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** 2](#_Toc20687475)

[**資料介紹** 3](#_Toc20687476)

[ **變數介紹及說明** 3](#_Toc20687477)

[ **範例說明** 4](#_Toc20687478)

[**模型挑選** 8](#_Toc20687479)

[ **Stepwise Selection** 8](#_Toc20687480)

[ **Forward Selection** 9](#_Toc20687481)

[ **Backward Elimination** 10](#_Toc20687482)

[ **Best Subsets Regression** 12](#_Toc20687483)

[**迴歸模型分析** 13](#_Toc20687484)

[ **初步檢視迴歸分析** 13](#_Toc20687485)

[ **離群值(outlier)分析** 14](#_Toc20687486)

[ **檢查迴歸假設** 16](#_Toc20687487)

[**最終模型與解釋** 20](#_Toc20687488)

[**資料來源** 24](#_Toc20687489)

**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)** |
| ***Y*** | Youtube影片觀看次數(萬次) |
|  | Youtuber的頻道訂閱數(萬人) |
|  | Youtuber的臉書追蹤人數(萬人) |
|  | 影片的廣告數(個) |
|  | 片長(分鐘) |
|  | 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**

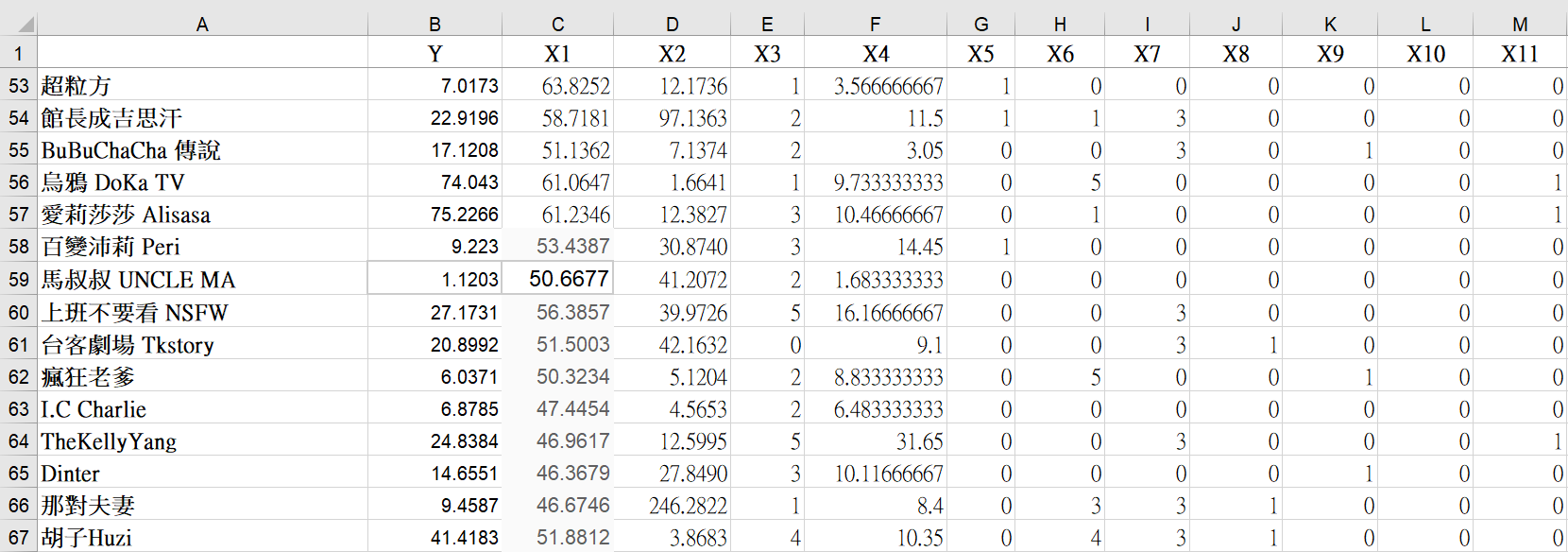
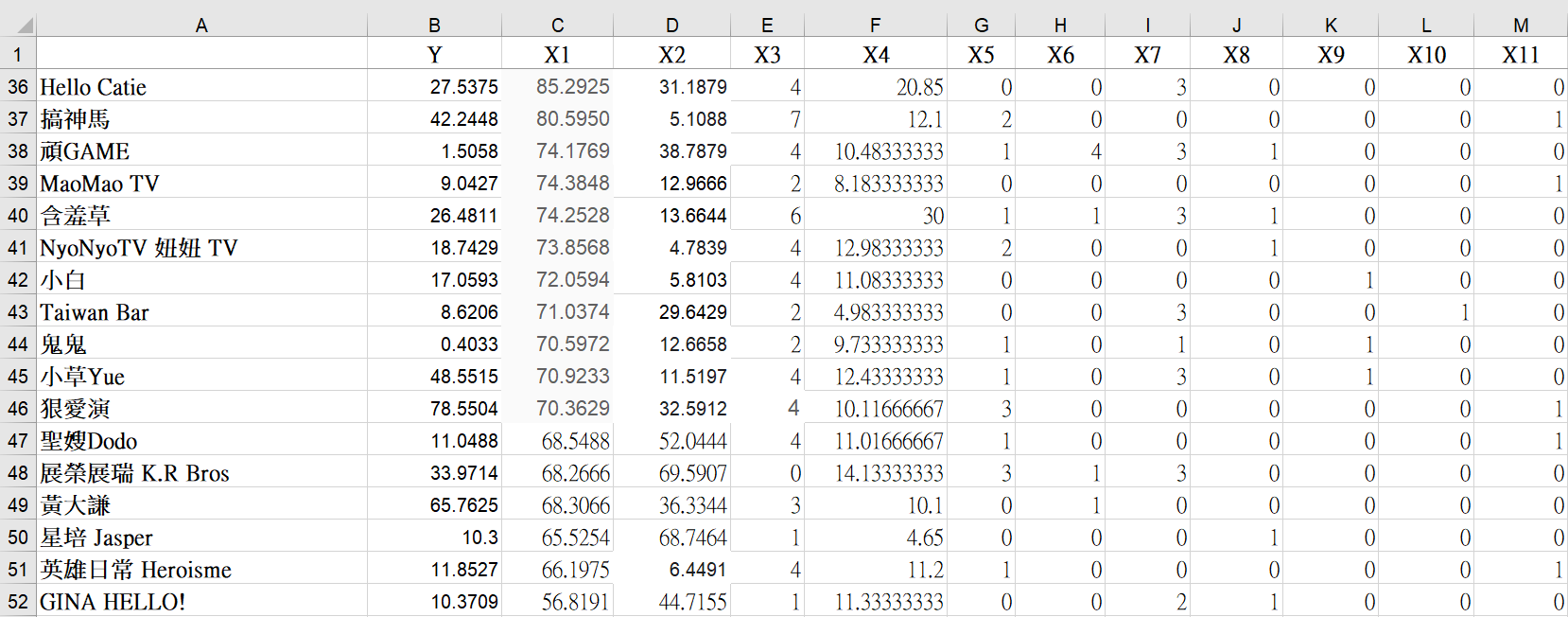
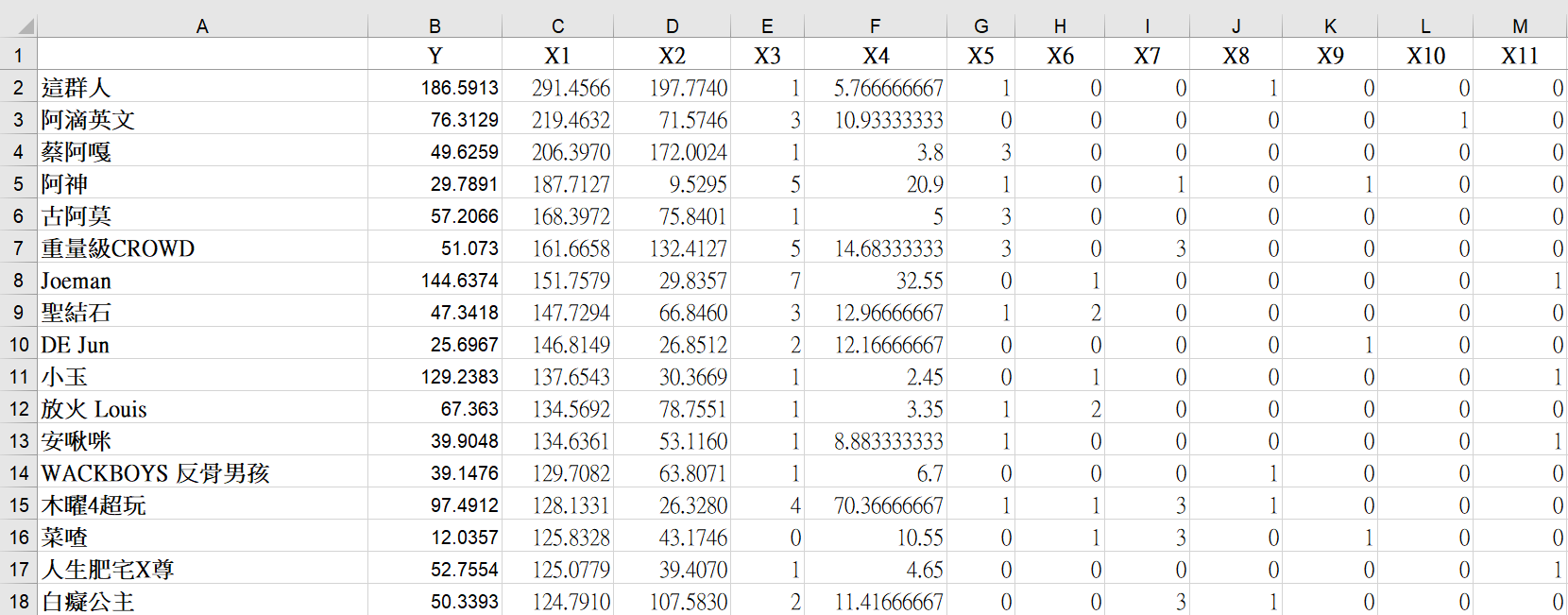
****首先，開啟臉書，並在搜尋的地方輸入Youtuber名稱，進入其首頁後，在右下角即可看到此Youtuber的臉書追蹤人數()。

