應用統計專題

Youtube 觀看次數大解析

指導老師

李信宏 教授

組員

S0522108 洪羽柔

S0522112 艾品璇

S0522131 陳晏琦

S0522143 顏均翰

2019 年 6 月 12 日

目錄

研究動機

身處在資訊流通快速、人手一台智慧型手機的時代下，隨著各年齡層使用社群媒體的時間大幅提升，許多的應用程式紛紛被廣為轉載、使用，而 Youtube 成為了使用較頻繁的平台之一。由於 Youtube 這個平台相較於電視節目，可以免費自己開立一個專屬的頻道，因此各種類型的影片開始被拍攝放在平台上給大家觀賞，不論是教育型、生活型、寵物型…等影片，都迅速累積了許多忠實觀眾，透過觀看 Youtube 影片能夠打發時間、放鬆心情之外，還能利用 Youtube 來吸收各方面的知識，甚至許多的新聞、電視節目，除了在電視上播放外，也會使用 Youtube 這個平台來讓更多的族群能夠觀賞到，逐漸的政治、經濟開始與 Youtube 產生連結，進而發現 Youtuber 這個「職業」，可以藉由拍攝影片來累積觀看次數，且觀看次數的多寡也與 Youtuber 的收入有所關聯。除此之外，也能帶來額外的效益，像是廣告的置入、品牌代言、知名度的提升，都會是在拍攝影片之餘能夠獲得的利益，因此想探討什麼原因會影響到觀看次數，藉此知道，相對應的收入該是如何？因此，上述的原因是引起了想要研究 Youtube 的動機。藉此利用課堂上所學習到的迴歸分析技術，來深入探討影響 Youtube 影片觀看次數的原因。

資料介紹

資料擷取時間：以 2019 年 5 月 10 日下午 3 點左右為基準

- **變數介紹及說明**

綜合維基百科、新知筆記 Knowledge Notes 的【台灣前 100 名 Youtube 訂閱排行榜，你認識幾個呢?（2019 年 2 月)】影片，取出 66 名較知名的臺灣 Youtuber，為了不讓影片觀看次數浮動太大，因此擷取 3 月底上傳的一支影片，並透過直觀的想法選取八個可能影響影片觀看次數的原因，以下為所選擇的變數及其說明：

變數	**變數說明(單位)**
$\boldsymbol{Y}$	Youtube 影片觀看次數(萬次)
$\boldsymbol{X_1}$	Youtuber 的頻道訂閱數(萬人)
$\boldsymbol{X_2}$	Youtuber 的臉書追蹤人數(萬人)
$\boldsymbol{X_3}$	影片的廣告數(個)
$\boldsymbol{X_4}$	片長(分鐘)
$\boldsymbol{X_5}$	副頻道數(個)
$\boldsymbol{X_6}$	影片合作人數(人)
$\boldsymbol{X_7}$	Hashtag 數(個)
$\boldsymbol{X_8, X_9, X_{10}, X_{11}}$	頻道類型區分五種 $(X_8, X_9, X_{10}, X_{11})\begin{cases}(1,0,0,0) & \text{娛樂}\\(0,1,0,0) & \text{遊戲}\\(0,0,1,0) & \text{教育}\\(0,0,0,1) & \text{人物與日誌}\\(0,0,0,0) & \text{其他}\end{cases}$

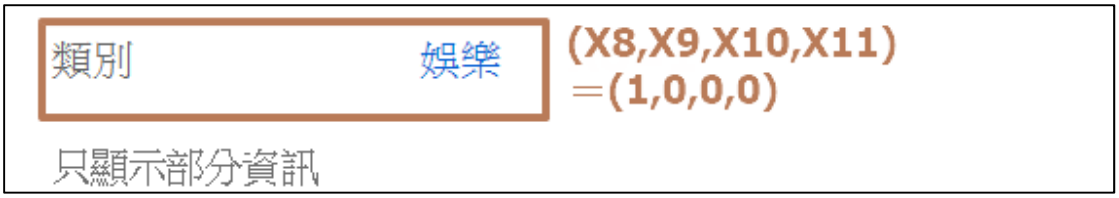
- **範例說明**

首先，開啟臉書，並在搜尋的地方輸入 Youtuber 名稱，進入其首頁後，在右下角即可看到此 Youtuber 的臉書追蹤人數(X_2)。

其次，開啟 Youtube 輸入欲尋找的 Youtuber，進入其首頁，就可以看到其 Youtube 訂閱人數(X_1)，再來點選【頻道】，也可看出其副頻道數(X_5)。

接著，在 Youtuber 的首頁點擊【影片】，從中找出目標影片並點開，即可看到此影片的觀看次數(Y)，還有片長(X_4)，而時間軸上的黃色點個數即為廣告個數(X_3)，接著從片名中可看出此影片的合作人數(X_6)，且片名上方之藍色井字號即為 Hashtag 數(X_7)。

