迴歸報告

綜合所得探討

統碩一 110354026 張順益

目錄

研究動機:

由於時常可以在電視新聞上面看到哪一個里的所得最高，然而卻僅僅只排出排名，卻沒有深入地進行探討，因此我在此次報告中，想要深入地去探討全國各村里的綜合所得。

資料介紹:

107 年全國內政大數據資料(以里為單位)(7600 筆)

反應變數:

綜合所得平均(以千為單位)

解釋變數:

1.扶養比: $\frac{\text{幼年人口+老年人口}}{\text{青壯年人口}}$(幼年:14 歲以下、老年:15 歲以上、青壯年:15 歲以上 64 歲以下)

2.人口密度

3.單獨生活戶數比

4. 15 歲以上學歷碩博比例

5.社會增加率 : $\frac{\text{遷入人口數}-\text{遷出人口數}}{\text{期中人口數}}$

6.出生率

7.性別比

8.原住民人口比

9.離婚率

10.死亡率

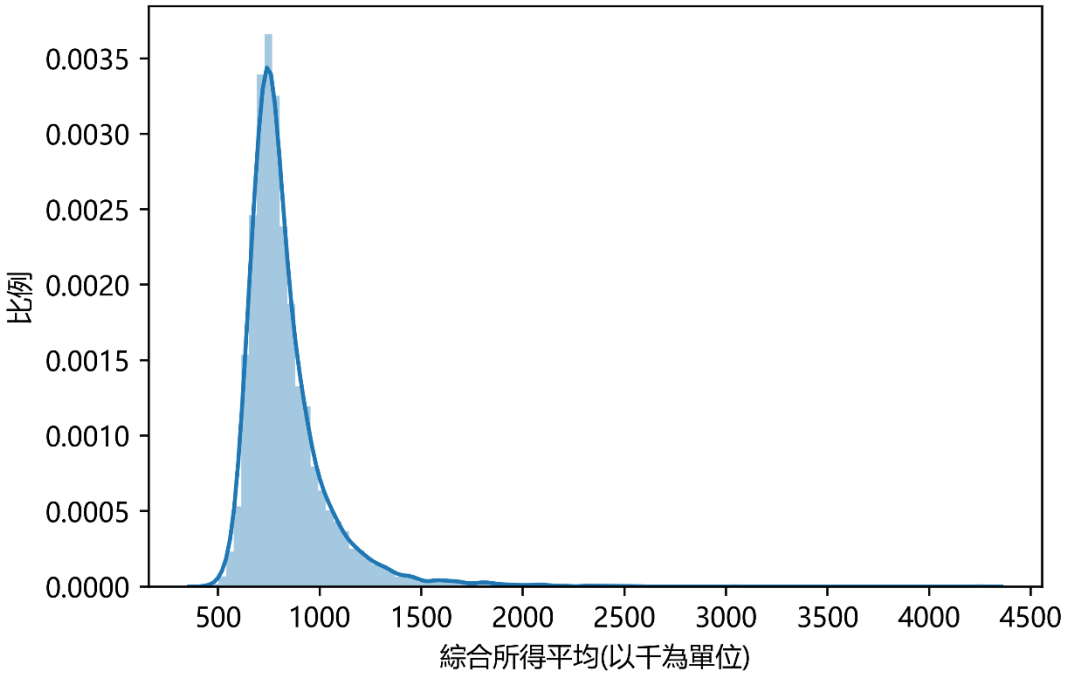
單獨生活戶數筆以及 15 歲以上學歷碩博比例，原資料都是給個數，然而我不希望變數受到總戶數或者是總人口數影響因此我各自將其轉換為比例。

資料處理及模型配適診斷:

先進行敘述統計，看一下各個變數的概況，再藉由 count 這一列了解變數是否有缺失值。

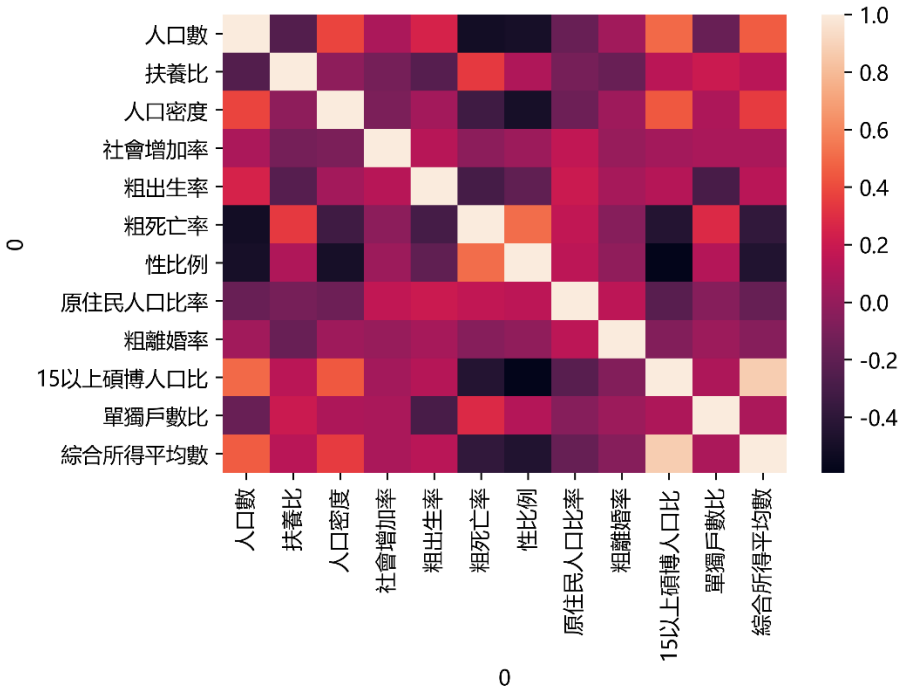
	人口數	扶養比	人口密度	單獨生活戶數	總戶數	15歲以上博士人口數	15歲以上碩士人口數	15-64歲人口數	65歲以上人口數	社會增加率	粗出生率
count	7760.000000	7760.000000	7760.000000	7760.000000	7760.000000	7760.000000	7760.000000	7760.000000	7760.000000	7644.000000	7759.000000
mean	3039.810825	39.845640	10437.060008	370.803093	1125.576933	17.869845	173.034794	2204.534536	442.463531	-0.940955	7.066595
std	2560.336406	7.679664	15826.780879	350.116907	986.015703	32.452129	231.556288	1886.385122	301.694625	20.474424	2.999129
min	56.000000	17.450000	0.310000	3.000000	23.000000	0.000000	0.000000	37.000000	13.000000	-221.380000	0.000000
25%	1193.000000	34.520000	447.740000	138.000000	425.000000	2.000000	34.000000	836.500000	229.000000	-11.340000	5.100000
50%	2272.000000	38.795000	2349.750000	260.000000	810.500000	7.000000	87.000000	1627.000000	370.000000	-3.130000	6.960000
75%	4274.250000	44.020000	14632.262500	496.000000	1582.250000	20.000000	226.000000	3103.000000	583.000000	6.272500	8.865000
max	43713.000000	109.960000	120638.880000	5008.000000	16692.000000	538.000000	3591.000000	32415.000000	3411.000000	619.960000	25.040000

粗死亡率	性比例	原住民人口比率	粗離婚率	綜合所得平均數
7759.000000	7760.000000	7315.000000	7759.000000	7759.000000
9.288219	104.099564	4.666554	2.253712	832.493491
4.610384	12.166975	16.211229	1.302259	208.517311
0.000000	69.330000	0.030000	0.000000	463.000000
6.055000	95.457500	0.360000	1.450000	712.000000
8.260000	102.380000	0.680000	2.160000	780.000000
11.460000	110.630000	1.440000	2.920000	890.000000
52.630000	233.020000	99.290000	13.770000	4252.000000