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()，再來點選【頻道】，也可看出其副頻道數()。

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()，而時間軸上的黃色點個數即為廣告個數()，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()。

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。

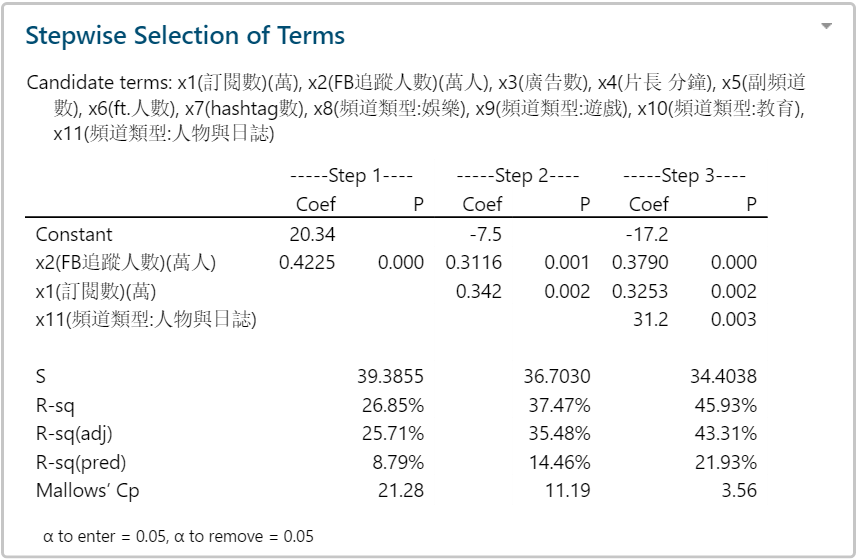
以下是擷取所得66筆資料的結果：



**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**

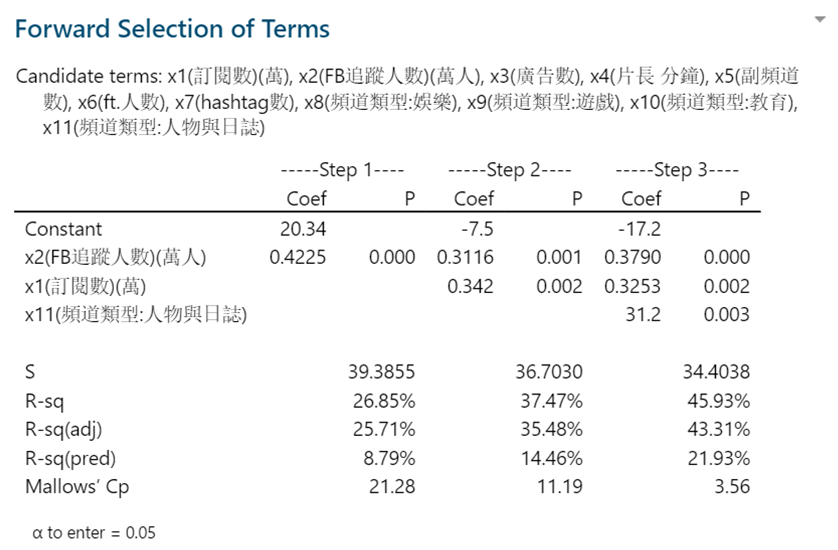
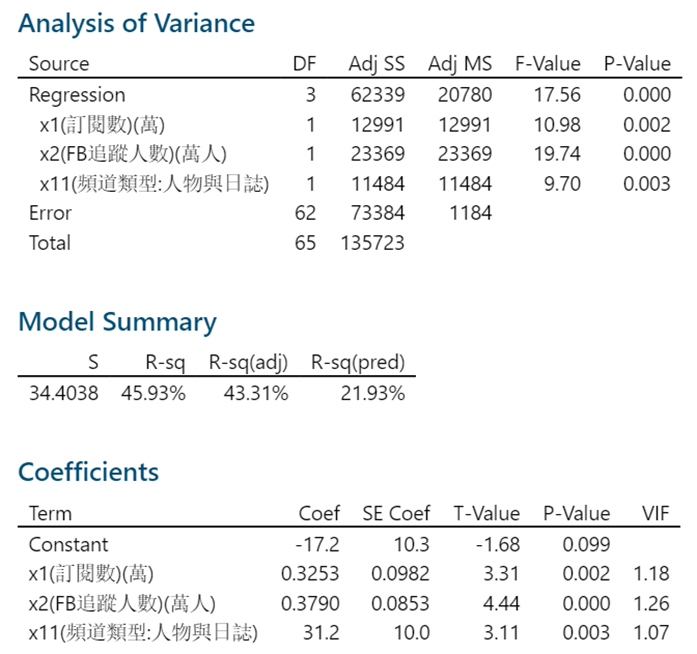
****此方法的選入門檻（-to-enter）及移出門檻（-to-remove）皆設為0.05。以下由Minitab統計軟體所呈現的選取結果。

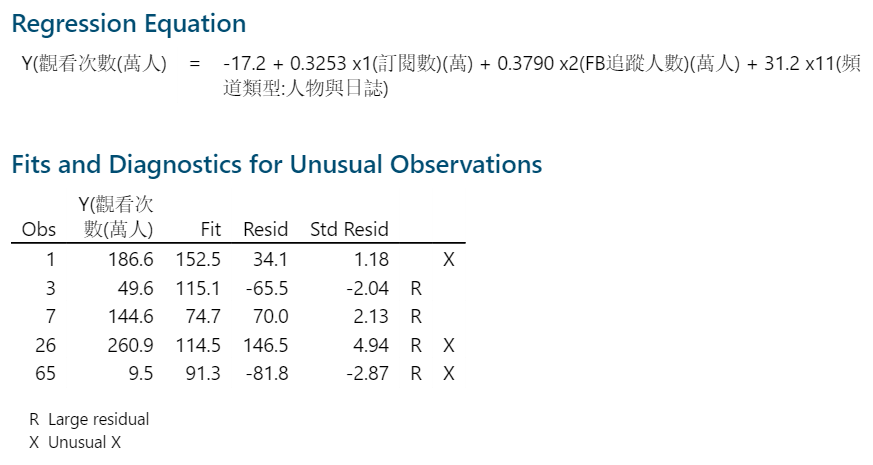


****

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（）、臉書追蹤人數（）、頻道類型-人物與日誌（），且過程中沒有刪除任何前一階段已加入的變數。

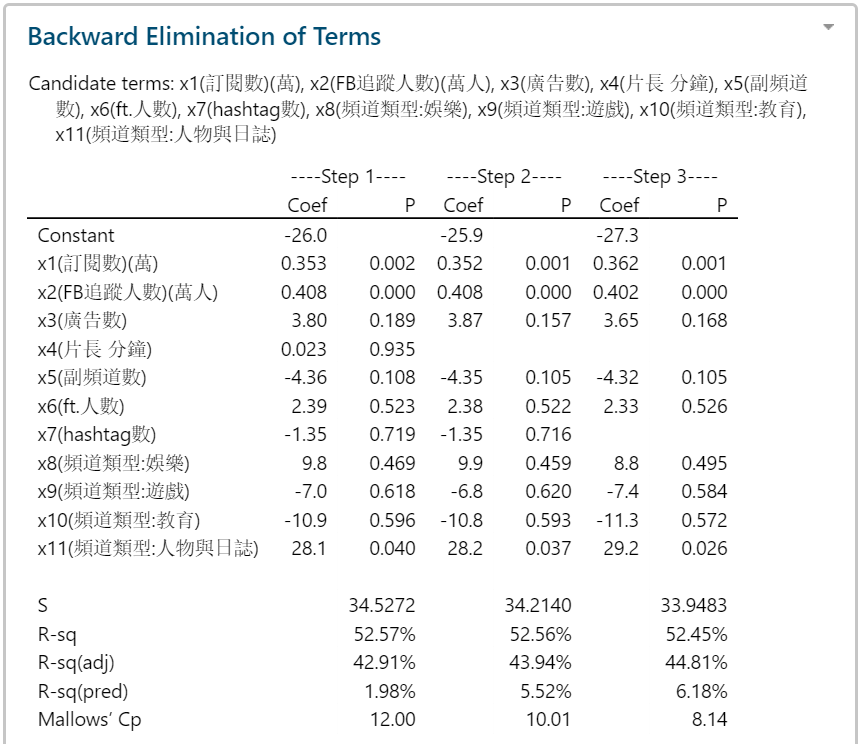
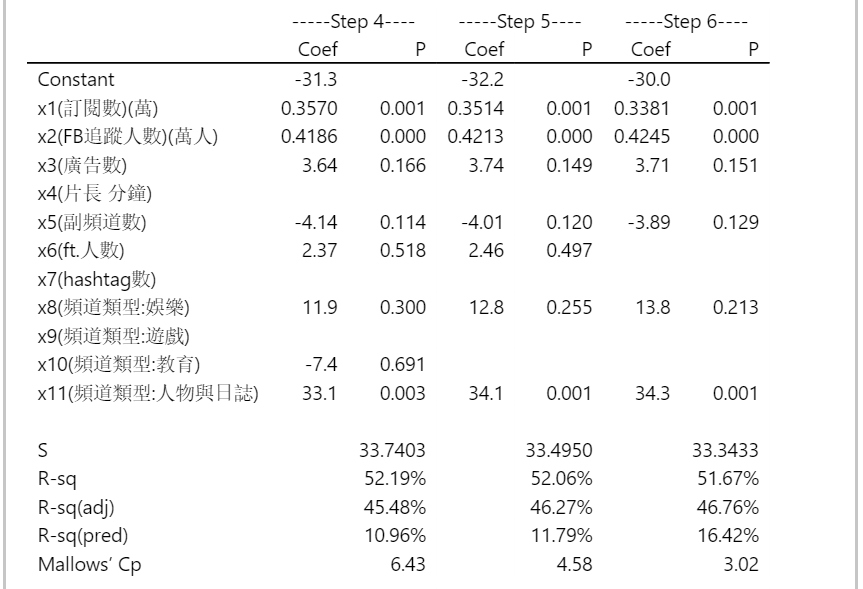
* **Forward Selection**

