Spring 2020, NCUE

Applied Statistics I – Final Project Report

Topic

Diamond Price Forecast Analysis and Regression Modeling

Member

S0722020 數三甲 梁家瑀

摘要

本研究欲使用 SPSS 軟體分析鑽石相關數據，並藉由統計圖表及相應方法得出的結果加以解讀來研究對鑽石價格產生影響的因素為何並藉此建構出價格與鑽石其他屬性之關係，最終目的在於利用這些資料去配置出可預測鑽石價格之複迴歸模型。

1. 資料介紹

1.1 資料來源及簡述

這份資料來自於 kaggle 平台，共有 10 個維度，共計 53,940 筆資料。

資料來源網址：https://www.kaggle.com/shivam2503/diamonds

1.2 欲處理之議題

此數據集包含近 54,000 顆鑽石的價格及其相關屬性，目的在於利用其它屬性來對鑽石價格加以預測。

2. 資料變數介紹

本數據集含有 10 個維度的資料，因此我們在此設定應變數Y為鑽石價格，而其餘自變數$X_{1\sim9}$分別代表鑽石的重量(單位：克拉)、切工品質、顏色、淨度、長度(單位：公釐)、寬度(單位：公釐)、高度(單位：公釐)、深度比例及檯面寬比例，由圖 1 所示。

- Y ：鑽石價格(單位：美金)
- X_1：鑽石重量(單位：克拉)
- X_2：切工品質(分為五類：Ideal, Premium, Very Good, Good, Fair)
- X_3：鑽石顏色(由好到壞分為七類： D～J)
- X_4：鑽石淨度
  (由好到壞分為八類： IF, VVS1, VVS2, VS1, VS2, SI1, SI2, I1)
- X_5：鑽石長度 x (單位：公釐(mm))
- X_6：鑽石寬度 y (單位：公釐(mm))
- X_7：鑽石高度 z (單位：公釐(mm))
- X_8：鑽石深度比例
- X_9：鑽石檯面寬比例

圖 1、資料變數設定

2.1 虛擬變數設定

本研究在資料處理步驟上針對類別型資料X_2(切工品質)、X_3(鑽石顏色)、X_4(鑽石淨度)進行虛擬變數的轉換，如公式(1)、公式(2)、公式(3)所示。

$$X_2 \rightarrow (X_{21}, X_{22}, X_{23}, X_{24}) = \begin{cases} (1,0,0,0) & if\ quality\ is\ "Ideal" \\ (0,1,0,0) & if\ quality\ is\ "Premium" \\ (0,0,1,0) & if\ quality\ is\ "VeryGood" \\ (0,0,0,1) & if\ quality\ is\ "Good" \\ (0,0,0,0) & if\ quality\ is\ "Fair" \end{cases} \quad (1)$$

$$X_3 \to (X_{31}, X_{32}, X_{33}, X_{34}, X_{35}, X_{36}) = \begin{cases} (1,0,0,0,0,0) & if\ color\ is\ "D" \\ (0,1,0,0,0,0) & if\ color\ is\ "E" \\ (0,0,1,0,0,0) & if\ color\ is\ "F" \\ (0,0,0,1,0,0) & if\ color\ is\ "G" \\ (0,0,0,0,1,0) & if\ color\ is\ "H" \\ (0,0,0,0,0,1) & if\ color\ is\ "I" \\ (0,0,0,0,0,0) & if\ color\ is\ "J" \end{cases} \quad (2)$$

$$X_4 \to (X_{41}, X_{42}, X_{43}, X_{44}, X_{45}, X_{46}, X_{47}) = \begin{cases} (1,0,0,0,0,0,0) & if\ clarity\ is\ "IF" \\ (0,1,0,0,0,0,0) & if\ clarity\ is\ "VVS1" \\ (0,0,1,0,0,0,0) & if\ clarity\ is\ "VVS2" \\ (0,0,0,1,0,0,0) & if\ clarity\ is\ "VS1" \\ (0,0,0,0,1,0,0) & if\ clarity\ is\ "VS2" \\ (0,0,0,0,0,1,0) & if\ clarity\ is\ "SI1" \\ (0,0,0,0,0,0,1) & if\ clarity\ is\ "SI2" \\ (0,0,0,0,0,0,0) & if\ clarity\ is\ "I1" \end{cases} \quad (3)$$

3. 敘述統計

3.1 變數轉換

首先觀察鑽石價格Y的直方圖，如圖 2 所示。其圖形過於偏左且資料差距過大，因此將Y取底數為 10 之對數以方便分析其迴歸模型，其中圖 3 為對數價格$log_{10}Y$的直方圖。

在此提供Y的 P-P 圖和$log_{10}Y$的 P-P 圖以呈現轉換前後的常態分配情形，分別為圖 4、圖 5 所示。相較之下，取對數後的Y較接近常態分配。

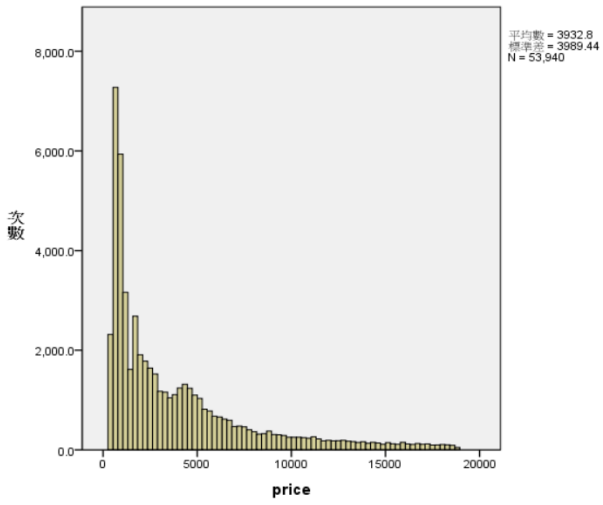

圖 2、Y的直方圖

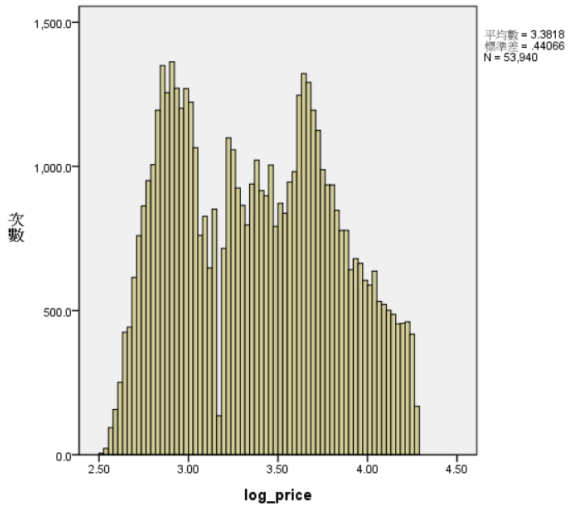

圖 3、$log_{10}Y$的直方圖

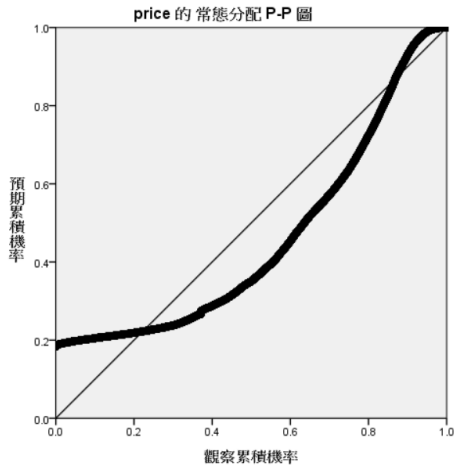

圖 4、Y的 P-P 圖

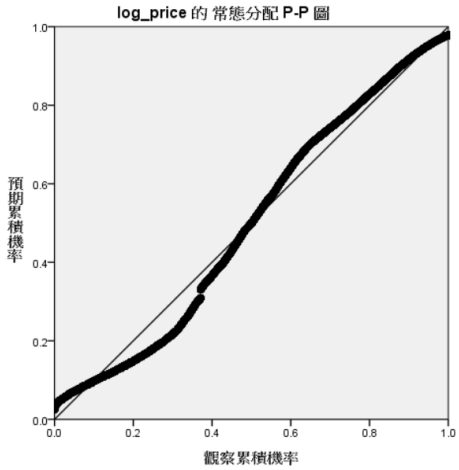

圖 5、$log_{10}Y$的 P-P 圖

3.2 連續型資料 – 相關性

透過圖 6 呈現出各連續型資料X_1、$X_{5\sim9}$與對數價格$log_{10}Y$及各連續型資料之間的相關係數。由圖可知深度比例(X_8)和對數價格之間的線性關係並不顯著，而重量(X_1)、長度(X_5)、寬度(X_6)與高度(X_7)和價格有著高度的線性關係，其中這 4 個自變數之間也存在著高度的線性關係，因此可能會有共線性的問題在此當中。

