應用統計報告

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 3](#_Toc15549328)

[資料介紹 4](#_Toc15549329)

[ 變數介紹及說明 4](#_Toc15549330)

[ 範例說明 5](#_Toc15549331)

[ 列舉66筆數據 6](#_Toc15549332)

[模型挑選 8](#_Toc15549333)

[ 逐步選擇法(Stepwise Selection) 8](#_Toc15549334)

[ 向前選擇法(Forward Selection) 10](#_Toc15549335)

[ 後退消除法(Backward Elimination) 12](#_Toc15549336)

[ 最佳子集迴歸(Best Subsets Regression) 15](#_Toc15549337)

[ 模型說明 16](#_Toc15549338)

[模型迴歸分析 17](#_Toc15549339)

[ 初步迴歸分析 17](#_Toc15549340)

[ 刪除離群值(outlier)之迴歸分析 19](#_Toc15549341)

[ 應變數之線性轉換 22](#_Toc15549342)

[最終模型與解釋 26](#_Toc15549343)

[ 結論 31](#_Toc15549344)

[心得感想 32](#_Toc15549345)

[資料來源 34](#_Toc15549346)

**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**

為了不讓影片觀看次數太過極端，因此擷取上傳一個月左右的影片來分析，並透過最直接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)** |
| ***Y*** | Youtube影片觀看次數(萬次)  綜合維基百科、新知筆記 Knowledge Notes的【台灣前100名Youtuber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Youtuber3月底上傳的一支影片。 |
|  | Youtuber的頻道訂閱數(萬人) |
|  | Youtuber的facebook追蹤人數(萬人) |
|  | Youtuber影片的廣告數(個)  時間軸上的一個黃點就代表置入一支廣告 |
|  | 片長(分鐘) |
|  | Youtuber副頻道數(個) |
|  | 影片合作人數(人) |
|  | Hashtag數(個) |
|  | 頻道類型區分五種 |

* **範例說明**

首先將說明收集變數之數據的方法：

步驟一：

****開啟facebook，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facebook追蹤人數()。

步驟二：

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()，再來點選【頻道】，可看出此Youtube的副頻道數()。

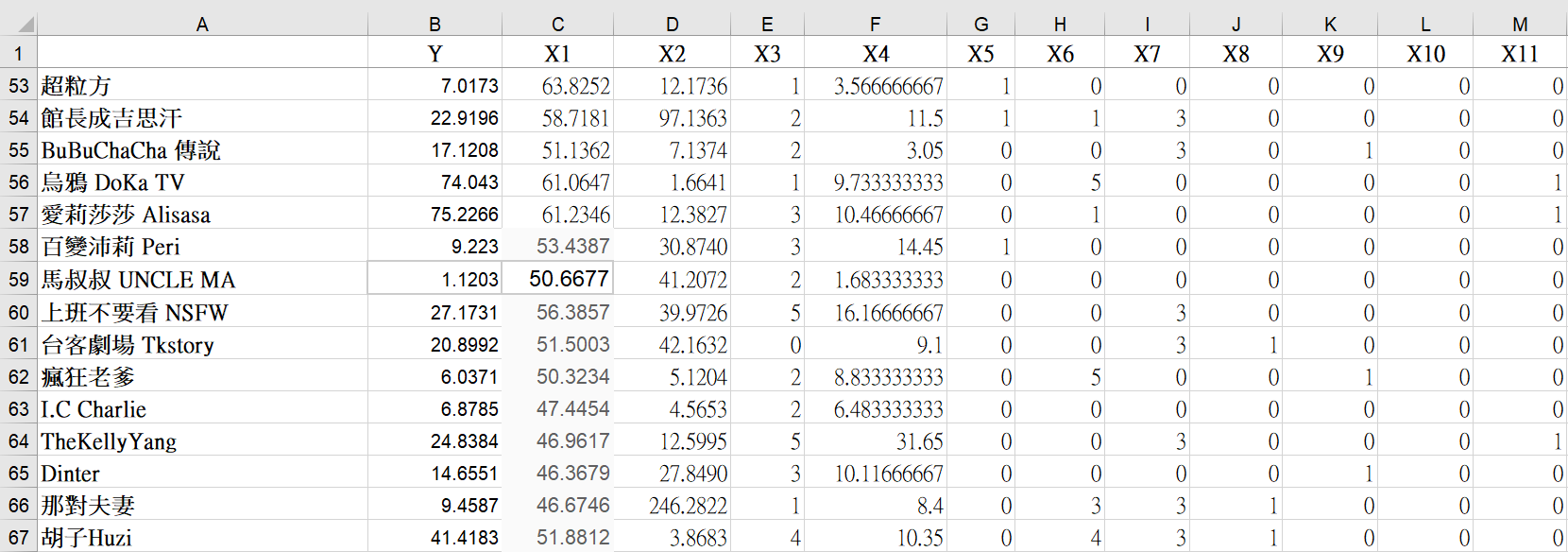
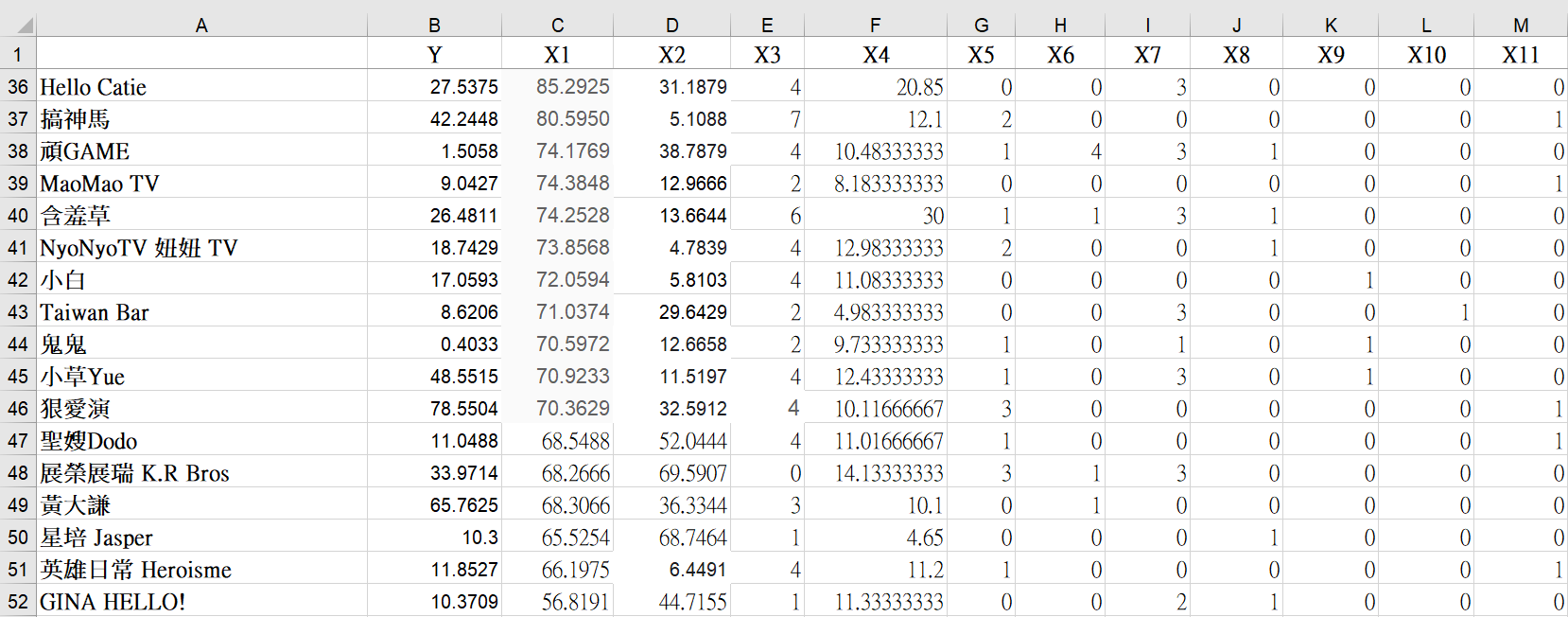
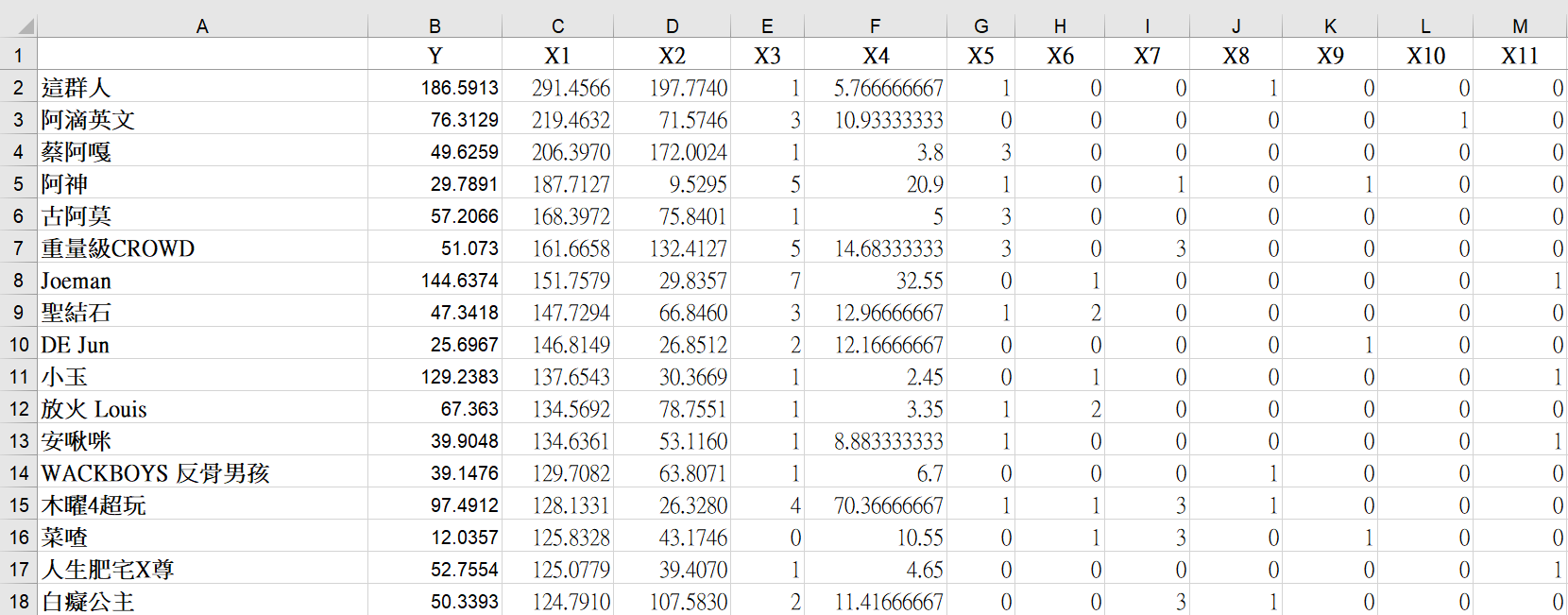
步驟三：

