行政院國家科學委員會補助

大專學生參與專題研究計畫研究成果報告

＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊
＊ 計畫 ＊
＊ ：一個知識管理現象的數學證明 ＊
＊ 名稱 ＊
＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊

執行計畫學生：林柏佐

學生計畫編號：97-2815-C-003 -013 -M

研 究 期 間 ：97 年 7 月 1 日至 98 年 2 月底止，計 8 個月

指 導 教 授 ：陳界山

執 行 單 位：國立臺灣師範大學數學系(所)

中華民國 98 年 3 月 30 日

1. 研究動機與研究問題

現在各個大專院校都在積極爭取「邁向頂尖大學」5 年 500 億計畫，各個學校在第一階段分配金額如下，台大 30 億、成大 17 億、清大 10 億、交大 8 億、陽明 5 億、中央及中山各 6 億、中興 4 億、政大、台科大、長庚、元智各 3 億。從這樣的分配中，我們可以看的出來教育部將重點學校分配大量的金額，而其他學校則是些微補助，那是否這樣的補助方式對整體是最好的呢？有沒有數學根據能夠證明這件事？

現在我們考慮以下情形，如果有 n 所學校要分配 5 年 500 億，令 x_i 代表每所學校成功的 probability(類似 random variable 的概念). $\Delta = \sum_{i=1}^{n}(x_i - \mu)^2$, 這裡 $\mu = \frac{\sum_{i=1}^{n} x_i}{n}, 0 < x_i < 1, \Delta$ 就很類似 variation 的概念，也就是說其離散的程度。

Define $H = 1 - \prod_{i=1}^{n}(1 - x_i)$ 代表的意思就是，會成功的 probability(及 $1-$ 皆失敗的 probability).

本計劃就是要來探討 H 與 Δ 的關係，進而判斷教育部的暨定政策是不是符合數學證明的結果。

2. 研究方法及步驟

本來想要證明 $\frac{\partial H}{\partial \Delta} > 0$ (即 increasing), 也就是說當 Δ 越大(離散程度越大)，H 也越大(即成功的 probability 越大)，採用 analysis 的手法，利用 concave analysis, 但在過程中碰到了很大的阻礙，因為其 Δ 並非一般分析手法可以解決的，因此在嘗試利用數學證明的方法行不通的情形下，改而去分析教育部的補助經費與其學校辦學績效，是否有很好的關聯，由於有些學校評鑑很難判別，本想要去教育部找一些相關資料(或是各大學自行公佈的資料)，但由於搜尋有困難，最後改採取從英國泰晤士報公佈的全世界大學排名，來判別其是否這樣的補助是不錯可行的。因此我們預期補助越多，名次將會越低，有明顯的改善。

底下我利用統計軟體 R, 利用 regression 的方法，來作 analysis.

X 代表各個學校的排名(rank)，下標代表該年度

Y 代表各個學校補助的金額，下標代表第一期或第二期，單位為億元新台幣

底下輸入的順序依序為

台灣大學、成功大學、清華大學、交通大學、中央大學、中山大學、陽明大學、中興大學、台灣科技大學、政治大學(只考慮公立學校，原因為一是兩次都有補助，二是大部份可以查到世界排名)

底下將不再重複贅述。

如有需要說明，將會在後面加#加註說明

```
> x2005<-c(114,335,206,427,488,481,257,501,266,396)
> y1<-c(30,17,10,8,6,6,5,4,3,3)
> fm2005=lm(y1~x2005) #y1 對 x2005 做 regression
> summary(fm2005)

Call:
lm(formula = y1 ~ x2005)

Residuals:
```

