國立中山大學 應用數學所

迴歸分析 期末書面報告

世界幸福探索性數據分析
World Happiness Exploratory Data Analysis

姓名 : 温宏岳

學號 : M112040010

民國 111 年 12 月 26 日

目錄

壹、 研究動機與目的

本研究旨在探討世界幸福人均國內生產毛額（Log GDP per capita）、社會支持資源（Social support）、預期健康壽命（Healthy life expectancy at birth）、人生抉擇自由度（Freedom to make life choices）、慷慨程度（Generosity）、貪腐程度（Perceptions of corruption）等指標之相關性與影響情形。本研究以 Kaggle 的資料做描述性統計、獨立樣本 t 檢定、Pearson 相關分析、探索性因素分析與多元逐步迴歸等進行分析。最後找出最佳的配適模型，以利往後做最佳的迴歸分析。

貳、 資料背景介紹

世界幸福報告（英語：World Happiness Report）為聯合國為衡量幸福的可持續發展方案，於網路出版的國際調查報告。

1. 計算方式 :

《世界幸福報告》的排名是使用蓋洛普世界民意調查的數據，這是由民意調查中提出的生活評估問題的回答，它要求受訪者想出一個階梯，對他們來說最好的是 10，最壞的是 0，評分的細項共有 10 點 :

(1) 人均 GDP (Log GDP per capita)
(2) 社會支持資源 (Social support)
(3) 預期的健康壽命 (Healthy Life Expectancy)
(4) 人生抉擇的自由 (Freedom to make life choices)
(5) 慷慨程度 (Generosity)
(6) 正面影響 (Positive affect)
(7) 負面影響 (Negative affect)
(8) 家庭收入報告的基尼係數 (GINI of household income reported)
(9) 世界銀行提供的基尼指數 (GINI index from the World Bank)
(10) 對於政府機關的信任程度 (Institutional trust)

這十點細項將表現出以下六大因素，這六大因素（GDP 水平、預期壽命、慷慨、社會支持、自由和腐敗）中的每一個皆有助於評估每個國家。

2. 殘差 :

除了上面這些指標外，因為統計方法的緣故，分數裡還加了一項「 Dystopia + 殘差 」這個數字。在這裡， Dystopia 是個「作為標準的虛擬國家」，擁有上面六項指標的最低分數；「殘差」則是代表實際值和這個統計模型預估值的差異：以迴歸分析算出來的值跟實際

值的差異（高為正數、低為負數），再加上 dystopia 的實際值，簡單來說，可以視為這個統計模型中「無法解釋的部分」。

因此，一開始就可需移除這些「無法解釋的部分」的資料。

3. 樣本來源 ：

每年每個國家的典型樣本為 1,000 人，聯合國使用最近三年的回復來提供最新的生活評估，所以如果一個典型的國家每年進行調查，樣本量將是 3,000，其樣本數夠多，可以減少隨機抽樣誤差；然而，目前還有許多國家沒有進行年度調查。

4. 數據“浪潮”：

蓋洛普將每個日曆年收集的調查作為當年調查浪潮的一部分。在絕大多數情況下，浪潮對應於日曆年，但也有一些例外。一些在 2022 年初完成的調查被認為是 2021 年浪潮的一部分。

並非每個國家每年都接受調查。因此，調查波的規模每年也不同。

參、 變數介紹

Feature:

- 人均國內生產毛額（Log GDP per capita）— 數值:0 ~ 100
- 社會支持資源（Social support）— 數值: 0 ~ 1
- 預期健康壽命（Healthy life expectancy at birth）— 數值: 0 ~ 100
- 人生抉擇自由度（Freedom to make life choices）— 數值: 0 ~ 1
- 慷慨程度（Generosity）— 透過捐贈來評估程度— 數值: -1 ~ 1
- 貪腐程度（Perceptions of corruption）— 對政府的信任度— 數值:
- 反烏托邦程度（Ladder Score InDystopia）— 它是一種不得人心、令人恐懼的假想社群或社會，是與理想社會相反的，一種極端惡劣的社會最終形態— 數值: 0 ~ 1

  (由於在資料中數值皆相同，但還是有參考程度，因此我會放入模型中做預測)

Target:

- 幸福指數（Ladder Score）— 數值: 0 ~ 10

肆、 資料匯入與預處理

a. **資料型態（圖 1）**：除了 CountryName、RegionalIndicato 為物件，其他皆為浮點數

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 149 entries, 0 to 148
Data columns (total 20 columns):
 #   Column                            Non-Null Count  Dtype
---  ------                            --------------  -----
 0   CountryName                       149 non-null    object
 1   RegionalIndicator                 149 non-null    object
 2   LadderScore                       149 non-null    float64
 3   StandardErrorOfLadderScore        149 non-null    float64
 4   upperwhisker                      149 non-null    float64
 5   lowerwhisker                      149 non-null    float64
 6   LoggedGDPPerCapita                149 non-null    float64
 7   SocialSupport                     149 non-null    float64
 8   HealthyLifeExpectancy             149 non-null    float64
 9   FreedomToMakeLifeChoices          149 non-null    float64
 10  Generosity                        149 non-null    float64
 11  PerceptionsOfCorruption           149 non-null    float64
 12  LadderScoreInDystopia             149 non-null    float64
 13  ExplainedbyLogGDPpercapita        149 non-null    float64
 14  ExplainedbySocialsupport          149 non-null    float64
 15  ExplainedbyHealthylifeexpectancy  149 non-null    float64
 16  ExplainedbyFreedomtomakelifechoices 149 non-null  float64
 17  ExplainedbyGenerosity             149 non-null    float64
 18  ExplainedbyPerceptionsofcorruption 149 non-null   float64
 19  Dystopiaresidual                  149 non-null    float64
dtypes: float64(18), object(2)
memory usage: 23.4+ KB


len =  149
```

（圖 1）

b. **移除無關值(圖 2)**：下圖藍色框線中的資料

```
RangeIndex: 149 entries, 0 to 148
Data columns (total 20 columns):
 #   Column                            Non-Null Count  Dtype
---  ------                            --------------  -----
 0   CountryName                       149 non-null    object
 1   RegionalIndicator                 149 non-null    object
 2   LadderScore                       149 non-null    float64
 3   StandardErrorOfLadderScore        149 non-null    float64
 4   upperwhisker                      149 non-null    float64
 5   lowerwhisker                      149 non-null    float64
 6   LoggedGDPPerCapita                149 non-null    float64
 7   SocialSupport                     149 non-null    float64
 8   HealthyLifeExpectancy             149 non-null    float64
 9   FreedomToMakeLifeChoices          149 non-null    float64
 10  Generosity                        149 non-null    float64
 11  PerceptionsOfCorruption           149 non-null    float64
 12  LadderScoreInDystopia             149 non-null    float64
 13  ExplainedbyLogGDPpercapita        149 non-null    float64
 14  ExplainedbySocialsupport          149 non-null    float64
 15  ExplainedbyHealthylifeexpectancy  149 non-null    float64
 16  ExplainedbyFreedomtomakelifechoices 149 non-null  float64
 17  ExplainedbyGenerosity             149 non-null    float64
 18  ExplainedbyPerceptionsofcorruption 149 non-null   float64
 19  Dystopiaresidual                  149 non-null    float64
dtypes: float64(18), object(2)
```

（圖 2）

c. **檢查資料是否有缺失值（圖 3）：** 無

