

HW4

高嘉好、柯堯城、吳承恩、趙友誠

2024-10-27

Table of contents

資料簡介	2
資料前處理	2
資料整理	2
遺失值比例圖	4
候選人支持率分析表	5
三號候選人的競選策略 (需在何地、對何人進行拉票)	6
三號候選人之里 heatmap	6
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)	8
三號候選人支持率預測模式	12
資料不平衡處理	12

```
if(!require(rio)){          #read sav file
  install.packages("rio")
  library(rio)
}
if(!require(labelled)){     #remove attribute of sav data
  install.packages("labelled")
  library(labelled)
}
if(!require(Hmisc)){        #describe
  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)
}
if(!require(sf)){          #render map
```

```

install.packages("sf")
library(sf)
}
if(!require(twmap)){      #map data
  remotes::install_github("shihjyun/twmap")
  library(twmap)
}
if(!require(showtext)){  #show zw-tw in ggplot2
  install.packages("showtext")
  library(showtext)
}
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	v2: 33 個里, v3: 20 個里
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

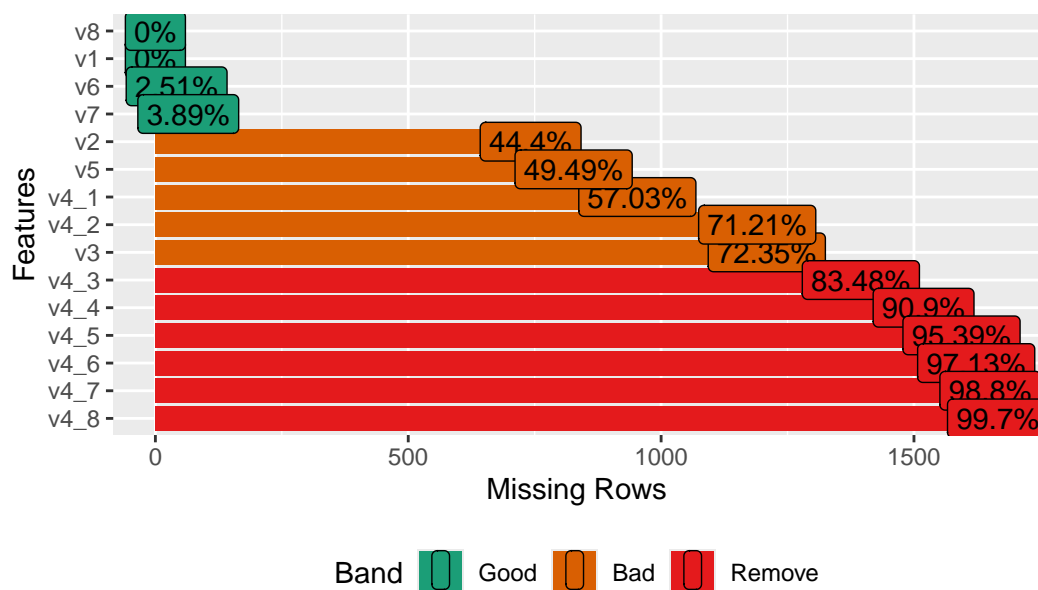
Variables	Missing
V1	98,99
V2、V3	44,98,99
V4_1~V4_8	91,98,99
V5	91,98,99
V6	6,99
V7	95,99
V8	99

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")
```

Fig 1: Missing Value



候選人支持率分析表

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

```
# 計算總體支持度
count5.total <- sapply(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$v1==i,]
  n.temp <- dim(tempdata)[1]
  return(sapply(1:11, function(x){
    if(x==11) return(sum(is.na(tempdata$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$v8==i,]
  n.temp <- dim(tempdata)[1]
  return(sapply(1:11, function(x){
    if(x==11) return(sum(is.na(tempdata$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$v6==i,]
  n.temp <- dim(tempdata)[1]
  return(sapply(1:11, function(x){
    if(x==11) return(sum(is.na(tempdata$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$v7==i,]
  n.temp <- dim(tempdata)[1]
  return(sapply(1:11, function(x){
    if(x==11) return(sum(is.na(tempdata$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 <- 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 = "",title="",
  rgroup = c(" 總計"," 分區"," 性別"," 年齡"," 學歷"),
  n.rrgroup = c(1,2,2,5,5),
  caption = " 候選人支持度整理表"
)

