Psy 5011: Categorical Data Analysis **(Due: 26/05/20)**

Homework #7

Instructor: Yao, Kaiping Grace *Name:* Hao-Cheng Lo, *Id:* D08227104

Problem 7.2:

(a)

There is strong evidence that the estimated conditional DP odds ratio is $e^{-.867} = 0.4198$ at each each level of V, indicating the odds of defendants receiving death penalty for defendants being Whites were less than half the odds for defendants being Blacks, for each race of victims.

(b)

The residual deviance is 0.37984 (df = 1) and P-value = .538 > .05. Hence, this homogeneous-association-model fits data quite well.

(c)

```
DP <- read.table("http://users.stat.ufl.edu/~aa/cat/data/DeathPenalty.dat", header=TRUE)
fit <- glm(count ~ D + V + P + D:V + D:P + P:V, family=poisson, data=DP)
summary(fit) # homogeneous association model
# Coefficients:
#               Estimate Std. Error z value Pr(>|z|)
# (Intercept)    4.93578    0.08471  58.265  < 2e-16 ***
# Dwhite        -2.17465    0.26377  -8.245  < 2e-16 ***
# Vwhite        -1.32980    0.18479  -7.196 6.19e-13 ***
# Pyes          -3.59610    0.50691  -7.094 1.30e-12 ***
# Dwhite:Vwhite  4.59497    0.31353  14.656  < 2e-16 ***
# Dwhite:Pyes   -0.86780    0.36707  -2.364   0.0181 *
# Vwhite:Pyes    2.40444    0.60061   4.003 6.25e-05 ***
#
#     Null deviance: 1225.07955  on 7  degrees of freedom
# Residual deviance:    0.37984  on 1  degrees of freedom
# AIC: 52.42
```

The results above are same as Table 7.13.

```
DP2 <- data.frame(V = c("white","white","black","black"), D = c("white","black","white","black"),
    yes = c(53,11,0,4), no = c(414,37,16,139))
fit2 <- glm(yes/(no+yes) ~ V + D, family = binomial, weights = no+yes, data = DP2)
summary(fit2)
# Coefficients:
#             Estimate Std. Error z value Pr(>|z|)
# (Intercept)  -3.5961     0.5069  -7.094 1.30e-12 ***
# Vwhite        2.4044     0.6006   4.003 6.25e-05 ***
# Dwhite       -0.8678     0.3671  -2.364   0.0181 *
#
#     Null deviance: 22.26591  on 3  degrees of freedom
# Residual deviance:  0.37984  on 1  degrees of freedom
# AIC: 19.3
```

The estimated effects of $D = -.8678$ and $V = 2.4044$ relate to loglinear model estimates by the relations $\beta_i^D = \lambda_{i1}^{DP} - \lambda_{i2}^{DP} \rightarrow -.8678 = -.8678 - 0$ and $\beta_k^V = \lambda_{k1}^{VP} - \lambda_{k2}^{VP} \rightarrow 2.4044 = 2.4044 - 0$.

Which model is plausible depends on the purpose. If we regard these three variables as response variables and care about the associations among this categorical variables, we may adopt the loglinear model. On the other hand, if we regard one variable as response variable and rest of them as explanatory variables and care about the effects of explanatory variables on the response variable, we may adopt the logistic regression model.

Problem 7.4:

```
mbti <- read.table("http://users.stat.ufl.edu/~aa/intro-cda/data/MBTI.dat", header=TRUE)
fit3 <- glm(n ~ EI + SN + TF + JP + EI:SN + EI:TF + EI:JP + SN:TF + SN:JP + TF:JP, family=poisson,
    data=mbti)
summary(fit3) # homogeneous association model
# Coefficients:
#              Estimate Std. Error z value Pr(>|z|)
# (Intercept)  3.44760    0.13793  24.994  < 2e-16 ***
# EIi         -0.02907    0.15266  -0.190 0.848952
# SNs          1.21082    0.14552   8.320  < 2e-16 ***
# TFt         -0.64194    0.16768  -3.828 0.000129 ***
# JPp          0.93417    0.14594   6.401 1.54e-10 ***
# EIi:SNs      0.30212    0.14233   2.123 0.033780 *
# EIi:TFt      0.19449    0.13121   1.482 0.138258       \\(ii)
# EIi:JPp      0.01766    0.13160   0.134 0.893261       \\(ii)
# SNs:TFt      0.40920    0.15243   2.684 0.007265 **
# SNs:JPp     -1.22153    0.14547  -8.397  < 2e-16 ***   \\(i)
# TFt:JPp     -0.55936    0.13512  -4.140 3.48e-05 ***
# ---
# Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1
#     Null deviance: 399.944  on 15  degrees of freedom
# Residual deviance:  10.162  on  5  degrees of freedom
# AIC: 125
```