相關

		log_price	carat	depth	table	x	y	z
log_price	Pearson 相關	1	.920**	.001	.158**	.961**	.938**	.938**
	顯著性 (雙尾)		.000	.841	.000	.000	.000	.000
	個數	53932	53932	53932	53932	53932	53932	53932
carat	Pearson 相關	.920**	1	.028**	.182**	.978**	.954**	.956**
	顯著性 (雙尾)	.000		.000	.000	.000	.000	.000
	個數	53932	53932	53932	53932	53932	53932	53932
depth	Pearson 相關	.001	.028**	1	-.296**	-.025**	-.029**	.095**
	顯著性 (雙尾)	.841	.000		.000	.000	.000	.000
	個數	53932	53932	53932	53932	53932	53932	53932
table	Pearson 相關	.158**	.182**	-.296**	1	.196**	.185**	.152**
	顯著性 (雙尾)	.000	.000	.000		.000	.000	.000
	個數	53932	53932	53932	53932	53932	53932	53932
x	Pearson 相關	.961**	.978**	-.025**	.196**	1	.975**	.971**
	顯著性 (雙尾)	.000	.000	.000	.000		.000	.000
	個數	53932	53932	53932	53932	53932	53932	53932
y	Pearson 相關	.938**	.954**	-.029**	.185**	.975**	1	.952**
	顯著性 (雙尾)	.000	.000	.000	.000	.000		.000
	個數	53932	53932	53932	53932	53932	53932	53932
z	Pearson 相關	.938**	.956**	.095**	.152**	.971**	.952**	1
	顯著性 (雙尾)	.000	.000	.000	.000	.000	.000	
	個數	53932	53932	53932	53932	53932	53932	53932

**. 在顯著水準為0.01時 (雙尾)，相關顯著。

圖 6、各連續型資料間與對數價格$log_{10}Y$的相關係數表

3.3 類別型資料 – 分布情形

透過圖 7、圖 8 及圖 9 分別為類別型資料$X_{2\sim4}$與對數價格$log_{10}Y$的盒形圖。由圖可知類別型資料和對數價格的相關性都不高，且在圖形上的走勢有著和一般認知相違背的行為存在，如「顏色等級越好，價格越低」、「淨度等級越高，價格越低」等，這個部分目前較難透過圖形觀察出整體的趨勢，在之後的模型配置過程會在對其做解釋。

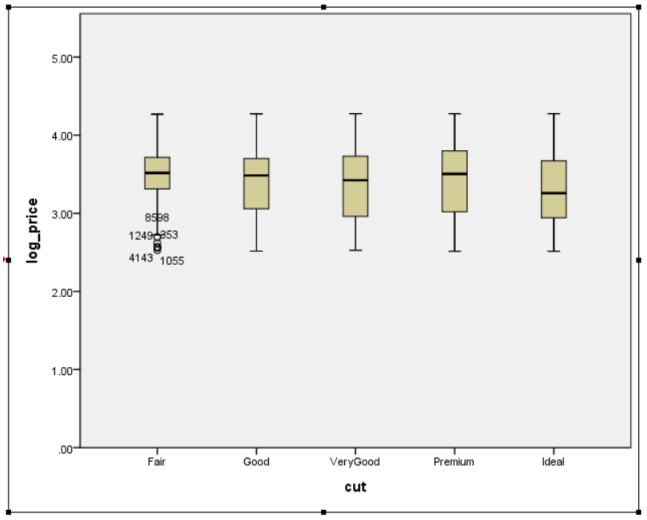

圖 7、切工品質(X_2)的盒形圖

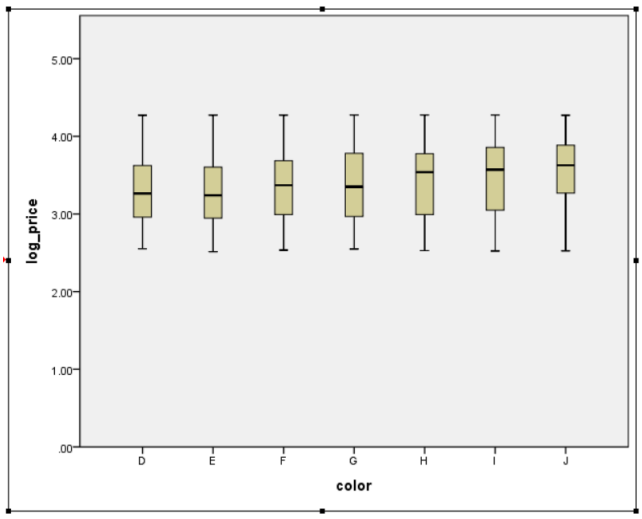

圖 8、鑽石顏色(X_3)的盒形圖

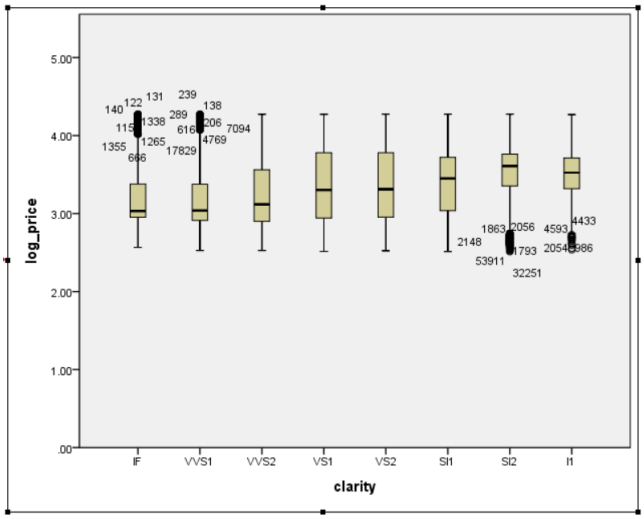

圖 9、鑽石淨度(X_4)的盒形圖

4. 推論統計

4.1 迴歸方法

在配置初始模型時，本研究利用以下三種方法配置出複回歸模型之雛型，針對每種方式給出的模型做為參考依據來決定所欲配置之模型所需納入的變數為何。

■ **Backward Selection(向後法)**

利用 Backward Selection 所得出之模型摘要如圖 10 所示。由圖可知，藉由向後法得出之模型選入了所有的自變數，而這些自變數對$log_{10}Y$的解釋力高達 97%。

模式摘要

模式	R	R 平方	調過後的 R 平方	估計的標準誤
1	.985[a]	.970	.970	.07625

a. 預測變數:(常數), Clarity_SI2, Color_I, Cut_VeryGood, depth, Clarity_IF, Clarity_VVS1, Color_D, Clarity_VVS2, Cut_Good, Color_H, Clarity_VS1, Color_F, y, table, Color_E, Clarity_VS2, Cut_Premium, Color_G, Cut_Ideal, carat, Clarity_SI1, z, x

圖 10、向後法所選入之變數及模式摘要

■ **Forward Selection (向前法)**

利用 Forward Selection 所得出之模型摘要如圖 11、圖 12 所示。由圖可知，藉由向前法得出之模型選入與向後法一致。

選入/刪除的變數[a]

