投票資料分析報告

第八組:吳宜軒、鄭敬恒、邱勃誠、李振偉、康成駿

2025-04-17

目錄

1. 資料集介紹與前處理

- 本筆資料為調查台南市中西區或北區選民欲投票的候選人，並延伸調查各候選人的支持率與選民的熱衷程度，最後針對特定候選人給予其競選的建議。

資料變數轉換

- 將v4_1~8的變數做轉換，定義成一個新變數:政治熱衷程度(political_interest)，其值的計算方式為一個被訪人知道自己的候選區有幾位候選人

```
FullData = read_sav("C:/Users/BoCheng/Desktop/Quarto Workspace/poll.sav")
FullData_ori = FullData

k = rowSums(FullData[,4:11] < 11, na.rm = TRUE)
political_interest = k
FullData = cbind(FullData,political_interest)
```

定義變數型態與描述性統計

Variable	Data Type	Definition	Note
v1	factor	戶籍在台南市哪一區	01" 北區" , 02" 中西區" , 98 " 拒答/戶籍不在臺南市以上幾區/家中無有投票權的人" , 99 " 遺漏值或跳答"
v2	factor	戶籍在北區哪個里?	01" 北華里" , 02" 元寶里" , 03" 中樓里" , ..., 33" 文元里" , 44" 不知道" , 98" 拒答" , 99" 遺漏值或跳答"
v3	factor	戶籍在中西區哪個里?	01" 郡王里" , 02" 赤嵌里" , 03" 法華里" , ..., 20" 南門里" , 44" 不知道" , 98" 拒答" , 99" 遺漏值或跳答"
v5	factor	若明天要投票，會將票投給誰?	01" (1號)" , 02" (2號)" , 03" (3號)" , ..., 10" (10號)" , 98" 無反應" , 99" 尚未決定" , 99" 遺漏值或跳答"
v6	factor	年齡區間	01" 20-29歲" , 02" 30-39歲" , 03" 40-49歲" , 04" 50-59歲" , 05" 60歲及以上" , 06" 不知道或拒答" ,99" 遺漏值或跳答"
v7	factor	目前最高的學歷	01" 小學或以下" , 02" 初中、國中" , 03" 高中、高職" , 04" 專科" , 05" 大學以上" , 95" 拒答" , 99" 遺漏值或跳答"
v8	factor	受訪者性別	01" 男性" , 02" 女性" , 99" 遺漏值或跳答"
political	factor	政治熱衷程度	計算每位選民認識候選人的個數, 並factor化，將其定義為政治的熱衷程度(0-8)

```
FullData = FullData[,-c(4:11)]
FullData = FullData %>% mutate(across(c(v1, v2, v3, v5, v6, v7, v8), as.factor))
FullData$political_interest = as.factor(FullData$political_interest)
```

```
latex(describe(FullData), file = '')
```

FullData
8 Variables 1671 Observations

---

v1

n	missing	distinct
1671	0	2

Value	1	2
Frequency	1107	564
Proportion	0.662	0.338

---

v2

n	missing	distinct
1671	0	36

lowest : 1 2 3 4 5 , highest: 32 33 44 98 99

---

v3

n	missing	distinct
1671	0	23

lowest : 1 2 3 4 5 , highest: 19 20 44 98 99

---

v5

n	missing	distinct
1671	0	13

Value	1	2	3	4	5	6	7	8	9	10	91	98	99
Frequency	158	9	205	79	33	98	195	6	8	53	10	269	548
Proportion	0.095	0.005	0.123	0.047	0.020	0.059	0.117	0.004	0.005	0.032	0.006	0.161	0.328

---

v6

n	missing	distinct
1671	0	6

Value	1	2	3	4	5	6
Frequency	52	94	201	336	946	42
Proportion	0.031	0.056	0.120	0.201	0.566	0.025

---

v7

n	missing	distinct
1671	0	6

Value	1	2	3	4	5	95
Frequency	292	165	431	198	520	65
Proportion	0.175	0.099	0.258	0.118	0.311	0.039

---

v8

n	missing	distinct
1671	0	2

Value	1	2
Frequency	682	989
Proportion	0.408	0.592

---

political_interest

n	missing	distinct
1671	0	9

Value	0	1	2	3	4	5	6	7	8
Frequency	953	237	205	124	75	29	28	15	5
Proportion	0.570	0.142	0.123	0.074	0.045	0.017	0.017	0.009	0.003

