W271 Assignment 2 Solution

Contents


```
library(tidyverse)
library(sandwich)
library(lmtest)
library(Hmisc)
library(car)
library(stargazer)
```

1 Placekicking Data: Binary Logistic Regression (3 points – one for each sub-question)

Does the strategy of *icing the kicker* reduce the probability of success for a field goal? The idea is this: In American football, there is a play where a person kicks the ball through the uprights to score points. This is a high-pressure event, and there is a theory that making the kicker stand on the field and think about it will make the kicker nervous, and so make them more likely to miss their attempt.

```
pk <- read_csv('./data/placekick.BW.csv')

str(pk)
```

```
## spc_tbl_ [2,003 x 10] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ GameNum    : chr [1:2003] "2002-0101" "2002-0101" "2002-0101" "2002-0101" ...
##  $ Kicker     : chr [1:2003] "Bryant" "Bryant" "Cortez" "Cortez" ...
##  $ Good       : chr [1:2003] "Y" "Y" "N" "Y" ...
##  $ Distance   : num [1:2003] 29 33 25 23 48 33 36 34 45 48 ...
##  $ Weather    : chr [1:2003] "Sun" "Sun" "Sun" "Sun" ...
##  $ Wind15     : num [1:2003] 0 0 0 0 0 0 0 0 0 0 ...
##  $ Temperature: chr [1:2003] "Nice" "Nice" "Nice" "Nice" ...
##  $ Grass      : num [1:2003] 1 1 1 1 1 1 1 0 0 0 ...
##  $ Pressure   : chr [1:2003] "N" "N" "N" "N" ...
##  $ Ice        : num [1:2003] 0 0 0 0 0 0 0 0 0 0 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   GameNum = col_character(),
##   ..   Kicker = col_character(),
##   ..   Good = col_character(),
##   ..   Distance = col_double(),
##   ..   Weather = col_character(),
##   ..   Wind15 = col_double(),
##   ..   Temperature = col_character(),
##   ..   Grass = col_double(),
##   ..   Pressure = col_character(),
##   ..   Ice = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

1.1 Linear Model, with Linear Effects

Use the `distance`, `weather`, `wind15`, `temperature`, `grass`, `pressure` and `ice` as explanatory variables in a logistic regression model that predicts success. Estimate the model, and interpret each of the indicator variables that are used in the model.

```
# Examine the data structure
#str(pk)
#describe(pk)

levels(factor(pk$Weather))
```

```
## [1] "Clouds"   "Inside"   "SnowRain" "Sun"
```

```
mod.glm1 = glm(formula = factor(Good) ~ Distance + factor(Weather) + Wind15
               + factor(Temperature) + Grass + factor(Pressure) + Ice,
               family = binomial(link = logit), data = pk)
summary(mod.glm1)
```

```
##
## Call:
## glm(formula = factor(Good) ~ Distance + factor(Weather) + Wind15 +
##     factor(Temperature) + Grass + factor(Pressure) + Ice, family = binomial(link = logit),
##     data = pk)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6804   0.2599   0.4360   0.7148   1.8698
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                5.740185   0.369597  15.531   <2e-16 ***
## Distance                  -0.109600   0.007188 -15.249   <2e-16 ***
## factor(Weather)Inside     -0.083030   0.214711  -0.387   0.6990
## factor(Weather)SnowRain   -0.444193   0.217852  -2.039   0.0415 *
## factor(Weather)Sun        -0.247582   0.139642  -1.773   0.0762 .
## Wind15                    -0.243777   0.175527  -1.389   0.1649
## factor(Temperature)Hot     0.250013   0.247540   1.010   0.3125
## factor(Temperature)Nice    0.234932   0.181461   1.295   0.1954
## Grass                     -0.328435   0.160050  -2.052   0.0402 *
## factor(Pressure)Y          0.270174   0.262809   1.028   0.3039
## Ice                       -0.876133   0.451251  -1.942   0.0522 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2104.0  on 2002  degrees of freedom
## Residual deviance: 1791.3  on 1992  degrees of freedom
## AIC: 1813.3
##
## Number of Fisher Scoring iterations: 5
```

$$
\begin{aligned}
&logit(\hat{\pi}(Good)) \\
&\quad = 5.74 - 0.11 Distance - 0.08 WeatherInside - 0.44 WeatherSnowRain - 0.25 WeatherSun \\
&\quad\quad - 0.24 Wind15 + 0.25 TemperatureHot + 0.23 TemperatureNice - 0.33 Grass + 0.27 PressureY \\
&\quad\quad - 0.88 Ice
\end{aligned}
$$

First, notice that only coefficients of `distance`, `(Weather)SnowRain`, and `Grass` are statistically significant (associated with a p-value < 0.05), so our model suggests that these variables do in fact, influence the probability of success for a field goal. And because they are negative numbers, we can say that they decrease the probability of success for a field goal.

One-yard increase is associated with 10.9% decrease in log-odds. The log-odds are smaller when `Weather = SnowRain` or `Grass = 1` compared to the reference group.

1.2 Sun Shine Daydream

The authors use the `Weather==Sun` as the base level category for `Weather`. This is not the default that R uses. Change either the data, or how you estimate the model so that `Weather==Sun` is the base category and other types of weather are the contrasts. Interpret the results.

