Categorical Data Analysis - Assignment 3

Hyungyeong Hong (2022KU0104)

5/27/2022

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_3")
rm(list = ls(all = TRUE))
```

[Questions 1-4]

The second horseshoe crab data file is uploaded on the Blackboard. (crab2.dat) In the data file, y = 1 if a female crab is monandrous, and y = 0 if polyandrous. (For the details, read Slide#26 in Chapter 5.)

[Question 1]

Fit a logistic regression using all the explanatory variables except for the 'year'. Here, Fcolor and Fsurf are categorical variables. (Do not consider the two as continuous.)

- *a) Loading the given dataset*

```
Crabs <- read.table("Crabs2.dat", header = TRUE); head(Crabs, 5)
```

```
##   y Year Fcolor Fsurf  FCW AMCW AMcolor AMsurf
## 1 0 1993      1     1 20.2 18.1       5      5
## 2 0 1993      1     3 23.0 13.5       5      5
## 3 0 1993      5     5 22.6 16.7       5      3
## 4 0 1993      5     5 20.8 16.0       5      4
## 5 0 1993      3     4 23.3 15.3       5      3
```

- *b) Fitting the Logistic Regression Model*
- Here, we fit a logistic regression model by using all the explanatory variables except for the 'year' variable.
- Fcolor and Fsurf are categorical variables.
[*The followings are the descriptions for each variable in the dataset*]
- 1. y = whether a female horseshoe crab is monandrous or polyandrous
- 2. Fcolor = female crab's color(1 = dark, 3 = mediun, 5 = light)
- 3. Fsurf = female crab's surface condition(values 1, 2, 3, 4, 5 where lower values representing worse condition)
- 4. FCW = female crab's carapace width
- 5. AMCW = attached male's carapace width
- 6. AMcolor = attached male's color(1 = dark, 3 = mediun, 5 = light)
- 7. AMsurf = attached male's surface condition(values 1, 2, 3, 4, 5 where lower values representing worse condition)

```
logistic.fit <- glm(y ~ factor(Fcolor) + factor(Fsurf) + FCW + AMCW + AMcolor + AMsurf,
                    family = binomial(link = "logit"), data = Crabs)
```

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

```
##
## Call:
## glm(formula = y ~ factor(Fcolor) + factor(Fsurf) + FCW + AMCW +
##     AMcolor + AMsurf, family = binomial(link = "logit"), data = Crabs)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2856  -0.8311  -0.7181   1.2420   1.9263
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      0.78174    1.21159   0.645  0.51878
## factor(Fcolor)3  0.79719    0.18436   4.324 1.53e-05 ***
## factor(Fcolor)5  0.27965    0.20238   1.382  0.16703
## factor(Fsurf)2  -0.04317    0.24772  -0.174  0.86167
## factor(Fsurf)3  -0.50957    0.23167  -2.200  0.02784 *
## factor(Fsurf)4  -0.60156    0.24827  -2.423  0.01539 *
## factor(Fsurf)5  -0.84875    0.28250  -3.004  0.00266 **
## FCW             -0.04433    0.03937  -1.126  0.26024
## AMCW            -0.05096    0.05573  -0.914  0.36047
## AMcolor          0.07409    0.05777   1.282  0.19970
## AMsurf          -0.02402    0.06647  -0.361  0.71779
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1633.8  on 1344  degrees of freedom
## Residual deviance: 1581.8  on 1334  degrees of freedom
## AIC: 1603.8
##
## Number of Fisher Scoring iterations: 4
```

[Question 2]

Conduct a likelihood ratio test for comparing the model in Q1 with a null model under the significance level 0.05. (Read Note2 carefully.)

a) Setting the hypotheses

- We need to conduct a LR test in order to compare the fitted logistic regression model and the null model.
- Therefore, the hypotheses are $H_0 : \beta_1 = \cdots = \beta_p = 0 \quad vs \quad H_1 : at\ least\ one\ \beta_j \neq 0$.
- Or this can be written as, $H_0 : null\ model\ is\ better \quad vs \quad H_1 : current\ model\ is\ better$.

b) Calculating the test statistic

- We first get the null deviance, the residual deviance and the degrees of freedom of the null model and the current(fitted) model from the fitted logistic regression model above.

