一.蒐集資料

從老師給的網站蒐集了 2010 到 2020 年這十年之間南部地區的資料，只蒐集十年的原因是因為再跟早以前的數據跟現在會有差異的；而地區的選擇我就直接選一個大區域，我會覺得南部地區的生活可能會比較相近，可以看成是大生活區，而南部地區包含:嘉義市、嘉義縣、台南市、高雄市、屏東縣，所以總共有 55 筆資料。

二.變數選擇

- 粗出生率(‰) (CBR)
- 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)

除了粗出生率是老師指定的以外，其餘變數都是選擇我自己認為可能跟其有相關且自己比較有興趣的，這邊我就沒有特別用選擇變數的方法來看，純粹想了解自己對於變數的敏感度。所以就可以做初步的迴歸模型，也就是粗出生率為反映變數，其他變數為解釋變數的線性模型，然後去觀察這個模型所給的報表。

程式碼:

```
library(readxl)

data <- read_excel("D:/dataset/報告用.xlsx")
```

fm <-lm(CBR~ER+CR+YAR+LOWR, data)

summary(fm)

```
Call:
lm(formula = CBR ~ ER + CR + YAR + LOWR, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.95914 -0.44194  0.07901  0.42903  1.53026

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.759e+01  7.922e+00  -3.483 0.001040 **
ER           1.188e-01  3.090e-02   3.845 0.000342 ***
CR          -6.842e-04  6.174e-04  -1.108 0.273096
YAR          4.145e-01  1.167e-01   3.553 0.000842 ***
LOWR        -7.377e-01  2.461e-01  -2.998 0.004226 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6788 on 50 degrees of freedom
Multiple R-squared:  0.5916,    Adjusted R-squared:  0.5589
F-statistic:  18.1 on 4 and 50 DF,  p-value: 2.998e-09
```

可以從報表看出犯罪人口率(CR)的 p-value 為 0.273096 是很高的，這邊在統計上的意思就是這個變數是與反應變數是無關的，所以會考慮把他從模型拿掉。所以我就再建立一個沒有這個變數的模型，一樣去觀察報表

程式碼:

fm1 <-lm(CBR~ER+YAR+LOWR, data)

summary(fm1)

```
Call:
lm(formula = CBR ~ ER + YAR + LOWR, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-1.8332 -0.4540  0.1351  0.4537  1.6428

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
```

比較這兩個模型，我們可以發現兩個模型的 R-squared 其實是差不多的，這個意思就是就算拿掉 CR 這個變數，模型的解釋力也不會下降，所以後續我就會使用沒有 CR 這個變數的模型來做診斷。再做模型的診斷之前，我們必

須先了解我們的變數之間是否有共線性。

```r
library(car)
fm1 <-lm(CBR~ER+YAR+LOWR, data)
vif(fm1)
```

```
      ER      YAR     LOWR
1.521229 2.612805 1.925871
```

從程式可以看出 output 的部分 VIF 值都蠻小的，所以可以斷定變數之間是沒有共線性的。

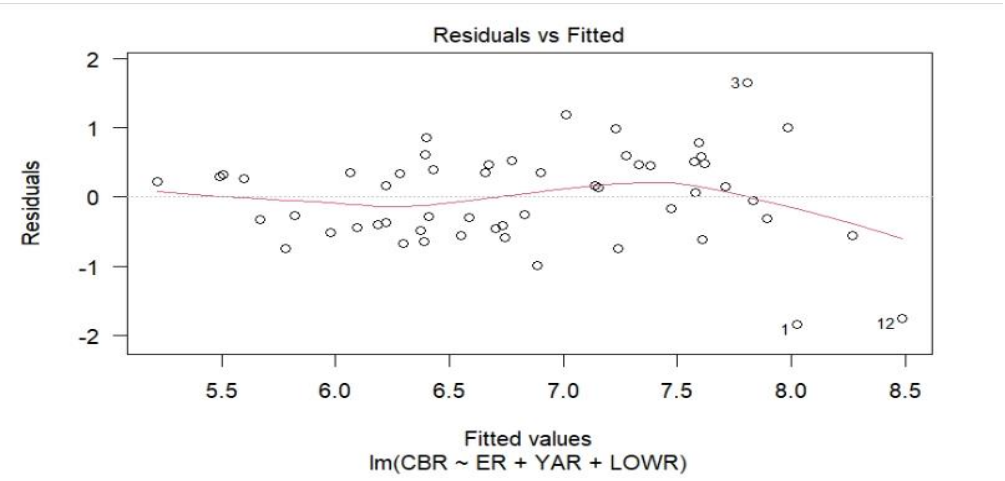
三.Diagnostics

選擇好模型之後就可以開始做診斷，想了解是否有滿足線性模型的假設

那我這邊會做三個部分的診斷，然後都是利用圖形去判斷:

我們利用程式:plot(fm1,las=1)可以獲得 residual plot 和 Q-Q plot

(1)判斷誤差是否等於常數:



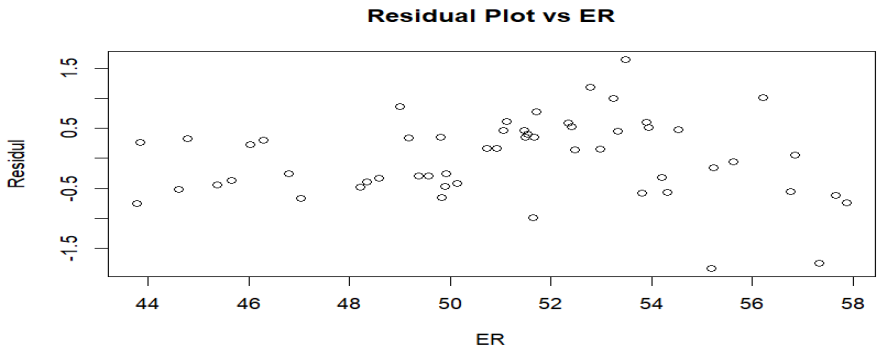
從圖形可以看出，大部分的點都落在兩個水平線之間，而且分布蠻均勻的，分散性也沒有太大問題，所以我會認定沒有違背方差是常數的假設。

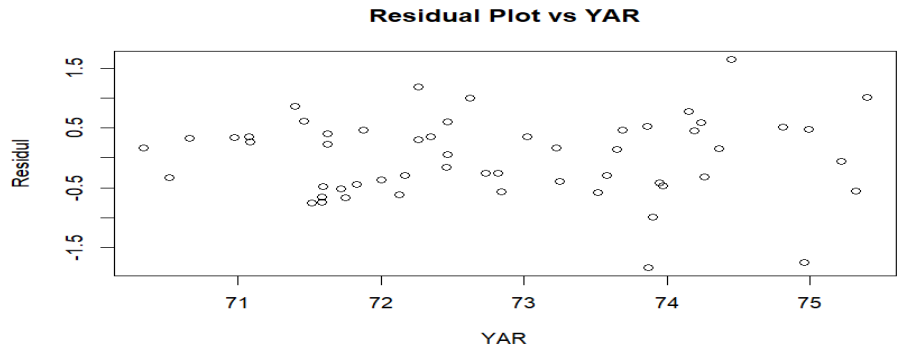
(2)變數是否與反應變數有非線性關係

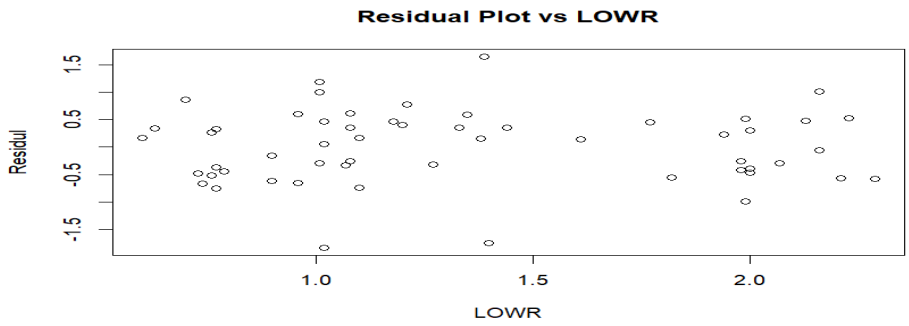
程式:

plot(data$ER,resid(fm1),xlab="ER", ylab="Residul", main="Residual Plot vs ER")

然後變數有 ER、YAR、LOWR



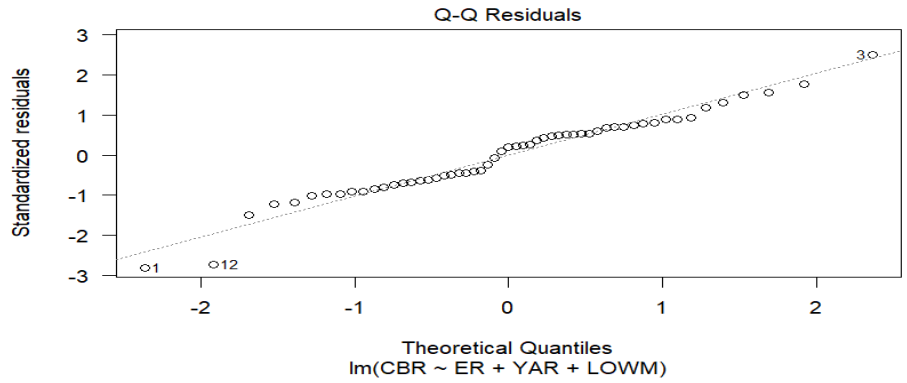




這三張圖，大部分的點也都是落在兩個水平線之間，分佈沒有特別奇怪，所以我會判斷這三個變數都沒有非線性的關係，都蠻符合線性模型的。

(3)誤差是否為常態分佈:

這邊是利用上面的程式獲得的 Q-Q Plot 來分析



我們從圖中可以看到大部分的點都是落在直線上，而且兩端也沒什麼偏離，所以我會判斷誤差是有滿足常態分佈的。

四.結論

從上面的分析來看，我的模型是沒有什麼太大的問題，所以我們暫且可以相信這個模型的報表是可信的，我們可以從中得出:

```
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 **
```

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(%)

出來的係數的相關性(正或負)跟自己所想的也是一樣的。

但即使是這樣，因為我選的樣本數不是很多，而且 R-squared 也不是很高，所以如果以後要用這個模型做分析的話，我們不能把結論下的太死。