Chapter 8: Study Sheet & Homework

Sheng-Mao

2025/04/08

Models for Match Pairs

- 一個人回答二個問題
- 想知道這二個回答的比例是否一致
- 例如：問一個人二道題
  - 1）是否願意為環保降低生活水平(CLS)
  - 2）是否願意為環保增加稅金(PHT)
  - 資料（1144人回答結果）

	CLS-Yes	**CLS-No**	**Total**
PHT-Yes	227	132	359
PHT-No	107	678	785
Total	334	810	1144

  - 母體（一個人回答的機率）

	CLS-Yes	**CLS-No**	**Total**
PHT-Yes	π_{11}	π_{12}	$\pi_{1\cdot}$
PHT-No	π_{21}	π_{22}	$\pi_{2\cdot}$
Total	$\pi_{\cdot 1}$	$\pi_{\cdot 2}$	1

  - 母體（一群人回答的平均次數）

	CLS-Yes	**CLS-No**
PHT-Yes	μ_{11}	μ_{12}
PHT-No	μ_{21}	π_{22}

  - 問題：回答二題的比例是否一致？
  - 數學式：H_0：$\pi_{1\cdot} = \pi_{\cdot 1}$是否成立？ Or H_0：$\pi_{12} = \pi_{21}$是否成立？
  - 若$\pi_{1\cdot} = \pi_{\cdot 1}$則$\pi_{2\cdot} = \pi_{\cdot 2}$，也就是說二個分配相等。

McNemar Test for Two Proportions

- 若H_0：$\pi_{12} = \pi_{21}$成立，則回答(CLS=NO, PHT=YES)與回答(CLS=YES, PHT=NO)的人要差不多
- 在H_0：$\pi_{12} = \pi_{21}$成立之下

$$N_{12} \sim Bin(N_{12} + N_{21}, 1/2)$$

- R Code:

```
mcnemar.test(A)
```

Marginal Models (Population-Average Model)

- 原始的問題其實是 $\pi_{1\cdot} = \pi_{\cdot 1}$。
- 我們可以對這二個機率建模。例如:針對第i個人的回答
  - linear model

$$\pi_{1\cdot} = \Pr(Y_{i1} = 1) = \alpha_0$$
$$\pi_{\cdot 1} = \Pr(Y_{i2} = 1) = \alpha_0 + \alpha_1$$

    或者

$$\Pr(Y_{ij} = 1) = \alpha_0 + \alpha_1 x_{ij} \text{ where } x_{ij} = I(j = 1)$$

  - logistic model

$$\text{logit}(\Pr(Y_{ij} = 1)) = \beta_0 + \beta_1 x_{ij}$$

- 針對此二模型，檢定 $H_0 : \alpha_1 = 0$ 或 $H_0 : \beta_1 = 0$ 等同於檢定 $H_0 : \pi_{1\cdot} = \pi_{\cdot 1}$。
- 資料：因為模型是regression/logistic regression的樣子，所以資料也要整理成迴歸所需的樣子

```
## 'data.frame':    2288 obs. of  3 variables:
##  $ person  : int  1 1 2 2 3 3 4 4 5 5 ...
##  $ question: int  1 0 1 0 1 0 1 0 1 0 ...
##  $ y       : int  1 1 1 1 1 1 1 1 1 1 ...
```

- 因為一個人回答二題 (Y_{i1}, Y_{i2})，所以此視此二變數獨立並不恰當。我們假設此二變數存在某種 correlation。因此使用GEE方法分析。

```
  library(gee)
  fit_identity <- gee(y~question, id=person, family=binomial(link="identity"), data=0
pinions_reg)
  fit_logit <- gee(y~question, id=person, family=binomial(link="logit"), data=Opinion
s_reg)
```

Conditional Models (Subject-Specific Model)

- 前述 marginal model

$$\text{logit}\,\Pr(Y_{ij} = 1) = \alpha_0 + \alpha_1 x_{ij} \quad (1)$$

- 假設所有人有相同的截距和斜率。
- 觀查：
  - 環保主義者二題都答YES的比例比較高。不信地球暖化者則都回答NO的比例較高。所以最少α_0應因人而異。
  - 所以（1） 錯了嗎？也沒有，只是我們視α_0為眾人的平均。
  - 我們可以把「α_0應因人而異」寫入模型嗎？可以的。

$$\text{logit}\,\Pr(Y_{ij} = 1) = \alpha_i + \beta_1 x_{ij} \quad (2)$$

  - 模型（2）就是 conditional model。有點複雜，之後會再回來。

Marginal Homogeneity for Baselined-Category Logit Model

- 前述問題的答案都是是非題 Yes or No，很自然的，我們可以回答選擇題 multiple choice。
- 一樣的情境，每個人回答二題 (Y_{i1}, Y_{i2})，二題的答案都是 $1, \ldots, c$ 其中之一。