```
df.isnull().sum(axis=0)
```

```
CountryName                          0
RegionalIndicator                    0
LadderScore                          0
StandardErrorOfLadderScore           0
upperwhisker                         0
lowerwhisker                         0
LoggedGDPPerCapita                   0
SocialSupport                        0
HealthyLifeExpectancy                0
FreedomToMakeLifeChoices             0
Generosity                           0
PerceptionsOfCorruption              0
LadderScoreInDystopia                0
ExplainedbyLogGDPpercapita           0
ExplainedbySocialsupport             0
ExplainedbyHealthylifeexpectancy     0
ExplainedbyFreedomtomakelifechoices  0
ExplainedbyGenerosity                0
ExplainedbyPerceptionsofcorruption   0
Dystopiaresidual                     0
dtype: int64
```

（圖 3）

伍、 複迴歸模型

Score 會透過 R^2 來判定我們模型的精準程度；如果訓練集的分數很高，但測試集的分數卻很低，那就是過度擬和。

以下分別用不同方式做複迴歸模型，並從中挑取調整後 R^2 最大的模型，即為較合適的模型。

a. 從理論公式推導
b. 使用 sklearn linear_model 計算(雖然無法直接從文檔中找到任何計算 adjusted R^2 方式的函數。)
c. 使用 statsmodels 計算

	Model	R^2	Adj R^2
0	a	0.7558471374226854	0.7437260733231024
1	b	0.7558471374226855	0.7437260733231026
2	c	0.7558471374226855	0.7455308192856158

（圖 4）

=> 由 3 種方式(圖 4)得知，調整後的 R^2(Adj R^2)中最好的模型是 c，也就是使用 statsmodels 的模型，因此在之後的模型皆以此方式做配適。

陸、 OLS 線性關係判斷

線性回歸模型，顧名思義，首先要保證自變量與因變量之間存在線性關係。關於線性關係的判斷，我們可以通過圖形或 Pearson 相關係數來識別。

一般情況下的評判標準 ：

- 當相關係數 低於 0.4 ，則表明變量之間存在弱相關關係；
- 當相關係數在 0.4~0.6 之間 ，則說明變量之間存在中度相關關係；
- 當相關係數在 0.6 以上時 ，則反映變量之間存在強相關關係。

a. 創建線性迴歸最小平方法模型(圖 5)

OLS Regression Results

Dep. Variable:	LadderScore	**R-squared:**	0.756
Model:	OLS	**Adj. R-squared:**	0.746
Method:	Least Squares	**F-statistic:**	73.27
Date:	Tue, 13 Dec 2022	**Prob (F-statistic):**	5.06e-41
Time:	00:10:50	**Log-Likelihood:**	-116.50
No. Observations:	149	**AIC:**	247.0
Df Residuals:	142	**BIC:**	268.0
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
LoggedGDPPerCapita	0.2795	0.087	3.219	0.002	0.108	0.451
SocialSupport	2.4762	0.668	3.706	0.000	1.155	3.797
HealthyLifeExpectancy	0.0303	0.013	2.274	0.024	0.004	0.057
FreedomToMakeLifeChoices	2.0105	0.495	4.063	0.000	1.032	2.989
Generosity	0.3644	0.321	1.134	0.259	-0.271	0.999
PerceptionsOfCorruption	-0.6051	0.291	-2.083	0.039	-1.179	-0.031
LadderScoreInDystopia	-0.9207	0.259	-3.548	0.001	-1.434	-0.408

Omnibus:	12.908	**Durbin-Watson:**	1.614
Prob(Omnibus):	0.002	**Jarque-Bera (JB):**	13.688
Skew:	-0.667	**Prob(JB):**	0.00107
Kurtosis:	3.650	**Cond. No.**	1.05e+03

(圖 5)

[1] 標準誤差假設正確指定了誤差的共變異數矩陣。

[2] 條件數很大，1.05e+03。 這可能表明存在強多重共線性或其他數值問題。

此時的迴歸模型為下:

$$
\begin{aligned}
LadderScore &= 0.2795 * LoggedGDPPerCapita \\
&+ 2.4762 * SocialSupport \\
&+ 0.0303 * HealthyLifeExpectancy \\
&+ 2.0105 * FreedomToMakeLifeChoices \\
&+ 0.3644 * Generosity \\
&- 0.6051 * PerceptionsOfCorruption \\
&- 0.9207 * LadderScoreInDystopia
\end{aligned}
$$

b. LadderScore 與自變量之間的相關係數(圖 6)

