迴歸分析期末報告

全台鄉鎮市區中低收入戶之因素影響分析

指導教授:鄭宗記教授
學生:陳槐廷

一、變數名稱

Y	中低收入戶
X1	20-24 歲大學生人數
X2	平均收入(以千為單位)
X3	撫養比
X4	金融及保險業
X5	醫院數
X6	結婚對數
X7	老人戶數
X8	總增加人口
X9	科技業
X10	總戶數

二、敘述統計

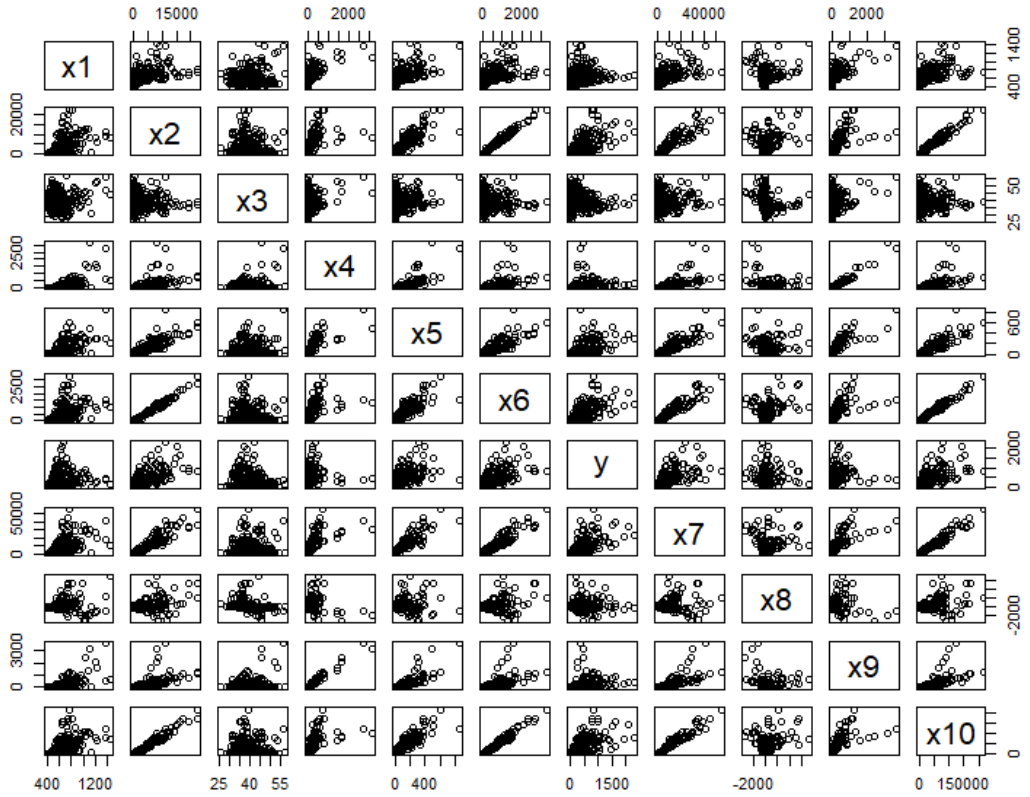
(1)最小值，第一四分位距，中位數，平均數、第三四分位距、最大值

x1	x2	x3	x4	x5
Min. : 388.0	Min. : 40	Min. :25.79	Min. : 0.00	Min. : 0.00
1st Qu.: 513.8	1st Qu.: 525	1st Qu.:35.93	1st Qu.: 3.00	1st Qu.: 5.75
Median : 557.5	Median : 1324	Median :39.13	Median : 10.00	Median : 16.00
Mean : 602.1	Mean : 2822	Mean :39.66	Mean : 97.91	Mean : 62.88
3rd Qu.: 645.2	3rd Qu.: 3673	3rd Qu.:43.51	3rd Qu.: 62.25	3rd Qu.: 67.25
Max. :1442.0	Max. :22349	Max. :56.90	Max. :3166.00	Max. :862.00

