Project 2

David Nguyen

2021-03-02

```r
# packages used
library(dplyr)
library(car)
library(mcprofile)
library(boot) # for inv.logit
```

Flagstick in or out? Edoardo Molinari putting experiment.

1) (47 total points) This problem is a continuation of the setting for #2 on Project #1. The entire data set used for Bilder (2020) is used now for this problem. This data is available in the flagstick.csv file on the graded materials web page of the course website. Below is the data:

```r
# read in full golf data set
set1 <- read.csv("data/flagstick.csv")
head(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

where

- Flagstick: In or out
- BallSpeed: The ball speed at the hole is coded as 1 = low, 2 = medium, 3 = high
- EntryLine: The entry line of the ball corresponding to the flagstick : 1 = center, 2 = slightly off-center, 3 = grazing
- Success: Number of observed successes
- Trials: Number of trials

For this project, treat BallSpeed and EntryLine as ordinal in nature so that they contribute only one term to any model. Complete the following.

a) (2 points) To correspond with the material discussed so far in our course, we will need to create a new form of the flagstick variable so that is has values of 0 and 1. Run the following code to add a new variable named FlagstickOut and use this variable in place of Flagstick for the remainder of the project. Print the data frame to verify the code worked.

```r
# dummy coding for flagstick
set1$FlagstickOut <- ifelse(test = set1$Flagstick == "Out", yes = 1, no = 0)
set1 %>% knitr::kable()
```

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

b) (4 points) Estimate and state the logistic regression model that includes flagstick, ball speed, and entry line in the model as linear terms. Use this model for the remainder of the project.

```r
# BallSPeed and EntryLine as numeric
mod.fit <- glm(Success/Trials ~ FlagstickOut + BallSpeed + EntryLine,
    weights = Trials, family = binomial(link = "logit"), data = set1)
summary(mod.fit)
```

```
##
## Call:
## glm(formula = Success/Trials ~ FlagstickOut + BallSpeed + EntryLine,
##     family = binomial(link = "logit"), data = set1, weights = Trials)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -8.4893  -0.7154   0.3609   1.1759   8.2563
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   16.1414     0.7726  20.894   <2e-16 ***
## FlagstickOut   0.4101     0.1790   2.291    0.022 *
## BallSpeed     -3.9746     0.2031 -19.567   <2e-16 ***
## EntryLine     -3.3164     0.1715 -19.336   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1729.26  on 17  degrees of freedom
## Residual deviance:  237.78  on 14  degrees of freedom
## AIC: 273.32
##
## Number of Fisher Scoring iterations: 6
```

$logit(\hat{\pi}) = 16.1414 + 0.4101(\text{Flagstick}_{out}) - 3.9746(\text{BallSpeed}) - 3.3164(\text{EntryLine})$

c) For each explanatory variable, complete the following items. Use a one-unit decrease for the ball speed and entry line variables when computing their odds ratios and associated intervals.

- i) (4 points) Estimate the odds ratio

```r
ORhat_fout <- exp(mod.fit$coefficients[2]) %>% as.numeric() %>% round(2) # exp(beta_1)
ORhat_bs <- exp(-mod.fit$coefficients[3]) %>% as.numeric() %>% round(2) # 1/exp(beta_2)
ORhat_el <- exp(-mod.fit$coefficients[4])  %>% as.numeric() %>% round(2) # 1/exp(beta_3)
```

The estimated odds of a success are $e^{\hat{\beta}_1} = 1.51$ as large when the flagstick is taken out compared to when the flagstick is left in and all other variables are fixed.

The estimated odds of a success are $e^{-\hat{\beta}_2} = 53.23$ as large for every one-unit decrease in ball speed when all other variables are fixed.

The estimated odds of a success are $e^{-\hat{\beta}_3} = 27.56$ as large for every one-unit decrease in entry line when all other variables are fixed.

- ii) (4 points) Compute a profile LR interval for the odds ratio.

