HWA2, Analysis of Categorical Data

1. (2pt) Consider a $2 \times 2 \times K$ table where

$$\pi_{ik} \equiv P(Y = 1 \mid X = i, Z = k) \quad = \quad \Phi\left(\alpha + \beta x_i + \beta_k^Z\right), \quad k = 1, ..., K,$$

with $x_1 = 1$ and $x_2 = 0$. Here $\Phi()$ is the cumulative distribution function of a standard normal random variable.Which of the following statements is/are correct, and why?

   (a) If $\beta \neq 0$, the model implies homogeneous XY assocation.

   (b) β is the log odds ratio in partial table k.

   (c) If $\beta = 0$, the model implies $X \perp Y \mid Z$?

   Solution: The conditional odds ratio is

$$\frac{\pi_{1k}(1 - \pi_{2k})}{\pi_{2k}(1 - \pi_{1k})} \quad = \quad \frac{\Phi\left(\alpha + \beta + \beta_k^Z\right)\left[1 - \Phi\left(\alpha + \beta_k^Z\right)\right]}{\Phi\left(\alpha + \beta_k^Z\right)\left[1 - \Phi\left(\alpha + \beta + \beta_k^Z\right)\right]},$$

   which depends on k. Hence, we do not neccesarily have homogenous association.
   We can also see that the odds ratio cannot be simplified to $\exp(\beta)$. Hence, β is not the log odds ratio.
   If $\beta = 0$, then the odds ratio becomes

$$\frac{\pi_{1k}(1 - \pi_{2k})}{\pi_{2k}(1 - \pi_{1k})} \quad = \quad \frac{\Phi\left(\alpha + \beta_k^Z\right)\left[1 - \Phi\left(\alpha + \beta_k^Z\right)\right]}{\Phi\left(\alpha + \beta_k^Z\right)\left[1 - \Phi\left(\alpha + \beta_k^Z\right)\right]} = 1.$$

   Hence, $X \perp Y \mid Z$. Alternatively,

$$\begin{aligned} P(Y = 1 \mid Z = k) &= P(X = 1, Y = 1 \mid Z = k) + P(X = 2, Y = 1 \mid Z = k) \\ &= P(Y = 1 \mid X = 1, Z = k) P(X = 1 \mid Z = k) + \\ &\quad P(Y = 1 \mid X = 2, Z = k) P(X = 2 \mid Z = k) \\ &= \Phi\left(\alpha + \beta + \beta_k^Z\right) P(X = 1 \mid Z = k) + \Phi\left(\alpha + \beta_k^Z\right) P(X = 2 \mid Z = k). \end{aligned}$$

   If $\beta = 0$, then

$$\begin{aligned} P(Y = 1 \mid Z = k) &= \Phi\left(\alpha + \beta_k^Z\right) P(X = 1 \mid Z = k) + \Phi\left(\alpha + \beta_k^Z\right) P(X = 2 \mid Z = k) \\ &= \Phi\left(\alpha + \beta_k^Z\right), \end{aligned}$$

   which is the same as

$$P(Y = 1 \mid X = i, Z = k) \quad = \quad \Phi\left(\alpha + \beta_k^Z\right).$$

   It means that $X \perp Y \mid Z$.

2. (2pt) Consider a multiway $2 \times 2 \times K \times M$ table

$$\log\left(\frac{\pi_{ikm}}{1 - \pi_{ikm}}\right) \quad = \quad \alpha + \beta x_i + \beta_k^Z + \beta_m^W, \quad k = 1, ..., K, \; m = 1, ..., M,$$

   where $x_1 = 1$, $x_2 = 0$, and

$$\pi_{ikm} \quad = \quad P(Y = 1 \mid X = i, Z = k, W = m).$$

   There are in total KM partial tables.