在Youtuber的首頁點擊【影片】，從中找出三月底所上傳的影片並點開，即可看到此影片的觀看次數(Y)，還有片長()，而時間軸上的黃色點個數即為Youtuber影片的廣告個數()，接著從片名中可看出此影片的合作人數()，且片名上方之藍色井字號即為Hashtag數()。

步驟四：

在影片下面的詳細資訊最後一行可得知此影片的類別()。

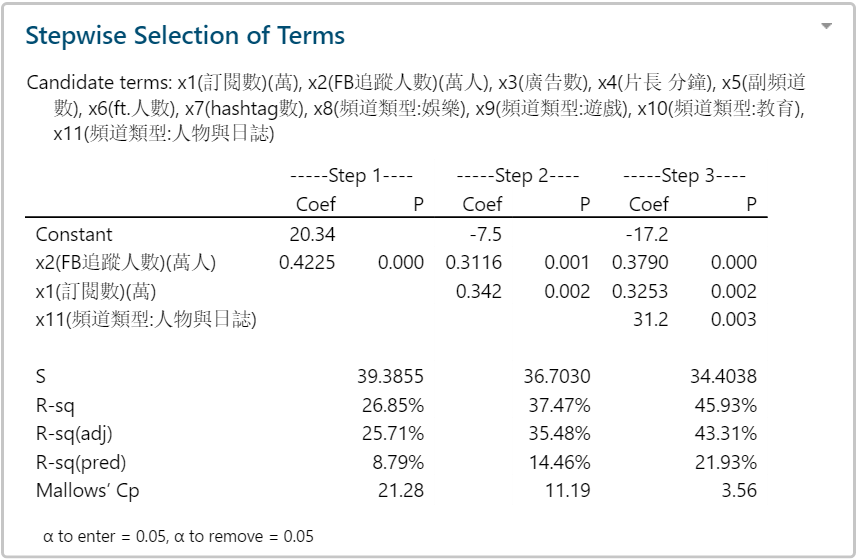
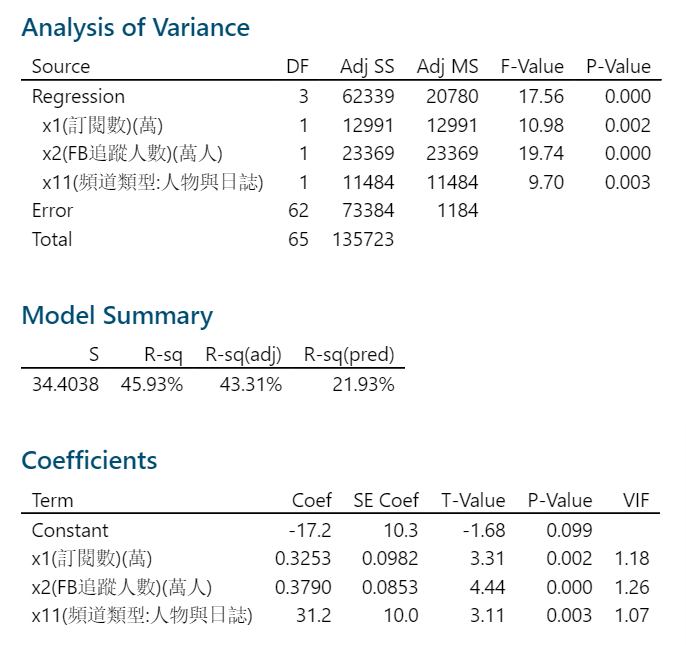
* **列舉66筆數據**



**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，以下將會使用逐步選擇法（Stepwise Selection）、向前選擇法（Forward Selection）、後退消除法（Backward Elimination），以及最佳子集迴歸（Best Subsets Regression）做個別分析，然後再從這四種方法所得結果綜合出一個可能是最適當的模型。

* **逐步選擇法****(Stepwise Selection)**

****一開始將進入門檻（）及移出門檻（）皆設為0.05。第一步先將P-Value最小且小於0.05的變數放入模型中；第二步在從未挑選的變數中，把P-Value最小且不超過進入門檻的變數加入模型，並檢查放入新變數後，模型裡變數之P-Value是否有因此超過移出門檻，若超過就將超過的變數移出；重複步驟二，直到剩餘的變數其P-Value皆大於進入門檻。