最後，在影片下面的詳細資訊最後一行可得知此影片的類別。



以下是擷取所得 66 筆資料的結果：

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		Y	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11
2	這群人	186.5913	291.4566	197.7740	1	5.766666667	1	0	0	1	0	0	0
3	阿滴英文	76.3129	219.4632	71.5746	3	10.93333333	0	0	0	0	0	1	0
4	蔡阿嘎	49.6259	206.3970	172.0024	1	3.8	3	0	0	0	0	0	0
5	阿神	29.7891	187.7127	9.5295	5	20.9	1	0	1	0	1	0	0
6	古阿莫	57.2066	168.3972	75.8401	1	5	3	0	0	0	0	0	0
7	重量級CROWD	51.073	161.6658	132.4127	5	14.68333333	3	0	3	0	0	0	0
8	Joeman	144.6374	151.7579	29.8357	7	32.55	0	1	0	0	0	0	1
9	聖結石	47.3418	147.7294	66.8460	3	12.96666667	1	2	0	0	0	0	0
10	DE Jun	25.6967	146.8149	26.8512	2	12.16666667	0	0	0	0	1	0	0
11	小玉	129.2383	137.6543	30.3669	1	2.45	0	1	0	0	0	0	1
12	放火 Louis	67.363	134.5692	78.7551	1	3.35	1	2	0	0	0	0	0
13	安啾咪	39.9048	134.6361	53.1160	1	8.883333333	1	0	0	0	0	0	1
14	WACKBOYS 反骨男孩	39.1476	129.7082	63.8071	1	6.7	0	0	0	1	0	0	0
15	木曜4超玩	97.4912	128.1331	26.3280	4	70.36666667	1	1	3	1	0	0	0
16	菜喳	12.0357	125.8328	43.1746	0	10.55	0	1	3	0	1	0	0
17	人生肥宅X尊	52.7554	125.0779	39.4070	1	4.65	0	0	0	0	0	0	1
18	白癡公主	50.3393	124.7910	107.5830	2	11.41666667	0	0	3	1	0	0	0
19	魚乾	6.223	121.0740	35.2289	1	8.533333333	0	0	2	0	1	0	0
20	黃阿瑪的後宮生活	113.4823	117.4216	133.4785	1	7.8	0	0	0	0	0	0	0
21	HowFun	59.4892	110.3562	50.3829	1	2.95	0	0	3	1	0	0	0
22	滴妹	47.6495	108.0146	21.3899	1	7.2	0	1	0	0	0	0	1
23	理科太太 Li Ke Tai Tai	19.7645	106.2399	51.2783	0	16.58333333	0	0	1	0	0	1	0
24	三原JAPAN Sanyuan_JAPAN	60.4288	105.9070	36.6164	1	8.116666667	2	0	0	0	0	0	1
25	黃氏兄弟	18.7787	105.4942	9.5141	0	47.63333333	1	0	0	0	0	0	1
26	千千進食中	70.9482	104.7303	46.0894	2	11.5	0	0	0	0	0	0	1
27	sandy mandy	260.9366	105.2306	257.1839	1	3.333333333	0	0	2	1	0	0	0
28	在不瘋狂就等死x狂人娛樂	59.2295	103.0585	191.5539	1	13.35	13	2	1	0	0	0	0
29	舞秋風	1.9546	101.5444	22.4836	1	5.433333333	2	0	0	0	1	0	0
30	老皮	6.8438	48.4893	54.0968	4	117.6833333	1	0	0	0	1	0	0
31	啾啾鞋	15.6119	97.6745	11.8181	2	7.3	2	0	0	1	0	0	0
32	巧克力	5.8814	95.4939	15.2452	5	14.6	2	0	3	0	1	0	0
33	the劉沛	10.6631	91.9489	10.9691	3	12.2	2	0	3	0	1	0	0
34	Stopkiddinstudio	5.3362	89.6924	14.5038	2	2.316666667	1	0	0	0	0	1	0
35	古娃娃 WawaKu	31.491	88.7744	39.2893	2	8.366666667	1	0	0	0	0	0	1
36	Hello Catie	27.5375	85.2925	31.1879	4	20.85	0	0	3	0	0	0	0
37	搞神馬	42.2448	80.5950	5.1088	7	12.1	2	0	0	0	0	0	1
38	頑GAME	1.5058	74.1769	38.7879	4	10.48333333	1	4	3	1	0	0	0
39	MaoMao TV	9.0427	74.3848	12.9666	2	8.183333333	0	0	0	0	0	0	1
40	含羞草	26.4811	74.2528	13.6644	6	30	1	1	3	1	0	0	0
41	NyoNyoTV 妞妞 TV	18.7429	73.8568	4.7839	4	12.98333333	2	0	0	1	0	0	0
42	小白	17.0593	72.0594	5.8103	4	11.08333333	0	0	0	0	1	0	0
43	Taiwan Bar	8.6206	71.0374	29.6429	2	4.983333333	0	0	3	0	0	1	0
44	鬼鬼	0.4033	70.5972	12.6658	2	9.733333333	1	0	1	0	1	0	0
45	小草Yue	48.5515	70.9233	11.5197	4	12.43333333	1	0	3	0	1	0	0
46	狠愛演	78.5504	70.3629	32.5912	4	10.11666667	3	0	0	0	0	0	1
47	聖嫂Dodo	11.0488	68.5488	52.0444	4	11.01666667	1	0	0	0	0	0	1
48	展榮展瑞 K.R Bros	33.9714	68.2666	69.5907	0	14.13333333	3	1	3	0	0	0	0
49	黃大謙	65.7625	68.3066	36.3344	3	10.1	0	1	0	0	0	0	0
50	星培 Jasper	10.3	65.5254	68.7464	1	4.65	0	0	0	1	0	0	0
51	英雄日常 Heroisme	11.8527	66.1975	6.4491	4	11.2	1	0	0	0	0	0	1
52	GINA HELLO!	10.3709	56.8191	44.7155	1	11.33333333	0	0	2	1	0	0	0
53	超粒方	7.0173	63.8252	12.1736	1	3.566666667	1	0	0	0	0	0	0
54	館長成吉思汗	22.9196	58.7181	97.1363	2	11.5	1	1	3	0	0	0	0
55	BuBuChaCha 傳說	17.1208	51.1362	7.1374	2	3.05	0	0	3	0	1	0	0
56	烏鶇 DoKa TV	74.043	61.0647	1.6641	1	9.733333333	0	5	0	0	0	0	1
57	愛莉莎莎 Alisasa	75.2266	61.2346	12.3827	3	10.46666667	0	1	0	0	0	0	1
58	百變沛莉 Peri	9.223	53.4387	30.8740	3	14.45	1	0	0	0	0	0	0
59	馬叔叔 UNCLE MA	1.1203	50.6677	41.2072	2	1.683333333	0	0	0	0	0	0	0
60	上班不要看 NSFW	27.1731	56.3857	39.9726	5	16.16666667	0	0	3	0	0	0	0
61	台客劇場 Tkstory	20.8992	51.5003	42.1632	0	9.1	0	0	3	1	0	0	0
62	瘋狂老爹	6.0371	50.3234	5.1204	2	8.833333333	0	5	0	0	1	0	0
63	I.C Charlie	6.8785	47.4454	4.5653	2	6.483333333	0	0	0	0	0	0	0
64	TheKellyYang	24.8384	46.9617	12.5995	5	31.65	0	0	3	0	0	0	1
65	Dinter	14.6551	46.3679	27.8490	3	10.11666667	0	0	0	0	1	0	0
66	那對夫妻	9.4587	46.6746	246.2822	1	8.4	0	3	3	1	0	0	0
67	胡子Huzi	41.4183	51.8812	3.8683	4	10.35	0	4	3	1	0	0	0

模型挑選

收集好 66 筆數據後，首先要來挑選適當的變數模型。以下將會使用 Stepwise Selection、Forward Selection、Backward Elimination，以及 Best Subsets Regression 做個別分析，然後再從這四種方法比較得出一個可能是最適當的模型。

- **Stepwise Selection**

此方法的選入門檻（α-to-enter）及移出門檻（α-to-remove）皆設為 0.05。以下由 Minitab 統計軟體所呈現的選取結果。

Stepwise Selection of Terms

Candidate terms: x1(訂閱數)(萬), x2(FB追蹤人數)(萬人), x3(廣告數), x4(片長 分鐘), x5(副頻道數), x6(ft.人數), x7(hashtag數), x8(頻道類型:娛樂), x9(頻道類型:遊戲), x10(頻道類型:教育), x11(頻道類型:人物與日誌)