```
df.corrwith(df['LadderScore'])
```

```
LadderScore                    1.000000
StandardErrorOfLadderScore    -0.470787
LoggedGDPPerCapita             0.789760
SocialSupport                  0.756888
HealthyLifeExpectancy          0.768099
FreedomToMakeLifeChoices       0.607753
Generosity                    -0.017799
PerceptionsOfCorruption       -0.421140
LadderScoreInDystopia               NaN
dtype: float64
```

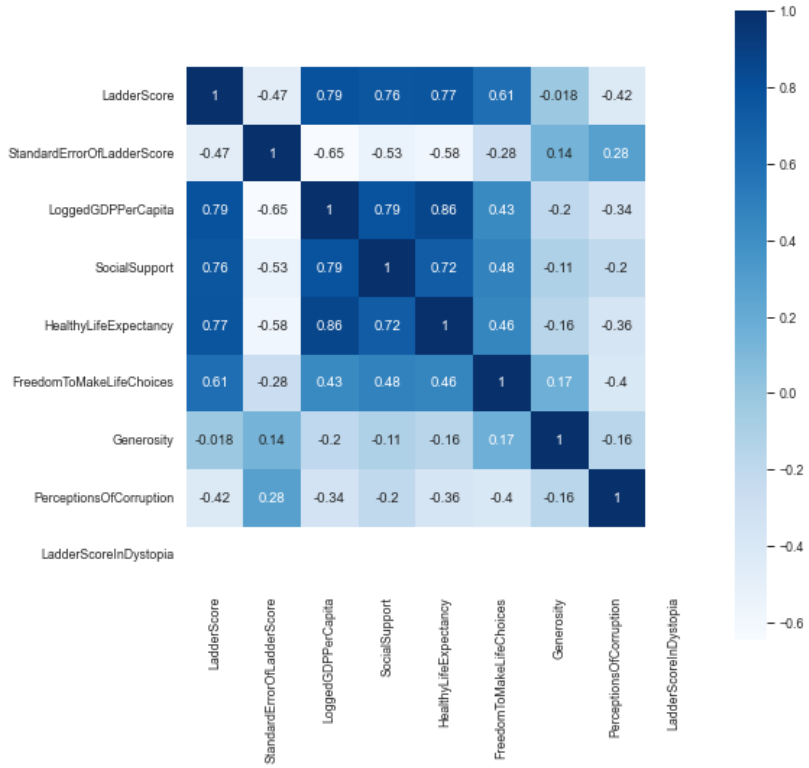
(圖 6)

經過(圖 6)對比發現，LadderScore 與 Generosity 之間的為弱相關關係，可以不考慮將該變量納入模型。當然，變量之間不存在線性關係並不代表不存在任何關係，可能是二次函數關係、對數關係等，所以一般還需要進行檢驗和變量轉換。

相關係數較大的是 Logged GDP per capita、Healthy life expectancy、Social support、Freedom to make life choices。

柒、 多重共線性判斷

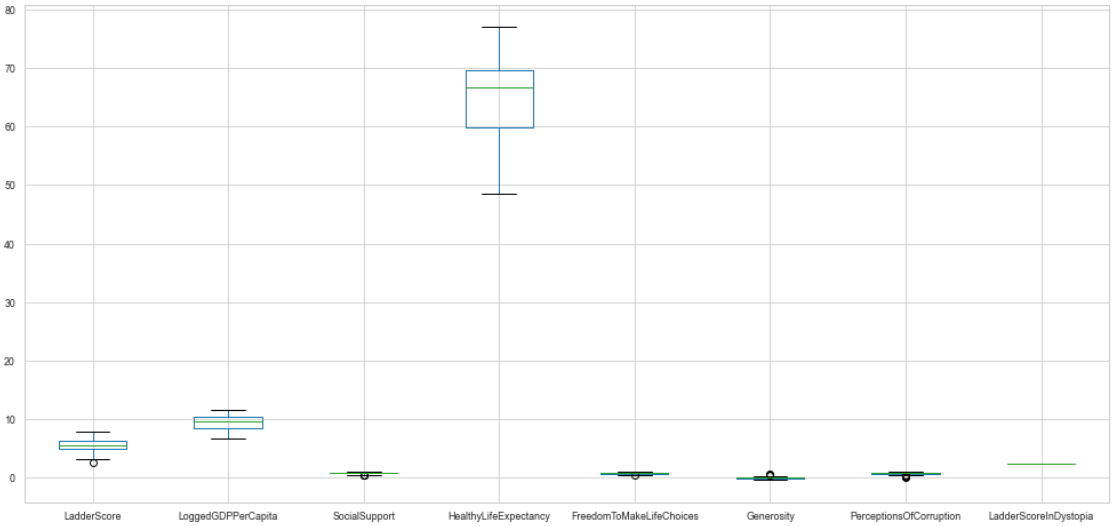
a. 相關性(使用 heatmap（圖 7))



（圖 7）

b. Boxplot(圖 8)

極端異常值，即超出四分位數差 3 倍距離的異常值，用實心點表示；較爲溫和的異常值，即處於 1.5 倍-3 倍四分位數差之間的異常值，用空心點表示。



（圖 8）

由上圖(圖 8)，可發現幾乎沒有太多的異常值，因此不用多做處理。

c. 一開始建立模型(圖 9)-模型顯著性和參數顯著性判斷
(使用切分資料做訓練(0.8)和預測(0.2))

OLS Regression Results

Dep. Variable:	LadderScore	R-squared:	0.752
Model:	OLS	Adj. R-squared:	0.739
Method:	Least Squares	F-statistic:	56.63
Date:	Tue, 13 Dec 2022	Prob (F-statistic):	1.14e-31
Time:	00:10:53	Log-Likelihood:	-97.408
No. Observations:	119	AIC:	208.8
Df Residuals:	112	BIC:	228.3
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.3658	0.108	-3.389	0.001	-0.580	-0.152
LoggedGDPPerCapita	0.2185	0.111	1.968	0.052	-0.002	0.438
SocialSupport	2.8474	0.799	3.565	0.001	1.265	4.430
HealthyLifeExpectancy	0.0424	0.016	2.689	0.008	0.011	0.074
FreedomToMakeLifeChoices	1.6991	0.568	2.990	0.003	0.573	2.825
Generosity	0.3891	0.358	1.088	0.279	-0.320	1.098
PerceptionsOfCorruption	-0.5927	0.333	-1.779	0.078	-1.253	0.067
LadderScoreInDystopia	-0.8889	0.262	-3.389	0.001	-1.409	-0.369

Omnibus:	7.976	Durbin-Watson:	2.318
Prob(Omnibus):	0.019	Jarque-Bera (JB):	7.660
Skew:	-0.592	Prob(JB):	0.0217
Kurtosis:	3.375	Cond. No.	7.23e+17

（圖 9）

[1] 標準誤差假設正確指定了誤差的共變異矩陣。
[2] 最小特徵值為 9.97e-31。 這可能表明存在強多重共線性問題或設計矩陣是奇特的。

```
[('Lagrange multiplier statistic', 13.146280708740417),
 ('p-value', 0.0686234890786861),
 ('f-value', 2.3182675255960543),
 ('f p-value', 0.037918093587370354)]
```

（圖 10）

=> 通過上面(圖 9、圖 10)結果我們清楚看到：
如果用 p-value 來看，p-value = 0.0686234890786861 大於 α = 0.05，因此 not reject H_0 。

但由 7 個回歸係數的 t 統計量 p 值除了 Generosity、PerceptionsOfCorruption、LoggedGDPPerCapita 其餘的都<0.05，說明剩下的迴歸係數較顯著，因此需要 Drop 掉 P 值最大的 Generosity 重新建模，來處理他造成的強多重共線性問題。

d. 第一次重新建模(Drop : Generosity)

OLS Regression Results

Dep. Variable:	LadderScore	R-squared:	0.752
Model:	OLS	Adj. R-squared:	0.739
Method:	Least Squares	F-statistic:	56.63
Date:	Tue, 13 Dec 2022	Prob (F-statistic):	1.14e-31
Time:	00:10:53	Log-Likelihood:	-97.408
No. Observations:	119	AIC:	208.8
Df Residuals:	112	BIC:	228.3
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.3658	0.108	-3.389	0.001	-0.580	-0.152
LoggedGDPPerCapita	0.2185	0.111	1.968	0.052	-0.002	0.438
SocialSupport	2.8474	0.799	3.565	0.001	1.265	4.430
HealthyLifeExpectancy	0.0424	0.016	2.689	0.008	0.011	0.074
FreedomToMakeLifeChoices	1.6991	0.568	2.990	0.003	0.573	2.825
Generosity	0.3891	0.358	1.088	0.279	-0.320	1.098
PerceptionsOfCorruption	-0.5927	0.333	-1.779	0.078	-1.253	0.067
LadderScoreInDystopia	-0.8889	0.262	-3.389	0.001	-1.409	-0.369

Omnibus:	7.976	Durbin-Watson:	2.318
Prob(Omnibus):	0.019	Jarque-Bera (JB):	7.660
Skew:	-0.592	Prob(JB):	0.0217
Kurtosis:	3.375	Cond. No.	7.23e+17

（圖 11）

[1] 標準誤差假設正確指定了誤差的共變異矩陣。

[2] 最小特徵值為 9.98e-31。 這可能表明存在強多重共線性問題或設計矩陣是奇特的。

