Homework3

Yufei Yin

13

The Larry Bird free throw shooting data discussed in Section 1.2 can be examined in a logistic regression context. The purpose here is to estimate the probability of success for the second free throw attempt given what happened on the first free throw attempt. Complete the following:

(a)

The **c.table** object created in **Bird.R** contains the contingency table format of the data. To create the data frame format needed for **glm()**, we can re-enter the data in a new data frame with

```
bird <- data.frame(First = c("made", "missed"),
                   success = c(251 ,48),
                   trials = c(285 , 53))
bird
```

```
##    First success trials
## 1   made     251    285
## 2 missed      48     53
```

or transform it directly with

```
c.table<-array(data = c(251, 48, 34, 5), dim = c(2,2), dimnames = list(First = c("made", "missed"),
             Second = c("made", "missed")))
bird1 <- as.data.frame(as.table(c.table))
trials <- aggregate(formula = Freq ~ First , data = bird1 , FUN = sum)
success <- bird1[bird1$Second == "made", ]
bird2 <- data.frame(First = success$First ,
                    success = success$Freq ,
                    trials = trials$Freq )
bird2
```

```
##    First success trials
## 1   made     251    285
## 2 missed      48     53
```

The second method is more general and can work with larger contingency tables like those examined in Chapters 3 and 4. Implement each line of code in the second method above and describe what occurs.

```
bird1 <- as.data.frame(as.table(c.table))
bird1
```

```
##    First Second Freq
## 1   made   made  251
## 2 missed   made   48
## 3   made missed   34
```

```
## 4 missed missed    5
```

as.table() function coerce the array to class table, **as.data.frame()** function convert it to dataframe showed above.

```
trials <- aggregate(formula = Freq ~ First , data = bird1 , FUN = sum)
trials
```

```
##    First Freq
## 1   made  285
## 2 missed   53
```

aggregate() function returns the sum of made and missed on the first free throw attempt.

```
success <- bird1[bird1$Second == "made", ]
success
```

```
##    First Second Freq
## 1   made   made  251
## 2 missed   made   48
```

success is the subset of bird1, we want to get the rows from bird1 with the condition that column Second is "made".

```
bird2 <- data.frame(First = success$First ,
                    success = success$Freq ,
                    trials = trials$Freq )
bird2
```

```
##    First success trials
## 1   made     251    285
## 2 missed      48     53
```

we use **data.frame()** function to create a new data frame contains the previous data.

(b)

Estimate a logistic regression model for the probability of success on the second attempt, where First is the explanatory variable.

```
bird1$First <- ifelse(bird1[,1] == "made",1, 0)
bird1$Second <- ifelse(bird1[,2]== "made",1, 0)
bird1
```

```
##   First Second Freq
## 1     1      1  251
## 2     0      1   48
## 3     1      0   34
## 4     0      0    5
```

```
mod.fit <- glm(formula = Second ~ First,
               family = binomial(link = logit),
               weights = Freq,
               data = bird1)
mod.fit$coefficients
```

```
## (Intercept)       First
##   2.2617631  -0.2626707
```

$\text{logit}(\hat{\pi})$ = 2.2617631 - 0.2626707 * FIRST

(c)

Estimate the odds ratio comparing the two levels of First. Calculate both Wald and profile LR intervals for the odds ratio. Compare these calculated values with those obtained in the Larry Bird example of Section 1.2.5. If you use confint() to compute the intervals, make sure to use the correct value for the parm argument (same name as the indicator variable given by summary(mod.fit)).

```
# estimate the odds ratio
exp(mod.fit$coefficients[2])
```

```
##     First
## 0.7689951
```

```
1/exp(mod.fit$coefficients[2])
```

```
##    First
## 1.300398
```

odds ratio comparing the odds of a successful second free throw attempt when the first is made to when the first is missed. The result we got is same as Section 1.2.5 page 43.

Because the estimated odds ratio is less than 1, we decided to invert it to help with its interpretation. The estimated odds of a successful second free throw attempt are 1.30 times as large as when the first free throw is missed than when the first free throw is made.

```
# Wald interval
exp(confint.default(object = mod.fit, parm = "First", level = 0.95))
```

```
##           2.5 %   97.5 %
## First 0.2862484 2.065875
```

```
rev(1/exp(confint.default(object = mod.fit, parm = "First", level = 0.95)))
```

```
## [1] 0.4840564 3.4934693
```

with 95% confidence, the odds of a successful free throw attempt are between 0.48 and 3.49 times as large as when the first free throw is missed than when the first free throw is made.

The same as Section 1.2.5 page 43.

```
# profile LR interval
exp(confint(object = mod.fit, parm = "First", level = 0.95))
```

```
##     2.5 %    97.5 %
## 0.2537934 1.9072330
```

```
rev(1/exp(confint(object = mod.fit, parm = "First", level = 0.95)))
```

```
##    97.5 %     2.5 %
## 0.5243198 3.9402133
```

with 95% confidence, the odds of a successful free throw attempt are between 0.52 and 3.94 times as large as when the first free throw is missed than when the first free throw is made.