查看缺失值與重複值可能

```
# 1.       row-wise
total_na = sum(rowSums(is.na(FullData_ori)))

# 2.     row
dup_count = sum(duplicated(FullData_ori))
total_na;dup_count
```

```
[1] 0
```

```
[1] 153
```

```
#
# # 3.
# dup_data = FullData_ori %>% filter(duplicated(.))
#
# # 4.       &
# dup_summary = dup_data %>%
#   group_by(across(1:15)) %>%           #   15
#   summarise(times = n(), .groups = "drop")
#
# # 5.
# dup_summary = dup_summary %>%
#   rowwise() %>%
#   mutate(location = {
#     match_row = apply(FullData_ori[, 1:15], 1, function(x) all(x == c_across(1:15), na.rm = TRUE))
#     which(match_row)[1]
#   }) %>%
#   ungroup()
# # 6.     CSV
# write_excel_csv(dup_summary, "RE.csv")
```

小結論：

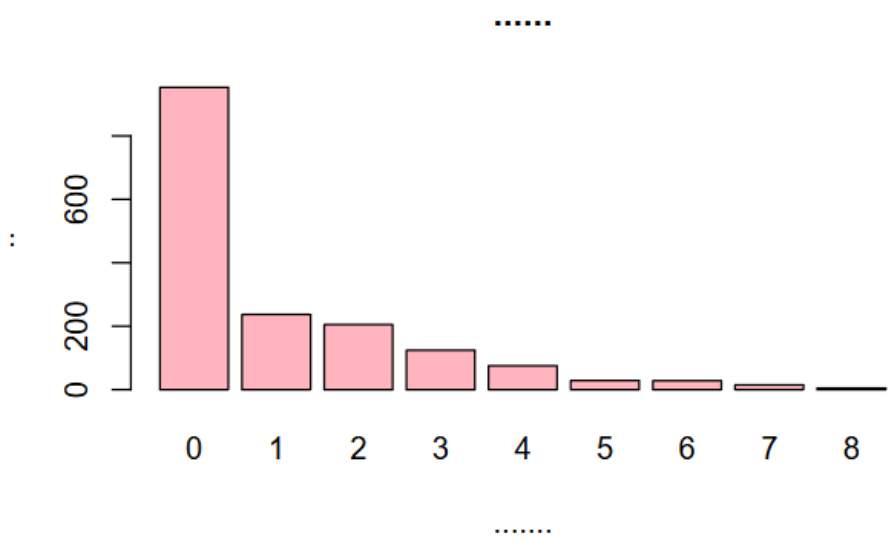
- 本資料集無缺失，有153筆重複樣本，但因為無法透過問卷調查的回答選項得知他是不是真的為相同受訪者，所以還是把他們當作不同來進行後續分析，並有將對重複資料進行分組 & 計算出現次數的結果先做匯出。

2. 候選人支持度分析

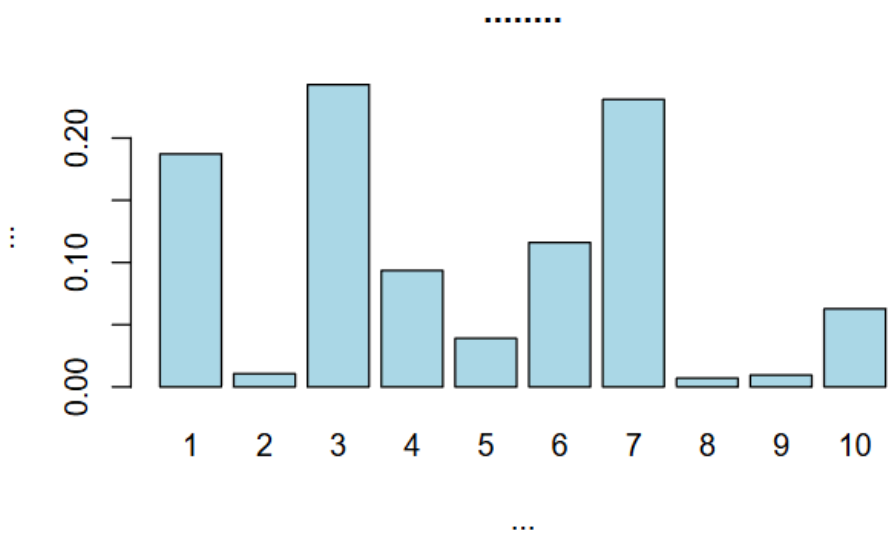
支持度計算與視覺化

- **支持度定義: 扣除”無反應”、“尚未決定”、“遺漏值或跳答”等選項後，各候選人在v5(“假設明天要投票，被訪者會投給誰”)的計數中，除以所有人得到的總票數**

```
barplot(table(FullData$political_interest), main = "    ",
        xlab = "    ", ylab = " ", col = "lightpink")
```




```
support = table(FullData$v5)[-11:-13]/sum(table(FullData$v5)[-11:-13])
barplot(support, xlab = "  ", ylab = "  ",
        main = "     ", col = "lightblue")
```