```
dev.null <- logistic.fit$null.deviance
dev.resid <- logistic.fit$deviance
df.null <- logistic.fit$df.null
df.resid <- logistic.fit$df.residual
```

- Then, we compute the test statistic.
- The likelihood ratio test statistic is calculated from the null deviance, and the residual deviance. Its value

can be calculated as,

```
LRT.stat <- dev.null - dev.resid; LRT.stat
```

```
## [1] 51.95703
```

- For the likelihood ratio test, we also need the degrees of freedom for the chi-squared distribution.
- The degrees of freedom for the likelihood ratio test is calculated from the degrees of freedom of the null model and the current model. Its value can be calculated as,

```
LRT.df <- df.null - df.resid; LRT.df
```

```
## [1] 10
```

c) Finding the p-value

```
1-pchisq(LRT.stat, df = LRT.df)
```

```
## [1] 1.162205e-07
```

d) Conclusion

- Since the p-value is less than 0.05, reject the null hypothesis $H_0 : \beta_1 = \cdots = \beta_p = 0$ or H_0 : *null model is better*, equivalently.
- Therefore, we can conclude that at least one predictor has an effect or equivalently, the current(fitted) model is better.

```
1-pchisq(LRT.stat, df = LRT.df) < 0.05
```

```
## [1] TRUE
```

[Question 3]

We want to select a smaller model beginning from the model in Q1. (That is, we will perform a backward elimination for the variable selection.)

[Question 3-(a)]

Which variables are contained in an optimal model selected by AIC?

- From a backward elimination by AIC, the model having Fcolor and Fsurf as predictors seems to be optimal.

```
backward.AIC <- MASS::stepAIC(logistic.fit)
```

```
## Start:  AIC=1603.84
## y ~ factor(Fcolor) + factor(Fsurf) + FCW + AMCW + AMcolor + AMsurf
##
##                   Df Deviance    AIC
## - AMsurf           1   1582.0 1602.0
## - AMCW             1   1582.7 1602.7
## - FCW              1   1583.1 1603.1
## - AMcolor          1   1583.5 1603.5
## <none>                 1581.8 1603.8
## - factor(Fsurf)    4   1596.6 1610.6
## - factor(Fcolor)   2   1606.1 1624.1
##
## Step:  AIC=1601.97
## y ~ factor(Fcolor) + factor(Fsurf) + FCW + AMCW + AMcolor
##
##                   Df Deviance    AIC
## - AMCW             1   1582.8 1600.8
## - FCW              1   1583.3 1601.3
```

```
## - AMcolor         1   1583.5 1601.5
## <none>                1582.0 1602.0
## - factor(Fsurf)   4   1597.0 1609.0
## - factor(Fcolor)  2   1606.6 1622.6
##
## Step:  AIC=1600.79
## y ~ factor(Fcolor) + factor(Fsurf) + FCW + AMcolor
##
##                  Df Deviance    AIC
## - AMcolor         1   1584.3 1600.3
## - FCW             1   1584.4 1600.4
## <none>                1582.8 1600.8
## - factor(Fsurf)   4   1597.6 1607.6
## - factor(Fcolor)  2   1607.4 1621.4
##
## Step:  AIC=1600.34
## y ~ factor(Fcolor) + factor(Fsurf) + FCW
##
##                  Df Deviance    AIC
## - FCW             1   1585.9 1599.9
## <none>                1584.3 1600.3
## - factor(Fsurf)   4   1599.2 1607.2
## - factor(Fcolor)  2   1611.0 1623.0
##
## Step:  AIC=1599.93
## y ~ factor(Fcolor) + factor(Fsurf)
##
##                  Df Deviance    AIC
## <none>                1585.9 1599.9
## - factor(Fsurf)   4   1601.0 1607.0
## - factor(Fcolor)  2   1612.6 1622.6
```

[**Question 3-(b)**]

Which variables are contained in an optimal model selected by BIC?
- From a backward elimination by BIC, the model having Fcolor as the only predictor seems to be optimal.