```
[('Lagrange multiplier statistic', 12.658862253462022),
 ('p-value', 0.048784687876246846),
 ('f-value', 2.690306808735979),
 ('f p-value', 0.02457361031936004)]
```

（圖 12）

=> 通過上面結果我們清楚看到：

如果用 p-value 來看，The p-value = 0.048784687876246846 小於 α = 0.05，因此 reject H_0 。

剩下 6 個回歸係數的 t 統計量 p 值除了 LoggedGDPPerCapita、PerceptionsOfCorruption、其餘的都<0.05，說明剩下的迴歸係數較顯著，因此需要 Drop 掉 P 值最大的 LoggedGDPPerCapita 繼續重新建模。

e. 第二次重新建模(Drop ： Generosity、LoggedGDPPerCapita)

```
fit3 = smf.ols('LadderScore~ SocialSupport + HealthyLifeExpectancy + \
                FreedomToMakeLifeChoices + PerceptionsOfCorruption + \
                LadderScoreInDystopia'
                ,data = Train.drop('LoggedGDPPerCapita', axis = 1)).fit()

fit3.summary()
```

OLS Regression Results

Dep. Variable:	LadderScore	**R-squared:**	0.742
Model:	OLS	**Adj. R-squared:**	0.733
Method:	Least Squares	**F-statistic:**	82.02
Date:	Mon, 13 Jun 2022	**Prob (F-statistic):**	1.22e-32
Time:	17:09:59	**Log-Likelihood:**	-99.749
No. Observations:	119	**AIC:**	209.5
Df Residuals:	114	**BIC:**	223.4
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.3122	0.106	-2.951	0.004	-0.522	-0.103
SocialSupport	3.7287	0.665	5.610	0.000	2.412	5.045
HealthyLifeExpectancy	0.0607	0.012	5.195	0.000	0.038	0.084
FreedomToMakeLifeChoices	1.5791	0.550	2.872	0.005	0.490	2.668
PerceptionsOfCorruption	-0.7708	0.326	-2.366	0.020	-1.416	-0.125
LadderScoreInDystopia	-0.7586	0.257	-2.951	0.004	-1.268	-0.249

Omnibus:	6.199	**Durbin-Watson:**	2.225
Prob(Omnibus):	0.045	**Jarque-Bera (JB):**	5.773
Skew:	-0.524	**Prob(JB):**	0.0558
Kurtosis:	3.257	**Cond. No.**	7.15e+17

（圖 13）

[1] 標準誤差假設正確指定了誤差的共變異矩陣。
[2] 最小特徵值為 9.98e-31。 這可能表明存在強多重共線性問題或設計矩陣是奇特的。

```
[('Lagrange multiplier statistic', 13.21341343978292),
 ('p-value', 0.02145885728468162),
 ('f-value', 3.559830175817713),
 ('f p-value', 0.008941679488950775)]
```

（圖 14）

=> 如果用 p-value 來看，The p-value = 0.02145885728468162 小於 α = 0.05，因此 reject $H0$ 。

=> 通過模型反饋的結果我們可知，模型是通過顯著性檢驗的，即剩下 6 個迴歸係數的 t 統計量 p 值遠遠小於 0.05 這個閾值的，說明需要拒絕原假設(即認為模型的所有回歸係數都不全為 0)。

捌、 強影響點診斷

a. VIF :如果自變量 X 與其他自變量共線性強，那麼回歸方程的 R2 就會較高，導致 VIF 也高。一般，有自變量 VIF 值大於 10，則說明存在嚴重多重共線性，可以選擇刪除該變量或者用其他類似但 VIF 低的變量代替。

可以看到 AT 的方差膨脹因子大於 10，可以刪除該變量。

b. 多重共線性的處理方法 :

多重共線性對於線性回歸是種災難，並且我們不可能完全消除，而只能利用一些方法來減輕它的影響。對於多重共線性的處理方式，有以下幾種思路：

(1)提前篩選變量：可以利用相關檢驗來或變量聚類的方法。注意：決策樹和隨機森林也可以作為提前篩選變量的方法，但是它們對於多重共線性幫助不大，因為如果按照特徵重要性排序，共線性的變量很可能都排在前面。

(2)子集選擇：包括逐步回歸和最優子集法。因為該方法是貪婪算法，理論上大部分情況有效，實際中需要結合第一種方法。

(3)收縮方法：正則化方法，包括嶺回和 LASSO 回歸。LASSO 回歸可以實現篩選變量的功能。

(4)維數縮減：包括主成分迴歸(PCR)和偏最小平方法迴歸(PLS)方法。

c. 檢查 VIF 並確認刪除異常值是否會對模型造成不好的影響:

	VIF Factor	feature
0	0.000000	Intercept
1	5.104890	LoggedGDPPerCapita
2	2.972200	SocialSupport
3	4.099348	HealthyLifeExpectancy
4	1.585807	FreedomToMakeLifeChoices
5	1.180982	Generosity
6	1.367122	PerceptionsOfCorruption
7	0.000000	LadderScoreInDystopia

（圖 15）

