FCU ePaper

逢甲大學學生報告 ePaper

報告題名：

電影票房之迴歸分析：

以『西遊記之大鬧天宮』為例

The Analysis of Regression on the box office of movie：

Take The Monkey King for example

作者：謝宜君、呂朋、沈允人、林宣穎、郭又禎

系級：統計學系三年甲班

學號：D0130375、D0107309、D0129902、D0172252、D0130171

開課老師：高秀蘭

課程名稱：迴歸分析

開課系所：統計學系

開課學年：103 學年度 第 1 學期

商學院

中文摘要

在生活中，看電影是很普遍的一種休閒活動之餘，在報紙裡有時候會討論到一部電影的票房成績，因此我們想知道是什麼樣的因素會影響到票房。在此報告裡，我們藉由一部電影『西遊記之大鬧天宮』為例，探討一部電影的總票房與平均票價、放映場次、觀影人次和上座率，四個解釋變數與應變數之間的關聯。

在這個報告裡，主要目的是找出所有解釋變數中，對於總票房這個應變數最有解釋能力的模型。首先，我們先對各個變數做散佈圖及相關係數表，以判讀自變數是否與應變數有線性的關係。在確立線性關係後，利用三種選取法，分別為前進選擇法(Forward Selection) 、後退選取法(Backward Method) 與逐步迴歸法(Stepwise Method)，篩選解釋變數。由於三種方法選取的最後模型皆為相同，因此直接採用此模型為最後模型。

模型確立完畢後，要回歸原本的假設前題，因此模型必須要通過殘差的三個檢定(常態性、變異數齊一性、獨立性)，最後才可能為最佳模型。

在不考慮其他因素下，最後的結果為當平均票價下降、觀影人次增加與上座率的上升，都會使總票房越高。

關鍵字：西遊記之大鬧天宮、電影票房、迴歸、殘差分析、選取法

Abstract

In our life, go to see a movie is a common leisure activity. We sometimes read newspaper could see the report about the movie's box office, so we want to know what kind of factors impacts on the box office. On this report, we took "The Monkey King" for example, to discuss the relation of explanatory variable (average fare, screen, visitors, attendance rate) and dependent variable.

On this report, we wanted to find the best explanatory capability of the explanatory variable to influence the box office. First, we did the scatter plot and correlation coefficient to recognize whether the explanatory variable and dependent variable are linearity or not. After we confirmed their linearity, then we used three methods(forward method, backward method and stepwise method) to sift out the explanatory variable. Due to the same of the final model, so we adopted that to be the final model.

After confirmed the model, we back to our hypothesis, so the model had to pass through the residual analysis (normally, homoscedasticity, independently), then we could probably find the best model.

We did not consider other factors, the conclusion is that when the average fare decrease or the visitors and attendance rate increase may tend to the box office getting higher.

Keyword：box office; The Monkey King; regression; residual analysis; selection methods

