報告題名：

全國大專院校(大學與獨立學院)學生人數逐年變化之系統模型建制與評估

作者：黃志偉
系級：土木工程學系研究所
學號：M9406413

開課老師：林正紋老師

課程名稱：統計學應用與實務

開課系所：土木工程學系研究所

開課學年：九十四學年度 第一學期

Summary

隨著學校的增加以及錄取率的提升，大學生的學生人數每年都在增加，因此藉著統計學的方法建立一個系統模型，用來預測未來大學生(大學與獨立學院)的學生人數。

利用 Multiple Regression 和冪級數預測未來大學生(大學與獨立學院)的學生人數。以學生人數為 response variable y，學校數、學系數、班級數分別為 explanatory variable x1 x2 x3，觀察彼此之間的相關係數，並繪製 Normal quantile plots 及 scatterplots，接著進行線性與非線性之迴歸分析，最後得到統計學推論以用來預測學生人數逐年之變化，而提供相關決策者之一參考數據。

Key Words：系統模型、Multiple Regression、冪級數、統計學推論

Contents

Part I. Project Design

1. 目的與動機：

由於隨著學校的增加以及錄取率的提升，大學生的學生人數每年都在增加，因此藉著統計學的方法預測未來大學生的學生人數。

2. 方法與項目：

利用 Multiple Regression 預測未來大學生的學生人數。

考慮學校數、學系數以及班級數與大學生的學生人數之間的關係(相關係數)，再判斷各因素是否需要取捨。

應用軟體：Stata 與 Excel。

概要流程：

(a) 計算各項的值(平均值、標準差、最大值、最小值)以及 Normal quantile plot。

(b) 計算與比較各項之間的相關係數。

(c) Regress、predict hat、regression line (線性)。

(d) t statistic = b / SE_b。

(e) 比較(P-value)。

(f) 重新 regress，並與之前比較、predict hat、regression line (線性)。

(g) 展開二次方項。

(h) 導入二次方項並計算各項的值(平均值、標準差、最大值、最小值)。

(i) 計算各項之間的相關係數。

(j) Regress。

(k) 展開三次方項。

(l) 導入三次方項並計算各項的值(平均值、標準差、最大值、最小值)。

(m) 計算各項之間的相關係數。

(n) Regress。

(o) 觀察結果與判斷決定。

(p) predict hat、regression line (非線性)。

(q) Correlate。

(r) 畫圖。

(s) 討論。

(t) 預測。

3. 樣本數目：7 (87~93學年度)。

Part II. Data Collection and Organization

大學+獨立學院(Univ. & College)				
學年度	學生人數	學校數	學系數	班級數
87	409705	84	1842	8008
88	470030	105	2195	8940
89	564059	127	2602	10592
90	677171	135	3098	12688
91	770915	139	3700	14892
92	837602	142	4059	16796
93	894528	145	4406	29454

資料來源：教育部統計處官方網站。

http://www.edu.tw/EDU_WEB/Web/STATISTICS/index.htm [1]

session	y	x1	x2	x3
87	409705	84	1842	8008
88	470030	105	2195	8940
89	564059	127	2602	10592
90	677171	135	3098	12688
91	770915	139	3700	14892
92	837602	142	4059	16796
93	894528	145	4406	29454

y：大學生人數

x1：學校數

x2：學系數

x3：班級數

以下的統計分析方法部分採自於 Moore DS, McCabe GP 書中之 11.2 [2]