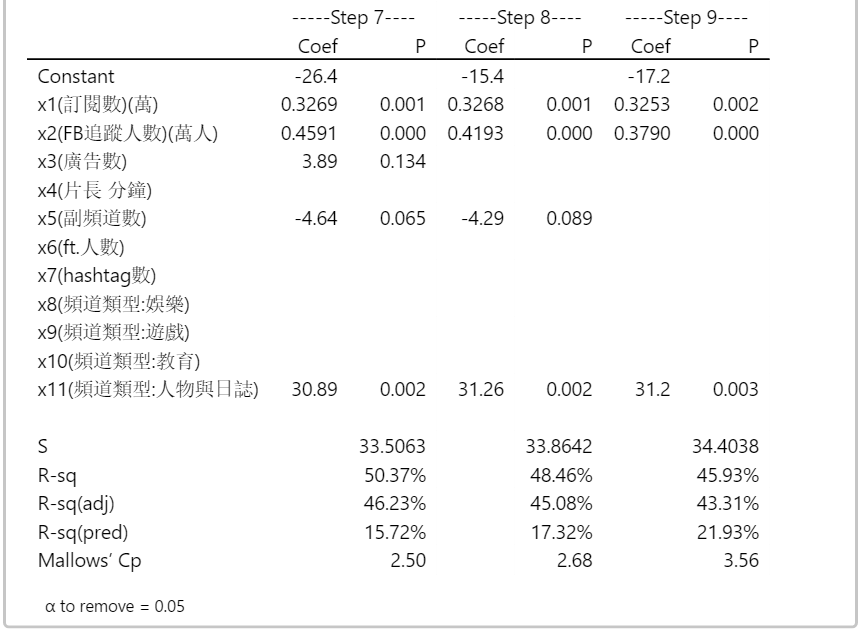
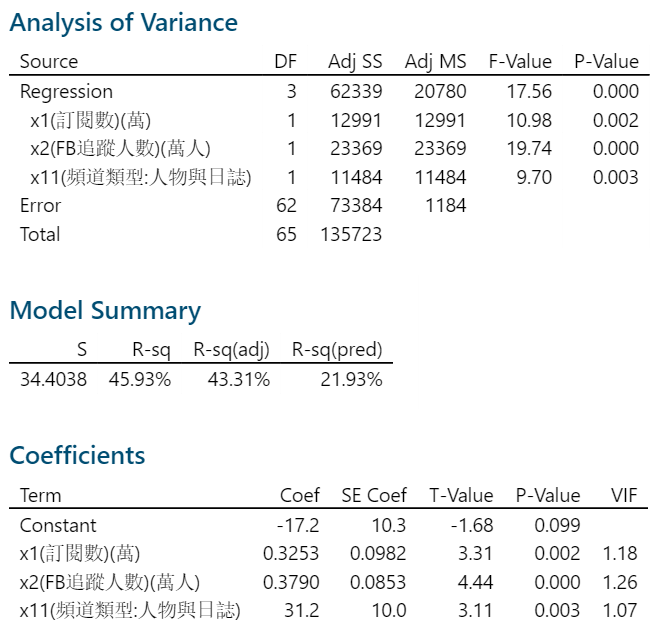
同樣的先將設進入門檻（-to-enter）定為0.05。