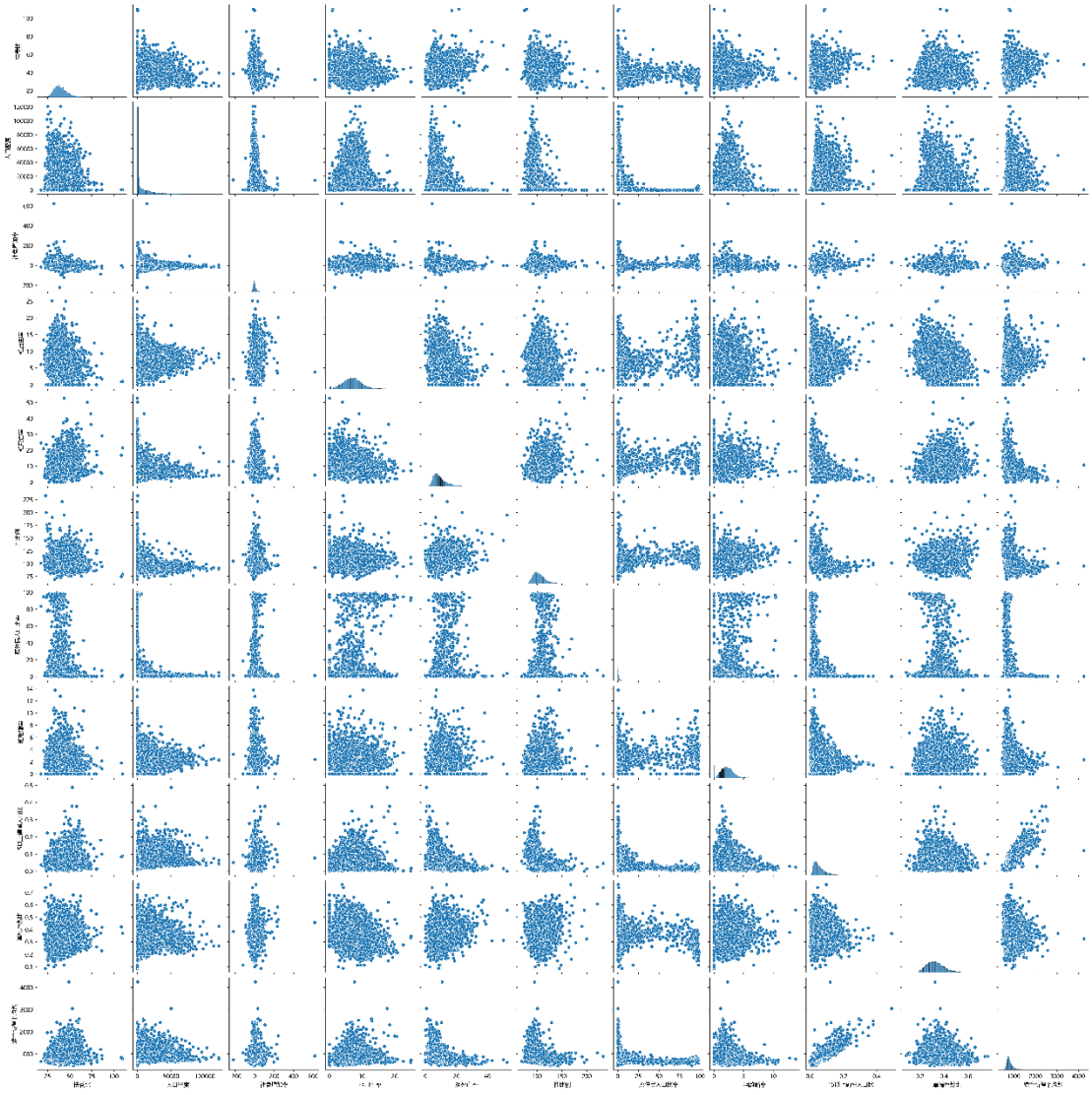
(圖一)

藉由圖一可以知道綜合平均所得為右偏分配並且大部分集中在 750-1000 左右。



(圖二)(相關係數熱圖)

由圖二可以知道綜合所得平均樹跟 15 歲以上碩博人口正相關係數是較高的，且各個變數間的負相關係數普遍並不大，然而某些解釋變數間的相關係數是偏高的。



(圖三)

藉由圖三可以約略了解個變數之間的分布情況，大概知道其相關程度。

人口數	0
扶養比	0
人口密度	0
社會增加率	116
粗出生率	1
粗死亡率	1
性比例	0
原住民人口比率	445
粗離婚率	1
15以上碩博人口比	0
單獨戶數比	0
綜合所得平均數	1

