應用迴歸期末報告

姓名:趙立騰

系級:統計碩一

學號:110354017

一、變數介紹

本次報告選擇年度的死亡數當作應變數，以出生率、醫療院所家數、碩士人口數、性比率、單獨生活戶數等等作為解釋變數。

空間範圍:全國

空間統計單元:鄉鎮市區

樣本數:368

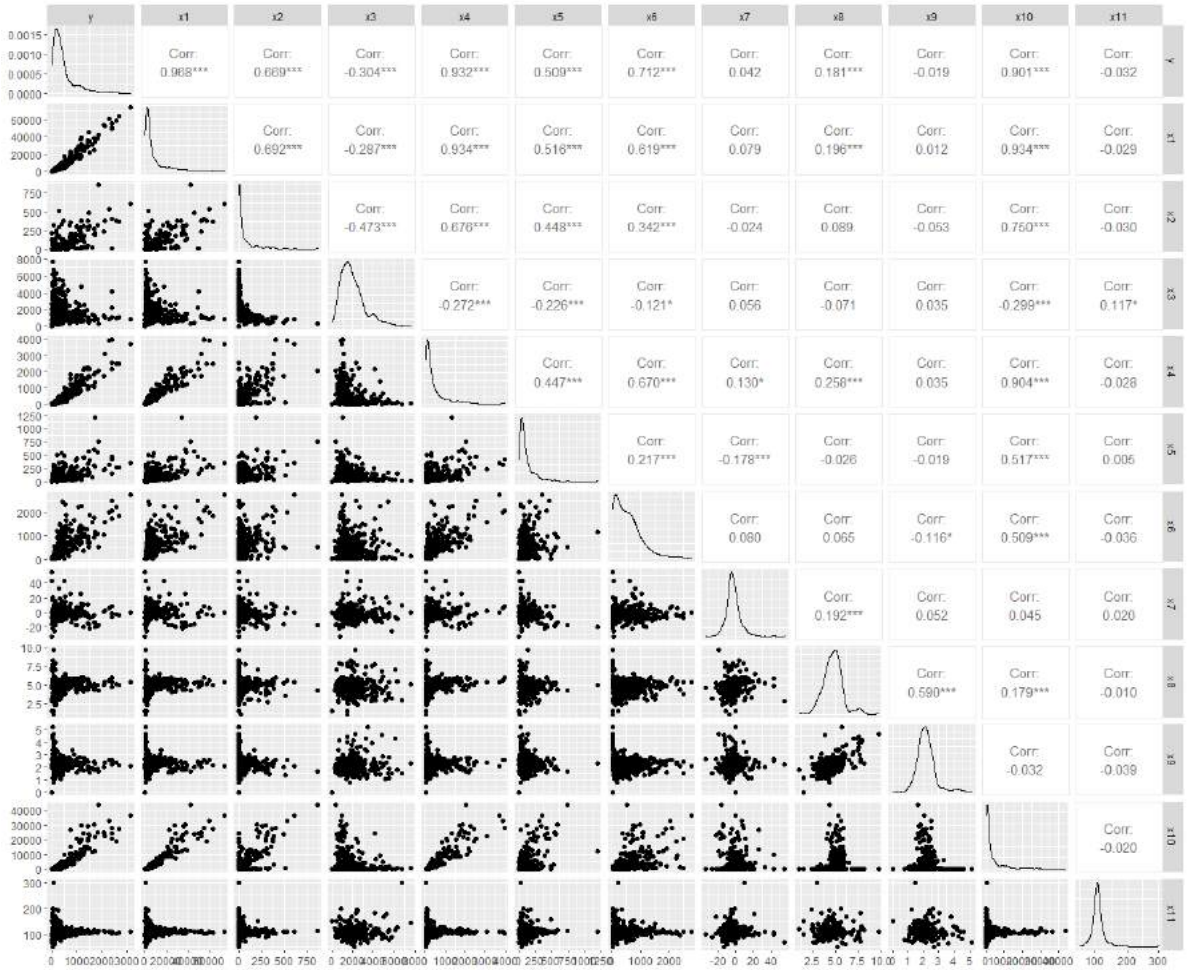
	變數名稱
Y	死亡數
X1	單獨生活戶數
X2	醫療院所家數
X3	醫療院所平均每家服務人數
X4	出生數
X5	需關懷之獨居老人人數
X6	不識字人口數
X7	社會增加率
X8	粗結婚率
X9	粗離婚率
X10	碩士人口數
X11	性比率

從這些資料自己是認為，死亡數應該與單獨生活戶數與出生數和需關懷之獨居老人人數呈正相關，與碩士人口數和結婚率等等就比較沒有那麼明顯的相關性，以下先對資料做一些初步的了解。

```
> summary(data2)
```