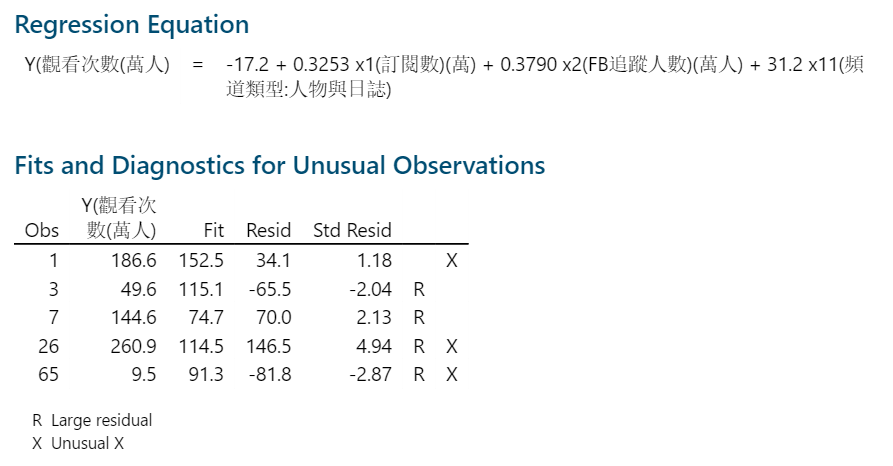
****

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**

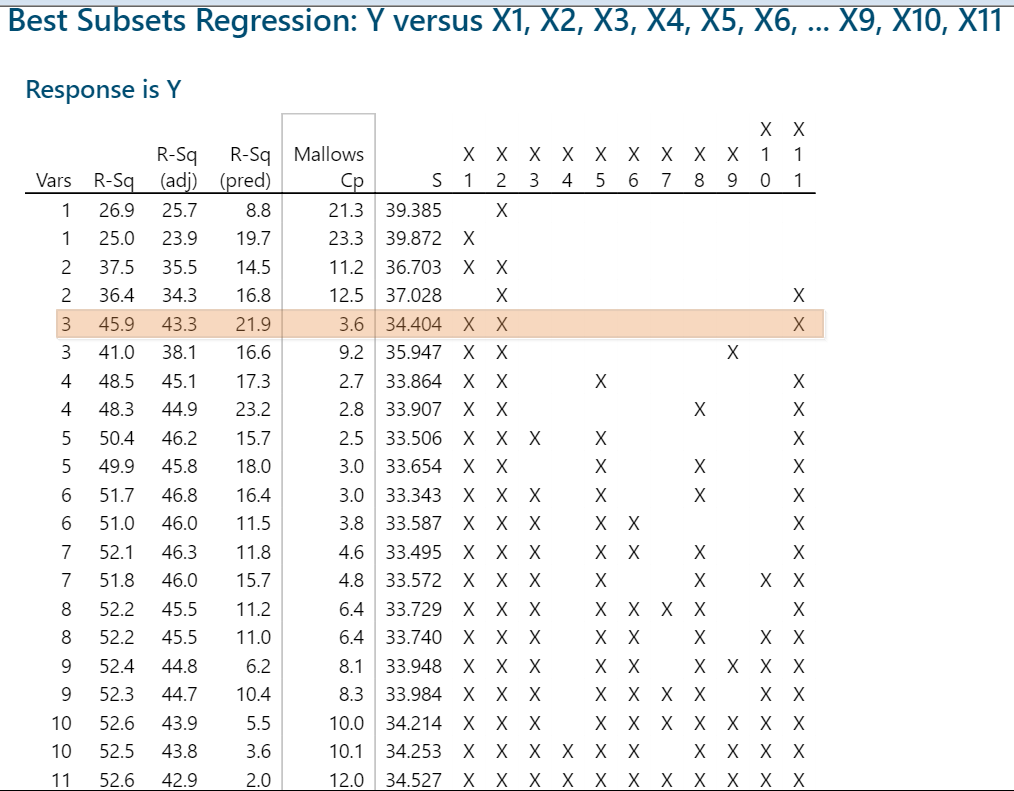
****應用此法將移出門檻（-to-remove）設定成0.05。

****

****

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（）、臉書追蹤人數（），以及頻道類型-人物與日誌(）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。

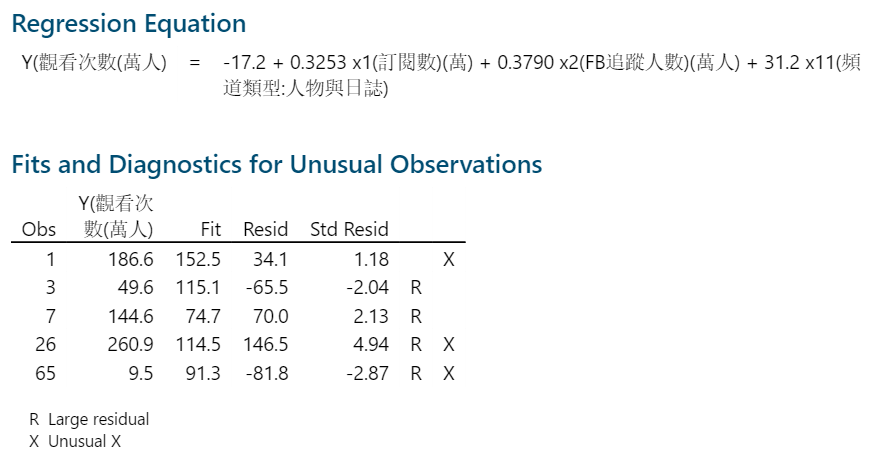
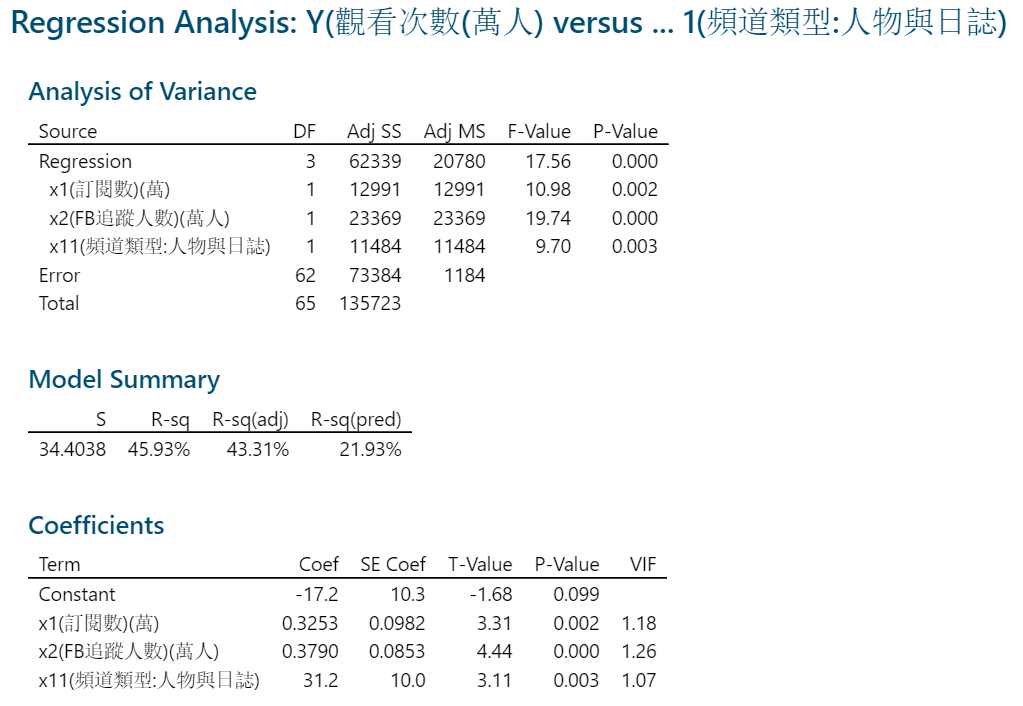