```r
# create contrast matrix
K <- matrix(c(0, 1, 0, 0,
              0, 0, -1, 0,
              0, 0, 0, -1),
            nrow = 3, ncol = 4,
            byrow = TRUE)
# compute profile LR intervals on logit scale
linear.combo <- mcprofile(object = mod.fit, CM = K)
ci.log.or <- confint(linear.combo, level = 0.95, adjust = "none")

# output df of intervals converted to OR scale
comparisons <- c("Flag out vs. in", "unit decrease ball speed", "unit decrease entry line")
data.frame(comparisons, OR = exp(ci.log.or$confint)) %>% knitr::kable(digits = 2)
```

comparisons	OR.lower	OR.upper
Flag out vs. in	1.06	2.15
unit decrease ball speed	36.29	80.53
unit decrease entry line	19.89	38.97

- iii) (6 points) Interpret the profile LR interval.

With 95% confidence, the odds of success are between 1.06 and 2.15 times as large when the flagstick is removed compared to when it is left in when ball speed and line of entry are fixed. The lower bound of the interval is close to 1, which indicates there is marginal evidence that removing the flagstick increases the odds of success.

With 95% confidence, the odds of success are between 36.29 and 80.53 times as large when the ball speed is decreased by one unit (i.e., high to medium or medium to low) when flag stick and entry line are fixed. The interval clearly excludes one which provides strong evidence that decreased ball speed improves the odds of success.

With 95% confidence, the odds of success are between 36.29 and 80.53 times as large when the entry line is decreased by one unit (i.e., from grazing to slightly off-center or slightly off-center to center) when ball speed and flag stick are fixed. The interval clearly excludes one which provides strong evidence that putting toward the center of the hole improves the odds of success compared to putting towards an edge.

- iv) (6 points) Perform LRTs to evaluate the importance of each explanatory variable in the model. Make sure to fully state the hypotheses in a symbolic form.

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

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Success/Trials
##              LR Chisq Df Pr(>Chisq)
## FlagstickOut     5.31  1     0.0212 *
## BallSpeed     1066.22  1     <2e-16 ***
## EntryLine      788.99  1     <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Flag stick:

$H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$

There is statistical evidence ($\Lambda = 5.31$, $p = 0.0212$) that the presence of a flag stick in the hole has an effect on the probability of success given that ball speed and entry line are in the model.

Ball Speed:

$H_0 : \beta_2 = 0$ vs $H_1 : \beta_2 \neq 0$

There is strong statistical evidence ($\Lambda = 1066.22$, $p < 2 \times 10^{-16}$) that ball speed has an effect on the probability of success given that flag stick and entry line are in the model.

Entry line

$H_0 : \beta_3 = 0$ vs $H_1 : \beta_3 \neq 0$

There is strong statistical evidence ($\Lambda = 788.99$, $p < 2 \times 10^{-16}$) that line of entry has an effect on the probabilty of success given that ball speed and flag stick are in the model.

d) (4 points) Estimate the probability of success for slightly off-center putts that approach the hole at a medium speed when the flagstick is out. Compute the corresponding 95% confidence interval for the probability of success using the appropriate interval. Interpret the interval.

```
# LR interval calculation
# get CI and estimate on logit link scale
K <- matrix(c(1, 1, 2, 2), nrow = 1, ncol = 4, byrow = TRUE)
linear.combo <- mcprofile(mod.fit, CM = K)
ci.log.or <- confint(linear.combo, level = 0.95, adjust = "none")

