# Discrete analysis_HW6

ID : 111024517 Name : 鄭家豪

## Problem 1

**Question 1.**
For the data in the Question 2 of assignment 5, develop a score-based model and find some good fitting scores.

**Sol.**

Assume that the social class of the man in 1971 and 1981 are homogeneous. Based on the assigned score method, we only consider *linear-by-linear association* model using logit link.

$$\log E(Y_{ij}) = \log \mu_{ij} = \log n + \alpha_i + \beta_j + \gamma u_i v_j,$$

where $u_i's$, $v_j's$ are known scores, and $\gamma$ is an unknown parameter that represents the amount of association.

• Evenly spcaced socres

We assign evenly spaced scores — one(indicating "V") to six(indicating "I") for both class71 and class81:

```
data = read.table("cmob.txt")
data$class71 <- ordered( factor(data$class71,
                        levels = rev( unique(data$class71) )) )
data$class81 <- ordered( factor(data$class81,
                        levels = rev( unique(data$class81) )) )
data$score71 <- unclass(data$class71)
data$score81 <- unclass(data$class81)

fit_score <- glm(y ~ class71+class81+ I(score71*score81), family="poisson", data=data)
summary(fit_score)$coef["I(score71 * score81)",]
```

```
    Estimate   Std. Error      z value     Pr(>|z|)
7.377918e-01 6.296498e-03 1.171750e+02 0.000000e+00
```

We see that p-value is small and $\hat{\gamma}$ is positive, which, given the way that we have assigned evenly spcaced scores, means that a higher level of classes is associated with a greater probability of social mobility of men.

• Another Scores

We consider that the social class "IIIN" and "IIIM" are equality. Then, we assign scores — one("V"), two("IV"), three("IIIM" or "IIIN"), four("II"), and five("I") for both class71 and class81:

```
new_score <- c(1,2,3,3,4,5)
fit_score2 <- glm(y ~ class71+class81+ I(new_score[score71]*new_score[score81]),
                  family="poisson", data=data)
summary(fit_score2)$coef["I(new_score[score71] * new_score[score81])",]
```

```
   Estimate   Std. Error      z value     Pr(>|z|)
 1.28895311   0.01147818 112.29598566   0.00000000
```

We see that p-value is small and $\hat{\gamma}$ is positive, the same conclusion for this assigned score model is same with the equally spaced score model.

Finally, we compare two score-based model to the independence model:

```
nomod <- glm(y ~ class71+class81, data, family= "poisson")
anova(nomod, fit_score, fit_score2, test="Chi")
```

```
Analysis of Deviance Table

Model 1: y ~ class71 + class81
Model 2: y ~ class71 + class81 + I(score71 * score81)
Model 3: y ~ class71 + class81 + I(new_score[score71] * new_score[score81])
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1        25      42127
2        24      14719  1  27407.9 < 2.2e-16 ***
3        24      19448  0  -4728.8
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We find that using the ordinal information gives us more power to detect an association. The appropriate model is equally spaced score model.

## Problem 2

**Question 2.**
Suppose that for the data in the Question 3 of assignment 4, the marginal totals of each age groups had been fixed in advance by design.

a. Treat marital status as a nominal variable and build a multinomial model to describe how marital status changes with age.
b. Give an understandable interpretation of the important parameter estimates.
c. Under your model, predict the probability that a Dane, aged 55, is divorced.

**Sol.**

We first check whether the marginal total for each age group is greater than 5 to apply the asymptotic property.

```
data = read.table("maritaldane.txt",header = T)
data[c("Age", "Total")]
```

```
    Age Total
1 17-21    18
2 21-25    24
3 25-30    26
4 30-40    32
5 40-50    32
6 50-60    28
7 60-70    16
8   70+     9
```

Now, before fitting the multinomial logit model, we should take the midpoint of the age interval as the predictor.

