HW4

高嘉妤、柯堯珹、吳承恩、趙友誠

2024-10-26

Table of contents


```
if(!require(rio)){
  install.packages("rio")
  library(rio)
}
if(!require(labelled)){
  install.packages("labelled")
  library(labelled)
}
if(!require(Hmisc)){
  install.packages("Hmisc")
  library(Hmisc)
}
if(!require(dplyr)){
  install.packages("dplyr")
  library(dplyr)
}
if(!require(ggplot2)){
  install.packages("ggplot2")
  library(ggplot2)
}
if(!require(MASS)){
  install.packages("MASS")
  library(MASS)
}
pollsav <- import("poll.sav")
str(pollsav)
```

資料簡介

Dimension of the Data : ***1671 samples × 15 columns***

Table 1: 變數解釋

Variables	Explanation	remark
V1	District	1: 北區, 2: 中西區
V2、V3	Li	
V4_1~V4_8	Candidate known	1~10 號
V5	Candidate supported	1~10 號
V6	Age	1:20 到 29 歲,2:30 到 39 歲,3:40 到 49 歲,4:50 到 59 歲,5:60 歲以上
V7	Education level	1: 小學, 2: 國中, 3: 高中, 4: 專科, 5: 大學以上
V8	Sex	1:male, 2:female

資料前處理

資料整理

```
pollcsv <- data.frame(
  apply(pollsav,2,function(col){
    as.factor(
      remove_attributes(col,
                        attributes = c("label","format.spss",
                                       "display_width","labels")))
})) # 因為 sav 格式的" 屬性" 會造成 describe
pollcsv <- remove_attributes(pollcsv, "dimnames")
n <- dim(pollcsv)[1]
latex(describe(pollcsv), file="")
```

pollcsv
15 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 10 11 12 13, highest: 7 8 9 98 99

v3

n	missing	distinct
1671	0	23

lowest : 1 10 11 12 13, highest: 7 8 9 98 99

v4_1

n	missing	distinct
1671	0	12

Value	1	10	2	3	4	5	6	7	8	9	91	98
Frequency	328	11	5	214	43	27	38	47	4	1	14	939
Proportion	0.196	0.007	0.003	0.128	0.026	0.016	0.023	0.028	0.002	0.001	0.008	0.562

v4_2

n	missing	distinct
1671	0	10

Value	10	2	3	4	5	6	7	8	9	99
Frequency	15	6	189	59	32	75	99	2	4	1190
Proportion	0.009	0.004	0.113	0.035	0.019	0.045	0.059	0.001	0.002	0.712

---

v4_3

n	missing	distinct
1671	0	9

Value	10	3	4	5	6	7	8	9	99
Frequency	19	6	60	36	61	91	1	2	1395
Proportion	0.011	0.004	0.036	0.022	0.037	0.054	0.001	0.001	0.835

---

v4_4

n	missing	distinct
1671	0	8

Value	10	4	5	6	7	8	9	99
Frequency	20	4	28	41	52	3	4	1519
Proportion	0.012	0.002	0.017	0.025	0.031	0.002	0.002	0.909

---

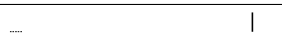
v4_5

n	missing	distinct
1671	0	7

Value	10	5	6	7	8	9	99
Frequency	15	3	14	38	4	3	1594
Proportion	0.009	0.002	0.008	0.023	0.002	0.002	0.954

---

v4_6



n	missing	distinct
1671	0	6

Value	10	6	7	8	9	99
Frequency	20	3	12	6	7	1623
Proportion	0.012	0.002	0.007	0.004	0.004	0.971

---

v4_7

n	missing	distinct
1671	0	5

Value	10	7	8	9	99
Frequency	12	3	2	3	1651
Proportion	0.007	0.002	0.001	0.002	0.988

