MAS3906 Project

Q1.

We define the categorical variable *sex* as follows:

$$g = \begin{cases} 0, & male \\ 1, & female \end{cases}$$

First, we fit a logistic regression model for the binomial response data using:

```
> model = glm(cbind(hyper, total - hyper) ~ factor(sex) + smoking + obesity
+ snoring, family = binomial, data = snoring)
```

with the linear predictor being:

$$\eta_i = \beta_0 + \beta_1 g_i + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i}$$

And the summary is as follows:

```
Call:
glm(formula = cbind(hyper, total - hyper) ~ factor(sex) + smoking +
    obesity + snoring, family = binomial, data = snoring)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4569  -0.4381  -0.0414   0.3405   2.9874

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.34774    0.25053  -9.371  < 2e-16 ***
factor(sex)2  0.09573    0.19450   0.492  0.62260
smoking       0.01459    0.21854   0.067  0.94679
obesity       0.82539    0.22033   3.746  0.00018 ***
snoring       0.76947    0.24217   3.177  0.00149 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From the above R output, the following $\hat{\beta}$ values are significant:

$$\widehat{\beta_0} = -2.348, \widehat{\beta_4} = 0.825, \widehat{\beta_5} = 0.769$$

corresponding to *obesity* (highly significant) and *snoring* (significant at 1% level).

In terms of explaining the presence of hypertension, it does appear that *obesity* is the most important variables, followed by *snoring*. Currently there is no evidence of any effect of *sex* and *smoking*.

Then we fit an ordered model to do analysis of deviance using:

```
> model.order = glm(cbind(hyper, total - hyper) ~ obesity + snoring +
factor(sex) + smoking, family = binomial, data = snoring)
> anova(model.order, test = "Chisq")
```

Here is the Analysis of Deviance Table:

```
             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                         14     44.123
obesity       1  17.7480        13     26.375 2.522e-05 ***
snoring       1  11.1458        12     15.230 0.0008422 ***
factor(sex)   1   0.2380        11     14.992 0.6256598
smoking       1   0.0044        10     14.987 0.9468257
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Based on the above table, we draw the following conclusions:

- For the row corresponding to *obesity*, the related model is
$$\eta_i = \beta_0 + \beta_3 x_{3i}$$
The p-value is 2.522×10^{-5} for testing $H_0: \beta_3 = 0 \; vs \; H_1: \beta_3 \neq 0$.
We reject H_0, concluding *hypertension* depends on *obesity*.

- For the row corresponding to *snoring*, the related model is
$$\eta_i = \beta_0 + \beta_3 x_{3i} + \beta_4 x_{4i}$$
The p-value is 0.0008 for testing $H_0: \beta_4 = 0 \; vs \; H_1: \beta_4 \neq 0$.
Again, we reject H_0 and include *snoring* in the model.

- For *factor(sex)* the model is
$$\eta_i = \beta_0 + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_1 g_i$$
The p-value is 0.626 for testing $H_0: \beta_1 = 0 \; vs \; H_1: \beta_1 \neq 0$.
We do not reject H_0, so *hypertension* does not appear to depend on *sex*. We do not include *sexing* the model.

- Similar arguments indicate we do not need *smoking* in the model, although strictly speaking we might need to test this with *factor(sex)* excluded.

Therefore, We include only *obesity* and *snoring* as covariates in the final model, and we use the following code to give a summary:

```
> model.final = glm(cbind(hyper, total - hyper) ~ obesity + snoring, family
= binomial, data = snoring)
> summary(model.final)

Call:
glm(formula = cbind(hyper, total - hyper) ~ obesity + snoring,
    family = binomial, data = snoring)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.30565  -0.54291   0.01857   0.34509   3.05284

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.2815     0.2092 -10.906  < 2e-16 ***
obesity       0.8193     0.2198   3.728 0.000193 ***
snoring       0.7462     0.2349   3.177 0.001488 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 44.123  on 14  degrees of freedom
Residual deviance: 15.230  on 12  degrees of freedom
AIC: 70.501
```

The final fitted model is
$$\eta_i = -2.2815 + 0.8193 \times x_{3i} + 0.7462 \times x_{4i}$$
From this model we conclude that since $\beta_3 = 0.8193 > 0$, *obesity* is positively related to *hypertension*.
Finally, we have $\beta_4 = 0.7462 > 0$, which means the more snoring, the higher probability of *hypertension*.

Q2.