```
response  <- as.matrix(data[,c('Single', 'Married', 'Divorced')])
age = c(19, 23, 27.5, 35, 45, 55, 65, 75)

library(nnet)
mmod <- multinom(response ~ age )
```

```
# weights:  9 (4 variable)
initial  value 203.243273
iter  10 value 155.213073
iter  10 value 155.213072
final  value 155.213072
converged
```

```
mmodi <- step(mmod)
```

```
Start:  AIC=318.43
response ~ age

trying - age
# weights:  6 (2 variable)
initial  value 203.243273
final  value 184.477247
converged
       Df      AIC
<none>  4 318.4261
- age   2 372.9545
```

```
pchisq( deviance(mmodi) , df = sum(data$Total) - mmodi$edf, lower.tail = F)
```

```
[1] 7.089467e-09
```

Based on the deviance relative to null model, the variable age is significant.

**(a)**

```
result=predict(mmodi,data.frame(age=age),type="probs")
row.names(result) = data$Age
result
```

```
          Single   Married   Divorced
17-21 0.64449469 0.3350588 0.02044655
21-25 0.57163979 0.3980652 0.03029504
25-30 0.48500251 0.4692152 0.04578229
30-40 0.34240231 0.5730054 0.08459232
40-50 0.18635798 0.6475827 0.16605932
50-60 0.08749267 0.6313121 0.28119522
60-70 0.03626482 0.5433544 0.42038075
70+   0.01352787 0.4208743 0.56559787
```

As the above predicted probability results, we can observe that
– For single status, the probability will decrease as age increases.

– For Married status, the probability will be unimodal.

– For Divorced status, the probability will increase as age increases.

**(b)**

```
summary(mmodi)
```

```
Call:
multinom(formula = response ~ age)

Coefficients:
          (Intercept)        age
Married    -2.042430 0.07306683
Divorced   -5.888008 0.12828185

Std. Errors:
          (Intercept)        age
Married    0.5421681 0.01561171
Divorced   0.9217066 0.02069053

Residual Deviance: 310.4261
AIC: 318.4261
```

The intercept terms model the probabilities of marital status for an age of zero, however the predicted probability of status for zero age is ridiculous. The slope terms represent the log-odds of moving from the baseline category of "Single" to "Married" and "Divorced", respectively.
Specifically, for example age=35, the log-odds in favor of category "Married" over category "Divorced" is

$$\log\left(\frac{P(Y = Married|age = 35)}{P(Y = Divorced|age = 35)}\right) = -2.0424 + 5.8880 + 35 \times (0.0731 - 0.1283) = 1.9136.$$

**(c)**

```
predict(mmodi,data.frame(age=55),type="probs")
```

```
    Single    Married   Divorced
0.08749267 0.63131211 0.28119522
```

The desired predicted probability is approximately 0.2812.

## Problem 3

**Question 3.**
Here is a data on the balance of subjects, which were observed for two different surfaces and for restricted and unrestricted vision. Balance was assessed qualitatively on an *ordinal* 4-point scale based on observation by the experimenter. Subjects were expected to be better balanced (show less sway) when standing on the normal surface than on foam, and when their eyes were open rather than closed or when their vision was restricted by a dome. The variables in the dataset are:

| Variable | Description |
|---|---|
| Subject | 1 to 40 |
| Sex | male or female |
| Age | Age of subject in years |
| Height | Height in cm |
| Weight | Weight in kg |
| Surface | normal or foam |
| Vision | eyes open, eyes closed, or closed dome |
| CTSIB | Qualitive measure of balance, 1 (stable) - 4 (unstable) |

a. Determine which factors affect balance.
b. Describe the nature of the effects.
c. Predict the response for 25 year old female, height 170cm, weight 65kg on a normal surface with eyes closed.

You may ignore the fact that there are repeated observations on the same subjects (which would likely cause dependence in the response). You may also collapse categories with very small numbers of observations.

**Sol.**

### (a)

First, we should check the number of each category, and collapse categories with a small number into adjacent categories. Then, the observation of blance category 4 is collapsed to blance 3.