---

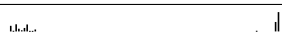
v4_8

n	missing	distinct
1671	0	3

Value	10	8	99
Frequency	4	1	1666
Proportion	0.002	0.001	0.997

---

v5



n	missing	distinct
1671	0	13

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

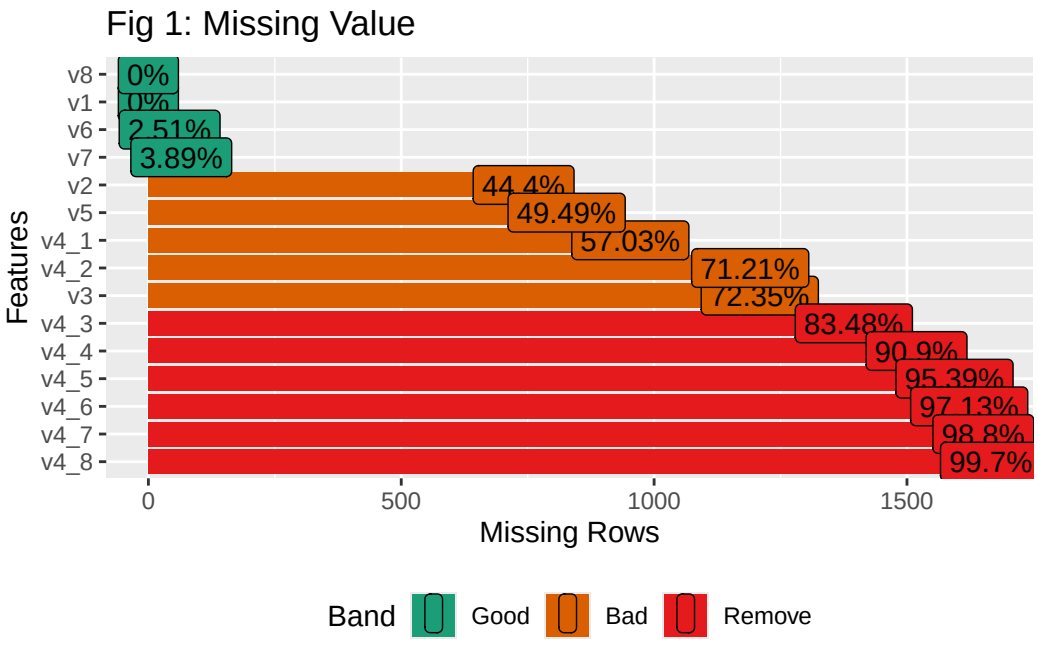
Table 2: 遺失值定義

Variables	Missing
V1	98,99
V2、V3	44,98,99
V4_1~V4_8	91,98,99
V5	91,98,99
V6	6,99
V7	95,99
V8	99

遺失值比例圖

將定義的遺失值轉換成 NA 並以遺失值比例圖 (by variable) 的方式呈現。考量到遺失值的性質，我們並未刪除任何資料，決定後續對不同變數分析時再移除。

```
pollcsv <- data.frame(
  t(apply(pollcsv,MARGIN = 1, FUN = function(row){
    row[row==99 | row==98 | row==95 | row==91 | row==44] <- NA
    return(row)
  }))
)
pollcsv$v6[pollcsv$v6==6] <- NA
DataExplorer::plot_missing(pollcsv, title = "Fig 1: Missing Value")
```



候選人支持率分析表

支持度定義: 支持度 = $\frac{\text{第五題出現次數}}{\text{樣本數}}$

```
# 計算總體支持度
count5.total <- unlist(lapply(1:11,function(x){
  if(x==11) return(sum(is.na(pollcsv$v5))/n)
  else return(sum(pollcsv$v5[!is.na(pollcsv$v5)]==x)/n)
} ))
# 計算分區支持度（北區中西區）v1
support.district <- do.call(rbind, lapply(1:2,function(i){
  tempdata <- pollcsv[pollcsv$v1==i,]
  n.temp <- dim(tempdata)[1]
  return(unlist(
    lapply(1:11, function(x){
      if(x==11) return(sum(is.na(tempdata$v5))/n.temp)
      else return(sum(tempdata$v5[!is.na(tempdata$v5)]==x)/n.temp)
    })))
}))
# 計算性別支持度 v8
support.sex <- do.call(rbind, lapply(1:2,function(i){
  tempdata <- pollcsv[pollcsv$v8==i,]
  n.temp <- dim(tempdata)[1]
  return(unlist(
    lapply(1:11, function(x){
      if(x==11) return(sum(is.na(tempdata$v5))/n.temp)
      else return(sum(tempdata$v5[!is.na(tempdata$v5)]==x)/n.temp)
    })))
}))
# 計算年齡支持度 v6
support.age <- do.call(rbind, lapply(1:5,function(i){
  tempdata <- pollcsv[pollcsv$v6==i,]
  n.temp <- dim(tempdata)[1]
  return(unlist(
    lapply(1:11, function(x){
      if(x==11) return(sum(is.na(tempdata$v5))/n.temp)
      else return(sum(tempdata$v5[!is.na(tempdata$v5)]==x)/n.temp)
    })))
}))
# 計算教育程度支持度 v7
support.edu <- do.call(rbind, lapply(1:5,function(i){
  tempdata <- pollcsv[pollcsv$v7==i,]
  n.temp <- dim(tempdata)[1]
  return(unlist(
    lapply(1:11, function(x){
      if(x==11) return(sum(is.na(tempdata$v5))/n.temp)
      else return(sum(tempdata$v5[!is.na(tempdata$v5)]==x)/n.temp)
    })))
}))
table.support <- rbind(
  count5.total,
  support.district,
  support.sex,
  support.age,
  support.edu
```

