Midterm Examine

711378912 蔡宜諠

2025-04-10

Q1: 台灣人口組成問題。

研究表明，台灣人口組成比例為：閩南人（71%）、客家人（12%）、外省人（12%）、原住民族（2.6%）及新住民（2.4%）。若某團體內部調查結果如下：

學歷	閩南人	客家人	外省人	原住民族	新住民
大學及以下	40	25	2	3	1
大學以上	84	20	10	3	3

```
# 建立交叉表
edu_pop <- matrix(c(40,25,2,3,1,
                    84,20,10,3,3),
                  nrow = 2, byrow = TRUE)
colnames(edu_pop) <- c("閩南人", "客家人", "外省人", "原住民族", "新住民")
rownames(edu_pop) <- c("大學及以下", "大學以上")
edu_pop <- as.table(edu_pop)
edu_pop
```

```
##             閩南人 客家人 外省人 原住民族 新住民
## 大學及以下     40     25      2        3      1
## 大學以上       84     20     10        3      3
```

統計檢定

- H_0: 閩南人比例為 71%
- H_a: 閩南人比例不為 71%

```
n <- sum(edu_pop)                 # (40+25+2+3+1) + (84+20+10+3+3) 全部人口數
x <- sum(edu_pop[, "閩南人"])      # 40 + 84 閩南語人口數
cat(n, x)
```

```
## 191 124
```

```
pihat <- x / n
pi0 <- 0.71
alpha <- 0.05
wald_stat <- (pihat - pi0) / sqrt(pihat * (1 - pihat) / n)
score_stat <- (pihat - pi0) / sqrt(pi0 * (1 - pi0) / n)
cat("Wald test statistic =", wald_stat)
```

```
## Wald test statistic = -1.760357
```

```
cat("Score test statistic =", score_stat)
```

```
## Score test statistic = -1.851345
```

雙尾檢定的臨界值：$z_{0.975} = \pm 1.96$

- Wald 檢定統計量為 -1.760357，未落在拒絕域（即 |t| < 1.96），因此**無法拒絕**虛無假設 H_0。
- Score 檢定統計量為 -1.851345，也未達到拒絕標準，因此同樣**無法拒絕** H_0。
- 統計上**沒有足夠證據證明**認為該團體中閩南人比例與台灣整體比例（71%）有顯著差異。

請利用Odds Ratio的概念來檢視是否外省人比閩南人更重視學歷。

比較「外省人」與「閩南人」在學歷上的傾向是否有顯著差異。
定義「重視學歷」為：是否擁有 **大學以上學歷**。

- H_0：$\theta = 1$（外省人與閩南人對於學歷的勝算比相同）
- H_a：$\theta \neq 1$（兩族群對於學歷的勝算比不同）

```r
# 取出所需資料
# 若有 0 要補 0.5，但這題不用
B2 <- edu_pop[c("大學以上", "大學及以下"), c("閩南人", "外省人")]

theta_hat <- B2[1,1] * B2[2,2] / (B2[1,2] * B2[2,1])
log_theta_hat <- log(theta_hat)
var_log_theta_hat <- sum(1 / B2)
ub <- log_theta_hat + qnorm(.975) * sqrt(var_log_theta_hat)
lb <- log_theta_hat + qnorm(.025) * sqrt(var_log_theta_hat)
elb <- exp(lb); eub <- exp(ub)

cat("Odds Ratio =", theta_hat, "\n")
```

```
## Odds Ratio = 0.42
```

根據 Odds Ratio 的計算，外省人相對於閩南人取得大學以上學歷的勝算比為 0.42，
表示外省人取得高學歷的傾向**低於**閩南人。由於勝算比小於 1，**沒有證據證明外省人比閩南人更重視學歷**。

請檢定是否學歷與人口組成獨立。

- H_0: 學歷與人口組成**獨立**
- H_a: 學歷與人口組成**不獨立**

```r
chisq_result <- chisq.test(edu_pop)
```

```
## Warning in chisq.test(edu_pop): Chi-squared approximation may be incorrect
```

```r
chisq_result
```