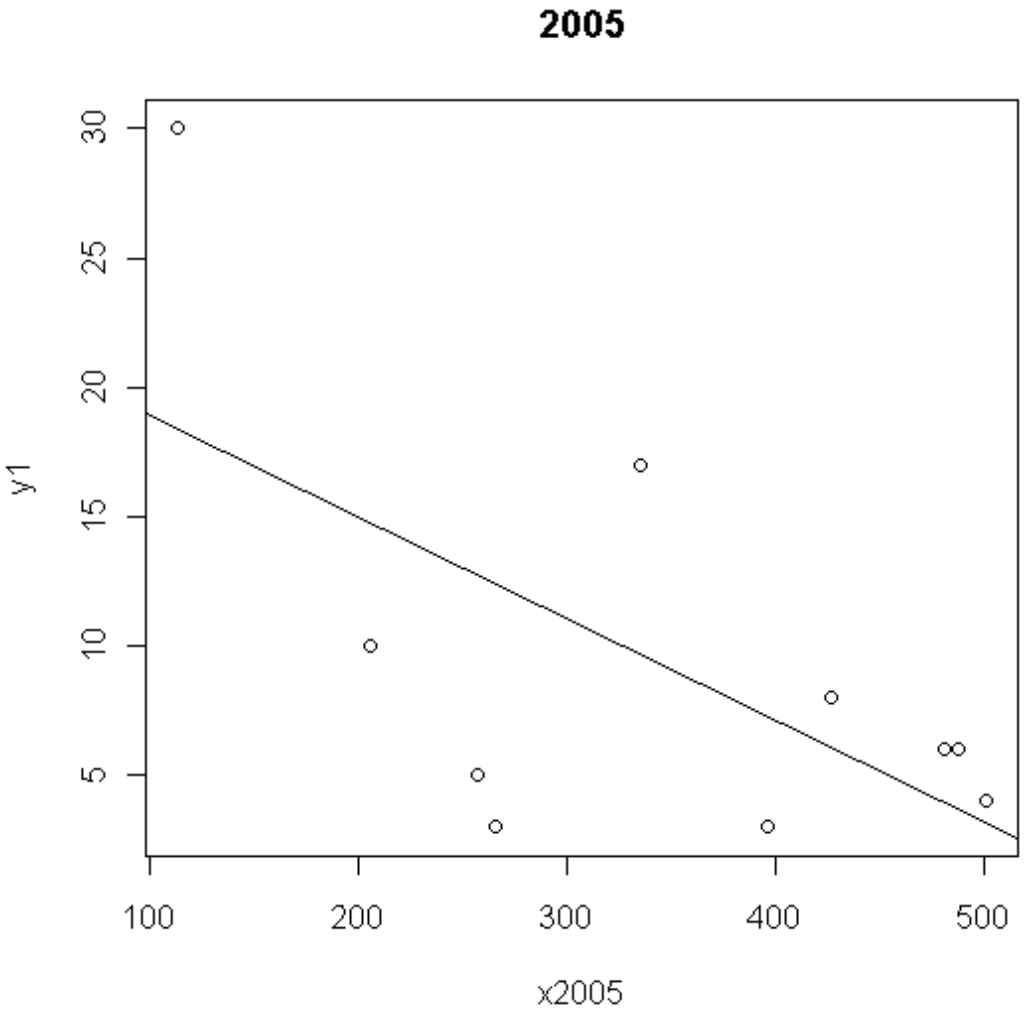
```
     Min       1Q Median       3Q      Max
-9.383 -4.624    1.388    2.261 11.651

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept) 22.82267       6.47852      3.523    0.00781 **
x2005             -0.03925       0.01754    -2.237    0.05564 .
---
Signif. codes:    0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.002 on 8 degrees of freedom
Multiple R-squared: 0.3849,         Adjusted R-squared: 0.308
F-statistic: 5.006 on 1 and 8 DF,    p-value: 0.05564
```

因為 slope p – value=0.05564 並非有很顯著的 linear relation

```
> plot(x2005,y1,main="2005")
> abline(coef(fm2005))
```




```
> x2006<-c(108,386,343,441,497,505,392,536,454,511)
> fm2006=lm(y1~x2006) #y1 對 x2006 做 regression
> summary(fm2006)

Call:
lm(formula = y1 ~ x2006)

Residuals:
     Min       1Q   Median       3Q      Max
```

