迴歸期末報告

M122040017 吳俞憲

期末目標

WORDART

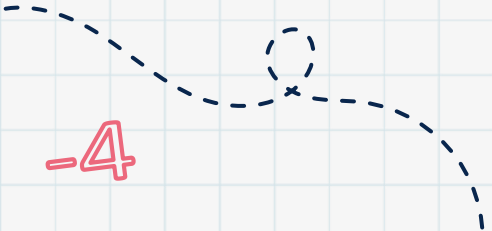
1. 增加模型的解釋力

2. 處理模型沒有滿足假設的問題

Table of contents

01

回顧期中模型



資料蒐集

蒐集南部地區:嘉義市、嘉義縣、台南市、高雄市、屏東縣,這五個縣市2010-2020這十年間的資料,而只找了4個變數為解釋變數來做線性模型。

就業者之年齡別結構-25-44歲(%)　(ER)

犯罪人口率(人/十萬人)　(CR)

青壯年人口比率(15-64歲)(%)　(YAR)

低收入戶人口數占總人口比率(%)　(LOWR)

期中最終模型

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -31.50904    7.10517  -4.435 4.93e-05 ***
ER            0.11897    0.03097   3.841 0.000340 ***
YAR           0.45667    0.11051   4.132 0.000134 ***
LOWR         -0.75284    0.24622  -3.058 0.003549 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6803 on 51 degrees of freedom
Multiple R-squared:  0.5815,    Adjusted R-squared:  0.5569
F-statistic: 23.62 on 3 and 51 DF,  p-value: 1.006e-09
```

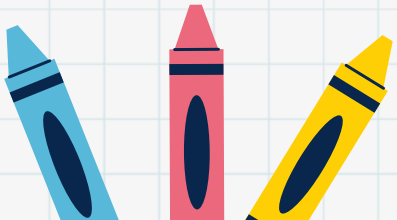
ER:就業者之年齡別結構-25-44歲(%)

YAR:青壯年人口比率(15-64歲)(%)

LOWR: 低收入戶人口數占總人口比率(%)

02

篩選變數與考慮非線性關係



變數篩選

利用原本期中選出的三個變數加上額外加上八個變數來做篩選,且資料從2000取到2022年總共有105筆資料。

Forward selection

```
Step:  AIC=35.55
CBR ~ AR + YAR + DM + UR

       Df Sum of Sq    RSS    AIC
<none>              118.03 35.554
+ CAR   1    5.0497 112.98 35.617
+ LOWR  1    3.0605 114.97 37.450
+ ER    1    1.9198 116.11 38.486
+ DI    1    0.3693 117.66 39.879
+ LT    1    0.2189 117.81 40.013
+ UV    1    0.2183 117.81 40.014
+ FD    1    0.0017 118.03 40.207
```

變數篩選

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 50.45207    4.83994  10.424  < 2e-16 ***
AR          -0.82237    0.06526 -12.601  < 2e-16 ***
YAR         -0.27625    0.06109  -4.522 1.69e-05 ***
DM          -0.10046    0.03013  -3.334  0.00120 **
UR          -0.52191    0.16999  -3.070  0.00275 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.086 on 100 degrees of freedom
Multiple R-squared:  0.7324,    Adjusted R-squared:  0.7216
F-statistic: 68.41 on 4 and 100 DF,  p-value: < 2.2e-16
```

AR : 平均每人居住面積(坪)

YAR : 青壯年人口比率(15-64歲)(%)（上一個模型有的變數)

DM : 家庭收支-平均消費傾向(%)

UR ：失業率(%)