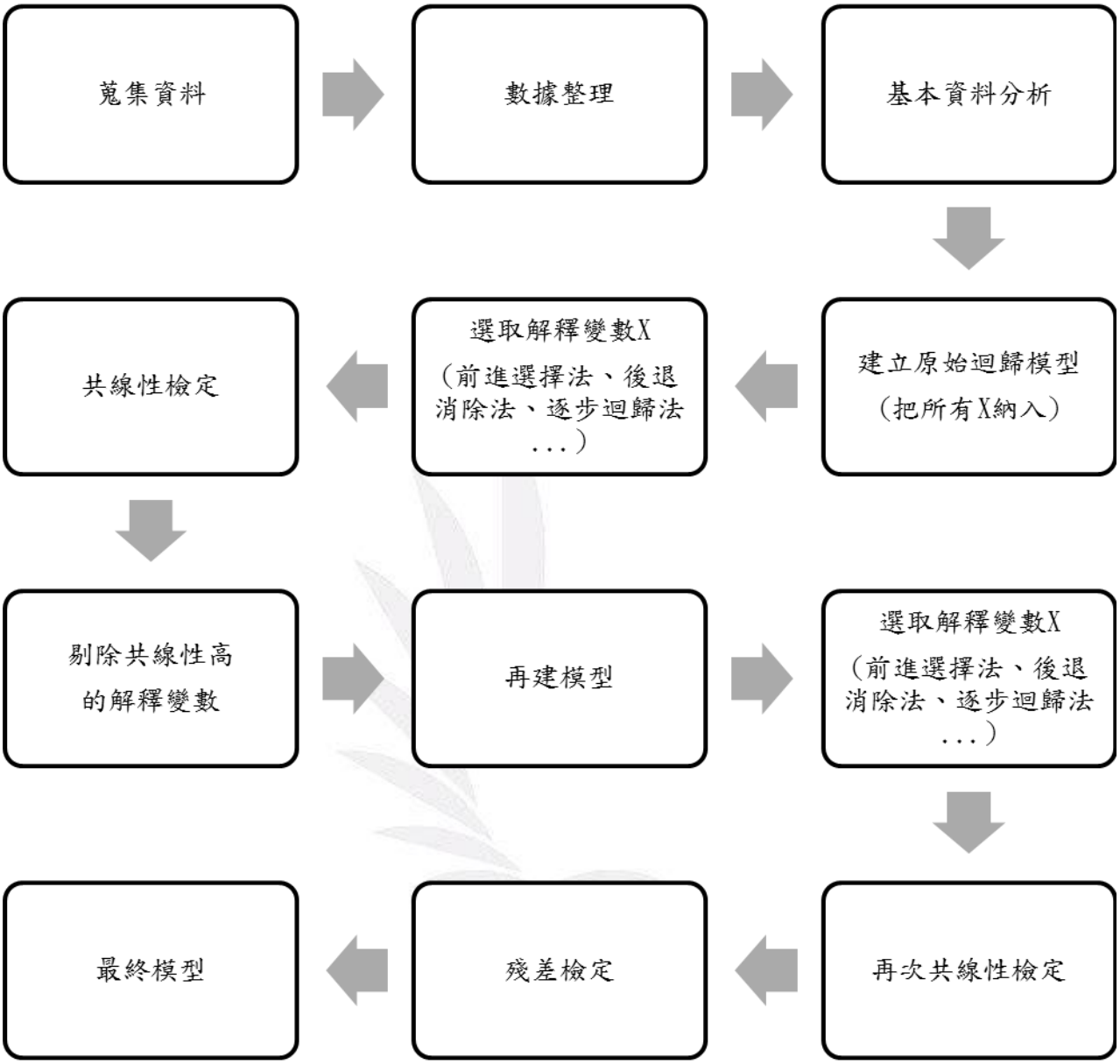
目　錄

第壹章、緒論

1.1 研究背景、動機與目的

電影最早開始於 1894 年，由盧米埃兄弟製成第一部完整的放映機和第一部電影『工人離開盧米埃工廠』，早期的電影都為無聲黑白電影，直到 1927 年由華納兄弟公司拍攝上映的音樂故事片『爵士鼓手』，開啟了有聲電影的時代， 1930 年彩色電影的引入更是電影發展的重要躍進。如今幾乎所有電影都為有聲彩色電影，且廣泛整個世界，成為我們休閒時的好選擇，電影大致上可分為動作、科幻、喜劇、愛情、紀錄、驚悚和武俠片...等。早期因受到技術限制，所以一部片都只有短短幾分鐘，講的故事也很簡單，直到後來電影敘事能力的發展與科技的進步，電影的長度才逐漸增加約 90~120 分鍾，甚至更長。電影從最早期利用視覺暫留，膠捲，錄像帶到現在的數位播放器，把聲音和影像捕捉起來後，再加上編輯工作便完成一部電影，是一種融合表演、視覺及聽覺的藝術。　　因為看電影為組員們的共同興趣，所以我們選擇此主題作研究。同時在觀看之餘，也可應用課程所學之統計專長，對電影的相關數據做分析。 經過討論後我們決定以電影『西遊記之大鬧天宮』為例，探討此部電影在中國上映的一個多月中，如何得到消費者的支持而走向戲院觀看。我們選擇 4 個變數，分別為平均票價、放映場次、觀影人數以及上座率，以此數據來做迴歸分析，探討這些變數對總票房的影響。

1.2 研究流程圖


蒐集資料
數據整理
基本資料分析
建立原始迴歸模型
（把所有X納入）
選取解釋變數X
（前進選擇法、後退消除法、逐步迴歸法...）
共線性檢定
剔除共線性高的解釋變數
再建模型
選取解釋變數X
（前進選擇法、後退消除法、逐步迴歸法...）
再次共線性檢定
殘差檢定
最終模型

1.3 資料出處與介紹變數

1.3.1 資料出處

數據收集源自於中國的網站—電影票房，我們選擇某一熱門電影「西遊記之大鬧天宮」為例子來做研究。

1.3.2 介紹變數

(1) 反應變數(Y)

➢總票房：

意指電影在影院上映期間賣出票的總額。一部電影的成功與否，主要就是看其票房的銷售情況。像我們常聽到有些電影或演員被稱做票房毒藥，就是指這部影片或這個演員出演的電影上座率很差，沒什麼票房收入。同理，如果說的是某某電影為票房冠軍，那就代表此電影是最賣座的。

(2)解釋變數(X)

➢X1：平均票價

每家電影院各有自己的票價，平均票價就是把所有的票價相加後再除以影院個數。

➢X2：放映場次

電影在上映期間，在影院播放的次數。

➢X3：觀影人次

電影在上映期間，進影院觀看的人次。

➢X4：上座率

計算方式為總人數除以放映場次的總座位數。是衡量一部影片口碑、票房、熱度等一系列參數的指標。但實際上，上座率的權威性並不高，很多時候可能會誤導觀眾對影片實際表現的判斷。上座率與每個電影院中的每個影廳座位數有關，而當影院排片人員把一部電影排在不同影廳時，上座率的計算會更加複雜。

第貳章、基本統計資料分析

2.1 基本敘述統計量

2.1.1 基本統計量

Y：總票房

X1：平均票價　X2：放映場次　X3：觀影人次　X4：上座率

基本統計量表							
變數	標籤	N	平均值	標準差	總和	最小值	最大值
Y	總票房	37	1412775	2569619	52272689	203	11075600
X1	平均票價	37	43.56	7.870004	1611.6	25	54.6
X2	放映場次	37	829.5	1146.458	30692	2	3843
X3	觀影人次	37	27020	47365.59	999743	7	202742
X4	上座率	37	20.21	12.38918	747.59	2.53	67.93

