目錄

一.期末目標

1.提升模型解釋力

2.處理模型沒有滿足假設的問題

因為期中的模型做出來的解釋力很差,所以這次的目標會放在提升模型的 R-squared,且要去解決模型沒有滿足重要假設的問題。

二.回顧期中模型

當時資料蒐集是蒐集南部地區包含:嘉義市、嘉義縣、台南市、高雄市、屏東縣。而模型是沒有透過變數篩選的

```
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(%)

可以發現期中最終模型的 R-squared 只有 0.5569，是一個不太好的結果，可能是因為變數太少且沒有特過變數篩選的原因。

三.變數篩選及查看線性關係

1.變數篩選

想要增加模型解釋力最簡單的方法就是增加變數，那如果想增加變數的話資料就不能太少，

.所以我這次一樣取了 5 個縣市，但年分從 2000 年取到 2020 年總共有 105 筆資料，那變數是選原本期中的三個變數再額外增加八個變數來做變數篩選，篩選的方法為 forward selection 標準為 AIC，以下是我取到最好的模型。

```
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(%)

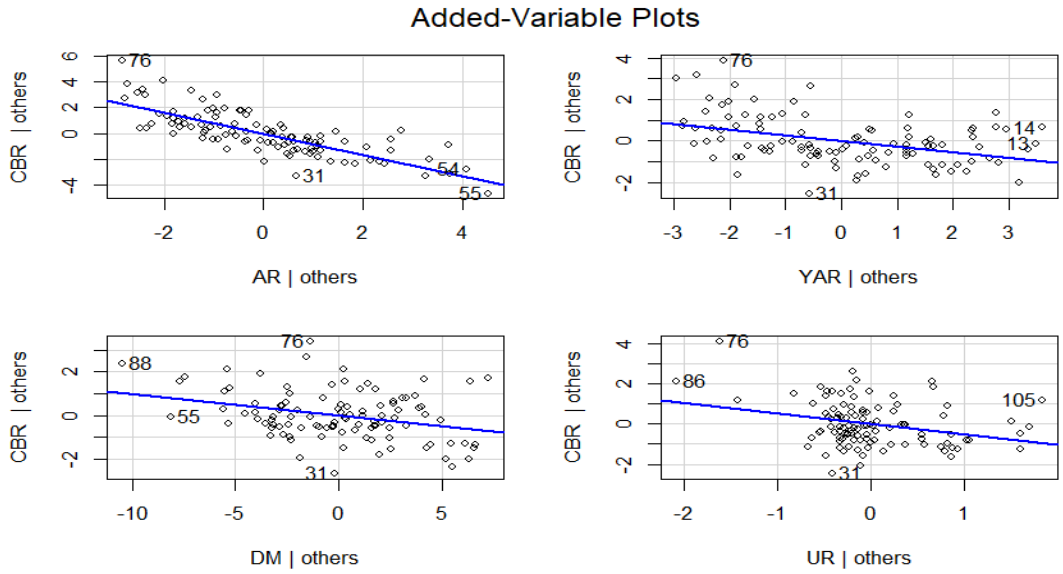
UR : 失業率(%)

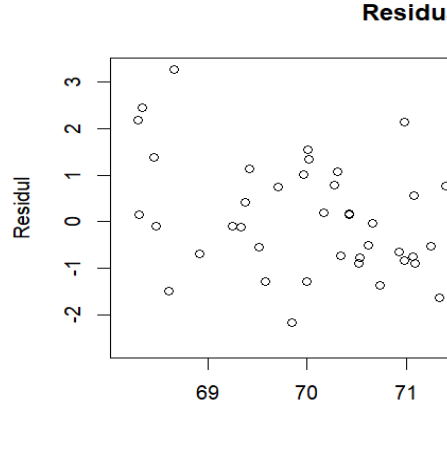
VIF 值:

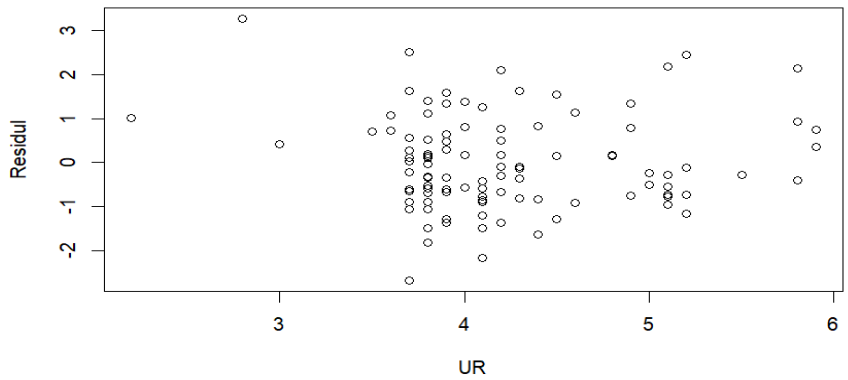
```
       AR       YAR        DM        UR
1.224343 1.096531 1.049637 1.078050
```

可以看到 R-squared 為 07216 已經有一個還不錯的解釋力了，且變數都沒有共線性問題那跟期中的模型重複的變數只有一個，那這也可以說明為什麼期中模型會比較差的原因。

2.查詢非線性關係


Added-Variable Plots
CBR | others
AR | others
YAR | others
DM | others
UR | others


Residual Plot vs YAR
Residul
YAR


Residual Plot vs UR
Residul
UR

從圖中可以看出 YAR 跟跟 UR 都有呈現非線性的關係，所以我有嘗試過增加這兩個變數的平方項、立方項、根號跟 log 但發現只有平方項對模型有顯著的影響，所以最終只加了平方項。以下是增加之後的模型

```
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
```

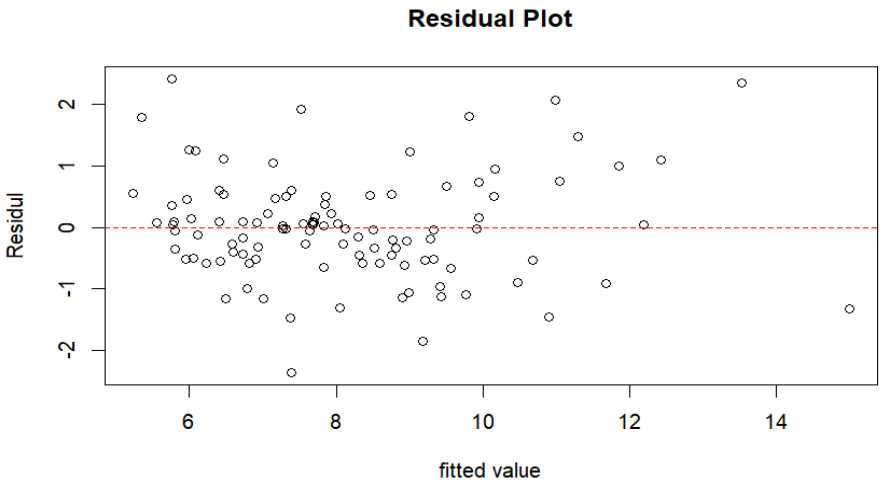
可以看到加入兩個變數的平方項 R-squared 上升的很明顯

接著在增加交互項，那因為有沒有交互項的關係很難從圖片看出，所以我這邊就兩個變數的交互項慢慢去嘗試，然後加入對模型提升最多的交互項，那這邊是取 YAR 跟 UR 的交互項。以下是增加之後的模型

```
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
```