```
# Examine the current level in the Weather variable
levels(factor(pk$Weather))
```

```
## [1] "Clouds"   "Inside"   "SnowRain" "Sun"
```

```
# Relevel the variable using factor() function
pk$Weather = factor(pk$Weather, levels = c("Sun", "Clouds", "Inside", "SnowRain"),
                    labels = c("Sun", "Clouds", "Inside", "SnowRain"))
# Re-estimate the logistic regression, calling it mod.glm1b
mod.glm1b <- glm(formula = factor(Good) ~ Distance + Weather + Wind15 +
                   factor(Temperature) + Grass + factor(Pressure) + Ice,
                 family = binomial(link = logit), data = pk)
summary(mod.glm1b)
```

```
##
## Call:
## glm(formula = factor(Good) ~ Distance + Weather + Wind15 + factor(Temperature) +
##     Grass + factor(Pressure) + Ice, family = binomial(link = logit),
##     data = pk)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6804   0.2599   0.4360   0.7148   1.8698
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                5.492602   0.370141  14.839   <2e-16 ***
## Distance                  -0.109600   0.007188 -15.249   <2e-16 ***
## WeatherClouds              0.247582   0.139642   1.773   0.0762 .
## WeatherInside              0.164553   0.215062   0.765   0.4442
## WeatherSnowRain           -0.196611   0.219015  -0.898   0.3693
## Wind15                    -0.243777   0.175527  -1.389   0.1649
## factor(Temperature)Hot     0.250013   0.247540   1.010   0.3125
## factor(Temperature)Nice    0.234932   0.181461   1.295   0.1954
## Grass                     -0.328435   0.160050  -2.052   0.0402 *
## factor(Pressure)Y          0.270174   0.262809   1.028   0.3039
## Ice                       -0.876133   0.451251  -1.942   0.0522 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2104.0  on 2002  degrees of freedom
## Residual deviance: 1791.3  on 1992  degrees of freedom
## AIC: 1813.3
##
## Number of Fisher Scoring iterations: 5
```

In this model, also the coefficients of `distance`, and `Grass` are statistically significant with the same sign and interpretation.

1.3 Likelihood Ratio Tests

Perform likelihood ratio tests for all explanatory variables to evaluate their importance within the model. Discuss and interpret the results of these tests.

Let's use our original model, *mod.glm1*, for thie exercise.

```
library(car)
# Conduct LRTs on all of the explanatory variables
Anova(mod.glm1, test="LR" )
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: factor(Good)
##                      LR Chisq Df Pr(>Chisq)
## Distance              294.341  1    < 2e-16 ***
## factor(Weather)         5.670  3    0.12884
## Wind15                  1.898  1    0.16833
## factor(Temperature)     1.723  2    0.42254
## Grass                   4.314  1    0.03781 *
## factor(Pressure)        1.088  1    0.29682
## Ice                     3.698  1    0.05448 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Estimate the Profile likelihood C.I.
mod.glm1.ci <- confint(object = mod.glm1, level = 0.95)
```

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

```
# Print Profile likelihood C.I. for the estimated odds ratios
exp(mod.glm1.ci)
```