查看共線性

VIF

AR	YAR	DM	UR
1.224343	1.096531	1.049637	1.078050

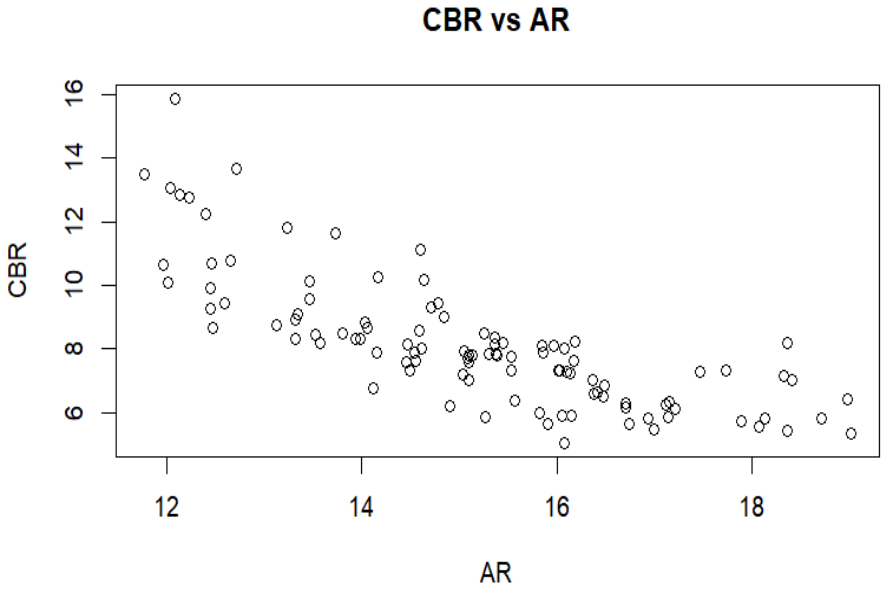
AR : 平均每人居住面積(坪)

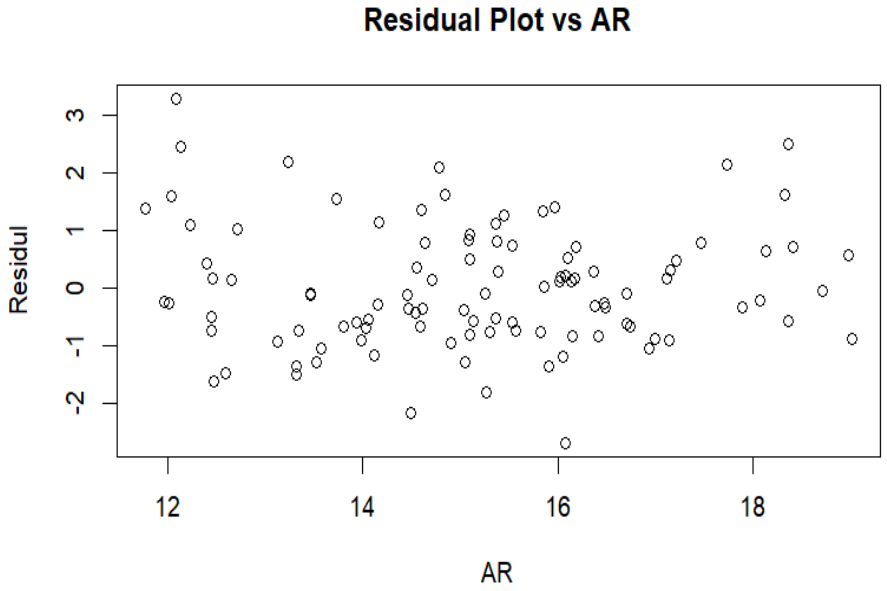
YAR : 青壯年人口比率(15-64歲)(%)（上一個模型有的變數)

DM : 家庭收支-平均消費傾向(%)

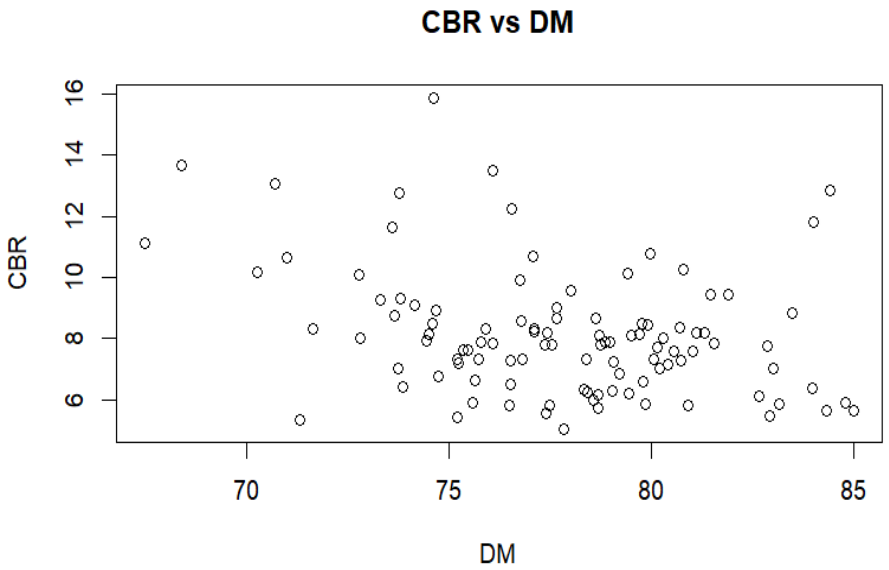
UR ：失業率(%)

