Lab 10

Logistic Regression

NFL field goals

In American football, if you can kick the football through the field goal you will get three points. This is the example we use to illustrate in the lecture. Now we will use this data again to see how to implement Logistic regression and how to interpret your results.

```r
fileName <- "https://raw.githubusercontent.com/ysamwang/btry6020_sp22/main/lectureData/fg_data.csv"
fg_data <- read.csv(fileName)
head(fg_data)
```

```
##   fg_result distance wind  rain
## 1         1       21    8 FALSE
## 2         1       26    8 FALSE
## 3         1       52    8 FALSE
## 4         1       41   12  TRUE
## 5         0       52   12  TRUE
## 6         1       39   12  TRUE
```

There are 4099 observations with the following variables:

- fg_result: was the kick succesful or not
- distance: distance of the attempt in yards
- wind: wind speed at time of kick in mph
- rain: was it raining or not?

We would like to explore the association between fg_result with distance, wind and rain. Since the fg_result is binary variable which only takes value 0 or 1, we will choose the binomial regression to model the NFL data.

Mathematical model

$$\theta(x) = E(Y|X = x)$$

$$log\left(\frac{\theta(x)}{1-\theta(x)}\right) = b_0 + b_1 x_d + b_2 x_w + b_3 x_r$$

Concepts:

- Probability of "success": $\theta(x)$, given covariates are x, ranging from $(0, 1)$.
- Odds: $\frac{\theta(x)}{1-\theta(x)}$, ranging from $(0, \infty)$.
- Logit function (log odds): $log\left(\frac{\theta(x)}{1-\theta(x)}\right)$, ranging from $(-\infty, \infty)$.
- Odds ratio: $\frac{\theta(x_2)/1-\theta(x_2)}{\theta(x_1)/1-\theta(x_1)}$; x_1 and x_2 are two individuals.

Implementation in R

```r
mod_binom <- glm(fg_result ~ distance+ wind + rain,
           family = "binomial", data = fg_data)
```

```r
summary(mod_binom)
```

```
##
## Call:
## glm(formula = fg_result ~ distance + wind + rain, family = "binomial",
##     data = fg_data)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  6.818507   0.382270  17.837  < 2e-16 ***
## distance    -0.117351   0.007871 -14.909  < 2e-16 ***
## wind        -0.035539   0.012832  -2.770  0.00561 **
## rainTRUE    -0.438537   0.261281  -1.678  0.09327 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1895.7  on 2065  degrees of freedom
## Residual deviance: 1589.5  on 2062  degrees of freedom
##   (2033 observations deleted due to missingness)
## AIC: 1597.5
##
## Number of Fisher Scoring iterations: 5
```

Interpretation

The fitted model is

$$log\left(\frac{\theta(x)}{1-\theta(x)}\right) = 6.8185 - 0.1174 * x_d - 0.0355 * x_w - 0.4385 * x_r$$

Suppose x_1 and x_2 are individuals whose covariates values which are the all the same, except that the wind is different by 1: $x_{2,w} = x_{1,w} + 1$.

$$\begin{aligned}
&\log\left(\frac{\theta\left(\mathbf{x}_2\right)}{1-\theta\left(\mathbf{x}_2\right)}\right) - \log\left(\frac{\theta\left(\mathbf{x}_1\right)}{1-\theta\left(\mathbf{x}_1\right)}\right) \\
=&6.8185 - 0.1174 * x_{2,d} - 0.0355 * x_{2,w} - 0.4385 * x_{2,r} - (6.8185 - 0.1174 * x_{1,d} - 0.0355 * x_{1,w} - 0.4385 * x_{1,r}) \\
=&-0.0355 * x_{2,w} - (-0.0355 * x_{1,w}) \\
=&-0.0355 * (x_{1,w} + 1) + 0.0355 * x_{1,w} \\
=&-0.0355
\end{aligned}$$