x6	y	x7	x8	x9
Min. : 1.0	Min. : 0	Min. : 47	Min. :-3385.00	Min. : 0.00
1st Qu.: 67.5	1st Qu.: 89	1st Qu.: 1912	1st Qu.: -287.00	1st Qu.: 4.75
Median : 151.0	Median : 196	Median : 3787	Median : -139.00	Median : 17.50
Mean : 357.6	Mean : 315	Mean : 6614	Mean : 38.56	Mean : 137.41
3rd Qu.: 457.0	3rd Qu.: 440	3rd Qu.: 7179	3rd Qu.: 16.50	3rd Qu.: 88.50
Max. :3278.0	Max. :2294	Max. :55398	Max. : 7180.00	Max. :3617.00

x10
Min. : 135
1st Qu.: 5082
Median : 10718
Mean : 24002
3rd Qu.: 28116
Max. :215279

(2)散佈圖



三、模型選取

1.full model

$$y = \boldsymbol{\beta}_0 + \boldsymbol{x}_1\boldsymbol{\beta}_1 + \boldsymbol{x}_2\boldsymbol{\beta}_2 + \boldsymbol{x}_3\boldsymbol{\beta}_3 + \boldsymbol{x}_4\boldsymbol{\beta}_4 + \boldsymbol{x}_5\boldsymbol{\beta}_5 + \boldsymbol{x}_6\boldsymbol{\beta}_6 + \boldsymbol{x}_7\boldsymbol{\beta}_7 + \boldsymbol{x}_8\boldsymbol{\beta}_8 + \boldsymbol{x}_9\boldsymbol{\beta}_9 + \boldsymbol{x}_{10}\boldsymbol{\beta}_{10}$$

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	576.269314	123.069292	4.682	4.03e-06	***
x1	-0.446059	0.122342	-3.646	0.000306	***
x2	0.068249	0.028847	2.366	0.018521	*
x3	-5.825430	2.563169	-2.273	0.023635	*
x4	0.416824	0.227441	1.833	0.067685	.
x5	2.011037	0.367318	5.475	8.27e-08	***
x6	0.094386	0.262374	0.360	0.719256	
x7	0.019012	0.011998	1.585	0.113956	
x8	-0.059828	0.019904	-3.006	0.002835	**
x9	-0.780675	0.209463	-3.727	0.000225	***
x10	-0.007141	0.004875	-1.465	0.143822	

```
Residual standard error: 223.8 on 357 degrees of freedom
Multiple R-squared:  0.5914,    Adjusted R-squared:   0.58
F-statistic: 51.68 on 10 and 357 DF,  p-value: < 2.2e-16
```

可得知上方模型 **p-value<0.05,**模型為顯著但解釋變異為 **58%**則不理想所以會在進行變數選取。

2.變數選取

(一)逐步迴歸

By *Forward Selection*

模型	AIC
$\mathbf{y\sim 1}$	4302.49
$\mathbf{y\sim x_2}$	4078.03
$\mathbf{y\sim x_2 + x_1}$	4052.57
$\mathbf{y\sim x_2 + x_1 + x_8}$	4026.98
$\mathbf{y\sim x_2 + x_1 + x_8 + x_9}$	4016.19
$\mathbf{y\sim x_2 + x_1 + x_8 + x_9 + x_5}$	3992.18
$\mathbf{y\sim x_2 + x_1 + x_8 + x_9 + x_5 + x_3}$	3990.44
$\mathbf{y\sim x_2 + x_1 + x_8 + x_9 + x_5 + x_3 + x_4}$	3989.84

By *Backward Elimination*

模型	AIC
$y\sim x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	3993.09
$y\sim x_1 + x_2 + x_3 + x_4 + x_5 + x_7 + x_8 + x_9 + x_{10}$	3991.23

By *Both*

模型	AIC
$y\sim x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	3993.09
$y\sim x_1 + x_2 + x_3 + x_4 + x_5 + x_7 + x_8 + x_9 + x_{10}$	3991.23