```
##                                2.5 %      97.5 %
## (Intercept)              152.9569791 651.7836742
## Distance                   0.8834276   0.9086871
## factor(Weather)Inside      0.6040049   1.4026042
## factor(Weather)SnowRain    0.4199751   0.9877853
## factor(Weather)Sun         0.5932796   1.0259643
## Wind15                     0.5573810   1.1100160
## factor(Temperature)Hot     0.7916314   2.0917846
## factor(Temperature)Nice    0.8823269   1.7986655
## Grass                      0.5239985   0.9818637
## factor(Pressure)Y          0.7933060   2.2306039
## Ice                        0.1721834   1.0172552
```

From a statistical significance perspective, the variables *Distance*, *Grass*, and *Ice* are all significant, thought *Ice* is only marginally significant.

The other p-values are all greater than 0.10, where Weather and Wind15 are somewhat closer to 0.10 than Pressure and Temperature. We can say for these four variables that there is not sufficient evidence that they affect the probability of success for a field goal when $\alpha = 0.05$.

It is important to note that each hypothesis test is conditional on the other variables remaining in the model.

1.4 Should you kick or not?

Suppose that you are trying to make an assessment about whether to kick a field goal in *The Game* – the annual rivalry game played between the Cal Bears and Stanford ... (What is their mascot? A tree?)

Suppose that Cal is down by two points (so `Pressure = Y`), that the distance is 35 yards, and that it is a typical autumn evening in Berkeley, so `Wind15 = 0`, `Weather=Sun`, and `Temperature=Nice`. Cal plays on a turf stadium, and Stanford is out of timeouts, so cannot ice the kicker. What are the chances that Cal makes the kick? Compute the 95% confidence interval

```
alpha = 0.5

## Create the dataframe
data <- data.frame(Distance = 35, Weather = "Sun", Wind15 = 0,
                   Temperature = "Nice", Grass = 0 , Pressure = "Y", Ice = 0)

# Obtain the linear predictor
linear.pred = predict(object = mod.glm1, newdata = data, type = "link", se = TRUE)

# Then, compute pi.hat
pi.hat = exp(linear.pred$fit)/(1+exp(linear.pred$fit))
#pi.hat

# Compute Wald Confidence Interval (in 2 steps)
# Step 1: compute the CI of linear predictor
CI.lin.pred = linear.pred$fit + qnorm(p = c(alpha/2, 1-alpha/2))*linear.pred$se.fit
#CI.lin.pred

# Step 2: compute the CI of probability of succcess
CI.pi = exp(CI.lin.pred)/(1+exp(CI.lin.pred))
#CI.pi

# Store all the components in a data frame
#str(predict.data)
round(data.frame(pi.hat, lower=CI.pi[1], upper=CI.pi[2]),2)
```

```
##   pi.hat lower upper
## 1    0.9  0.88  0.91
```

The 95% Wald confidence interval for the probability that Cal makes the kick is between 0.88 and 0.91, So the probability of success for the kick is quite high when Cal is down (`Pressure = Y`), plays on a turf stadium, the distance is 35 yards, `Wind15 = 0`, `Weather=Sun`, `Temperature=Nice`, and cannot ice the kicker.

2 Binary Logistic Regression (5 points – one for each sub-question)

For this question, we use the Mroz dataset from *car* library to study factors that are related to married female participation in the labor market.

```
glimpse(Mroz)
```

```
## Rows: 753
## Columns: 8
## $ lfp  <fct> yes, yes, yes, yes, yes, yes, yes, yes, yes, yes, yes, yes, yes, ~
## $ k5   <int> 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ k618 <int> 0, 2, 3, 3, 2, 0, 2, 0, 2, 2, 1, 1, 2, 2, 1, 3, 2, 5, 0, 4, 2, 0,~
## $ age  <int> 32, 30, 35, 34, 31, 54, 37, 54, 48, 39, 33, 42, 30, 43, 43, 35, 4~
## $ wc   <fct> no, no, no, no, yes, no, yes, no, no, no, no, no, no, no, no, no,~
## $ hc   <fct> no, no, no, no, no, no, no, no, no, no, no, yes, yes, no, yes, no~
## $ lwg  <dbl> 1.2101647, 0.3285041, 1.5141279, 0.0921151, 1.5242802, 1.5564855,~
## $ inc  <dbl> 10.910001, 19.500000, 12.039999, 6.800000, 20.100000, 9.859000, 9~
```

In this dataset, lfp is a binary variable indicating labor force participation by a married woman during 1975. lfp is equal to one if the woman reports working for a wage outside the home during the year and zero otherwise. We assume that married female labor force participation depends on the following seven potential explanatory variables included in this data set:

- k5: number of kids below the age of 5
- k18: number of kids between 6 and 18
- age: wife's age (in years)
- wc: wife's college attendance
- hc: husband's college attendance
- lwg: log of wife's estimated wage rate
- inc: family income excluding the wife's wage ($1000)

2.1 Estimate a binary logistic regression

Estimate a binary logistic regression with `lfp`, which is a binary variable recoding the participation of the females in the sample, as the dependent variable. The set of explanatory variables includes `age`, `inc`, `wc`, `hc`, `lwg`, `totalKids`, and a quadratic term of `age`, called `age_squared`, where `totalKids` is the total number of children up to age 18 and is equal to the sum of `k5` and `k618`.

> We first create a new varirables, such as the total number of kids and the quadratic term of age. Then, we estimate a binary logistic regression using the `glm()` function and display the estimation result.