模式	選入的變數	刪除的變數	方法
1	x	.	向前選擇法 (準則:F-選入的機率 <= .050)
2	Clarity_SI2	.	向前選擇法 (準則:F-選入的機率 <= .050)
3	Clarity_SI1	.	向前選擇法 (準則:F-選入的機率 <= .050)
4	Color_I	.	向前選擇法 (準則:F-選入的機率 <= .050)
5	carat	.	向前選擇法 (準則:F-選入的機率 <= .050)
6	depth	.	向前選擇法 (準則:F-選入的機率 <= .050)
7	Color_H	.	向前選擇法 (準則:F-選入的機率 <= .050)
8	Clarity_VVS1	.	向前選擇法 (準則:F-選入的機率 <= .050)
9	Clarity_IF	.	向前選擇法 (準則:F-選入的機率 <= .050)
10	Clarity_VVS2	.	向前選擇法 (準則:F-選入的機率 <= .050)
11	Color_D	.	向前選擇法 (準則:F-選入的機率 <= .050)
12	Color_E	.	向前選擇法 (準則:F-選入的機率 <= .050)
13	Color_F	.	向前選擇法 (準則:F-選入的機率 <= .050)
14	Color_G	.	向前選擇法 (準則:F-選入的機率 <= .050)
15	Clarity_VS1	.	向前選擇法 (準則:F-選入的機率 <= .050)
16	Clarity_VS2	.	向前選擇法 (準則:F-選入的機率 <= .050)
17	Cut_Ideal	.	向前選擇法 (準則:F-選入的機率 <= .050)
18	table	.	向前選擇法 (準則:F-選入的機率 <= .050)
19	Cut_VeryGood	.	向前選擇法 (準則:F-選入的機率 <= .050)
20	Cut_Premium	.	向前選擇法 (準則:F-選入的機率 <= .050)
21	Cut_Good	.	向前選擇法 (準則:F-選入的機率 <= .050)
22	y	.	向前選擇法 (準則:F-選入的機率 <= .050)
23	z	.	向前選擇法 (準則:F-選入的機率 <= .050)

a. 依變數: log_price

圖 11、向前法所選入之變數

模式摘要

模式	R	R 平方	調過後的 R 平方	估計的標準誤
1	.958[a]	.918	.918	.12635
2	.963[b]	.926	.926	.11950
3	.965[c]	.932	.932	.11521
4	.968[d]	.936	.936	.11125
5	.970[e]	.940	.940	.10800
6	.971[f]	.942	.942	.10592
7	.972[g]	.944	.944	.10408
8	.973[h]	.946	.946	.10227
9	.974[i]	.948	.948	.10003
10	.975[j]	.951	.951	.09747
11	.976[k]	.953	.953	.09575
12	.977[l]	.954	.954	.09436
13	.978[m]	.956	.956	.09263
14	.980[n]	.961	.961	.08724
15	.981[o]	.962	.962	.08602
16	.985[p]	.969	.969	.07721
17	.985[q]	.969	.969	.07702
18	.985[r]	.970	.970	.07688
19	.985[s]	.970	.970	.07674
20	.985[t]	.970	.970	.07663
21	.985[u]	.970	.970	.07640
22	.985[v]	.970	.970	.07630
23	.985[w]	.970	.970	.07625

圖 12、向前法之模式摘要

■ **Forward Stepwise Selection (逐步迴歸向前法)**

利用 Forward Stepwise Selection 所得出之模型摘要如圖 13、圖 14 所示。由圖可知，藉由逐步迴歸向前法得出之模型選入與向後法和向前法一致。

選入/刪除的變數[a]

模式	選入的變數	刪除的變數	方法
1	x	.	逐步迴歸分析法 (準則:F-選入的機率 <= .050，F-刪除的機率 >= .100)。
2	Clarity_SI2	.	逐步迴歸分析法 (準則:F-選入的機率 <= .050，F-刪除的機率 >= .100)。
3	Clarity_SI1	.	逐步迴歸分析法 (準則:F-選入的機率 <= .050，F-刪除的機率 >= .100)。
4	Color_I	.	逐步迴歸分析法 (準則:F-選入的機率 <= .050，F-刪除的機率 >= .100)。
5	carat	.	逐步迴歸分析法 (準則:F-選入的機率 <= .050，F-刪除的機率 >= .100)。
6	depth	.	逐步迴歸分析法 (準則:F-選入的機率 <= .050，F-刪除的機率 >= .100)。
7	Color_H	.	逐步迴歸分析法 (準則:F-選入的機率 <= .050，F-刪除的機率 >= .100)。
8	Clarity_VVS1	.	逐步迴歸分析法 (準則:F-選入的機率 <= .050，F-刪除的機率 >= .100)。
9	Clarity_IF	.	逐步迴歸分析法 (準則:F-選入的機率 <= .050，F-刪除的機率 >= .100)。
10	Clarity_VVS2	.	逐步迴歸分析法 (準則:F-選入的機率 <= .050，F-刪除的機率 >= .100)。
11	Color_D	.	逐步迴歸分析法 (準則:F-選入的機率 <= .050，F-刪除的機率 >= .100)。
12	Color_E	.	逐步迴歸分析法 (準則:F-選入的機率 <= .050，F-刪除的機率 >= .100)。
13	Color_F	.	逐步迴歸分析法 (準則:F-選入的機率 <= .050，F-刪除的機率 >= .100)。
14	Color_G	.	逐步迴歸分析法 (準則:F-選入的機率 <= .050，F-刪除的機率 >= .100)。
15	Clarity_VS1	.	逐步迴歸分析法 (準則:F-選入的機率 <= .050，F-刪除的機率 >= .100)。
16	Clarity_VS2	.	逐步迴歸分析法 (準則:F-選入的機率 <= .050，F-刪除的機率 >= .100)。
17	Cut_Ideal	.	逐步迴歸分析法 (準則:F-選入的機率 <= .050，F-刪除的機率 >= .100)。
18	table	.	逐步迴歸分析法 (準則:F-選入的機率 <= .050，F-刪除的機率 >= .100)。
19	Cut_VeryGood	.	逐步迴歸分析法 (準則:F-選入的機率 <= .050，F-刪除的機率 >= .100)。
20	Cut_Premium	.	逐步迴歸分析法 (準則:F-選入的機率 <= .050，F-刪除的機率 >= .100)。
21	Cut_Good	.	逐步迴歸分析法 (準則:F-選入的機率 <= .050，F-刪除的機率 >= .100)。
22	y	.	逐步迴歸分析法 (準則:F-選入的機率 <= .050，F-刪除的機率 >= .100)。
23	z	.	逐步迴歸分析法 (準則:F-選入的機率 <= .050，F-刪除的機率 >= .100)。

a. 依變數: log_price

圖 13、逐步迴歸向前法所選入之變數

模式摘要

模式	R	R 平方	調過後的 R 平方	估計的標準誤
1	.958[a]	.918	.918	.12635
2	.963[b]	.926	.926	.11950
3	.965[c]	.932	.932	.11521
4	.968[d]	.936	.936	.11125
5	.970[e]	.940	.940	.10800
6	.971[f]	.942	.942	.10592
7	.972[g]	.944	.944	.10408
8	.973[h]	.946	.946	.10227
9	.974[i]	.948	.948	.10003
10	.975[j]	.951	.951	.09747
11	.976[k]	.953	.953	.09575
12	.977[l]	.954	.954	.09436
13	.978[m]	.956	.956	.09263
14	.980[n]	.961	.961	.08724
15	.981[o]	.962	.962	.08602
16	.985[p]	.969	.969	.07721
17	.985[q]	.969	.969	.07702
18	.985[r]	.970	.970	.07688
19	.985[s]	.970	.970	.07674
20	.985[t]	.970	.970	.07663
21	.985[u]	.970	.970	.07640
22	.985[v]	.970	.970	.07630
23	.985[w]	.970	.970	.07625