表 1. 基本統計量表

2.1.2 散佈圖

(1) X1：平均票價

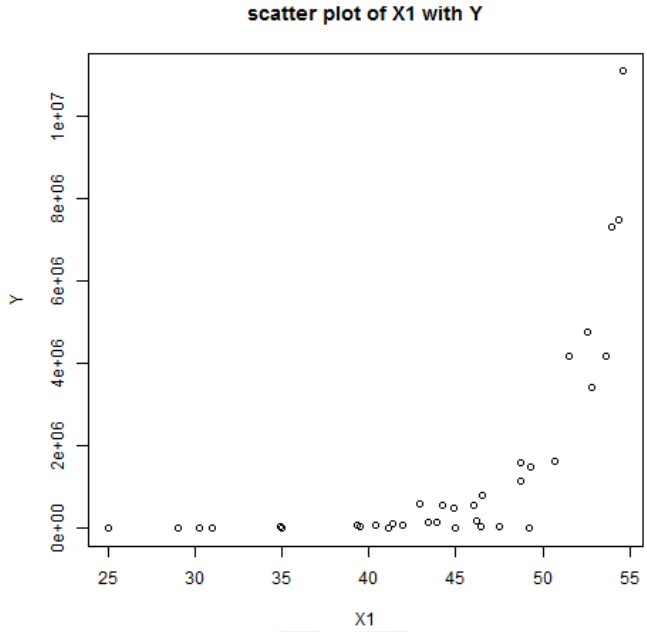


圖 1. 平均票價與 Y 的散佈圖

由此圖 1 的散佈圖可知，票價偏低時，與總票房的線性關係並不明顯；但票價到了 50~60 元間，總票房卻急劇上升，因此判斷票價與總票房大致呈現正相關。

(2) X2：放映場次

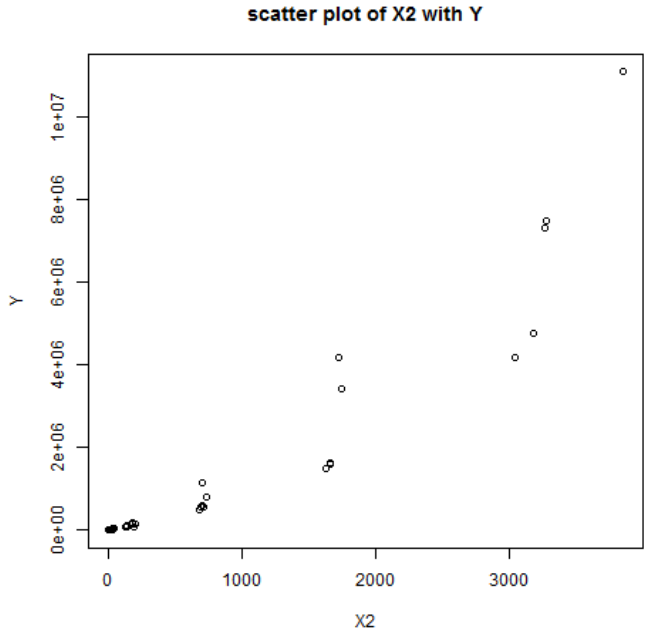


圖 2. 放映場次與 Y 的散佈圖

由此圖 2 的散佈圖可知，放映場次與總票房大致可推測呈現正相關。代表隨著

放映場次增加，總票房就愈高。

(3) X3：觀影人次

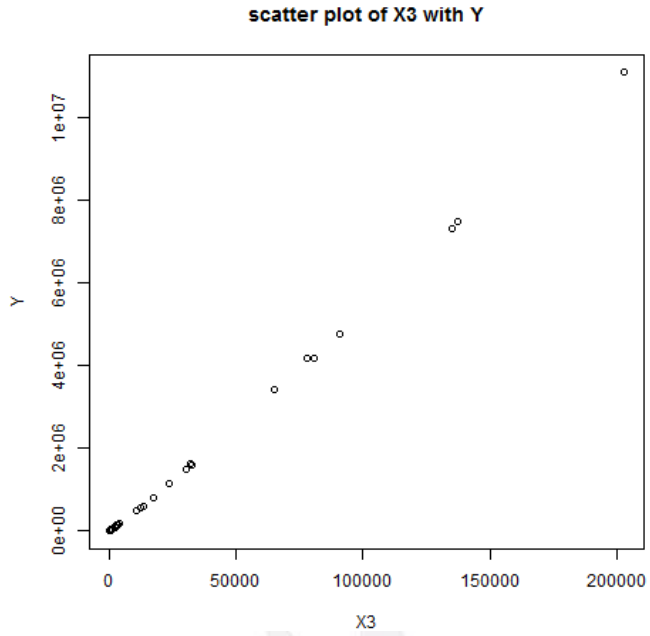


圖 3. 觀影人次與 Y 的散佈圖

由此圖 3 的散佈圖可知，觀影人次與總票房明顯地呈現正相關。可推測兩變數有高度的線性關係。

(4) X4：上座率

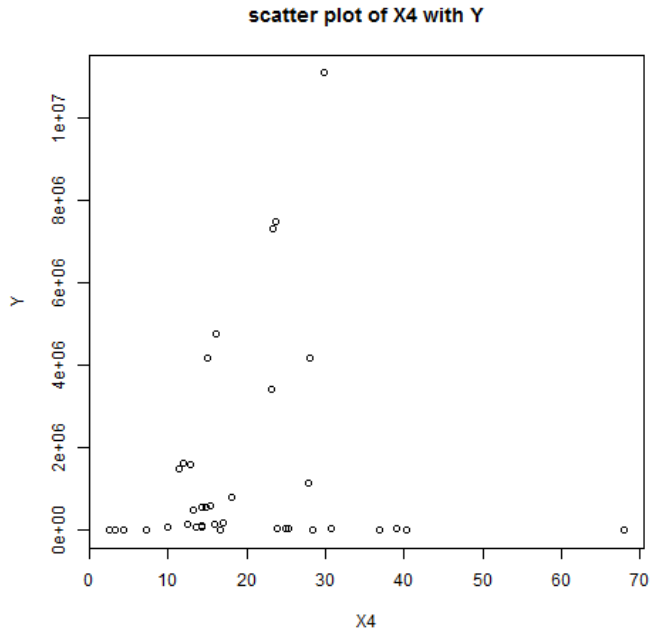


圖 4. 上座率與 Y 的散佈圖

由此圖 4 的散佈圖可知，上座率與總票房較無線性關係。我們猜想是因為每個戲院廳內座位不同，而導致基準的不同，因此難以判別線性關係。

2.2 相關性

相關係數表					
	Y	X1	X2	X3	X4
Y（總票房）	1.0000	0.6406250	0.93509691	0.99966781	0.09688067
X1（平均票價）	0.6406250	1.0000	0.71970985	0.65023299	0.21836676
X2（放映場次）	0.93509691	0.71970985	1.0000	0.94301386	-0.01903902
X3（觀影人次）	0.99966781	0.65023299	0.94301386	1.0000	0.09005876
X4（上座率）	0.09688067	0.21836676	-0.01903902	0.09005876	1.0000