```
# Create new explanatory variables

# Total number of kids
Mroz['totalKids'] <- Mroz$k5 + Mroz$k618
# Quadratic term of age (i.e. age squared)
Mroz['age_squared'] <- Mroz$age^2

# Estimate a bineary logistic regression with the variables specified in the questions
mroz.glm1 <- glm(lfp ~ age + age_squared + inc + wc+ hc + lwg + totalKids,
                 family = 'binomial', data = Mroz)


# Note that another way to include a quadratic term is to include
```

```r
#the transformation in the glm() function directly:

#glm(lfp ~ age + I(age^2) + inc + wc + hc+  lwg + totalKids, family = 'binomial', data = Mroz)

# Display the estimation results
summary(mroz.glm1)
```

```
##
## Call:
## glm(formula = lfp ~ age + age_squared + inc + wc + hc + lwg +
##     totalKids, family = "binomial", data = Mroz)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8342  -1.1669   0.6773   1.0079   2.0614
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.294073   2.281551  -2.320 0.020320 *
## age          0.318014   0.109463   2.905 0.003670 **
## age_squared -0.004114   0.001272  -3.233 0.001224 **
## inc         -0.034561   0.007922  -4.363 1.28e-05 ***
## wcyes        0.666013   0.218074   3.054 0.002258 **
## hcyes        0.098260   0.198970   0.494 0.621417
## lwg          0.549976   0.145506   3.780 0.000157 ***
## totalKids   -0.222490   0.063849  -3.485 0.000493 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1029.75  on 752  degrees of freedom
## Residual deviance:  952.02  on 745  degrees of freedom
## AIC: 968.02
##
## Number of Fisher Scoring iterations: 4
```

Using the usual z statistics, all variables except `hcyes` are statistically significant. The positive coefficient of `age` and negative coefficient of 'age_squared' indicates that the probability of labor force participation by a married woman increase with an increase in age but with a diminishing rate.

The negative coefficients of 'inc' and 'totalKids' imply a negative association between family income and the number of kids and women's probability of labor force participation.

The positive coefficient of `lwg` mean that wife's estimated wage rate is positively correlated with the probability of labor force participation by women. Also, the positive coefficient of `wcyes` indicates a higher probability of labor force participation for educated females.

2.2 Evaluate statistical significance

Is the age effect statistically significant?

To test the statistical significance of the age effect, we will apply LRT using R's `anova()` function, and to do so, we will estimate a "restricted" model with the age variables, which include both `age` and `age_squared` in the "full" model. We will call the restricted model `mroz.glm2`. Note also that because age is entered the logistic regression as a quadratic function, testing the statistical significance of the age effect include testing multiple hypotheses.

The model being estimated, surpressing the subscript for individuals, is

$$log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 age + \beta_2 age\_squared + \beta_3 inc + \beta_4 wc + \beta_5 hc + \beta_6 lwg + \beta_7 totalKids$$

where π denotes the probability that a female participating in the labor force. That is, $P(lfp_i = 1)$

$$H_0 : \beta_1 = 0 \text{ and } \beta_2 = 0$$

$$H_1 : (\beta_1 \neq 0 \text{ and } \beta_2 = 0), \text{ or } (\beta_1 = 0 \text{ and } \beta_2 \neq 0), \text{ or } (\beta_1 \neq 0 \text{ and } \beta_2 \neq 0)$$

Note: I just explicitly write out all the alternative hypotheses. In most case, the following expression is being used

$$H_0 : \beta_1 = 0 \text{ and } \beta_2 = 0$$

$$H_1 : H_0 \text{ is not true}$$