	-----Step 1----		-----Step 2----		-----Step 3----	
	Coef	P	Coef	P	Coef	P
Constant	20.34		-7.5		-17.2	
x2(FB追蹤人數)(萬人)	0.4225	0.000	0.3116	0.001	0.3790	0.000
x1(訂閱數)(萬)			0.342	0.002	0.3253	0.002
x11(頻道類型:人物與日誌)					31.2	0.003
S		39.3855		36.7030		34.4038
R-sq		26.85%		37.47%		45.93%
R-sq(adj)		25.71%		35.48%		43.31%
R-sq(pred)		8.79%		14.46%		21.93%
Mallows' Cp		21.28		11.19		3.56

α to enter = 0.05, α to remove = 0.05

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-17.2	10.3	-1.68	0.099	
x1(訂閱數)(萬)	0.3253	0.0982	3.31	0.002	1.18
x2(FB追蹤人數)(萬人)	0.3790	0.0853	4.44	0.000	1.26
x11(頻道類型:人物與日誌)	31.2	10.0	3.11	0.003	1.07

Regression Equation

Y(觀看次數(萬人) = -17.2 + 0.3253 x1(訂閱數)(萬) + 0.3790 x2(FB追蹤人數)(萬人) + 31.2 x11(頻道類型:人物與日誌)

由上述的 Stepwise Selection 所得結果，可看出此方法依序選擇的變數有 Youtuber 的頻道訂閱數（X_1）、臉書追蹤人數（X_2）、頻道類型−人物與日誌（X_{11}），且過程中沒有刪除任何前一階段已加入的變數。

- **Forward Selection**

同樣的先將設進入門檻（α-to-enter）定為 0.05。

Forward Selection of Terms

Candidate terms: x1(訂閱數)(萬), x2(FB追蹤人數)(萬人), x3(廣告數), x4(片長 分鐘), x5(副頻道數), x6(ft.人數), x7(hashtag數), x8(頻道類型:娛樂), x9(頻道類型:遊戲), x10(頻道類型:教育), x11(頻道類型:人物與日誌)

	-----Step 1----		-----Step 2----		-----Step 3----	
	Coef	P	Coef	P	Coef	P
Constant	20.34		-7.5		-17.2	
x2(FB追蹤人數)(萬人)	0.4225	0.000	0.3116	0.001	0.3790	0.000
x1(訂閱數)(萬)			0.342	0.002	0.3253	0.002
x11(頻道類型:人物與日誌)					31.2	0.003
S		39.3855		36.7030		34.4038
R-sq		26.85%		37.47%		45.93%
R-sq(adj)		25.71%		35.48%		43.31%
R-sq(pred)		8.79%		14.46%		21.93%
Mallows' Cp		21.28		11.19		3.56

α to enter = 0.05

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-17.2	10.3	-1.68	0.099	
x1(訂閱數)(萬)	0.3253	0.0982	3.31	0.002	1.18
x2(FB追蹤人數)(萬人)	0.3790	0.0853	4.44	0.000	1.26
x11(頻道類型:人物與日誌)	31.2	10.0	3.11	0.003	1.07

Regression Equation

Y(觀看次數(萬人) = -17.2 + 0.3253 x1(訂閱數)(萬) + 0.3790 x2(FB追蹤人數)(萬人) + 31.2 x11(頻道類型:人物與日誌)

從 Forward Selection 的結果顯示中，可發現其最後模型結果和 Stepwise Selection 是一樣的。

- **Backward Elimination**

  應用此法將移出門檻（α-to-remove）設定成 0.05。

Backward Elimination of Terms

Candidate terms: x1(訂閱數)(萬), x2(FB追蹤人數)(萬人), x3(廣告數), x4(片長 分鐘), x5(副頻道數), x6(ft.人數), x7(hashtag數), x8(頻道類型:娛樂), x9(頻道類型:遊戲), x10(頻道類型:教育), x11(頻道類型:人物與日誌)

	----Step 1----		----Step 2----		----Step 3----	
	Coef	P	Coef	P	Coef	P
Constant	-26.0		-25.9		-27.3	
x1(訂閱數)(萬)	0.353	0.002	0.352	0.001	0.362	0.001
x2(FB追蹤人數)(萬人)	0.408	0.000	0.408	0.000	0.402	0.000
x3(廣告數)	3.80	0.189	3.87	0.157	3.65	0.168
x4(片長 分鐘)	0.023	0.935				
x5(副頻道數)	-4.36	0.108	-4.35	0.105	-4.32	0.105
x6(ft.人數)	2.39	0.523	2.38	0.522	2.33	0.526
x7(hashtag數)	-1.35	0.719	-1.35	0.716		
x8(頻道類型:娛樂)	9.8	0.469	9.9	0.459	8.8	0.495
x9(頻道類型:遊戲)	-7.0	0.618	-6.8	0.620	-7.4	0.584
x10(頻道類型:教育)	-10.9	0.596	-10.8	0.593	-11.3	0.572
x11(頻道類型:人物與日誌)	28.1	0.040	28.2	0.037	29.2	0.026
S		34.5272		34.2140		33.9483
R-sq		52.57%		52.56%		52.45%
R-sq(adj)		42.91%		43.94%		44.81%
R-sq(pred)		1.98%		5.52%		6.18%
Mallows' Cp		12.00		10.01		8.14

	-----Step 4----		-----Step 5----		-----Step 6----	
	Coef	P	Coef	P	Coef	P
Constant	-31.3		-32.2		-30.0	
x1(訂閱數)(萬)	0.3570	0.001	0.3514	0.001	0.3381	0.001
x2(FB追蹤人數)(萬人)	0.4186	0.000	0.4213	0.000	0.4245	0.000
x3(廣告數)	3.64	0.166	3.74	0.149	3.71	0.151
x4(片長 分鐘)						
x5(副頻道數)	-4.14	0.114	-4.01	0.120	-3.89	0.129
x6(ft.人數)	2.37	0.518	2.46	0.497		
x7(hashtag數)						
x8(頻道類型:娛樂)	11.9	0.300	12.8	0.255	13.8	0.213
x9(頻道類型:遊戲)						
x10(頻道類型:教育)	-7.4	0.691				
x11(頻道類型:人物與日誌)	33.1	0.003	34.1	0.001	34.3	0.001
S		33.7403		33.4950		33.3433
R-sq		52.19%		52.06%		51.67%
R-sq(adj)		45.48%		46.27%		46.76%
R-sq(pred)		10.96%		11.79%		16.42%
Mallows' Cp		6.43		4.58		3.02