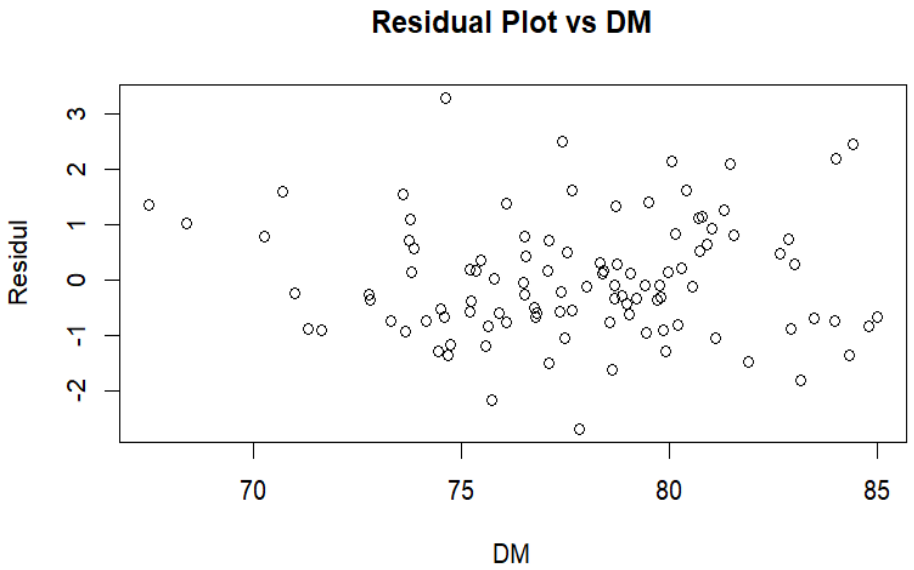
查看非線性關係(AR)


