Categorical Data Analysis

Exam, Form: A

Name: _____

Student Number: _____

TA: _____

Date: _____

Section 1. (20%) 是非題, 若是錯誤敍述, 請寫下簡要理由.

1. Using a survey of college students, we study the association between opinion about whether it should be legal to (1) use marijuana, (2) drink alcohol if you are 18 years old.We may get a different value for the odds ratio if we treat opinion about marijuana use as the response variable than if we treat alcohol use as the response variable.

2. Suppose that income (high, low) and gender are conditionally independent, given type of job (secretarial, construction, service, professional, etc.). Then, income and gender are also independent in the 2 marginal table (i.e., ignoring, rather than controlling, type of job).

3. With a GLM (Generalized Linear Model), Y does not need to have a normal distribution and one can model a function of the mean of Y instead of just the mean itself, but in order to get ML estimates the variance of Y must be constant at all values of predictors.

4. An ordinary regression (or ANOVA) model that treats the response Y as normally distributed is a special case of a GLM (Generalized Linear Model), with normal random component, identity link function and ML estimation.

5. Interchanging two rows or interchanging two columns in a contingency table has no effect on the value of the X^2 or G^2 chi-squared statistics. Thus, these tests treat both the rows and the columns of the contingency table as ordinal scale.

Section 2. (60%) 資料分析.

6. 一個早期的小型臨床研究 (Mendenhall 等人, 1984), 比較外科手術與放射線治療控制鼻咽癌的治療結果, 研究共納入 位受試者, 有 23 位受試者接受外科手術治療, 另外 18 位受試者接受放射線治療, 主要的反應變數爲鼻咽癌是否可以得到適當的控制. 在 5 位接受外科手術治療的受試者中, 有 4 位鼻咽癌受到控制, 而在 4 位接受放射線治療中, 有 3 位鼻咽癌受到控制, 研究結果以 2×2 二維列聯表摘要於表 1. 請使用 Fisher's exact test 檢定, 並計算 p-value.

$$H_0 : OR = 1 \quad H_A : \ OR \neq 1. \tag{1}$$

Table 1: 鼻咽癌臨床研究

	鼻咽癌		
組別	有控制	無控制	mcc總人數
手術	4	1	5
放射線	3	1	4
總和	7	2	9

7. 鐵達尼號郵輪 (Titanic) 於 912 年 4 月處女航, 從英國南安普敦 (Southampton) 出發, 計劃中的目的地爲美國紐約 (New York), 最終載著 1324 乘客和 892 名甲板工作人員駛向紐約. 1912 年 4 月 14 日, 船上時間晚 11 點 40 分, 鐵達尼號撞上冰山, 2 小時 40 分鐘後, 即 4 月 15 日凌晨 2 點 20 分沉沒, 約 1500人隨著她的消失而葬身大洋中. 資料檔案 **titanicR.xls** 呈現鐵達尼號當時的 1314 位乘客資料, 有些缺失值, 表 2 呈現部分乘客資料, 包含乘坐的艙等 (PClass), 年紀 (Age), 性別 (Sex), 以及最後存活 (Survived) 的狀況, 且 Survived $= 1$ 表示乘客最後獲救且存活. 表 3 呈現 2 個 logistic 迴歸模型, 檢視最後存活機率 $\pi = P(\text{survived} = 1)$ 與解釋變數關係, 模型分別爲

$$\mathcal{M}_1 \Rightarrow \text{logit}(\pi) = \text{age} + \text{PClass} + \text{Sex} \tag{2}$$

$$\mathcal{M}_0 \Rightarrow \text{logit}(\pi) = \text{age} + \text{Sex}. \tag{3}$$

   (a) 請說明 $\mathcal{M}_1$ 模型中, 乘坐的艙等 (PClass) 對最後存活機率的影響.

   (b) 請比較 $\mathcal{M}_1$ 模型 與 $\mathcal{M}_0$ 模型, 請寫下虛無假說, 對立假說, 與檢定.