	-----Step 7----		-----Step 8----		-----Step 9----	
	Coef	P	Coef	P	Coef	P
Constant	-26.4		-15.4		-17.2	
x1(訂閱數)(萬)	0.3269	0.001	0.3268	0.001	0.3253	0.002
x2(FB追蹤人數)(萬人)	0.4591	0.000	0.4193	0.000	0.3790	0.000
x3(廣告數)	3.89	0.134				
x4(片長 分鐘)						
x5(副頻道數)	-4.64	0.065	-4.29	0.089		
x6(ft.人數)						
x7(hashtag數)						
x8(頻道類型:娛樂)						
x9(頻道類型:遊戲)						
x10(頻道類型:教育)						
x11(頻道類型:人物與日誌)	30.89	0.002	31.26	0.002	31.2	0.003
S		33.5063		33.8642		34.4038
R-sq		50.37%		48.46%		45.93%
R-sq(adj)		46.23%		45.08%		43.31%
R-sq(pred)		15.72%		17.32%		21.93%
Mallows' Cp		2.50		2.68		3.56

α to remove = 0.05

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-17.2	10.3	-1.68	0.099	
x1(訂閱數)(萬)	0.3253	0.0982	3.31	0.002	1.18
x2(FB追蹤人數)(萬人)	0.3790	0.0853	4.44	0.000	1.26
x11(頻道類型:人物與日誌)	31.2	10.0	3.11	0.003	1.07

Regression Equation

Y(觀看次數(萬人) = -17.2 + 0.3253 x1(訂閱數)(萬) + 0.3790 x2(FB追蹤人數)(萬人) + 31.2 x11(頻道類型:人物與日誌)

觀察 Backward Elimination 的結果數據，可發現此法依序刪除片長、hashtag 數、頻道類型(遊戲)、…等，得出最終模型為 Youtuber 的頻道訂閱數（X_1）、臉書追蹤人數（X_2），以及頻道類型-人物與日誌(X_{11}）這三個變數，明顯看出此結果與上述的 Forward Selection 及 Stepwise Selection 一模一樣。

- **Best Subsets Regression**

由一個自變數增加至十一個自變數，當中 Minitab 程式自動選取兩個模型，並提供該模型的 R-sq、R-sq(adj)、R-sq(pred)、Cp，以方便比較各個模型。

Best Subsets Regression: Y versus X1, X2, X3, X4, X5, X6, ... X9, X10, X11

Response is Y

Vars	R-Sq	R-Sq (adj)	R-Sq (pred)	Mallows Cp	S	X 1	X 2	X 3	X 4	X 5	X 6	X 7	X 8	X 9	X 10	X 11
1	26.9	25.7	8.8	21.3	39.385		X									
1	25.0	23.9	19.7	23.3	39.872	X										
2	37.5	35.5	14.5	11.2	36.703	X	X									
2	36.4	34.3	16.8	12.5	37.028		X									X
3	45.9	43.3	21.9	3.6	34.404	X	X									X
3	41.0	38.1	16.6	9.2	35.947	X	X							X		
4	48.5	45.1	17.3	2.7	33.864	X	X			X						X
4	48.3	44.9	23.2	2.8	33.907	X	X						X			X
5	50.4	46.2	15.7	2.5	33.506	X	X	X		X						X
5	49.9	45.8	18.0	3.0	33.654	X	X			X			X			X
6	51.7	46.8	16.4	3.0	33.343	X	X	X		X			X			X
6	51.0	46.0	11.5	3.8	33.587	X	X	X		X	X					X
7	52.1	46.3	11.8	4.6	33.495	X	X	X		X	X		X			X
7	51.8	46.0	15.7	4.8	33.572	X	X	X		X			X		X	X
8	52.2	45.5	11.2	6.4	33.729	X	X	X		X	X	X	X			X
8	52.2	45.5	11.0	6.4	33.740	X	X	X		X	X		X		X	X
9	52.4	44.8	6.2	8.1	33.948	X	X	X		X	X		X	X	X	X
9	52.3	44.7	10.4	8.3	33.984	X	X	X		X	X	X	X		X	X
10	52.6	43.9	5.5	10.0	34.214	X	X	X		X	X	X	X	X	X	X
10	52.5	43.8	3.6	10.1	34.253	X	X	X	X	X	X		X	X	X	X
11	52.6	42.9	2.0	12.0	34.527	X	X	X	X	X	X	X	X	X	X	X

先從 R-sq(adj)較大但自變數相對較少的模型中做挑選，代表可用相對較少的成本，達到差不多的效果。因此首選圖中所框起的模型，該C_p值為 3.6，比較接近其 p 值(p=4)，其他的模型變數增加但 R-sq(adj)卻沒有明顯的提升，舉例來說，變數個數為 6 的第一筆資料，雖然多了三個變數，但它的 R-sq(adj)僅僅上升了 3.5，但，且C_p值與 p 值差了 4。再根據 Stepwise Selection、Forward Selection 和 Backward Elimination 三種挑選模型的方式中，分析出的模型應變數 Y（影片觀看次數)都是與 X_1(頻道訂閱數)、X_2（臉書追蹤人數)和 X_{11}(頻道類型-人物與日誌）這三個自變數有關。因此最終選的模型為：

$$\hat{Y} = -17.2 + 0.3253X_1 + 0.3790X_2 + 31.2X_{11}$$

迴歸模型分析

選擇模型後，將透過迴歸分析，藉由給定頻道訂閱數、臉書追蹤人數、頻道類型是否為人物與日誌，來預估影片觀看次數。

- **初步檢視迴歸分析**

對模型進行迴歸分析以檢查此模型是否顯著。

Regression Equation

Y(觀看次數(萬人)) = -17.2 + 0.3253 x1(訂閱數)(萬) + 0.3790 x2(FB追蹤人數)(萬人) + 31.2 x11(頻道類型:人物與日誌)

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-17.2	10.3	-1.68	0.099	
x1(訂閱數)(萬)	0.3253	0.0982	3.31	0.002	1.18
x2(FB追蹤人數)(萬人)	0.3790	0.0853	4.44	0.000	1.26
x11(頻道類型:人物與日誌)	31.2	10.0	3.11	0.003	1.07

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	62339	20780	17.56	0.000
x1(訂閱數)(萬)	1	12991	12991	10.98	0.002
x2(FB追蹤人數)(萬人)	1	23369	23369	19.74	0.000
x11(頻道類型:人物與日誌)	1	11484	11484	9.70	0.003
Error	62	73384	1184		
Total	65	135723			

藉由 F 檢定(P-Value=0.000)可以發現此模型是顯著的，也就是說觀看次數與 X_1(頻道訂閱數)、X_2（臉書追蹤人數)和 X_{11}(頻道類型-人物與日誌）這三個自變數有關係。就個別的 t 檢定來看，X_1、X_2、X_{11}的 P-Value 分別為 0.002、0.000 和 0.003，可知自變數與應變數之間是有關係的。又各別的 VIF 值大小為 1.18(X_1)、1.26(X_2)、1.07(X_{11})也可以得知自變數之間沒有存在共線性關係。