Part III. Data Analysis and Inference

(a) 計算各項的值(平均值、標準差、最大值、最小值)：

```
. log
(closed)

. edit
(4 vars, 7 obs pasted into editor)

. summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
y	7	660572.9	185778.3	409705	894528
x1	7	125.2857	22.61794	84	145
x2	7	3128.857	968.5973	1842	4406
x3	7	14481.43	7310.5	8008	29454

由於 Data 只有 7 組，因此繪出各項的 Normal quantile plot 圖以觀察是否爲常態分佈。

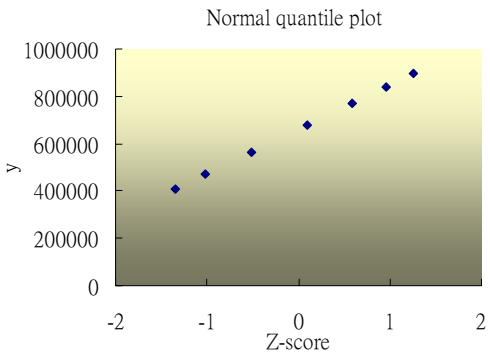

圖 1

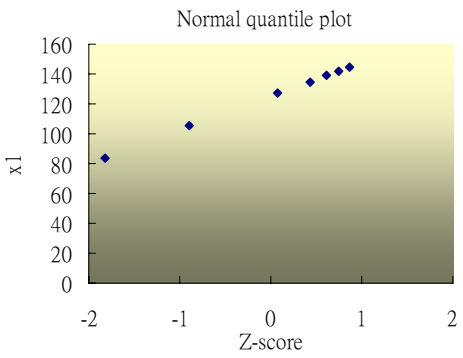

圖 2

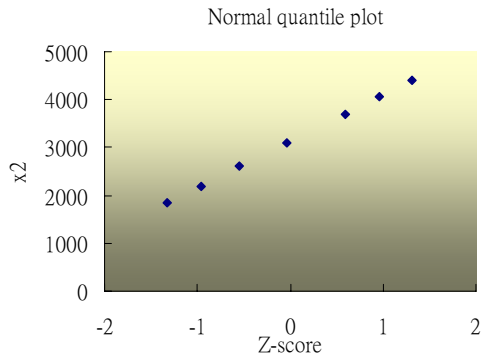

圖 3

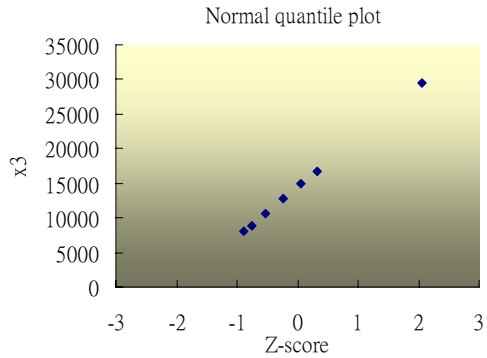

圖 4

由圖 1 至圖 4 顯示，均爲常態分佈。

如果不是常態分佈的話(有 outlier)，造成 t distribution 的誤差大，因而導致 Multiple Regression 的統計結果誤差也大。

(b) 計算與比較各項之間的相關係數：

