Psy 5011: Categorical Data Analysis **(Due: 02/06/20)**

Homework #8

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

Problem 8.1:

```
tab <- matrix(c(159,22,8,14),nrow = 2,ncol = 2)
mcnemar.test(tab, correct = FALSE)
#   McNemar's Chi-squared test
#
# data:  tab
# McNemar's chi-squared = 6.5333, df = 1, p-value = 0.01059
```

From the results above, the McNemar chi-squared $= 6.53$ (P-value $= 0.011$), indicating that low birth weight (cases) than normal birth weight (controls) are more likely to be smokers.

Problem 8.3:

```
Opinions <- read.table("http://users.stat.ufl.edu/~aa/cat/data/Opinions.dat", header=TRUE)
fit_marginal <- gee(y ~ question, id=person, family=binomial(link=logit), data=Opinions)
print(summary(fit_marginal))
# Coefficients:
#               Estimate Naive S.E.    Naive z Robust S.E.   Robust z
# (Intercept) -0.8858933 0.06505597 -13.617401  0.06502753 -13.623357
# question     0.1035319 0.09107816   1.136737  0.06397794   1.618244
fit_subject <- glmer(y ~ (1|person) + question, family=binomial, nAGQ=50, data=Opinions)
print(summary(fit_subject))
# Random effects:
#  Groups Name        Variance Std.Dev.
#  person (Intercept) 8.143    2.854
# Number of obs: 2288, groups:  person, 1144
#
# Fixed effects:
#             Estimate Std. Error z value Pr(>|z|)
# (Intercept)  -1.8343     0.1624 -11.295   <2e-16 ***
# question      0.2100     0.1301   1.614    0.106
```

Marginal model focuses on the marginal distributions of the two responses, and its β is the log odds ratio comparing the marginal distributions (i.e. the log odds ratio for the combined sample). Subject-specific model focuses on the partial table of the two responses conditional on the subject, and its β is the log odds ratio conditional on the subject. To illustrate that, using the environment data, we can find the above results where for the marginal model has $\hat{\beta} = \log((359/785)/(334/810)) = 0.104$. and the subject-specific model has $\hat{\beta} = \log(132/107) = 0.21$.

Problem 8.7:

```
Survey <- data.frame(I=c(1,1,1,0,0,0),II=c(-1,0,0,1,1,0),III=c(0,-1,0,-1,0,1),IV=c(0,0,-1,0,-1,-1),
    nij=c(49,3,288,2,186,12),nji=c(129,1,81,0,9,4))
symm <- glm(nij/(nij+nji) ~ -1, family=binomial, weights=nij+nji, data=Survey)
print(summary(symm))
# Residual deviance: 365.81  on 6  degrees of freedom
res <- rstandard(symm,type="pearson")
#         1         2         3         4         5         6
# -5.996254  1.000000 10.775990  1.414214 12.675233  2.000000
QS <- glm(nij/(nij+nji) ~ -1 + I + II + III + IV, family=binomial, weights=nij+nji, data=Survey)
print(summary(QS)) # quasi-symmetry model
# Residual deviance:   5.5347  on 3  degrees of freedom
```

For (a), Symmetry model has residual deviance $= 365.81$ ($df = 6$) which is such a bad fit. From the results we can observe that if one can do more shifts from II to I (resstandard $= 6.0$), I to IV (resstandard $= 10.8$), II to IV (resstandard $= 12.7$) then it would be symmetry.

For (b), QS model has residual deviance $= 5.53$ ($df = 3$) which indicating fits well. QS compares to symmetry model resulting in difference $= 365.81 - 5.53 = 360.3$ ($df = 6 - 3 = 3$ and P-value $< .001$), providing strong evidence against marginal homogeneity. There are some reasons that includes changes in II (Catholic) or IV (none or other) and large sample size.

Problem 8.11:

```
coffee <- data.frame(first=c(1,1,1,1,1,2,2,2,2,2,3,3,3,3,3,4,4,4,4,4,5,5,5,5,5),second=c
    (1,2,3,4,5,1,2,3,4,5,1,2,3,4,5,1,2,3,4,5,1,2,3,4,5),count=c
    (93,17,44,7,10,9,46,11,0,9,17,11,155,9,12,6,4,9,15,2,10,4,12,2,27),diag=c
    (1,0,0,0,0,0,2,0,0,0,0,0,3,0,0,0,0,0,4,0,0,0,0,0,5))
ordinary <- glm(count~factor(first)+factor(second),family=poisson,data=coffee)
print(summary(ordinary))
# Residual deviance: 346.38  on 16  degrees of freedom
quasi <- glm(count~factor(first)+factor(second)+factor(diag),family=poisson,data=coffee)
print(summary(quasi))
# Residual deviance:  13.786  on 11  degrees of freedom
```

For the ordinary model, we find that model fits bad with r.d. $= 346.4$ ($df = 16$). Yet, for the quasi-independence model, we find that model fits quite well with r.d. $= 13.8$ ($df = 11$). Hence, we can say that the second choice of coffe brand is independent to the first one, under quasi-independence model..