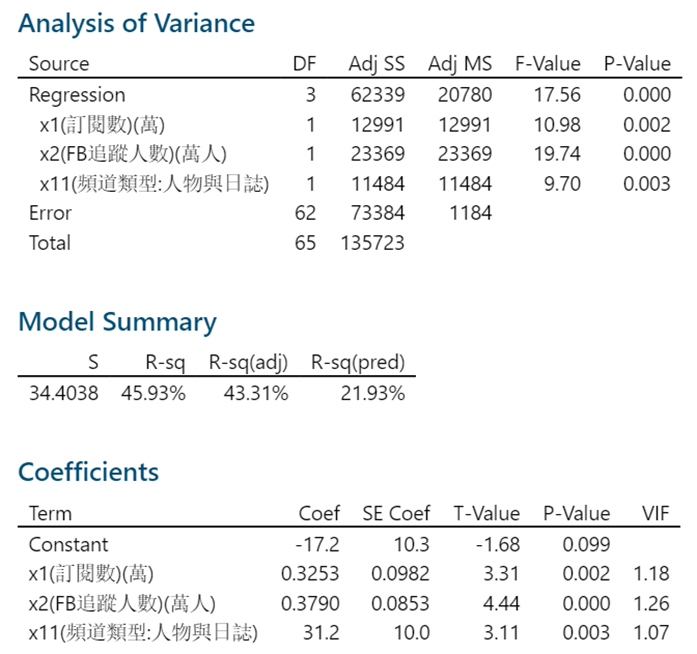
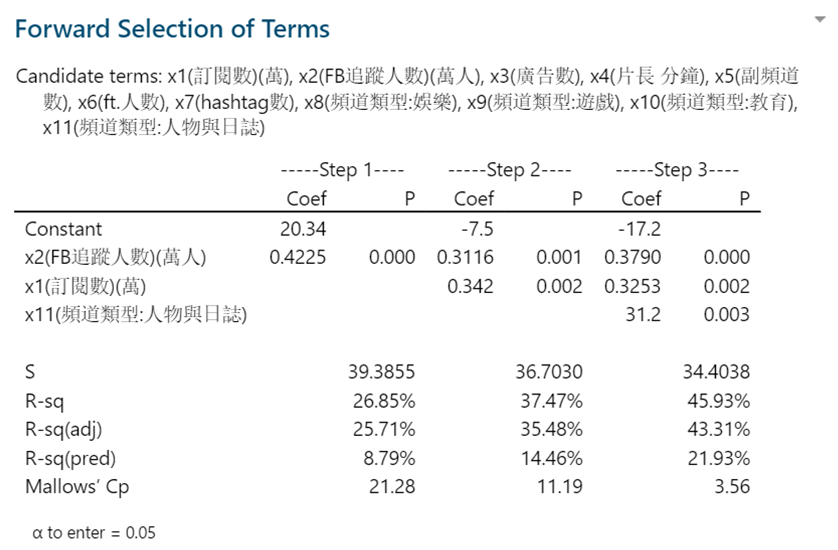


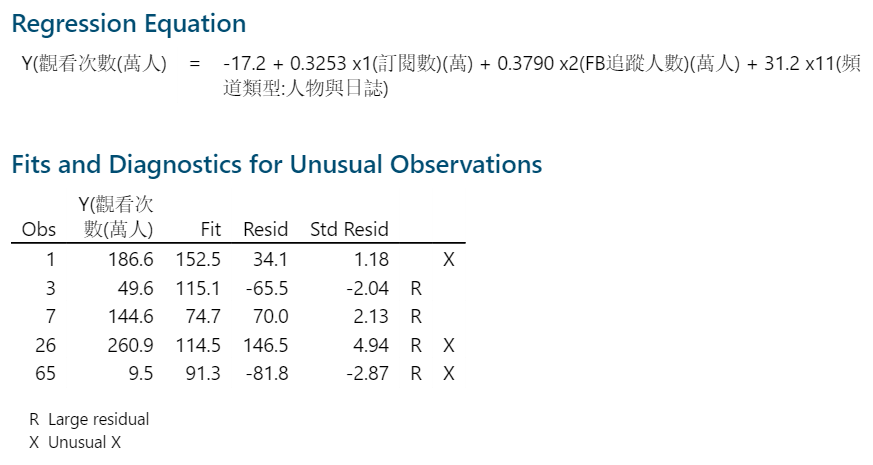
****

結果分析：

由上述的逐步選擇法（Stepwise Selection）所得結果，可看出此方法最後模型包含的變數有Youtuber的頻道訂閱數（）、facebook追蹤人數（），與頻道類型：人物與日誌（），且R-sq(adj)為43.31％。

* **向前選擇法(Forward Selection)**

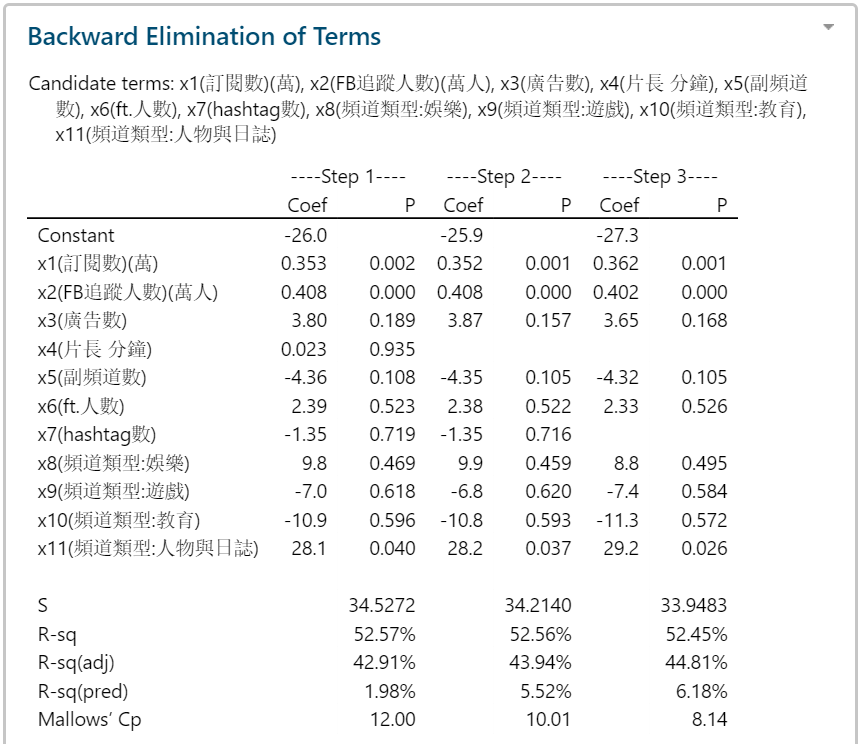
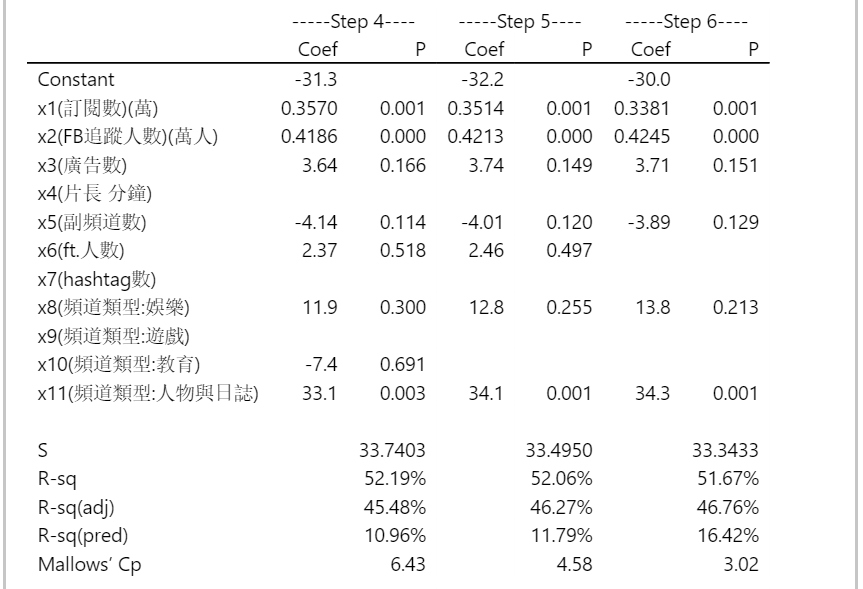
同樣的先將設進入門檻（）定為0.05。此方法將會把P-Value最小且不超過進入門檻的變數加進模型，然後重複這步驟，如果剩餘變數P-Value都大過進入門就停止。