```
異常值數量的比例 =  0.04697986577181208
```

（圖 16）

=> 結果顯示(圖 15、16),所有自變量的 VIF 均低於 10,異常值比例也極低，說明自變量之間並不存在多重共線性的隱患。

d. 第三次重新建模(刪除異常值):

OLS Regression Results

Dep. Variable:	LadderScore	R-squared:	0.760
Model:	OLS	Adj. R-squared:	0.743
Method:	Least Squares	F-statistic:	43.38
Date:	Tue, 13 Dec 2022	Prob (F-statistic):	1.91e-23
Time:	00:10:54	Log-Likelihood:	-41.998
No. Observations:	89	AIC:	98.00
Df Residuals:	82	BIC:	115.4
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.1028	0.108	-0.953	0.343	-0.317	0.112
LoggedGDPPerCapita	0.3202	0.089	3.580	0.001	0.142	0.498
SocialSupport	1.9628	0.762	2.577	0.012	0.448	3.478
HealthyLifeExpectancy	0.0026	0.014	0.187	0.852	-0.025	0.030
FreedomToMakeLifeChoices	2.6519	0.531	4.992	0.000	1.595	3.709
Generosity	0.3410	0.323	1.054	0.295	-0.303	0.985
PerceptionsOfCorruption	-0.6485	0.270	-2.403	0.018	-1.185	-0.112
LadderScoreInDystopia	-0.2498	0.262	-0.953	0.343	-0.771	0.272

Omnibus:	2.560	Durbin-Watson:	1.915
Prob(Omnibus):	0.278	Jarque-Bera (JB):	2.553
Skew:	-0.375	Prob(JB):	0.279
Kurtosis:	2.645	Cond. No.	2.81e+17

(圖 17)

[1] 標準誤差假設正確指定了誤差的協方差矩陣。

[2] 最小特徵值為 1.5e-29。 這可能表明有強多重共線性問題或設計矩陣是奇特的。

=> 通過模型反饋(圖 17)的結果我們可知，模型跟想像中的一樣，造成 P value 變大的情況，因此其實不需要做刪除異常值的第三次重新建模。

e. 比較

	Model	MSE	RMSE	R2_score	AIC	BIC
0	一開始建立模型	0.21186796214884354	0.4602911710524584	0.7547648927487713	208.81508934894586	228.26895380072656
1	第一次重新建模	0.2122926966195769	0.46075231591341664	0.7542732667264735	208.06519458409602	224.73993554276518
2	第二次重新建模	0.2676812064772173	0.5173791708961787	0.6901616048326304	209.49718702713366	223.3928044926913
3	第三次重新建模	0.30662190512232085	0.553734507794413	0.6069360030908807	97.99555405655951	115.4160086446845

(圖 18)

=> 通過上圖(圖 18)結果我們可以發現經過第三次重新的建模，雖然 AIC、BIC 也下降了，但反而造成 MSE、RMSE 的提升，R2_score 的下降。這不是我們所期望的。

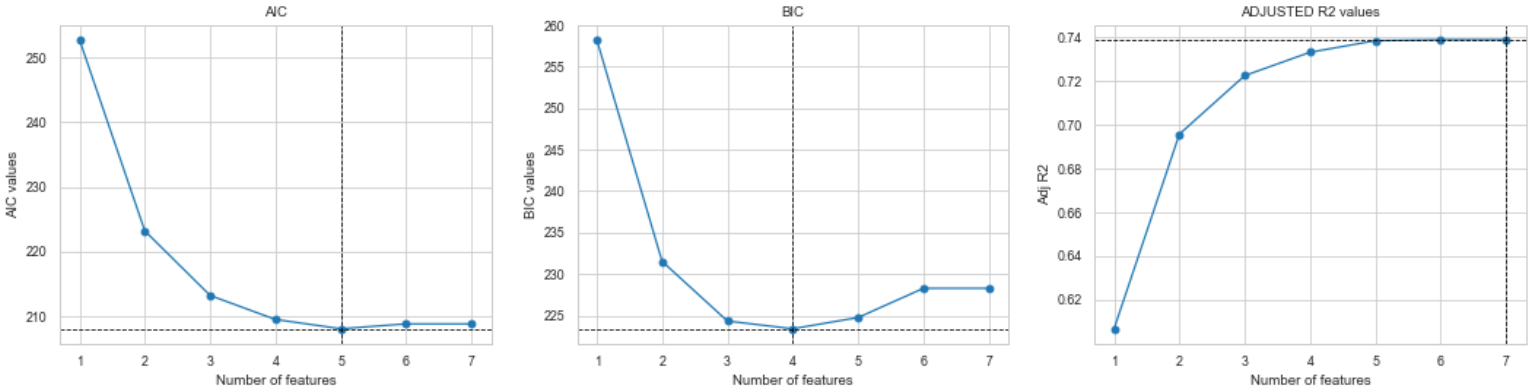
玖、 Model 之間的比較

a. Best subset selection

(Using AIC & BIC & Adj r squared)

	numb_features	RSS	R_squared	AIC	BIC	adj_r2	features
0	1	56.332067	0.610053	252.712085	258.270332	0.606721	(V1,)