(a) Does the model imply homogenous XY association?
Solution: The conditional odds ratio is

$$\begin{aligned}\frac{\pi_{1km}\left(1-\pi_{2km}\right)}{\pi_{2km}\left(1-\pi_{1km}\right)} &= \frac{\frac{\exp\left(\alpha+\beta+\beta_k^Z+\beta_m^W\right)}{1+\exp\left(\alpha+\beta+\beta_k^Z+\beta_m^W\right)}\frac{1}{1+\exp\left(\alpha+\beta_k^Z+\beta_m^W\right)}}{\frac{\exp\left(\alpha+\beta_k^Z+\beta_m^W\right)}{1+\exp\left(\alpha+\beta_k^Z+\beta_m^W\right)}\frac{1}{1+\exp\left(\alpha+\beta+\beta_k^Z+\beta_m^W\right)}} \\ &= \exp\left(\beta\right).\end{aligned}$$

Hence, it implies homogeneous XY association.

(b) If $\beta_k^Z=0$ and $\beta_m^W=0$ for all k and m, do we have $Y \perp (Z,W) \mid X$?
Solution: Note that

$$\begin{aligned}P\left(Y=y \mid X=i\right) &= \sum_k\sum_m P\left(Y=y \mid X=i, Z=k, W=m\right) P\left(Z=k, W=m \mid X=i\right) \\ &= \sum_k\sum_m \frac{\exp\left(\alpha+\beta x_i+\beta_k^Z+\beta_m^W\right)}{1+\exp\left(\alpha+\beta x_i+\beta_k^Z+\beta_m^W\right)} P\left(Z=k, W=m \mid X=i\right) \\ &= \frac{\exp\left(\alpha+\beta x_i\right)}{1+\exp\left(\alpha+\beta x_i\right)}\sum_k\sum_m P\left(Z=k, W=m \mid X=i\right) \\ &= \frac{\exp\left(\alpha+\beta x_i\right)}{1+\exp\left(\alpha+\beta x_i\right)}\end{aligned}$$

Hence, $Y \perp (Z,W) \mid X$.

(c) If only $\beta_k^Z=0$ for all k, is Y conditionally independent of any variable(s)?
Solution: Note that

$$\begin{aligned}P\left(Y=y \mid X=i, W=m\right) &= \sum_k P\left(Y=y \mid X=i, Z=k, W=m\right) P\left(Z=k \mid X=i, W=m\right) \\ &= \sum_k \frac{\exp\left(\alpha+\beta x_i+\beta_m^W\right)}{1+\exp\left(\alpha+\beta x_i+\beta_m^W\right)} P\left(Z=k \mid X=i, W=m\right) \\ &= \frac{\exp\left(\alpha+\beta x_i+\beta_m^W\right)}{1+\exp\left(\alpha+\beta x_i+\beta_m^W\right)}\sum_k P\left(Z=k \mid X=i, W=m\right) \\ &= \frac{\exp\left(\alpha+\beta x_i+\beta_m^W\right)}{1+\exp\left(\alpha+\beta x_i+\beta_m^W\right)}.\end{aligned}$$

Hence, $Y \perp Z \mid (X,W)$.

3. (1pt) Consider the model

$$\log\left(\frac{\pi_{ik}}{1-\pi_{ik}}\right) = \alpha+\beta x_i,$$

for a $2\times 2\times K$ model. We said in the slides that $Y \perp Z \mid X$. Under what conditions will Y be marginally independent of Z?
Solution: Note that the conditional ZY association and marginal ZY association will be the same under the collapsibility conditions ($Z \perp X \mid Y$ or $X \perp Y \mid Z$). Take $X \perp Y \mid Z$ as an example. Note that the conditional XY association is

$$\log\left(\frac{\pi_{ik}/\left(1-\pi_{ik}\right)}{\pi_{i+1,k}/\left(1-\pi_{i+1,k}\right)}\right) = \alpha+\beta x_i-\left(\alpha+\beta x_{i+1}\right)=\beta\left(x_i-x_{i+1}\right).$$

If we have $\beta=0$, then the conditional XY association is 1 which means that $X \perp Y \mid Z$. Hence, by the collapsibility conditions, the conditional ZY association is the same as the marginal ZY association. The conditional ZY association given $X=i$ is

$$\log\left(\frac{\pi_{ik}/\left(1-\pi_{ik}\right)}{\pi_{i,k+1}/\left(1-\pi_{i,k+1}\right)}\right) = \alpha+\beta x_i-\left(\alpha+\beta x_i\right)=0.$$