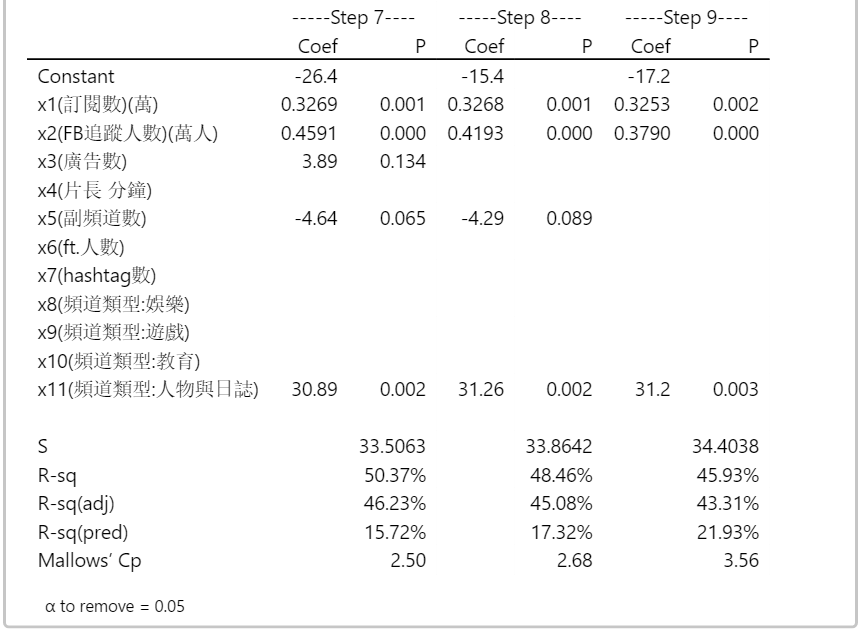
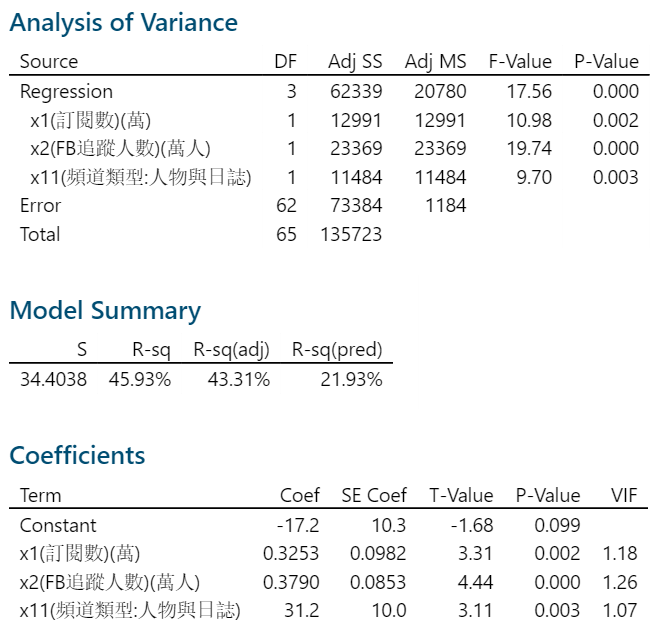
****

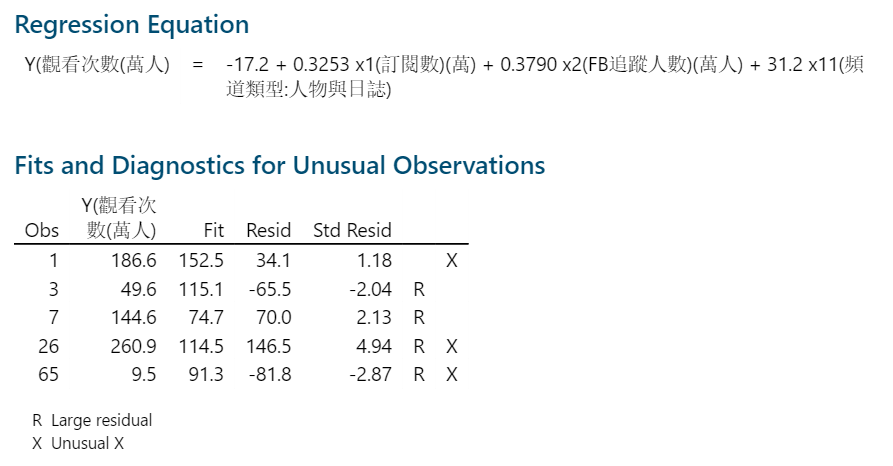
結果分析：

從向前選擇法（Forward Selection）的結果顯示中，可發現其最後模型結果和逐步選擇法是一樣的，都是Youtuber的頻道訂閱數（）、facebook追蹤人數（），及頻道類型：人物與日誌（），且R-sq(adj)同樣為43.31％。

* **後退消除法(Backward Elimination)**

****剛開始先把移出門檻（）設定成0.05。起初先將所有變數皆放在模型裡，接著把P-Value最大且超過移出門檻的變數移除，然後重複此行為，當模型中的變數其P-Value皆不超過移出門檻時，即可終止。

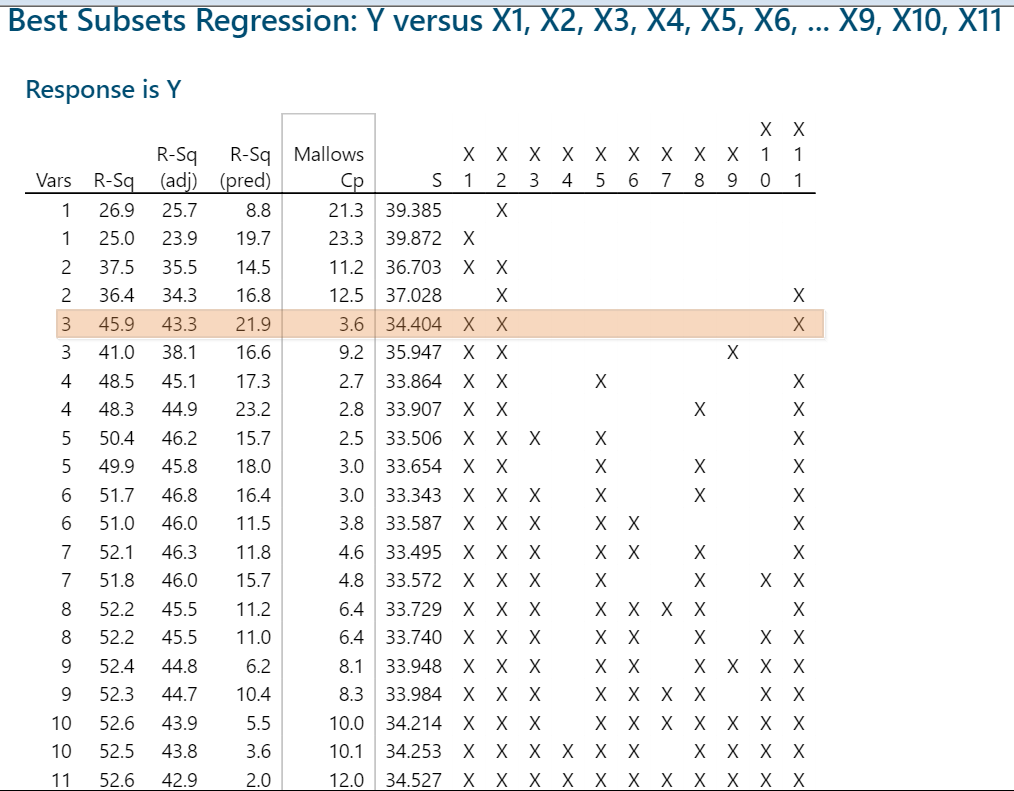
****

****

結果分析：

觀察後退消除法（Backward Elimination）的結果數據後，可以明顯看出此結果與上述的逐步選擇法及向前選擇法一模一樣，無論是在最終模型或是R-sq(adj)，皆為Youtuber的頻道訂閱數（）、facebook追蹤人數（），以及頻道類型：人物與日誌（）這三個變數且有43.31％的R-sq(adj)。