- **離群值(outlier)分析**

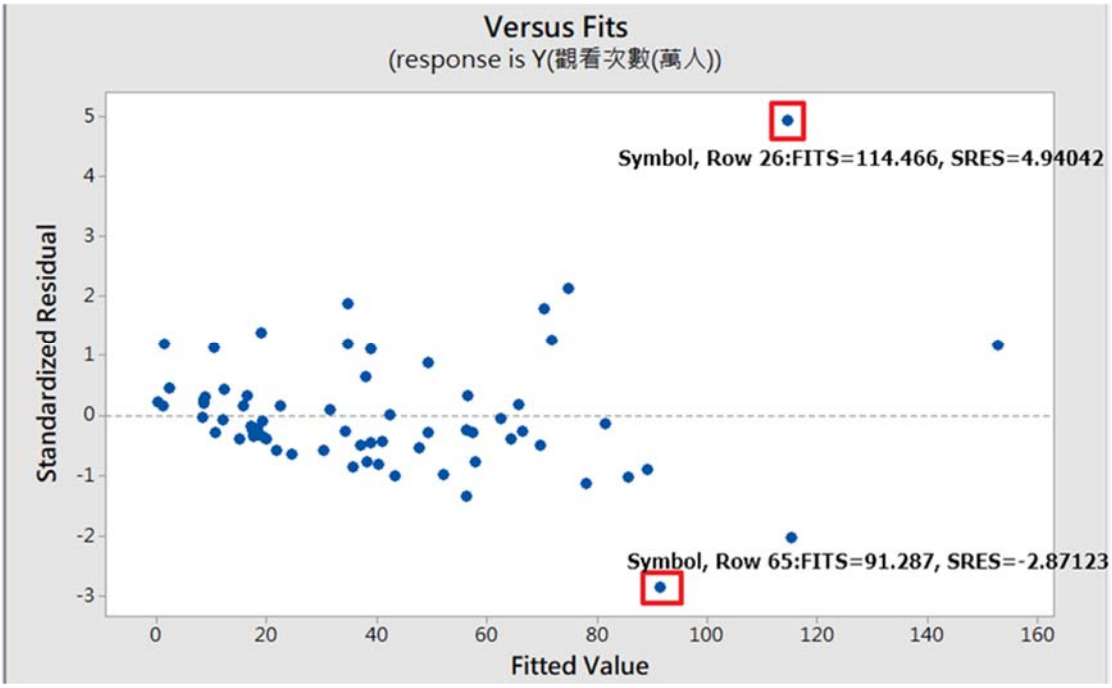
接下來，將觀察是否有離群值的存在。

Fits and Diagnostics for Unusual Observations

Obs	Y(觀看次數(萬人))	Fit	Resid	Std Resid		
1	186.6	152.5	34.1	1.18		X
3	49.6	115.1	-65.5	-2.04	R	
7	144.6	74.7	70.0	2.13	R	
26	260.9	114.5	146.5	4.94	R	X
65	9.5	91.3	-81.8	-2.87	R	X

R Large residual
X Unusual X

從數據可看出第 3、7、26、65 筆資料的標準化殘差較大，因此將透過殘差圖來觀察這四筆資料是否為離群值。



從圖中，可以發現有兩點，分別是第 26 筆資料(standard residual=4.94)和第 65 筆資料(standard residual=-2.87)，與其他點相比之下較為偏離，且這兩筆資料的標準化殘差也落在(-2,2)的區域外，相對而言這兩筆數據的殘差較為極端，因此可能是離群值。

再透過觀察背景資料發現，第 26 筆資料為兩位年輕女生跳舞的頻道，有別其他 65 筆資料的原始背景，都沒有以跳舞為主軸的頻道，再加上該影片的拍攝主題為當紅韓團 BLACKPINK - 'Kill This Love' Dance cover，且 Kill This Love 這首歌相當膾炙人口，可以發現這首歌在 Youtube 官方平台的影片觀看次數高達 3 億次，並可以透過搜尋發現，翻跳這首歌的影片觀看次數也都相當可觀，因此視第 26 筆資料是會影響迴歸模型的離群值，決定將它刪除。

另外，也刪掉了第 65 筆資料，而此影片為該 Youtuber 的回顧影片，影片內容是把之前他們上傳過的多支影片，將較為好笑或是相關性較高的片段，又再加以剪接成一支新的影片，主要是在做重點回顧，再加上透過觀察其他其他 65 筆資料的原始背景，發現沒有其他支影片是屬於回顧類型，因此也將它視為離群值刪除。

接著再做一次迴歸分析檢查是否還存在其他潛在之離群值。

Fits and Diagnostics for Unusual Observations

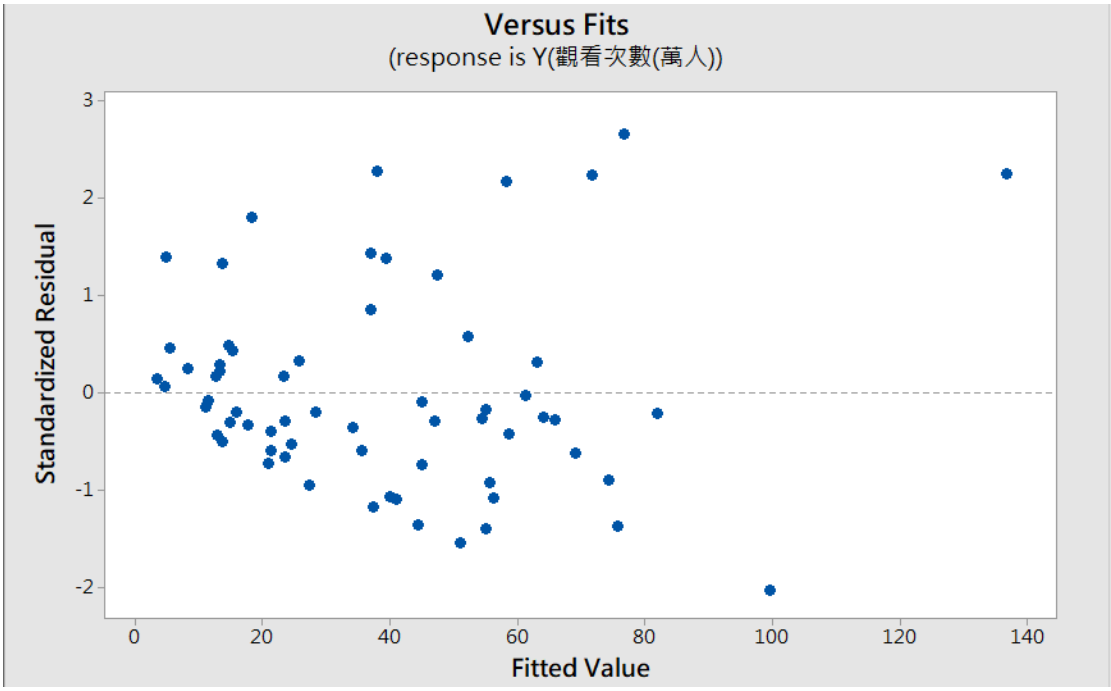
Obs	Y(觀看次數(萬人)	Fit	Resid	Std Resid		
1	186.59	136.60	49.99	2.25	R	X
3	49.63	99.45	-49.83	-2.03	R	
7	144.64	76.58	68.05	2.66	R	
10	129.24	71.50	57.74	2.24	R	
14	97.49	37.92	59.58	2.27	R	
19	113.48	58.00	55.48	2.18	R	
26	59.23	65.74	-6.51	-0.29		X