表 2. 相關係數表

2.2.1 判斷方式：

相關係數(r)為兩變數間的相關程度，其值介於-1 至 1 之間，值為正代表為正相關；反之，為負代表為負相關，而相關係數的大小可將相關程度分為高度、中度、低度。

（1）0 ≦｜r｜＜0．3 →低度相關
（2）0．3 ≦｜r｜≦0．7 →中度相關
（3）0．7＜｜r｜≦1 →高度相關

2.2.2 結論：

由上表與相關系數的判定標準，可發現解釋變數與應變數之間的相關程度，除了 X4（上座率）為低度相關外，其他皆呈現中度或高度正相關。在此，可得知 X4（上座率）與總票房較無線性關係。

在解釋變數之間，發現 X4（上座率）與其他解釋變數的相關程度都為低度相關。其中，與 X2(放映場次)呈現負相關(-0.01903902)。而撇除掉 X4（上座率）這個解釋變數後，可知道剩下的變數彼此之間的相關性很大，也就代表著變數之間可能隱藏著共線性問題。尤其，發生在 X2（放映場次）對 X3（觀影人次）相關係數為最大(0.94301386)，因此，在建立模型後，會對模型進行共線性的檢驗，以確保篩選模型的正確性。

第參章、原始模型檢定

3.1 建立迴歸模型

首先，我們將所有解釋變數列入模型內。假設由平均票房(X1)、放映場次(X2)、觀影人次(X3)、上座率(X4)預測總票房(Y)，建立此模型。

Model_1： $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \hat{\beta}_4 X_{4i}$ **i=1,2,…,37**

3.2 參數估計

參數估計						
變數	標籤	DF	參數估計值	標準誤差	t 值	Pr>|t|
Intercept	Intercept		39008.0534	38792.2641	1.006	0.322
X1	平均票價	1	-1629.8289	1049.8135	-1.552	0.130
X2	放映場次	1	-140.7785	16.9378	-8.312	1.7e-09
X3	觀影人次	1	57.6166	0.3626	158.879	< 2e-16
X4	上座率	1	234.0942	485.8700	0.482	0.633

表 3. 參數估計

由表 3 的參數估計可知道：

$$\hat{\beta}_0 = 39008.0534、\hat{\beta}_1 = -1629.8289、$$
$$\hat{\beta}_2 = -140.7785、\hat{\beta}_3 = 57.6166、\hat{\beta}_4 = 234.0942$$

Model_1：

$$\hat{Y}_i = 39008.0534 - 1629.8289X_{1i} - 140.7785X_{2i} + 57.6166X_{3i} + 234.0942X_{4i}$$

i=1,2,…,37

3.3 模型適合度檢定

變異數分析表如下表：

變異數分析表				
來源	DF	Sum Sq	Mean Sq	Pr>F
Reg	4	2.376712e+14	5.94178e+13	< 2.2e-16
Error	32	3.1392e+10	9.8101e+08	

表 4. 變異數分析表

統計假設如下：

H_0：$\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

H_1：β_k不完全為 0，k = 1,2,3,4

由上表 4 的變異數分可得知，因為 P-value< 2.2e-16 < α = 0.05，所以拒絕H_0的假設，代表β_1、β_2、β_3、β_4不完全為 0，即我們可以推論解釋變數(平均票價、放映場次、觀影人次、上座率)與反應變數(總票房)有線性迴歸相關。

3.4 模型解釋能力

模型解釋分析表			
R-Square	0.9999	Adj R-Square	0.9999

表 5. 模型解釋分析表

在此模型下，Multiple R-Square = 0.9999，表示 Y 的總差異可以由 X1、X2、X3、X4 解釋 99.99%；而經過校正過後的 R-Square = 0.9999，相較之下，R-Square 比調整過後一樣，代表 X1、X3、X4 對 Y 有很大的解釋能力。

3.5 參數檢定

我們得到β值後，開始檢定各個解釋變數平均票價(X1)、觀影人次(X3)、上座率(X4)與反應變數(Y)總票房之間是否有線性相關。

(1)我們想要判斷平均票價(X1)與總票房(Y)是否存在線性關係，首先我們假定其他變數固定的情況下，設立假設:

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

因為 P-value= 0.130>α=0.05，所以拒絕H_0的假設，表示$\beta_1 = 0$。即平均票價(X1)與總票房(Y)不存在線性關係。

(2) 我們想要判斷放映場次(X2)與總票房(Y)是否存在線性關係，首先我們假定其他變數固定的情況下，設立假設:

$H_0 : \beta_2 = 0$

$H_1 : \beta_2 \neq 0$

因為 P-value= 1.7e-09<α=0.05，所以拒絕H_0的假設，表示$\beta_2 \neq 0$。即放映場次(X2)與總票房(Y)存在線性關係。

(3)我們想要判斷觀影人次(X3)與總票房(Y)是否存在線性關係，首先我們假定其他變數固定的情況下，設立假設:

$H_0 : \beta_3 = 0$

$H_1 : \beta_3 \neq 0$

因為 P-value< 2e-16<α=0.05，所以拒絕H_0的假設，表示$\beta_3 \neq 0$。即觀影人次(X3)與總票房(Y)存在線性關係。

(4)我們想要判斷上座率(X4)與總票房(Y)是否存在線性關係，首先我們假定其他變數固定的情況下，設立假設:

$H_0 : \beta_4 = 0$

$H_1 : \beta_4 \neq 0$

因為 P-value= 0.633>α=0.05，所以拒絕H_0的假設，表示$\beta_4 = 0$。即上座率(X4)與總票房(Y)不存在線性關係。

第肆章、模型的選擇方式

4.1 前進選擇法(Forward Selection)

前進選擇法是以完全沒有解釋變數進入模型的狀況下，逐一挑選變數進入模型。第一個進入迴歸模型的解釋變數，是參數估計值最為顯著，也就是 P-value 最小的變數進入。接著，在進入模型裡的每一個解釋變數都必須達到設定標準(在這裡的設定為α=0.05)，以顯著性高的變數為優先，以此類推…直到沒有變數符合標準為止。此外，這個選擇法有一個很重要的規則，是一旦進入模型內的變數就不會被剔除。此時還在模型內的變數就為最終模型。