-5.7400 -2.9160 0.9469 2.0122 5.8948

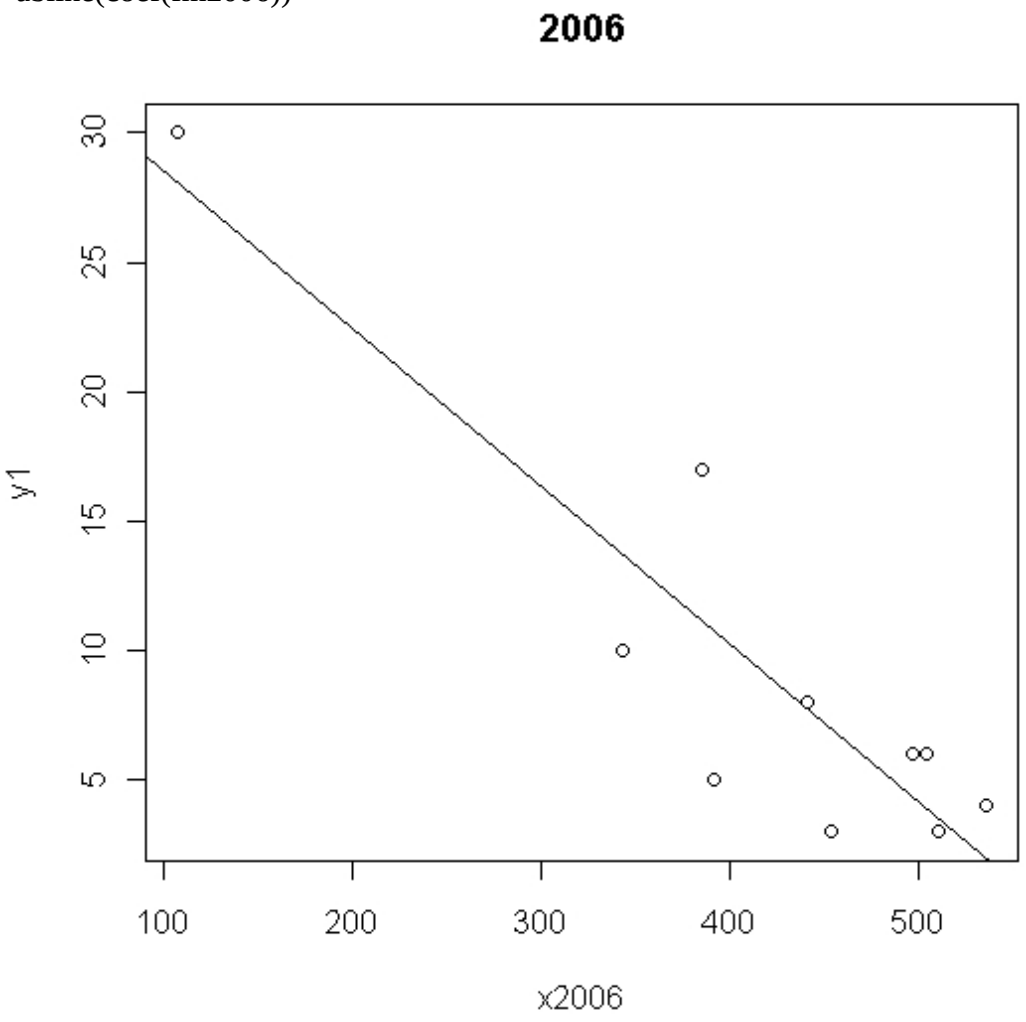
Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	34.600160	4.331329	7.988	4.41e-05	***
x2006	-0.060868	0.009981	-6.098	0.00029	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.756 on 8 degrees of freedom
Multiple R-squared: 0.823, Adjusted R-squared: 0.8008
F-statistic: 37.19 on 1 and 8 DF, p-value: 0.0002901
因為 slope p – value=0.0002901 有很顯著的 linear relation

```
> plot(x2006,y1,main="2006")
> abline(coef(fm2006))
```




```
> x2007<-c(102,336,334,450,398,450,450,450,450,450)
> fm2007=lm(y1~x2007) #y1 對 x2007 做 regression
> summary(fm2007)
```

Call:
lm(formula = y1 ~ x2007)

Residuals:

