence problem?

The warning message stating that "glm.fit: fitted probabilities numerically 0 or 1 occurred" suggests convergence problems. It indicates that there may be complete seperation.

ii) Show that a larger number of iterations and stricter convergence criteria lead to different regression parameter estimates.

```
mod.fit.b2 <- glm(Success / Trials ~ Flagstick*BallSpeed + EntryLine, weights = Trials, data = set1,
                  family = binomial(link = "logit"), control = list(maxit = 50, epsilon = 1e-12))
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(mod.fit.b2)
```

```
##
## Call:
## glm(formula = Success/Trials ~ Flagstick * BallSpeed + EntryLine,
##     family = binomial(link = "logit"), data = set1, weights = Trials,
##     control = list(maxit = 50, epsilon = 1e-12))
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5650   0.0000   0.0000   0.0000   0.5057
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)              6.792e+01  4.263e+06       0        1
## FlagstickOut             1.974e-02  5.479e+06       0        1
## BallSpeed2              -3.377e+01  3.975e+06       0        1
## BallSpeed3              -3.635e+01  3.975e+06       0        1
## EntryLine2              -3.435e+01  2.667e+06       0        1
## EntryLine3              -3.596e+01  2.667e+06       0        1
## FlagstickOut:BallSpeed2  1.245e+00  5.479e+06       0        1
## FlagstickOut:BallSpeed3 -3.013e+01  6.094e+06       0        1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1729.2568  on 17  degrees of freedom
## Residual deviance:    2.8629  on 10  degrees of freedom
## AIC: 46.397
##
## Number of Fisher Scoring iterations: 30
```

```
# print table showing that estimates have changed
tibble("parameter" = names(coef(mod.fit.b1)),
       "default settings" = coef(mod.fit.b1),
                    "more iterations" = coef(mod.fit.b2),
       "difference" = `default settings` - `more iterations`) %>%
  knitr::kable()
```

parameter	default settings	more iterations	difference
(Intercept)	54.7384797	67.9170839	-13.1786042
FlagstickOut	-0.0000001	0.0197365	-0.0197366
BallSpeed2	-27.3770080	-33.7670956	6.3900876
BallSpeed3	-29.9641227	-36.3542103	6.3900876
EntryLine2	-27.5650612	-34.3535778	6.7885166
EntryLine3	-29.1707756	-35.9592922	6.7885166
FlagstickOut:BallSpeed2	1.2646529	1.2449163	0.0197366
FlagstickOut:BallSpeed3	-23.3243468	-30.1326000	6.8082532

From the table of parameter estimates, we can see that increasing the number of iterations and decreasing the convergence tolerance changed the parameter estimates.

c) (2 points) A small adjustment needs to be made to some data values so that a logistic regression model can be estimated properly. Add 0.5 to each 0 value for a number of successes. Subtract 0.5 to each 100 value for a number of successes. Below is the code to make this adjustment. Use the data with these adjustments for the remainder of the project.

```
const <- 0.5
Success2 <- ifelse(test = set1$Success == 0, yes = const, no = set1$Success)
```

```r
set1$Success2 <- ifelse(test = Success2 == 100, yes = 100 - const, no = Success2)
set1 %>% knitr::kable()
```

Flagstick	BallSpeed	EntryLine	Success	Trials	Success2
Out	1	1	100	100	99.5
Out	2	1	100	100	99.5
Out	3	1	81	100	81.0
Out	1	2	100	100	99.5
Out	2	2	73	100	73.0
Out	3	2	0	100	0.5
Out	1	3	100	100	99.5
Out	2	3	38	100	38.0
Out	3	3	0	100	0.5
In	1	1	100	100	99.5
In	2	1	100	100	99.5
In	3	1	100	100	99.5
In	1	2	100	100	99.5
In	2	2	45	100	45.0
In	3	2	7	100	7.0
In	1	3	100	100	99.5
In	2	3	14	100	14.0
In	3	3	0	100	0.5