# inv-logit transform to get probabilities
pi.hat.d <- inv.logit(unlist(ci.log.or$estimate)) %>% as.numeric() %>% round(2)
ci.pi.hat.d <- inv.logit(unlist(ci.log.or$confint)) %>% as.numeric() %>% round(2)
```

$\hat{\pi}_{\text{out,medium,slightly off-center}} = 0.88$

With 95% confidence, the probability of success is between 0.84 and 0.91for slightly off-center putts that approach the hole at a medium speed when the flagstick is out. This means that putts made under this condition have a high probability of successs.

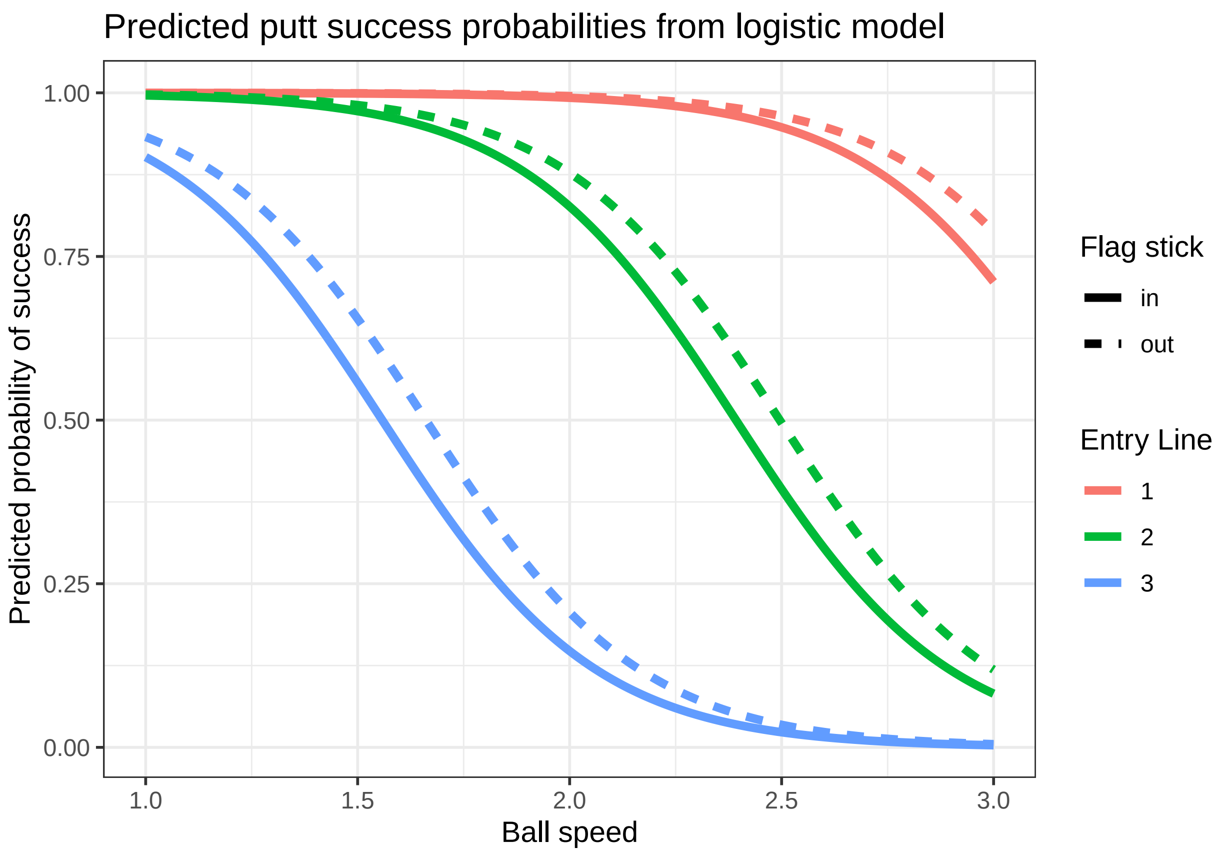
```
# compute Wald interval for comparison to LR interval
wald.d <- wald(linear.combo)
wald.ci.d <- confint(wald.d, level = 0.95, adjust = "none")
all.equal(ci.log.or$confint, wald.ci.d$confint) # yes they are pretty much the same
```

```
## [1] "Component \"lower\": Mean relative difference: 0.005228336"
## [2] "Component \"upper\": Mean relative difference: 0.004204231"
```

e) (8 points) Plot the estimated probability of success from the model with ball speed on the x-axis. Include separate lines on the plot for the different combinations of flagstick and entry line. These lines should have different colors (and line types too, if needed) and a legend corresponding to these lines needs to be included. Interpret the plot relative to your results from the odds ratios and the LRTs.