在 R 軟體中，我們利用 step 函數來幫助我們篩選變數，其中，step 函數是以 AIC 作為篩選標準，AIC 越小越好。

令 **X1**：平均票價　**X2**：放映場次　**X3**：觀影人次　**X4**：上座率

Y ~ X1 + X2 + X3

STEP	變數	DF	AIC	F 值	PR(>F)
1	+ X3	1	824.45	52654.9899	< 2.2e-16
2	+ X2	1	769.39	124.9340	6.259e-13
3	+ X1	1	768.95	2.2530	0.1429

表 6. Forward Selection Procedure

Step1：由系統選擇出顯著性最大的解釋變數 X3(觀影人次)，P-value < 2.2e-16。
Step2：接著從剩下的 3 個解釋變數(X1、X2、X4)中，依序選取顯著性到達設定標準且最大的變數，依序為 X2(放映場次)、X1(平均票價)，剩下的 X4(上座率)並不顯著，所以不放入模型中。利用前進選取法選出的最後模型為

$$\hat{Y}_i = 36180.6665 - 1428.1146X_{1i} - 144.486X_{2i} + 57.6849X_{3i}$$

i=1,2,…,37

4.2 後退選取法(Backward Method)

後退選取法是將所有解釋變數加入模型內，將不顯著解釋變數從模型內剔除。踢出的規則是所有解釋變數中，最不顯著(P-value 最大)的解釋變數優先剔除，直到在模型內的所有解釋變數皆具有顯著性時，停止。此外，這個選擇方法有一個重要原則，一旦解釋變數被剔除，就不再進入模型內。此時還在模型內的變數就為最終模型。

在 R 軟體中，我們利用 step 函數來幫助我們篩選變數，其中，step 函數是以 AIC 作為篩選標準，AIC 越小越好。

令 **X1**：平均票價　**X2**：放映場次　**X3**：觀影人次　**X4**：上座率

Y ~ X1 + X2 + X3

STEP	剔除變數	DF	AIC	F 值	PR(>F)
1	- X4	1	768.95	0.2321	0.6332

表 7. Backward Method

Step1：由系統選擇出沒有顯著性的解釋變數 X4(上座率)，P-value = 0.6332 大於 0.05，將其剔除。由於剩下的 3 個變數(X1、X2、X3)皆有顯著性，故不剔除。利用後退選取法選出的最終模型為

$$\hat{Y}_i = 36180.6665 - 1428.1146X_{1i} - 144.486X_{2i} + 57.6849X_{3i}$$

i=1,2,…,37

4.3 逐步迴歸法(Stepwise Method)

逐步迴歸法是前面的後退選取法和前進選擇法的綜合運用。第一步是運用前進選擇法，將模型外有顯著性的解釋變數放入模型內；第二步，再利用後退選取法檢定模型內的所有變數，將沒有顯著性的變數踢出模型外，一直重複這動作，直到模型外的變數都不具有顯著性或者此變數已經進入過模型中時，停止。此時還在模型內的變數就為最終模型。

令 X1：平均票價　X2：放映場次　X3：觀影人次　X4：上座率

Y ~ X1 + X2 + X3

STEP	變數	DF	AIC	F 值	PR(>F)
1	+ X3	1	824.45	52654.9899	< 2.2e-16
2	+ X2	1	769.39	124.9340	6.259e-13
3	+ X1	1	768.95	2.2530	0.1429

表 8. Stepwise Method

系統利用前進選擇法的方式，首先選擇顯著性最大的變數 X3(觀影人次)。接著再依序放入有顯著性的解釋變數 X2(放映場次)、X1(平均票價)。在過程中，放入解釋變數時，皆沒有需要剔除的對象，因此系統停止動作。最後，逐步迴歸法所選出的最終模型為

$$\hat{Y}_i = 36180.6665 - 1428.1146X_{1i} - 144.486X_{2i} + 57.6849X_{3i}$$

i=1,2,…,37

4.4 小結

綜合以上三種方法所選出來的解釋變數皆相同。因此我們選定 X1：平均票價、X2：放映場次、X3：觀影人次，共三個解釋變數，選為最終模型，其迴歸方程式為

$$\hat{Y}_i = 36180.6665 - 1428.1146X_{1i} - 144.486X_{2i} + 57.6849X_{3i}$$

i=1,2,…,37

4.5 共線性檢定

由於在相關係數表裡知道 X2(放映場次)與 X3(觀影人次)的相關係數高達 0.94301386，因此對於最終模型進行共線性檢定，如下：

共線性檢定			
變數名稱	**X1(平均票價)**	**X2(放映場次)**	**X3(觀影人次)**
VIF 值	2.106590	10.981432	9.170603

表 9.共線性檢定

(1) 判定方式：

VIF 值>10，代表此模型有嚴重的共線性問題。

(2) 結論：

由上表可得知 X2(放映場次)的 VIF 值為 10.981432 大於 10，代表 X2(放映場次)這個變數有嚴重的共線性問題。雖然 X3(觀影人次) VIF 值沒有到達 10，但是也很接近。因此，我們決定將 VIF 值最大的 X2(放映場次)從模型中剔除，只留下 X3(觀影人次)，以解決模型之共線性問題。

4.6 結論

為解決共線性之問題，我們將 X2(放映場次)這個變數剔除，重新進行選取模型。

第伍章、再建模型

5.1 再建迴歸模型檢定

將 X2(放映場次)剔除後，建立新的模型。重新假設由平均票房(X1)、觀影人次(X3)、上座率(X4)來預測總票房(Y)。

$$\text{Model\_2：}\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_3 X_{3i} + \hat{\beta}_4 X_{4i} \quad \text{i=1,2,...,37}$$