(圖四)

圖四為各變數其缺失情況，而由於我後面會採取迴歸建模，因此我在此處採取 KNN(K=5)，進行補值。

首先我先將所有的變數都放入跑多元迴歸，在這裡可以看出離婚率以及單獨戶數筆並不顯著，並且從係數上可以看出雖然 15 歲以上學歷碩博比為比例但其係數相當大，因此即便比例只增加 1%對於綜合所得的影響依然蠻大的。

Dep. Variable:	綜合所得平均數	**R-squared:**	0.774
Model:	OLS	**Adj. R-squared:**	0.773
Method:	Least Squares	**F-statistic:**	2649.
Date:	Fri, 14 Jan 2022	**Prob (F-statistic):**	0.00
Time:	12:13:38	**Log-Likelihood:**	-46683.
No. Observations:	7760	**AIC:**	9.339e+04
Df Residuals:	7749	**BIC:**	9.347e+04
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	330.4325	16.118	20.501	0.000	298.837	362.028
扶養比	0.5039	0.172	2.938	0.003	0.168	0.840
人口密度	-0.0002	8.59e-05	-2.597	0.009	-0.000	-5.46e-05
社會增加率	0.2228	0.058	3.847	0.000	0.109	0.336
粗出生率	1.9393	0.423	4.581	0.000	1.110	2.769
粗死亡率	-1.9496	0.340	-5.740	0.000	-2.615	-1.284
性比例	1.9078	0.130	14.690	0.000	1.653	2.162
原住民人口比率	0.2113	0.077	2.757	0.006	0.061	0.362
粗離婚率	0.9877	0.890	1.110	0.267	-0.757	2.732
15以上碩博人口比	5021.4984	41.814	120.090	0.000	4939.531	5103.466
單獨戶數比	2.2686	16.627	0.136	0.891	-30.325	34.862

之後我透過逐步迴歸(AIC)backward 以及 forward 皆找到變數為扶養比、人口密度、原住民人口比例、社會增加率、出生率、死亡率、性比例、以及 15 歲以上學歷碩博比。

	Df	Sum of Sq	RSS	AIC
<none>			72308146	66437
- 扶養比	1	46834	72354980	66440
- 人口密度	1	66607	72374753	66442
- 原住民人口比率	1	100197	72408343	66445
- 社會增加率	1	115882	72424028	66447
- 粗出生率	1	208870	72517016	66456
- 粗死亡率	1	422471	72730617	66477
- 性比例	1	2113305	74421451	66643
- X15以上碩博人口比	1	144192639	216500785	74341

之後我在對所選取的變數進行迴歸，此時可以看出所有變數皆為顯著，且 R square 依然維持在 0.774。

Dep. Variable:	綜合所得平均數	**R-squared:**	0.774
Model:	OLS	**Adj. R-squared:**	0.773
Method:	Least Squares	**F-statistic:**	3312.
Date:	Fri, 14 Jan 2022	**Prob (F-statistic):**	0.00
Time:	14:11:20	**Log-Likelihood:**	-46684.
No. Observations:	7760	**AIC:**	9.339e+04
Df Residuals:	7751	**BIC:**	9.345e+04
Df Model:	8		
Covariance Type:	nonrobust		