Min	1Q	Median	3Q	Max
-3.0606	-1.6110	-0.2853	1.1390	4.0851

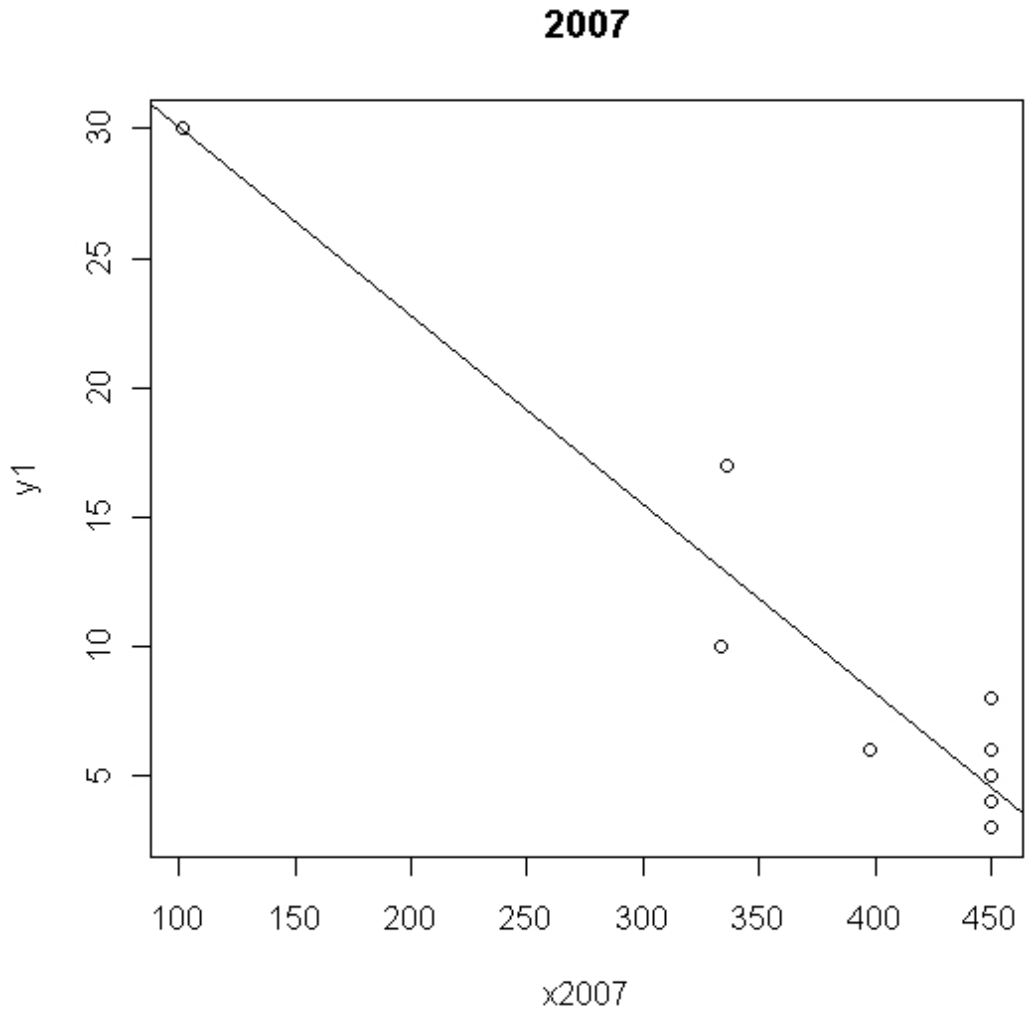
Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	37.389364	3.043316	12.286	1.79e-06	***
x2007	-0.072841	0.007589	-9.599	1.15e-05	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.523 on 8 degrees of freedom
Multiple R-squared: 0.9201, Adjusted R-squared: 0.9101
F-statistic: 92.13 on 1 and 8 DF, p-value: 1.152e-05
因為 slope $p-$value≒0, 有很顯著的 linear relation

```
> plot(x2007,y1,main="2007")
> abline(coef(fm2007))
```




```
> y2<-c(30,17,12,9,7,6,5,4.5,2,2)
> x2008<-c(124,354,281,450,450,450,341,501,450,501)
> fm2008=lm(y2~x2008) #y2 對 x2008 做 regression
> summary(fm2008)

Call:
lm(formula = y2 ~ x2008)

Residuals:
      Min        1Q    Median        3Q       Max
-7.5903 -2.8193  0.8669  3.0557  5.2394
```