d) (4 points) Bilder (2020) used the logistic regression model that included terms for flagstick, entry line, ball speed, and the interaction between flagstick and ball speed. Estimate and state this model. Use this model for the remainder of the project.

```r
mod.fit <- glm(Success2 / Trials ~ Flagstick*BallSpeed + EntryLine, weights = Trials,
               data = set1, family = binomial(link = "logit"))
```

```
## Warning in eval(family$initialize): non-integer #successes in a binomial glm!
```

```r
summary(mod.fit)
```

```
##
## Call:
## glm(formula = Success2/Trials ~ Flagstick * BallSpeed + EntryLine,
##     family = binomial(link = "logit"), data = set1, weights = Trials)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4268  -0.4509  -0.2132   0.1917   1.4161
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)               1.218e+01  1.015e+00  12.005  < 2e-16 ***
## FlagstickOut             -7.375e-15  1.161e+00   0.000   1.0000
## BallSpeed2               -6.136e+00  8.433e-01  -7.276 3.44e-13 ***
## BallSpeed3               -8.406e+00  8.930e-01  -9.414  < 2e-16 ***
## EntryLine2               -6.295e+00  5.846e-01 -10.768  < 2e-16 ***
## EntryLine3              -7.793e+00  6.073e-01 -12.832  < 2e-16 ***
## FlagstickOut:BallSpeed2  1.239e+00  1.183e+00   1.047   0.2950
## FlagstickOut:BallSpeed3 -2.325e+00  1.299e+00  -1.790   0.0734 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1657.224  on 17  degrees of freedom
## Residual deviance:   18.506  on 10  degrees of freedom
## AIC: 60.351
##
## Number of Fisher Scoring iterations: 7
```

Estimated model: $\text{logit}(\hat{\pi}) = 12.18 - 7.38 \times 10^{-15}\text{Flagstick}_{out} - 6.14\text{BallSpeed}_2 - 8.41\text{BallSpeed}_3 - 6.30\text{EntryLine}_2 - 7.79\text{EntryLine}_3 + 1.24(\text{Flagstick}_{out}, \text{BallSpeed}_2) - 2.33(\text{Flagstick}_{out}, \text{BallSpeed}_3)$

e) (18 points) The main explanatory variable of interest is flagstick. Complete the following to develop an interpretation of this variable.

i) State the odds of a success for when the flagstick is out of the hole. Do the same for when the flagstick is in the hole. Write these expressions using the population model.

Since we have an interaction in our model between flagstick and ballspeed, we need to examine the odds ratio of flagstick out vs in at the three different levels of ball speed. Assume entry line is fixed at center (it will cancel out when odds ratios are formed).

- Ball speed is low:
  - Flagstick out: $\exp\left(\beta_0 + \beta_1\right)$
  - Flagstick in: $\exp\left(\beta_0\right)$
- Ball speed is medium:
  - Flagstick out: $\exp\left(\beta_0 + \beta_1 + \beta_2 + \beta_6\right)$
  - Flagstick in: $\exp\left(\beta_0 + \beta_2\right)$
- Ball speed is high:
  - Flagstick out: $\exp\left(\beta_0 + \beta_1 + \beta_3 + \beta_7\right)$
  - Flagstick in: $\exp\left(\beta_0 + \beta_3\right)$

ii) State the odds ratio that compares the odds of a success for flagstick out to the odds of a success for flagstick in. Write this expression using the population model.

- Ball speed is low: $\hat{O}R_{low} = \frac{\exp(\beta_0+\beta1)}{\exp(\beta_0)} = \exp(\beta_1)$
- Ball speed is medium: $\hat{O}R_{med} = \frac{\exp(\beta_0+\beta_1+\beta_2+\beta_6)}{\exp(\beta_0+\beta_2)} = \exp(\beta_1 + \beta_6)$
- Ball speed is high: $\hat{O}R_{high} = \frac{\exp(\beta_0+\beta_1+\beta_3+\beta_7)}{\exp(\beta_0+\beta_3)} = \exp(\beta_1 + \beta_7)$

