Categorical Data Analysis - Assignment 2

Hyungyeong Hong (2022KU0104)

4/14/2022

Note 1) Show your code and output
Note 2) Conducting a test includes
a) write the hypotheses, b) calculate the test statistic, c) find the p-value, and d) write the conclusion

```
setwd("/Users/hyungyeonghong/Desktop/2022_Spring/CDA_2022/Assignment_2")
rm(list = ls(all = TRUE))
```

[Questions 1-2]

The horseshoe crab data file (shown partly in Table 3.2) is on the Blackboard. (crab.dat) In the data file, y = 1 if a crab has at least one satellite and y = 0 otherwise.

- Loading the data

```
Crabs <- read.table("Crabs.dat", header = TRUE); head(Crabs)
```

```
##   crab sat y weight width color spine
## 1    1   8 1   3.05  28.3     2     3
## 2    2   0 0   1.55  22.5     3     3
## 3    3   9 1   2.30  26.0     1     1
## 4    4   0 0   2.10  24.8     3     3
## 5    5   4 1   2.60  26.0     3     3
## 6    6   0 0   2.10  23.8     2     3
```

- Checking the y variable of the data

```
unique(Crabs$y)
```

```
## [1] 1 0
```

[Question 1]

Using weight as the predictor, fit the linear regression model. That is, fit GLM using the identity link. Write the estimated regression equation. Find the fitted probability of having at least one satellite when weight is 5.20kg.

a) Fitting the Linear Regression Model
- We fit the linear regression model by using the identity link, where weight is used as a predictor.

```
fit.linear <- glm(y ~ weight, family = gaussian(link = "identity"), data = Crabs)
```

b) Estimated Regression Equation
- The summary of the linear regression model fit is,

```
summary(fit.linear)
```

```
##
## Call:
## glm(formula = y ~ weight, family = gaussian(link = "identity"),
##     data = Crabs)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8878  -0.4683   0.1606   0.3704   0.6689
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.14487    0.14715  -0.984    0.326
## weight       0.32270    0.05876   5.492 1.42e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1977574)
##
##     Null deviance: 39.780  on 172  degrees of freedom
## Residual deviance: 33.817  on 171  degrees of freedom
## AIC: 214.56
##
## Number of Fisher Scoring iterations: 2
```

- Therefore the estimated regression equation is $\hat{P}(Y = 1) = -0.14487 + 0.32270x$, where x indicates the weight.

c) Finding the Fitted Probability
- The fitted probability of having at least one satellite when the weight is 5.20kg is,

```
alpha.hat <- fit.linear$coefficients[1]; names(alpha.hat) <- NULL
beta.hat <- fit.linear$coefficients[2]; names(beta.hat) <- NULL

alpha.hat + beta.hat*5.20
```

```
## [1] 1.533186
```

- Or, we can use the `predict()` function to find the fitted probability of having at least one satellite when the weight is 5.20kg.

```
predict(fit.linear, data.frame(weight = 5.20), type = "response")
```

```
##        1
## 1.533186
```

- Since probability cannot be greater than 1, this result is problematic.

[Question 2]

Fit the logistic regression model. Write the estimated regression equation. Find the fitted probability of having at least one satellite when the weight is 5.20kg.

a) Fitting the Logistic Regression Model

```
fit.logistic <- glm(y ~ weight, family = binomial(link = "logit"), data = Crabs)
```

b) Estimated Regression Equation
- Result of the logistic regression model fit

```
summary(fit.logistic)
```

```
##
## Call:
## glm(formula = y ~ weight, family = binomial(link = "logit"),
##     data = Crabs)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1108  -1.0749   0.5426   0.9122   1.6285
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.6947     0.8802  -4.198 2.70e-05 ***
## weight        1.8151     0.3767   4.819 1.45e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 225.76  on 172  degrees of freedom
## Residual deviance: 195.74  on 171  degrees of freedom
## AIC: 199.74
##
## Number of Fisher Scoring iterations: 4
```