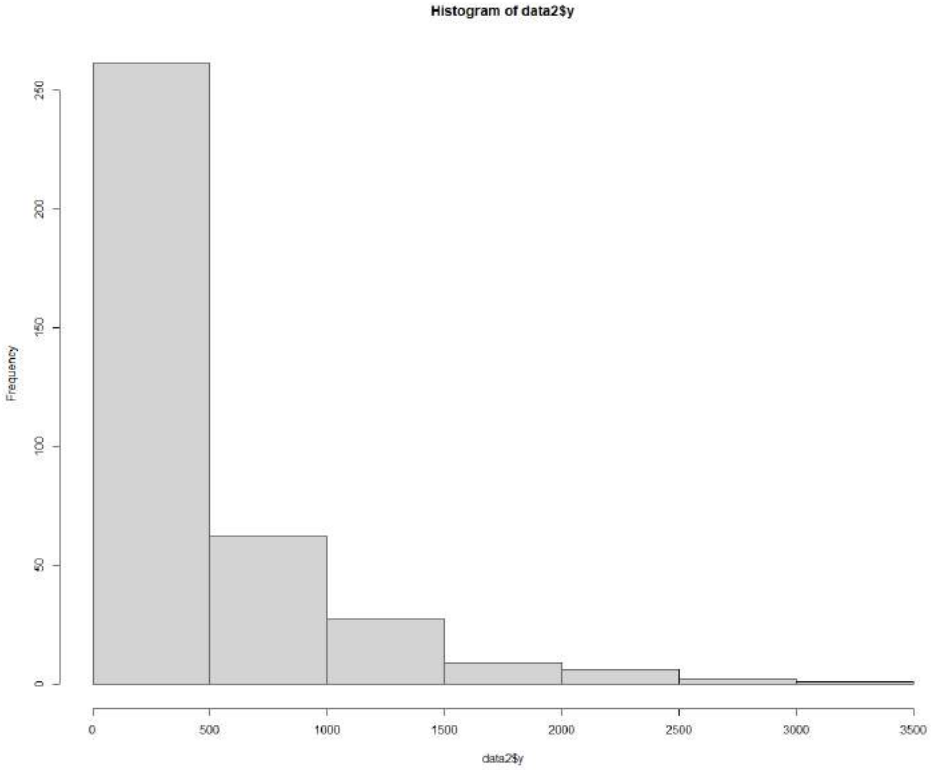
	Y	X1	X2	X3
Min	2	18	0	0
1st Qu.	157	1800	5.75	1136
Median	311.5	3540	16	1730
mean	470.5	8266	62.65	2018
3rd Qu.	547	9126	66.5	2518
Max	3142	74118	858	7638
	X4	X5	X6	X7
Min	5	1	2	-33.64
1st Qu.	83.75	39	139.5	-7.265
Median	196.5	74	433	-3.85
mean	449.05	114.1	545.2	-2.6519
3rd Qu.	539.75	134.2	764.2	0.5075
Max	3966	1216	2768	55.26
	X8	X9	X10	X11

Min	1.08	0	37	50
1st Qu.	4.098	1.83	370.8	98.64
Median	4.77	2.12	1128	106.91
mean	4.741	2.164	3938.8	109.04
3rd Qu.	5.322	2.422	3961	115.5
Max	9.69	5.15	43719	300

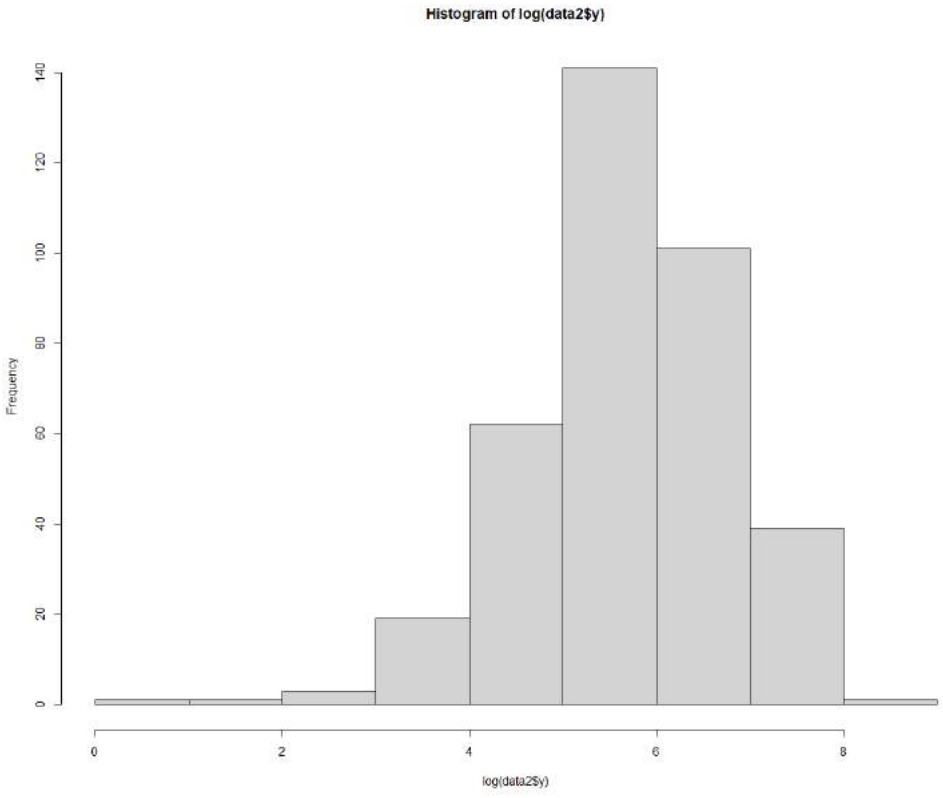


從上圖可以看出應變數與死亡數、醫療院所家數、出生數和不識字人口數都有蠻高的相關性，跟我們的猜想蠻接近的，反而是獨居老人的人數與應變數的相關性沒有我們所想的那麼明顯。