Table 2: 鐵達尼號郵輪乘客存活: 部分資料

PClass	Age	Sex	Survived
1st	29	female	1
1st	2	female	0
1st	30	male	0
2nd	30	male	0
2nd	28	female	1
2nd	18	male	0
2nd	NA	male	0
2nd	34	male	0
3rd	51	male	0
3rd	18	male	0
3rd	45	female	1

Table 3: 鐵達尼號郵輪乘客存活: 邏輯迴歸分析

```
# MODEL 1
t.main<-glm(cbind(Survived, Not.Survived)~ Age + Sex + PClass,
            family=binomial, data=titanic.group.data,x=T)
summary(t.main)
             Estimate Std. Error z value    Pr(>|z|)
(Intercept)  3.759662   0.397567   9.457     < 2e-16 ***
Age         -0.039177   0.007616  -5.144 0.000000269 ***
Sexmale     -2.631357   0.201505 -13.058     < 2e-16 ***
PClass2nd   -1.291962   0.260076  -4.968 0.000000678 ***
PClass3rd   -2.521419   0.276657  -9.114     < 2e-16 ***
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 644.52  on 273  degrees of freedom
Residual deviance: 314.08  on 269  degrees of freedom

################################
# MODEL 2
t.main.p<-glm(cbind(Survived, Not.Survived)~ Age + Sex ,
             family=binomial, data=titanic.group.data,x=T)
summary(t.main.p)
             Estimate Std. Error z value     Pr(>|z|)
(Intercept)  1.306157   0.231052   5.653 0.0000000158 ***
Age         -0.006352   0.006187  -1.027        0.305
Sexmale     -2.465996   0.178456 -13.819      < 2e-16 ***
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 644.52  on 273  degrees of freedom
Residual deviance: 414.53  on 271  degrees of freedom
```

8. 老年人經常因行動遲緩而跌倒受傷, 造成髖關節骨股頸位移性骨折, 通常需進行半髖人工關節置換手術, 而半髖人工關節置換手術的併發症中少見的是人工關節脫臼, 然而, 潛在的人工關節感染卻可能是人工關節脫臼的一項危險因子. 由於半髖人工關節置換術後脫臼並不常見, 進行前瞻性研究有其困難, 因此在一個回溯性研究中, 檢視一家醫學中心過去 10 年所有進行半髖人工關節置換手術, 紀錄每一位病患的是否發生人工關節脫臼與人工關節感染, 研究的目的是探討人工關節感染的人是否有較高的人工關節脫臼機率. 研究資料總共有 980 觀察病患, 這些病患依照脫臼和感染這 2 個變數分類, 研究調查結果可以 2×2 二維列聯表摘要於表 4. 請選擇適並計算當的測量描述脫臼和感染這 2 個變數的關連性, 並進行檢定, 說明分析結果.

Table 4: 半髖人工關節置換手術後髖關節脫臼與感染

	關節脫臼		
關節感染	有 (對照組)	無 (控制組)	總合
有	10	19	29
沒有	11	940	951
總合	21	959	980

Section 3. (20%) 統計理論.

9. 假設一個研究資料的二元反應變數爲 $y_{ij} = 0$ 或 1, 在解釋變數 X_i 之下的觀測資料爲 (y_{ij}, x_i), $i = 1, 2, \ldots, C$, $j = 1, 2, \ldots, n_i$, $N = \sum_i n_i$, 則令隨機變數 Y_i 的觀測值 y_i 爲每個 X_i 之下的反應變數觀測總數 $y_i = \sum_{j=1}^{n_i} y_{ij}$, 令隨機變數 Z_i 的觀測值爲 $z_i = \frac{y_i}{n_i}$, $Y_i = n_i Z_i$, Y_i 爲二項分配, $Y_i \sim Bin(n_i, \pi_i(X_i))$, 則 $n_i Z_i$ 也是二項分配, $n_i Z_i \sim Bin(n_i, \pi_i(X_i))$. 隨機變數 Z_i 爲在每個 X_i 之下, 爲反應變數的樣本分率 (sample proportion).

   (a) 請將隨機變數 Z_i 的機率分配以廣義線性模型中的指數族分配 (exponential family) 寫出, 包含 θ_i, $b(\theta_i)$, $c(z_i, \phi)$, $a(\phi)$ 等的函數型式.

   (b) 計算隨機變數 Z_i 的平均值與變異數.

Answer Key for Exam A

Section 1. (20%) 是非題, 若是錯誤敍述, 請寫下簡要理由.