table.support	1	2	3	4	5	6	7	8	9	10	沒決定
總計											
	9.5%	0.5%	12.3%	4.7%	2%	5.9%	11.7%	0.4%	0.5%	3.2%	49.5%
分區											
北區	5.1%	0.6%	14.7%	2.9%	2.6%	7.5%	12.9%	0.3%	0.4%	2.7%	50.3%
中西區	18.1%	0.4%	7.4%	8.3%	0.7%	2.7%	9.2%	0.5%	0.7%	4.1%	47.9%
性別											
男性	9.8%	0.9%	12.9%	5.6%	2.5%	7.3%	11.6%	0.7%	0.3%	4%	44.4%
女性	9.2%	0.3%	11.8%	4.1%	1.6%	4.9%	11.7%	0.1%	0.6%	2.6%	53%
年齡											
20 到 29 歲	3.2%	1.1%	5.3%	3.2%	0%	1.1%	11.7%	1.1%	0%	1.1%	72.3%
30 到 39 歲	5.9%	1.5%	8.8%	1.5%	2.2%	4.4%	11.8%	1.5%	0.7%	2.9%	58.8%
40 到 49 歲	4.5%	1.2%	12.8%	4.5%	3.3%	5.3%	16%	0%	0.8%	1.2%	50.2%
50 到 59 歲	10.6%	0.8%	13.8%	5%	2.6%	5.8%	11.4%	0.3%	0.5%	1.9%	47.4%
60 歲以上	9.6%	0%	10.6%	4.5%	1.2%	5.7%	8.6%	0.2%	0.3%	3.8%	55.5%
學歷											
小學	8.7%	0%	7.6%	1.4%	0.6%	3.4%	5%	0.3%	0%	1.1%	72%
國中	7.8%	0%	11.3%	2.6%	1.3%	2.2%	7.4%	0%	0%	3%	64.3%
高中	9.1%	0%	12.9%	5%	2.6%	6.5%	9.5%	0.4%	0.8%	3.2%	50%
專科	7.2%	0.4%	11.8%	3.8%	2.3%	6.1%	7.6%	0%	0%	2.3%	58.6%
大學以上	7.2%	1.4%	9.7%	5.3%	1.5%	5.6%	15.4%	0.5%	0.7%	3.4%	49.2%

```
)
table.support <- data.frame(
  apply(table.support, 2, function(col) paste0(round(col,3)*100,"%"))
)
rownames(table.support) <- c(
  "",
  " 北區"," 中西區",
  " 男性"," 女性",
  "20 到 29 歲","30 到 39 歲","40 到 49 歲","50 到 59 歲","60 歲以上",
  " 小學"," 國中"," 高中"," 專科"," 大學以上 ")
colnames(table.support) <- c(1:10," 沒決定")
latex(table.support, file = "",
      rgroup = c(" 總計"," 分區"," 性別"," 年齡"," 學歷"),
      n.rgroup = c(1,2,2,5,5)
)
```

三號候選人的競選策略 (需在何地、對何人進行拉票)

受訪者政治熱衷程度之統計模型 (需說明使用此模型之理由)

三號候選人支持率預測模式

資料不平衡處理