Also,

$$\log\left(\frac{\theta\left(\mathbf{x}_2\right)}{1-\theta\left(\mathbf{x}_2\right)}\right) - \log\left(\frac{\theta\left(\mathbf{x}_1\right)}{1-\theta\left(\mathbf{x}_1\right)}\right) = log\left(\frac{\theta(x_2)/1-\theta(x_2)}{\theta(x_1)/1-\theta(x_1)}\right) \Rightarrow \quad \frac{\theta(x_2)/1-\theta(x_2)}{\theta(x_1)/1-\theta(x_1)} = exp(-0.0355)$$

Interpretation: If observation 1 and observation 2 have all the same covariates, but $x_{1,w}$ increases by 1 unit to $x_{2,w}$, then the odds for Y_2 are $exp(-0.0355)$ times **smaller** (i.e., multiplicative) than the odds for Y_1

Questions

- Based on the results above, interprete the coefficients for distance and rain.

- Can you determine "smaller" or "larger" in the interpretation by just looking at the coefficient?
- What conclusion can you draw by looking at the p values on the summary?

Prediction

If a kick is from 35 yards, the wind speed is 10 mph, and it is not raining, then we estimate that

$$log\left(\frac{\theta(x)}{1-\theta(x)}\right) = 6.8185 - 0.1174 * 35 - 0.0355 * 10 - 0.4385 * 0 = 2.3545$$

$$\frac{\theta(x)}{1-\theta(x)} = exp(2.3545) = 10.5329$$

$$P(success) = \theta(x) = \frac{exp(2.3545)}{1 + exp(2.3545)} = 0.9133$$

```
## We can use the predict function to get
newdata <- data.frame(distance = 35, wind = 10, rain = FALSE)

# on log-odds scale
predict(object = mod_binom, newdata = newdata, type="link")
```

```
##       1
## 2.35584
```

```
# on "probability of success" scale
predict(object = mod_binom, newdata = newdata, type="response")
```

```
##         1
## 0.9133973
```

Confidence interval

```
# Method 1: Profile likelihood confidence intervals.
# Perform better under the small to moderate sample sizes
confint(mod_binom)
```

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

```
##                   2.5 %      97.5 %
## (Intercept)  6.08836899  7.58791161
## distance    -0.13312726 -0.10224994
## wind        -0.06051182 -0.01015531
## rainTRUE    -0.93809447  0.08998582
```

```
# Method 2: Wald type confidence intervals
cbind(summary(mod_binom)$coefficients[,1]-1.96*summary(mod_binom)$coefficients[,2], summary(mod_binom)$c
```

```
##                   [,1]        [,2]
## (Intercept)  6.0692586  7.56775639
## distance    -0.1327784 -0.10192314
## wind        -0.0606893 -0.01038880
## rainTRUE    -0.9506486  0.07357484
```

Poisson Regression

In this dataset, we record some information regarding games, including competing teams, game season, how much advantage of one team over another in the game. And the numbers of penalties which occurred in the game is our interest.

```r
penalty_data <- read.csv("https://raw.githubusercontent.com/ysamwang/btry6020_sp22/main/lectureData/pena
head(penalty_data)
```

```
##           game_id home_team away_team abs_spread div_game reg_playoff
## 1 2018_01_ATL_PHI       PHI       ATL        1.0        0         REG
## 2 2018_01_BUF_BAL       BAL       BUF        7.5        0         REG
## 3  2018_01_CHI_GB        GB       CHI        6.5        1         REG
## 4 2018_01_CIN_IND       IND       CIN        1.0        0         REG
## 5 2018_01_DAL_CAR       CAR       DAL        2.5        0         REG
## 6  2018_01_HOU_NE        NE       HOU        6.0        0         REG
##   penalty_count
## 1            26
## 2            19
## 3            13
## 4            15
## 5            19
## 6            12
```

There are 1088 observations with the following variables:

- game_id: unique id for game
- home_team: name of home team
- away_team: name of away team
- abs_spread: the absolute value of the betting spread. Roughly speaking, this is the number of points the favored team is expected to win by. A larger value means the game is not expected to be close. We might expect games that are not expected to be close to have less penalties because the refs are less concerned
- div_game: Is the game between two teams in the same division (potentially rivals)
- reg_playoff: is the game a regular season game or a playoff game
- penalty_count: the number of penalties which occurred in the game