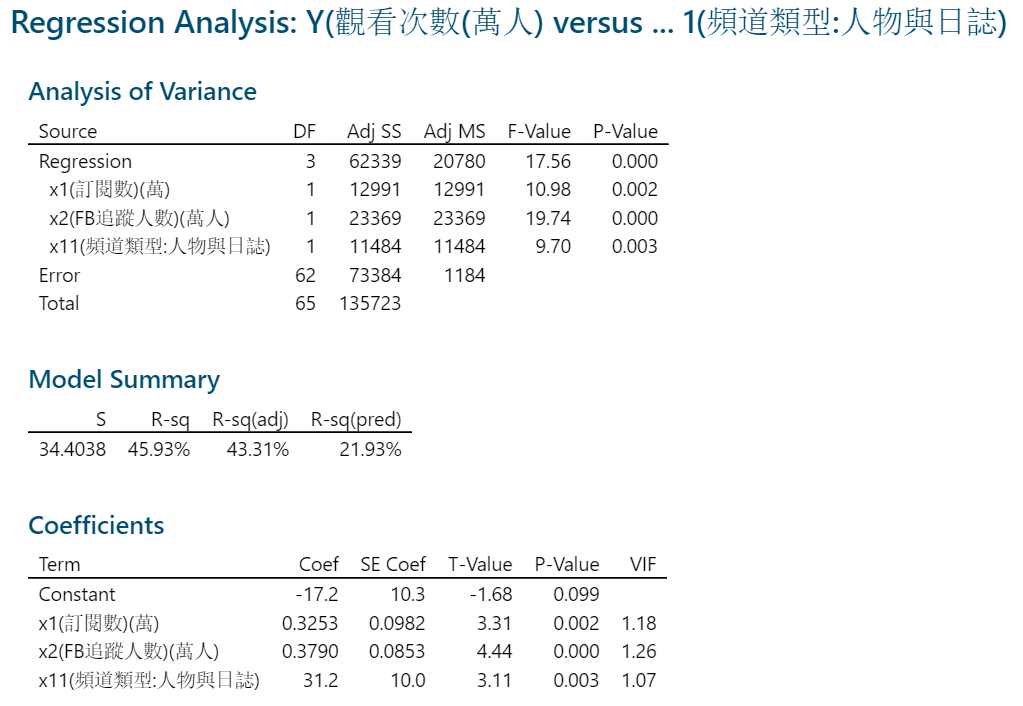
先從R-sq(adj)較大但自變數相對較少的模型中做挑選，代表可用相對較少的成本，達到差不多的效果。因此首選圖中所框起的模型，該值為3.6，比較接近其p值(p=4)，其他的模型變數增加但R-sq(adj)卻沒有明顯的提升，舉例來說，變數個數為6的第一筆資料，雖然多了三個變數，但它的R-sq(adj)僅僅上升了3.5，但，且值與p值差了4。再根據Stepwise Selection、Forward Selection和Backward Elimination 三種挑選模型的方式中，分析出的模型應變數*Y* (影片觀看次數)都是與 (頻道訂閱數)、 (臉書追蹤人數)和 (頻道類型-人物與日誌)這三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