As a result, the marginal ZY association is also 1.
Alternatively, note that

$$\begin{aligned} P(Y=y \mid Z=k) &= \sum_i P(X=i, Y=y \mid Z=k) \\ &= \sum_i P(Y=y \mid X=i, Z=k) P(X=i \mid Z=k) \\ \text{(use logit model)} &= \sum_i \frac{\exp[y(\alpha+\beta x_i)]}{1+\exp(\alpha+\beta x_i)} P(X=i \mid Z=k) \end{aligned}$$

but

$$\begin{aligned} P(Y=y) &= \sum_i \sum_k P(X=i, Y=y, Z=k) \\ &= \sum_i \sum_k P(Y=y \mid X=i, Z=k) P(X=i, Z=k) \\ \text{(use logit model)} &= \sum_i \sum_k \frac{\exp[y(\alpha+\beta x_i)]}{1+\exp(\alpha+\beta x_i)} P(X=i, Z=k) \\ &= \sum_i \frac{\exp[y(\alpha+\beta x_i)]}{1+\exp(\alpha+\beta x_i)} \sum_k P(X=i, Z=k) \\ &= \sum_i \frac{\exp[y(\alpha+\beta x_i)]}{1+\exp(\alpha+\beta x_i)} P(X=i) \end{aligned}$$

which are not the same in general. But if we have $P(X=i \mid Z=k) = P(X=i)$ for any i and k , that is $X \perp Z$, then we will have $Z \perp Y$.

4. (1pt) Consider a multiway table relating occupational aspirations to socioeconomic status and different groups. The dataset can be downloaded from Studium. Test conditional independence of socioeconomic status and occupational aspiration given groups using a logit model, and using the Cochran-Mantel-Haenszel test.
Solution:

```
## We need to arrage data as a data frame
Data.GLM <- NULL
group <- NULL
socio <- NULL
for(k in 1 : 12){
  group <- c(group, rep(as.character(k), 2))
  socio <- c(socio, c("High", "Low"))
  Data.GLM <- rbind(Data.GLM, t(Task4[, , k]))
}
Data.GLM <- data.frame(group = group,
                       socio = socio,
                       high = Data.GLM[, "Highaspiration"],
                       low = Data.GLM[, "Lowaspiration"])
```

We can test conditional independence using the Wald test, under the assumption of homogeneous association.

```
Logit <- glm(cbind(high, low) ~ socio + group, data = Data.GLM, family = binomial)
summary(Logit)
```

```
##
## Call:
## glm(formula = cbind(high, low) ~ socio + group, family = binomial,
##     data = Data.GLM)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.70891  -0.60589  -0.00132   0.87006   2.48995
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.91379    0.13003   7.027 2.11e-12 ***
## socioLow    -1.69984    0.07737 -21.971  < 2e-16 ***
## group10     -0.24829    0.15988  -1.553  0.12042
## group11      0.95338    0.20581   4.632 3.62e-06 ***
## group12     -0.05698    0.20424  -0.279  0.78025
## group2      -0.53829    0.17463  -3.082  0.00205 **
## group3       0.70800    0.16060   4.409 1.04e-05 ***
## group4      -0.03613    0.16144  -0.224  0.82291
## group5       0.91193    0.21099   4.322 1.54e-05 ***
## group6       0.08930    0.23407   0.381  0.70283
## group7       0.03103    0.17225   0.180  0.85705
## group8      -0.77890    0.17234  -4.520 6.19e-06 ***
## group9       0.92655    0.16244   5.704 1.17e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1331.509  on 23  degrees of freedom
## Residual deviance:   25.194  on 11  degrees of freedom
## AIC: 174.94
##
## Number of Fisher Scoring iterations: 4
```

The Wald test suggests that socioeconomic status and occupational aspiration are not conditionally independent given groups. To apply the Cochran-Mantel-Haenszel test,

```
mantelhaen.test(Task4)
```

```
##
##  Mantel-Haenszel chi-squared test with continuity correction
##
## data:  Task4
## Mantel-Haenszel X-squared = 528.8, df = 1, p-value < 2.2e-16
## alternative hypothesis: true common odds ratio is not equal to 1
## 95 percent confidence interval:
##  4.641810 6.287857
## sample estimates:
## common odds ratio
##          5.402503
```

It is seen that we still reject the null hypothesis of conditional independence.