From the results above, we can find that the deviance = 10.162, $df = 5$ and P-value = .071, so this model fits not bad. As for the (i) which marked in the results, the estimated conditional log odds ratio of SN and JP is -1.22 at each level of any combination of remain variables, which is significant. As for the (ii) which marked in the results, the estimated contional associations of EI and TF and of EI and JP do not reach the significant level.

Problem 7.6:

(a)

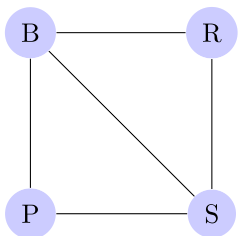
```
BPRS <- read.table("http://users.stat.ufl.edu/~aa/cat/data/BPRS.dat", header=TRUE)
BPRS$B <- factor(BPRS$B);BPRS$P <- factor(BPRS$P);BPRS$R <- factor(BPRS$R);BPRS$S <- factor(BPRS$S)
fit4 <- glm(count~B+P+R+S, family=poisson, data=BPRS)
fit5 <- glm(count~B+P+R+S+B*P+B*R+B*S+P*R+P*S+R*S, family=poisson, data=BPRS)
fit6 <- glm(count~B+P+R+S+B*P+B*R+B*S+P*R+P*S+R*S+B*P*R+B*R*S+P*R*S, family=poisson, data=BPRS)
fit4.s <- summary(fit4); fit5.s <- summary(fit5); fit6.s <- summary(fit6)
reslt76 <- data.frame(Model=c("single","twoway","threeway"),Dev=c(fit4.s$deviance,fit5.s$deviance,
    fit6.s$deviance),df=c(fit4.s$df.residual,fit5.s$df.residual,fit6.s$df.residual),AIC=c(fit4.s$aic
    ,fit5.s$aic,fit6.s$aic))
# Model     Dev         df  AIC
# single    277.0846    18  413.5247   \\P-value < .00001
# twoway    6.963067     9  161.4031   \\P-value = .64097
# threeway  2.041843     4  166.4819   \\P-value = .72807
```

By observing the table in the results, the single-factor model, which implies mutual independence of the four variables, fits very poorly. The homogeneous association two-factor model fits much better (deviance = 6.9, *df* = 9). The three factor model fits well as well (deviance = 2.4, *df* = 4). Further, we can inspect the AICs of last two models. We can find that the AIC of two-factor model is smaller than it of three-factor model. Hence, we select the two-factor model.

```
fit5.s
# Coefficients:
#              Estimate Std. Error z value Pr(>|z|)
# (Intercept)   4.5650     0.0950  48.054  < 2e-16 ***
# B2           -1.7824     0.1802  -9.892  < 2e-16 ***
# P2           -0.1729     0.1335  -1.295 0.195261
# P3           -0.7375     0.1522  -4.845 1.27e-06 ***
# R2           -0.2290     0.1327  -1.726 0.084379 .
# S2           -2.3672     0.2029 -11.668  < 2e-16 ***
# B2:P2         0.3048     0.1893   1.610 0.107377
# B2:P3         0.9288     0.1936   4.797 1.61e-06 ***
# B2:R2         0.5979     0.1626   3.678 0.000235 ***
# B2:S2         1.1468     0.1532   7.488 6.99e-14 ***
# P2:R2         0.2583     0.1729   1.494 0.135261
# P3:R2         0.3441     0.1883   1.827 0.067658 .
# P2:S2         0.7199     0.1952   3.688 0.000226 ***
# P3:S2         0.8018     0.2031   3.948 7.87e-05 ***
# R2:S2         1.1459     0.1698   6.749 1.49e-11 ***
```

Two factor model indicates conditional association of any level of arbitrary two variables at each combination of levels of the remaining variables are the same. For example, in the results shown above, there is a strong evidence that the odds of thinking sex relations before marriage is always or almost always wrong than wrong only sometimes or not wrong at all for those do religious service attendance at least several times a year are $e^{1.1459} = 3.14$ times the odds for those do religious service attendance at most a few times a year, at each combination of levels of political views and levels of birth control availability to teenagers between the ages of 14 and 16.