圖 14、逐步迴歸向前法之模式摘要

4.2 變數挑選

藉由上述三種方式得到之模型皆相同，因為逐步迴歸向前法能夠清楚地表示各變數被選入前後的差異，因此最後本研究由逐步迴歸向前法得出之結果去做篩選變數。

由圖 14 可知，在模型選入變數執行到第 17 次左右時，R 平方的改變量已無較大改變，這表示對於第 16 號模型來說，當模型選入了前 16 個變數後再選入後面的 17～23 號變數已經對模型沒有太大的幫助。因此可考慮將 R 平方改變量小於 0.001 的變數剔除，由表 1 所示，這樣的作法表示我們可以利用更少的成本得到具相同解釋力的模型。

變數名稱	R 平方改變量
X_{32}(顏色"E")	0.001
X_{44}(淨度"VS1")	0.001
X_9(檯面寬比例)	0.001
X_{21}(切工品質"Ideal")	< 0.001
X_{23} (切工品質"VeryGood")	< 0.001
X_{22} (切工品質"Premium")	< 0.001
X_{24}(切工品質"Good")	< 0.001
X_6(寬度)	< 0.001
X_7(高度)	< 0.001

表 1、R 平方改變量小於 0.001 之變數

由於變數X_{32}(顏色"E")和X_{44}(淨度"$VS1$")為虛擬變數的其中一部分，因此在顏色與淨度的其它虛

擬變數沒有被剔除下，我們也不將其剔除。而由切工品質所設定的虛擬變數整組皆被選入剔除候補，因此選擇將其拿掉，最終剔除的變數有X_2(切工品質)、X_6(寬度)、X_7(高度)、X_9(檯面寬比例)。

4.3 模型選擇

將部分變數剔除後再建立一次迴歸模型，得到結果如圖 15 所示。由圖可得，儘管前一步驟刪除許多變數但剩餘變數對模型的解釋力僅下降0.001，因此這個剔除的動作的確大幅度降低成本，且對模型解釋力無太大影響，因此選定由剩餘的這些變數(X_1、$X_{31\sim36}$、$X_{41\sim47}$、X_5、X_8)配出之模型作為本研究之初始模型。

模式摘要

模式	R	R 平方	調過後的 R 平方	估計的標準誤
1	.985[a]	.969	.969	.07721

a. 預測變數:(常數), Clarity_SI2, Color_I, depth, Clarity_IF, Clarity_VVS1, Color_D, Clarity_VVS2, Color_H, Clarity_VS1, Color_F, carat, Color_E, Clarity_VS2, Color_G, Clarity_SI1, x

圖 15、經變數篩選後之初始模型摘要

5. 模型分析

5.1 整體模型之顯著性

選定初始模型後，接著分析此模型是否有較大的問題存在。根據圖 16 得到的初始模型之 ANOVA 表可得知在顯著性那欄為0.000 < 0.05，因此拒絕由假設檢定所選定之H_0：係數皆為零的假設，也就是說在此階段選入之變數與$log_{10}Y$的確有顯著的關係。

Anova[a]

模式		平方和	df	平均平方和	F	顯著性
1	迴歸	10152.364	16	634.523	106451.865	.000[b]
	殘差	321.416	53923	.006		
	總數	10473.781	53939			

a. 依變數: log_price

b. 預測變數:(常數), Clarity_SI2, Color_I, depth, Clarity_IF, Clarity_VVS1, Color_D, Clarity_VVS2, Color_H, Clarity_VS1, Color_F, carat, Color_E, Clarity_VS2, Color_G, Clarity_SI1, x

圖 16、初始模型之 ANOVA 表

5.2 共線性診斷

接著透過統計量 VIF 值診斷此模型的變數之間是否存在共線性的問題，由圖 17 可以得知，有部分變數之 VIF 值超過 10 以上，因此此模型選入之變數的確存在共線性問題。

係數[a]

模式		未標準化係數		標準化係數	t	顯著性	B 的 95.0% 信賴區間		共線性統計量	
		B 之估計值	標準誤差	Beta 分配			下界	上界	允差	VIF
1	(常數)	-1.154	.018		-65.726	.000	-1.189	-1.120		
	carat	-.262	.003	-.282	-79.059	.000	-.269	-.256	.045	22.357
	depth	.020	.000	.065	83.008	.000	.020	.020	.928	1.078
	x	.530	.001	1.348	377.415	.000	.527	.532	.045	22.416
	Color_D	.228	.002	.172	127.782	.000	.225	.232	.316	3.166
	Color_E	.203	.002	.177	119.127	.000	.199	.206	.256	3.899
	Color_F	.189	.002	.163	111.220	.000	.185	.192	.264	3.786
	Color_G	.159	.002	.147	95.914	.000	.156	.163	.242	4.130
	Color_H	.116	.002	.095	67.950	.000	.112	.119	.294	3.407
	Color_I	.061	.002	.042	34.022	.000	.058	.065	.378	2.648
	Clarity_IF	.489	.003	.199	141.991	.000	.482	.496	.290	3.444
	Clarity_VVS1	.449	.003	.256	141.304	.000	.443	.456	.173	5.783
	Clarity_VVS2	.420	.003	.278	135.841	.000	.414	.426	.136	7.364
	Clarity_VS1	.367	.003	.299	122.161	.000	.361	.373	.095	10.509
	Clarity_VS2	.337	.003	.321	114.074	.000	.332	.343	.072	13.907
	Clarity_SI1	.275	.003	.267	93.306	.000	.269	.281	.069	14.405
	Clarity_SI2	.201	.003	.172	67.868	.000	.195	.207	.089	11.238

a. 依變數: log_price

圖 17、模型係數表與 VIF 值

5.3 殘差分析

再來進行殘差分析，由圖 18 可見此模型之標準化殘差 P-P 圖呈現 Heavy Tail 的狀態，這表示殘差有特大值或特小值存在，而此現象在圖 19 的標準化殘差散佈圖中也可發現左上角有很異常的資料點存在，因此本研究會先著手於離群值的分析而非模型的配置。

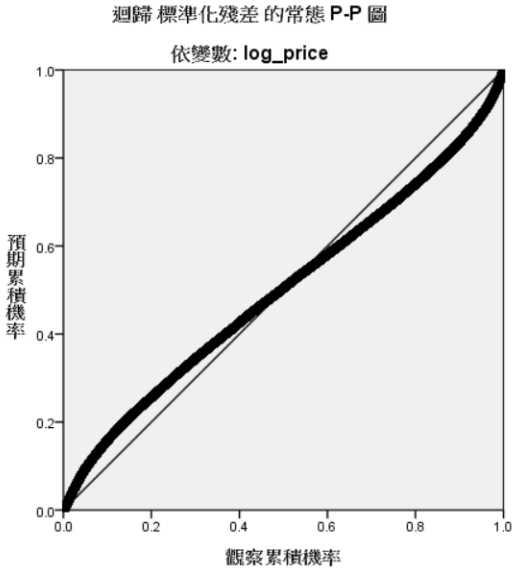

圖 18、標準化殘差之 P-P 圖

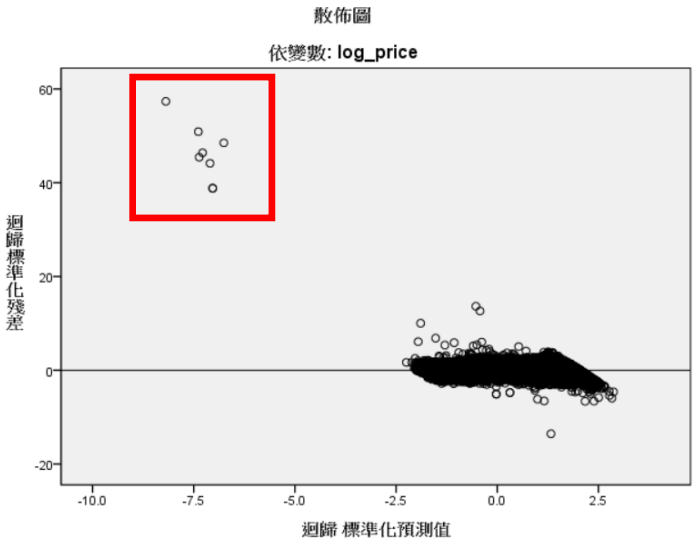

圖 19、標準化殘差散佈圖

6. 模型矯正

6.1 共線性問題

將選入之變數再次使用逐步迴歸向前法分析，得出每次選入變數之 VIF 值之差異，如圖 20、圖 21 和圖 22 所示。當模型選入變數執行至第 4 次與第 5 次時，可由圖 20 所知，當X_1(重量)被選入時，它與X_5(長度)產生共線性問題。同樣在執行至第 15 次與第 16 次時，可由圖 21 和圖 22 所知，當X_{45}(淨度"VS2")被選入時，它與X_{44}(淨度"VS1")、X_{46}(淨度"SI1")和X_{47}(淨度"SI2")也會產生共線性問題。