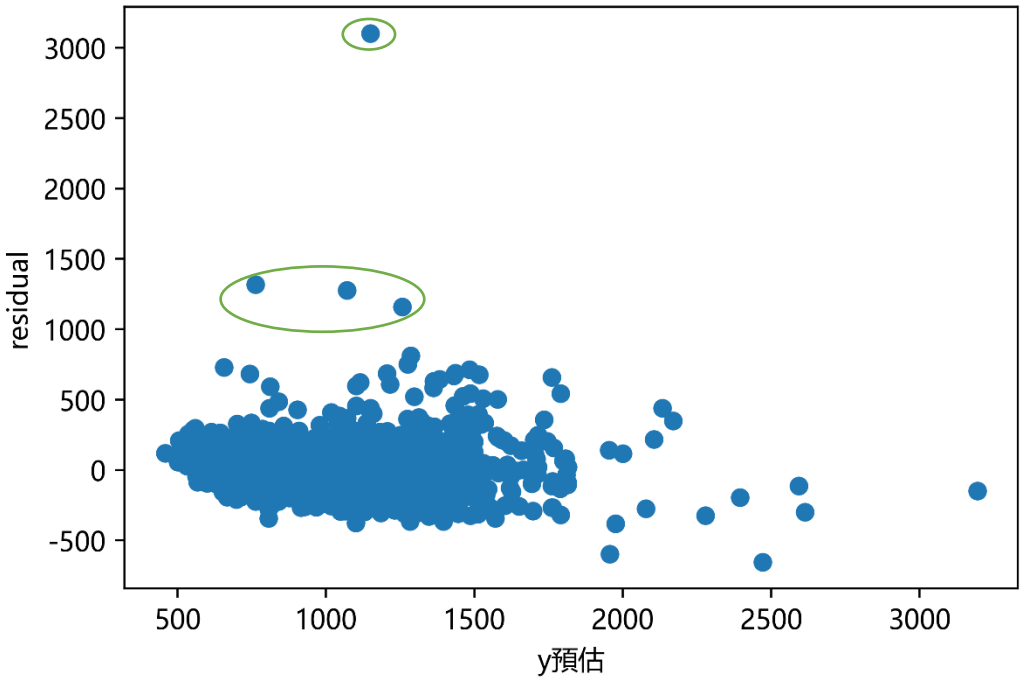
	coef	std err	t	P>|t|	[0.025	0.975]
const	334.3233	15.558	21.488	0.000	303.825	364.822
扶養比	0.4820	0.170	2.830	0.005	0.148	0.816
人口密度	-0.0002	8.47e-05	-2.559	0.011	-0.000	-5.07e-05
原住民人口比率	0.2212	0.076	2.906	0.004	0.072	0.370
社會增加率	0.2224	0.057	3.873	0.000	0.110	0.335
粗出生率	1.9274	0.413	4.669	0.000	1.118	2.737
粗死亡率	-1.9471	0.328	-5.934	0.000	-2.590	-1.304
性比例	1.9076	0.129	14.795	0.000	1.655	2.160
15以上碩博人口比	5020.2626	40.856	122.877	0.000	4940.174	5100.351

之後我考量到每個里的人數並不相同，因此我使用權重迴歸，並且將權重設為各里的人數，可以發現各變數依然都顯著但 R square 變為 0.837

Dep. Variable:	綜合所得平均數	**R-squared:**	0.837
Model:	WLS	**Adj. R-squared:**	0.837
Method:	Least Squares	**F-statistic:**	4992.
Date:	Fri, 14 Jan 2022	**Prob (F-statistic):**	0.00
Time:	14:22:02	**Log-Likelihood:**	-48083.
No. Observations:	7760	**AIC:**	9.618e+04
Df Residuals:	7751	**BIC:**	9.625e+04
Df Model:	8		
Covariance Type:	nonrobust		