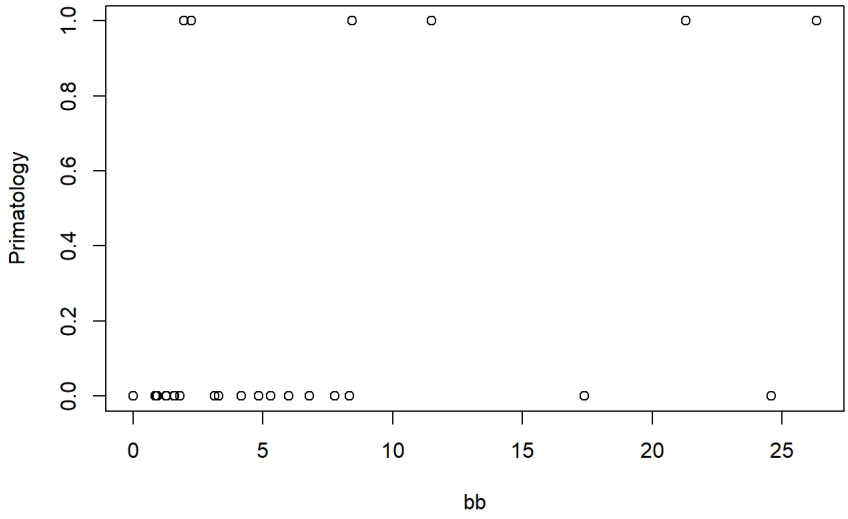
```
##
##  Pearson's Chi-squared test
##
## data:  edu_pop
## X-squared = 10.631, df = 4, p-value = 0.03104
```

卡方檢定統計量為 10.631，p 值為 0.031。
由於 p 值小於 0.05，拒絕虛無假設，表示學歷與人口組成之間**具有統計上的顯著關聯性(不獨立)**。

Q2: 靈長類的腦容量是否相對比較大？

為回答此問題，R library(MASS) 裡的 Animals 資料集收集了28種哺乳動物之腦重量及身體重。 以下程式可調動此資料集。bb 是腦身比，Primatology 是是否為靈長類的指標（1:靈長類、0:非靈長類）。 請用 logistic regression 分析此資料，並回達「靈長類的腦容量是否相對比較大」之問題。

```r
library(MASS)
data("Animals")
bb <- Animals$brain/Animals$body ## 腦身比
Primatology <- c(0,0,0,0,0,0,0,1,0,1,0,0,1,1,0,0,1,0,0,0,0,0,0,1,0,0,0,0) ## 靈長類指標
plot(bb, Primatology)
```




```
# 邏輯斯回歸模型
model <- glm(Primatology ~ bb, family = binomial)
summary(model)
```

```
##
## Call:
## glm(formula = Primatology ~ bb, family = binomial)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.19956    0.72535  -3.032  0.00243 **
## bb           0.11768    0.06099   1.929  0.05367 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 29.096  on 27  degrees of freedom
## Residual deviance: 24.998  on 26  degrees of freedom
## AIC: 28.998
##
## Number of Fisher Scoring iterations: 4
```

分析結果顯示，腦身比對是否為靈長類的影響接近顯著，但 p 值為 0.0537，略高於 0.05。
因此，**在統計上沒有足夠證據證明靈長類的腦容量相對較大**。
不過從圖形與係數方向來看，腦身比較高的動物仍較有可能是靈長類，未來可透過更多樣本進一步驗證。

$$\text{logit}(\pi) = -2.200 + 0.1177 \cdot \text{腦身比}，\text{其中}(\pi)\text{為預測為靈長類的機率}$$