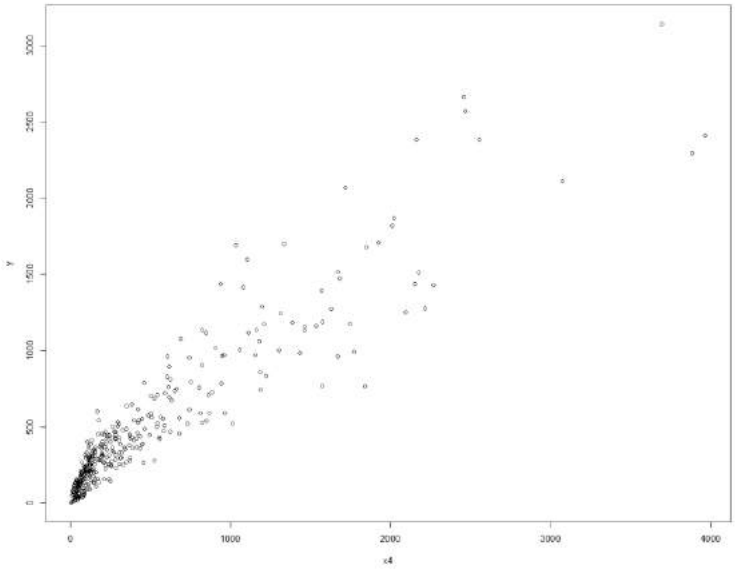
再來我想分別對一些變數再去觀察他們之間的關係。



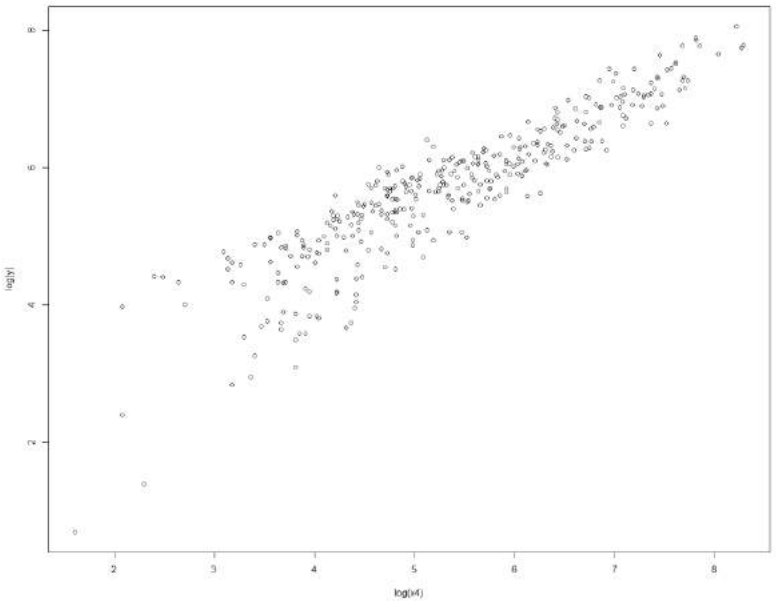
首先看到 y 的直方圖，是一組明顯右偏的資料，為了讓他的資料看起來對稱一點我想對他取 log。



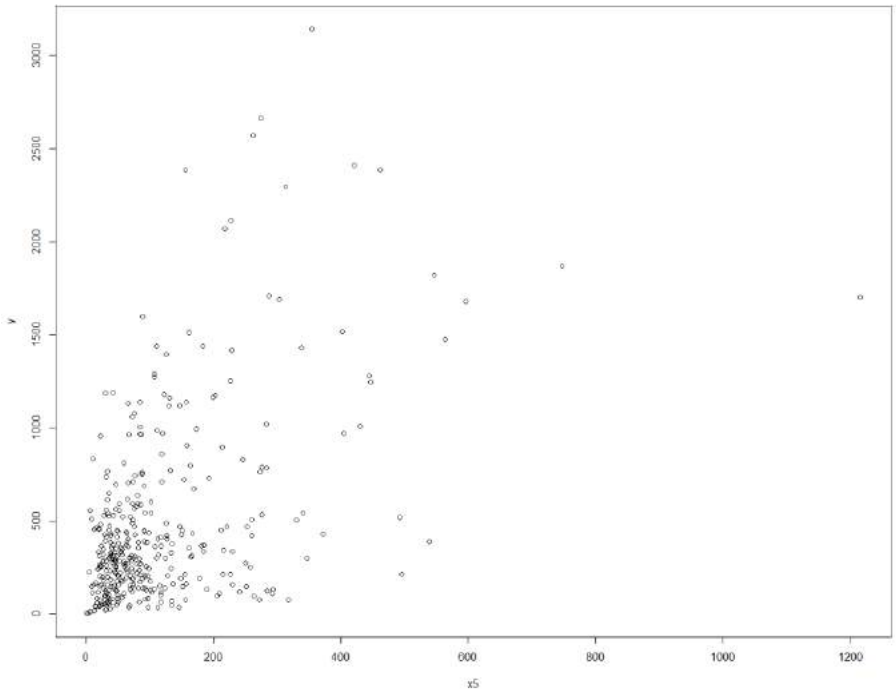
從 ggpairs 可以看出應變數與出生數高度相關



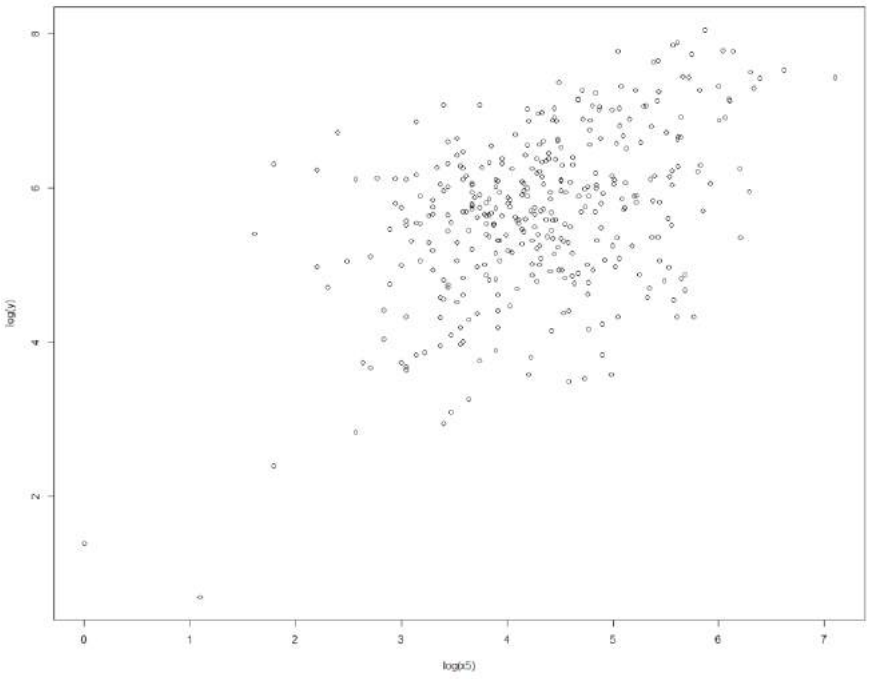
但因兩資料都是極右偏分布，所以資料幾乎都集中在左下部分，我打算對死亡數及出生數都取 log 去觀察看看。