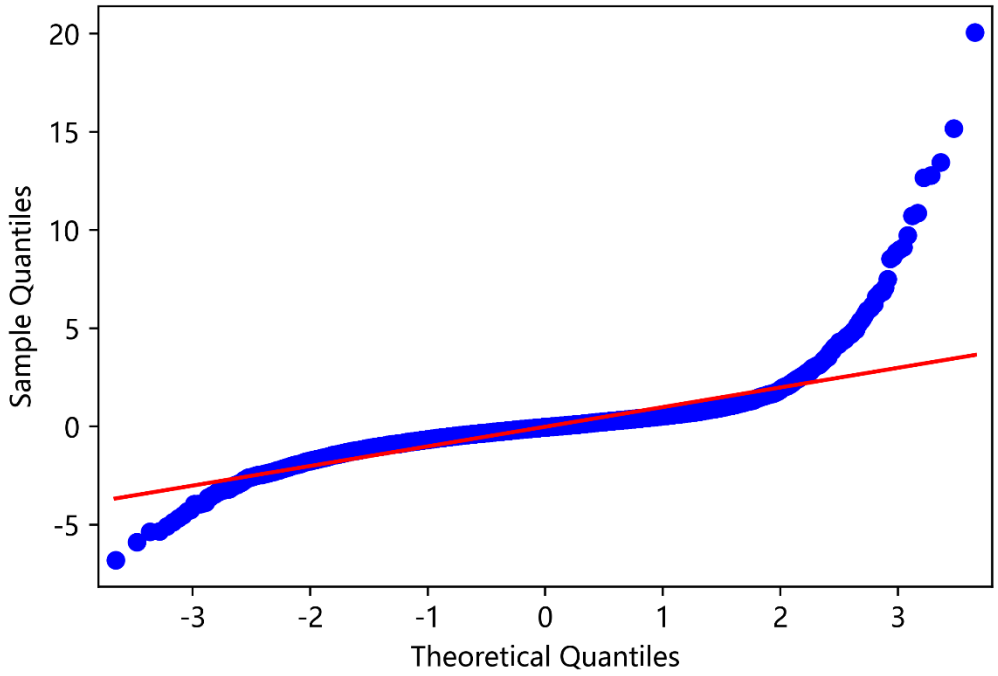
	coef	std err	t	P>|t|	[0.025	0.975]
const	261.3139	19.357	13.500	0.000	223.369	299.259
扶養比	0.6511	0.197	3.307	0.001	0.265	1.037
人口密度	-0.0002	7.42e-05	-2.641	0.008	-0.000	-5.05e-05
原住民人口比率	0.3783	0.130	2.921	0.003	0.124	0.632
社會增加率	0.1389	0.055	2.537	0.011	0.032	0.246
粗出生率	1.8444	0.493	3.738	0.000	0.877	2.812
粗死亡率	-4.3994	0.471	-9.347	0.000	-5.322	-3.477
性比例	2.5806	0.177	14.566	0.000	2.233	2.928
15以上碩博人口比	5346.9168	37.886	141.132	0.000	5272.650	5421.184

再來我對權重迴歸進行殘差檢定



(圖五)

由圖五可以看出可能會有離群值的存在。



(圖六)

藉由圖六 Q-Q plot 可以看出模型並不符合常態分配。

且由於我的資料筆數較多因此我不選擇 shapiro test 做常態檢定改為使用 Kolmogorov-Smirnov test 進行常態檢定，並且發現確實非常態。

```
KstestResult(statistic=0.5036666386163267, pvalue=0.0)
```

所以我對 y 進行 boxcox 轉換並且 λ 為-1.767(由於此為學術報告非實務所以我先以最佳解為主)，且將 cook 距離>4/(n-p-1),p 為變數個數，再進行一次迴歸以及權重迴歸，可以看出一般迴歸的 R squre 上升但性別比變為不顯著，而權重迴歸其性別比以及扶養比也變為不顯著，且 R square 略為下降。