* **最佳子集迴歸****(Best Subsets Regression)**

****電腦由一個自變數增加至挑選之十一個自變數，每增加一個自變數，電腦會選擇R-sq較大之兩個模型，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。

結果分析：

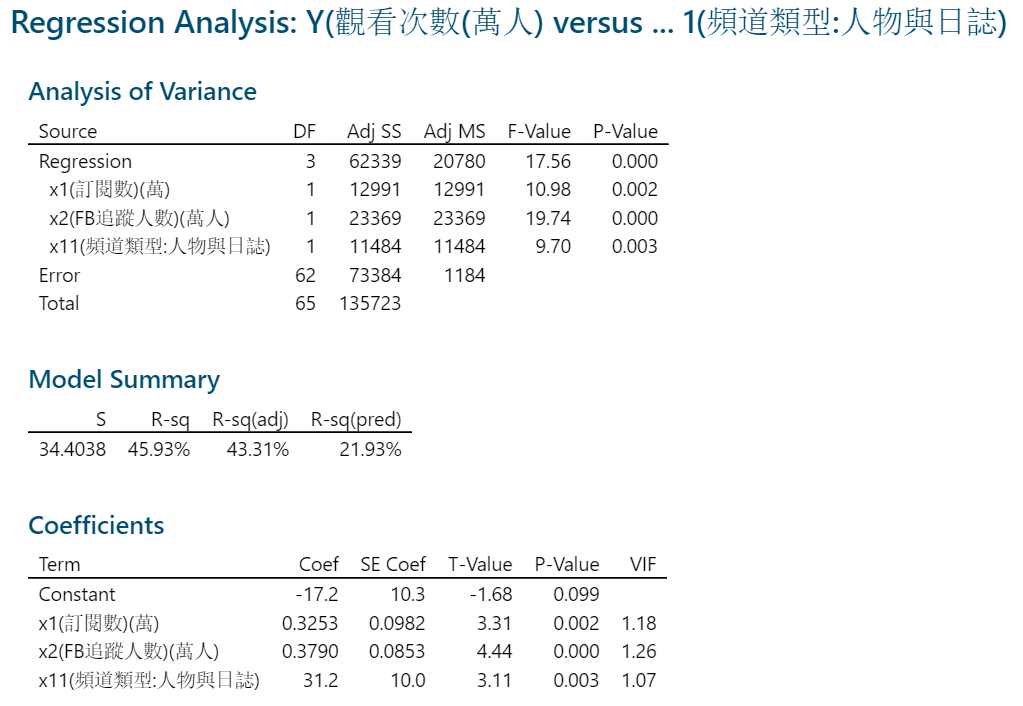
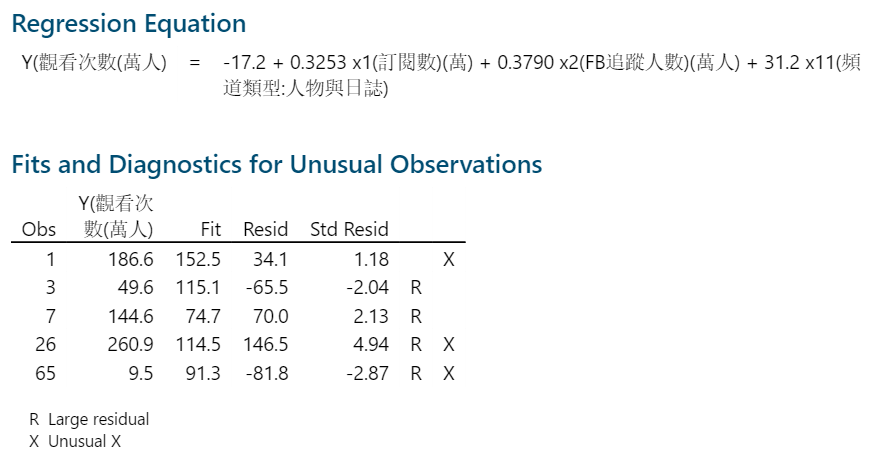
先從R-sq(adj)較大但自變數相對較少的模型中做挑選，之後作研究即可用相對較少的收集成本，達到差不多的效果，因此可發現擁有三、四個變數之模型的R-sq(adj)不輸於擁有全部自變數的模型，再來觀察Cp值，圖中所框起模型的Cp值滿接近自變數加上常數的個數。

* **模型說明**

根據 逐步選擇法(Stepwise Selection)、向前選擇法(Forward Selection) 和 後退消除法(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 **(頻道訂閱數)、 (facebook追蹤人數)**和 **(頻道類型：人物與日誌)** 這三個自變數有關。而透過最佳子集迴歸(Best Subsets Regression)的數據顯示中，我們可以觀察到此模型的值為3.6，相較於其他模型的值，相對接近其p值(p＝4)。加上，其他的模型因為值與其p值相差較大，或是變數增加但R-sq(adj)值卻沒有明顯的提升等因素考量，例如：變數個數為6的第一筆資料，雖然它的R-sq(adj)上升了3.5，但多了三個變數，且值與p值差了4，因此我們最終選的模型為：

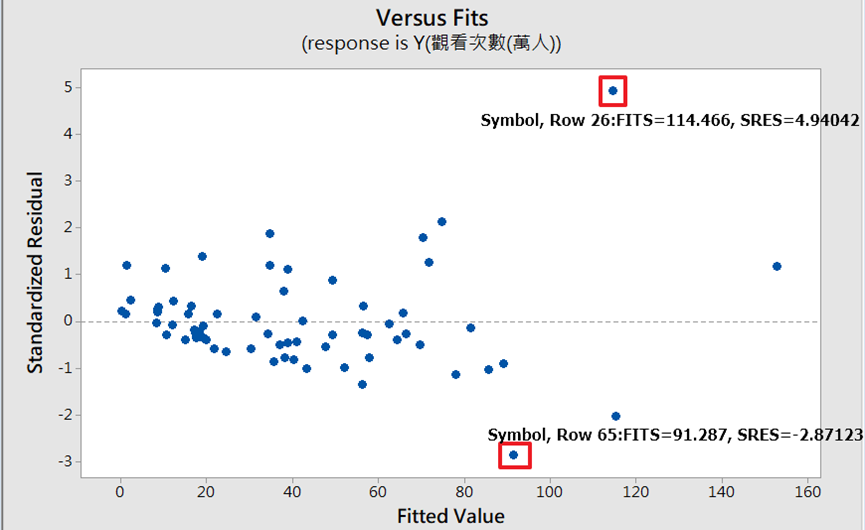
**模型迴歸分析**

選擇模型後，將對66筆數據進行迴歸分析，觀察自變數與應變數間的相關性及變化量，並且透過迴歸分析，藉由自變數來估計應變數的值，也就是說，給定頻道訂閱數、facebook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。

* **初步迴歸分析**

分析：

藉由F檢定可以發現有很大的機會模型中自變數的係數至少有一個非零，且根據t檢定可以知道自變數與應變數之間是有關係的。再者，變數各別的P-Value都相當小，因此更加確認模型是顯著的。從各別的VIF值得知自變數之間沒有存在共線性關係。從分析數據可看出第3、7、26、65筆資料的標準化殘差較大，因此將在後頭藉由擬合值比較圖來觀察這四筆資料是否為離群值。

* **擬合值比較圖(Versus Fits)**

分析：

藉由此圖，可以發現有其中兩點，分別是第26筆資料和第65筆資料，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，因此可做為往後判斷是否為離群值的依據。而這些點的分佈有向右開的趨勢(right-opening magaphone)，可以得知標準化殘差的變異數並非常數(nonconstant variance)，這表示其變異數可能會和應變數(*Y*)、單個或是多個自變數()，甚至是其他因素有關聯。