CBR vs AR
CBR
AR
6
8
10
12
14
16
12
14
16
18

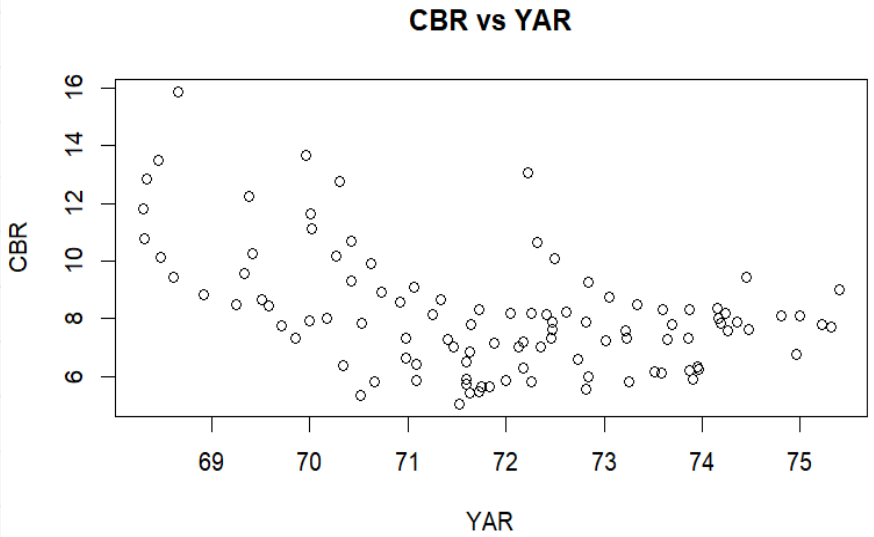

Residual Plot vs AR
Residul
AR
-2
-1
0
1
2
3
12
14
16
18

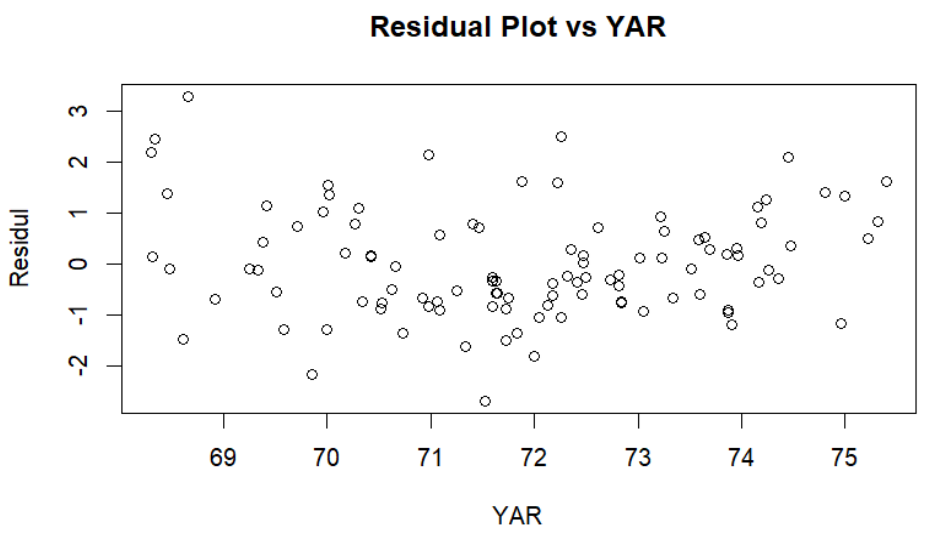
查看非線性關係(DM)


CBR vs DM
CBR
DM
6
8
10
12
14
16
70
75
80
85

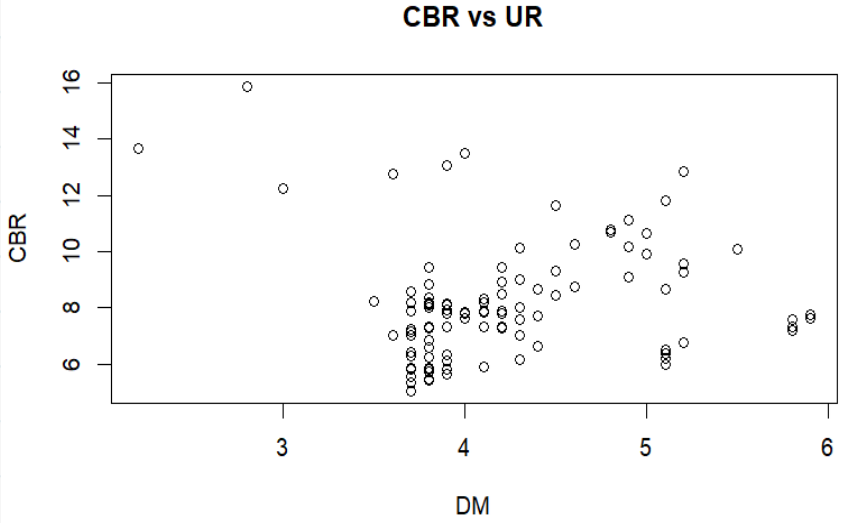

Residual Plot vs DM
Residul
DM
-2
-1
0
1
2
3
70
75
80
85

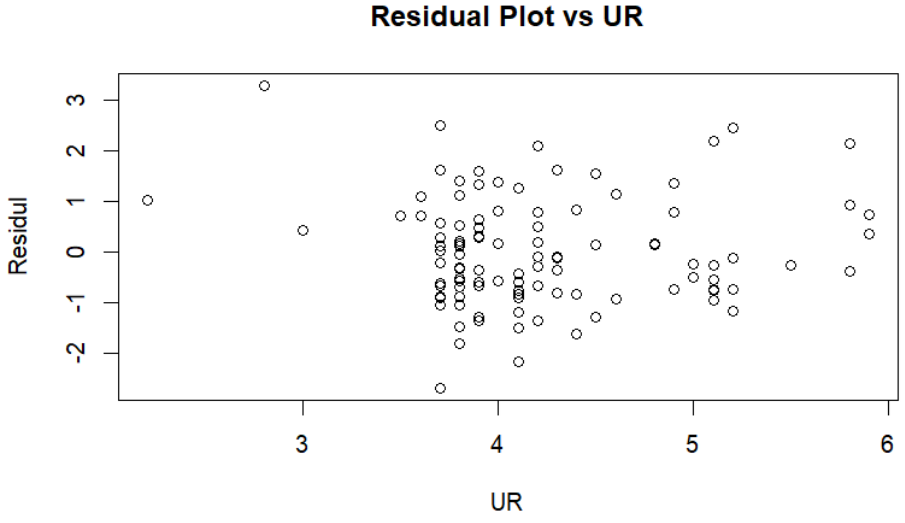
查看非線性關係(YAR)