1. Using a survey of college students, we study the association between opinion about whether it should be legal to (1) use marijuana, (2) drink alcohol if you are 18 years old.We may get a different value for the odds ratio if we treat opinion about marijuana use as the response variable than if we treat alcohol use as the response variable. **(False)**

2. Suppose that income (high, low) and gender are conditionally independent, given type of job (secretarial, construction, service, professional, etc.). Then, income and gender are also independent in the 2 marginal table (i.e., ignoring, rather than controlling, type of job). **(False)**

3. With a GLM (Generalized Linear Model), Y does not need to have a normal distribution and one can model a function of the mean of Y instead of just the mean itself, but in order to get ML estimates the variance of Y must be constant at all values of predictors. **(False)**

4. An ordinary regression (or ANOVA) model that treats the response Y as normally distributed is a special case of a GLM (Generalized Linear Model), with normal random component, identity link function and ML estimation. **(False)**

5. Interchanging two rows or interchanging two columns in a contingency table has no effect on the value of the X^2 or G^2 chi-squared statistics. Thus, these tests treat both the rows and the columns of the contingency table as ordinal scale. **(False)**

Section 2. (60%) 資料分析.

6. 一個早期的小型臨床研究 (Mendenhall 等人, 1984), 比較外科手術與放射線治療控制鼻咽癌的治療結果, 研究共納入 位受試者, 有 23 位受試者接受外科手術治療, 另外 18 位受試者接受放射線治療, 主要的反應變數爲鼻咽癌是否可以得到適當的控制. 在 5 位接受外科手術治療的受試者中, 有 4 位鼻咽癌受到控制, 而在 4 位接受放射線治療中, 有 3 位鼻咽癌受到控制, 研究結果以 2×2 二維列聯表摘要於表 1. 請使用 Fisher's exact test 檢定, 並計算 p-value.

$$H_0 : OR = 1 \quad H_A : \ OR \neq 1. \tag{1}$$

Table 5: 鼻咽癌臨床研究

	鼻咽癌		
組別	有控制	無控制	mcc總人數
手術	4	1	5
放射線	3	1	4
總和	7	2	9

```
> (fisher.test(matrix(c(4,1,3,1),byrow=T,ncol=2)))
p-value = 1

> (fisher.test(matrix(c(4,1,3,1),byrow=T,ncol=2), alternative = "greater"))
p-value = 0.7222

> (fisher.test(matrix(c(4,1,3,1),byrow=T,ncol=2), alternative = "less"))
p-value = 0.8333
```

7. 鐵達尼號郵輪 (Titanic) 於 912 年 4 月處女航, 從英國南安普敦 (Southampton) 出發, 計劃中的目的地爲美國紐約 (New York), 最終載著 1324 乘客和 892 名甲板工作人員駛向紐約. 1912 年 4 月 14 日, 船上時間晚 11 點 40 分, 鐵達尼號撞上冰山, 2 小時 40 分鐘後, 即 4 月 15 日凌晨 2 點 20 分沉沒, 約 1500人隨著她的消失而葬身大洋中. 資料檔案 **titanicR.xls** 呈現鐵達尼號當時的 1314 位乘客資料, 有些缺失值, 表 2 呈現部分乘客資料, 包含乘坐的艙等 (PClass), 年紀 (Age), 性別 (Sex), 以及最後存活 (Survived) 的狀況, 且 Survived $= 1$ 表示乘客最後獲救且存活. 表 3 呈現 2 個 logistic 迴歸模型, 檢視最後存活機率 $\pi = P(\text{survived} = 1)$ 與解釋變數關係, 模型分別爲

$$\mathcal{M}_1 \Rightarrow \text{logit}(\pi) = \text{age} + \text{PClass} + \text{Sex} \tag{2}$$

$$\mathcal{M}_0 \Rightarrow \text{logit}(\pi) = \text{age} + \text{Sex}. \tag{3}$$

   (a) 請說明 $\mathcal{M}_1$ 模型中, 乘坐的艙等 (PClass) 對最後存活機率的影響.

   (b) 請比較 $\mathcal{M}_1$ 模型 與 $\mathcal{M}_0$ 模型, 請寫下虛無假說, 對立假說, 與檢定.