係數[a]

模式		未標準化係數		標準化係數	t	顯著性	共線性統計量	
		B 之估計值	標準誤差	Beta 分配			允差	VIF
1	(常數)	1.225	.003		432.489	.000		
	x	.376	.000	.958	775.955	.000	1.000	1.000
2	(常數)	1.185	.003		435.046	.000		
	x	.387	.000	.984	811.400	.000	.927	1.079
	Clarity_SI2	-.113	.001	-.097	-79.780	.000	.927	1.079
3	(常數)	1.181	.003		449.607	.000		
	x	.391	.000	.996	840.986	.000	.903	1.107
	Clarity_SI2	-.140	.001	-.119	-97.712	.000	.849	1.178
	Clarity_SI1	-.078	.001	-.075	-63.938	.000	.910	1.098
4	(常數)	1.167	.003		458.418	.000		
	x	.396	.000	1.007	870.328	.000	.882	1.134
	Clarity_SI2	-.144	.001	-.123	-103.871	.000	.847	1.180
	Clarity_SI1	-.078	.001	-.076	-66.756	.000	.910	1.098
	Color_I	-.101	.002	-.069	-62.556	.000	.977	1.024
5	(常數)	.769	.007		104.519	.000		
	x	.501	.002	1.275	265.987	.000	.048	20.623
	Clarity_SI2	-.144	.001	-.123	-106.968	.000	.847	1.180
	Clarity_SI1	-.082	.001	-.080	-72.278	.000	.907	1.103
	Color_I	-.093	.002	-.064	-59.308	.000	.969	1.031
	carat	-.255	.004	-.275	-57.393	.000	.049	20.568

圖 20、模型 4 至模型 5 之 VIF 值改變量

15	(常數)	-.714	.019		-37.426	.000		
	x	.533	.002	1.357	341.063	.000	.045	22.405
	Clarity_SI2	-.114	.001	-.097	-94.820	.000	.673	1.485
	Clarity_SI1	-.043	.001	-.042	-39.961	.000	.652	1.533
	Color_I	.059	.002	.040	29.473	.000	.378	2.648
	carat	-.282	.004	-.303	-76.445	.000	.045	22.295
	depth	.018	.000	.059	67.325	.000	.933	1.072
	Color_H	.111	.002	.091	58.469	.000	.294	3.405
	Clarity_VVS1	.128	.002	.073	77.864	.000	.803	1.245
	Clarity_IF	.167	.002	.068	75.962	.000	.880	1.137
	Clarity_VVS2	.100	.001	.066	69.155	.000	.776	1.288
	Color_D	.225	.002	.169	113.103	.000	.316	3.165
	Color_E	.199	.002	.174	104.717	.000	.257	3.897
	Color_F	.184	.002	.159	97.202	.000	.264	3.783
	Color_G	.155	.002	.143	83.959	.000	.242	4.128
	Clarity_VS1	.048	.001	.039	39.239	.000	.716	1.396

圖 21、模型 15 之 VIF 值

16	(常數)	-1.154	.018		-65.726	.000		
	x	.530	.001	1.348	377.415	.000	.045	22.416
	Clarity_SI2	.201	.003	.172	67.868	.000	.089	11.238
	Clarity_SI1	.275	.003	.267	93.306	.000	.069	14.405
	Color_I	.061	.002	.042	34.022	.000	.378	2.648
	carat	-.262	.003	-.282	-79.059	.000	.045	22.357
	depth	.020	.000	.065	83.008	.000	.928	1.078
	Color_H	.116	.002	.095	67.950	.000	.294	3.407
	Clarity_VVS1	.449	.003	.256	141.304	.000	.173	5.783
	Clarity_IF	.489	.003	.199	141.991	.000	.290	3.444
	Clarity_VVS2	.420	.003	.278	135.841	.000	.136	7.364
	Color_D	.228	.002	.172	127.782	.000	.316	3.166
	Color_E	.203	.002	.177	119.127	.000	.256	3.899
	Color_F	.189	.002	.163	111.220	.000	.264	3.786
	Color_G	.159	.002	.147	95.914	.000	.242	4.130
	Clarity_VS1	.367	.003	.299	122.161	.000	.095	10.509
	Clarity_VS2	.337	.003	.321	114.074	.000	.072	13.907

圖 22、模型 16 之 VIF 值

所謂共線性是指當 2 個以上的自變數互不獨立時，即這些變數具有共線性。若變數之間韓有共線性會使迴歸模型中存在著重複的自變數，使得模型的建構不準確。因此在本研究採取的作法為，透過

觀察簡單線性迴歸(如圖 23、圖 24)所配置出的 R 平方值(即該變數對$log_{10}Y$的解釋力)，選擇留下具較高解釋力的變數X_5(長度)，而X_1(重量)則剔除。

模式摘要[b]

模式	R	R 平方	調過後的 R 平方	估計的標準誤
1	.958[a]	.918	.918	.12635

a. 預測變數:(常數), x

b. 依變數: log_price

圖 23、X_5(長度)對$log_{10}Y$的解釋力

模式摘要[b]

模式	R	R 平方	調過後的 R 平方	估計的標準誤
1	.920[a]	.847	.847	.17249

a. 預測變數:(常數), carat

b. 依變數: log_price

圖 24、X_1(重量)對$log_{10}Y$的解釋力

而因為X_{45}(淨度"VS2")、X_{44}(淨度"VS1")、X_{46}(淨度"SI1")和X_{47}(淨度"SI2")為淨度的部分虛擬變數，因此不能將其所剔除，所以在此選擇保留這些虛擬變數。

6.2 離群值分析

6.2.1 殘差異常資料點

由圖 19 得知有部分資料點之標準化殘差過大，因此從資料中將其調出來觀察，如圖 25 所示。由圖可見這些資料的X_5(長度)皆為 0，因此懷疑這些資料為缺失項，而這部分也經 kaggle 平台上求證過，的確是資料缺失，於是選擇將這些資料剔除。

carat	cut	color	clarity	depth	table	price	x	y	z	index
1.07	5	3	SI2	61.6	56	4954	0	6.62	0	11183
1	3	5	VS2	63.3	53	5139	0	0	0	11964
1.14	1	4	VS1	57.5	67	6381	0	0	0	15952
1.56	5	4	VS2	62.2	54	12800	0	0	0	24521
1.2	4	1	VVS1	62.1	59	15686	0	0	0	26244
2.25	4	5	SI2	62.8	59	18034	0	0	0	27430
0.71	2	3	SI2	64.1	60	2130	0	0	0	49557
0.71	2	3	SI2	64.1	60	2130	0	0	0	49558

圖 25、殘差過大之資料點數據

6.2.2 具影響力之離群值

在剔除資料缺失的資料後，重新審視剩餘資料的標準化殘差散佈圖，如圖 26 所示，發現圖形右側存在線性的趨勢。因此在仔細對資料做檢查後發現有部分資料存在明顯不尋常之處，如「體質皆一致，價格卻差 5 倍」等。這種現象是由於資料上存在無法被量化的因素或是沒有被提供之資訊所導致，如依據產地不同、鑽石職人的輩分不同等因素所導致價格上會有異常。