5. (2pt) Consider a multiway table relating occupational aspirations to gender, residence type, IQ, and socioeconomic status. The dataset can be downloaded from Studium.

   (a) Fit binomial data models with different link functions: logit, probit, and cloglog. Only main effects are present in the model. Which link function do you prefer?
   Solution:

```
## Fit different models
Logit <- glm(cbind(Highaspiration, Lowaspiration) ~ gender + residence + IQ + socio,
             data = Task5, family = binomial(link = "logit"))
Probit <- glm(cbind(Highaspiration, Lowaspiration) ~ gender + residence + IQ + socio,
             data = Task5, family = binomial(link = "probit"))
Cloglog <- glm(cbind(Highaspiration, Lowaspiration) ~ gender + residence + IQ + socio,
             data = Task5, family = binomial(link = "cloglog"))
```

   We can compare different models using information criterion

```
## AIC
c(AIC(Logit), AIC(Probit), AIC(Cloglog))

## [1] 171.4678 172.8714 180.4333

## BIC
c(BIC(Logit), BIC(Probit), BIC(Cloglog))

## [1] 178.5362 179.9397 187.5016
```

   Both AIC and BIC favor the logit link. We can also consider the residuals, which is skipped here.

   (b) Suppose that we fitted the following models

```
Logit <- glm(cbind(Highaspiration, Lowaspiration) ~ gender * residence * IQ * socio,
             data = Task5, family = binomial(link = "logit"))
Probit <- glm(cbind(Highaspiration, Lowaspiration) ~ gender * residence * IQ * socio,
             data = Task5, family = binomial(link = "probit"))
Cloglog <- glm(cbind(Highaspiration, Lowaspiration) ~ gender * residence * IQ * socio,
             data = Task5, family = binomial(link = "cloglog"))
```

   Which model fits the data best?
   Solution: All models are saturated models. Hence, they are the same.

   (c) Suppose that the logit link is used. The effect of gender (binary) to occupational aspirations (binary) is our focus, and other variables are confounders. Test the hypothesis that the conditional odds ratio

$$\frac{\pi_{1lst}/(1-\pi_{1lst})}{\pi_{2lst}/(1-\pi_{1lst})}$$

   is the same across all partial tables, where

$$\pi_{ilst} = P(Y=1 \mid \text{gender}=i, \text{residence}=l, \text{IQ}=s, \text{socio}=t).$$

   Be explicit which model you have fitted, if any.
   Solution: We will fit the following logit model

$$\log\left(\frac{\pi_{ilst}}{1-\pi_{ilst}}\right) = \alpha + \beta x_i + \beta_l^R + \beta_s^{IQ} + \beta_t^S + \beta_{ls}^{R,IQ} + \beta_{lt}^{RS} + \beta_{st}^{IQ,S} + \beta_{lst}^{R,IQ,S},$$

   where $x_1 = 1$ and $x_2 = 0$, with all interactions among the confounders. Note that

$$\frac{\pi_{ilst}}{1-\pi_{ilst}} = \exp\left(\alpha + \beta x_i + \beta_l^R + \beta_s^{IQ} + \beta_t^S + \beta_{ls}^{R,IQ} + \beta_{lt}^{RS} + \beta_{st}^{IQ,S} + \beta_{lst}^{R,IQ,S}\right)$$