小結論：

- 支持度資料已排除”無反應”、“尚未決定”、“遺漏值或跳答”等選項
- 3號、7號、1號為支持率較高的前三者。

3. 3號候選人競選策略建議

```
v5_3 = FullData[which(FullData$v5==3),]
```

地理分布

```
geo_north = table(v5_3$v2)
geo_midwest = table(v5_3$v3)
geo_north; geo_midwest
```

```
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
 1  2  0  1  8  1  7 23  5  9  2  5  1  2  4  1  2  2  4  1  5  2  6  5  7  4
27 28 29 30 31 32 33 44 98 99
 8  3  2  1  7 11  8 11  2 42
```

```
  1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20
  1   1   1   5   4   0   2   0   0   1   2   2   1   1   0   1   8   1   5   0
 44  98  99
  5   1 163
```

地區分析結論：

- 北區支持者以08 “大港里”為多數
- 中西區支持者以17 “西和里”為多數

人口特徵分析

```
age = table(v5_3$v6)
edu = table(v5_3$v7)
sex = table(v5_3$v8)
age; edu; sex
```

```
  1   2   3   4   5   6
  5  12  31  52 105   0
```

```
 1  2  3  4  5 95
27 26 64 31 57  0
```

```
  1   2
 88 117
```

人口特徵結論：

- 支持者年齡已”60歲及以上”占多數
- 支持者的教育程度偏高中高職族群較多。
- 女性支持者較多。

小結論：

- 目前北區的”大港里”，中西區的”西和里”有最多支持者，3號候選人應鞏固其票倉；而北區的”北華里”、“東興里”、“公園里”、“長勝里”、“力行里”、“永祥里”、“雙安里”皆只有1票；“元寶里”、“賢北里”、“振興里”、“重興里”、“仁愛里”、“大光里”、“立人里”皆只有2票，可以多去拉票。
- 目前中西區的”西和里”為多數，應該鞏固其票倉，而中西區的遺漏值太多，應該先改進這部份再來看哪個票倉太少，進而多去拉票
- 年齡越大好像越支持3號候選人，所以應該先轉向爭取年紀較小的選民(ex.20-40歲的青壯年)支持；教育程度則是要提升”專科”支持(因為高中和大學以上目前支持人數都足夠，而考慮到投票年齡限制，尋求專科支持能帶來最大效益);性別差不多平衡，可以繼續保持

4. 政治熱衷程度的分析

有序邏輯斯模型

```
library(tidyverse)
library(MASS)

FullData = FullData %>% mutate(across(c(v1,v6,v7,v8), factor))
model_4 = polr(political_interest ~ v1 + v6 + v7 + v8, data = FullData, Hess = TRUE)
summary(model_4)
```

```
Call:
polr(formula = political_interest ~ v1 + v6 + v7 + v8, data = FullData,
    Hess = TRUE)

Coefficients:
       Value Std. Error t value
v12  -0.1930    0.10295 -1.8749
v62   0.6400    0.34620  1.8487
v63   0.7980    0.31531  2.5310
v64   0.9338    0.30536  3.0580
v65   0.7978    0.30066  2.6534
v66  -0.4569    0.75199 -0.6076
v72   0.4045    0.19738  2.0495
v73   0.6725    0.16010  4.2006
v74   0.7374    0.19436  3.7941
v75   0.6612    0.17119  3.8622
v795 -0.7700    0.48096 -1.6009
v82  -0.2406    0.09891 -2.4321

Intercepts:
    Value   Std. Error t value
0|1  1.3443  0.3345     4.0191
1|2  1.9921  0.3362     5.9247
2|3  2.7291  0.3396     8.0358
3|4  3.4240  0.3446     9.9358
4|5  4.1591  0.3542    11.7421
```

```
5|6   4.6523   0.3652    12.7375
6|7   5.5466   0.4033    13.7538
7|8   6.9425   0.5591    12.4164

Residual Deviance: 4539.828
AIC: 4579.828
```

```r
p_4 = pnorm(abs(coef(summary(model_4))[, "t value"]), lower.tail = FALSE)*2
OR_4 = exp(coef(model_4))
print(cbind(ODDS = round(OR_4,2), p = round(p_4,4)))
```

```
     ODDS      p
v12  0.82 0.0608
v62  1.90 0.0645
v63  2.22 0.0114
v64  2.54 0.0022
v65  2.22 0.0080
v66  0.63 0.5435
v72  1.50 0.0404
v73  1.96 0.0000
v74  2.09 0.0001
v75  1.94 0.0001
v795 0.46 0.1094
v82  0.79 0.0150
0|1  0.82 0.0001
1|2  1.90 0.0000
2|3  2.22 0.0000
3|4  2.54 0.0000
4|5  2.22 0.0000
5|6  0.63 0.0000
6|7  1.50 0.0000
7|8  1.96 0.0000
```