在透過人工檢查後，將從 520 筆離群值資料剔除中明顯可看出異常之資料(共計 230 筆資料)，至於其它離群資料在相較之下沒有明顯不同處，因此未將其視為具影響的離群值

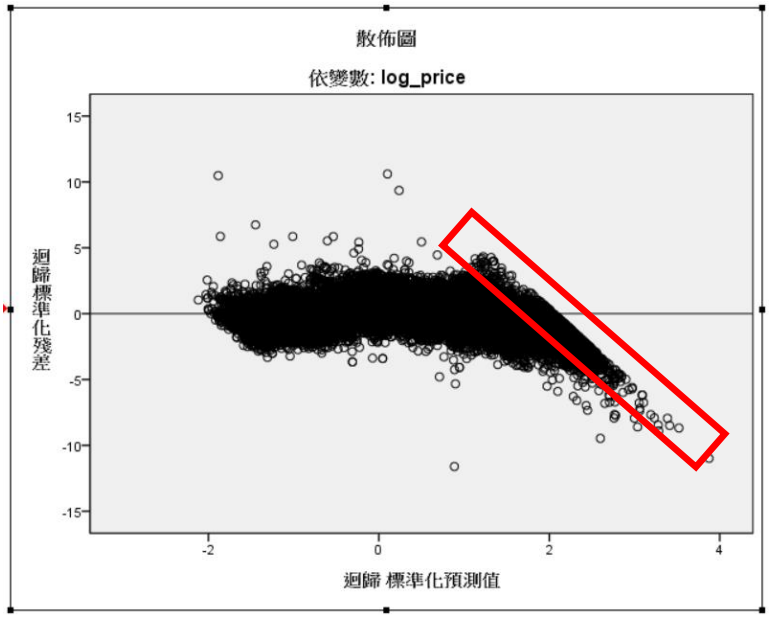

圖 26、剔除資料缺失後之標準化殘差散佈圖

6.3 矯正後模型

解決共線性問題與離群值問題後，重新配置模型得出之結果如圖 27 所示。由圖可見經過矯正後的模型其 R 平方往上提升至 0.974，較原先配出之初始模型要好一些。

模式摘要[b]

模式	R	R 平方	調過後的 R 平方	估計的標準誤	Durbin-Watson 檢定
1	.987[a]	.974	.974	.07046	1.897

a. 預測變數:(常數), Clarity_SI2, Color_F, depth, Clarity_IF, Clarity_VVS1, Color_I, Clarity_VVS2, Color_D, Clarity_VS1, Color_H, x, Color_E, Clarity_VS2, Color_G, Clarity_SI1

b. 依變數: log_price

圖 27、矯正後模型摘要

7. 模型檢視

配置出大致的模型後，本階段要對此模型作檢驗，檢查項目分別為自變數對應變數的顯著性、迴歸模型之四大假設以及高影響點的檢查。

7.1 整體模型之顯著性

根據圖 28 得到的初始模型之 ANOVA 表可得知在顯著性那欄為$0.000 < 0.05$，因此拒絕由假設檢定所設定之H_0：係數皆為零的假設，也就是說選入之變數與$log_{10}Y$的確有顯著的關係。

Anova[a]

模式		平方和	df	平均平方和	F	顯著性
1	迴歸	10072.296	15	671.486	135243.455	.000[b]
	殘差	266.552	53686	.005		
	總數	10338.848	53701			

a. 依變數: log_price

b. 預測變數:(常數), Clarity_SI2, Color_F, depth, Clarity_IF, Clarity_VVS1, Color_I, Clarity_VVS2, Color_D, Clarity_VS1, Color_H, x, Color_E, Clarity_VS2, Color_G, Clarity_SI1

圖 28、矯正後模型之 ANOVA 表

7.2 迴歸診斷

7.2.1 迴歸假設－獨立性

由圖 29 可知此模型的 DW 值為1.897(≈ 2)，這邊用到 Durbin-Watson Test[3]，此檢定法是說明當 DW 值接近 2 時，表示樣本中沒有檢測到自相關，因此此模型之殘差確實有符合獨立性假設。

模式摘要[b]

模式	R	R 平方	調過後的 R 平方	估計的標準誤	Durbin-Watson 檢定
1	.987[a]	.974	.974	.07046	1.897

a. 預測變數:(常數), Clarity_SI2, Color_F, depth, Clarity_IF, Clarity_VVS1, Color_I, Clarity_VVS2, Color_D, Clarity_VS1, Color_H, x, Color_E, Clarity_VS2, Color_G, Clarity_SI1

b. 依變數: log_price

圖 29、矯正後模型之 DW 值

7.2.2 迴歸假設－變異數同質性

由圖 30 可知刪去具影響力的資料點後，殘差散佈圖上仍能觀察到右側明顯有一部份資料呈現遞減的直線關係，這個部分以一個外行人去看資料也許很難觀察到其原因所在，可能需要有相關知識的人才能為此現象做解釋。但除紅色區域所為出之資料點，綠色部分所納入的資料點幾乎都平均散布在$e = 0$這條線的上下，所以大部分的樣本點之殘差的變異數都有符合同質性假設。若將來有機會學習到更進階的統計模型時，也許能再針對兩群資料各自配置適當的模型。

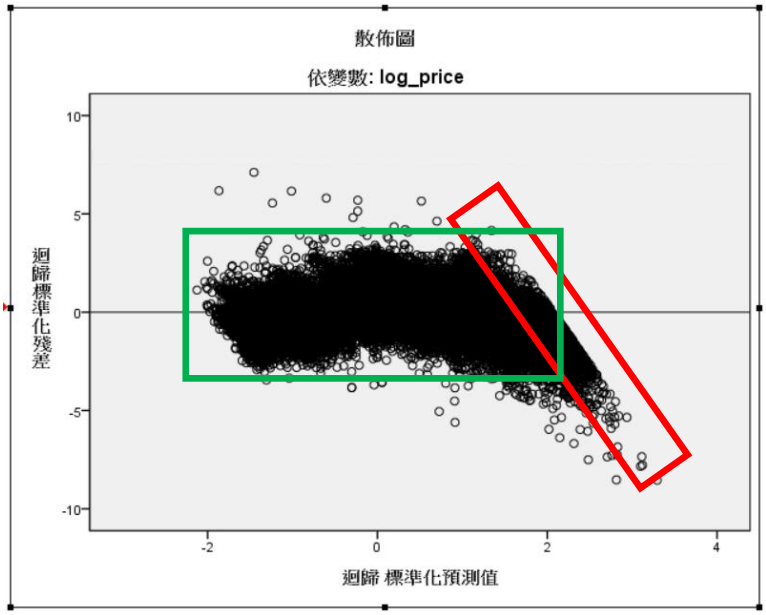

圖 30、矯正後模型之標準化殘差散佈圖

7.2.3 迴歸假設－常態性

由圖 31 可知此模型的標準化殘差 P-P 圖分布於 45 度線上，也就表示此模型之殘差確實有服從常態分配。

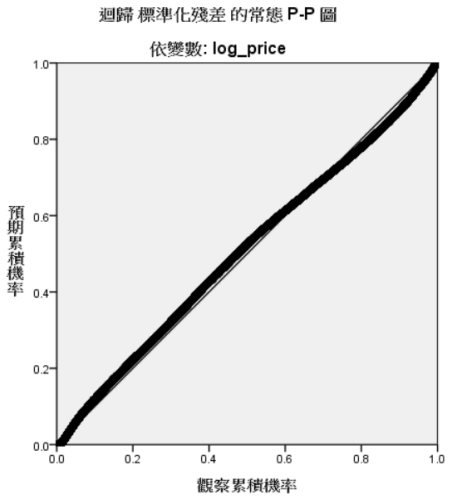

圖 31、矯正後模型之標準化殘差 P-P 圖

7.3 檢查是否存在高影響點

分別對資料點的 Cook's D 值遞減排序及遞增排序，分別由圖 32 及圖 33 所示。可得每筆資料的 Cook's D 值皆介於 $0 \sim 0.00792$ 之間，沒有資料的 Cook's D 值大於 1，於是我們得知本資料不存在高影響點。