根據上方三種逐步迴歸，

則採用向前選取法的模型、因為三種方法選去出來的模型中以向前前選取法的模型 AIC 達最低則選擇$\mathbf{y\sim x_2 + x_1 + x_8 + x_9 + x_5 + x_3 + x_4}$

去除掉 $x_6 = 結婚對數$ $x_7 = 老人戶數 x_{10} = 總戶數$，其實從 full model 中檢測共線性，可知這個三個變數 VIF 個別為 120.528486、70.446979、185.351506 其實也蠻合理的。

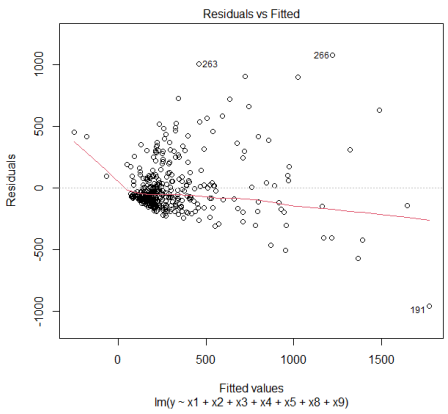
```
      x1       x2       x3       x4       x5       x6       x7       x8       x9
2.464063 85.735877  1.366815 32.759450 11.579150 120.528486 70.446979  2.937558 42.393389
     x10
185.351506
```

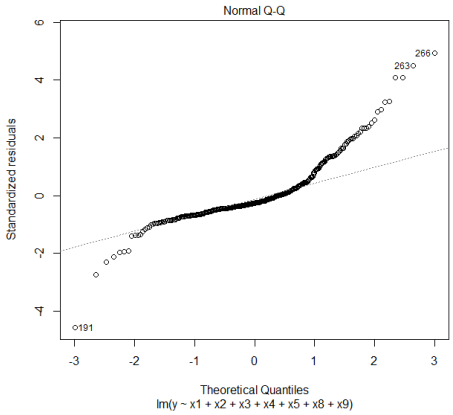
(二)殘差檢定

$\mathbf{y}\sim\mathbf{x}_2+\ \boldsymbol{x}_1+\boldsymbol{x}_8+\boldsymbol{x}_9+\boldsymbol{x}_5+\boldsymbol{x}_3+\boldsymbol{x}_4$以此模型做殘差檢定:

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 548.583503 116.838614   4.695 3.79e-06 ***
x1           -0.448098   0.117656  -3.809 0.000164 ***
x2            0.059561   0.008095   7.358 1.28e-12 ***
x3           -4.812922   2.410376  -1.997 0.046605 *
x4            0.352579   0.220954   1.596 0.111430
x5            2.010257   0.354813   5.666 3.00e-08 ***
x8           -0.079920   0.014034  -5.695 2.57e-08 ***
x9           -0.718923   0.202190  -3.556 0.000427 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 223.7 on 360 degrees of freedom
Multiple R-squared:  0.5884,    Adjusted R-squared:  0.5804
F-statistic: 73.51 on 7 and 360 DF,  p-value: < 2.2e-16
```

在刪減變數後 R suquare 看起來跟 fullmodel 沒甚麼變化，模型也顯著，再來就是看殘差有沒有符合常態。





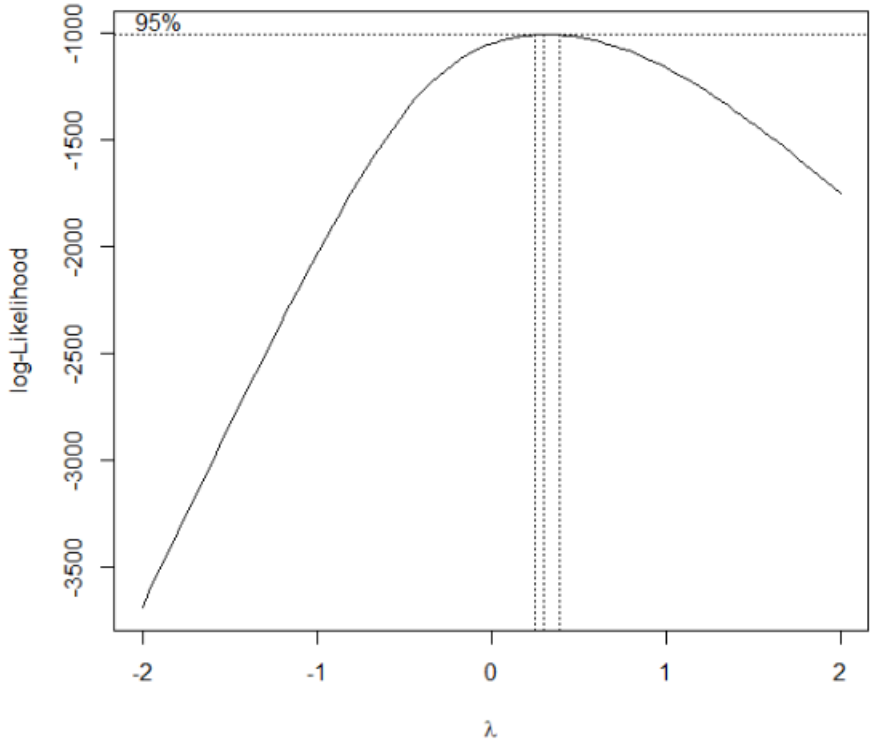
殘差圖呈線性且常態圖左偏且又右偏，且用 shapiro-test<0.05 不符合常態所以先用 BoxCox 對 Y 進行變數變換