```
n = nrow(Crabs)
backward.BIC <- MASS::stepAIC(logistic.fit, k=log(n))
```

```
## Start:  AIC=1661.09
## y ~ factor(Fcolor) + factor(Fsurf) + FCW + AMCW + AMcolor + AMsurf
##
##                  Df Deviance    AIC
## - factor(Fsurf)   4   1596.6 1647.0
## - AMsurf          1   1582.0 1654.0
## - AMCW            1   1582.7 1654.7
## - FCW             1   1583.1 1655.2
## - AMcolor         1   1583.5 1655.5
## <none>                1581.8 1661.1
## - factor(Fcolor)  2   1606.1 1670.9
##
## Step:  AIC=1647.03
## y ~ factor(Fcolor) + FCW + AMCW + AMcolor + AMsurf
##
```

```
##                   Df Deviance    AIC
## - AMsurf           1   1597.0 1640.2
## - AMCW             1   1597.2 1640.5
## - FCW              1   1598.1 1641.3
## - AMcolor          1   1598.6 1641.8
## <none>                 1596.6 1647.0
## - factor(Fcolor)   2   1626.9 1662.9
##
## Step:  AIC=1640.2
## y ~ factor(Fcolor) + FCW + AMCW + AMcolor
##
##                   Df Deviance    AIC
## - AMCW             1   1597.6 1633.6
## - FCW              1   1598.5 1634.5
## - AMcolor          1   1598.6 1634.6
## <none>                 1597.0 1640.2
## - factor(Fcolor)   2   1628.5 1657.3
##
## Step:  AIC=1633.59
## y ~ factor(Fcolor) + FCW + AMcolor
##
##                   Df Deviance    AIC
## - AMcolor          1   1599.2 1628.0
## - FCW              1   1599.4 1628.2
## <none>                 1597.6 1633.6
## - factor(Fcolor)   2   1629.2 1650.8
##
## Step:  AIC=1628.03
## y ~ factor(Fcolor) + FCW
##
##                   Df Deviance    AIC
## - FCW              1   1601.0 1622.6
## <none>                 1599.2 1628.0
## - factor(Fcolor)   2   1631.4 1645.8
##
## Step:  AIC=1622.61
## y ~ factor(Fcolor)
##
##                   Df Deviance    AIC
## <none>                 1601.0 1622.6
## - factor(Fcolor)   2   1633.8 1641.0
```

[Question 4]

Can you conduct a likelihood ratio test for comparing the two models you selected in Q3? If yes, perform a LR test. (In this case, report the value of test statistic and p-value.) If no, explain why you cannot.

- From the backward elimination with AIC as its criterion, the model having Fcolor and Fsurf as its predictors is selected.
- From the backward elimination with BIC as its criterion, the model having Fcolor as the only predictor is selected.
- Since the two models are nested, we can compare these models by likelihood ratio test.

- *fitting the model selected by AIC*

```
fit.AIC <- glm(y ~ factor(Fcolor) + factor(Fsurf), family = binomial(link = "logit"), data = Crabs)
summary(fit.AIC)
```

```
##
## Call:
## glm(formula = y ~ factor(Fcolor) + factor(Fsurf), family = binomial(link = "logit"),
##     data = Crabs)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1762  -0.8363  -0.7461   1.1948   1.9249
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -0.83271    0.21659  -3.845 0.000121 ***
## factor(Fcolor)3  0.82978    0.18167   4.568 4.93e-06 ***
## factor(Fcolor)5  0.29920    0.19964   1.499 0.133946
## factor(Fsurf)2  -0.03798    0.24699  -0.154 0.877793
## factor(Fsurf)3  -0.50960    0.23054  -2.210 0.027071 *
## factor(Fsurf)4  -0.60301    0.24715  -2.440 0.014694 *
## factor(Fsurf)5  -0.84935    0.28142  -3.018 0.002543 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1633.8  on 1344  degrees of freedom
## Residual deviance: 1585.9  on 1338  degrees of freedom
## AIC: 1599.9
##
## Number of Fisher Scoring iterations: 4
```

- *fitting the model selected by BIC*

```
fit.BIC <- glm(y ~ factor(Fcolor), family = binomial(link = "logit"), data = Crabs)
summary(fit.BIC)
```

```
##
## Call:
## glm(formula = y ~ factor(Fcolor), family = binomial(link = "logit"),
##     data = Crabs)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0100  -0.7670  -0.7449   1.3546   1.6838
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -1.07304    0.14259  -7.525 5.27e-14 ***
## factor(Fcolor)3  0.66567    0.17277   3.853 0.000117 ***
## factor(Fcolor)5 -0.06716    0.16953  -0.396 0.691971
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 1633.8  on 1344  degrees of freedom
## Residual deviance: 1601.0  on 1342  degrees of freedom
## AIC: 1607
##
## Number of Fisher Scoring iterations: 4
```

- *a) Setting the hypotheses*
- Since the model selected by BIC is smaller than the model selected by AIC, we can set the hypotheses as $H_0 : BIC\ model\ is\ better \quad vs \quad H_1 : AIC\ model\ is\ better$

- *b) Calculating the test statistic*
- We first get the residual deviance and residual degrees of freedom from each model.