參數解釋與模型結論：

- 我們選擇使用有序邏輯斯迴歸模型探討政治熱衷程度的影響因素。因為所有變數放入會產生太多dummy variable，讓模型無法收斂，所以在自變數x的選擇上，我們放入v1(北區or中西區), v6(年齡), v7(教育程度), v8(性別)，應變數y則是我們自行定義的political_interest，其中0~9的值則定義為熱衷程度
- 由於我們模型得出t-value，我們再自行轉換到p-value，以檢查變數哪些是顯著的，同時新增odds ratio的還原以檢視變數之間的倍率比較
- 擬合的結果顯示，v6(年齡)與v7(教育程度)在多數分類中對政治熱中程度具有正向且顯著的影響。例如，某些年齡層，如v63(年齡40-49歲) 與 v64(年齡50-59歲)的人比參考年齡層v61(20-29歲)更有 2.22倍 倍以上的機率有更高的政治熱衷程度；同樣，教育程度的某些分類，如v74(專科)也比參考組v71(小學或以下)展現了2.09倍以上的機率。此外，性別或特定區域分類中的某些組別如v82(女性)，則顯示對政治的關心程度較低。

5. 3號候選人支持預測模型與不平衡資料處理

加權邏輯斯模型

```r
support_3 = ifelse(FullData$v5 == 3, 1, 0)
FullData$support_3 = factor(support_3)
weight = ifelse(FullData$support_3 == 1, 1466/205, 1)
model_5 = glm(support_3 ~ v1 + v6 + v7 + v8, data = FullData, weights = weight, family = quasibinomial(link = "l
summary(model_5)
```

```
Call:
glm(formula = support_3 ~ v1 + v6 + v7 + v8, family = quasibinomial(link = "logit"),
    data = FullData, weights = weight)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.02604    0.35805  -0.073  0.94203
v12          -0.72458    0.11432  -6.338 2.99e-10 ***
v62           0.25572    0.36474   0.701  0.48333
v63           0.41968    0.32942   1.274  0.20285
v64           0.31403    0.32079   0.979  0.32777
v65          -0.04806    0.31843  -0.151  0.88004
v66         -14.01122  380.81747  -0.037  0.97065
v72           0.58699    0.19893   2.951  0.00321 **
v73           0.35915    0.16343   2.198  0.02812 *
v74           0.34760    0.20281   1.714  0.08674 .
v75          -0.08044    0.18291  -0.440  0.66017
v795        -15.09466  338.80208  -0.045  0.96447
v82          -0.06048    0.10482  -0.577  0.56400
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasibinomial family taken to be 1.725732)

    Null deviance: 4064.6  on 1670  degrees of freedom
Residual deviance: 3832.6  on 1658  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 15
```

模型結論：

- 影響選民支持 3 號候選人的因素，使用加權邏輯斯迴歸模型處理資料中類別不平衡問題。其中weight設定為”沒有投給3號/投給3號的”
- 結果發現，「區域(v12)」與「教育程度(v72-v74)」為影響支持意向的主要因素。特定區域如 v12(中西區)居民顯著不支持 3 號候選人，而特定教育分類如 v72(國中) 與 v73(高中)則與支持意向呈正向關係。年齡與性別在本模型中未呈現顯著效果。

轉換成勝算比(倍數概念)

```
OR_5 = exp(coef(model_5))
round(OR_5, 2)
```

```
(Intercept)         v12         v62         v63         v64         v65
       0.97        0.48        1.29        1.52        1.37        0.95
        v66         v72         v73         v74         v75        v795
       0.00        1.80        1.43        1.42        0.92        0.00
        v82
       0.94
```

結論：

- 教育程度為初中、國中、高中、高職、專科較支持三號候選人。
- 戶籍在中西區者支持度明顯較低，相比北區有顯著差距。
- 年齡與性別對支持與否的影響並不明顯（或樣本不足）。

建議:

- **主攻的受眾: 加強對教育程度為初中、國中、高中、高職、專科背景選民的鞏固，且應優先鎖定這群人作為核心支持群體。**
- **區域策略：中西區的支持度偏低，建議針對此區進行針對性的政見溝通或補強形象。可能考慮增加在中西區的曝光率或推出地方型政策吸引當地選民。**