CBR vs YAR
CBR
6
8
10
12
14
16
69
70
71
72
73
74
75
YAR


Residual Plot vs YAR
Residul
-2
-1
0
1
2
3
69
70
71
72
73
74
75
YAR

查看非線性關係(UR)


CBR vs UR
CBR
DM
6
8
10
12
14
16
3
4
5
6


Residual Plot vs UR
Residul
UR
-2
-1
0
1
2
3
3
4
5
6

增加平方項

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 794.88215  147.33290   5.395 4.76e-07 ***
AR           -0.68657    0.06210 -11.056  < 2e-16 ***
YAR         -20.78844    4.09469  -5.077 1.83e-06 ***
UR           -4.52408    1.25647  -3.601  0.00050 ***
DM           -0.12438    0.02741  -4.537 1.62e-05 ***
I(YAR^2)      0.14282    0.02847   5.017 2.35e-06 ***
I(UR^2)       0.45421    0.14168   3.206  0.00182 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9386 on 98 degrees of freedom
Multiple R-squared:  0.8042,    Adjusted R-squared:  0.7922
F-statistic:  67.1 on 6 and 98 DF,  p-value: < 2.2e-16
```

R-squared從原本的0.7216上升到0.7922

增加交互向項

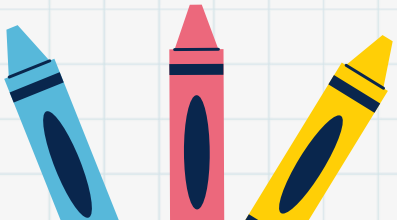
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 699.40684  142.77771   4.899 3.86e-06 ***
AR           -0.72466    0.06006 -12.066  < 2e-16 ***
YAR         -19.26314    3.91635  -4.919 3.56e-06 ***
UR           15.12642    5.90301   2.562 0.011930 *
DM           -0.14498    0.02674  -5.421 4.33e-07 ***
I(YAR^2)      0.14154    0.02705   5.232 9.68e-07 ***
I(UR^2)       0.69521    0.15214   4.569 1.44e-05 ***
YAR:UR       -0.30549    0.08987  -3.399 0.000982 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8918 on 97 degrees of freedom
Multiple R-squared:  0.8251,    Adjusted R-squared:  0.8124
F-statistic: 65.36 on 7 and 97 DF,  p-value: < 2.2e-16
```