```
mroz.glm2 <- glm(lfp ~ inc + wc + hc + lwg + totalKids, family = 'binomial', data = Mroz)
# Display both Model 1 and Model 2
stargazer(mroz.glm1, mroz.glm2, type = 'text')
```

```
##
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                                 lfp
##                           (1)            (2)
## ------------------------------------------------
## age                     0.318***
##                         (0.109)
##
## age_squared            -0.004***
##                         (0.001)
##
## inc                    -0.035***      -0.035***
##                         (0.008)        (0.008)
##
## wcyes                   0.666***       0.645***
##                         (0.218)        (0.215)
##
## hcyes                    0.098          0.117
##                         (0.199)        (0.194)
##
## lwg                     0.550***       0.583***
```

```
##                       (0.146)        (0.145)
##
## totalKids            -0.222***       -0.087*
##                       (0.064)        (0.053)
##
## Constant              -5.294**        0.263
##                       (2.282)        (0.226)
##
## ------------------------------------------------
## Observations            753             753
## Log Likelihood        -476.011       -486.040
## Akaike Inf. Crit.     968.022         984.080
## ================================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

```
# Apply LRT
anova(mroz.glm1, mroz.glm2, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: lfp ~ age + age_squared + inc + wc + hc + lwg + totalKids
## Model 2: lfp ~ inc + wc + hc + lwg + totalKids
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1       745     952.02
## 2       747     972.08 -2  -20.057 4.412e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using $\alpha = 0 : 05$, we would reject the null hypothesis. Thus, age has a statistically significant relationship with the probability of labor force participation by women.

2.3 Interpret an effect

What is the effect of a decrease in age by 5 years on the odds of labor force participation for a female who was 45 years of age.

Recall our model:

$$log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 age + \beta_2 age\_squared + \beta_3 inc + \beta_4 wc + \beta_5 hc + \beta_6 lwg + \beta_7 totalKids$$

The odds ratio for an increase in age by 5 is expressed in the following formula:

$$OR = exp(5\beta_1 + 5\beta_2(2 \times age + 5))$$

which depends on the level of age.

Let's compute the numerical change of the odds ratio by inserting the estimates to the formula above from the model stored in mroz.glm1, which is used here because we have tested that the age effect is significant.

```
c = -5
age = 45

OR.change = exp(c*(coefficients(mroz.glm1)[['age']] +
                     coefficients(mroz.glm1)[['age_squared']]*(2*age + c)))

OR.change
```

```
## [1] 1.171602
```

Therefore, the estimated odds of labor force participation (lfp) of females who are 45 years of age increase by 1.18 times for five years increase in age.

2.4 Construct a confidence interval

Estimate the 95% profile likelihood confidence interval of the probability of labor force participation for females who were 40 years old, had income equal to 20, did not attend college, her husband attend college, had log wage equal to 1, and did not have children.

```
library(mcprofile)

# Define the contrast matrix
K = matrix(data = c(1, 40, 40^2, 20, 0, 1, 1, 0), nrow = 1, ncol = 8)

# Calculate -2log(Lambda)
linear.combo = mcprofile(object = mroz.glm1, CM = K)

# CI for the linear prredictor
ci.logit.profile <- confint(object = linear.combo, level = 0.95)
ci.logit.profile
```

```
##
##    mcprofile - Confidence Intervals
##
## level:        0.95
## adjustment:   single-step
##
##    Estimate lower upper
## C1    0.801 0.346  1.26
```

```
names(ci.logit.profile)
```

```
## [1] "estimate"    "confint"     "CM"          "quant"       "alternative"
## [6] "level"       "adjust"
```

```
# CI for probability
exp(ci.logit.profile$confint)/(1 + exp(ci.logit.profile$confint))
```

```
##       lower     upper
## 1 0.5857033 0.7795871
```

Thus, the 95% profile likelihood confidence interval of the probability of labor force participation for females who were 40 years old, had income equal to 20, did not attend college, her husband attended college, had log wage equal to 1, and did not have children is $0.586 < \pi < 0.779$