```
dev.small <- fit.BIC$deviance
dev.large <- fit.AIC$deviance
df.small <- fit.BIC$df.residual
df.large <- fit.AIC$df.residual
```

- Then, we compute the test statistic.
- Here, the likelihood ratio test statistic is calculated from the residual deviance of the smaller model(BIC) and the residual deviance of the larger model(AIC). Its value can be calculated as,

```
LRT.stat <- dev.small - dev.large; LRT.stat
```

```
## [1] 15.06612
```

- As mentioned before, we also need the degrees of freedom for the chi-squared distribution in order to conduct the likelihood ratio test.
- The degrees of freedom for the likelihood ratio test is calculated from the degrees of freedom of the smaller model and the larger model. Its value can be calculated as,

```
LRT.df <- df.small - df.large; LRT.df
```

```
## [1] 4
```

c) Finding the p-value

```
1-pchisq(LRT.stat, df = LRT.df)
```

```
## [1] 0.004566022
```

d) Conclusion
- Since the p-value is less than 0.05, reject the null hypothesis $H_0 : BIC\ model\ is\ better$.
- Therefore, we can conclude that the model selected from backward elimination by AIC is better than the model selected from backward elimination by BIC.

```
1-pchisq(LRT.stat, df = LRT.df) < 0.05
```

```
## [1] TRUE
```

Categorical Data Analysis – Assignment 3

(Answers for question # 5 ~ question # 8)

2022KV0104 홍현덩

[Question #5]

Baseline-category logit model predicting preference for the US president (Democrats, Republican and Independent) using x= annual income (in $10,000 dollars)

$\Rightarrow \log\left(\frac{\hat{\pi}_D}{\hat{\pi}_I}\right) = 3.3 - 0.2x$, $\log\left(\frac{\hat{\pi}_R}{\hat{\pi}_I}\right) = 1.0 + 0.3x$

As we can see in the fitted model, the baseline category is I

[Question #5-(a)]

i) $\log\left(\frac{\hat{\pi}_R}{\hat{\pi}_D}\right) = \log\left(\frac{\hat{\pi}_R / \hat{\pi}_I}{\hat{\pi}_D / \hat{\pi}_I}\right) = \log\left(\frac{\hat{\pi}_R}{\hat{\pi}_I}\right) - \log\left(\frac{\hat{\pi}_D}{\hat{\pi}_I}\right)$

$= (1.0 + 0.3x) - (3.3 - 0.2x)$

$= (1.0 - 3.3) + (0.3 + 0.2x) = \boxed{-2.3 + 0.5x}$

ii) $\frac{\hat{\pi}_R}{\hat{\pi}_D} = e^{-2.3 + 0.5x} = e^{-2.3} \cdot e^{0.5x}$

There is a multiplicative effect of $e^{0.5} = 1.649$ on the estimated odds that the Republican is preferred rather than the Democrat.

[Question #5-(b)]

$\hat{\pi}_R = \frac{\exp(1.0 + 0.3x)}{1 + \exp(1.0 + 0.3x) + \exp(3.3 - 0.2x)}$,

$\hat{\pi}_D = \frac{\exp(3.3 - 0.2x)}{1 + \exp(1.0 + 0.3x) + \exp(3.3 - 0.2x)}$

(#5-(b) cont'd)

$$\hat{\pi}_R > \hat{\pi}_D \Leftrightarrow \exp(1.0+0.3x) > \exp(3.3-0.2x)$$

$$\Leftrightarrow \quad 1.0+0.3x > 3.3-0.2x$$

$$\Leftrightarrow \quad 0.5x > 2.3$$

$$\Leftrightarrow \quad x > \frac{2.3}{0.5} = 4.6$$

Therefore the range of x that satisfies $\hat{\pi}_R > \hat{\pi}_D$ is $\boxed{x > 4.6}$

[Question #5-(c)]

$$\log\left(\frac{\hat{\pi}_D}{\hat{\pi}_I}\right) = 3.3-0.2x, \quad \log\left(\frac{\hat{\pi}_R}{\hat{\pi}_I}\right) = 1.0+0.3x$$