```

三號候選人的競選策略(需在何地、對何人進行拉票)

三號候選人之里 heatmap

```

# 計算三號候選人對於里的支持度
support.li_north <- data.frame(
  support = sapply(1:33, function(i){
    tempdata <- pollcsv[pollcsv$v2==i,]
    n.temp <- dim(tempdata)[1]
    return(sum(tempdata$v5[!is.na(tempdata$v5)]==3)/n.temp)}
),
  VILLNAME = names(attr(pollsav$v2,"labels"))[1:33]
)
support.li_midwest <- data.frame(
  support = sapply(1:20, function(i){
    tempdata <- pollcsv[pollcsv$v3==i,]
    n.temp <- dim(tempdata)[1]

```

Table 3: 候選人支持度整理表

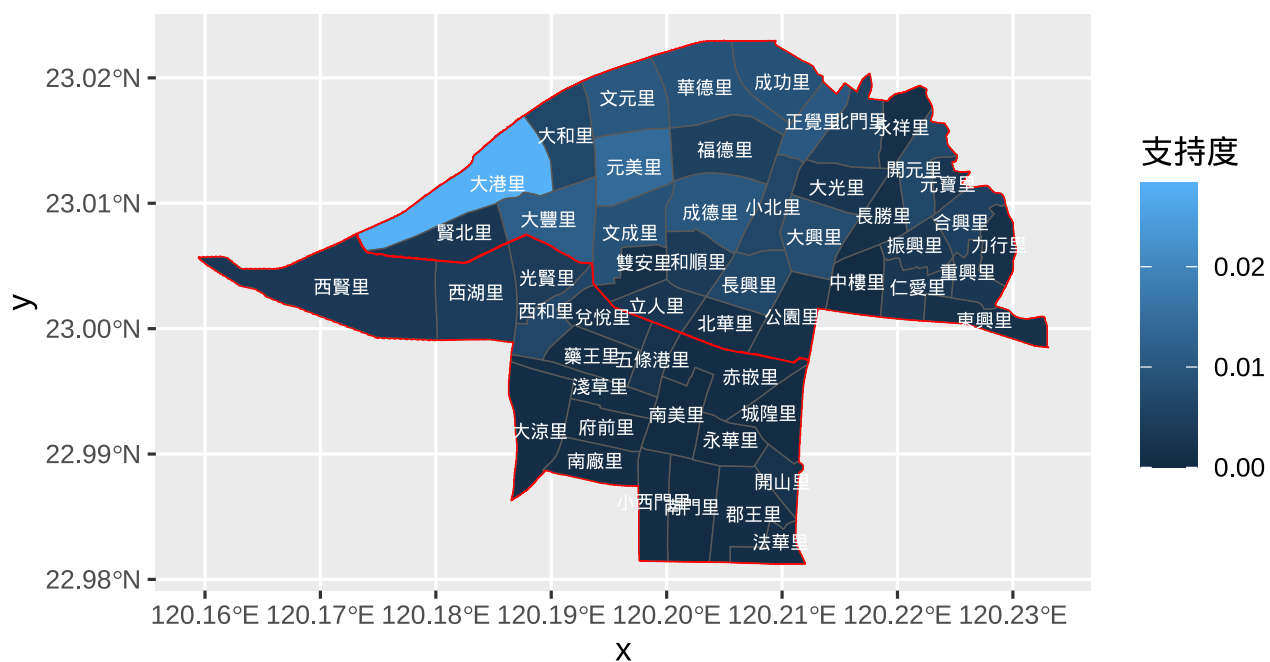
	1	2	3	4	5	6	7	8	9	10	沒決定
總計	9.5%	0.5%	12.3%	4.7%	2%	5.9%	11.7%	0.4%	0.5%	3.2%	49.5%
分區											
北區	5.1%	0.6%	14.7%	2.9%	2.6%	7.5%	12.9%	0.3%	0.4%	2.7%	50.3%
中西區	18.1%	0.4%	7.4%	8.3%	0.7%	2.7%	9.2%	0.5%	0.7%	4.1%	47.9%
性別											
男性	9.8%	0.9%	12.9%	5.6%	2.5%	7.3%	11.6%	0.7%	0.3%	4%	44.4%
女性	9.2%	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%

```

    return(sum(tempdata$v5[!is.na(tempdata$v5)]==3)/n.temp)
  }),
  VILLNAME = names(attr(pollsav$v3,"labels"))[1:20]
)
myMap <- tw_village[
  tw_village$COUNTYNAME == " 臺南市" &
  (tw_village$TOWNNAME==" 中西區"| tw_village$TOWNNAME==" 北區"),]
myMap <- merge(x = myMap, y = rbind(support.li_midwest, support.li_north), by = "VILLNAME")
showtext_auto()
ggplot(data = myMap) +
  geom_sf(aes(fill = support)) + # 填充區域
  geom_sf(
    data = summarize(
      group_by(myMap,TOWNNAME),
      geometry = st_union(st_buffer(geometry,dist = 0.01))) , fill = NA, color = 'red') +
    #st_buffer 是為了解決 union 之後內部還有線條的問題
  geom_sf_text(aes(label=VILLNAME), size = 2, color = "white")+
  ggtitle("Fig 2: 三號候選人支持度熱區圖")+
  labs(fill = " 支持度")+
  theme_gray(base_family="Arial", base_size = 10)