```
# create df of predicted probabilities from model
new.data <- expand.grid(BallSpeed = seq(1,3, by = 0.01),
                        EntryLine = 1:3,
                        FlagstickOut = 0:1)
new.data$prediction <- predict(mod.fit, newdata = new.data, type = "response")
new.data <- new.data %>% mutate(FlagStick = ifelse(FlagstickOut == 1, "out", "in"))

# plot
new.data %>%
  ggplot(aes(x = BallSpeed)) +
  geom_line(aes(y = prediction,
                group = interaction(FlagStick, factor(EntryLine)),
                linetype = FlagStick, col = factor(EntryLine)),
            size = 1.5) +
  labs(title = "Predicted putt success probabilities from logistic model",
       y = "Predicted probability of success",
       x = "Ball speed",
       col = "Entry Line",
       linetype = "Flag stick") + theme_bw()
```



The plotted model predictions are consistent with statistical inferences shown previously:

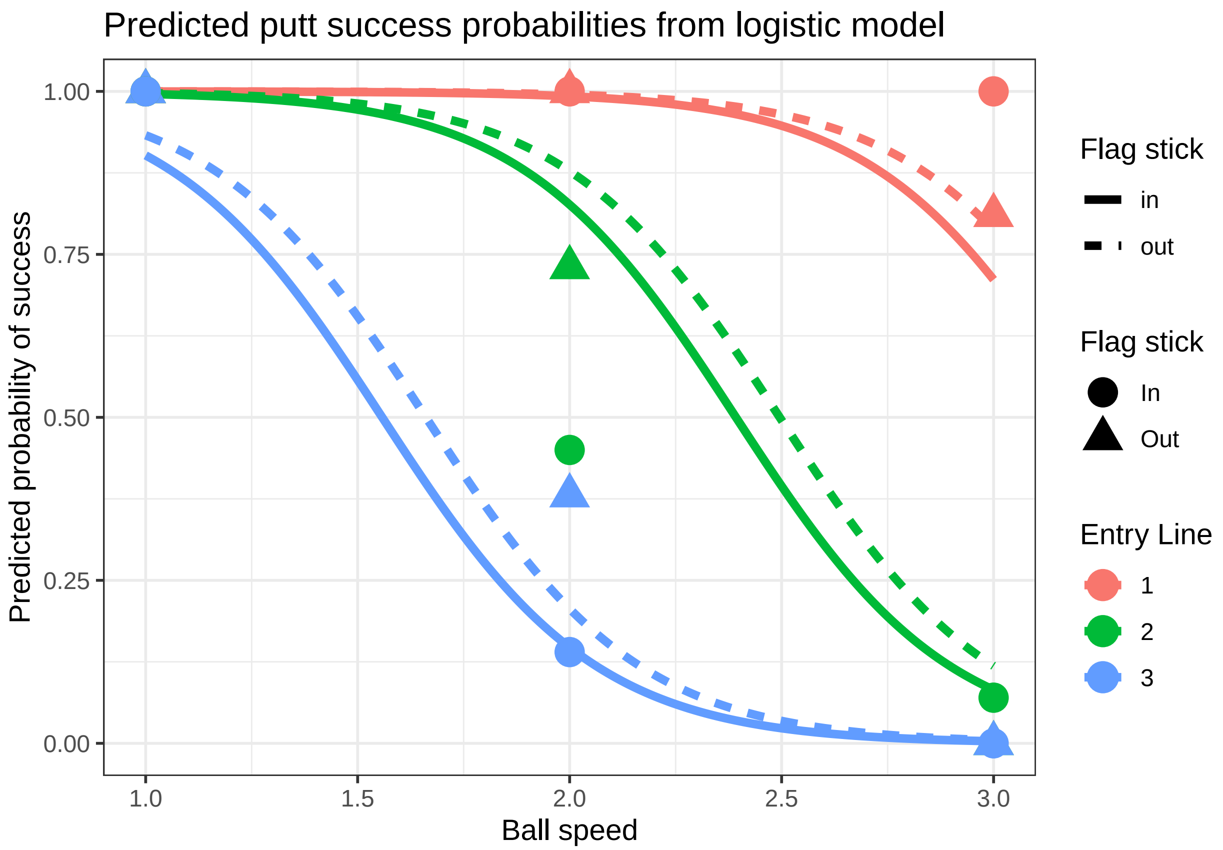
- the probability of success decreases with ball speed
- the probability of success is lower when the flag stick is in than when it is out
- the probability of success decreases the further off-center the line of entry is
- the effects of ball speed and entry line are larger than the effect of the whether the flag stick is in or not

f) (2 points) Pages 2.65 – 2.70 of the course notes discuss a bubble plot as a way to evaluate how well the model fits the data when the there is a binomial response variable. Why would the bubble aspect of this plot not be useful to evaluate the model for this problem? Explain.

The reason that a bubble plot is not useful for these data is that the number of trials is 100 for each covariate pattern which means all the bubbles would be the same size. Since the main purpose of a bubble plot is to provide a visual weighting of how influential an observed proportion should be relative to the number of trials for that covariate pattern, a bubble plot will not be useful for our data.

g) (4 points) Add the observed proportions for each explanatory variable combination to the plot in part e). One way to add these proportions to the plot is by using the points() function. Choose appropriate colors for these plotting characters when representing these proportions that enable direct comparison to lines plotted for the model. Assess the fit of the model relative to these observed proportions.