- Therefore the estimated logistic regression equation is $\log[\frac{\hat{P}(Y=1)}{1-\hat{P}(Y=1)}] = -3.6947 + 1.8151x$, where x indicates the weight.

c) Finding the Fitted Probability

- By using the formula $\hat{P}(Y=1) = \frac{\exp(\hat{\alpha}+\hat{\beta}x)}{1+\exp(\hat{\alpha}+\hat{\beta}x)}$, we can calculate the fitted probability of having at least one satellite when the weight is 5.20kg.

```
alpha.hat <- fit.logistic$coefficients[1]; names(alpha.hat) <- NULL
beta.hat <- fit.logistic$coefficients[2]; names(beta.hat) <- NULL

exp(alpha.hat + beta.hat*5.20) / (1 + exp(alpha.hat + beta.hat*5.20))
```

```
## [1] 0.9968084
```

- Or, we can use the `predict()` function to find the fitted probability of having at least one satellite when the weight is 5.20kg.

```
predict(fit.logistic, data.frame(weight = 5.20), type = "response")
```

```
##         1
## 0.9968084
```

[Questions 3-6]

See the slide # 40 in Chapter 3 and utilize the provided code.

```
rm(list = ls(all = TRUE))
```

From the 2016 General Social Survey, when we cross-classify political ideology (with 1 being most liberal and 7 being most conservative) by political party affiliation for subjects ages 18-27,

we get

	1	2	3	4	5	6	7
Democrat	5	18	19	25	7	7	2
Republican	1	3	1	11	10	11	1

When we use R to model the effect of political ideology on the probability of being a Democrat, we get the result:

```
y <- c(5,18,19,25,7,7,2)
n <- c(6,21,20,36,17,18,3)
x <- c(1,2,3,4,5,6,7)

fit <- glm(y/n ~ x, family = binomial, weights = n)
```

[Question 3]

Report the estimated equation and interpret the direction of the estimated effect.

a) Estimated Equation

- The summary of logistic regression fit is,

```
summary(fit)
```

```
##
## Call:
## glm(formula = y/n ~ x, family = binomial, weights = n)
##
## Deviance Residuals:
##       1        2        3        4        5        6        7
## -0.8058  -0.3360   1.8917  -0.0154  -1.2160  -0.2041   1.3886
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   3.1870     0.7002   4.552 5.33e-06 ***
## x            -0.5901     0.1564  -3.772 0.000162 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 24.7983  on 6  degrees of freedom
## Residual deviance:  7.7894  on 5  degrees of freedom
## AIC: 30.516
##
## Number of Fisher Scoring iterations: 4
```

- Thus the estimated equation is $\log[\frac{\hat{P}(Y=1)}{1-\hat{P}(Y=1)}] = 3.1870 - 0.5901x$, where x represents the political ideology and $P(Y=1)$ represents the probability of being a Democrat.

b) Direction of the Estimated Effect

- From the estimated equation $\log[\frac{\hat{P}(Y=1)}{1-\hat{P}(Y=1)}] = 3.1870 - 0.5901x$, we can get the formula for the odds which looks like $\frac{\hat{P}(Y=1)}{1-\hat{P}(Y=1)} = e^{3.1870-0.5901x} = e^{3.1870} \times e^{-0.5901x}$.
- Therefore, a one unit increase in x has a multiplicative impact of $e^{-0.5901}$, which leads to the decrement of

estimated odds of being a Democrat $\Leftrightarrow \frac{\hat{P}(Y=1)}{1-\hat{P}(Y=1)}$.

[Question 4]

Calculate the 95% Wald confidence interval for the effect of political ideology.