```

Fig 2: 三號候選人支持度熱區圖



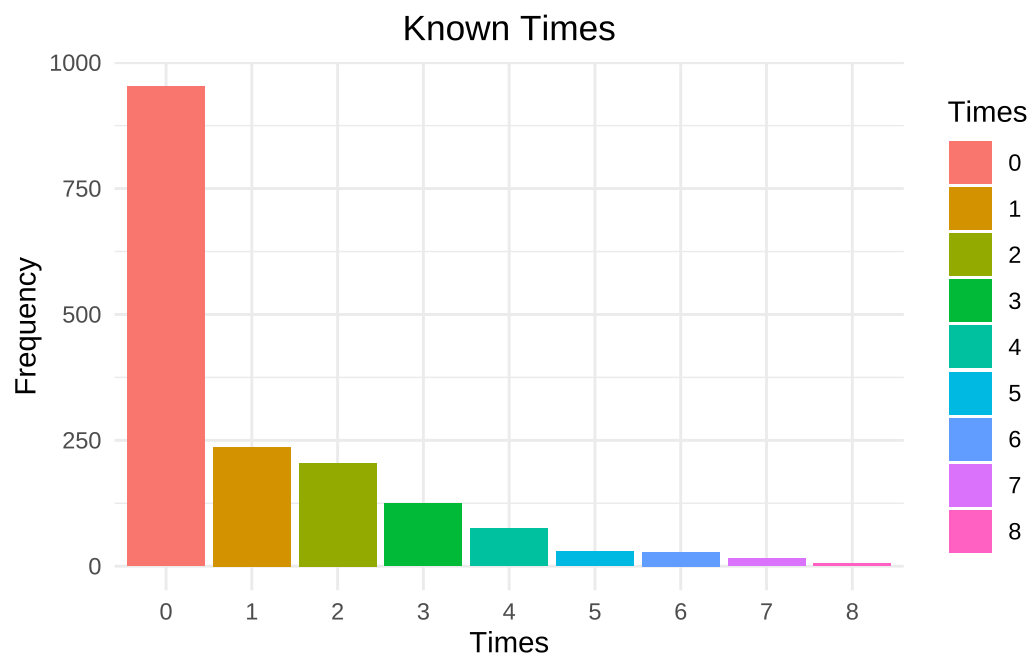
受訪者政治熱衷程度之統計模型 (需說明使用此模型之理由)

```
pollcsv$known_count <- rowSums(!is.na(pollcsv[,c("v4_1", "v4_2", "v4_3", "v4_4", "v4_5", "v4_6", "v4_7")])

count_da <- sapply(0:8,function(x){
  sum(pollcsv$known_count==x)
})

count_data <- data.frame(
  Times = factor(0:8),
  Values = count_da
)

ggplot(count_data, aes(x = Times, y = Values , fill = Times ))+
  geom_bar(stat = 'identity')+
  scale_x_discrete(breaks = 0:8)+
  labs(title='Known Times', x = 'Times', y = 'Frequency')+
  theme_minimal()+
  theme(plot.title = element_text(hjust = 0.5))
```

```
poiv4<-glm(known_count~v1+v6+v7+v8, data = pollcsv, family = poisson())
summary(poiv4)
```