Table 6: 鐵達尼號郵輪乘客存活: 部分資料

PClass	Age	Sex	Survived
1st	29	female	1
1st	2	female	0
1st	30	male	0
2nd	30	male	0
2nd	28	female	1
2nd	18	male	0
2nd	NA	male	0
2nd	34	male	0
3rd	51	male	0
3rd	18	male	0
3rd	45	female	1

Table 7: 鐵達尼號郵輪乘客存活: 邏輯迴歸分析

```
# MODEL 1
t.main<-glm(cbind(Survived, Not.Survived)~ Age + Sex + PClass,
            family=binomial, data=titanic.group.data,x=T)
summary(t.main)
             Estimate Std. Error z value    Pr(>|z|)
(Intercept)  3.759662   0.397567   9.457     < 2e-16 ***
Age         -0.039177   0.007616  -5.144 0.000000269 ***
Sexmale     -2.631357   0.201505 -13.058     < 2e-16 ***
PClass2nd   -1.291962   0.260076  -4.968 0.000000678 ***
PClass3rd   -2.521419   0.276657  -9.114     < 2e-16 ***
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 644.52  on 273  degrees of freedom
Residual deviance: 314.08  on 269  degrees of freedom

################################
# MODEL 2
t.main.p<-glm(cbind(Survived, Not.Survived)~ Age + Sex ,
             family=binomial, data=titanic.group.data,x=T)
summary(t.main.p)
             Estimate Std. Error z value     Pr(>|z|)
(Intercept)  1.306157   0.231052   5.653 0.0000000158 ***
Age         -0.006352   0.006187  -1.027        0.305
Sexmale     -2.465996   0.178456 -13.819      < 2e-16 ***
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 644.52  on 273  degrees of freedom
Residual deviance: 414.53  on 271  degrees of freedom
```

```
> regTermTest(t.main, "PClass", method="LRT")
Working LR =  100.4448 p= < 2.22e-16

> anova(t.main.p, t.main, "Chisq")
Model 1: cbind(Survived, Not.Survived) ~ Age + Sex
Model 2: cbind(Survived, Not.Survived) ~ Age + Sex + PClass
  Resid. Df Resid. Dev Df Deviance
1       271     414.53
2       269     314.08  2   100.44
```

8. 老年人經常因行動遲緩而跌倒受傷, 造成髖關節骨股頸位移性骨折, 通常需進行半髖人工關節置換手術, 而半髖人工關節置換手術的併發症中少見的是人工關節脫臼, 然而, 潛在的人工關節感染卻可能是人工關節脫臼的一項危險因子. 由於半髖人工關節置換術後脫臼並不常見, 進行前瞻性研究有其困難, 因此在一個回溯性研究中, 檢視一家醫學中心過去 10 年所有進行半髖人工關節置換手術, 紀錄每一位病患的是否發生人工關節脫臼與人工關節感染, 研究的目的是探討人工關節感染的人是否有較高的人工關節脫臼機率. 研究資料總共有 980 觀察病患, 這些病患依照脫臼和感染這 2 個變數分類, 研究調查結果可以 2×2 二維列聯表摘要於表 4. 請選擇適並計算當的測量描述脫臼和感染這 2 個變數的關連性, 並進行檢定, 說明分析結果.

Table 8: 半髖人工關節置換手術後髖關節脫臼與感染

	關節脫臼		
關節感染	有 (對照組)	無 (控制組)	總合
有	10	19	29
沒有	11	940	951
總合	21	959	980

```
> hip.dis.tab<-matrix(c(10,19,11,940),nrow=2,byrow=T)
> chisq.test(hip.dis.tab,correct=F)
X-squared = 149.0511, df = 1, p-value < 2.2e-16

> chisq.test(hip.dis.tab)
X-squared = 133.582, df = 1, p-value < 2.2e-16

> fisher.test(hip.dis.tab)
p-value = 2.694e-11
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  14.90394 131.06932
sample estimates:
odds ratio
  44.12354
```

Section 3. (20%) 統計理論.

9. 假設一個研究資料的二元反應變數爲 $y_{ij} = 0$ 或 1, 在解釋變數 X_i 之下的觀測資料爲 (y_{ij}, x_i), $i = 1, 2, \ldots, C$, $j = 1, 2, \ldots, n_i$, $N = \sum_i n_i$, 則令隨機變數 Y_i 的觀測值 y_i 爲每個 X_i 之下的反應變數觀測總數 $y_i = \sum_{j=1}^{n_i} y_{ij}$, 令隨機變數 Z_i 的觀測值爲 $z_i = \frac{y_i}{n_i}$, $Y_i = n_i Z_i$, Y_i 爲二項分配, $Y_i \sim Bin(n_i, \pi_i(X_i))$, 則 $n_i Z_i$ 也是二項分配, $n_i Z_i \sim Bin(n_i, \pi_i(X_i))$. 隨機變數 Z_i 爲在每個 X_i 之下, 爲反應變數的樣本分率 (sample proportion).