- Extracting the estimated coefficient and the standard error from the result of model fit

```
beta.hat <- fit$coefficients[2]; names(beta.hat) <- NULL
std.err <- summary(fit)$coef[2,2]
```

- The lower bound for the 95% Wald CI is,

```
LB <- beta.hat - qnorm(0.025, lower.tail = FALSE)*std.err; LB
```

```
## [1] -0.8967066
```

- The upper bound for the 95% Wald CI is,

```
UB <- beta.hat + qnorm(0.025, lower.tail = FALSE)*std.err; UB
```

```
## [1] -0.2835034
```

- Therefore, the 95% Wald CI for the effect of political ideology is $[-0.8967066, -0.2835034]$.

[Question 5]

Conduct the Wald test for the effect of x. (See the note 2 above.)

a) Setting the hypotheses
- Since we need to test the effect of x, we should test whether the coefficient of x is zero or not.
- Therefore the hypotheses are, $H_0 : \beta = 0 \quad vs \quad H_1 : \beta \neq 0$.

b) Calculating the test statistic

```
beta.hat <- fit$coefficients[2]; names(beta.hat) <- NULL
std.err <- summary(fit)$coef[2,2]

Wald.stat <- beta.hat / std.err; Wald.stat
```

```
## [1] -3.772272
```

c) Finding the p-value

```
2*pnorm(Wald.stat)
```

```
## [1] 0.0001617676
```

d) Conclusion
- Since the p-value is less than 0.05, reject the null hypothesis $H_0 : \beta = 0$.
- Therefore, we can conclude that there is a significant association between political ideology and political party affiliation.

[Question 6]

Conduct the likelihood-ratio test for the effect of x. (See the note 2 above.)

a) Setting the hypotheses
- Since we need to test the effect of x, we should test whether the coefficient of x is zero or not.
- Therefore the hypotheses are, $H_0 : \beta = 0 \quad vs \quad H_1 : \beta \neq 0$.

b) Calculating the test statistic

```
LR.stat <- car::Anova(fit)$`LR Chisq`; LR.stat
```

```
## [1] 17.00891
```

c) Finding the p-value

```
1 - pchisq(LR.stat, df = 1) # or) pchisq(LR.stat, df = 1, lower.tail = FALSE)
```

```
## [1] 3.720474e-05
```

d) Conclusion

- Since the p-value is less than 0.05, reject the null hypothesis $H_0 : \beta = 0$.
- Therefore, we can conclude that there is a significant association between political ideology and political party affiliation.

[Questions 7-9]

The following output shows the R code for the data and analyzing the cancer remission study data. The study investigated characteristics associated with y = whether a cancer patient achieved remission (1=yes, 0=no). An important explanatory variable is a labeling index (LI=percentage of "labeled" cells) that measures proliferative activity of cells after a patient receives an injection of tritiated thymidine.

```
------------------------------------------------------------------------------
> LI <- c(8,8,10,10,12,12,12,14,14,14,16,16,16,18,20,20,20,22,22,24,26,28,32,34,
+         38,38,38)
> y <- c(0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,0,1,0,0,1,1,0,1,1,1,0)
> summary(glm(y ~ LI, family=binomial))
             Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  -3.77714     1.37862   -2.740   0.00615
LI            0.14486     0.05934    2.441   0.01464
---
    Null deviance: 34.372  on 26  degrees of freedom
Residual deviance: 26.073  on 25  degrees of freedom

>  confint(glm(y ~ LI, family=binomial))
                2.5 %     97.5 %
LI            0.04252    0.28467
------------------------------------------------------------------------------
```

```
rm(list = ls(all = TRUE))

LI <- c(8, 8, 10, 10, 12, 12, 12, 14, 14, 14, 16, 16, 16, 18,
        20, 20, 20, 22, 22, 24, 26, 28, 32, 34, 38, 38, 38)
y <- c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 1,
       0, 0, 1, 1, 0, 1, 1, 1, 0)

summary(glm(y ~ LI, family = binomial))
```

```
##
## Call:
## glm(formula = y ~ LI, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9448  -0.6465  -0.4947   0.6571   1.6971
##
## Coefficients:
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.77714    1.37862  -2.740  0.00615 **
## LI           0.14486    0.05934   2.441  0.01464 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 34.372  on 26  degrees of freedom
## Residual deviance: 26.073  on 25  degrees of freedom
## AIC: 30.073
##
## Number of Fisher Scoring iterations: 4
```