```
        Shapiro-Wilk normality test

data:  model$residuals
W = 0.85065, p-value < 2.2e-16
```

Box-cox 轉換:

因為做 boxcox 轉換需 Y 值>0 所以先把 Y 值平移

利用 boxcox 可得 λ=0.3030303



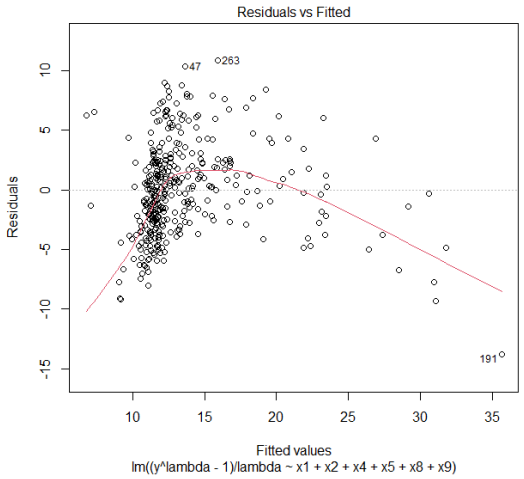
可得模型:$(y^{\lambda}-\mathbf{1})/\lambda \sim \mathbf{x}_2 + x_1 + x_8 + x_9 + x_5 + x_3 + x_4$

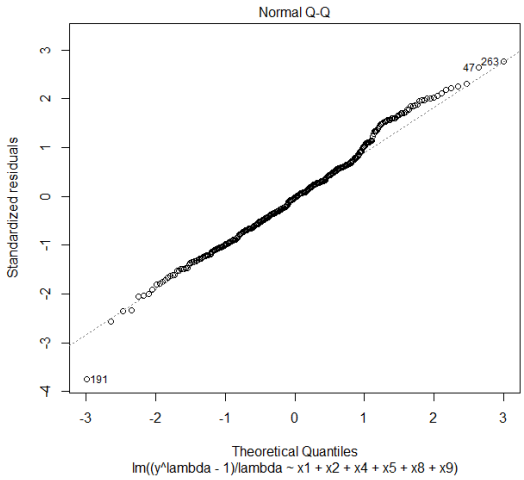
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 18.2340474  2.0055375   9.092  < 2e-16 ***
x1          -0.0127176  0.0021379  -5.949 6.57e-09 ***
x2           0.0013588  0.0001468   9.258  < 2e-16 ***
x3          -0.0384268  0.0402028  -0.956  0.33982
x4           0.0125189  0.0042021   2.979  0.00309 **
x5           0.0303211  0.0066109   4.587 6.29e-06 ***
x8          -0.0014018  0.0003392  -4.133 4.49e-05 ***
x9          -0.0163053  0.0039721  -4.105 5.04e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.648 on 349 degrees of freedom
Multiple R-squared:  0.5713,    Adjusted R-squared:  0.5627
```

