STAT 475/675: Midterm Exam 2 (Mar 15, 2018)

Name: *Student ID #:*

Question 1. [12 points] The data in Table 1 were from a prospective cohort study. The study aimed to evaluate the association between coronary heart disease (CHD) and serum cholesterol level (SCL).

Table 1: Data OF A Coronary Heart Disease (CHD) Study

	Serum Cholesterol Level (SCL, mg/100 ml)			
CHD Status	<190	190-219	220-249	≥ 250
yes	41	68	106	172
not	1022	1203	1119	1125

Q1.1 Below is part of the R outputs of the logistic regression analysis of the data, taking CHD status (Y) as the response with two categories "yes" and "not" and SCL level (X) as the explanatory variable with four categories "<190", "190-219", "220-249" and "≥ 250".

```
Call: glm(formula = chd ~ scl, family = binomial, weights = counts)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    3.2159     0.1593  20.191  < 2e-16 ***
sclSCL190-219 -0.3429     0.2023  -1.695     0.09 .
sclSCL220-249 -0.8592     0.1889  -4.548 5.43e-06 ***
sclSCL>249    -1.3379     0.1791  -7.471 7.97e-14 ***
```

(i) Give an approximate 95% confidence interval (CI) for the probability of having CHD among people with SCL below 190 (mg/100ml).

(ii) Give an approximate 95% CI for the odds ratio (OR) of having CHD among people with SCL above 249 (mg/100ml) vs below 190 (mg/100ml).

(i) ∵ a 95% CI of the intercept term (α) is $\hat{\alpha} \pm (1.96) \cdot \hat{SE}(\hat{\alpha})$

⇒ An approximate 95% CI of the prob is $(0.029, 0.052)$

(ii) ∵ a 95% CI of the main effect "SCL>249" vs "SCL<190" in the model is $\hat{\beta}_4^{scl} \pm (1.96)\, \hat{SE}(\hat{\beta}_4^{scl})$

⇒ An approximate 95% CI of the OR is $(2.683, 5.414)$

Q1.2 Below is part of the R output of the Poisson (loglinear) regression analysis of Table 1 data under the model of independence, taking the cell counts (V) as the response and both the CHD status (Y) with two categories "yes" and "not" and SCL (X) with four categories "<190", "190-219", "220-249" and "$\geq$ 250" as the two explanatory variables.

```
Call: glm(formula = counts ~ scl + chd, family = poisson)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.05301    0.05761  52.995  < 2e-16 ***
sclSCL190-219 0.17871    0.04156   4.300 1.71e-05 ***
sclSCL220-249 0.14185    0.04192   3.384 0.000715 ***
sclSCL>249    0.19896    0.04137   4.809 1.52e-06 ***
chdno         2.44650    0.05299  46.171  < 2e-16 ***
```

(i) Give the fitted model.

(ii) Give an approximate 95% CI for the expected count of people with SCL below 190 (mg/100ml) having CHD.

(iii) Give the maximum likelihood estimate of the OR of having CHD among people with SCL above 249 (mg/100ml) vs below 190 (mg/100ml).

(iv) Why is the estimate of (iii) not in agreement with the CI obtained in **Q1.1**(ii)?

(i) The model considered is "$V \sim$ loglinear (X, Y)", that is

$$\log \mu(x,y) = \alpha + \beta_x^{SCL} + \beta_y^{CHD}, \text{ with } x=1,2,3,4 \text{ for SCL<190, 190-219, 220-249, >249}$$

and $y = 0, 1$ for yes, not CHD

$\Rightarrow$ the fitted model is

$$\log \hat{\mu}(x,y) = \hat{\alpha} + \hat{\beta}_x^{SCL} + \hat{\beta}_y^{CHD} \text{ with } \hat{\alpha} = 3.05$$

$\hat{\beta}_1^{SCL} \equiv 0$, $\hat{\beta}_2^{SCL} = 0.18$, $\hat{\beta}_3^{SCL} = 0.14$, $\hat{\beta}_4^{SCL} = 0.20$

and $\hat{\beta}_0^{CHD} \equiv 0$, $\hat{\beta}_1^{CHD} = 2.45$

(ii) $\because$ an approximate 95% CI of α is $\hat{\alpha} \pm (1.96)\,SE(\hat{\alpha})$

$\therefore$ An approximate 95% CI of the count is $(18.92, 23.71)$

(iii) The MLE of the OR is 1.

(iv) The current analysis is under the model of independence, which assumes SCL $\perp\!\!\!\perp$ CHD. The CI in Q1.1(ii) indicates a strong association between the two variables.

Question 2. [8 points] Table 2 provides the information of the CHD study in **Question 1** concerning the study subjects' age (Z) and sex (W) as the two additional variables may confound or modify the association of the CHD status (Y) with the SCL (X).