   (a) 請將隨機變數 Z_i 的機率分配以廣義線性模型中的指數族分配 (exponential family) 寫出, 包含 θ_i, $b(\theta_i)$, $c(z_i, \phi)$, $a(\phi)$ 等的函數型式.

   (b) 計算隨機變數 Z_i 的平均值與變異數.

   考慮 Y_i 的二項式分配爲

$$n_i z_i \sim Bin(n_i, \pi_i) \tag{4}$$

   而白努力分配 (Bernoulli) 爲二項式分配的一個特例, $n_i = 1$. 若假設對 C 個獨立觀測值且反應變數 z_i 爲 指數族分配 (**exponential family**), 其概似函數與對數概似函數爲

$$f(z_i; \theta_i, \phi) = \exp\left[\frac{z_i\theta_i - b(\theta_i)}{a(\phi)} + c(z_i, \phi)\right] \tag{5}$$

$$\boldsymbol{\ell}(\underline{\boldsymbol{\beta}}) = \sum_i \log f(z_i; \theta_i, \phi) = \sum_i \ell_i(\underline{\boldsymbol{\beta}}) \tag{6}$$

   考慮廣義線性模式中的指數族分配型式, 並利用 $y_i = n_i z_i$ 爲二項分配, 則 z_i 的機率分配可寫成

$$\begin{aligned}
& f(z_i; \theta_i, \phi) \\
&= n_i \binom{n_i}{n_i z_i} \pi^{n_i z_i} (1 - \pi_i)^{n_i - n_i z_i} \\
&= \exp\left[n_i z_i \log \pi_i + (n_i - n_i z_i) \log(1 - \pi_i) + \log\left(n_i \binom{n_i}{n_i z_i}\right)\right] \\
&= \exp\left[\frac{z_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + \log(1 - \pi_i)}{1/n_i} + \log\left(n_i \binom{n_i}{n_i z_i}\right)\right] \\
&= \exp\left[\frac{z_i\theta_i - b(\theta_i)}{a(\phi)} + c(z_i, \phi)\right].
\end{aligned} \tag{7}$$

可以得到廣義線性模式中的指數族分配型式, 分別爲

$$\theta_i = \log(\frac{\pi_i}{1-\pi_i}) \tag{8}$$

$$\pi_i = \frac{e^{\theta_i}}{1+e^{\theta_i}} \tag{9}$$

$$b(\theta_i) = -\log(1-\pi_i) = \log[1+e^{\theta_i}] \tag{10}$$

$$c(z_i, \phi) = \log\left(n_i \binom{n_i}{n_i z_i}\right) \tag{11}$$

$$a(\phi)/w_i = = 1/n_i \tag{12}$$

$$\phi = 1 \tag{13}$$

$$w_i = n_i. \tag{14}$$

因此 Z_i 的平均值與變異數爲

$$\mathcal{E}(Z_i) = \frac{\partial b(\theta_i)}{\partial \theta_i} = \frac{e^{\theta_i}}{1+e^{\theta_i}} = \pi_i \tag{15}$$

$$\mathbf{Var}(Z_i) = b''(\theta_i)a(\phi) = \frac{e^{\theta_i}(1+e^{\theta_i}) - e^{\theta_i}(e^{\theta_i})}{(1+e^{\theta_i})^2} a(\phi) \tag{16}$$

$$= \frac{e^{\theta_i}}{(1+e^{\theta_i})^2} a(\phi) = \frac{\pi_i(1-\pi_i)}{n_i} \tag{17}$$

再一次假設對於所有觀測值的 $a(\phi)$ 都相同, 且現在令 $\pi_i = \pi(\underline{\mathbf{x}}_i), i = 1, 2, \ldots, C$ 滿足

$$\pi_i = \pi(\underline{\mathbf{x}}_i^T \underline{\boldsymbol{\beta}}) = \text{expit}(\underline{\mathbf{x}}_i^T \underline{\boldsymbol{\beta}}) = \frac{e^{\underline{\mathbf{x}}_i^T \underline{\boldsymbol{\beta}}}}{1+e^{\underline{\mathbf{x}}_i^T \underline{\boldsymbol{\beta}}}} \tag{18}$$

$$\eta_i = g(\mu_i) = \theta_i = \text{logit}(\pi_i) = \log \frac{\pi_i}{(1-\pi_i)} = \underline{\mathbf{x}}_i^T \underline{\boldsymbol{\beta}} \tag{19}$$

這便是廣義線性模型的型式.