and

$$\frac{\pi_{1lst}/(1-\pi_{1lst})}{\pi_{2lst}/(1-\pi_{2lst})} = \exp(\beta).$$

Hence, we have homogeneous XY association. To fit the model, we use

```
Logit <- glm(cbind(Highaspiration, Lowaspiration) ~ gender + residence * IQ * socio,
             family = binomial, data = Task5)
Logit

##
## Call:  glm(formula = cbind(Highaspiration, Lowaspiration) ~ gender +
##     residence * IQ * socio, family = binomial, data = Task5)
##
## Coefficients:
##                           (Intercept)                             genderMale
##                              1.346019                               0.350041
##                         residenceRural                  residenceSmall urban
##                             -0.877172                              -0.147850
##                                 IQLow                               socioLow
##                             -1.880609                              -1.856629
##                  residenceRural:IQLow           residenceSmall urban:IQLow
##                              0.322171                               0.065860
##              residenceRural:socioLow        residenceSmall urban:socioLow
##                              0.560341                              -0.005975
##                         IQLow:socioLow        residenceRural:IQLow:socioLow
##                              0.025819                              -0.143079
## residenceSmall urban:IQLow:socioLow
##                             -0.120925
##
## Degrees of Freedom: 23 Total (i.e. Null);  11 Residual
## Null Deviance:     1747
## Residual Deviance: 23.3  AIC: 171.1
```

We will investigate whether the model fits the data well. One alternative is to use the residual deviance. Note that

```
qchisq(0.95, 11)

## [1] 19.67514
```

Hence, we reject the null hypothesis that the model fits the data as well as the saturated model at $\alpha = 0.05$.

6. (1pt) Consider the model

$$\log\left(\frac{P(Y_i=1)}{1-P(Y_i=1)}\right) = \alpha + \beta x_i, \quad i = 1, ..., n,$$

where $x_i = 1$ or 0. Let β be the focus parameter. Derive the expression of the conditional likelihood for β.
Solution: By the factorization theorem of sufficient statistics,

$$P(Y_1 = y_1, \cdots, Y_n = y_n) = \frac{\exp\left[\left(\sum_{i=1}^N y_i\right)\alpha + \left(\sum_{i=1}^N y_i x_i\right)\beta\right]}{\prod_{i=1}^N [1+\exp(\alpha+\beta x_i)]}$$

implies that $\left(\sum_{i=1}^N y_i, \sum_{i=1}^N y_i x_i\right)$ is a sufficient statistic for (α, β). Let

$$S = \left\{(y_1^*, \cdots, y_N^*) : \sum_{i=1}^N y_i^* = t_0\right\}.$$

Then,

$$\begin{aligned}
& P\left(Y_1 = y_1, \cdots, Y_n = y_n \mid \sum_{i=1}^N y_i = t_0\right) \\
= & \frac{P\left(Y_1 = y_1, \cdots, Y_n = y_n, \sum_{i=1}^N y_i = t_0\right)}{P\left(\sum_{i=1}^N y_i = t_0\right)} \\
= & \frac{\frac{\exp\left[t_0\alpha + \left(\sum_{i=1}^N y_i x_i\right)\beta\right]}{\prod_{i=1}^N [1+\exp(\alpha+\beta x_i)]}}{\sum_S \frac{\exp\left[\left(\sum_{i=1}^N y_i^*\right)\alpha + \left(\sum_{i=1}^N y_i^* x_i\right)\beta\right]}{\prod_{i=1}^N [1+\exp(\alpha+\beta x_i)]}} \\
= & \frac{\exp\left[\left(\sum_{i=1}^N y_i x_i\right)\beta\right]}{\sum_S \exp\left[\left(\sum_{i=1}^N y_i^* x_i\right)\beta\right]}.
\end{aligned}$$

which depends only on β_1.

7. (1pt) Consider the three-way table $2 \times 2 \times K$ in Task7.RData, where the response variable is response. Under the logit link, estimate the effect of gender.
Solution: The response is either 0 or 1. It is seen that K is large and the row sums in each partial table are $(1, 1)$. Hence, we will use conditional ML, since ML will be biased.

```
library(survival)
clogit(response ~ gender + strata(group), data = Task7)

## Call:
## clogit(response ~ gender + strata(group), data = Task7)
##
##          coef exp(coef) se(coef)     z        p
## gender 1.2262    3.4082   0.1625 7.547 4.45e-14
##
## Likelihood ratio test=68.13  on 1 df, p=< 2.2e-16
## n= 1000, number of events= 440
```