iii) Compute the appropriate estimated odds ratios and corresponding profile LR intervals needed to interpret flagstick. Apply a Bonferroni adjustment when calculating these intervals. Use a familywise error rate of 0.05.

```r
# create contrast matrix
K <- matrix(c(0, 1, 0, 0, 0, 0, 0, 0,
              0, 1, 0, 0, 0, 0, 1, 0,
              0, 1, 0, 0, 0, 0, 0, 1),
            nrow = 3, ncol = 8,
            byrow = TRUE)
# compute profile LR intervals on logit scale
# include bonferroni correction
linear.combo <- mcprofile(object = mod.fit, CM = K)
ci.log.or <- confint(linear.combo, level = 0.95, adjust = "bonferroni")

# output df of point and interval estimates converted to OR scale
comparisons <- c("out vs. in at low speed", "out vs. in at medium speed", "out vs. in at high speed")
data.frame(comparisons, exp(ci.log.or$estimate), exp(ci.log.or$confint)) %>%
  mutate(OR = paste(round(Estimate,2), " (", round(lower,2), ", ", round(upper,2), ")", sep = "")) %>%
  select(comparisons, OR) %>%
  knitr::kable()
```

comparisons	OR
out vs. in at low speed	1 (0.04, 25.23)
out vs. in at medium speed	3.45 (2.02, 6.02)
out vs. in at high speed	0.1 (0.02, 0.33)

iv) Interpret the confidence intervals.

With 95% confidence, the odds of success is between 0.04 and 25.23 times as large when the flagstick is out vs in at low ball speed. Since this interval includes 1, we cannot conclude that the flagstick affects the odds of success when the ball speed is low.

With 95% confidence, the odds of success is between 2.02 and 6.02 times as large when the flagstick is out vs in at medium ball speed. Since the lower bound of this interval is greater than 1, we can conclude that removing the flagstick improves the odds of success for medium speed putts.

With 95% confidence, the odds of success is between 0.02 and 0.33 times as large when the flagstick is out vs in at high ball speed. Since the upper bound of this interval is smaller than 1, we can conclude that removing the flagstick reduces the odds of success for high speed putts.

v) What aspect of these calculations and/or interpretations coincide with the definition of an interaction?

Interactions between categorical variables in logistic regression means that the log odds of success at a specific level for one covariate can change depending on the level of a seperate, interacting, covariate. In this specific case, the interaction between flagstick and ball speed means that the effect of flagstick on the odds of success in our model can change depending on ball speed. In this analysis we found that taking the flagstick out can have no effect, a positive effect, or a negative effect on odds of putting success depending on the level of ball speed.

f) (5 points) Perform a LRT to assess the importance of the interaction term in the model. Make sure to fully state the hypotheses for the test.

```r
Anova(mod.fit, test = "LR")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Success2/Trials
##                      LR Chisq Df Pr(>Chisq)
## Flagstick                6.15  1    0.01313 *
## BallSpeed             1066.72  2  < 2.2e-16 ***
## EntryLine              851.07  2  < 2.2e-16 ***
## Flagstick:BallSpeed     50.34  2  1.171e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$H_0 : \beta_6 = \beta_7 = 0$

$H_A : \beta_6 \text{ or } \beta_7 \neq 0$

There is strong statistical evidence that there is an interactive effect of flagstick and ball speed on the putting success ($p = 1.17 \times 10^{-11}$). This is consistent with out previous finding that the effect of flagstick on the odds of success can switch sign depending on ball speed.

g) (6 points) For the slightly off-center putts that approach the hole at a medium speed, complete the following.

h) State the probability of a success for when the flagstick is out of the hole. Do the same for when the flagstick is in the hole. Write these expressions using the population model.

- $\hat{\pi}_{\text{out,medium,slightly off-center}} = \text{logit}^{-1}(\beta_0 + \beta_1 + \beta_2 + \beta_4 + \beta_6)$
- $\hat{\pi}_{\text{in,medium,slightly off-center}} = \text{logit}^{-1}(\beta_0 + \beta_2 + \beta_4)$

ii) Estimate the probabilities of success when the flagstick is out and when the flagstick is in. If you use predict() for these calculations, you will need to specify the ball speed and entry line values are factor values rather than numerical. One way to do this is to use the as.factor() function when giving these variables their values.