1	2	43.220108	0.700818	223.182071	231.519442	0.695660	(V1, V4)
2	3	39.081067	0.729470	213.202611	224.319105	0.722412	(V2, V3, V4)
3	4	37.251543	0.742134	209.497187	223.392804	0.733086	(V2, V3, V4, V6)
4	5	36.192543	0.749465	208.065195	224.739936	0.738379	(V1, V2, V3, V4, V6)
5	6	35.814327	0.752083	208.815089	228.268954	0.738802	(V1, V2, V3, V4, V5, V6)
6	7	35.814327	0.752083	208.815089	228.268954	0.738802	(V1, V2, V3, V4, V5, V6, V7)

(圖 19)



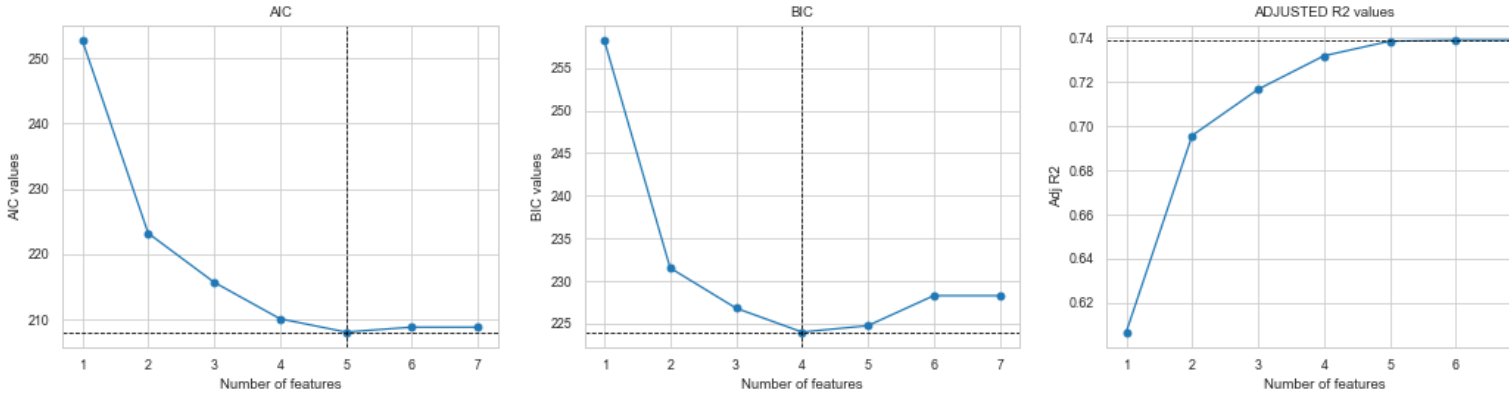
(圖 20)

b. Forward selection

(Using AIC & BIC & Adj r squared)

	numb_features	R_squared	AIC	BIC	adj_r2	features
0	1	0.610053	252.712085	258.270332	0.606721	[V1]
1	2	0.700818	223.182071	231.519442	0.695660	[V1, V4]
2	3	0.723775	215.681701	226.798195	0.716569	[V1, V4, V2]
3	4	0.740831	210.097262	223.992879	0.731737	[V1, V4, V2, V3]
4	5	0.749465	208.065195	224.739936	0.738379	[V1, V4, V2, V3, V6]
5	6	0.752083	208.815089	228.268954	0.738802	[V1, V4, V2, V3, V6, V5]
6	7	0.752083	208.815089	228.268954	0.738802	[V1, V4, V2, V3, V6, V5, V7]

(圖 21)



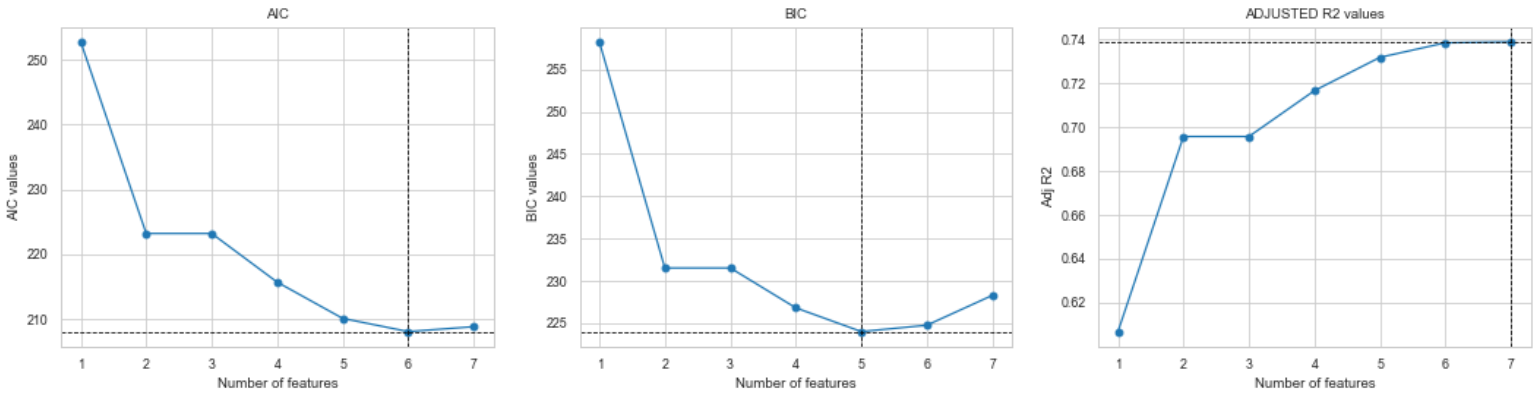
（圖 22）

c. Backward selection

(Using AIC & BIC & Adj r squared)

	numb_features	R_squared	AIC	BIC	adj_r2	features
0	7	0.752083	208.815089	228.268954	0.738802	[V1, V2, V3, V4, V5, V6, V7]
1	6	0.749465	208.065195	224.739936	0.738379	[V1, V2, V3, V4, V6, V7]
2	5	0.740831	210.097262	223.992879	0.731737	[V1, V2, V3, V4, V7]
3	4	0.723775	215.681701	226.798195	0.716569	[V1, V2, V4, V7]
4	3	0.700818	223.182071	231.519442	0.695660	[V1, V4, V7]
5	2	0.700818	223.182071	231.519442	0.695660	[V1, V4]
6	1	0.610053	252.712085	258.270332	0.606721	[V1]

（圖 23）



（圖 24）

d. 挑選:

	Selection of Model	AIC	BIC	Adj R Squared
0	bsb	208.065194584096	223.39280449269134	0.7388017805999297
1	fwd	208.06519458409602	223.99287934062863	0.7388017805999297
2	bwd	208.06519458409596	223.9928793406286	0.7388017805999297

(圖 25)

=> 通過上圖(圖 25)結果我們可以發現使用 Best subset selection 所選的 AIC、BIC 是最小的，調整後的 R^2(Adj R^2)皆相同。

因此，分別對 AIC、BIC 重新建模，再與前面的模型做比較。