(b)



Employeeing the rule of *collapsibility for multiway contingency tables*, we consider the sets $\{P\}$, $\{B, S\}$, and $\{R\}$. For this model, every path between P and R involves a variable in $\{B, S\}$. Collapsing over the variables R, the PB and PS conditional associations are the same as with the model (PB, PS, BS). That is, collapsing R or not, the conditional odds ratio of PB or PS are the same at each level of the remaining variable. Similarly, RB and RS have same rationale. Noted that *no* any of two variable have identical conditional association and marginal association, when collapsing remaining **both** variables.

Problem 7.6:

(c)

```r
BPRS2<-data.frame(b=c(1,1,1,1,1,1,2,2,2,2,2,2),
                  p=c(1,1,2,2,3,3,1,1,2,2,3,3),
                  r=c(1,2,1,2,1,2,1,2,1,2,1,2),
                  right=c(99,73,73,87,51,51,15,25,20,37,19,36),
                  wrong=c(8,24,20,20,6,33,4,22,13,60,12,88))
fit7 <- glm(right/(right+wrong) ~ b + factor(p) + r, family=binomial, weights=wrong+right,data=BPRS2)
summary(fit7)
# Coefficients:
#             Estimate Std. Error z value Pr(>|z|)
# (Intercept)   4.6695     0.3842  12.154  < 2e-16 ***
# b            -1.3743     0.1579  -8.706  < 2e-16 ***
# factor(p)2   -0.4246     0.2030  -2.091 0.036510 *
# factor(p)3   -0.7755     0.2049  -3.785 0.000154 ***
# r            -0.9635     0.1741  -5.535 3.11e-08 ***
# ---
# Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1
#
# (Dispersion parameter for binomial family taken to be 1)
#
#     Null deviance: 190.922  on 11  degrees of freedom
# Residual deviance:  15.944  on  7  degrees of freedom
# AIC: 77.334
```

With a deviance = 15.944 on 7 degrees of freedom (P-value $< .05$), the logistic model indicates some lack of fit.

Problem 7.8:

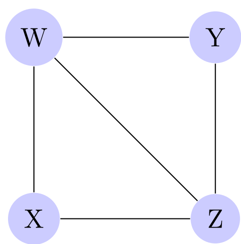
Corresponding loglinear Model:

$$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ}$$

Conditionally independent: XY and ZY.

Marginal association same as conditional association: XZ, ZY, and XY.

Problem 7.10:



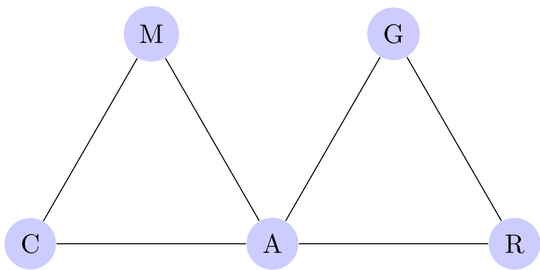
Variables X and Y are conditionally independent given both W and Z.

Problem 7.12:

(a)

Obeying the rule of *collapsibility for multiway contingency tables*, we consider the sets $\{M\}$, $\{C, A, G\}$, and $\{R\}$. For this model, every path between M and R involves a variable in $\{C, A, G\}$. Hence, AM conditional odds ratio is unchanged by collapsing over race. On the contrary, we can not find a mutually exclusive partition way to let $A \in S_1$, $M \in S_2$, and $G \in S_3$, such that S_1 seperates S_2 and S_3 or S_2 seperates S_1 and S_3. Hence, AM conditional odds ratio is not unchanged by collapsing over gender.

(b)



We can trivially find that every path between $\{C, M\}$ and $\{G, R\}$ involves A, so $\{G, R\}$ are separated from $\{C, M\}$ by A. All conditional associations among A, C, and M are then identical to those in model (AC, AM, CM), collapsing over G and R. Relating to the rule of *collapsibility for multiway contingency tables*, this is because C is independent of both G and R, and so is M, the AC, AM, and CM conditional associations at the level of remaining one keeps identical mo matter collapsing R and G or not.

Problem 7.14:

(a)

True.

(b)

False. Loglinear models are more appropriate.

(c)

True.