```r
pred.g <- data.frame(Flagstick = c("Out", "In"),
                     BallSpeed = c("2", "2"),
                     EntryLine = c("2", "2"))
pred.g$`Est. Pr(Success)` <- predict(mod.fit, newdata = pred.g, type = "response")
pred.g %>% knitr::kable(digits = 2)
```

Flagstick	BallSpeed	EntryLine	Est. Pr(Success)
Out	2	2	0.73
In	2	2	0.44

- $\hat{\pi}_{\text{out,medium,slightly off-center}} = 0.73$
- $\hat{\pi}_{\text{in,medium,slightly off-center}} = 0.44$

iii) (3 points extra credit) Construct a 99.44% Wald confidence interval for the difference in the probabilities of success when the flagstick is out vs. when the flagstick is in. Notes about this part:

- The confidence level here corresponds to the confidence level needed for each of 9 Bonferroni-adjusted confidence intervals given in Table 4 of Bilder (2020).

- I used the deltaMethod() function from the car package. For the g argument of this function, I found it a little easier to write a probability of success from a general logistic regression model as 1 - 1/[1 + exp(0 + 1x)] rather than exp(0 + 1x)/[1 + exp(0 + 1x)]

```r
# get wald interval
parNames <-  paste(rep("b",8), 0:7, sep = "")
difference <- c("(1 - 1/(1 + exp(b0 + b1 + b2 + b4 + b6))) - (1 - 1/(1 + exp(b0 + b2 + b4)))")
CI.inout.wald <- deltaMethod(object = mod.fit, g. = difference,
                            parameterNames = parNames, level = 0.9944)
rownames(CI.inout.wald) <- NULL
# print table
CI.inout.wald %>%
  tibble::add_column(comparison = "$\\pi_{out} - \\pi_{in}$", .before = 1) %>%
  knitr::kable(digits = 2)
```

comparison	Estimate	SE	0.28 %	99.72 %
$\pi_{out} - \pi_{in}$	0.29	0.05	0.15	0.43

We are 99.44% confident that $0.15 < \pi_{out} - \pi_{in} < 0.43$ when putts are medium speed and slightly off-center.

h) (3 points) Bilder (2020) present two different analyses using methods from Chapters 1 and 2 of our book. Provide one advantage of each analysis method over the other.

Analysis 1 does not require adding pseudo-observations to the data.

Analysis 2 allows us to more easily interpret the effects of the explanatory variables on putting success. That is, instead of calculating seperate probabilities for each combination of factor levels, we can also look at main effects where appropriate (entry line).

However, both approaches yield the same conclusions about flagstick placement for each of the scenarios which is reassuring.

i) (3 points) Page 59 of Bilder (2020) states the following conclusions based on the analysis of the data from EMGA:

> In or out? The conclusion depends on ball speed and entry line. For low-speed putts, there is not sufficient evidence that flagstick placement matters. For putts reaching the hole at a medium speed, putting with the flagstick out is the better strategy for off-center putts (not enough evidence either way for on-center putts). For putts reaching the hole at a high speed, leaving the flagstick in is better for putts that would hit the center of the flagstick, suggesting a similar effect as observed for bank shots in basketball. For high-speed, off-center putts, the evidence is not as strong that leaving the flagstick in is the better strategy.

Using Analysis #2 detailed in the paper, why were these conclusions reached? Fully explain your answer.

For low speed putts, the adjusted 95% confidence interval for the odds of of success comparing flagstick out vs in included 1 ($0.04 < \text{OR} < 25.2$) which indicates there is no statistically significant evidence that flagstick placement changes the odds of success. That is why the conclusion for this is “there is not sufficient evidence that flagstick placement matters.”

For medium speed putts, the lower bound of the adjusted 95% confidence interval for the odds of of success comparing flagstick out vs in was greater than 1 ($2.02 < \text{OR} < 6.02$). This is why there is sufficient evidence that taking the flagstick out is the better strategy for medium speed putts.

For high speed putts, the upper bound of the adjusted 95% interval for the odds ratio is below 1 ($0.02 < \mathrm{OR} < 0.33$). That means the odds of success when the flagstick in is greater than when the flagstick is out. That is why the conclusion is “leaving the flagstick in is better for putts that would hit the center of the flagstick, suggesting a similar effect as observed for bank shots in basketball.”