經過變數變換後的確擁有了更好的線性關係及分布狀況。



死亡數及需關懷之獨居老人人數也有相同的問題，所以我對他做相同的變數變換。



也與之前得到相同的結論。

二、建立模型

1.共線性

一開始我先將所有解釋變數都丟入模型內來觀察

```
r1 <- lm(y~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11,dat
a = data2)
```

但考慮到變數間可能會有共線性的問題而導致 R square 異常，因此

使用 VIF 來檢測模型中是否有共線性的問題。

	X1	X2	X3	X4	X5
	13.033360	2.774078	1.338742	11.387408	1.541771
X6	X7	X8	X9	X10	X11
2.289669	1.140789	1.799796	1.683180	11.502472	1.020327

因此，先將 VIF 最大的 X1 拿掉，使得模型變成

```
r0 <- lm(y~x2+x3+x4+x5+x6+x7+x8+x9+x10+x11,data =
 data2)
```

接下來再做一次相同的檢測

		X2	X3	X4	X5
		2.766473	1.335372	9.411462	1.492767
X6	X7	X8	X9	X10	X11
2.181905	1.138809	1.788866	1.660065	7.942891	1.020288

可以看出將 X1 去除後，所有變數的 VIF 都小於 10 了

2.變數選取

將共線性的問題處理完後，接下來我想挑選變數，分別去採用向前選取法、向候選取法及逐步選取法

向前選取法:

AIC=3633.45

```
Call:
lm(formula = y ~ x4 + x10 + x6 + x5 + x3 + x7, data
 = data2)
```

向後選取法:

AIC=3633.45

```
Call:
lm(formula = y ~ x3 + x4 + x5 + x6 + x7 + x10, data
 = data2)
```

逐步選取法:

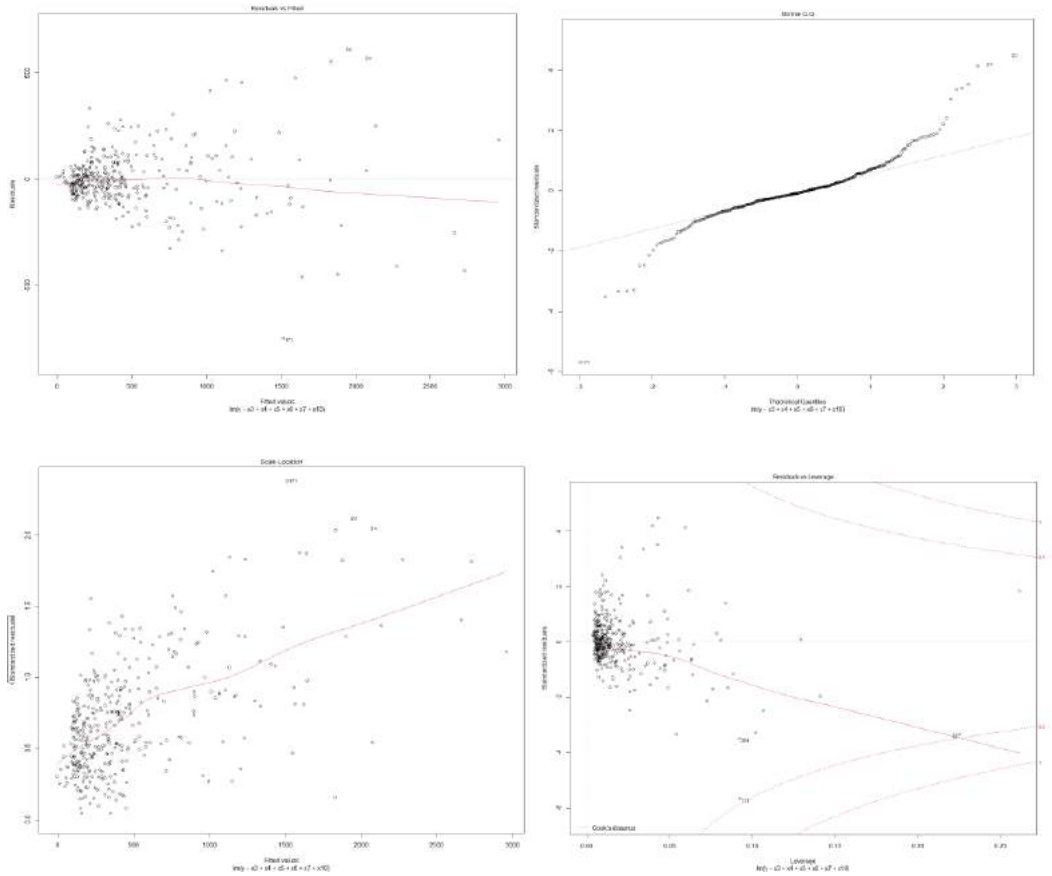
AIC=3633.45

```
Call:
lm(formula = y ~ x3 + x4 + x5 + x6 + x7 + x10, data
 = data2)
```

可看出三種方法所得到的 AIC 及模型皆相同，接下來將以此模型繼續做分析，也先看一下這個模型的一些資料

	Estimate	Std.Error	T value	Pr(>|t|)
Intercept	91.604171	18.592208	4.927	1.27e-0.6***
X3	-0.015409	0.006208	-2.482	0.0135*
X4	0.305159	0.033826	9.022	<2e-16***
X5	0.289357	0.069684	4.152	4.11e-05***
X6	0.233183	0.020244	11.519	<2e-16***
X7	-1.572064	0.834249	-1.884	0.0603.
X10	0.027591	0.002744	10.054	<2e-16***
Residual standard error:138 on 361 degrees of freedom				
Multiple R-aquare:0.9924 Adjusted R-square:0.9211				
F-statistic:714.7 on 6 and 361 DF , p-value: < 2.2e-16				