COO_2	COO_2
.00792	.00000
.00761	.00000
.00741	.00000
.00712	.00000
.00685	.00000
.00611	.00000
.00579	.00000
.00556	.00000
.00538	.00000
.00498	.00000

圖 32 (左圖)、Cook's D 值遞減排序

圖 33 (右圖)、Cook's D 值遞增排序

7.4 最終模型

檢視完上述問題後，得知此模型的配適性是不錯的。因此藉由模型係數表(如圖 34)得到最終模型為公式(4)所示，其中根據不同的顏色(X_3)與淨度(X_4)，共計有56種模型。

$$log_{10}\hat{Y} = -0.531 + \alpha X_{3i} + \beta X_{4j} + 0.427X_5 + 0.016X_8 \ , i = 1,2,3 \dots,6 \ ; \mathrm{j} = 1,2,3 \dots,7 \qquad (4)$$

其中

$$\alpha = \begin{cases} 0.244 & if\ i = 1\ (i.e.\ color\ is\ \mathrm{D}) \\ 0.219 & if\ i = 2\ (i.e.\ color\ is\ \mathrm{E}) \\ 0.204 & if\ i = 3\ (i.e.\ color\ is\ \mathrm{F}) \\ 0.173 & if\ i = 4\ (i.e.\ color\ is\ \mathrm{G}) \\ 0.126 & if\ i = 5\ (i.e.\ color\ is\ \mathrm{H}) \\ 0.067 & if\ i = 6\ (i.e.\ color\ is\ \mathrm{I}) \\ 0 & if\ color\ is\ "J" \end{cases}, \beta = \begin{cases} 0.488 & if\ i = 1\ (i.e.\ clarity\ is\ "IF") \\ 0.451 & if\ i = 2\ (i.e.\ clarity\ is\ "VS1") \\ 0.425 & if\ i = 3\ (i.e.\ clarity\ is\ "VS2") \\ 0.375 & if\ i = 4\ (i.e.\ clarity\ is\ "VVS1") \\ 0.344 & if\ i = 5\ (i.e.\ clarity\ is\ "VVS2") \\ 0.284 & if\ i = 6\ (i.e.\ clarity\ is\ "SI1") \\ 0.205 & if\ i = 7\ (i.e.\ clarity\ is\ "SI2") \\ 0 & if\ clarity\ is\ "I1" \end{cases}$$

係數[a]

模式		未標準化係數		標準化係數	t	顯著性	共線性統計量	
		B 之估計值	標準誤差	Beta 分配			允差	VIF
1	(常數)	-.531	.014		-37.420	.000		
	depth	.016	.000	.051	73.187	.000	.982	1.018
	x	.427	.000	1.079	1366.962	.000	.771	1.296
	Color_D	.244	.002	.184	149.440	.000	.317	3.151
	Color_E	.219	.002	.193	141.012	.000	.257	3.887
	Color_F	.204	.002	.178	132.414	.000	.266	3.761
	Color_G	.173	.002	.161	114.462	.000	.243	4.113
	Color_H	.126	.002	.103	80.743	.000	.294	3.403
	Color_I	.067	.002	.046	40.344	.000	.377	2.651
	Clarity_IF	.488	.003	.198	153.214	.000	.286	3.494
	Clarity_VVS1	.451	.003	.259	153.368	.000	.168	5.936
	Clarity_VVS2	.425	.003	.283	148.422	.000	.132	7.569
	Clarity_VS1	.375	.003	.307	134.907	.000	.093	10.785
	Clarity_VS2	.344	.003	.329	125.658	.000	.070	14.288
	Clarity_SI1	.284	.003	.277	104.115	.000	.068	14.778
	Clarity_SI2	.205	.003	.175	74.429	.000	.087	11.479

a. 依變數: log_price

圖 34、最終模型之係數表

8. 模型評估與解釋

8.1 模型評估

8.1.1 連續型資料

由圖 35 得知長度(X_5)和對數價格$log_{10}Y$有著高度正相關，對應到模型的係數為0.427，其結果是吻合的。

接著看深度比例 (X_8)和對數價格$log_{10}Y$的線性相關並不高，因此對應到模型的係數為0.016 (≈ 0)，其結果也是吻合的。

相關

		log_price	depth	x
log_price	Pearson 相關	1	.000	.962**
	顯著性 (雙尾)		.943	.000
	個數	53702	53702	53702
depth	Pearson 相關	.000	1	-.026**
	顯著性 (雙尾)	.943		.000
	個數	53702	53702	53702
x	Pearson 相關	.962**	-.026**	1
	顯著性 (雙尾)	.000	.000	
	個數	53702	53702	53702

**. 在顯著水準為0.01時 (雙尾)，相關顯著。

圖 35、模型中連續型變數與應變數$log_{10}Y$之相關係數表

8.1.2 類別型資料 – X_3(顏色)

在前面敘述統計階段提到顏色與對數價格的走勢呈現「顏色越差，價格越高」的現象，而在模型配置完後，其係數對其顏色的關係呈現「顏色越好，相關程度越高」的狀態。這時本研究推測會產生這種矛盾之結果的理由有二：

(1) 資料筆數不均：

用軟體繪出各顏色類別個數之長條圖，如圖 36 所示。乍看之下資料筆數好像差不多，但由於本資料筆數高達 54,000 筆，所以直條圖呈現的個數每一單位就會差好幾千筆資料，因此才可能產出此違反以往認知的結果。

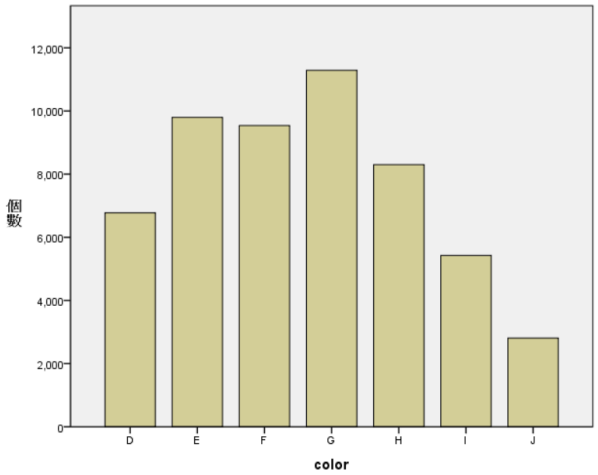

圖 36、X_3(顏色)的個數長條圖

(2) 對長度(X_5)的相關性呈遞減：

用軟體繪出各顏色類別和長度之相關係數圖，如圖 37 所示。可以得知，它們關係呈現「顏色越好，相關程度越高」，其中走勢為負相關。這表示顏色越差的鑽石挖出來的大小會越大，且加上長度(X_5)對$log_{10}Y$的解釋性高達 91.8%(如圖 23 所示)，因此鑽石價格很容易被長度(X_5)帶偏，所以顏色(X_3)對價格的影響力微乎其微。

		log_price	x	Color_D	Color_E	Color_F	Color_G	Color_H	Color_I
log_price	Pearson 相關	1	.962**	-.065**	-.095**	-.009*	.003	.054**	.076**
	顯著性 (雙尾)		.000	.000	.000	.029	.462	.000	.000
	個數	53702	53702	53702	53702	53702	53702	53702	53702
x	Pearson 相關	.962**	1	-.107**	-.134**	-.046**	-.022**	.094**	.145**
	顯著性 (雙尾)	.000		.000	.000	.000	.000	.000	.000
	個數	53702	53702	53702	53702	53702	53702	53702	53702

圖 37、X_3(顏色)對X_5(長度)之相關係數表

8.1.3 類別型資料－ X_4(淨度)

在前面敘述統計階段提到淨度與對數價格的走勢呈現「淨度等級越差，價格越高」的現象，會產生這種矛盾之結果的理由與上述兩個原因一致：

(1) 資料筆數不均：

用軟體繪出各淨度等級之個數的長條圖，如圖 38 所示。由於本資料筆數多達 54,000 筆，所以直條圖呈現的個數每一單位就會差好幾千筆資料，因此才可能產出此違反以往認知的結果。