- 我們想知道 Y_{i1} 與 Y_{i2} 的機率到底一不一樣？
- Marginal: 大家的截距都一樣
- Marginal Homogeneity:

$$\Pr(Y_{i1} = k) = \Pr(Y_{i2} = k), \quad k = 1, \dots, c$$

- Baselined-Category Logit Model

$$\log \frac{\Pr(Y_{ij} = k)}{\Pr(Y_{ij} = c)} = \alpha_k + \beta_k x_{ij}$$

- Test for marginal homogeneity:

$$H_0 : \beta_1 = \cdots = \beta_{c-1} = 0$$

- Again, Y_{i1} 與 Y_{i2} 是相關的，要用GEE的方法解。

```
library(multgee)
# full model (H_a 成立之下)
fit <- nomLORgee(y ~ purchase, id=person, LORstr="independence", data=Coffee)
# reduced model (H_0 成立之下)
fit_0 <- nomLORgee(y ~ 1, id=person, LORstr="independence", data=Coffee)
```

Symmetry and Quasi-Symmetry Models for Square Contingeny Tables

- Baselined-Category Logit Model 很好，但是需要 $2 \times (c - 1)$ 這麼多的參數，能否減少些？我們可以考慮 Symmetry model

$$\Pr(Y_{i1} = r, Y_{i2} = s) = \pi_{rs} = \pi_{sr} = \Pr(Y_{i1} = s, Y_{i2} = r)$$

- Symmetry model implies marginal homogeneity but verse is not true.
- Moreover

$$\log \frac{\pi_{rs}}{\pi_{sr}} = 0 \quad \text{for all} r \neq s$$

- Symmetry model is too restrictive. Let's try another one.
- Quasi-Symmetry Model

$$\log \frac{\pi_{rs}}{\pi_{sr}} = \beta_r - \beta_s$$

  - This model says

$$\frac{\pi_{12}}{\pi_{21}} = \frac{e^{\beta_1}/C_{12}}{e^{\beta_2}/C_{12}}$$

  - which implies

$$\pi_{12} = e^{\beta_1 + \Delta_{12}} \quad \text{and} \quad \pi_{21} = e^{\beta_2 + \Delta_{12}}$$

  - and thus

	$Y_1 = 1$	$Y_1 = 2$	$Y_1 = 3$
$Y_2 = 1$		$e^{\beta_1 + \Delta_{12}}$	$e^{\beta_1 + \Delta_{13}}$

	$Y_1 = 1$	$Y_1 = 2$	$Y_1 = 3$
$Y_2 = 2$	$e^{\beta_2+\Delta_{12}}$		$e^{\beta_2+\Delta_{23}}$
$Y_2 = 3$	$e^{\beta_3+\Delta_{13}}$	$e^{\beta_3+\Delta_{23}}$	

    - When $\beta_r = \beta_s$ holds for all r, s, ($\beta_1 = \beta_2 = \beta_3 = \beta$), the marginal homogeneity holds.

	$Y_1 = 1$	$Y_1 = 2$	$Y_1 = 3$
$Y_2 = 1$		$e^{\beta+\Delta_{12}}$	$e^{\beta+\Delta_{13}}$
$Y_2 = 2$	$e^{\beta+\Delta_{12}}$		$e^{\beta+\Delta_{23}}$
$Y_2 = 3$	$e^{\beta+\Delta_{13}}$	$e^{\beta+\Delta_{23}}$	

- The coffee example

```
##
##            Hight Pt Tasters Sanka Nescafe Brim
##   Hight Pt       93      17    44       7   10
##   Tasters         9      46    11       0    9
##   Sanka          17      11   155       9   12
##   Nescafe         6       4     9      15    2
##   Brim           10       4    12       2   27
```

- Data preparation for quasi-symmetry model

```
##    H  T  S  N  B nij nji
## 1  1 -1  0  0  0  17   9
## 2  1  0 -1  0  0  44  17
## 3  1  0  0 -1  0   7   6
## 4  1  0  0  0 -1  10  10
## 5  0  1 -1  0  0  11  11
## 6  0  1  0 -1  0   0   4
## 7  0  1  0  0 -1   9   4
## 8  0  0  1 -1  0   9   9
## 9  0  0  1  0 -1  12  12
## 10 0  0  0  1 -1   2   2
```

- Data Analysis

```
# quasi-symmetric model
QS <- glm(nij/(nij+nji) ~ -1 + H + T + S + N + B, family=binomial, weight=nij+nji, da
ta=Coffee2)
summary(QS)
```

Analyzing Rater Agreement

- 檢定二個醫師的診斷是否相同
- 相同 –> 診斷高相關 –> 診斷不獨立
- Ratings by two pathologists
    - Separately classified 118 slides on the presence and extent of carcinoma of the uterine cervix.