```
Using AIC criterion :
Best subset is model having features is 5 : 'V1', 'V2', 'V3', 'V4', 'V6'
Forward is model having features is 5 : 'V1', 'V4', 'V2', 'V3', 'V6'
Backward is model having features is 6 : 'V1', 'V2', 'V3', 'V4', 'V6', 'V7'
Best subset and Forward are the same.

Using BIC criterion :
Best subset is model having features is 4 : 'V2', 'V3', 'V4', 'V6'
Forward is model having features is 4 : 'V1', 'V4', 'V2', 'V3
Backward is model having features is 5 : 'V1', 'V2', 'V3', 'V4', 'V7'
They are not the same.

Using Adj R Squared criterion :
Best subset is model having features is 7 : 'V1', 'V2', 'V3', 'V4', 'V5', 'V6', 'V7'
Forward is model having features is 7 : 'V1', 'V4', 'V2', 'V3', 'V6', 'V5', 'V7'
Backward is model having features is 7 : 'V1', 'V2', 'V3', 'V4', 'V5', 'V6', 'V7'
They are the same.
```

(圖 26)

=> 上圖(圖 26)結果是分別所選出的 feature 值。

e. **第四次重新建模 Best subset using AIC**

(Drop : Generosity、LadderScoreInDystopia):

OLS Regression Results

Dep. Variable:	LadderScore	**R-squared:**	0.749
Model:	OLS	**Adj. R-squared:**	0.738
Method:	Least Squares	**F-statistic:**	67.61
Date:	Tue, 13 Dec 2022	**Prob (F-statistic):**	2.35e-32
Time:	00:11:05	**Log-Likelihood:**	-98.033
No. Observations:	119	**AIC:**	208.1
Df Residuals:	113	**BIC:**	224.7
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-2.3808	0.734	-3.244	0.002	-3.835	-0.927
LoggedGDPPerCapita	0.1996	0.110	1.818	0.072	-0.018	0.417
SocialSupport	2.9098	0.797	3.649	0.000	1.330	4.490
HealthyLifeExpectancy	0.0413	0.016	2.623	0.010	0.010	0.072
FreedomToMakeLifeChoices	1.8121	0.559	3.241	0.002	0.704	2.920
PerceptionsOfCorruption	-0.6498	0.329	-1.973	0.051	-1.302	0.003

Omnibus:	8.416	**Durbin-Watson:**	2.282
Prob(Omnibus):	0.015	**Jarque-Bera (JB):**	8.142
Skew:	-0.606	**Prob(JB):**	0.0171
Kurtosis:	3.419	**Cond. No.**	1.21e+03

(圖 27)

=> 通過上面結果(圖 27)我們清楚看到：

剩下 5 個回歸係數的 t 統計量 p 值除了 LoggedGDPPerCapita 其餘的都 <0.05，說明剩下的迴歸係數較顯著。

f. **第五次重新建模 Best subset using BIC**

(Drop : LoggedGDPPerCapita、Generosity、LadderScoreInDystopia):

OLS Regression Results

Dep. Variable:	LadderScore	R-squared:	0.742
Model:	OLS	Adj. R-squared:	0.733
Method:	Least Squares	F-statistic:	82.02
Date:	Tue, 13 Dec 2022	Prob (F-statistic):	1.22e-32
Time:	00:11:05	Log-Likelihood:	-99.749
No. Observations:	119	AIC:	209.5
Df Residuals:	114	BIC:	223.4
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-2.1557	0.731	-2.951	0.004	-3.603	-0.708
SocialSupport	3.7287	0.665	5.610	0.000	2.412	5.045
HealthyLifeExpectancy	0.0607	0.012	5.195	0.000	0.038	0.084
FreedomToMakeLifeChoices	1.5791	0.550	2.872	0.005	0.490	2.668
PerceptionsOfCorruption	-0.7708	0.326	-2.366	0.020	-1.416	-0.125

Omnibus:	6.199	Durbin-Watson:	2.225
Prob(Omnibus):	0.045	Jarque-Bera (JB):	5.773
Skew:	-0.524	Prob(JB):	0.0558
Kurtosis:	3.257	Cond. No.	1.08e+03

(圖 28)

=> 通過上面結果(圖 28)我們清楚看到：

剩下 4 個回歸係數的 t 統計量 p 值皆<0.05，說明這些迴歸係數都較為顯著。

g. **所有模型比較**

Model	使用的資料
一開始建立模型	切分資料做訓練(0.8)和預測(0.2)
第一次重新建模	Drop : Generosity
第二次重新建模	Drop : Generosity、LoggedGDPPerCapita
第三次重新建模	刪除異常值
第四次重新建模 Best subset using BIC	Drop : Generosity、LadderScoreInDystopia
第五次重新建模 Best subset using BIC	Drop : LoggedGDPPerCapita、Generosity、LadderScoreInDystopia

(圖 29)

	Model	MSE	RMSE	R2_score	AIC	BIC
		Min	Min	Max		
0	一開始建立模型	0.21186796214884354	0.4602911710524584	0.7547648927487713	208.81508934894586	228.26895380072656
1	第一次重新建模	0.2122926966195769	0.46075231591341664	0.7542732667264735	208.06519458409602	224.73993554276518
2	第二次重新建模	0.2676812064772173	0.5173791708961787	0.6901616048326304	209.49718702713366	223.3928044926913
3	第三次重新建模	0.30662190512232085	0.553734507794413	0.6069360030908807	97.99555405655951	115.4160086446845
4	第四次重新建模	0.21229269661957717	0.4607523159134169	0.7542732667264732	208.06519458409602	224.73993554276518
5	第五次重新建模	0.2676812064772171	0.5173791708961786	0.6901616048326308	209.4971870271337	223.39280449269134
					Min	

（圖 30）