R Large residual
X Unusual X

從分析的結果顯示出有幾筆資料的標準化殘差較其他資料來的大，但其殘差範圍大約都在$(-2, 2)$左右，還在可以接受的合理範圍內，並觀察這六筆資料的原始背景，發現沒有相較於其他筆資料有所特別的地方，因此並未將這六筆資料視為潛在的離群值。

- **檢查迴歸假設**

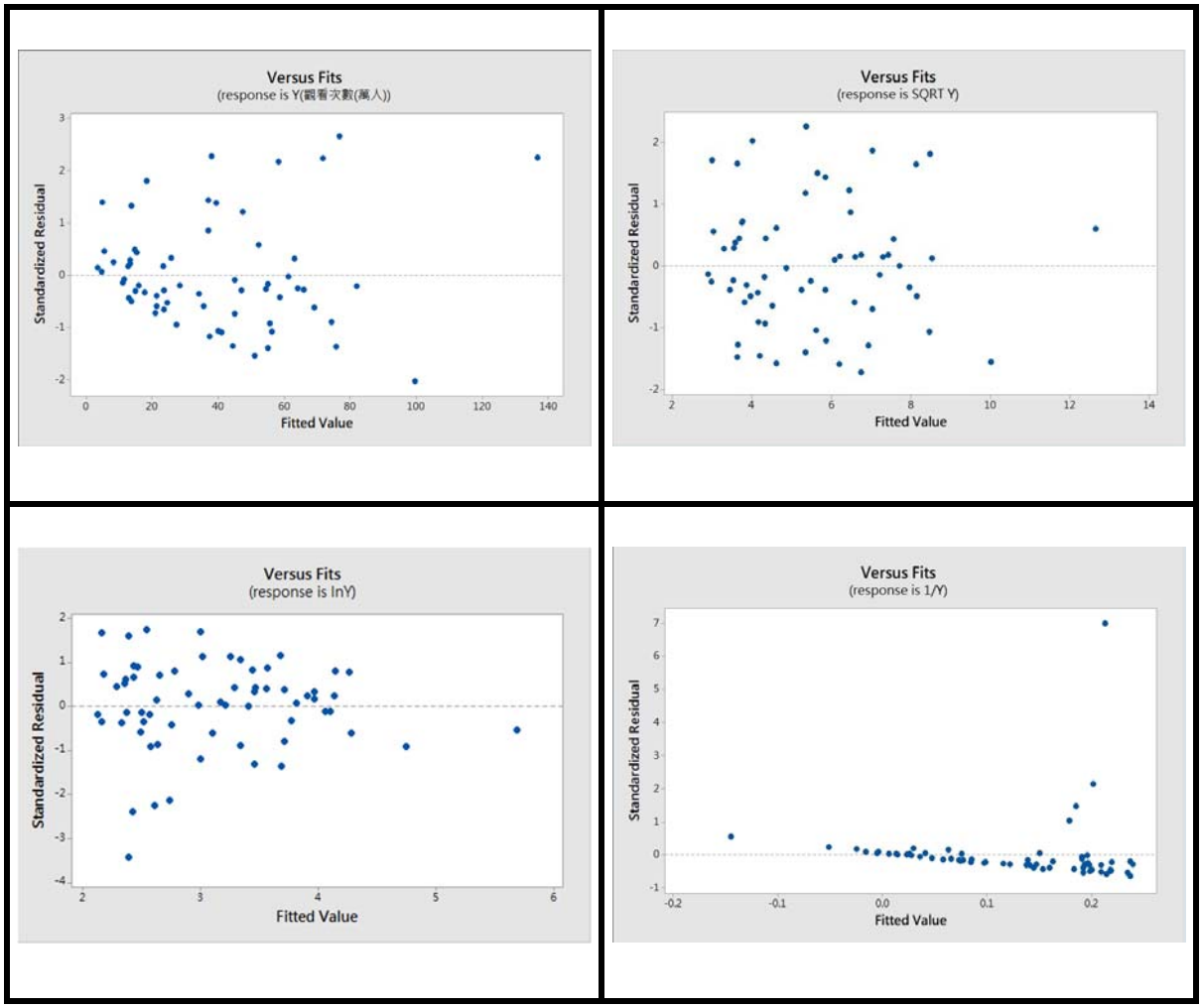
首先，藉由殘殘差圖來觀察標準化殘差的變異數是否為常數。



觀察點的分佈有逐漸向右展開的趨勢，因此可以得知標準化殘差的變異數並非常數，藉此，透過對 Y 做變數轉換以符合標準化殘差的變異數為常數的迴歸假設。

以下考慮三種常見的應變數轉換方式：分別是對 Y 開根號、對 Y 取 ln，以及取 Y 的倒數。重新進行迴歸分析後，得到的四個對應的殘差圖，其中左上圖為未轉換前的結果，右上圖是$\sqrt{Y}$，左下圖是 lnY，而右下圖則是$\frac{1}{Y}$的情形。

Standard residuals vs. fitted values 對照圖



由上圖可觀察出，lnY 和 $\frac{1}{Y}$ 的圖並未改善原本點的分佈，而$\sqrt{Y}$圖形中的點像一盤散沙，可推測標準化殘差的變異數為常數，代表此轉換之模型比較符合迴歸假設。

同樣地，觀察迴歸係數的估計與檢定。

Regression Equation

SQRT Y = 1.587 + 0.02565 x1(訂閱數)(萬) + 0.01808 x2(FB追蹤人數)(萬人) + 2.449 x11(頻道類型:人物與日誌)

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	240.86	80.285	18.89	0.000
x1(訂閱數)(萬)	1	62.02	62.017	14.59	0.000
x2(FB追蹤人數)(萬人)	1	24.59	24.593	5.79	0.019
x11(頻道類型:人物與日誌)	1	69.70	69.696	16.40	0.000
Error	60	255.04	4.251		
Total	63	495.89			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
2.06171	48.57%	46.00%	41.74%