因為x_3此變數不顯著所以把它移除掉

可得模型:$(y^{\lambda}-\mathbf{1})/\lambda \sim \boldsymbol{x}_1+\mathbf{x}_2+\boldsymbol{x}_4+\boldsymbol{x}_5+\boldsymbol{x}_8+\boldsymbol{x}_9$

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.6867256  1.1837270  14.097  < 2e-16 ***
x1          -0.0127079  0.0021376  -5.945 6.69e-09 ***
x2           0.0013943  0.0001420   9.821  < 2e-16 ***
x4           0.0123692  0.0041987   2.946  0.00343 **
x5           0.0289587  0.0064546   4.487 9.83e-06 ***
x8          -0.0013386  0.0003327  -4.024 7.02e-05 ***
x9          -0.0161520  0.0039684  -4.070 5.81e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.647 on 350 degrees of freedom
Multiple R-squared:  0.5702,    Adjusted R-squared:  0.5628
F-statistic: 77.39 on 6 and 350 DF,  p-value: < 2.2e-16
```





根據常態性假設檢定及變異數同質性檢定:

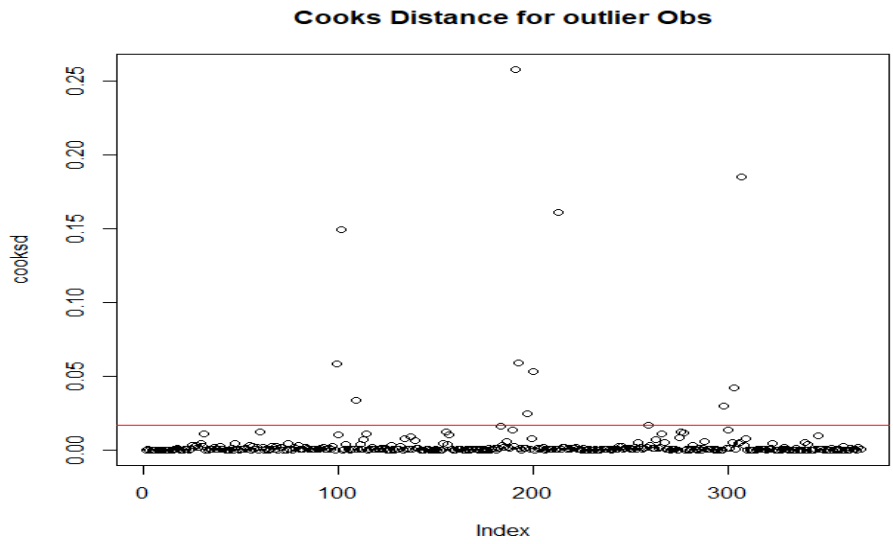
```
        Shapiro-Wilk normality test

data:  model3$residuals
W = 0.9922, p-value = 0.05097
```

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 10.98246, Df = 1, p = 0.00091978
```

但因為變異數同質性檢定 p-value<0.05，所以打算先移除 outlier 看看。

移除 outlier:

By cook's distance



根據 cook's distance 可得知第 99、101、 109、 191、 192、 197 、200、 213 、298、 303、 307 為 outlier。

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 1.01802, Df = 1, p = 0.31299
```

P-value>0.31299 所以符合變異數同質檢定

因此最終模型可得:

$$y = 16.686 - 0.0127\boldsymbol{x}_1 + 0.0013\boldsymbol{x}_2 + 0.0123\boldsymbol{x}_4 + 0.0289\boldsymbol{x}_5 - 0.0013\boldsymbol{x}_8 - 0.0161\boldsymbol{x}_9$$

原本想對 X2 變數進行變數變換，取根號後 R square 有從 56%增加到 66%但因為，再跑一次常態性檢定後就沒過了所以最終決定此模型為最終模型。