可以看到全部變數都是顯著的，且 R-squared 到達了 0.8124，是一個很好的結果了。

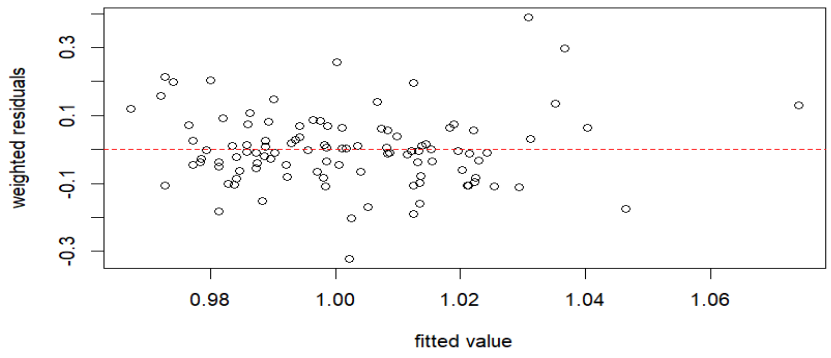
四.檢查模型假設



從 residual plot 圖來看有一個麥克風形狀的趨勢，所以我會認為有違反方差等於常數這個假設，那我會試著用兩種方法來處理這個問題。

方法一.WLSE

Weight = 1 / $\hat{y}^2$

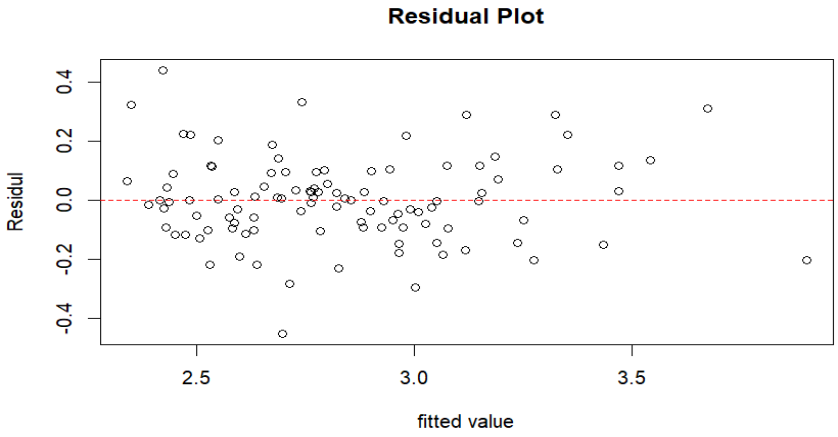



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 667.45693  155.86693   4.282 4.35e-05 ***
AR           -0.66347    0.05796 -11.446  < 2e-16 ***
YAR         -18.53346    4.26384  -4.347 3.41e-05 ***
UR           16.61237    6.27867   2.646 0.009508 **
DM           -0.13387    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
```

從圖片可以看出異方差這個問題解決了，但是 R-squared 下降到 0.7458

方法二: variance -stabilizing

$$y' = \sqrt{y}$$



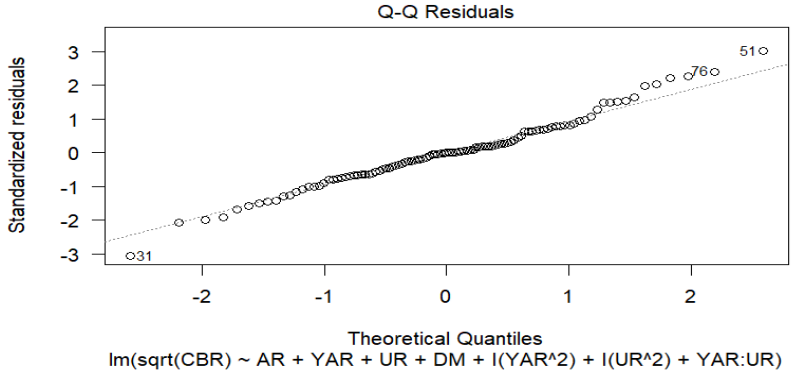

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 114.771361  24.028006   4.777 6.33e-06 ***
AR           -0.124979   0.010107 -12.365  < 2e-16 ***
YAR          -3.147283   0.659080  -4.775 6.36e-06 ***
UR            2.966780   0.993415   2.986 0.003573 **
DM           -0.024685   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
```

可以發現用方差穩定這個方法可以解決問題，且 R-squared 幾乎沒有下降。

所以最後會選擇這個方法來處理我的模型。

查看誤差是否為常態分配

這邊是利用方差穩定的方法來看的，QQ plot 的圖形有符合常態分配。



五.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)(%)

DM : 家庭收支-平均消費傾向(%)

UR : 失業率(%)

1. 有好的變數篩選與增加變數非線性關係對模型解釋力是有明顯的提升的
2. 比較兩個處理異方差的方法，方差穩定對這筆資料的幫助是比較好的，因為既解決了問題 R-squared 也幾乎沒有下降。
3. 最終的模型的 R-squared 為 0.8106，代表模型有很高的解釋力，且模型假設也都有符合，很好的達成我們期末的目標。