- The rating scale has the ordered categories: negative (1), atypical squamous hyperplasia (2), carcinoma in situ (3) and squamous or invasive carcinoma (4).
- The main diagonal represents observer agreement. Perfect agreement occurs when $\Sigma_i \pi_{ii} = 1$.

```
##        1  2  3  4 Total
## 1     22  2  2  0    26
## 2      5  7 14  0    26
## 3      0  2 36  0    38
## 4      0  1 17 10    28
## Total 27 12 69 10   118
```

- Use Poisson regression
- 定義第一位診斷 i、第二位診斷 j 的人數有 W_{ij}，則令 $W_{ij} \sim Poisson(\mu_{ij})$ 其中

$$\log \mu_{ij} = \lambda + \lambda_i^1 + \lambda_j^2 + \delta_i I(i = j)$$

- 當 $\delta_i = 0$ for all i 時，

$$\mu_{ij} = e^{\lambda+\lambda_i^1+\lambda_j^2} = e^{\lambda/2+\lambda_i^1} \times e^{\lambda/2+\lambda_j^2}$$

- 只有一位受檢者時，可視為

$$\Pr(\text{第一位診斷為} i, \text{第二位診斷為} j) = \Pr(\text{第一位診斷為} i) \times \Pr(\text{第二位診斷為} j)$$

Bradley-Terry Model

- 對於「配對」的比賽，且比每場比賽一定有勝敗（不能平手）的一系列比賽(tournament)。如何判預測選手或隊伍的實力？
- Let Π_{ij} denote the probability that player i is the victor when i and j play. Thus, $\Pi_{ji} = 1 - \Pi_{ij}$.
- The Bradley-Terry model (quasi-symmetry)

$$\text{logit}(\Pi_{ij}) = \log\left(\frac{\Pi_{ij}}{\Pi_{ji}}\right) = \beta_i - \beta_j$$

- The probability that player i wins equal 0.5 when $\beta_i = \beta_j$ and exceeds 0.5 when $\beta_i > \beta_j$.
- No intercept. One parameter of β_j is redundant.

```
##                Tai Tzu-Ying Chen Yufei An Seyoung Yamaguchi Akane
## Tai Tzu-Ying             NA         19          3              13
## Chen Yufei                8         NA         11              10
## An Seyoung               12          8         NA               9
## Yamaguchi Akane          11         20         12              NA
```

```
Tai <- c(1,1,1,0,0,0)
Chen <- c(-1,0,0,1,1,0)
An <- c(0,-1,0,-1,0,1)
Ya <- c(0,0,-1,0,-1,-1)
nij <- c(19,3,13,11,10,9)
nji <- c(8,12,11,8,20,12)
BWF <- data.frame(Tai, Chen, An, Ya, nij, nji)
BWF
```

```
##   Tai Chen An Ya nij nji
## 1   1   -1  0  0  19   8
## 2   1    0 -1  0   3  12
## 3   1    0  0 -1  13  11
## 4   0    1 -1  0  11   8
## 5   0    1  0 -1  10  20
## 6   0    0  1 -1   9  12
```

```
fit <- glm(nij/(nij+nji) ~ -1 + Tai + Chen + An + Ya,
           family=binomial, weights=nij+nji, data=BWF)
summary(fit)
```

```
##
## Call:
## glm(formula = nij/(nij + nji) ~ -1 + Tai + Chen + An + Ya, family = binomial,
##     data = BWF, weights = nij + nji)
##
## Coefficients: (1 not defined because of singularities)
##      Estimate Std. Error z value Pr(>|z|)
## Tai   -0.1320     0.2947  -0.448   0.6543
## Chen  -0.5597     0.2831  -1.977   0.0481 *
## An    -0.1179     0.3120  -0.378   0.7056
## Ya         NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 14.867  on 6  degrees of freedom
## Residual deviance: 10.317  on 3  degrees of freedom
## AIC: 36.974
##
## Number of Fisher Scoring iterations: 4
```

- 山口茜 > 安洗瑩 > 戴資穎 > 陳雨菲 （or 只有陳比較弱）

Homework # 5

找五位同學玩猜拳比賽。

1. 每次比賽採一戰一勝制，猜到第一次有人勝出者為勝，另一位則為敗。
2. 兩兩對戰最少五次，可以多戰幾次。
3. 分析此五人中，是否有人特別厲害。