3.殘差分析



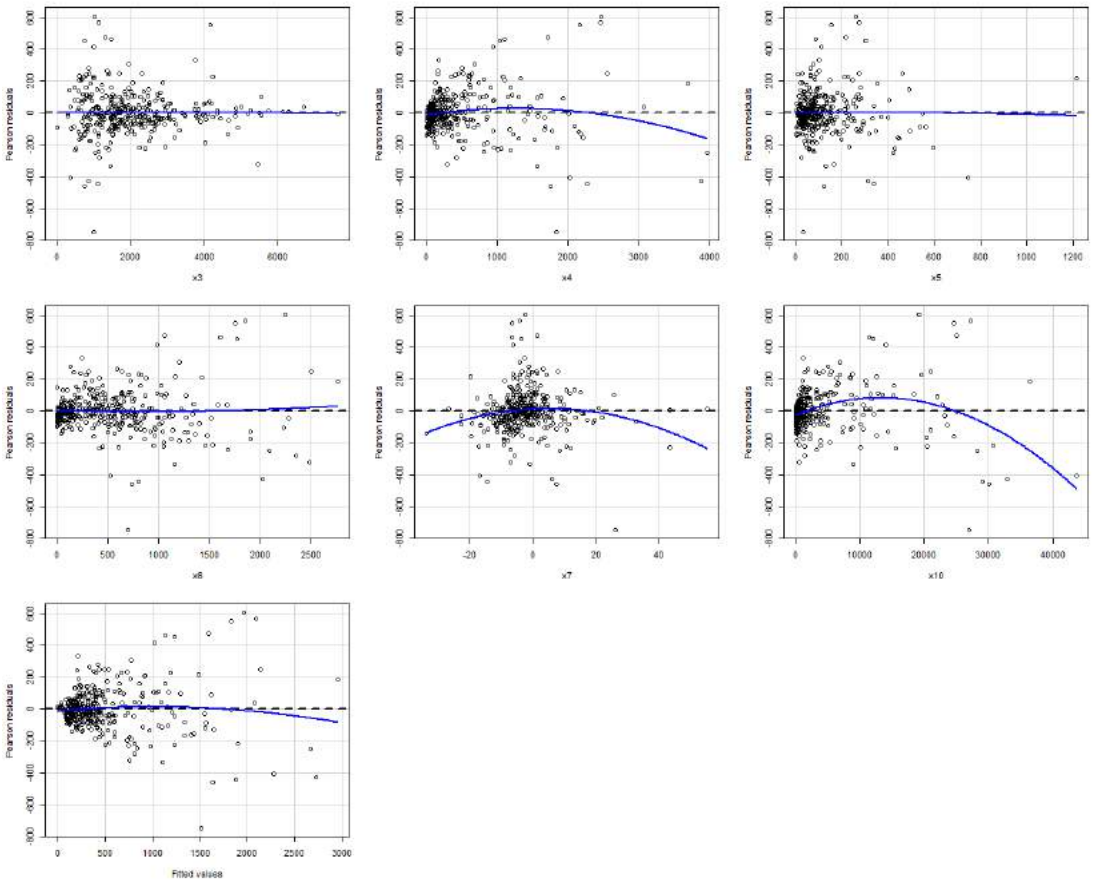
從上圖可看出殘差圖應該不是隨機分布，為了確認我還是先做檢定去確認自己的想法。

```
        One-sample Kolmogorov-Smirnov test

data:  scale(r5$residuals)
D = 0.11103, p-value = 0.0002296
alternative hypothesis: two-sided
```

由 p-value=0.0002296<0.05 可知，有足夠證據顯示 H0 不成立，即殘差並未服從常態分配。

Residual plot



變數變換

一開始我們觀察資料的分布後有一些右偏的現象，取 log 後的確都有蠻好的改變，因此這邊我們先對 y,x4,x5 取 log 後再去建模觀察。

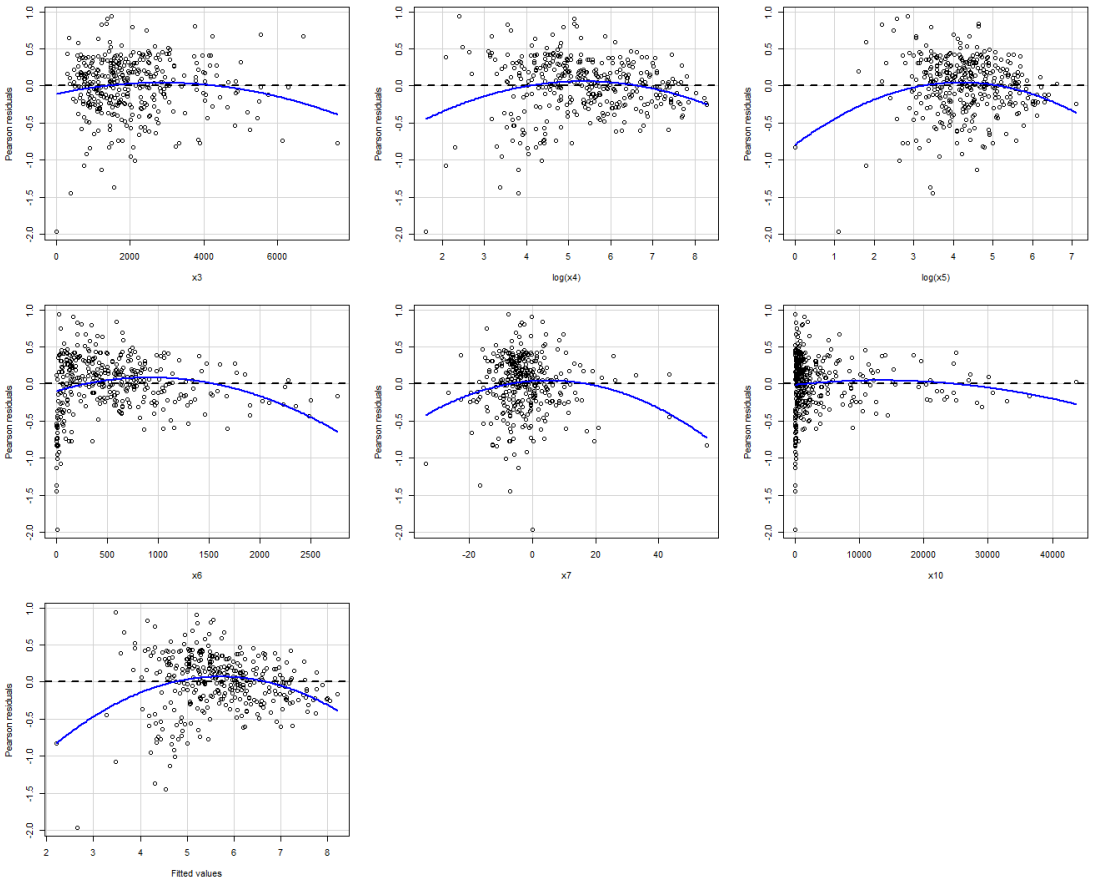
```
> lm10 <- lm(log(y)~x3+log(x4)+log(x5)+x6+x7+x10
  , data = data2)
```