Call:

```
glm(formula = known_count ~ v1 + v6 + v7 + v8, family = poisson(),
    data = pollcsv)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-1.07370	0.22022	-4.876	1.09e-06	***
v12	-0.14837	0.05231	-2.836	0.004562	**
v62	0.78280	0.22353	3.502	0.000462	***
v63	0.89771	0.21085	4.258	2.07e-05	***
v64	0.99635	0.20653	4.824	1.41e-06	***
v65	0.92553	0.20504	4.514	6.37e-06	***
v72	0.25686	0.10656	2.411	0.015926	*
v73	0.50898	0.08480	6.002	1.95e-09	***
v74	0.51496	0.09952	5.174	2.29e-07	***
v75	0.49976	0.08920	5.602	2.11e-08	***
v82	-0.17905	0.04880	-3.669	0.000243	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 3501.4 on 1600 degrees of freedom

Residual deviance: 3383.9 on 1590 degrees of freedom

(因為不存在，70 個觀察量被刪除了)

AIC: 5292

Number of Fisher Scoring iterations: 6

```
nbv4 <- glm.nb(known_count~v1+v6+v7+v8, data = pollcsv)
summary(nbv4)
```

Call:

```
glm.nb(formula = known_count ~ v1 + v6 + v7 + v8, data = pollcsv,
  init.theta = 0.6258750942, link = log)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-1.07825	0.30518	-3.533	0.000411	***
v12	-0.14018	0.08591	-1.632	0.102753	
v62	0.77007	0.31503	2.444	0.014506	*
v63	0.89173	0.29142	3.060	0.002213	**
v64	1.01004	0.28303	3.569	0.000359	***
v65	0.93793	0.27922	3.359	0.000782	***
v72	0.25893	0.16467	1.572	0.115855	
v73	0.50668	0.13195	3.840	0.000123	***
v74	0.51789	0.15992	3.239	0.001201	**
v75	0.50471	0.14010	3.603	0.000315	***
v82	-0.19371	0.08241	-2.350	0.018749	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(0.6259) family taken to be 1)

Null deviance: 1531.0 on 1600 degrees of freedom

Residual deviance: 1485.5 on 1590 degrees of freedom

(因為不存在，70 個觀察量被刪除了)

AIC: 4584.9

Number of Fisher Scoring iterations: 1

Theta: 0.6259

Std. Err.: 0.0473

2 x log-likelihood: -4560.9200

```
if(!require(AER)){
  install.packages("AER")
  library(AER)
}
dispersiontest(poiv4)
```

Overdispersion test

data: poiv4

z = 12.524, p-value < 2.2e-16

alternative hypothesis: true dispersion is greater than 1

sample estimates:

dispersion

2.321796

```
lrtest(poi4,nbv4) # 決定要用 Poisson 還是 Negative binomial
```

Likelihood ratio test

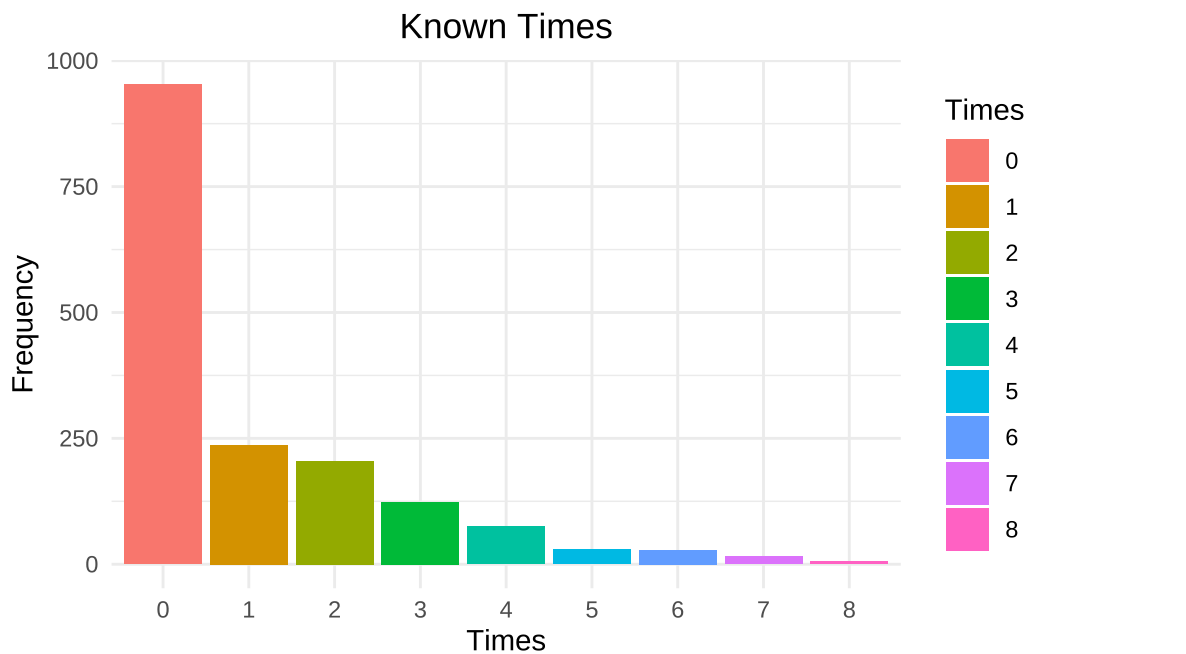
Model 1: known_count ~ v1 + v6 + v7 + v8