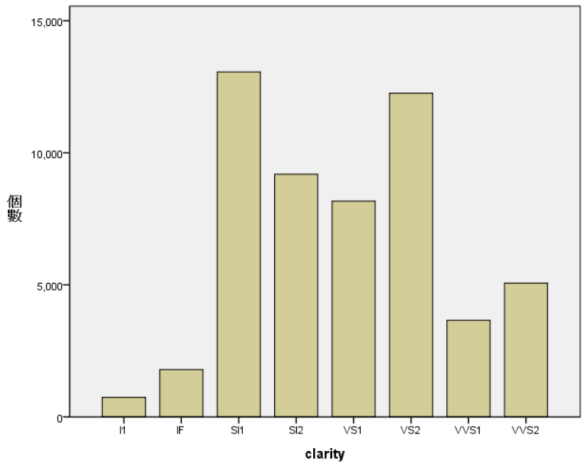

圖 38、X_4(淨度)的個數長條圖

(2) 對長度(X_5)的相關性呈遞減：

用軟體繪出各淨度等級和長度之相關係數圖，如圖 39 所示。可以得知，它們關係呈現「淨度等級越高，相關程度越高」，其中走勢亦為負相關。表示淨度等級越差的鑽石挖出來的大小也會越大，再加上長度(X_5)主掌了$log_{10}Y$高達 91.8%的解釋性 (如圖 23 所示)，因此鑽石價格很容易被長度(X_5)帶偏，因此在長度(X_5)的作用下淨度(X_4)對價格的影響力也是微乎其微。

		log_price	x	Clarity_IF	Clarity_VVS1	Clarity_VVS2	Clarity_VS1	Clarity_VS2	Clarity_SI1	Clarity_SI2
log_price	Pearson 相關	1	.962**	-.074**	-.123**	-.080**	-.024**	-.011*	.036**	.165**
	顯著性 (雙尾)		.000	.000	.000	.000	.000	.014	.000	.000
	個數	53702	53702	53702	53702	53702	53702	53702	53702	53702
x	Pearson 相關	.962**	1	-.129**	-.185**	-.147**	-.059**	-.033**	.083**	.267**
	顯著性 (雙尾)	.000		.000	.000	.000	.000	.000	.000	.000
	個數	53702	53702	53702	53702	53702	53702	53702	53702	53702

圖 39、X_4(淨度)對X_5(長度)之相關係數表

8.2 模型解釋

本研究得出之模型如公式(4)所示，在此階段會針對各變數變動一單位下會對$log_{10}Y$造成多大幅度的影響做出解釋，並利用原始資料來比對模型的可信度。

8.2.1 連續型資料－ X_5(長度)

由公式(4)所得出之模型來看，當顏色為 D 且淨度為 VS2 時，代入表 2 中的數據，並由所得最終模型$log_{10}\hat{Y} = -0.531 + 0.244X_{31} + 0.344X_{45} + 0.427X_5 + 0.016X_8$計算，得出當鑽石長度每增加一單位，其對數價格會變動為 $\frac{3.07191}{2.64651} \approx 1.16$倍。

對照表 2 原始資料本身的變動比例為 $\frac{2.95}{2.56} \approx 1.15$倍，兩者相異不大。

index	$log_{10}Y$ (價格)	X_{31} (顏色"D")	X_{45} (淨度"VS2")	X_5 (長度)	X_8 (深度比例)
31601	2.56	D	VS2	3.73	62.3
35164	2.95	D	VS2	4.73	62.2

表 2、針對 X_5(長度)變動之資料比對表

8.2.2 連續型資料－ X_8(深度比例)

由公式(4)所得出之模型來看，當顏色為 F 且淨度為 I1 時，代入表 3 中的數據，並由所得最終模型$log_{10}\hat{Y} = -0.531 + 0.204X_{33} + 0.427X_5 + 0.016X_8$計算，得出當鑽石深度比例每增加一單位，其對數價格會變動為 $\frac{3.38278}{3.29419} \approx 1.03$ 倍。

對照表 3 原始資料本身的變動比例為 $\frac{3.36}{3.20} \approx 1.05$倍，兩者相異不大。

index	$log_{10}Y$ (價格)	X_{33} (顏色"F")	X_4 (淨度"I1")	X_5 (長度)	X_8 (深度比例)
44212	3.20	F	I1	5.97	67
50583	3.36	F	I1	6.14	68

表 3、針對 X_8(深度比例)變動之資料比對表

8.2.3 類別型資料－ X_3(顏色)

由公式(4)所得出之模型來看，當顏色為 D & E 且淨度為 VS2 時，代入表 4 中的數據，並分別由所得最終模型$\begin{cases} log_{10}\hat{Y} = -0.531 + 0.244X_{31} + 0.344X_{45} + 0.427X_5 + 0.016X_8 \\ log_{10}\hat{Y} = -0.531 + 0.219X_{32} + 0.344X_{45} + 0.427X_5 + 0.016X_8 \end{cases}$計算，得出當鑽石顏色等級每下降一個等級，其對數價格會變動為 $\frac{3.77489}{3.75827} \approx 1.00$ 倍。

對照表 4 原始資料本身的變動比例為 $\frac{3.66}{3.79} \approx 0.97$ 倍，兩者相異不大。

index	$log_{10}Y$ (價格)	X_{31} & X_{32} (顏色"D & E")	X_{45} (淨度"VS2")	X_5 (長度)	X_8 (深度比例)
15421	3.79	D	VS2	6.21	65.6
9098	3.66	E	VS2	6.27	66.6

表 4、針對 X_3(顏色)變動之資料比對表

8.2.4 類別型資料－ X_4(淨度)

由公式(4)所得出之模型來看，當顏色為 I 且淨度為 VVS1 & VVS2 時，代入表 5 中的數據，並分別由所得最終模型$\begin{cases} log_{10}\hat{Y} = -0.531 + 0.067X_{36} + 0.451X_{42} + 0.427X_5 + 0.016X_8 \\ log_{10}\hat{Y} = -0.531 + 0.067X_{36} + 0.425X_{43} + 0.427X_5 + 0.016X_8 \end{cases}$計算，得出當鑽石淨度等級每下降一個等級，其對數價格會變動為 $\frac{2.85636}{2.92826} \approx 0.98$ 倍。

對照表 5 原始資料本身的變動比例為 $\frac{2.87}{2.78} \approx 1.03$倍，兩者相異不大。

index	$log_{10}Y$ (價格)	X_{36} (顏色"I")	X_{42} & X_{43} (淨度"VVS1 & VVS2")	X_5 (長度)	X_8 (深度比例)
15694	2.78	I	VVS1	4.58	61.6
30832	2.87	I	VVS2	4.48	61.4

表 5、針對 X_4(淨度)變動之資料比對表

9. 結論及後續探討可能

總之，鑽石價格和鑽石本身的顏色(X_3)、淨度(X_4)、長度(X_5)以及深度比例(X_8)存在顯著的關係。

尤其是長度，鑽石的長度主掌其價格約九成左右的解釋性。

根據網路上的資料[2]顯示，「4C」是判斷一顆鑽石價值與品質的衡量標準。所謂「4C」指的是重量（CARAT）、色澤（COLOR）、淨度（CLARITY）及切工（CUT），所以在業界也有人說此四項指標的總和就是一顆鑽石的價值。這個說法與本研究最後得出之模型組成相異並不是很大，且此模型能做到的價格預測也已經達到 97%，所以若 kaggle 平台上的數據來源可靠，我們的確可以依據此模型來預測鑽石的價格。

參考文獻

[1] Kutner：Applied Linear Regression Models 4/e ’04

[2] CHU, Singfat. Pricing the C's of Diamond Stones. *Journal of Statistics Education*, 2001, 9.2

[3] https://www.investopedia.com/terms/d/durbin-watson-statistic.asp