* **刪除離群值(outlier)之迴歸分析**

針對可能為離群值的數據，將探討是否有什麼背景資料沒被考慮到，因此導致較嚴重的誤差，如果存在這樣的因素，即可將此數據視為離群值刪除，然後再次做迴歸分析，看有沒有藉此改善我們的迴歸模型。

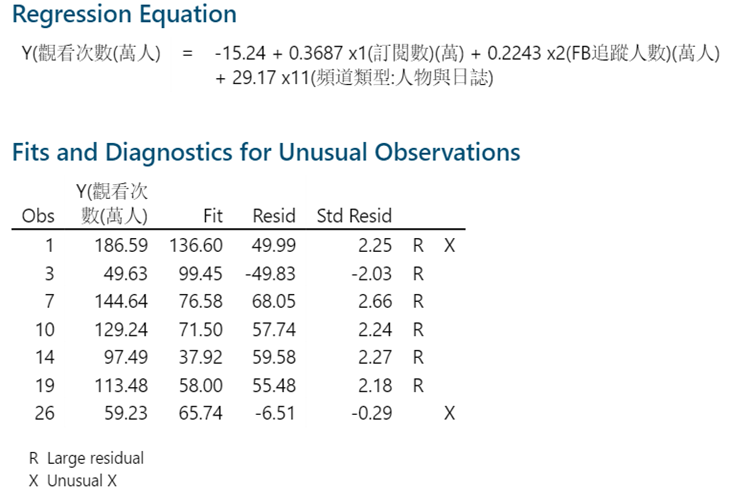
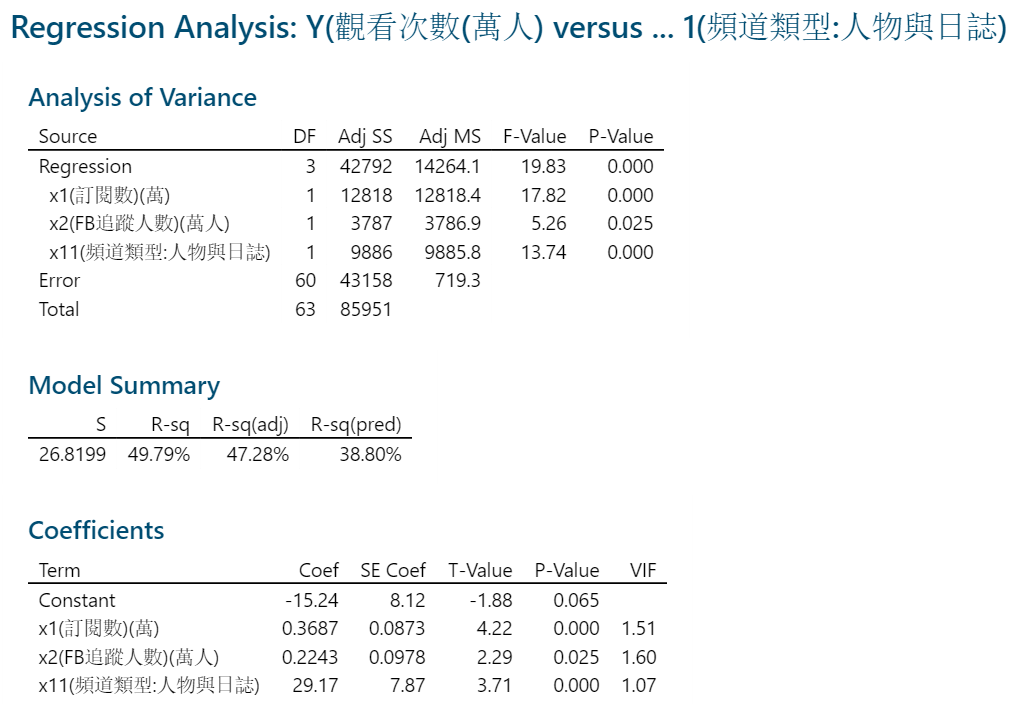
* **刪除原因**

第26筆資料為兩位年輕女生跳舞的頻道，觀察其餘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BLACKPINK - 'Kill This Love' Dance cover，而BLACKPINK為當紅韓團，當時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，並且可以透過搜尋發現，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，決定將它刪除以降低殘差。

另外，也刪掉了第65筆資料，而此影片為那對夫妻的回顧影片，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。

(已刪除資料的後方資料皆自動往前排序。)

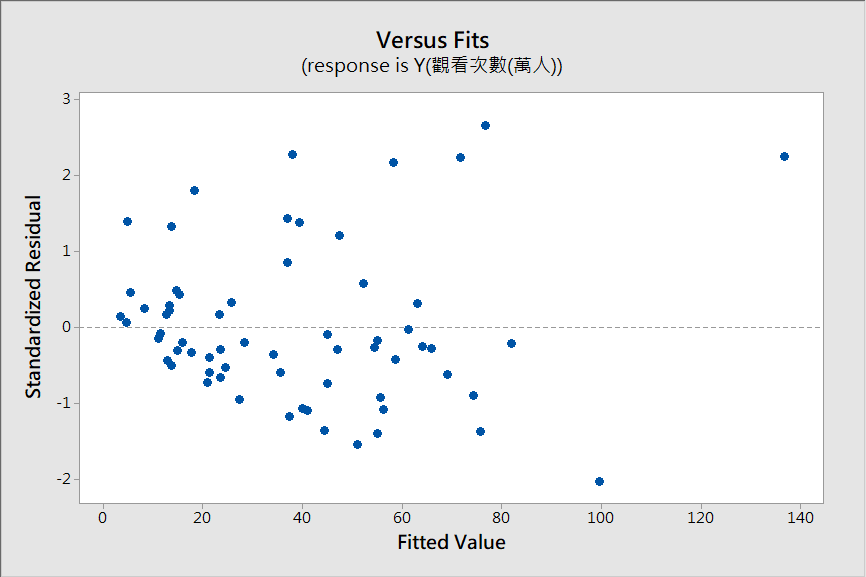
* **迴歸分析**

****透過對刪除離群值後的模型做迴歸分析，觀察刪除離群值對模型的影響，及檢查是否還存在其他潛在之離群值。

分析：

藉由F、T檢定得知刪掉離群值後的模型依然是顯著的。再來可發現R-sq(adj)從43.31%提升至47.28%。而透過VIF可以知道，自變數依然保持沒有共線性的關係。然而，從分析的結果顯示出第1、3、7、10、14、19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，相較於還未刪除第26筆和第65筆資料時，更縮小了一些，且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，因此不打算將這六筆資料視為潛在的離群值。

* **擬合值比較圖(Versus Fits)**

接著將透過擬合值比較圖（Versus Fits），來觀察是否標準化殘差的變異數為常數，因起初迴歸假設包含了標準化殘差的變異數為常數，如果不符假設，則需考慮做變數變換來改善。

分析：

刪掉離群值後，發現標準化殘差範圍縮小至(-2,3)左右，但觀察點的分佈依舊有右開的趨勢，因此可以得知標準化殘差的變異數並非常數，藉此，透過變數轉換以符合標準化殘差的變異數為常數的迴歸假設。