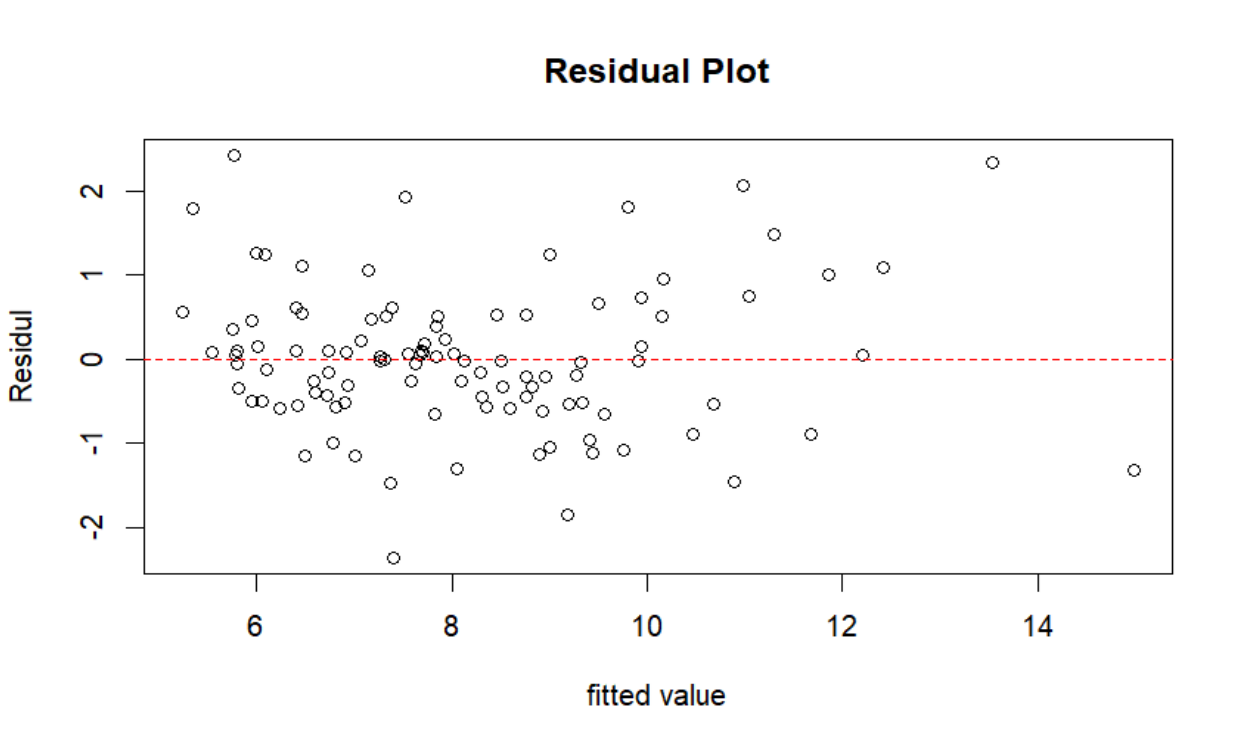
R-squared從0.7922上升到0.8124

03

檢查模型假設

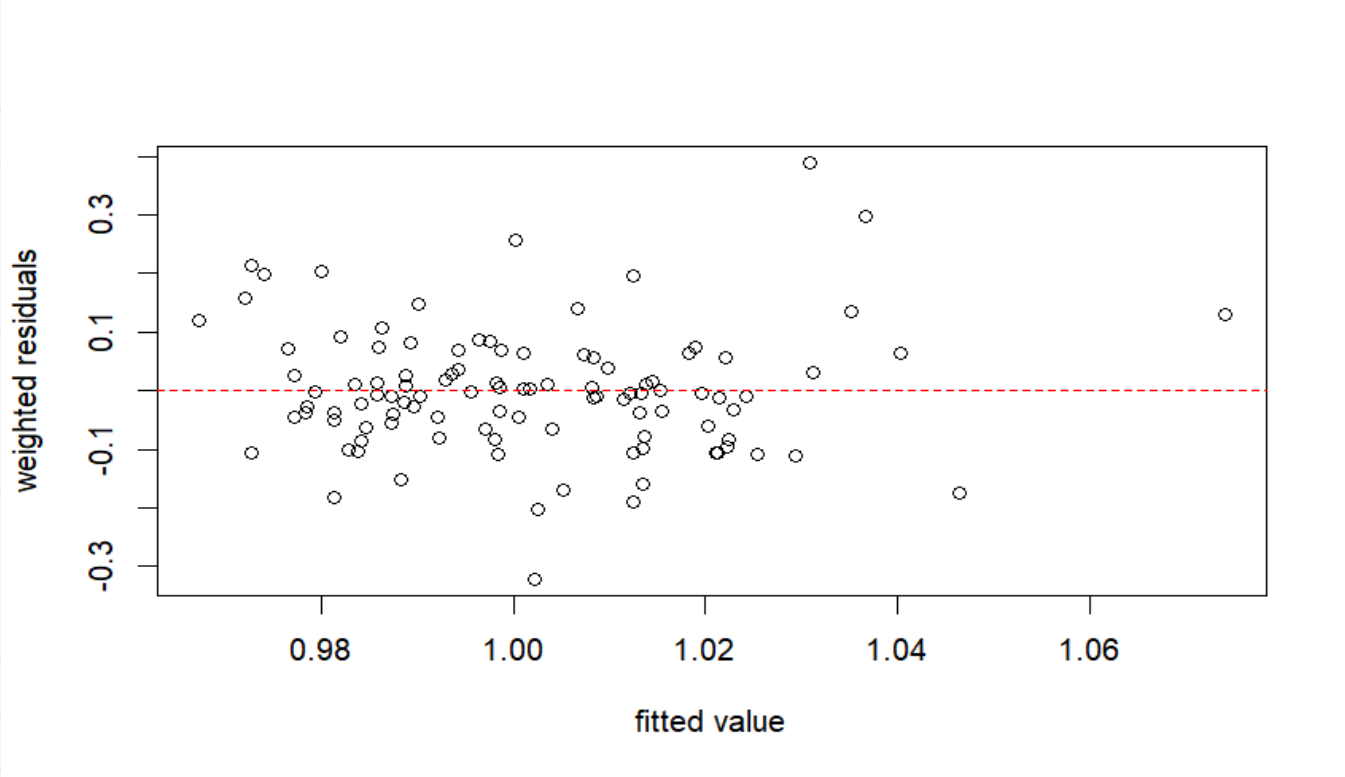


查看方差是否等於常數



方法一:WLSE

Weight = 1 /(ŷ)^2



方法一:WLSE

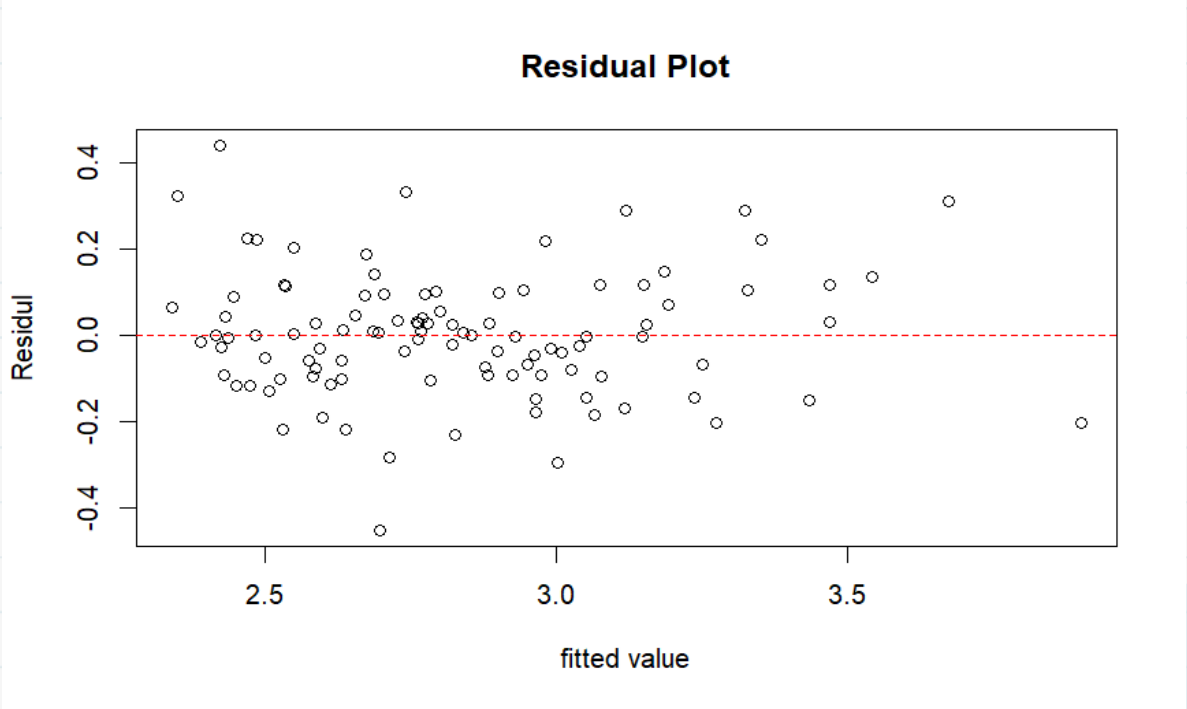
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 667.45693  155.86693   4.282 4.35e-05 ***
AR           -0.66347    0.05796 -11.446  < 2e-16 ***
YAR         -18.53346    4.26384  -4.347 3.41e-05 ***
UR           16.61237    6.27867   2.646 0.009508 **
DM           -0.13387    0.02529  -5.293 7.48e-07 ***
I(YAR^2)      0.13754    0.02932   4.690 8.94e-06 ***
I(UR^2)       0.79432    0.18877   4.208 5.76e-05 ***
YAR:UR       -0.33747    0.09373  -3.600 0.000503 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1114 on 97 degrees of freedom
Multiple R-squared:  0.7629,    Adjusted R-squared:  0.7458