[Question 7]

Find the value of LI which provides estimated probability of 0.5. (That is, $\hat{P}(Y = 1) = 0.5$)

$\log[\frac{\hat{P}(Y=1)}{1-\hat{P}(Y=1)}] = \hat{\alpha} + \hat{\beta}x$

$\Leftrightarrow \hat{\beta}x = \log[\frac{\hat{P}(Y=1)}{1-\hat{P}(Y=1)}] - \hat{\alpha}$

$\Leftrightarrow x = \frac{\log[\frac{\hat{P}(Y=1)}{1-\hat{P}(Y=1)}] - \hat{\alpha}}{\hat{\beta}}$

```
log.odds <- log(0.5/(1-0.5))

alpha.hat <- summary(glm(y ~ LI, family = binomial))$coefficients[1, 1]
beta.hat <- summary(glm(y ~ LI, family = binomial))$coefficients[2, 1]

(log.odds - alpha.hat) / beta.hat
```

```
## [1] 26.07384
```

- Therefore, the value of LI that provides estimated probability of 0.5 is 26.07384.

[Question 8]

When LI increases by 1, show that the estimated odds of remission multiply by 1.16.

$\log[\frac{\hat{P}(Y=1)}{1-\hat{P}(Y=1)}] = \hat{\alpha} + \hat{\beta}x$

$\Leftrightarrow \frac{\hat{P}(Y=1)}{1-\hat{P}(Y=1)} = e^{\hat{\alpha}+\hat{\beta}x}$

$\Leftrightarrow \frac{\hat{P}(Y=1)}{1-\hat{P}(Y=1)} = e^{\hat{\alpha}} \times e^{\hat{\beta}x}$

$\Leftrightarrow \frac{\hat{P}(Y=1)}{1-\hat{P}(Y=1)} = e^{\hat{\alpha}} \times (e^{\hat{\beta}})^{x}$

In our case, the value of $\hat{\beta}$ is 0.14486 and therefore as $x(=$ LI$)$ increases by 1, the estimated odds $\frac{\hat{P}(Y=1)}{1-\hat{P}(Y=1)}$ is multiplied by $e^{0.14486} = 1.16$.

- For example, let $x = 1$. Then the value of estimated odds is,

```
odds.1 <- exp(alpha.hat + beta.hat*1); odds.1
```

```
## [1] 0.02645588
```

- Suppose that x increases by 1 unit. Then, $x = 2$ and the value of estimated odds is,

```
odds.2 <- exp(alpha.hat + beta.hat*2); odds.2
```

```
## [1] 0.03057986
```

- If we divide the estimated odds when $x = 2$ by the estimated odds when $x = 1$, the result is approximately 1.16, which means that a unit increase in LI has multiplicative effect on the estimated odds.

```
odds.2 / odds.1
```

```
## [1] 1.155881
```

[Question 9]

Find the rate of change in $\hat{P}(Y = 1)$, when LI is 8.

$\hat{P}(Y = 1) = \frac{\exp(\hat{\alpha}+\hat{\beta}x)}{1+\exp(\hat{\alpha}+\hat{\beta}x)}$

```
pi.hat <- exp(alpha.hat + beta.hat*8) / (1 + exp(alpha.hat + beta.hat*8)); pi.hat
```

```
## [1] 0.06797405
```

- Since the estimated probability for $LI = 8$ is 0.06797405, the rate of change in $\hat{P}(Y = 1)$ is,

```
beta.hat * pi.hat * (1-pi.hat)
```

```
## [1] 0.009177602
```