From the model above, $\hat{\pi}_D = \exp(3.3-0.2x) \cdot \hat{\pi}_I$

and $\hat{\pi}_R = \exp(1.0+0.3x) \cdot \hat{\pi}_I$

$$\therefore \hat{\pi}_D + \hat{\pi}_R = 1 - \hat{\pi}_I = \left[\exp(3.3-0.2x) + \exp(1.0+0.3x)\right] \cdot \hat{\pi}_I$$

$$\Leftrightarrow \left[1+\exp(3.3-0.2x)+\exp(1.0+0.3x)\right]\hat{\pi}_I = 1$$

$$\therefore \hat{\pi}_I = \frac{1}{1+\exp(3.3-0.2x)+\exp(1.0+0.3x)}$$

Since the annual salary is \$65,000, $x = 6.5$

Therefore, the fitted probability $\hat{\pi}_I$ when $x = 6.5$ is,

$$\hat{\pi}_I = \frac{1}{1+\exp(3.3-0.2\times6.5)+\exp(1.0+0.3\times6.5)}$$

$$= \boxed{0.036}$$

[Question # 6]

[Question # 6-(a)]

Fitting the model to describe the effects of 1) size and 2) lake on primary food choice, considering lake George as a baseline lake.

↓

the prediction equations with fish as the baseline category are ...

(F: Fish, I: Invertebrate, R: Reptile, B: Bird, O: other)

baseline

(x_1: Hancock, x_2: Oklawaha, x_3: Trafford, S: Size)

$$\log\left(\frac{\hat{\pi}_I}{\hat{\pi}_F}\right) = -1.549 - 1.658x_1 + 0.937x_2 + 1.122x_3 + 1.458S$$

$$\log\left(\frac{\hat{\pi}_R}{\hat{\pi}_F}\right) = -3.315 + 1.243x_1 + 2.459x_2 + 2.935x_3 - 0.351S$$

$$\log\left(\frac{\hat{\pi}_B}{\hat{\pi}_F}\right) = -2.093 + 0.695x_1 - 0.653x_2 + 1.088x_3 - 0.631S$$

$$\log\left(\frac{\hat{\pi}_O}{\hat{\pi}_F}\right) = -1.904 + 0.826x_1 + 0.006x_2 + 1.516x_3 + 0.332S$$

[Question # 6-(b)]

The estimated odds for the choice between others and fish for the 3.5 meters alligator living in the lake Trafford.

sol) $\log\left(\frac{\hat{\pi}_O}{\hat{\pi}_F}\right) = -1.904 + 0.826x_1 + 0.006x_2 + 1.516x_3 + 0.332S$

$$\frac{\hat{\pi}_O}{\hat{\pi}_F} = \exp(-1.904 + 0.826x_1 + 0.006x_2 + 1.516x_3 + 0.332S)$$

since $S = 0$ and $(x_1 = 0, x_2 = 0, x_3 = 1)$,

$$\frac{\hat{\pi}_O}{\hat{\pi}_F} = \exp(-1.904 + 1.516) = \boxed{0.678}$$

[Question #6-(c)]

The estimated odds for the choice between reptiles and fish for the 1.5 meters alligator living in lake George.

sol) $\log\left(\frac{\hat{\pi}_R}{\hat{\pi}_F}\right) = -3.315 + 1.243x_1 + 2.459x_2 + 2.935x_3 - 0.351S$

$$\frac{\hat{\pi}_R}{\hat{\pi}_F} = \exp(-3.315 + 1.243x_1 + 2.459x_2 + 2.935x_3 - 0.351S)$$

since $S=1$ and $(x_1=0, x_2=0, x_3=0)$,

$$\frac{\hat{\pi}_R}{\hat{\pi}_F} = \exp(-3.315 - 0.351) = \boxed{0.026}$$

[Question #7]

Y = job satisfaction (4 ordered categories with 1: least satisfied)

x_1 = earnings compared to others with similar positions
(4 ordered categories with 1: much less, 4: much more)

x_2 = freedom to make decisions about how to do job
(4 ordered categories with 1: very true, 4: not at all true)

x_3 = work environment allows productivity
(4 ordered categories with 1: strongly agree, 4: strongly disagree)

↓ cumulative logit model

⇒ Prediction equation: $\text{logit}[P(\hat{Y} \le j)] = \hat{\alpha}_j - 0.54x_1 + 0.60x_2 + 1.19x_3$