In this question, we use the same notations as in the lecture and the only difference is $l \in \{1, 2\}$ here instead of $\{1, 2, 3\}$ in the lecture as we have removed the city London. And we set up the basic data using the following code:

```
> setwd("~/Desktop/MAS3906")
> da=read.table("ulcer2.txt",header=T)
> da$place = gl(2, 4)
> da$case = gl(2, 2, 8)
> da$blood = gl(2, 1, 8)
```

We are told that the only response variable is *blood* group, so the association relating to *case* and *place* must be included. Thus, the minimal model here which contains only marginal effects and any interactions built into the study design is:

$$\log(\mu_{jkl}) = \mu + \alpha_j + \beta_k + \gamma_l + (\alpha\gamma)_{jl}$$

(a) To answer this question we need to test the interaction between *case* and *blood*. The code and analysis of deviance table are as follows:

```
> ulcermod = glm(no ~ place + case + blood + place*case + case*blood, data
= da, family = poisson)
> anova(ulcermod, test = "Chisq")

Analysis of Deviance Table

Model: poisson, link: log

Response: no

Terms added sequentially (first to last)

           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                          7    20865.3
place       1    617.7         6    20247.6 < 2.2e-16 ***
case        1  19973.2         5      274.4 < 2.2e-16 ***
blood       1    250.0         4       24.4 < 2.2e-16 ***
place:case  1     16.2         3        8.2 5.698e-05 ***
case:blood  1      5.4         2        2.8   0.01997 *
```

We consider the model:

$$\log(\mu_{jkl}) = \mu + \alpha_j + \beta_k + \gamma_l + (\alpha\gamma)_{jl} + (\alpha\beta)_{jk}$$

We test $H_0 : \forall\, pair(j,k),\ (\alpha\beta)_{jk} = 0\ \ versus\ \ H_1 : \exists\, pair(j,k), (\alpha\beta)_{jk} \neq 0$. R gives $p = 0.01997 < 0.05$, we therefore marginally reject H_0 at the 5% level and conclude that the probability of having ulcers is associated with blood type.

(b) Then we test the 3-way interaction and here are the code and updated table:

```
> anova(update(ulcermod, . ~ . + blood*place + place*blood*case), test =
"Chisq")

Analysis of Deviance Table

Model: poisson, link: log

Response: no

Terms added sequentially (first to last)
```

```
                 Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                 7    20865.3
place             1    617.7         6    20247.6 < 2.2e-16 ***
case              1  19973.2         5      274.4 < 2.2e-16 ***
blood             1    250.0         4       24.4 < 2.2e-16 ***
place:case        1     16.2         3        8.2 5.698e-05 ***
case:blood        1      5.4         2        2.8   0.01997 *
place:blood       1      1.7         1        1.0   0.19111
place:case:blood  1      1.0         0        0.0   0.30725
```

We consider the maximal model:

$$\log(\mu_{jkl}) = \mu + \alpha_j + \beta_k + \gamma_l + (\alpha\gamma)_{jl} + (\alpha\beta)_{jk} + (\beta\gamma)_{kl} + (\alpha\beta\gamma)_{jkl}$$

We test $H_0 : \forall\, pair(j,k,l),\ (\alpha\beta\gamma)_{jkl} = 0\ \ versus\ \ H_1: \exists\, pair(j,k,l), (\alpha\beta\gamma)_{jkl} \neq 0$. The p-value is $0.30725 > 0.1$. Thus, we cannot reject H_0 and there is no evidence that the association between ulcers and blood type differs from city to city.

(c) To tell the nature of the significant effects, we need to look at the sign of the corresponding coefficients. The code and summary table are as follows:

```
> summary(update(ulcermod, . ~ . + blood*place + place*blood*case), test =
"Chisq")

Call:
glm(formula = no ~ place + case + blood + place:case + case:blood +
    place:blood + place:case:blood, family = poisson, data = da)

Deviance Residuals:
[1]  0  0  0  0  0  0  0  0

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         5.50533    0.06376  86.348  < 2e-16 ***
place2              0.16799    0.08661   1.940   0.0524 .
case2               2.73082    0.06580  41.500  < 2e-16 ***
blood2              0.38355    0.08267   4.639  3.5e-06 ***
place2:case2        0.16393    0.08920   1.838   0.0661 .
case2:blood2       -0.20078    0.08556  -2.347   0.0189 *
place2:blood2      -0.07546    0.11312  -0.667   0.5048
place2:case2:blood2  0.11914    0.11672   1.021   0.3074
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From above we can see that $\widehat{(\alpha\beta)}_{22}$ is significantly negative. So there are fewer controls of blood type B than expected, which means blood group B are more prone to ulcers than blood group A in this question.