* **應變數之線性轉換**

我們考慮三種常見的轉換方式：

分別是對Y開根號、對Y取ln、取Y的倒數。

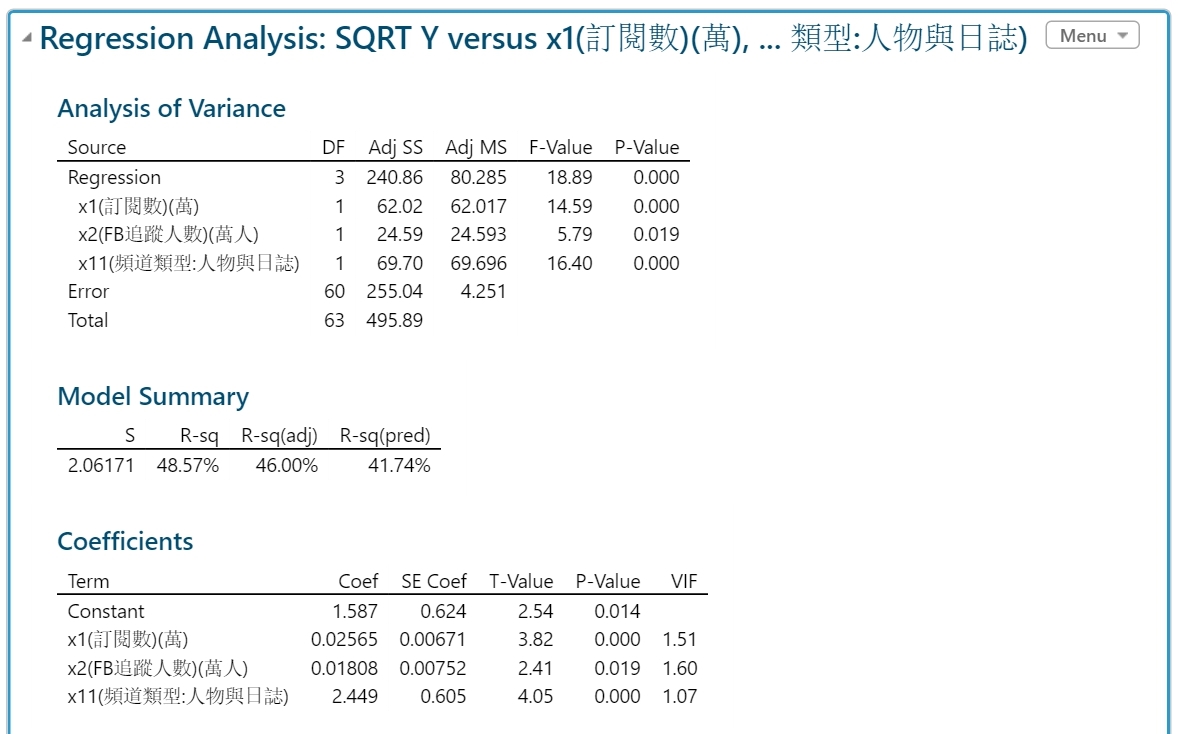
下列表格為Y與經過三種不同轉換後的Y的比較示意圖。

* **擬合值比較圖(Versus Fits)對照**

|  |  |
| --- | --- |
|  |  |
|  |  |

（左上：Y 、 右上： 、左下：lnY 、右下：）

由上圖可觀察出，ln Y和 的圖並未改善原本點的分佈，而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，代表此轉換之模型符合迴歸假設。

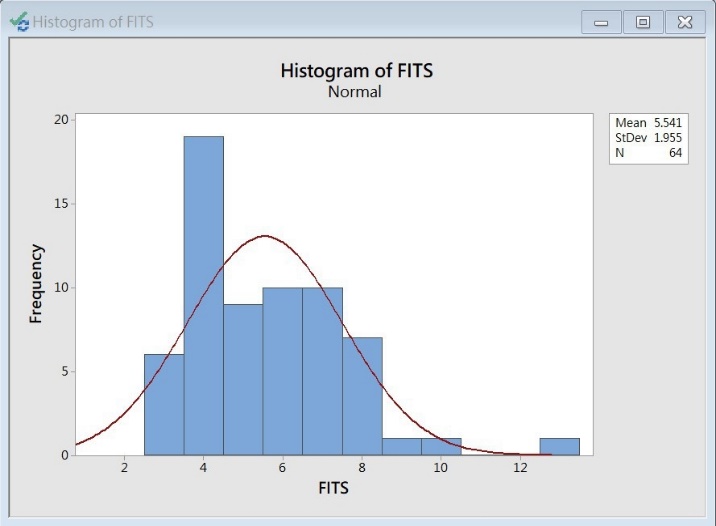
* **轉化後的迴歸分析**

**的迴歸分析**

分析：

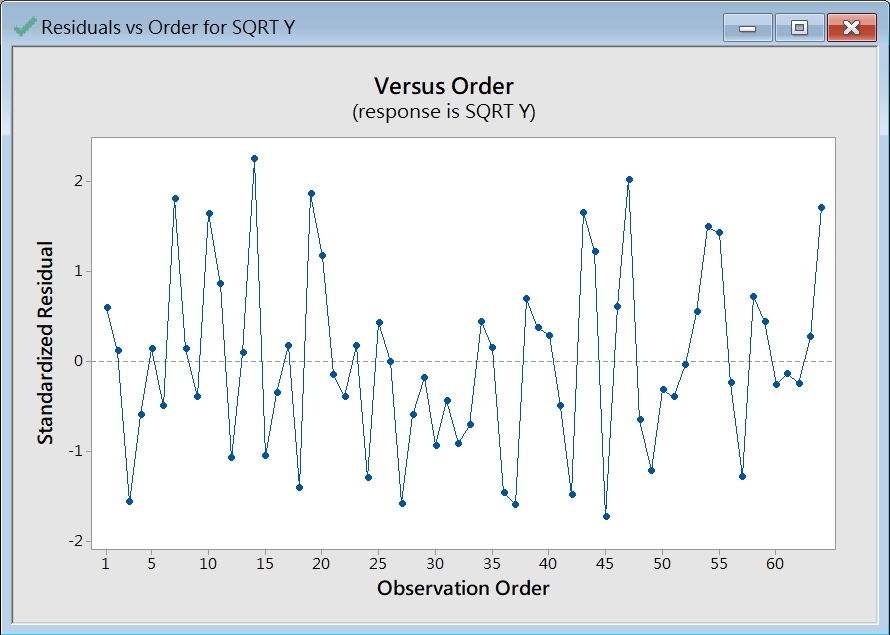
透過P-Value，我們得知轉換為的模型是顯著的，而且可以從VIF觀察到變數間依舊沒有共線性關係。

**擬合值之長條圖(Histogram of FITS)**

分析：

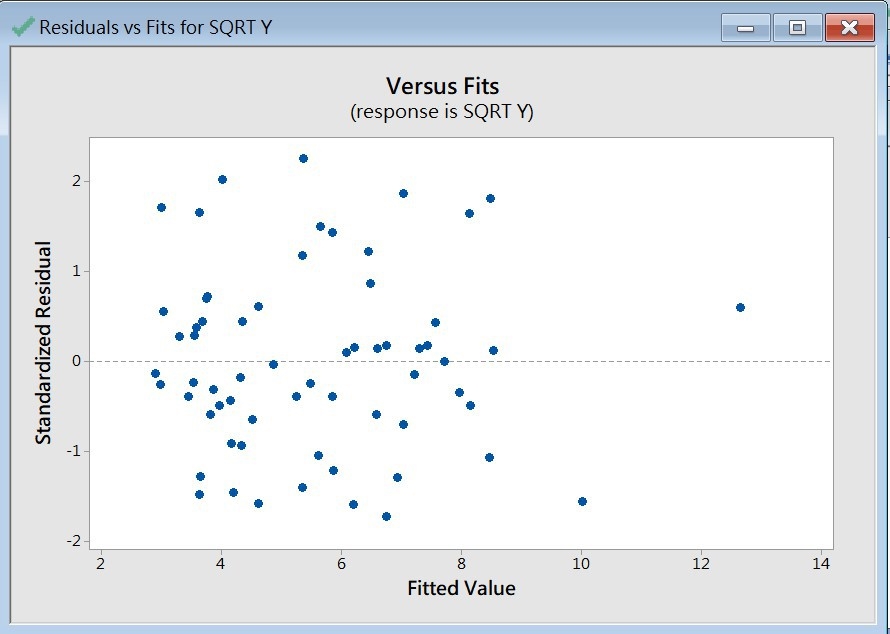
透過擬合值的長條圖中，可以觀察到擬合值並非完整呈現標準的常態樣貌，也就是未明顯表現出鐘形的模樣，例如：橫軸的FITS值的頻率相較於其他的高出了許多，或是FITS值在10~12之間有空缺，整體上也有向右偏的趨勢，可以知道資料並非呈現常態，但由於觀察資料在整體上的分布，並沒有某一個擬合值特別偏離其餘的資料，因此沒有進一步要對模型做其他變數變換的調整。