```
data= read.table("ctsib.txt", header = T)
table( data$CTSIB)
```

```
  1   2   3   4
114 292  73   1
```

```
data$CTSIB[which( data$CTSIB == 4 )] = 3
data$CTSIB = ordered( factor(data$CTSIB,
                        levels = rev( unique(data$CTSIB) )) )
unique(data$CTSIB)
```

```
[1] 1 2 3
Levels: 3 < 2 < 1
```

As a side note, we assign blance 1 is the highest level. Subsequently, we can fit proportional odds model using logit link for better interpretability.

```
## proportioanl odds
library(MASS)
pomod <- polr(CTSIB ~ .-Subject , data)
pomodi <- step(pomod,trace = 0)
summary(pomodi)
```

```
Call:
polr(formula = CTSIB ~ Sex + Height + Weight + Surface + Vision,
    data = data)

Coefficients:
              Value Std. Error t value
Sexmale      1.1463    0.38243   2.997
```

```
Height      -0.1016    0.02108  -4.821
Weight       0.0675    0.01434   4.706
Surfacenorm  4.4733    0.40573  11.025
Visiondome   0.4976    0.25911   1.920
Visionopen   3.7200    0.36566  10.174

Intercepts:
    Value    Std. Error t value
3|2 -12.1714   3.0213    -4.0286
2|1  -5.9584   2.9541    -2.0170

Residual Deviance: 501.3754
AIC: 517.3754
```

For proportional odds model, using an AIC-based variable selection method, we have 5 important factors: Sex, Height, Weight, Surface, and Vision.

```
pchisq( deviance(pomodi)-deviance(pomod) , pomod$edf-pomodi$edf,lower=F)
```

```
[1] 0.6413474
```

The simplification to model five factors is justifiable. For checking the proportional odds assumption, with respect to Height, we computed the log-odds difference between $\gamma_1$ and $\gamma_2$, where $\gamma_j = P(blance \leq 4 - j)$ with $\gamma_1 = 1$.

```
library(faraway)
pim <- with(data,prop.table(table(Height, CTSIB),1))
logit(pim[,1])-logit(pim[,1]+pim[,2])
```

```
      142       150       154       159     161.5       163       164     165.5
       NA -3.218876 -2.302585 -1.985915 -1.609438        NA        NA -2.708050
      166       167     167.5       168     168.5       170       172       173
-2.734368 -3.353407        NA        NA -1.386294 -4.744932        NA -3.091042
    174.5     175.5       176     176.5       177       180       181     182.5
       NA        NA -3.654194        NA        NA -3.091042 -2.302585 -3.218876
      183       184       185       187       190
-3.091042 -4.744932        NA -2.282382        NA
```

As a side note, the realization NA indicates that $\gamma_1 = 0$. The proportional odds assumption is not violated significantly, at least there is no trend in the above table.

**(b)**

From (a), we can say that the odds of moving from balance 1 to other categories of balance changes by a factors of $\exp(\widehat{\beta}_x)$ as $x$ changes by one unit. For example, holding Height, Weight, Surface, and Vision constant, the odds of moving from balance 1 to other categories of balance changes by a factors of $\exp(1.1463) = 3.146529$ from female to male.

**(c)**

The desired prediction probability is as follows:

```
result= predict(pomodi,data.frame(Sex="female", Height=170, Weight= 65,
                          Surface = "norm", Vision= "closed"), type="probs")
round(result, 4)
```

```
     3      2      1
0.0228 0.8981 0.0791
```

For the reference, we predict the probability when her eyes were open or when her vision was restricted by a dome, with holding other predictors being the same as before:

```
ref1= round( predict(pomodi,data.frame(Sex="female", Height=170, Weight= 65,
                          Surface = "norm", Vision= "open"), type="probs"),4 )
ref2 =round( predict(pomodi,data.frame(Sex="female", Height=170, Weight= 65,
                          Surface = "norm", Vision= "dome"), type="probs"),4 )
rbind(ref1, ref2)
```

```
          3      2      1
ref1 0.0006 0.2195 0.7799
ref2 0.0140 0.8622 0.1238
```

As expected, there have better balanced compared to the condition with Vision= “closed”.