5.1.1 參數估計

參數估計值						
變數	標籤	DF	參數估計值	標準誤差	t 值	Pr>|t|
Intercept	Intercept		1.595e+05	6.298e+04	2.532	0.016280
X1	平均票價	1	-6.215e+03	1.563e+03	-3.975	0.000361
X3	觀影人次	1	5.486e+01	2.545e-01	215.533	< 2e-16
X4	上座率	1	2.069e+03	7.575e+02	2.731	0.010057

表 10. 參數估計值

由上表可知道：$\hat{\beta}_0 = 159500$、$\hat{\beta}_1 = -6215$、$\hat{\beta}_3 = 54.86$、$\hat{\beta}_4 = 2069$

$$\text{Model\_2：}\hat{Y}_i = 159500 - 6215X_{1i} + 54.86X_{3i} + 2069X_{4i} \quad \text{i=1,2,...,37}$$

5.1.2 模型適合度檢定

利用挑選之後的三個變數利用 R 軟體進行適合度檢定，變異數分析表如下表：

變異數分析表				
來源	DF	Sum Sq	Mean Sq	Pr>F
Reg	3	2.376074e+14	7.920247e+13	< 2.2e-16
Error	33	9.9161e+10	3.0049e+09	

表 11. 變異數分析表

統計假設如下：

H_0：$\beta_1 = \beta_3 = \beta_4 = 0$

H_1：β_k不完全為 0，k = 1,3,4

由上表可得知，因為 P-value< 2.2e-16 < α = 0.05，所以拒絕H_0的假設，代表 β_1、β_3、β_4不完全為 0，即我們可以推論解釋變數(平均票價、觀影人次、上座率)與反應變數(總票房)有線性迴歸相關。

5.1.3 模型解釋能力

模型解釋分析表			
R-Square	0.9996	Adj R-Square	0.9995

表 12. 模型解釋分析表

在此模型下，Multiple R-Square＝0.9996，表示 Y 的總差異可以由 X1、X3、X4 解釋 99.96%；而經過校正過後的 R-Square＝0.9995，相較之下，R-Square 比調整過後少 0.01%，代表 X1、X3、X4 對 Y 有很大的解釋能力。

5.1.4 參數檢定

我們得到β值後，開始檢定各個解釋變數平均票價(X1)、觀影人次(X3)、上座率(X4)與反應變數(Y)總票房之間是否有線性相關。

(1)我們想要判斷平均票價(X1)與總票房(Y)是否存在線性關係，首先我們假定其他變數固定的情況下，設立假設:

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

因為 P-value=0.000361<α=0.05，所以拒絕H_0的假設，表示$\beta_1 \neq 0$。即平均票價(X1)與總票房(Y)存在線性關係。

(2)我們想要判斷觀影人次(X3)與總票房(Y)是否存在線性關係，首先我們假定其他變數固定的情況下，設立假設:

$H_0 : \beta_3 = 0$

$H_1 : \beta_3 \neq 0$

因為 P-value< 2e-16<α=0.05，所以拒絕H_0的假設，表示$\beta_3 \neq 0$。即觀影人次(X3)與總票房(Y)存在線性關係。

(3)我們想要判斷上座率(X4)與總票房(Y)是否存在線性關係，首先我們假定其他變數固定的情況下，設立假設:

$H_0 : \beta_4 = 0$

$H_1 : \beta_4 \neq 0$

因為 P-value=0.010057<α=0.05，所以拒絕H_0的假設，表示$\beta_4 \neq 0$。即上座率(X4)與總票房(Y)存在線性關係。

本節結論：

將每個解釋變數的參數值做檢定，結果皆為拒絕H_0的假設，代表著每個解釋變數都與反應變數皆存在線性關係。

5.2 再建模型的選擇

5.2.1 前進選擇法

依照 4.1 小節說明，我們再利用同樣規則進行選取變數。如下：

令 X1：平均票價　X3：觀影人次　X4：上座率

Y ~ X3 + X1 + X4

STEP	變數	DF	AIC	F 值	PR(>F)
1	+ X3	1	824.45	52654.9899	< 2.2e-16
2	+ X1	1	816.78	10.1596	0.003075
3	+ X4	1	811.24	7.4582	0.01006

表 13. Forward Selection Procedure

Step1：由系統選擇出顯著性最大的解釋變數 X3(觀影人次)，P-value < 2.2e-16
Step2：接著從剩下的 2 個解釋變數(X1、X4)中，依序選取具有顯著性的變數，依序為 X1(平均票價) X4(上座率)，兩者皆到達顯著水準，所以皆放入模型中。
利用前進選取法選出的最後模型為

$$\hat{Y}_i = 159500 - 6215X_{1i} + 54.86X_{3i} + 2069X_{4i} \quad i=1,2,\ldots,37$$

5.2.2 後退選取法

依照 4.2 小節說明，我們再利用同樣規則進行選取變數。如下：

令 X1：平均票價　X3：觀影人次　X4：上座率

Y ~ X1 + X3 + X4

	DF	AIC	F 值	PR(>F)
<None>		811.24		
-X4	1	816.78	7.4582	0.0100572
-X1	1	823.71	15.8033	0.0003608
-X3	1	1077.50	46454.51710	< 2.2e-16

表 14. Backward Method

3 個解釋變數皆到達顯著水準，P-value < 0.05，因此不需剔除任何變數。利用後退選取法選出的最後模型為

$$\hat{Y}_i = 159500 - 6215X_{1i} + 54.86X_{3i} + 2069X_{4i} \quad i=1,2,\ldots,37$$

5.2.3 逐步迴歸法

依照 4.3 小節說明，我們再利用同樣規則進行選取變數。如下：

令 X1：平均票價　X3：觀影人次　X4：上座率

STEP	變數	DF	AIC	F 值	PR(>F)
1	+ X3	1	824.45	52654.9899	< 2.2e-16
2	+ X1	1	816.78	10.1596	0.003075
3	+ X4	1	811.24	7.4582	0.01006