**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。

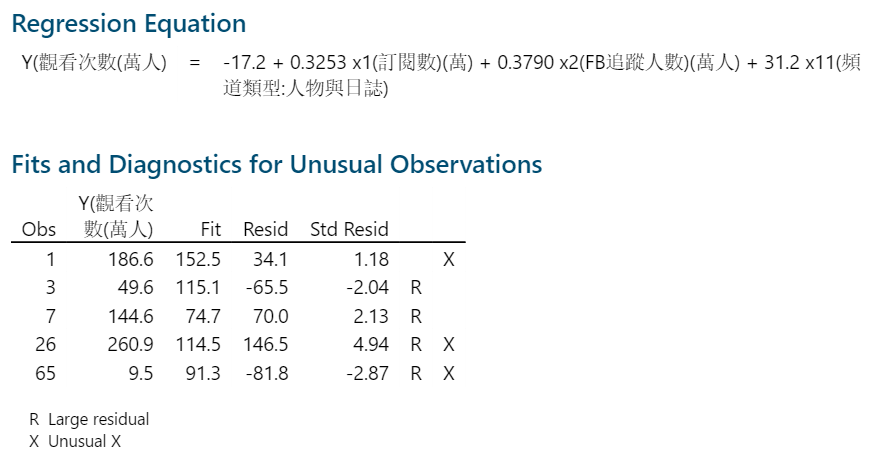
* **初步檢視迴歸分析**

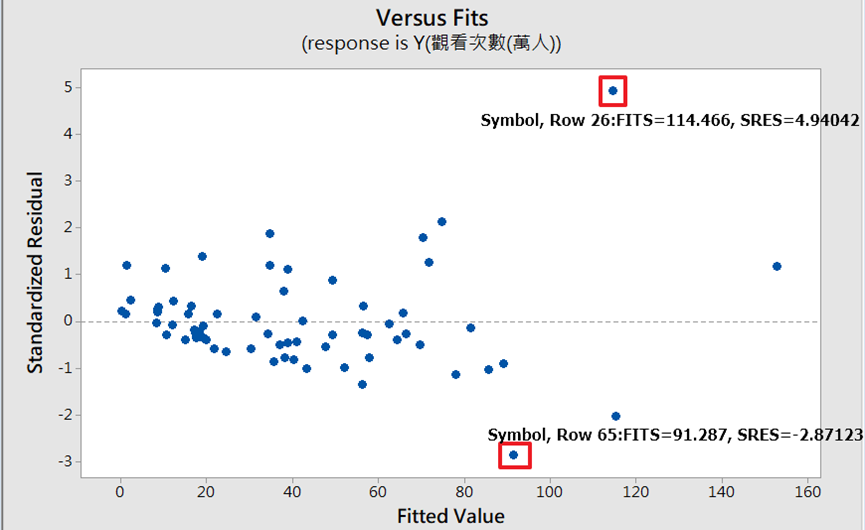
對模型進行迴歸分析以檢查此模型是否顯著。



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 (頻道訂閱數)、 (臉書追蹤人數)和 (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，、、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()、1.26()、1.07()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**