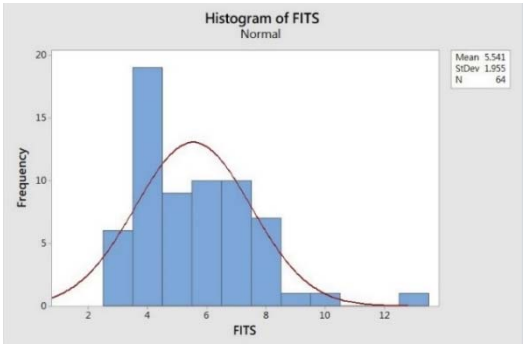
Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	1.587	0.624	2.54	0.014	
x1(訂閱數)(萬)	0.02565	0.00671	3.82	0.000	1.51
x2(FB追蹤人數)(萬人)	0.01808	0.00752	2.41	0.019	1.60
x11(頻道類型:人物與日誌)	2.449	0.605	4.05	0.000	1.07

得知轉換後的模型是顯著的，而且可以從 VIF 觀察到變數間依舊沒有共線性關係。

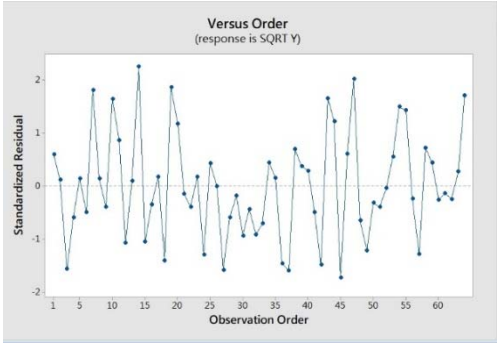
接著以此模轉換後的模型，繼續檢查其他的迴歸假設包括 fitted values 是否為常態分布？、標準殘差是否為隨機？、標準化殘差的變異數是否為常數？以及標準化殘差是否為常態分布等。

Histogram of fitted values



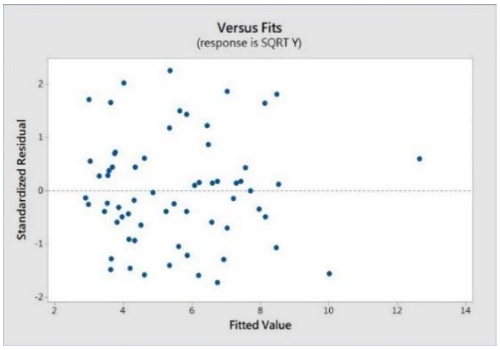
透過 fitted values 直方圖，可以觀察到除了 fitted value 值為 4 的部份，整體近似常態分佈，且由於觀察資料在整體上的分布，並沒有某一個 fitted values 特別偏離其餘的資料，因此沒有進一步要對模型做其他變數變換的調整。

Sequence plot of standard residuals



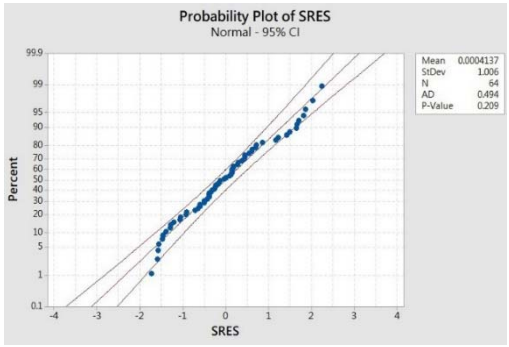
將標準化殘差依資料收集順序繪圖如上，觀察出此圖並沒有任何一個特殊趨勢，是任意上下起伏的，因此得知此份數據是隨機的，不會因為觀察次序的增加而標準化殘差就隨之有明顯的上升或下降。

Standard residuals vs. fitted values



觀察上圖的分布狀況，可以看到這些點的分布很散沙狀且相當均勻，殘差範圍也都大致落在(−2, 2)之間，相較於之前還未做變數變換時Y的擬合值比較圖，已經並未呈現出任何趨勢(例如：右開、左開或是線性的趨勢)，這就表示，其標準化殘差的變異數屬於常數，符合迴歸假設。

Probability plot of standard residuals



透過標準化殘差之機率圖，可以知道 P−Value 的值為 0. 209，大於顯著水準 α=0.05，且 95%的信賴區間帶幾乎含蓋所有殘差，可以知道標準化殘差是符合常態分配假設。

最終模型與解釋

藉由上述的迴歸分析和圖形，可發現$\sqrt{Y}$的結果皆符合迴歸假設的情況，因此最終選擇的模型為

$$\widehat{\sqrt{Y}} = 1.587 + 0.02565X_1 + 0.01808X_2 + 2.449X_{11}$$

其中迴歸係數$b_0 = 1.587$，$b_1 = 0.02565$，$b_2 = 0.01808$，$b_{11} = 2.449$

以下針對各個自變數，分別討論其對應變數 Y 的影響。

觀察訂閱人數(X_1)對於影片觀看次數(Y)的影響：

在固定臉書追蹤人數(X_2)和頻道類型(人物與日誌X_{11}) 的情況之下，

$\sqrt{Y} = b_0 + b_1X_1 + b_2X_2 + b_{11}X_{11}$

$\Rightarrow Y = (b_0 + b_1X_1 + b_2X_2 + b_{11}X_{11})^2$

考慮X_1增加 1 單位，設Y'為X_1代$(X_1 + 1)$所得結果

$Y' = [b_0 + b_1(X_1 + 1) + b_2X_2 + b_{11}X_{11}]^2$

$\Rightarrow Y' = [b_1 + (b_0 + b_1X_1 + b_2X_2 + b_{11}X_{11})]^2 = \left(b_1 + \sqrt{Y}\right)^2$

所以，當X_1增加 1 單位時Y的變化量為

$$Y' - Y = \left(b_1 + \sqrt{Y}\right)^2 - Y = b_1^2 + 2b_1\sqrt{Y} + Y - Y = b_1^2 + 2b_1\sqrt{Y}$$

$$= b_1^2 + 2b_1(b_0 + b_1X_1 + b_2X_2 + b_{11}X_{11})$$

舉例來說，當固定臉書追蹤人數(X_2)為 80 萬人，而頻道類型不為人物與日誌($X_{11} = 0$)時，若訂閱人數(X_1)為 80 萬人，則訂閱人數(X_1)增加 1 萬人時，影片觀看次數(Y)的變化為