=> 通過上面結果(圖 30)我們發現一開始建立的模型是 MSE、RMSE 有最小值，R2_score 有最大值與第四次重新建模為 AIC、BIC 有最小值。因此，我們可以分別對他們做殘差診斷(於後面壹拾壹的部分)來比較何者配適的為最佳。

壹拾、 ANOVA Table

原始建模

	df	sum_sq	mean_sq	F	PR(>F)
LoggedGDPPerCapita	1.0	39.479700	39.479700	210.213134	8.165657e-27
SocialSupport	1.0	4.534560	4.534560	24.144663	3.305200e-06
HealthyLifeExpectancy	1.0	0.930714	0.930714	4.955666	2.814180e-02
FreedomToMakeLifeChoices	1.0	6.595014	6.595014	35.115733	3.992819e-08
Generosity	1.0	0.380142	0.380142	2.024097	1.577851e-01
PerceptionsOfCorruption	1.0	2.201754	2.201754	11.723433	8.813401e-04
LadderScoreInDystopia	1.0	0.434498	0.434498	2.313524	1.312588e-01
Residual	105.0	19.719836	0.187808	NaN	NaN

（圖 31）

第一次重新建模

	df	sum_sq	mean_sq	F	PR(>F)
LoggedGDPPerCapita	1.0	88.128918	88.128918	275.155235	4.653080e-32
SocialSupport	1.0	8.650668	8.650668	27.009031	9.084253e-07
HealthyLifeExpectancy	1.0	4.597534	4.597534	14.354375	2.444682e-04
FreedomToMakeLifeChoices	1.0	5.644001	5.644001	17.621643	5.394397e-05
PerceptionsOfCorruption	1.0	1.247321	1.247321	3.894374	5.088927e-02
LadderScoreInDystopia	1.0	0.207096	0.207096	0.646593	4.230218e-01
Residual	113.0	36.192543	0.320288	NaN	NaN

（圖 32）

第二次重新建模

	df	sum_sq	mean_sq	F	PR(>F)
SocialSupport	1.0	85.702821	85.702821	262.274276	2.449714e-31
HealthyLifeExpectancy	1.0	14.892938	14.892938	45.576500	6.457762e-10
FreedomToMakeLifeChoices	1.0	4.784159	4.784159	14.640847	2.128068e-04
PerceptionsOfCorruption	1.0	1.829523	1.829523	5.598846	1.965953e-02
LadderScoreInDystopia	1.0	0.214852	0.214852	0.657506	4.191314e-01
Residual	114.0	37.251543	0.326768	NaN	NaN

（圖 33）

第三次重新建模

	df	sum_sq	mean_sq	F	PR(>F)
LoggedGDPPerCapita	1.0	39.479700	39.479700	210.213134	8.165657e-27
SocialSupport	1.0	4.534560	4.534560	24.144663	3.305200e-06
HealthyLifeExpectancy	1.0	0.930714	0.930714	4.955666	2.814180e-02
FreedomToMakeLifeChoices	1.0	6.595014	6.595014	35.115733	3.992819e-08
Generosity	1.0	0.380142	0.380142	2.024097	1.577851e-01
PerceptionsOfCorruption	1.0	2.201754	2.201754	11.723433	8.813401e-04
LadderScoreInDystopia	1.0	0.434498	0.434498	2.313524	1.312588e-01
Residual	105.0	19.719836	0.187808	NaN	NaN

（圖 34）

第四次重新建模

	df	sum_sq	mean_sq	F	PR(>F)
LoggedGDPPerCapita	1.0	88.128918	88.128918	275.155235	4.653080e-32
SocialSupport	1.0	8.650668	8.650668	27.009031	9.084253e-07
HealthyLifeExpectancy	1.0	4.597534	4.597534	14.354375	2.444682e-04
FreedomToMakeLifeChoices	1.0	5.644001	5.644001	17.621643	5.394397e-05
PerceptionsOfCorruption	1.0	1.247321	1.247321	3.894374	5.088927e-02
Residual	113.0	36.192543	0.320288	NaN	NaN

（圖 35）

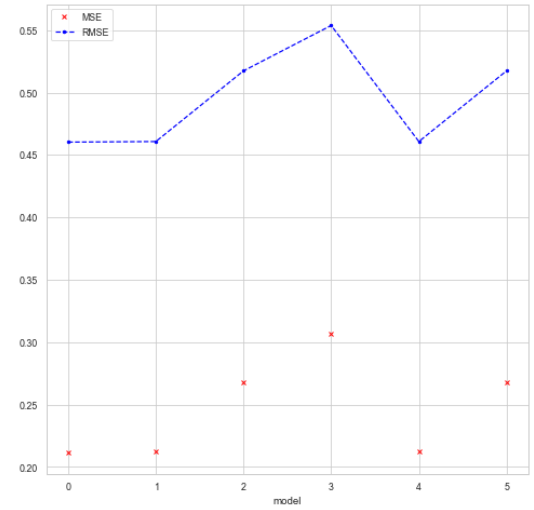
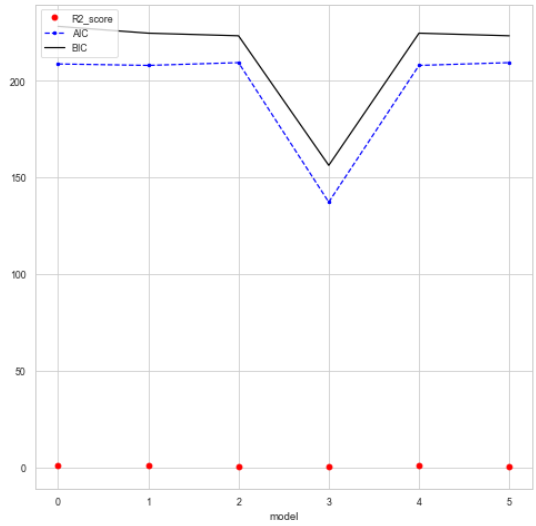
第五次重新建模

	df	sum_sq	mean_sq	F	PR(>F)
LoggedGDPPerCapita	1.0	39.479700	39.479700	210.213134	8.165657e-27
SocialSupport	1.0	4.534560	4.534560	24.144663	3.305200e-06
HealthyLifeExpectancy	1.0	0.930714	0.930714	4.955666	2.814180e-02
FreedomToMakeLifeChoices	1.0	6.595014	6.595014	35.115733	3.992819e-08
Generosity	1.0	0.380142	0.380142	2.024097	1.577851e-01
PerceptionsOfCorruption	1.0	2.201754	2.201754	11.723433	8.813401e-04
LadderScoreInDystopia	1.0	0.434498	0.434498	2.313524	1.312588e-01
Residual	105.0	19.719836	0.187808	NaN	NaN

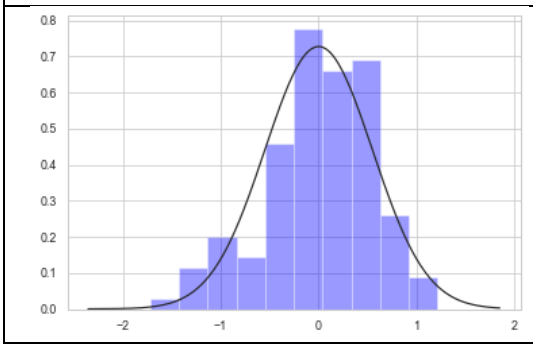
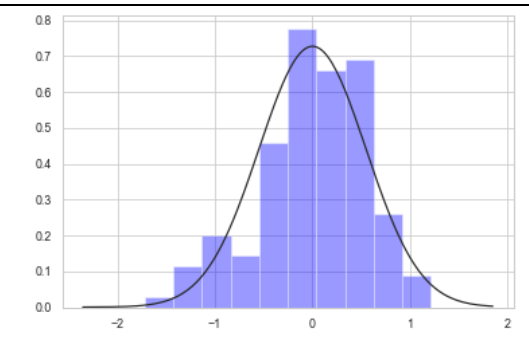
（圖 36）

壹拾壹、 殘差診斷

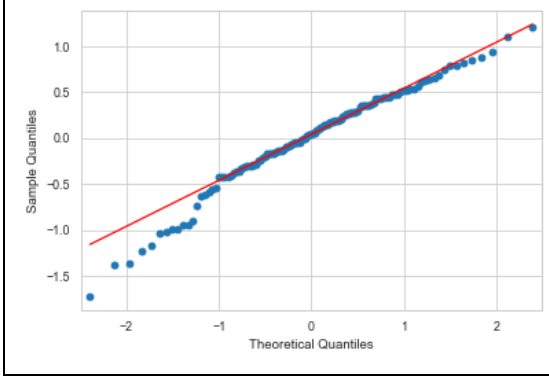
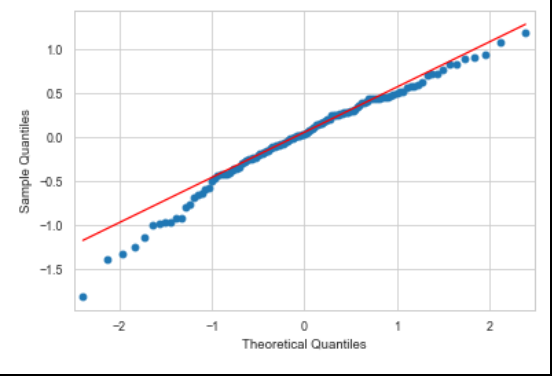
a. 模型可視化比較

MSE、RMSE、R2_score	AIC、BIC
	

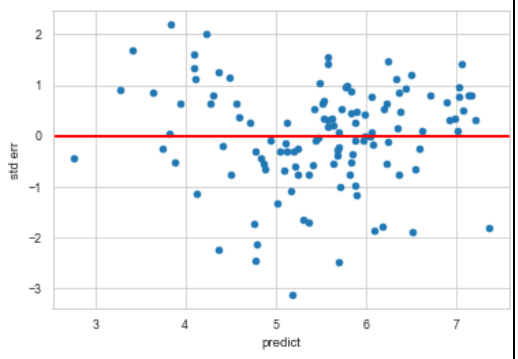
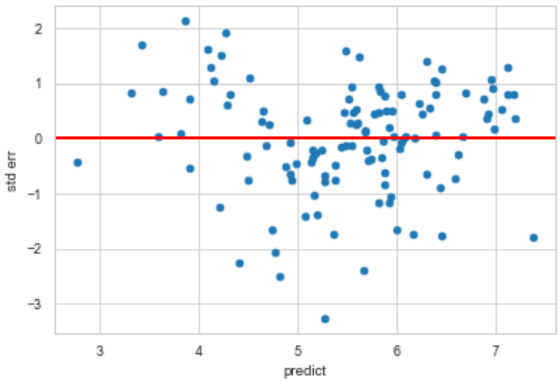
b. KS 檢驗

一開始建立的模型	第四次重新建模
	

c. QQ 圖

一開始建立的模型	第四次重新建模
	

壹拾貳、 標準化殘差檢定圖

一開始建立的模型	第四次重新建模
	

壹拾參、 結論

透過前面的所有分析後，KS 檢驗其實看不太出來，幾乎都接近常態，而 QQ 圖可以看到一開始建立的模型會比較貼近一些，再來標準化殘差檢定圖也差不多，所以獲得以下結論：綜合所有的分析後，一開始建立的模型會是 MSE、RMSE 較為好的，而第四次重新建模會是 AIC、BIC。

但如果是以 R2_score 為主的會是一開始建立的複迴歸模型為最佳。

壹拾肆、 參考資料

1. Kaggle : World Happiness Expanatory Data Analysis
   https://www.kaggle.com/datasets/mathurinache/world-happiness-report-20152021?select=2021.csv
2. 維基百科 : 世界幸福報告
3. 世界幸福報告官網