Model 2: known_count ~ v1 + v6 + v7 + v8

```
#Df LogLik Df Chisq Pr(>Chisq)
1 11 -2635.0
2 12 -2280.5 1 709.13 < 2.2e-16 ***
---
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```
ggplot(count_data, aes(x = Times, y = Values , fill = Times ))+ # 建立次數圖
  geom_bar(stat = 'identity')+
  scale_x_discrete(breaks = 0:8)+
  labs(title='Known Times', x = 'Times', y = 'Frequency')+
  theme_minimal()+
  theme(plot.title = element_text(hjust = 0.5))
```



```
if(!require(pscl)){
  install.packages("pscl")
  library(pscl)
}
zinb_model <- zeroinfl(known_count ~ v1 +v6+v7+v8, data = pollcsv, dist = "negbin")# 建立 zero-inflated
summary(zinb_model)
```

Call:

```
zeroinfl(formula = known_count ~ v1 + v6 + v7 + v8, data = pollcsv, dist = "negbin")
```

Pearson residuals:

Min	1Q	Median	3Q	Max
-0.8393	-0.7052	-0.5357	0.4613	4.6745

```

Count model coefficients (negbin with log link):
      Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.587285    0.399595  -1.470  0.14164
v12          -0.094520    0.079076  -1.195  0.23196
v62           1.043282    0.398262   2.620  0.00880 **
v63           1.081087    0.384695   2.810  0.00495 **
v64           1.185464    0.382191   3.102  0.00192 **
v65           1.240723    0.383341   3.237  0.00121 **
v72          -0.008993    0.158309  -0.057  0.95470
v73           0.162114    0.124789   1.299  0.19391
v74           0.201598    0.146656   1.375  0.16925
v75           0.254145    0.131547   1.932  0.05336 .
v82          -0.058140    0.072576  -0.801  0.42308
Log(theta)   1.476716    0.266200   5.547  2.9e-08 ***

Zero-inflation model coefficients (binomial with logit link):
      Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.7381      1.2734  -0.580  0.56219
v12           0.1319      0.1521   0.867  0.38574
v62           0.7652      1.2615   0.607  0.54415
v63           0.5669      1.2510   0.453  0.65042
v64           0.5615      1.2538   0.448  0.65428
v65           0.8606      1.2597   0.683  0.49447
v72          -0.4869      0.2819  -1.727  0.08408 .
v73          -0.6751      0.2189  -3.084  0.00204 **
v74          -0.6209      0.2714  -2.288  0.02215 *
v75          -0.4509      0.2308  -1.954  0.05070 .
v82           0.2961      0.1477   2.005  0.04496 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Theta = 4.3785
Number of iterations in BFGS optimization: 69
Log-likelihood: -2239 on 23 Df
lrtest(nbv4,zinb_model) # 決定要用 Negative 或 zero-inflated

Likelihood ratio test

Model 1: known_count ~ v1 + v6 + v7 + v8
Model 2: known_count ~ v1 + v6 + v7 + v8
      #Df LogLik Df  Chisq Pr(>Chisq)
1    12 -2280.5
2    23 -2238.9 11 83.162 3.599e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

三號候選人支持率預測模式

資料不平衡處理