$Y' - Y$

$= b_1^2 + 2b_1(b_0 + b_1X_1 + b_2X_2 + b_{11}X_{11})$

$= 0.02565^2 + 2 \times 0.02565$

$\times (1.587 + 0.02565 \times 80 + 0.01808 \times 80 + 2.449 \times 0)$

≈ 0.2615(萬次)，也就是影片觀看次數(Y)會增加 2615 次。

同理，若訂閱人數(X_1)為 90 萬人，則訂閱人數(X_1)增加 1 萬人時，影片觀看次數(Y) 的變化為

$$Y' - Y$$

$$= b_1^2 + 2b_1(b_0 + b_1X_1 + b_2X_2 + b_{11}X_{11})$$

$$= 0.02565^2 + 2 \times 0.02565$$

$$\times (1.587 + 0.02565 \times 90 + 0.01808 \times 80 + 2.449 \times 0)$$

≈ 0.2747(萬次) ，亦即影片觀看次數(Y)會增加 2747 次。

雖然訂閱人數 80 萬與 90 萬相差 10 萬，但當訂閱人數分別增加 1 萬人時，影片觀看次數其實只差 132 人，差異並不大。

觀察臉書追蹤人數(X_2)對於影片觀看次數(Y)的影響：

在固定訂閱人數(X_1)和頻道類型(人物與日誌X_{11}) 的情況之下，

$$\sqrt{Y} = b_0 + b_1X_1 + b_2X_2 + b_{11}X_{11}$$

$$\Rightarrow Y = (b_0 + b_1X_1 + b_2X_2 + b_{11}X_{11})^2$$

考慮X_2增加一單位，設Y'為X_2代$(X_2 + 1)$所得結果

$$Y' = [b_0 + b_1X_1 + b_2(X_2 + 1) + b_{11}X_{11}]^2$$

$$\Rightarrow Y' = [b_2 + (b_0 + b_1X_1 + b_2X_2 + b_{11}X_{11})]^2 = \left(b_2 + \sqrt{Y}\right)^2$$

因此，當X_2增加 1 單位時Y的變化量為

$$Y' - Y = \left(b_2 + \sqrt{Y}\right)^2 - Y = b_2^2 + 2b_2\sqrt{Y} + Y - Y = b_2^2 + 2b_2\sqrt{Y}$$

$$= b_2^2 + 2b_2(b_0 + b_1X_1 + b_2X_2 + b_{11}X_{11})$$

舉例來說，考量訂閱人數(X_1)為 70 萬人且頻道類型不為人物與日誌($X_{11} = 0$)時，若臉書追蹤人數(X_2)為 60 萬人，則當該頻道追蹤人數(X_2)增加 1 萬人時，影片觀看次數(Y)的變化量為

$$Y' - Y$$

$$= b_2^2 + 2b_2(b_0 + b_1X_1 + b_2X_2 + b_{11}X_{11})$$

$$= 0.01808^2 + 2 \times 0.01808$$

$$\times (1.587 + 0.02565 \times 70 + 0.01808 \times 60 + 2.449 \times 0)$$

≈ 0.1617(萬次)，也就是影片觀看次數(Y)大約增加 1617 次。

同理，若臉書追蹤人數(X_2)為 90 萬人，則當該頻道追蹤人數(X_2)增加 1 萬人時，影片觀看次數(Y) 的變化量為

$$Y' - Y$$

$$= b_2^2 + 2b_2(b_0 + b_1X_1 + b_2X_2 + b_{11}X_{11})$$

$$= 0.01808^2 + 2 \times 0.01808$$

$$\times (1.587 + 0.02565 \times 70 + 0.01808 \times 90 + 2.449 \times 0)$$

≈ 0.1812(萬次) ，亦即增加 1812 次。

可觀察出臉書追蹤人數 90 萬雖然為 60 萬的 1.5 倍，但當兩者的臉書追蹤人數增加 1 萬人，影片觀看次數並不會因此也增加 1.5 倍。

頻道類型為人物與日誌(X_{11})對影片觀看次數(Y)的影響：

$\sqrt{Y} = b_0 + b_1X_1 + b_2X_2 + b_{11}X_{11}$

$\Rightarrow Y = (b_0 + b_1X_1 + b_2X_2 + b_{11}X_{11})^2$

設Y_1為$X_{11} = 0$所得結果，$Y_1 = (b_0 + b_1X_1 + b_2X_2)^2$

設Y_2為$X_{11} = 1$所得結果，則

$Y_2 = (b_0 + b_1X_1 + b_2X_2 + b_{11})^2 = [b_{11} + (b_0 + b_1X_1 + b_2X_2)]^2$

$= \left(b_{11} + \sqrt{Y_1}\right)^2 = {b_{11}}^2 + 2b_{11}\sqrt{Y_1} + Y_1$

所以頻道類型是「人物與日誌($X_{11} = 1$)」較「非人物與日誌($X_{11} = 0$)」的影片觀看次數(Y)增加量為

$$Y_2 - Y_1 = \left(b_{11}{}^2 + 2b_{11}\sqrt{Y_1} + Y_1\right) - Y_1 = b_{11}{}^2 + 2b_{11}\sqrt{Y_1}$$

$$= b_{11}{}^2 + 2b_{11}(b_0 + b_1X_1 + b_2X_2)$$

舉例來說，當訂閱人數(X_1)為 100 萬人且臉書追蹤人數(X_2)為 70 萬人時，則頻道類型為人物與日誌($X_{11} = 1$) 相較頻道類型非人物與日誌($X_{11} = 0$)影片觀看次數(Y)的變化量為

$$Y_2 - Y_1 = b_{11}{}^2 + 2b_{11}(b_0 + b_1X_1 + b_2X_2)$$

$$= 2.449^2 + 2 \times 2.449 \times (1.587 + 0.02565 \times 100 + 0.01808 \times 70)$$

$$\approx 32.5090(\text{萬次})$$ ，大約增加 325090 次。由此可知，頻道類型是否為人物與日誌對觀看次數的影響較大。

總結來說，從迴歸分析結果，可以得知 Youtuber 的頻道訂閱數(X_1) 、 Youtuber 的臉書追蹤人數(X_2)以及頻道類型(人物與日誌X_{11})，的確顯著影響觀看次數 Y，從 R-sq(adj)= 46%來看，此三個自變數大約可以解釋約 46% Youtube 影片觀看次數(Y)的變化量。因此，想要有較高的觀看次數，就需設法提高頻道訂閱數和臉書追蹤人數，而且拍攝內容是屬於人物與日誌類型的影片。不過由於選取的 Youtuber 並非為剛起步之 Youtuber，也就是資料中的各個 Youtuber 皆是擁有至少 40 萬的訂閱數，因此該模型並不適合套用於剛創立的頻道。可惜的是此迴歸模型只能解釋不到一半的影片觀看次數，代表還有其他尚未探討到的因素會影響影片的觀看次數，期許在未來能夠發現其他因素，也希望將來的資料能夠擴展至全球 Youtuber 以發展出新的模型。

資料來源

YouTube：https://www.youtube.com/

臉書：https://www.臉書.com/

維基百科－台灣 YouTube 頻道訂閱人數排行榜：
https://zh.wikipedia.org/wiki/%E5%8F%B0%E7%81%A3YouTube%E9%A0%BB%E9%81%93%E8%A8%82%E9%96%B1%E4%BA%BA%E6%95%B8%E6%8E%92%E8%A1%8C%E6%A6%9C

新知筆記 Knowledge Notes－《台灣前 100 名 Youtuber 訂閱排行榜，你認識幾個呢?》：
https://www.youtube.com/watch?v=brTYxkUYIT0

TKU-機率計算：
http://netstat.stat.tku.edu.tw/prob.php