F-statistic: 44.58 on 7 and 97 DF,  p-value: < 2.2e-16
```

可以發現雖然問題解決了,但R-squared下降到了0.7458

方法二:variance -stabilizing

y’= y^1/2



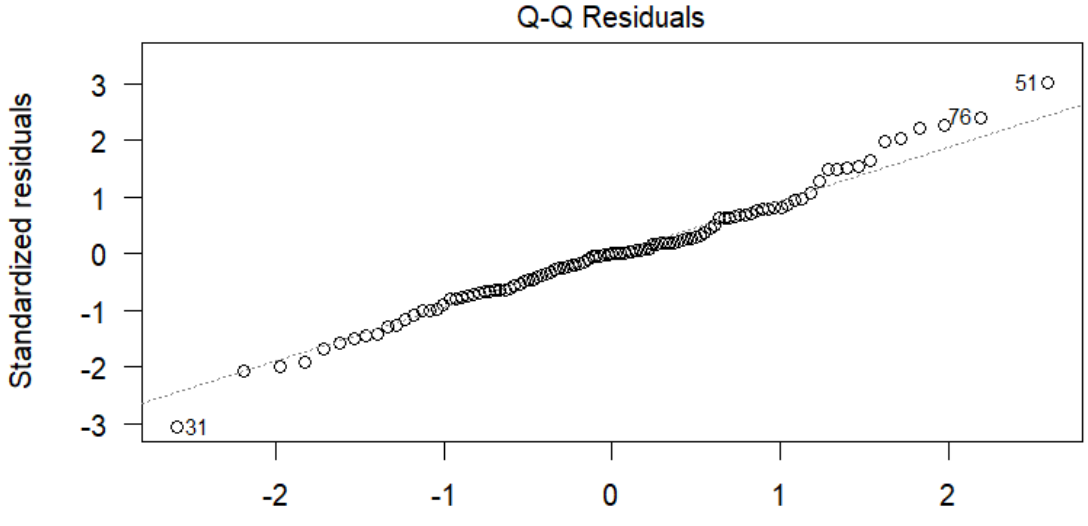
方法二:variance -stabilizing

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 114.771361  24.028006   4.777 6.33e-06 ***
AR           -0.124979   0.010107 -12.365  < 2e-16 ***
YAR          -3.147283   0.659080  -4.775 6.36e-06 ***
UR            2.966780   0.993415   2.986 0.003573 **
DM           -0.024685   0.004501  -5.485 3.29e-07 ***
I(YAR^2)      0.023322   0.004553   5.123 1.53e-06 ***
I(UR^2)       0.106763   0.025604   4.170 6.64e-05 ***
YAR:UR       -0.055857   0.015125  -3.693 0.000366 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1501 on 97 degrees of freedom
Multiple R-squared:  0.8233,    Adjusted R-squared:  0.8106
F-statistic: 64.58 on 7 and 97 DF,  p-value: < 2.2e-16
```