表 15. Stepwise Method

系統利用前進選擇法的方式，首先選擇顯著性最大的變數 X3(觀影人次)。接著再依序放入有顯著性的解釋變數 X1(平均票價)、X4(上座率)。在過程中，放入解釋變數時，皆沒有需要剔除的對象，因此系統停止動作。逐步迴歸法所選出的最終模型為

$$\hat{Y}_i = 159500 - 6215X_{1i} + 54.86X_{3i} + 2069X_{4i} \quad i=1,2,\ldots,37$$

5.2.4 小節

綜合以上三種方法所選出來的解釋變數皆相同，都為再建的原始模型。因此我們選定 X1：平均票價、X3：觀影人次、X4：上座率，共三個解釋變數，選為最終模型，其迴歸方程式為

$$\hat{Y}_i = 159500 - 6215X_{1i} + 54.86X_{3i} + 2069X_{4i} \quad i=1,2,\ldots,37$$

5.3 再次共線性檢定

共線性檢定			
變數名稱	X1(平均票價)	X3(觀影人次)	X4(上座率)
VIF 值	1.813402	1.741052	1.055249

表 16. 共線性檢定

由上表得知，在剔除 X2(放映場次)後，模型中的解釋變數的 VIF 值皆沒有超過 10 且所有變數的 VIF 值都下降，解決了共線性之問題。

第陸章、殘差分析

在迴歸分析中，在進行所有分析的假設前提是殘差值是一個常態分配，期望值為 0，變異數為σ^2，並且殘差之間相互獨立。因此，選定完模型後，必須進行常態性檢定、變異數齊一性檢定與獨立性檢定，如三個檢定通過，此模型才成立。

Model：$\hat{Y}_i = 159500 - 6215X_{1i} + 54.86X_{3i} + 2069X_{4i}$　　i=1,2,...,37

6.1　常態檢定

(1)在 R 軟體中，我們運用 shapiro.test 函數來進行殘差的常態性檢定。如下：

H_0：殘差項為常態分配

H_1：殘差項不為常態分配

SHAPIRO-WILK NORMALITY TEST			
W	0.9727	p-value	0.4863

表 11. 殘差常態性檢定

由上表可得知，p-value=0.4863>α=0.05，因此不拒絕H_0，即代表殘差符合常態分配的條件。

(2)Q-Q plot

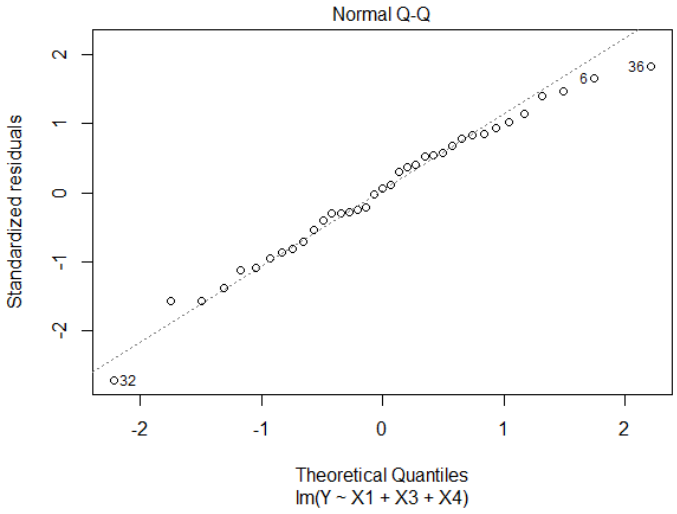


圖 5. Q-Q plot

上圖中大部分的殘差值皆接近虛線上，只有頭與尾的部分偏離得較嚴重，但也

可在此證明殘差是符合常態分配的。

6.2 變異數均齊性檢定

在 R 軟體中，我們運用 ncvTest 函數來進行殘差的變異數均齊性檢定。如下：

H_0：殘差項的變異數具均齊性
H_1：殘差項的變異數不具均齊性

NON-CONSTANT VARIANCE SCORE TEST					
Chisquare	2.759848	DF	1	p-value	0.09665744

表 12. 殘差變異數均齊性檢定

由上表可得知，p-value=0.09665744>α=0.05，因此不拒絕H_0，即代表殘差具有變異數均齊性的條件。

6.3 獨立性檢定

在 R 軟體中，我們運用 durbinWatsonTest 函數來進行殘差的獨立性檢定。如下：

H_0：殘差項之間互相獨立
H_1：殘差項之間互相不獨立

殘差獨立性檢定			
lag	Autocorrelation	D-W Statistic	p-value
1	0.1964232	1.599931	0.088

表 13. 殘差之間獨立性檢定

由上表可得知，p-value=0.088>α=0.05，因此不拒絕H_0，即代表殘差之間互相具有獨立性的條件。

第柒章、結語

這次的迴歸分析報告主要研究『西遊記之大鬧天宮』這部電影的「平均票價」、「放映場次」、「觀影人次」、「上座率」的四個變數與「總票房」之間的關係。

首先，在敘述統計量的章節裡，發現解釋變數之間的相關性滿大的，尤其「放映場次」與「觀影人次」這兩個變數高達 0.94，因此需注意兩變數會有共線性之問題。接著，我們利用前進選擇法、後退選取法與逐步迴歸法進行解釋變數的篩選，結果選出的變數為 X1「平均票價」、X2「放映場次」、X3「觀影人次」，即初步選模。以此模型進行接下來的迴歸分析。

在初步選模完畢後，由於最終模型包含 X2「放映場次」和 X3「觀影人次」，因此進行共線性的檢定，發現「放映場次」的 VIF 值皆超過 10，而「觀影人次」則是接近 10，所以最後為解決共線性之問題，將 VIF 值最大的「放映場次」剔除，再進行一次選模。