Q3: 異位性皮膚炎之預測

請以 lights.csv 之資料建構一logistic regression 用以預測得異位性皮膚炎之機率，需要執行變數選 擇。

- **反應變數：**
  - `Atopy`：是否患有異位性皮膚炎（1 = 有，0 = 無）
- **主要自變數：**
  - `Sex`：性別（1 = 男，0 = 女）
  - `Prematurity`：是否早產（1 = 是，0 = 否）
  - `Smoking`：懷孕期間母親是否抽菸（1 = 是，0 = 否）
  - `Total_IgE`：超敏指數
  - `Vita_C`：維他命C攝取量
- **過敏源自變數（皮膚測試結果，1 = 過敏反應，0 = 無）：**
  - `Milk`, `Dander`, `Mite`, `Grass`, `Fish`, `Crustacean`, `Egg`, `Vege`

```
allergy <- read.csv("D:/NTPU_class/categorical_analysis/code/lights.csv", header=TRUE) # 請用您自己的路徑
str(allergy)
```

```
## 'data.frame':    1447 obs. of  14 variables:
##  $ Sex        : int  1 0 0 0 0 0 0 1 1 0 ...
##  $ Prematurity: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Smoking    : int  1 0 0 0 1 1 0 0 0 0 ...
##  $ Atopy      : int  0 1 1 1 1 1 1 0 1 1 ...
##  $ Total_IgE  : num  34.4 8.62 184 1986 479 ...
##  $ Vita_C     : num  12.8 26.7 32.2 18.3 22.2 ...
##  $ Milk       : int  0 1 1 1 0 0 1 0 0 1 ...
##  $ Dander     : int  0 0 0 1 0 0 0 0 0 1 ...
##  $ Mite       : int  1 0 1 1 1 0 1 0 1 1 ...
##  $ Grass      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Fish       : int  0 1 0 1 0 0 0 0 0 0 ...
##  $ Crustacean : int  0 0 0 1 0 0 0 0 0 0 ...
##  $ Egg        : int  0 1 0 1 0 0 0 0 0 0 ...
##  $ Vege       : int  0 1 0 0 0 0 0 0 0 0 ...
```

```
# 建立 full model
full_model <- glm(Atopy ~ Sex + Prematurity + Smoking + Total_IgE + Vita_C +
                    Milk + Dander + Mite + Grass + Fish + Crustacean + Egg + Vege,
                  data = allergy, family = binomial)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
# 執行 Stepwise AIC 進行變數選擇
# 不想看到一大堆ㄋ warning 訊息
final_model <- suppressWarnings(step(full_model))
```

```
## Start:  AIC=805.73
## Atopy ~ Sex + Prematurity + Smoking + Total_IgE + Vita_C + Milk +
##     Dander + Mite + Grass + Fish + Crustacean + Egg + Vege
##
##               Df Deviance     AIC
## - Vita_C       1   777.79  803.79
## - Fish         1   777.80  803.80
## - Smoking      1   778.18  804.18
## - Sex          1   778.29  804.29
## - Prematurity  1   778.64  804.64
## - Dander       1   778.70  804.70
## - Egg          1   779.00  805.00
## <none>             777.73  805.73
## - Vege         1   780.98  806.98
## - Crustacean   1   782.19  808.19
## - Milk         1   783.65  809.65
## - Grass        1   798.24  824.24
## - Total_IgE    1   874.87  900.87
## - Mite         1   982.37 1008.37
##
## Step:  AIC=803.79
## Atopy ~ Sex + Prematurity + Smoking + Total_IgE + Milk + Dander +
##     Mite + Grass + Fish + Crustacean + Egg + Vege
##
##               Df Deviance     AIC
## - Fish         1   777.86  801.86
## - Smoking      1   778.22  802.22
## - Sex          1   778.38  802.38
## - Prematurity  1   778.73  802.73
## - Dander       1   778.76  802.76
## - Egg          1   779.06  803.06
## <none>             777.79  803.79
## - Vege         1   781.10  805.10
## - Crustacean   1   782.25  806.25
## - Milk         1   783.65  807.65
## - Grass        1   798.48  822.48
## - Total_IgE    1   874.92  898.92
## - Mite         1   982.37 1006.37
##
## Step:  AIC=801.86
## Atopy ~ Sex + Prematurity + Smoking + Total_IgE + Milk + Dander +
##     Mite + Grass + Crustacean + Egg + Vege
##
##               Df Deviance     AIC
## - Smoking      1   778.28  800.28
## - Sex          1   778.46  800.46
## - Prematurity  1   778.78  800.78
## - Dander       1   778.86  800.86
## - Egg          1   779.18  801.18
## <none>             777.86  801.86
## - Vege         1   781.29  803.29
## - Crustacean   1   782.33  804.33
## - Milk         1   783.90  805.90
## - Grass        1   799.47  821.47
## - Total_IgE    1   875.70  897.70
## - Mite         1   982.46 1004.46
##
## Step:  AIC=800.28
## Atopy ~ Sex + Prematurity + Total_IgE + Milk + Dander + Mite +
##     Grass + Crustacean + Egg + Vege
##
##               Df Deviance     AIC
## - Sex          1   778.88  798.88
## - Prematurity  1   779.18  799.18
## - Dander       1   779.25  799.25
## - Egg          1   779.65  799.65
## <none>             778.28  800.28
## - Vege         1   781.78  801.78
## - Crustacean   1   782.83  802.83
## - Milk         1   784.24  804.24
## - Grass        1   800.07  820.07
## - Total_IgE    1   876.10  896.10
## - Mite         1   982.46 1002.46
##
## Step:  AIC=798.88
## Atopy ~ Prematurity + Total_IgE + Milk + Dander + Mite + Grass +
##     Crustacean + Egg + Vege
##
##               Df Deviance     AIC
## - Prematurity  1   779.82  797.82
## - Dander       1   779.82  797.82
## - Egg          1   780.24  798.24
## <none>             778.88  798.88
## - Vege         1   782.34  800.34
```

```
## - Crustacean   1   783.29  801.29
## - Milk         1   784.60  802.60
## - Grass        1   800.28  818.28
## - Total_IgE    1   876.15  894.15
## - Mite         1   982.46 1000.46
##
## Step:  AIC=797.82
## Atopy ~ Total_IgE + Milk + Dander + Mite + Grass + Crustacean +
##     Egg + Vege
##
##              Df Deviance     AIC
## - Dander      1   780.83  796.83
## - Egg         1   781.09  797.09
## <none>            779.82  797.82
## - Vege        1   783.24  799.24
## - Crustacean  1   784.32  800.32
## - Milk        1   785.80  801.80
## - Grass       1   800.99  816.99
## - Total_IgE   1   876.82  892.82
## - Mite        1   985.02 1001.02
##
## Step:  AIC=796.83
## Atopy ~ Total_IgE + Milk + Mite + Grass + Crustacean + Egg +
##     Vege
##
##              Df Deviance    AIC
## - Egg         1   782.31 796.31
## <none>            780.83 796.83
## - Vege        1   784.16 798.16
## - Crustacean  1   785.30 799.30
## - Milk        1   785.83 799.83
## - Grass       1   801.78 815.78
## - Total_IgE   1   876.95 890.95
## - Mite        1   985.34 999.34
##
## Step:  AIC=796.31
## Atopy ~ Total_IgE + Milk + Mite + Grass + Crustacean + Vege
##
##              Df Deviance    AIC
## <none>            782.31 796.31
## - Crustacean  1   786.79 798.79
## - Milk        1   788.18 800.18
## - Vege        1   790.16 802.16
## - Grass       1   803.71 815.71
## - Total_IgE   1   878.61 890.61
## - Mite        1   986.28 998.28
```

```
# 查看最終模型結果
summary(final_model)
```

```
##
## Call:
## glm(formula = Atopy ~ Total_IgE + Milk + Mite + Grass + Crustacean +
##     Vege, family = binomial, data = allergy)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.816653   0.131847 -13.778  < 2e-16 ***
## Total_IgE     0.011694   0.001485   7.875  3.4e-15 ***
## Milk          0.785430   0.318124   2.469  0.01355 *
## Mite          3.652440   0.360220  10.139  < 2e-16 ***
## Grass         3.264065   1.051548   3.104  0.00191 **
## Crustacean   15.456366 579.560201   0.027  0.97872
## Vege          0.923415   0.322982   2.859  0.00425 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1885.26  on 1446  degrees of freedom
## Residual deviance:  782.31  on 1440  degrees of freedom
## AIC: 796.31
##
## Number of Fisher Scoring iterations: 17
```