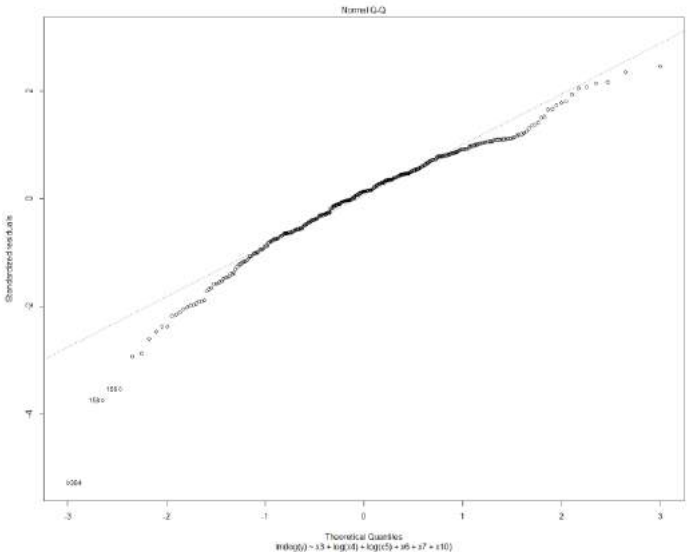
```
One-sample Kolmogorov-Smirnov test

data:  scale(lm10$residuals)
D = 0.067374, p-value = 0.07081
alternative hypothesis: two-sided
```

在經過變數變換後可看出，p-value=0.07081，表示無足夠證據顯示 H0 不成立，即殘差服從常態分布。

Residual plot





由 Residual plot 及 qqnorm 可看出模型優化許多。

```
Call:
lm(formula = log(y) ~ x3 + log(x4) + log(x5) + x6 + x7 + x10,
    data = data2)

Residuals:
     Min       1Q   Median       3Q      Max
-1.96148 -0.21897  0.04955  0.27094  0.94333

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.420e+00  1.618e-01   8.775  < 2e-16 ***
x3          -3.791e-05  1.764e-05  -2.149   0.0323 *
log(x4)      6.671e-01  2.899e-02  23.012  < 2e-16 ***
log(x5)      1.450e-01  2.431e-02   5.966 5.81e-09 ***
x6           2.624e-04  5.579e-05   4.703 3.65e-06 ***
x7          -1.257e-02  2.374e-03  -5.294 2.08e-07 ***
x10         -6.485e-06  4.575e-06  -1.418   0.1572
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3892 on 361 degrees of freedom
Multiple R-squared:  0.866,     Adjusted R-squared:  0.8638
F-statistic:   389 on 6 and 361 DF,  p-value: < 2.2e-16
```