(d)

Perform a hypothesis test of H0 : $\beta_1 = 0$ vs. Ha : $\beta_1 \neq 0$ using Wald and LR statistics. Compare these results to those found for the Larry Bird example in Section 1.2.3.

```
# Wald statistic
summary(mod.fit)
```

```
##
## Call:
## glm(formula = Second ~ First, family = binomial(link = logit),
##     data = bird1, weights = Freq)
##
## Deviance Residuals:
##       1        2        3        4
##   7.986    3.084  -12.024   -4.859
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   2.2618     0.4699   4.813 1.49e-06 ***
## First        -0.2627     0.5042  -0.521    0.602
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 241.76  on 3  degrees of freedom
## Residual deviance: 241.47  on 2  degrees of freedom
## AIC: 245.47
##
## Number of Fisher Scoring iterations: 5
```

The Wald test p-value is 0.602 is the same as section 1.2.3 page 36.

```
# Likelihood ratio test
library(car)
Anova(mod.fit)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Second
##       LR Chisq Df Pr(>Chisq)
## First  0.28575  1      0.593
```

The Likelihood ratio test p-value is 0.593 is the same as section 1.2.3 page 36.

(e)

Discuss why similarities and/or differences occur between the calculations here using logistic regression and the corresponding calculations in Chapter 1.

In the context of binary responses, the quantity we want to estimate is the probability of success, π. Let Y_i be independent binary response variables for observations i = 1, ... n, where a value of 1 denotes a success and a value of 0 denotes a failure. similar to Section 1.1, a Bernoulli distribution describes Y_i very well but we now allow the probability of success parameter π_i to be different for each observation. Thus, $P(Y_i = y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}$ for $y_i = 0$ or 1 is the PMF for Y_i.

To find the MLEs for π_i, the likelihood function is

$$L(\pi_1, ..., \pi_n | y_1, ...y_n) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

Maximum likelihood estimation is used to estimate the regression parameters $\beta_0, ...\beta_p$ of the logistic regression model.

the likelihood function is

$$L(\beta_0, ..., \beta_p | y_1, ...y_n) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

Noticed that we used the same likelihood function, so the calculations here using logistic regression and the corresponding calculations in Chapter 1 are similar.

19

Thorburn et al. (2001) examine hepatitis C prevalence among healthcare workers in the Glasgow area of Scotland. These healthcare workers were categorized into the following occupational groups:

(1) exposure prone (e.g., surgeons, dentists, surgical nurses),

(2) fluid contact (e.g., non-surgical nurses),

(3) lab staff,

(4) patient contact (e.g., pharmacists, social workers), and

(5) no patient contact (e.g., clerical).

The collected data are available in the healthcare_worker.csv file. Is there evidence of an occupational group effect on hepatitis status? If there is sufficient evidence of an effect, use the appropriate odds ratios to make comparisons among the groups. If there is not sufficient evidence of an effect, discuss why this may be a preferred result.

```
healthcare_worker <- read.csv("healthcare_worker.csv", stringsAsFactors = TRUE)
mod.fit <- glm(data = healthcare_worker,
               formula = Hepatitis/Size ~ Occup.group,
               family = binomial(link = logit),
               weights = Size)
Anova(mod.fit)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Hepatitis/Size
##             LR Chisq Df Pr(>Chisq)
## Occup.group    3.735  4     0.4431
```

According to the output of **Anova()** function, we have the result of likelihood ratio test. p-value is 0.4431 which is greater than 0.05, we reject the null hypothesis Occup.group is significant important in our model, so there is not sufficient evidence that an occupational group effect on hepatitis status. This may be a preferred result because from our data we did not observed that high-risk occupational group such as exposure prone are more likely to infect hepatitis.

8

Exercise 19 of Chapter 2 examined the prevalence of hepatitis C among healthcare workers in the Glasgow area of Scotland. Convert the data for this exercise to a contingency table structure where occupational groups are located on the rows and the presence and absence of hepatitis are located on the columns. Perform Pearson chi-square and LR tests for independence using these data. Are your results here similar to what was found for the data analysis in Chapter 2? Explain.

```
c.table <- array(data = c(healthcare_worker$Hepatitis,
                          healthcare_worker$Size-healthcare_worker$Hepatitis),
                 dim = c(5,2),
                 dimnames = list(Occup.group = healthcare_worker$Occup.group,
                                 Hepatitis = c("Presence", "Absence")))
c.table
```

```
##                     Hepatitis
## Occup.group          Presence Absence
##   Exposure prone            5    2200
##   Fluid contact            17    6190
##   Lab staff                 3     530
##   Patient contact           2    1236
##   No patient contact        3     468
```

```
library(package = vcd)
assocstats(x = c.table)
```

```
##                     X^2 df P(> X^2)
## Likelihood Ratio 3.7350  4  0.44305
## Pearson          4.5043  4  0.34204
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.021
## Cramer's V        : 0.021
```

Pearson chi-square test:

$\chi^2 = 4.5043$, p-value $= 0.34204$

Likelihood Ratio test:

-2log(Λ) $= 3.7350$, p-value $= 0.44305$

result here is similar to what was found for the data analysis in Chapter 2.

The reason is the same as we explained on chapter 2 question 13 part(e).