最後，在進行所有迴歸分析的假設前題的殘差檢定，分別為殘差常態性檢定、殘差變異數均齊性檢定、殘差間的獨立性檢定，符合三項殘差檢定後，為最佳線性迴歸模型。

在最佳模型 Model：$\hat{Y}_i = 159500 + -6215X_{1i} + 54.86X_{3i} + 2069X_{4i}$中，個參數代表的意義為：

β_0：當不考慮X_1、X_3、X_4的情況下，即X_1、X_3、X_4=0 時，Y 平均的值為 159500。

β_1：當其他變數X_3、X_4不變的情況下，每增加一單位的X_1 ，Y 的平均值會減少 6215。即不考慮觀影人次與上座率兩個變數時，平均票價每增加一單位，總票房之平均值就會下降 6215 人民幣。

β_3：當其他變數X_1、X_4不變的情況下，每增加一單位的X_3 ，Y 的平均值會增加 54.86。即不考慮平均票價與上座率兩個變數時，觀影人次每增加一單位，總票房之平均值就會增加 54.86 人民幣。

β_4：當其他變數X_1、X_3不變的情況下，每增加一單位的X_4 ，Y 的平均值會增加 2069。即不考慮平均票價與觀影人次兩個變數時，上座率每增加一單位，總票房之平均值就會增加 2069 人民幣。

在模型中校正後的R^2高達 99.95%接近 100%，對於模型的總變異的解釋能力是相當高。但此份資料唯一的不足是樣本數雖然有達到 30，但還是太少，所以我們推測這也可能是造成解釋能力偏高的原因之一。但撇除掉樣本數的因素，我們可以說，當平均票價下降、觀影人次增加與上座率的上升，都會使總票房越高。

第捌章、附錄

8.1 原始資料

Y	X1	X2	X3	X4
203	29	3	7	2.53
362	30.2	3	12	4.33
500	25	3	20	7.22
279	31	3	9	3.25
17800	35	15	510	40.38
12100	49.2	10	245	28.39
15400	31	18	496	36.9
33500	34.9	47	961	23.92
21100	41.1	36	512	16.64
33400	34.9	43	958	24.91
39000	47.5	37	821	25.28
45200	46.4	29	975	38.98
49800	39.5	48	1261	30.76
92700	40.4	137	2293	14.3
89800	41.9	131	2140	13.62
97000	41.4	138	2345	14.24
84100	39.3	189	2139	9.88
146100	43.9	178	3326	15.9
171400	46.2	183	3713	16.94
130100	43.4	206	2995	12.43
489500	44.9	685	10897	13.11
549900	44.2	713	12435	14.22
576100	42.9	707	13423	15.4
573500	46	695	12474	14.8
1148200	48.7	707	23558	27.75
797500	46.5	730	17139	18.13
1486000	49.3	1626	30117	11.43
1579300	48.7	1658	32432	12.9
1615100	50.7	1661	31854	11.99

3423400	52.8	1740	64892	23.21
4167200	53.6	1718	77691	28.08
4771600	52.6	3178	90750	16.06
7481000	54.4	3276	137479	23.6
7294100	54	3256	135195	23.39
10440500	55.1	3452	189487	30.97
11075600	54.6	3843	202742	29.79
7245	45	2	161	67.93

8.2 R 的程式碼

```
data0=read.csv("C:/Users/USER/Desktop/報告/data6.csv",header=T)
data0
names(data0)
str(data0)
data0$X4<- as.numeric(data0$X4)
str(data0)
attach(data0)
summary(data0)

path<-"C://Users//USER//Desktop//報告//"
png(paste(path,"data0_x1 散佈圖.png",sep=""))
plot(X1,Y,xlab='X1',ylab='Y',main='scatter plot of X1 with Y')
path<-"C://Users//USER//Desktop//報告//"
png(paste(path,"data0_x2 散佈圖.png",sep=""))
plot(X2,Y,xlab='X2',ylab='Y',main='scatter plot of X2 with Y')
ath<-"C://Users//USER//Desktop//報告//"
png(paste(path,"data0_x3 散佈圖.png",sep=""))
plot(X3,Y,xlab='X3',ylab='Y',main='scatter plot of X3 with Y')
path<-"C://Users//USER//Desktop//報告//"
png(paste(path,"data0_x4 散佈圖.png",sep=""))
plot(X4,Y,xlab='X4',ylab='Y',main='scatter plot of X4 with Y')

sink(paste(path,"data0_相關係數表.txt",sep=""))
cor(data0)
sink()

model_1=lm(Y ~ X1 + X2 + X3 + X4)
library(car)
vif(model_1)

data=data0[,-3]
data

names(data)
str(data)
```

```
data$X4<- as.numeric(data$X4)
str(data)
attach(data)

model_2=lm(Y ~ X1 + X3 + X4,data=data,x=T)
vif(model_2)
sink(paste(path,"data_參數表.txt",sep=""))
summary(model_2)
sink()

sink(paste(path,"data_anova.txt",sep=""))
anova(model_2)
sink()

null=lm(Y ~ 1,data=data)
null
full=lm(Y ~ X1 + X3 + X4,data=data)
full
step(null, scope=list(lower=null, upper=full),
direction="forward",test="F")
step(full, data=data, direction="backward",test="F")
step(null, scope = list(upper=full), data=data,
direction="both",test="F")

names(model_2)
resid2=model_2$residuals
shapiro.test(resid2)
ncvTest(model_2)
durbinWatsonTest(model_2)
hist(resid2)
plot(model_2)
```

8.3 工作分配

1. 文書報告：全組員
2. 簡報：全組員
3. 資料分析：謝宜君

8.4 參考文獻

1. 上座率的解釋
   http://daily.zhihu.com/story/3771681
2. 電影票房數據庫
3. 逢甲大學的優質報告：NBA 得分
4. 逢甲大學的優質報告：台灣人口數
5. R 軟體應用統計方法/陳景祥著/東華書局
6. INTRODUCTION TO REGRESSION MODELING
7. 變數選取的語法
   http://www.stat.columbia.edu/~martin/W2024/R10.pdf