common $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$ for $\forall j$
where $\hat{\beta}_1 = -0.54$, $\hat{\beta}_2 = 0.60$, $\hat{\beta}_3 = 1.19$

[Question #7-(a)]

(ex)
$$\log\left[\frac{P(\hat{Y}\le j \mid x_1=1, x_2, x_3) / P(\hat{Y} > j \mid x_1=1, x_2, x_3)}{P(\hat{Y}\le j \mid x_1=2, x_2, x_3) / P(\hat{Y} > j \mid x_1=2, x_2, x_3)}\right]$$

$$= \log\left[\frac{P(\hat{Y}\le j \mid x_1=1, x_2, x_3)}{P(\hat{Y} > j \mid x_1=1, x_2, x_3)}\right] - \log\left[\frac{P(\hat{Y}\le j \mid x_1=2, x_2, x_3)}{P(\hat{Y} > j \mid x_1=2, x_2, x_3)}\right]$$

$$= \cancel{\hat{\alpha}_j} - 0.54 + 0.60x_2 + 1.19x_3 - \cancel{\hat{\alpha}_j} + 2(0.54) + 0.60x_2 - 1.19x_3 = \boxed{0.54}$$

(Question #1-(a) cont'd)

i.e) $\log OR = 0.54$, $OR = e^{0.54} = 1.7116 > 1$

Therefore, job satisfaction tends to increase as x_1 increases.

[Question #1-(b)]

The values of x_1, x_2, x_3 at which a subject is most likely to have the highest job satisfaction

i) Job satisfaction tends to increase as x_1 increases since $\log OR = 0.54$

$\Leftrightarrow OR = e^{0.54} = 1.1716 > 1$

ii) Job satisfaction tends to decrease as

x_2 increases since $\log OR = -0.60 \Leftrightarrow OR = e^{-0.60} = 0.549 < 1$

$\because$ ex) $\log \left[\frac{P(\hat{Y} \le j \mid x_1, x_2 = 1, x_3) \,/\, P(\hat{Y} > j \mid x_1, x_2 = 1, x_3)}{P(\hat{Y} \le j \mid x_1, x_2 = 2, x_3) \,/\, P(\hat{Y} > j \mid x_1, x_2 = 2, x_3)} \right]$

$= 0.60 - 2 \cdot (0.60) = -0.60 \leftarrow$ this is the $\log OR$

iii) Job satisfaction tends to decrease as x_3 increases since

$\log OR = -1.19 \Leftrightarrow OR = e^{-1.19} = 0.304 < 1$

ex) $\log \left[\frac{P(\hat{Y} \le j \mid x_1, x_2, x_3 = 1) \,/\, P(\hat{Y} > j \mid x_1, x_2, x_3 = 1)}{P(\hat{Y} \le j \mid x_1, x_2, x_3 = 2) \,/\, P(\hat{Y} > j \mid x_1, x_2, x_3 = 2)} \right]$

$= 1.19 - 2(1.19) = -1.19 \leftarrow$ this is the $\log OR$

Therefore, a subject with $(x_1 = 4, x_2 = 1, x_3 = 1)$ is most likely to have the highest job satisfaction.

[Question #8]

Association b/w marital happiness and family income?

happiness : not, pretty, very — baseline

scores (1,2,3) for income categories.

⇒ baseline-category logit model

[Question #8-(a)]

$$\log\left(\frac{\hat{\pi}_n}{\hat{\pi}_v}\right) = -2.55518 - 0.22751x \quad , \quad \log\left(\frac{\hat{\pi}_p}{\hat{\pi}_v}\right) = -0.35129 - 0.09615x$$

where x = scores for income categories

[Question #8-(b)]

$$\hat{\pi}_v(x) = \frac{1}{1 + \exp(-2.55518 - 2.22751x) + \exp(-0.35129 - 0.09615x)}$$

where x = scores for income categories

[Question #8-(c)]

$$\log\left(\frac{\hat{\pi}_p}{\hat{\pi}_n}\right) = \log\left(\frac{\hat{\pi}_p / \hat{\pi}_v}{\hat{\pi}_n / \hat{\pi}_v}\right) = \log\left(\frac{\hat{\pi}_p}{\hat{\pi}_v}\right) - \log\left(\frac{\hat{\pi}_n}{\hat{\pi}_v}\right)$$

$$= -2.55518 - 0.22751x + 0.35129 + 0.09615x$$

$$= -2.20389 - 0.13136x$$