```r
new.data %>%
  ggplot(aes(x = BallSpeed)) +
  geom_line(aes(y = prediction,
                group = interaction(FlagStick, factor(EntryLine)),
                linetype = FlagStick, col = factor(EntryLine)),
            size = 1.5) +
  geom_point(data = set1,
             mapping = aes(x = BallSpeed, y = Success/Trials,
                           group = interaction(Flagstick, factor(EntryLine)),
                           shape = Flagstick, col = factor(EntryLine)), size = 5) +
  labs(title = "Predicted putt success probabilities from logistic model",
       y = "Predicted probability of success",
       x = "Ball speed",
       col = "Entry Line",
       linetype = "Flag stick",
       shape = "Flag stick") +
  theme_bw()
```



While basic aspects of the data are captured by the model, the model predictions appear to be a poor fit for specific covariate patterns. For instance, when the ball entry line is centered and ball speed is high (EntryLine = 1, BallSpeed = 3), the model predictions are too low compared to the observed proportions. When the entry line is slightly off-center and the ball speed is medium (EntryLine = 2, BallSpeed = 2) the model underestimates the difference between leaving the flag stick in vs out and overestimates the probability of success when the flagstick is in. The model predictions are fairly consistent with the observed data when the ball grazes the flag stick.

h) (3 points) Should one take the flagstick out or leave it in the hole when putting? Use your statistical inference results from this problem ONLY to explain your answer. Assume the model fits the data well.

The flag stick should be removed before putting. All else being equal (entry line and ball speed) the odds of success are between 1.06 and 2.15 times as large when the flag stick is removed compared to when it is left in. While the difference may be marginal, especially for lower speed and on-center putts, the odds of success are always predicted to be better when the flag stick is removed. Assuming the "cost" of removing the flag stick is very small, the golfer might as well remove the flag stick.