BOXCOX

除了自己對資料的觀察後去做轉換，我也利用 BOXCOX 及 power transform 對變數去做轉換來比較

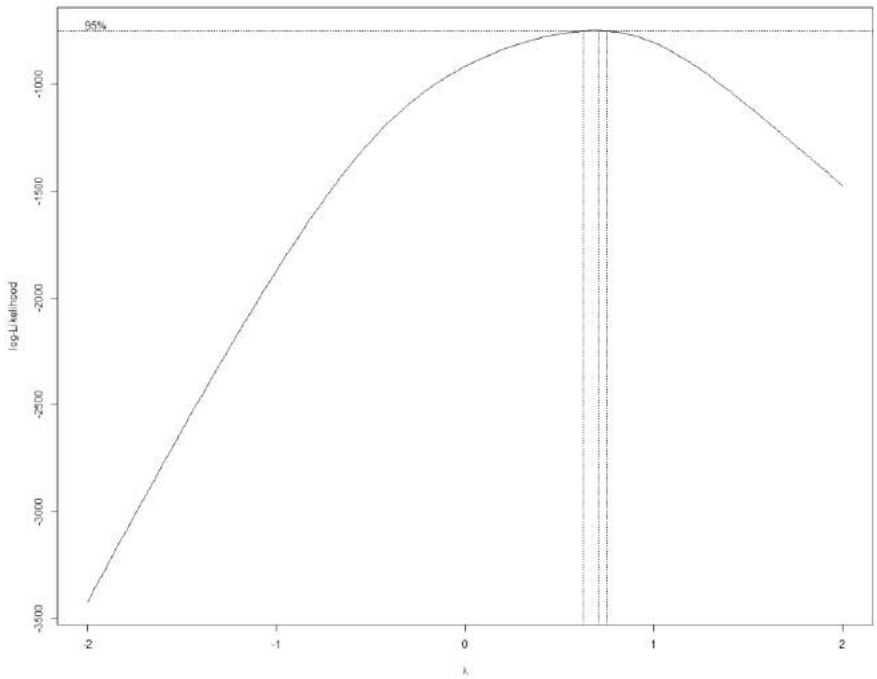
首先我先對 X 做 power transform

X3	X4	X5	X6	X7	X10
0.346540	0.076673	0.066394	0.362183	0.840143	0.065347

5	1	1	7	2	3

因此對 X4,X5,X10 取 log，X3 及 X6 都開根號

再來對 Y 做 BOXCOX



由上圖可知，lambda=0.7070707

取 lambda=0.5，Y=((Y^0.5)-1)/0.5

可得模型為:

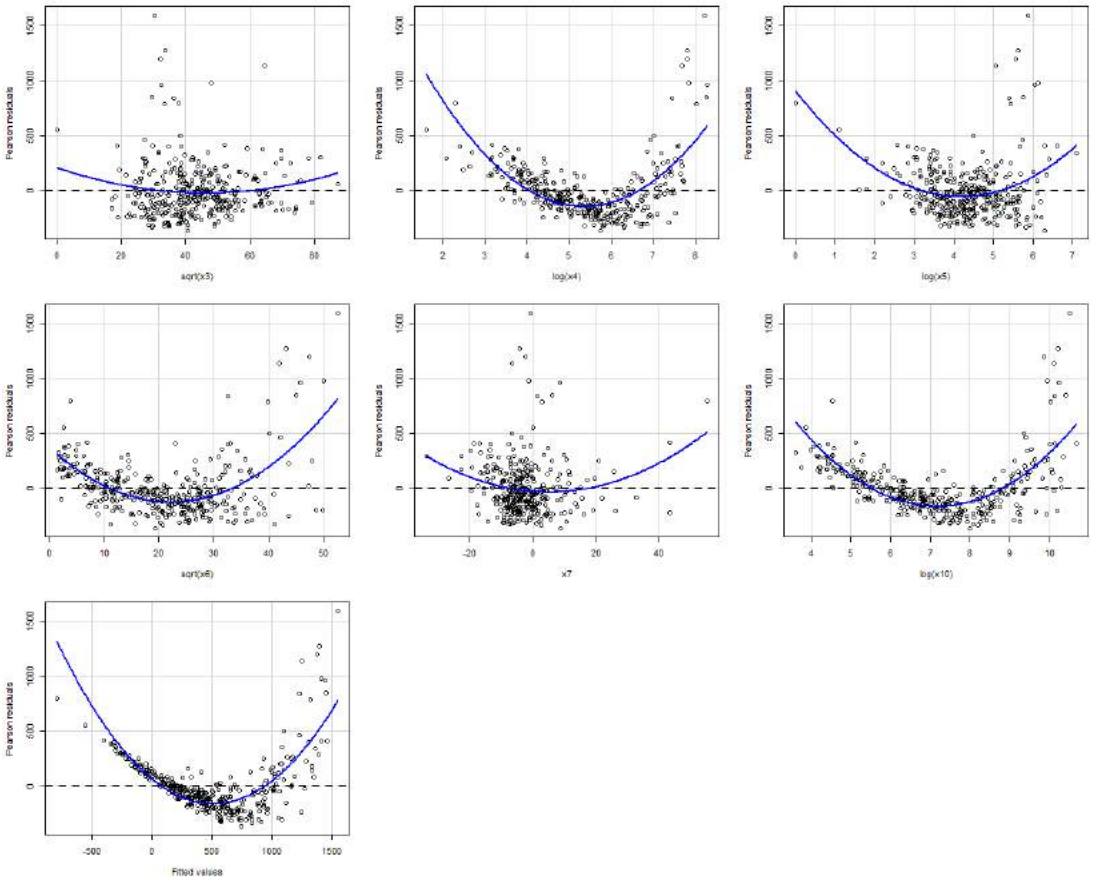
```
lm5 <- lm(y~sqrt(x3)+log(x4)+log(x5)+sqrt(x6)+x7+
log(x10),data = data2)
```

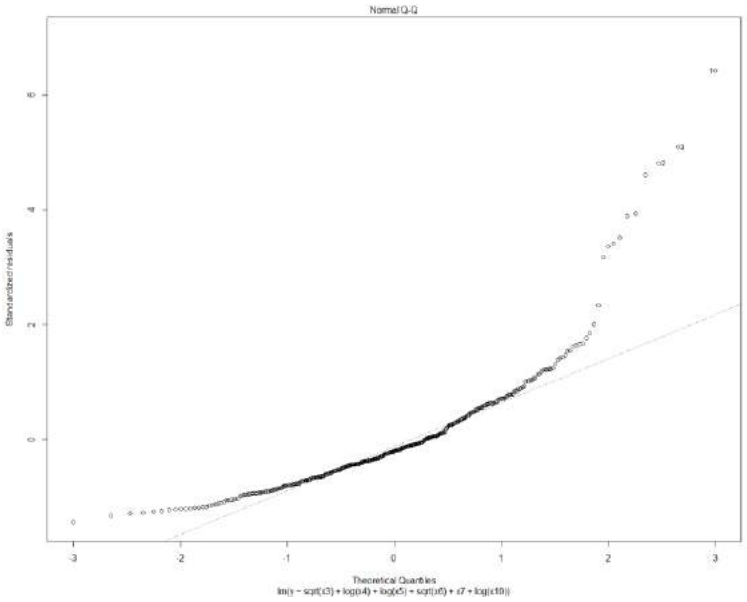
```
      One-sample Kolmogorov-Smirnov test

data:  scale(lm5$residuals)
D = 0.13127, p-value = 6.213e-06
alternative hypothesis: two-sided
```