Table 2: Data OF A Coronary Heart Disease (CHD) Study

		Serum Cholesterol Level (SCL, mg/100 ml)							
		<190		190-219		220-249		≥ 250	
Sex	Age Group	CHD		CHD		CHD		CHD	
	(years)	yes	no	yes	no	yes	no	yes	no
male	30-49	13	327	18	390	40	381	57	305
	50-62	13	110	33	143	35	139	49	134
female	30-49	6	536	5	547	10	402	18	339
	50-62	9	49	12	123	21	197	48	347

Q2.1 (i) Based on the provided information in Table 2, describe the group of study subjects with size 33 regarding their SCL, CHD status, age and sex. (ii) Give the sample OR of having CHD among females in age 30-49 with SCL above 249 (mg/100ml) vs below 190 (mg/100ml).

(i) The 33 subjects were male, aged 50-62 year-old, with SCL 190-219 mg/100ml and having CHD.

(ii) The sample OR is 4.743

Q2.2 Below is part of the R outputs of the logistic regression analysis of Table 2 data, taking CHD status (Y) as the response with two categories "yes" and "not" and all the other variables (SCL, age, and sex) as the explanatory variables.

```
Call: glm(formula = chd ~ scl + sex + age, family = binomial, weights = counts)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)     3.0832     0.1705  18.082  < 2e-16 ***
scl190-219     -0.2462     0.2059  -1.196  0.23176
scl220-249     -0.7040     0.1928  -3.652  0.00026 ***
scl>249        -1.1614     0.1843  -6.301 2.96e-10 ***
sexfemale       1.1000     0.1162   9.467  < 2e-16 ***
ageage(50-62)  -1.1345     0.1113 -10.195  < 2e-16 ***
```

(i) Give the MLE for the odds ratio (OR) of having CHD among females in age group 30-49 with SCL above 249 (mg/100ml) vs below 190 (mg/100ml).

(ii) Should the estimate in (i) be the same as the one in **Q2.1**(ii)? Why?

(i) The MLE of the OR is $\exp(-\hat{\beta}_4^{scl}) = 3.19$

(ii) In general they are different unless CHD and SCL's association is homogeneous across different sex and age groups.

Q2.3 Below is part of the R output of the Poisson (loglinear) regression analysis of the contingency table (Table 2) under the model of homogeneous association, including only all the main effects and two factor interactions of the four variables of Table 2.

```
Call: glm(formula = counts ~ scl * chd + scl * sex + scl * age + chd *
    sex + chd * age + sex * age, family = poisson)

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                      2.87934    0.16843  17.095  < 2e-16 ***
sclSCL190-219                    0.35124    0.20425   1.720 0.085486 .
sclSCL220-249                    0.64574    0.19201   3.363 0.000771 ***
sclSCL>249                       0.72725    0.18475   3.936 8.27e-05 ***
chdno                            3.02759    0.16959  17.852  < 2e-16 ***
sexfemale                       -0.74545    0.12879  -5.788 7.11e-09 ***
ageage(50-62)                   -0.60517    0.13409  -4.513 6.38e-06 ***
sclSCL190-219:chdno             -0.22496    0.20488  -1.098 0.272189
sclSCL220-249:chdno             -0.63836    0.19249  -3.316 0.000912 ***
sclSCL>249:chdno                -1.06516    0.18460  -5.770 7.93e-09 ***
sclSCL190-219:sexfemale         -0.08909    0.08422  -1.058 0.290169
sclSCL220-249:sexfemale         -0.16943    0.08541  -1.984 0.047284 *
sclSCL>249:sexfemale             0.13846    0.08666   1.598 0.110095
sclSCL190-219:ageage(50-62)      0.44416    0.10515   4.224 2.40e-05 ***
sclSCL220-249:ageage(50-62)      0.78567    0.10294   7.632 2.31e-14 ***
sclSCL>249:ageage(50-62)         1.27776    0.09998  12.780  < 2e-16 ***
chdno:sexfemale                  1.02919    0.11463   8.978  < 2e-16 ***
chdno:ageage(50-62)             -1.08370    0.11204  -9.673  < 2e-16 ***
sexfemale:ageage(50-62)          0.08672    0.06611   1.312 0.189561
```

(i) Give the MLE for the odds ratio (OR) of having CHD among females in age group 30-49 with SCL above 249 (mg/100ml) vs below 190 (mg/100ml).

(ii) Should the estimate in (i) be the same as the one in **Q2.1**(ii)? Why?

(iii) What additional finding(s) can this regression analysis lead to, compared the one in **Q2.2**?

(i) The MLE of the OR $= \exp(-\hat{\beta}_{42}^{XY}) \doteq 2.90$

(ii) In general they are different unless all the 3 factor interactions and the 4 factor interaction of the 4 variables are zero.

(iii) In addition to the associations of SCL, sex, and age with CHD shown in the analysis of Q2.2, the current analysis indicates the 3 variables SCL, sex, and age their own pair-wise associations