STAT8016 Assignment 2
Name: LAM Hin Tai
UID: 2004062587

Q1

Please refer to the hand-written page.

Q2

(a) This is a cross-sectional study as the total number of cases is not fixed, nor is the total number of each age group is fixed by the investigator.

(b) Rearranging the columns:

Age	No Dysentery	Dysentery	Row Total
< 1	30	20	50
1 to 5	95	105	200
6 to 10	120	80	200
11 to 18	130	70	200
19 to 50	420	130	550
> 50	75	25	100
Column Total	870	430	1300

The number of concordant pairs is

N_C = (30)(105+80+70+130+25) + (95)(80+70+130+25) + (120)(70+130+25)
+ (130)(130+25) + (420)(25)
= 98925

The number of discordant pairs is

N_D = (30)(105+80+70+130+25) + (95)(80+70+130+25) + (120)(70+130+25)
+ (130)(130+25) + (420)(25)
= 189425

The estimate of the Goodman and Kruskal's γ is

$$\hat{\gamma} = \frac{98925 - 189425}{98925 + 189425} = -0.31385$$

(c) The table is collapsed into this 2 x 2 table:

Age Group	No Dysentery	Dysentery	Row Total
Teenagers	375	275	650
Adults	495	155	650
Column Total	870	430	1300

The estimated relative risk of dysentery (for adults to teenagers) is

$$\hat{\rho} = \frac{155/650}{275/650} = 0.563636$$

The 95% C.I. for the relative risk of dysentery (for adults to teenagers) is

$$0.563636 \times \exp\left(\pm 1.96\sqrt{\frac{495}{155 \times 650} + \frac{375}{275 \times 650}}\right)$$

$$= 0.563636 \times \exp(\pm 0.164115)$$
$$= (0.47833, 0.66416)$$

(d) Since (1) the Goodman and Kruskal's γ is negative for the association between age and the dysentery attack rate, and (2) the relative risk of dysentery is less than 1 for adults to teenagers, we can say that an increasing age is negatively associated with the dysentery attack rate. That is, the younger the person is, the more likely the person will be affected by dysentery.

(e) Yes, we can estimate the difference in the proportions of dysentery between teenagers and adults, because the proportions have direct estimates using observed counts in cross-sectional studies and prospective studies. The proportion estimate of cases for teenagers is 155/650 while that for adults is 275/650.

(f) From the sample, we can see the proportions of dysentery cases look to be decreasing as age increases:

Age	Group	Proportion of Dysentery
< 1	0	0.4
1 to 5	1	0.525
6 to 10	2	0.4
11 to 18	3	0.35
19 to 50	4	0.23636
> 50	5	0.25

The hypothesis being tested is whether there is a trend in the proportions of dysentery cases across the age levels.

Let p_i be the proportion of dysentery cases in the i-th age group, i = 0, …, 5, then the hypotheses are:

$$H_0: \quad p_0 = p_1 = p_2 = p_3 = p_4 = p_5$$
$$H_1: \quad p_0 > p_1 > p_2 > p_3 > p_4 > p_5$$

Since the p-value for the one-sided test is < 0.0001, we reject the null hypothesis and conclude that the proportion of dysentery cases decreases as age increases.

(g) (i) R output:

```
> fit1 <- glm(formula = Dysentery ~ Group, family = binomial(link = "logit"),
data = dysentery)
> summary(fit1)

Call:
glm(formula = Dysentery ~ Group, family = binomial(link = "logit"), 
    data = dysentery)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.2723  -0.8857  -0.7743   1.2191   1.7865  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept)  0.22029    0.13786   1.598     0.11    
Group       -0.31788    0.04382  -7.255 4.03e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1650.3  on 1299  degrees of freedom
Residual deviance: 1596.4  on 1298  degrees of freedom
AIC: 1600.4

Number of Fisher Scoring iterations: 4
```

The fitted model is

$$\ln\left(\frac{p}{1-p}\right) = 0.22029 - 0.31788Group$$

The estimated coefficient $\hat{\beta}_1$ for Group is -0.31788, so exp(-0.31788) = 0.72769. This means that for one unit increase in the age group, there is a decrease of the odds of getting dysentery by 1 - 0.72769 = 27.23%.