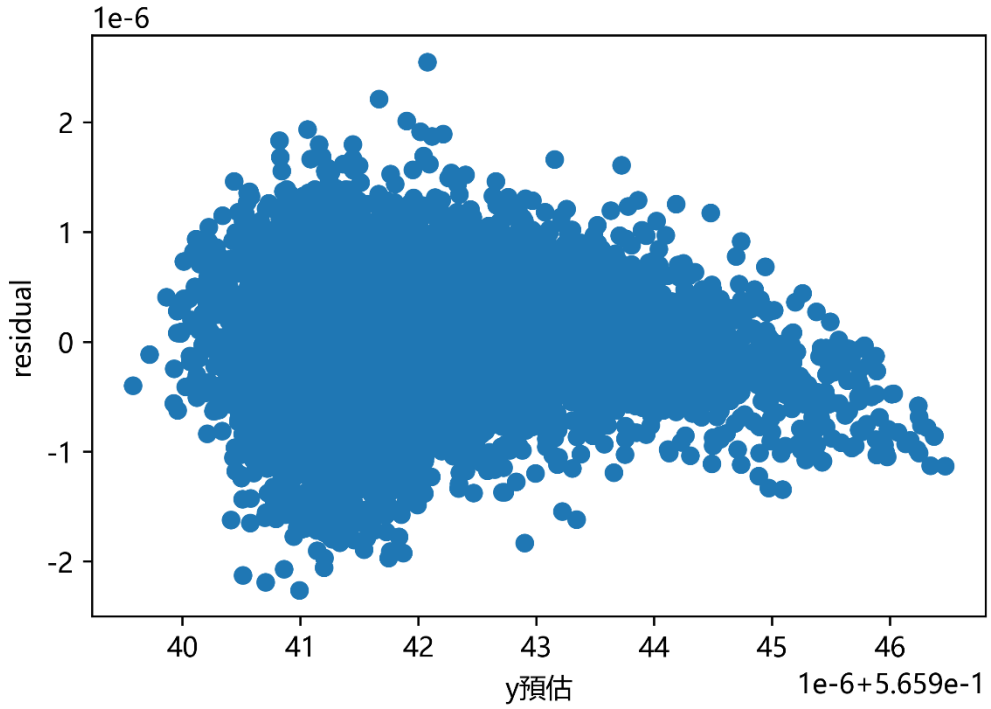
Dep. Variable:	y	**R-squared:**	0.772
Model:	OLS	**Adj. R-squared:**	0.772
Method:	Least Squares	**F-statistic:**	3078.
Date:	Fri, 14 Jan 2022	**Prob (F-statistic):**	0.00
Time:	15:22:15	**Log-Likelihood:**	94066.
No. Observations:	7266	**AIC:**	-1.881e+05
Df Residuals:	7257	**BIC:**	-1.881e+05
Df Model:	8		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	0.5659	1.09e-07	5.2e+06	0.000	0.566	0.566
扶養比	8.11e-10	1.08e-09	0.753	0.452	-1.3e-09	2.92e-09
人口密度	2.052e-12	5.19e-13	3.955	0.000	1.03e-12	3.07e-12
原住民人口比率	-6.471e-09	5.79e-10	-11.170	0.000	-7.61e-09	-5.34e-09
社會增加率	3.604e-09	4.12e-10	8.758	0.000	2.8e-09	4.41e-09
粗出生率	1.256e-08	2.61e-09	4.812	0.000	7.44e-09	1.77e-08
粗死亡率	-3.136e-08	2.18e-09	-14.417	0.000	-3.56e-08	-2.71e-08
性比例	2.144e-09	9.18e-10	2.336	0.020	3.45e-10	3.94e-09
15以上碩博人口比	2.807e-05	2.87e-07	97.740	0.000	2.75e-05	2.86e-05

Dep. Variable:	y	**R-squared:**	0.818
Model:	WLS	**Adj. R-squared:**	0.818
Method:	Least Squares	**F-statistic:**	4090.
Date:	Fri, 14 Jan 2022	**Prob (F-statistic):**	0.00
Time:	15:34:18	**Log-Likelihood:**	93843.
No. Observations:	7266	**AIC:**	-1.877e+05
Df Residuals:	7257	**BIC:**	-1.876e+05
Df Model:	8		
Covariance Type:	nonrobust		