Mathematical model

Log function

$$log\left(E(Y|X=x)\right) = b_0 + b_s * x_s + b_d * x_d + b_r * x_r$$

Implementation in R

```r
mod_possion <- glm(penalty_count ~ abs_spread + div_game
                   + reg_playoff, family = "poisson", data = penalty_data)
summary(mod_possion)
```

```
##
## Call:
## glm(formula = penalty_count ~ abs_spread + div_game + reg_playoff,
##     family = "poisson", data = penalty_data)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     2.345965   0.047211  49.691  < 2e-16 ***
## abs_spread     -0.007416   0.002347  -3.160  0.00158 **
## div_game       -0.039004   0.018133  -2.151  0.03147 *
## reg_playoffREG  0.233737   0.046657   5.010 5.45e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1528.0  on 1087  degrees of freedom
## Residual deviance: 1488.4  on 1084  degrees of freedom
## AIC: 6181.6
##
## Number of Fisher Scoring iterations: 4
```

Interpretation

The fitted model is

$$log\left(E[Y|X=x]\right) = 2.3460 - 0.0074 * x_s - 0.0390 * x_d + 0.2337 * x_r$$

Suppose x_1 and x_2 are individuals whose covariates values which are the all the same, except that x_2 is the game between two teams in the same division, while x_1 is not: $x_{2,d} = 1, x_{1,d} = 0$.

$$\begin{aligned}
&\log\left(E[Y|X=x_2]\right) - \log\left(E[Y|X=x_1]\right) \\
=&2.3460 - 0.0074 * x_{2,s} - 0.0390 * x_{2,d} + 0.2337 * x_{2,r} - (2.3460 - 0.0074 * x_{1,s} - 0.0390 * x_{1,d} + 0.2337 * x_{1,r}) \\
=& - 0.0390 * x_{2,d} + 0.0390 * x_{1,d} \\
=& - 0.0390 * (x_{1,d} + 1) + 0.0390 * x_{1,d} \\
=& - 0.0390
\end{aligned}$$

then we also have

$$\frac{E[Y|X=x_2]}{E[Y|X=x_1]} = exp(-0.0390)$$

Interpretation: Suppose two observations have all the same covariate values except differ in div_game (x_d) that x_2 is the game between two teams in the same division and x_1 is not, then the expected mean for the number of penalties with covariates x_1 is $exp(-0.0390)$ times (smaller) the expected mean for the number of penalties with covariates x_2.

Questions

- Based on the results above, interprete the coefficients for abs_spread and reg_playoffREG.
- What conclusion can you draw by looking at the p values on the summary?

Prediction

If in a game, the number of points the favored team is expected to win by 5(abs_spread), and this team play a regular season game with the rival in the same division, then we estimate that

$$log(E[Y|X=x]) = 2.3460 - 0.0074 * 5 - 0.0390 * 1 + 0.2337 * 1 = 2.5037$$

$$E[Y|X=x] = exp(2.5037) = 12.22765$$

This means the expected number of penalties which occurred in this game is around 12 times.

```r
## We can use the predict function
newdata <- data.frame(abs_spread = 5, div_game = 1, reg_playoff = "REG")

# predicted log of mean
predict(object = mod_possion, newdata = newdata, type="link")
```

```
##        1
## 2.503617
```

```r
# predicted mean
predict(object = mod_possion, newdata = newdata, type="response")
```

```
##        1
## 12.22664
```

Confidence interval

```r
# Method 1: Profile likelihood confidence intervals.
# Perform better under the small to moderate sample sizes
confint(mod_possion)
```

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

```
##                      2.5 %       97.5 %
## (Intercept)     2.25216371  2.437268425
## abs_spread     -0.01203173 -0.002831288
## div_game       -0.07460542 -0.003524408
## reg_playoffREG  0.14354741  0.326479820
```

```r
# Method 2: Wald type confidence intervals
cbind(summary(mod_possion)$coefficients[,1]-1.96*summary(mod_possion)$coefficients[,2], summary(mod_pos
```

```
##                       [,1]         [,2]
## (Intercept)     2.25343138  2.438498628
## abs_spread     -0.01201647 -0.002815976
## div_game       -0.07454428 -0.003463906
## reg_playoffREG  0.14228968  0.325184530
```