在經過變數變換後可看出，p-value=6.213e-06，表示有足夠證據顯示 H0 不成立，即殘差不服從常態分布。

Residual plot





由 Residual plot 及 qqnorm 這樣的變數變換似乎不是很好，或許是受到離群值影響導致這樣的結果。

```
Call:
lm(formula = y ~ sqrt(x3) + log(x4) + log(x5) + sqrt(x6) + x7 +
    log(x10), data = data2)

Residuals:
    Min      1Q  Median      3Q     Max
-356.12 -160.03  -51.04   98.61 1592.15

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1317.559    105.456 -12.494  < 2e-16 ***
sqrt(x3)       -2.465      1.091  -2.260   0.0244 *
log(x4)        76.617     33.560   2.283   0.0230 *
log(x5)        81.935     15.395   5.322 1.80e-07 ***
sqrt(x6)        7.455      1.817   4.103 5.04e-05 ***
x7             -3.778      1.533  -2.465   0.0142 *
log(x10)      136.853     25.062   5.461 8.85e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 251.7 on 361 degrees of freedom
Multiple R-squared:  0.7418,    Adjusted R-squared:  0.7375
F-statistic: 172.8 on 6 and 361 DF,  p-value: < 2.2e-16
```

從兩種變數變換的比較可看出，只有前者的殘差服從常態分配，且從 Adjust R-square=0.8638>後者的 0.7375 也可以看出前者的模型解釋能力較佳，且 BOXCOX 幾乎對所有變數都做了變數變換，較難去解釋，因此接下來的將採用前者的模型繼續去分析。

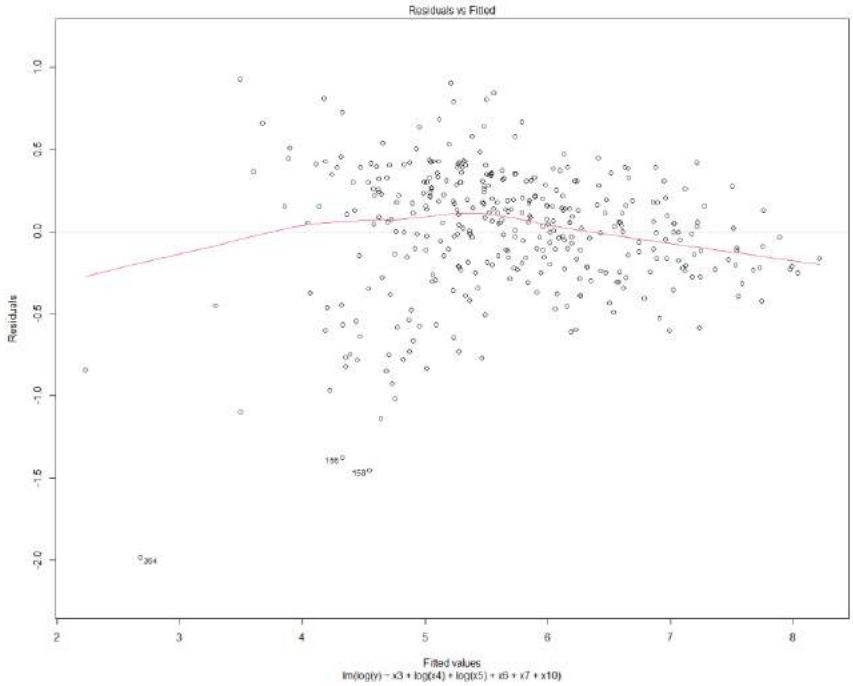
我們已經讓殘差服從常態了，接下來我將去檢定變異數的同質性

```
> ncvTest(lm10)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 115.2031, Df = 1, p = < 2.22e-16
```

從檢定結果可看出，p-value=<2.22e-16，表示有足夠證據顯示 H0 不成立，即模型具異質變異數，因此考慮此用 Weighted least square

Weighted least square

其中我將使用 HC3 來估計權重




```
> ncvTest(lm10_wt)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.01822564, Df = 1, p = 0.89261
```

從檢定結果可看出，p-value=0.89261>0.05，表示無足夠證據顯示 H0 不成立，即模型具同質變異數。

最後我再去確定做完 WLS 是否會影響模型的共線性，因此再對模型做一次 VIF 的診斷

X3	Log(x4)	Log(x5)	X6	X7	X10
1.674803	7.515889	2.432199	2.530103	3.900202	5.229310

可看出並沒有因為 WLS 而使模型具有共線性。

4.最終模型

```
> lm10_wt <- lm(log(y)~x3+log(x4)+log(x5)+x6+x7+x
10 , data = data2,weights = hc3wt)
```

	Estimate	Std.Error	T value	Pr(>|t|)
Intercept	1.456	2.678e-02	54.38	<2e-16***
X3	-3.738e-05	1.679e-06	-22.27	<2e-16***
Log(X4)	6.628e-01	3.405e-03	194.68	<2e-16***
Log(X5)	1.414e-01	3.805e-03	37.15	<2e-16***
X6	2.639e-04	7.706e-06	34.24	<2e-16***
X7	-1.270e-02	4.463e-04	-28.46	<2e-16***
X10	-6.042e-06	5.516e-07	-10.95	<2e-16***
Residual standard error:0.9854 on 361 degrees of freedom				
Multiple R-square:0.9985 Adjusted R-square:0.9985				
F-statistic:4.133e+04 on 6 and 361 DF , p-value: < 2.2e-16				

Adjust R-square 達到 0.9985，代表模型具有良好的解釋能力。

因為共線性的關係，把我們一開始認為可能會與 Y 有高度相關的 X1(單獨生活戶數)給拿掉了，再利用變數選取將一些變數也去除，且一開始從資料的觀察並不覺得 X10(碩士人口數)會進入模型，這

些是從敘述統計中不容易觀察到的，最終還是得到一個擁有相當優秀解釋能力的模型。