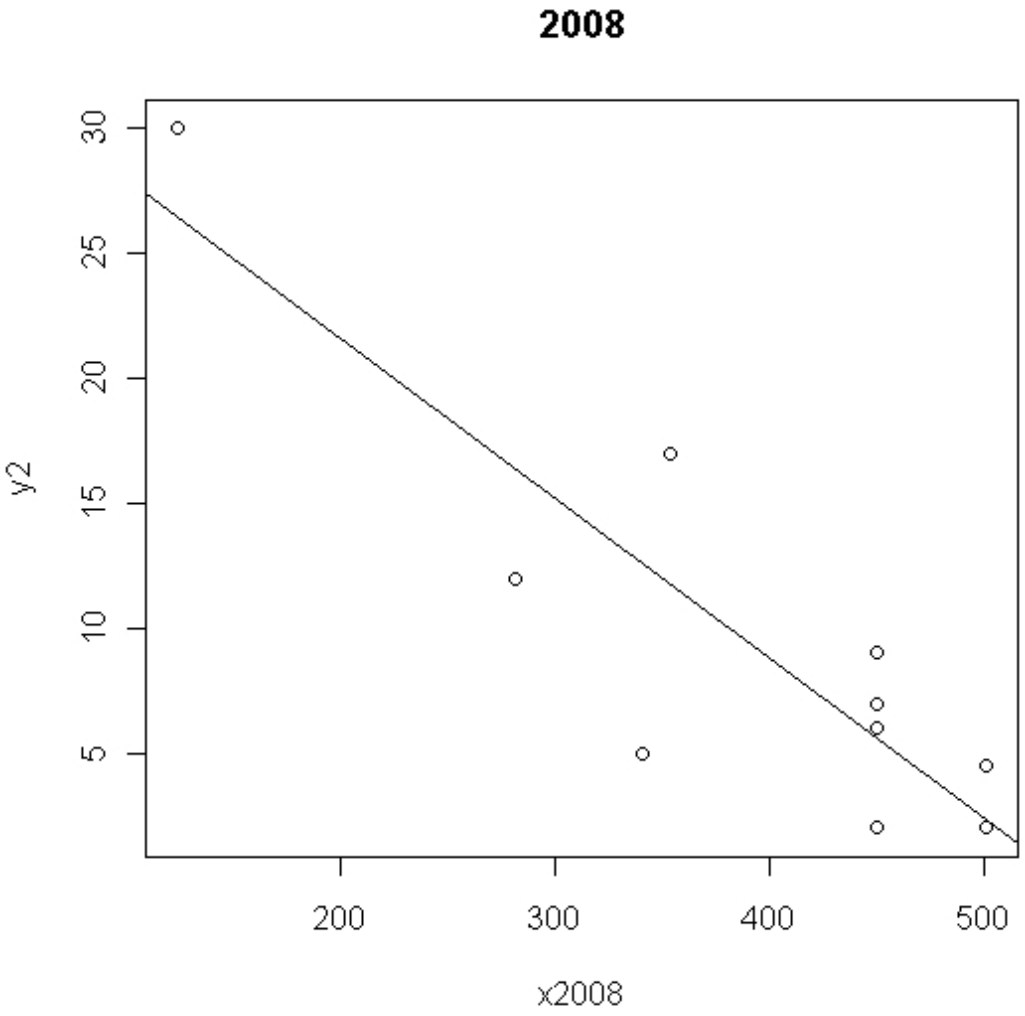
Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	34.35546	4.93870	6.956	0.000118	***
x2008	-0.06383	0.01217	-5.246	0.000777	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.308 on 8 degrees of freedom
Multiple R-squared: 0.7748, Adjusted R-squared: 0.7467
F-statistic: 27.53 on 1 and 8 DF, p-value: 0.0007772
因為 slope $p-$value=0.0007772 有很顯著的 linear relation

```
> plot(x2008,y2,main="2008")
> abline(coef(fm2008))
```



3. 結論

從 regression 的理論知道，教育部有越高的補助與世界大學的排名越前面高度的相關，因此可以知道教育部這樣的補助政策看起來是有道理的，至少在排名這個要求下，要集中在一些重點學校的政策是很合理的。

4. 建議與改進

首先當然是沒有一個很好的數學模型，或是理論性的證明去支持此建議，希望之後能進一步去探討此問題。還有在 regression 中，我們需要去假設其為 normal distribution, 並且每個 random variable 是 identical independent, 且必須是 equal variance;

但是 rank 這樣的 scale 比較接近是 ordinal 的資料，而且由於有些學校 rank 太過後面並沒有排到，故我將 400~500 名皆設為 450 名，500 名以上皆設定為 501 名，但是事實上其資迅並不齊全。改進的方式有三種，一種可能可以採用 EM(estimate-maximize) Algorithm, 把遺失的資料納近來考慮；再來就是可能用無母數的方法或許會更貼切，會比 normal distribution 的假設來的貼切；最後一種可能就適用 generalized linear regression model, 是著去找 link function, 來解決資料型態不是 normal 的假設。

當然還有一個最大的問題，就是蒐集資料的困難，因為教育部與各個學校不願意將此資料公佈給個人使用，使得我在蒐集資料上遇到不小的瓶頸，可是我覺得如果要大家都能夠檢視到底 5 年 500 億是否達到成效，應該將其成果公佈給社會大眾知道。

在這次的過程中，讓我學習到很多，覺得自己必須要多充實 statistical method, 尤其是在實務方面，可以學習例如像 Latent Class 等等，很多問題是真的去做了才知道哪裡有問題，希望自己在這方面能夠多多學習。