**次序比較圖(Versus Order)**

分析：

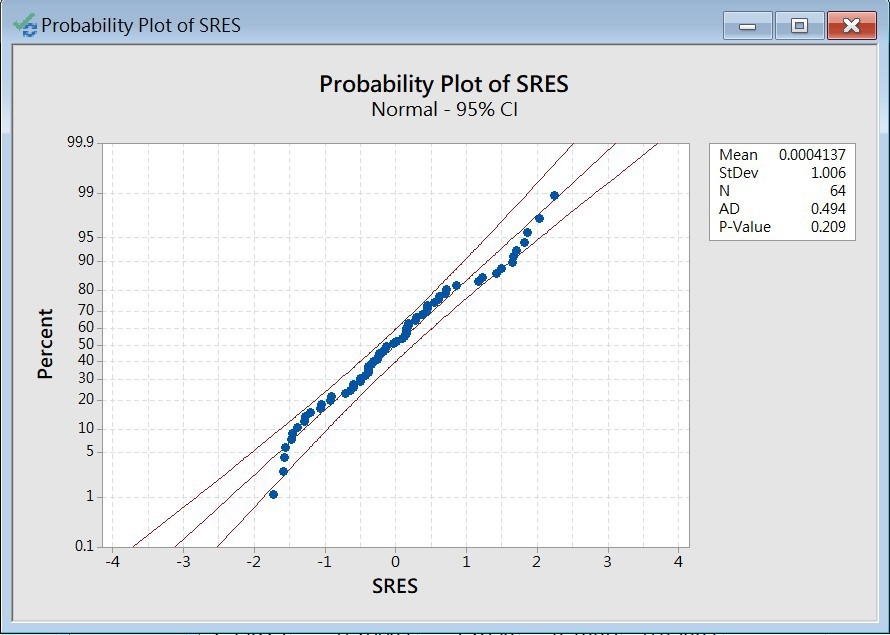
在做完變數變換後的標準化殘差範圍依舊大約位於(-2,2)左右，可觀察出此圖並沒有任何一個固定走向，是任意上下起伏的，因此得知此份數據是隨機的，不會因為排列次序的增加而標準化殘差就隨之有趨勢性的上升或下降。

**擬合值比較圖(Versus Fits)**

分析：

觀察 的擬合值比較圖的分布狀況，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。

**標準化殘差之機率圖(Probability Plot of SRES)**

分析：

透過觀察 標準化殘差之機率圖，可以知道P-Value的值增加至0.209，相對顯著，且再加上95%的信賴區間帶可以更明顯知道，標準化殘差幾乎都被信賴區間帶給覆蓋住，因此，符合標準化殘差是常態的假設。

**最終模型與解釋**

藉由上述的迴歸分析、圖形，可發現的表現不管是在四種圖形上還是迴歸分析，皆符合期待，並做此轉換後，修正了原本未符合假設的情況，因此最終選擇的模型為

**【模型解釋】**

* **觀察訂閱人數()增加1單位，影片觀看次數()會增加多少：**

固定 facebook追蹤人數() 和 頻道類型：人物與日誌() 的值

設為代所得結果

當增加1單位時的變化量為

範例 :

當 facebook追蹤人數()為80萬人

頻道類型不為人物與日誌()

若 訂閱人數()為80萬人，則訂閱人數()增加1萬人時，影片觀看次數()會增加約2615次。

<Sol.>

(萬次)

若 訂閱人數()為90萬人，則訂閱人數()增加1萬人時，影片觀看次數()會增加約2747次。

<Sol.>

(萬次)

* **觀察facebook追蹤人數增加1單位，影片觀看次數()會增加多少：**

固定 訂閱人數() 和 頻道類型：人物與日誌() 的值

設為代所得結果

當增加1單位時的變化量為

範例：

當 訂閱人數()為70萬人

頻道類型不為人物與日誌()

若 facebook追蹤人數()為60萬人，則當該頻道追蹤人數()增加1萬人時，影片觀看次數()會增加約1617次。

<Sol.>

若 facebook追蹤人數()為90萬人，則當該頻道追蹤人數()增加1萬人時，影片觀看次數()會增加約1812次。

<Sol.>

* **頻道類型是否為人物與日誌()對影片觀看次數()的影響：**

設所得結果，

設為所得結果，則

頻道類型是「人物與日誌()」較「非人物與日誌()」的影片觀看次數()增加量為

範例:

當 訂閱人數()為100萬人

facebook追蹤人數()為70萬人

則 頻道類型為人物與日誌() 相較 頻道類型非人物與日誌()影片觀看次數(Y)增加約325090次。

<Sol.>

* **結論**

從迴歸分析結果，可以得知 Youtuber的頻道訂閱數() 、 Youtuber的Facebook追蹤人數() 以及 頻道類型:人物與日誌() ，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Facebook追蹤人數，且拍攝人物與日誌類型的影片，相較於其他類型的影片會增加較高的觀看次數，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。

**心得感想**

【洪羽柔】

透過這次的報告，增加了我對統計的了解與熱情，並且接觸了從未觸碰過的MINITAB，更加明白迴歸分析要如何運用在生活實際應用裡，雖然一開始光是確定主題就遇到了很大的問題，而中間過程中也發生許多狀況，但在與組員們和老師的多次討論後，終於完成這次的報告，也許成果看似簡陋或是操作上輕而易舉，但其實都花了大家許多精力與時間，不過最重要的是我們都在這個過程中一起學習到團隊合作、溝通協調、問題處理能力、當然還有統計相關的背景知識，謝謝老師以及組員們的指導與合作。

【艾品璇】

在這次的分組報告有很不一樣的經驗，學習到許多事，例如透過這次報告，知道如何操作MINITAB，也藉由MINITAB加深了課堂上所學習到的知識，雖然過程並不是那麼的順利，從一開始找主題就四處碰壁，中間在操作上也發生一些錯誤，導致分析結果是錯的而重新來過，但經過老師耐心的講解，這些問題最後都解決了。也因為這次報告發現每個人都有各自擅長的地方，有的擅於美化ppt，有的善於修飾文字，很開心能和他們一起完成這次的報告。

【陳晏琦】

透過這次的專題報告，讓我學習到統計相關的資料分析，從一開始找主題、蒐集相關的資料，到使用MINITAB進一步的數據分析，從中的每一個過程，我一邊學習新的內容，一邊實際套用在我們的主題上。而Youtube觀看次數這個主題，對於我而言是相當生活化的，因為平時我也是看Youtube居多，因此在做相關分析時會更有興趣，也因此領悟到什麼因素會去影響觀看次數。雖然最後做出來的模型在R-sq的表現上不是非常突出，但在做報告的過程中，對於一些統計的圖表、專有名詞都更加了解，因為未來我也是要朝統計這方面繼續鑽研，因此這樣的學習經驗是我覺得收穫最多的地方。

【顏均翰】

第一次作統計相關的實驗報告，而在這報告過程中，發現原來要做出一個漂亮的迴歸，是一件不容易的事，但利用MINITAB軟體讓我們能夠去找出一個比較合適的迴歸，而我們也是初次學習MINITAB，很慶幸我的組員們!真得很厲害透過合作，一起思考，完成了這一份報告，雖然過程中有些不順遂，但透過詢問老師也一步一步帶我們慢慢解危，因此在這份期末報告學得了一些基本MINITAB的操作並能簡單分析相關內容，並也深深感謝我的組員們，辛苦了!

**資料來源**

YouTube：<https://www.youtube.com/>

FaceBook：<https://www.facebook.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>