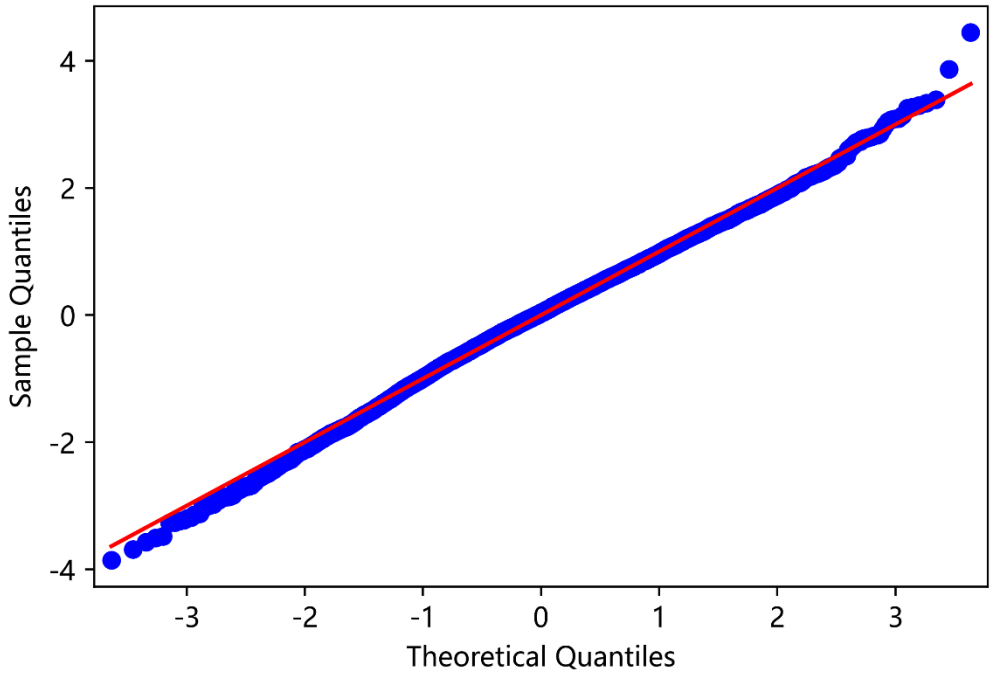
	coef	std err	t	P>|t|	[0.025	0.975]
const	0.5659	1.16e-07	4.88e+06	0.000	0.566	0.566
扶養比	-1.678e-09	1.06e-09	-1.580	0.114	-3.76e-09	4.04e-10
人口密度	2.213e-12	4.03e-13	5.486	0.000	1.42e-12	3e-12
原住民人口比率	-7.665e-09	8.25e-10	-9.295	0.000	-9.28e-09	-6.05e-09
社會增加率	3.57e-09	3.62e-10	9.869	0.000	2.86e-09	4.28e-09
粗出生率	1.611e-08	2.68e-09	6.020	0.000	1.09e-08	2.14e-08
粗死亡率	-4.181e-08	2.59e-09	-16.149	0.000	-4.69e-08	-3.67e-08
性比例	1.486e-09	1.05e-09	1.419	0.156	-5.67e-10	3.54e-09
15以上碩博人口比	2.672e-05	2.41e-07	111.044	0.000	2.63e-05	2.72e-05

之後我再做一次殘差檢定。



(圖七)

與圖五相比離群值的問題似乎減少了許多。



(圖八)

由圖七我可以看出相較於圖六表現是更好了但是貌似還是非常態，

從 k-s test 中可以看出依然並非常態

```
KstestResult(statistic=0.4999990983982532, pvalue=0.0)
```

於是我繼續嘗試對 x 轉換，我對 x 進行 yeo-johnson 轉換，然而其 R squre 下降很多，且經過 k-s test 測試後依然非常態，所以結果就沒有放上來。

Dep. Variable:	y	**R-squared (uncentered):**	0.281
Model:	WLS	**Adj. R-squared (uncentered):**	0.280
Method:	Least Squares	**F-statistic:**	354.2
Date:	Fri, 14 Jan 2022	**Prob (F-statistic):**	0.00
Time:	20:23:26	**Log-Likelihood:**	-6043.8
No. Observations:	7266	**AIC:**	1.210e+04
Df Residuals:	7258	**BIC:**	1.216e+04
Df Model:	8		
Covariance Type:	nonrobust		

而我也嘗試使用 lasso regression 進行變數選擇，然而其只選到人口密度這一個變數，因此我最終決定使用 stepwise regression

```
array([ 0.        ,  0.        ,  0.00446421,  0.        ,  0.        ,
       -0.        , -0.        , -0.        , -0.        ,  0.        ,
        0.        ])
```

(lasso regression 係數)

之後我對 y 進行過 boxcox 轉換的權重迴歸進行 Breusch-Pagan 異質性檢定，為變異數不同質。

P = 1.2496103900930718e-76

總結:

雖然迴歸模型以及權重迴歸的 R squre 的表現都不差，在進行過變數轉換之後，其在 qqplot 中也明顯更加像常態，然而其在檢定中依然並非常態，且也非變異數同質，因此這一個模型中的一些變數顯著等等運用到迴歸常態以及變異數同質假設的部分可能會有更高的偏誤。