挑選變數

雖然模型過程中出現警告訊息 glm.fit: fitted probabilities numerically 0 or 1 occurred，但不影響模型估計與變數解釋，先不管他。

本研究透過邏輯斯迴歸模型預測異位性皮膚炎（Atopy），並以 Stepwise AIC 方法選出最佳變數組合。 最終模型中選入以下顯著變數：

- **Total_IgE**（超敏指數）：係數為 0.0117，p < 0.001
- **Milk**（牛奶過敏）：係數為 0.7854，p = 0.0136
- **Mite**（蟎蟲過敏）：係數為 3.6524，p < 0.001
- **Grass**（草類過敏）：係數為 3.2641，p = 0.0019
- **Vege**（蔬菜過敏）：係數為 0.9234，p = 0.0043

模型總結

根據 stepwise AIC 篩選，最終邏輯斯迴歸模型如下：

$$\text{logit}(\pi) = -3.672 + 0.0117 \cdot \text{Total\_IgE} + 0.7854 \cdot \text{Milk} + 3.6524 \cdot \text{Mite} + 3.2641 \cdot \text{Grass} - 0.0244 \cdot \text{Crustacean} + 0.9234 \cdot \text{Vege}$$，其中(π)表

此模型可用於根據個體的超敏指數與過敏原反應，來預測其罹病機率。
模型中顯示，Total_IgE 越高，以及對牛奶、蟎蟲、草類與蔬菜過敏者，罹病風險皆顯著上升。

雖然 Crustacean（甲殼類）變數未達顯著水準（p = 0.9787），但仍被保留於最終模型中，代表其在 stepwise 選模過程中未被剔除。