(ii) R output:

```
> OR.1.to.3 <- exp((1-3)*fit1$coefficients["Group"])
> OR.1.to.3
   Group 
1.888457 
> fit1Summary <- summary(fit1)
> OR.SE <- fit1Summary$coefficients["Group","Std. Error"]
> OR.SE
[1] 0.04381787
> lowerCI <- OR.1.to.3 * exp((1-3) * 1.96 * OR.SE)
> lowerCI
   Group 
1.590413 
> upperCI <- OR.1.to.3 * exp((1-3) * -1.96 * OR.SE)
> upperCI
   Group 
2.242353 
> exp(confint.default(fit1, level=0.95)* -2)
                2.5 %    97.5 %
(Intercept) 1.104977 0.3749392
Group       2.242346 1.5904182
```

The odds ratio of dysentery infection for subjects in Group 1 to those in Group 3 is

$$OR(1|3) = \exp((1-3) \times -0.31788) = 1.888457$$

with standard error 0.04382.

The 95% C.I. for the odds of dysentery infection for subjects in Group 1 to those in Group 3 is

$$\exp[(1-3)\times -0.31788 \pm (1-3)\times 1.96 \times 0.04382] = (1.5904, 2.2424)$$

(iii) R output:

```
> Z <- fit1$coefficients["Group"] / OR.SE
> Z
    Group
-7.254573
> chiSq <- Z^2
> chiSq
   Group
52.62883
> chiSq.CV <- qchisq(0.95, 1)
> chiSq.CV
[1] 3.841459
```

H_0: $\beta_1 = 0$ vs H_1: $\beta_1 < 0$. Using the Wald test, the test statistic is

$$W = \left(\frac{-0.31788}{0.04382}\right)^2 = (-7.25457)^2 = 52.6288$$

This follows the Chi-squared(1) distribution under the null hypothesis, with critical value 3.8415 at 5% significance level. Therefore, we reject the null hypothesis and conclude that $\beta_1 < 0$.

This is consistent with the results in part (f), where we concluded that as the age group increases, the proportion of cases dysentery decreases – a negative association. Here the negative coefficient also shows that as age group increases, the odds of dysentery infection decreases (by 27.23%).

Q3

(a) The predicted probability of presence of CHD is

$$\hat{p} = \frac{\exp(-5.3095 + 0.1109AGE)}{1 + \exp(-5.3095 + 0.1109AGE)}$$

(b) The predicted probability of presence of CHD, for a male subject aged 60, is

$$\hat{p}_{AGE=60} = \frac{\exp(-5.3095 + 0.1109 \times 60)}{1 + \exp(-5.3095 + 0.1109 \times 60)} = 0.79323$$

(c) The odds ratio of CHD for male versus female is estimated as

$$\exp(-5.3095 + 6.127) = \exp(0.8175) = 2.26483$$

(d) No, AGE does not confound with gender and CHD in the calculation of part (c). The odds of CHD for male for each unit increase of AGE is the same as that for female, and so the AGE factors from each separate gender model cancel out each other. There is no change in the odds ratio associated with AGE across gender.

Appendix – R code

```
# STAT8016 Assignment 2 Q2
# Author: LAM Hin Tai
# UID: 2004062587

setwd("C:/Users/Tai/Documents/MStat 2018/S5 8016 Biostatistics/Assignment
2")
dysentery <- read.csv("dysentery.csv", header=TRUE)

# (g)(i)
fit1 <- glm(formula = Dysentery ~ Group, family = binomial(link = "logit"),
data = dysentery)
summary(fit1)


# (g)(ii)
OR.1.to.3 <- exp((1-3)*fit1$coefficients["Group"])
OR.1.to.3

# SE
fit1Summary <- summary(fit1)
OR.SE <- fit1Summary$coefficients["Group","Std. Error"]
OR.SE

lowerCI <- OR.1.to.3 * exp((1-3) * 1.96 * OR.SE)
lowerCI

upperCI <- OR.1.to.3 * exp((1-3) * -1.96 * OR.SE)
upperCI

# Alternatively by function call
exp(confint.default(fit1, level=0.95)* -2)


# (g)(iii)
Z <- fit1$coefficients["Group"] / OR.SE
Z
chiSq <- Z^2
chiSq

chiSq.CV <- qchisq(0.95, 1)
chiSq.CV
```