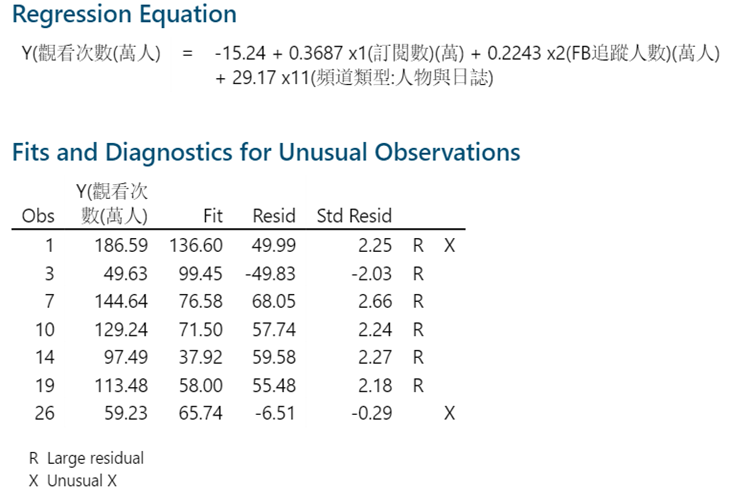
****接下來，將觀察是否有離群值的存在。

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。

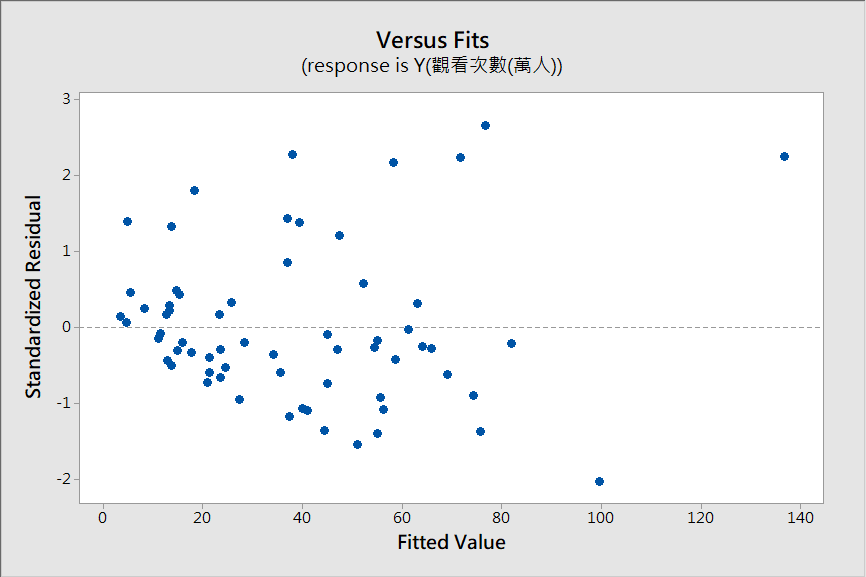
再透過觀察背景資料發現，第26筆資料為兩位年輕女生跳舞的頻道，有別其他65筆資料的原始背景，都沒有以跳舞為主軸的頻道，再加上該影片的拍攝主題為當紅韓團BLACKPINK - 'Kill This Love' Dance cover，且Kill This Love這首歌相當膾炙人口，可以發現這首歌在Youtube官方平台的影片觀看次數高達3億次，並可以透過搜尋發現，翻跳這首歌的影片觀看次數也都相當可觀，因此視第26筆資料是會影響迴歸模型的離群值，決定將它刪除。

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。

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。

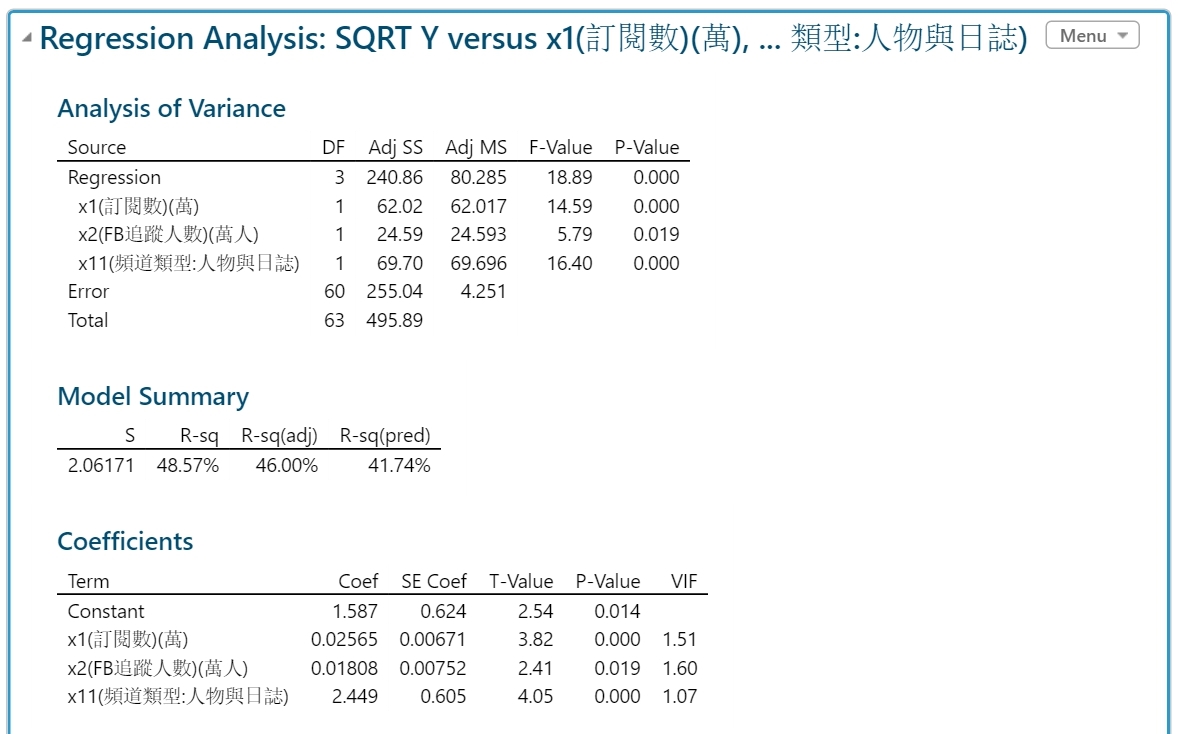
觀察點的分佈有逐漸向右展開的趨勢，因此可以得知標準化殘差的變異數並非常數，藉此，透過對Y做變數轉換以符合標準化殘差的變異數為常數的迴歸假設。