問題有解決,且R-squared下降不明顯

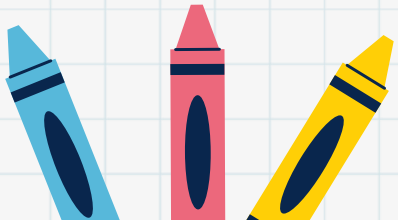
方法二:variance -stabilizing



模型符合常態分佈

04

結論



最終模型

```
Call:
lm(formula = sqrt(CBR) ~ AR + YAR + UR + DM + I(YAR^2) + I(UR^2) +
    YAR:UR, data = data)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 114.771361  24.028006   4.777 6.33e-06 ***
AR           -0.124979   0.010107 -12.365  < 2e-16 ***
YAR          -3.147283   0.659080  -4.775 6.36e-06 ***
UR            2.966780   0.993415   2.986 0.003573 **
DM           -0.024685   0.004501  -5.485 3.29e-07 ***
I(YAR^2)      0.023322   0.004553   5.123 1.53e-06 ***
I(UR^2)       0.106763   0.025604   4.170 6.64e-05 ***
YAR:UR       -0.055857   0.015125  -3.693 0.000366 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1501 on 97 degrees of freedom
Multiple R-squared:  0.8233,    Adjusted R-squared:  0.8106
F-statistic: 64.58 on 7 and 97 DF,  p-value: < 2.2e-16
```

AR : 平均每人居住面積(坪)

YAR : 青壯年人口比率(15-64歲)(%)（上一個模型有的變數)

DM : 家庭收支-平均消費傾向(%)

UR ：失業率(%)

結論

1.透過變數篩選與增加非線性關係確實對模型解釋力有很大的提升

2.比較兩種方法來處理方差不等於常數的假設,WLSE的R-squared下降的比較明顯,所以最終模型選擇做variance-stabilizing的

3.最終模型的R-squared為0.8106,代表模型具有很高的解釋力,有很好的達成期末的目標。