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，左下圖是lnY，而右下圖則是的情形。

**Standard residuals vs. fitted values對照圖**

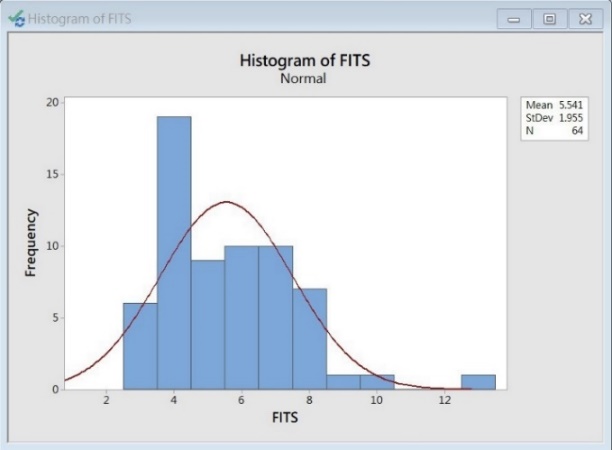
|  |  |
| --- | --- |
|  |  |
|  |  |

由上圖可觀察出，lnY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，代表此轉換之模型比較符合迴歸假設。

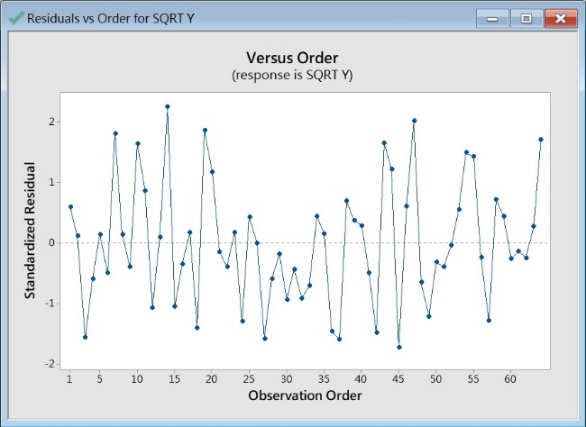
****同樣地，觀察迴歸係數的估計與檢定。

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。

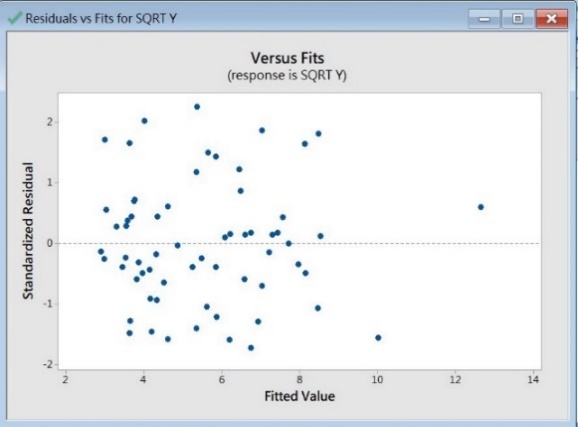
接著以此模轉換後的模型，繼續檢查其他的迴歸假設包括fitted values是否為常態分布？、標準殘差是否為隨機？、標準化殘差的變異數是否為常數？以及標準化殘差是否為常態分布等。

**Histogram of fitted values**

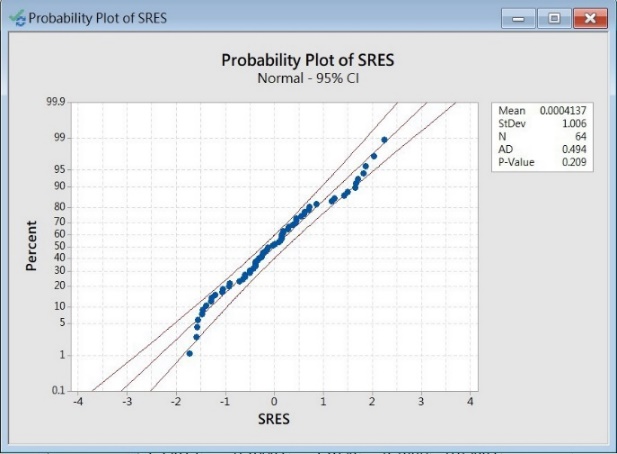
透過fitted values直方圖，可以觀察到除了fitted value值為4的部份，整體近似常態分佈，且由於觀察資料在整體上的分布，並沒有某一個fitted values特別偏離其餘的資料，因此沒有進一步要對模型做其他變數變換的調整。

**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，可發現的結果皆符合迴歸假設的情況，因此最終選擇的模型為

其中迴歸係數，，，

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()對於影片觀看次數()的影響：**

在固定臉書追蹤人數()和頻道類型(人物與日誌) 的情況之下，

考慮增加1單位，設為代所得結果

所以，當增加1單位時的變化量為

舉例來說，當固定臉書追蹤人數()為80萬人，而頻道類型不為人物與日誌()時，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()的變化為

(萬次)，也就是影片觀看次數()會增加2615次。

同理，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() 的變化為

(萬次) ，亦即影片觀看次數()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。

**觀察臉書追蹤人數對於影片觀看次數()的影響：**

在固定訂閱人數()和頻道類型(人物與日誌) 的情況之下，

考慮增加一單位，設為代所得結果

因此，當增加1單位時的變化量為

舉例來說，考量訂閱人數()為70萬人且頻道類型不為人物與日誌()時，若臉書追蹤人數()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()的變化量為

，也就是影片觀看次數()大約增加1617次。

同理，若臉書追蹤人數()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() 的變化量為

，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()對影片觀看次數()的影響：**

設所得結果，

設為所得結果，則

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

舉例來說，當訂閱人數()為100萬人且臉書追蹤人數()為70萬人時，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)的變化量為

，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() 、 Youtuber的臉書追蹤人數()以及頻道類型(人物與日誌)，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/](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>