```
. correlate
(obs=7)
```

	y	x1	x2	x3
y	1.0000			
x1	0.9243	1.0000		
x2	0.9981	0.9086	1.0000	
x3	0.8570	0.7006	0.8737	1.0000

y 與 x1、x2、x3 的相關係數 r 分別爲 0.9243、0.9981、0.8570 看起來不錯，其中以 y 與 x3 的相關係數爲最低，但 x1、x2、x3 彼此之間的相關係數(0.9086、0.7006、0.8737)還滿高的，其中以 x1 與 x2 的相關係數爲最高(比較不獨立)，所以此時還不能決定要刪除哪一項，因此進一步地觀察。

圖 5 至圖 10 爲各項之間的相關分佈圖。

. scatter y x1

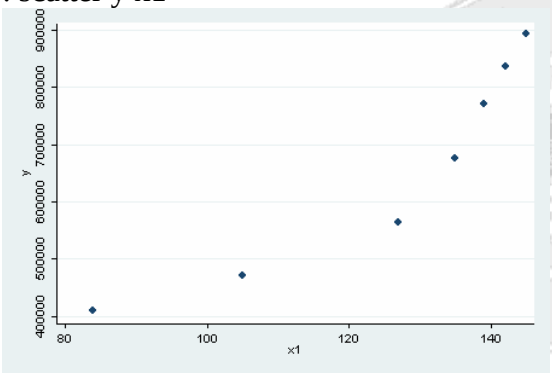

圖 5

. scatter y x2

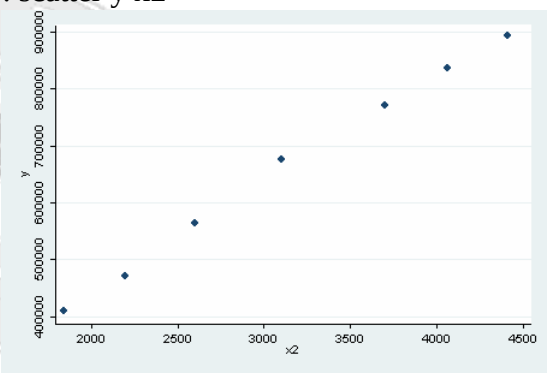

圖 6

. scatter y x3

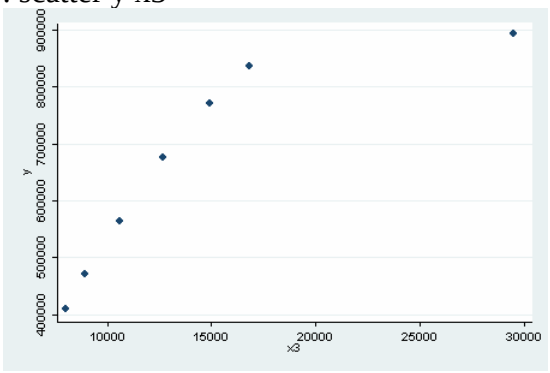

圖 7

. scatter x1 x2

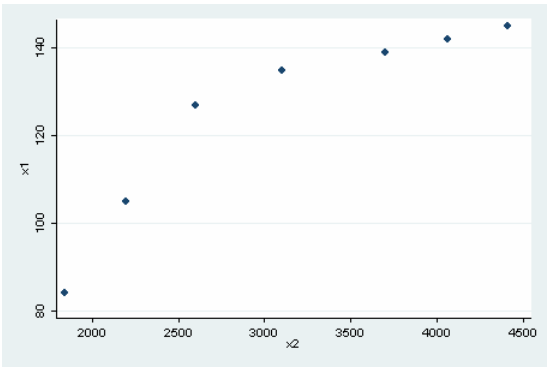

圖 8

. scatter x1 x3

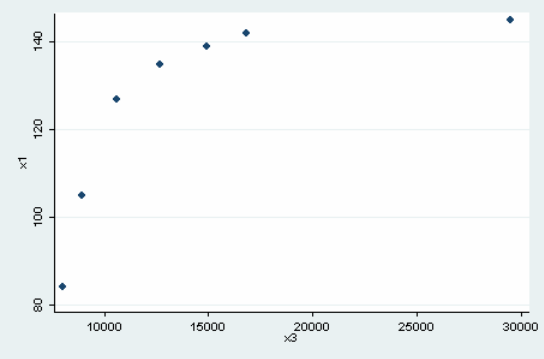

圖 9

. scatter x2 x3

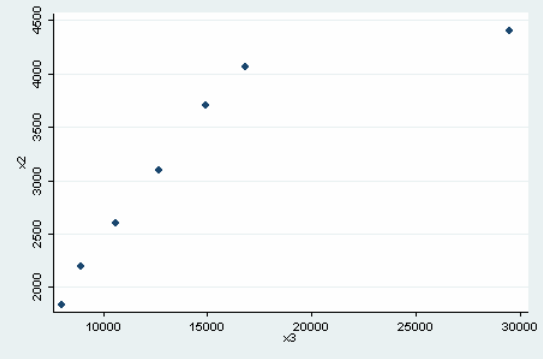

圖 10

(c)

```
. regress y x1 x2 x3
```

Source	SS	df	MS
Model	2.0668e+11	3	6.8894e+10
Residual	399123457	3	133041152
Total	2.0708e+11	6	3.4514e+10

Number of obs	7
F(3, 3)	517.84
Prob > F	0.0001
R-squared	0.9981
Adj R-squared	0.9961
Root MSE	11534

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x1	689.1547	560.8114	1.23	0.307	-1095.598	2473.907
x2	181.9079	19.20788	9.47	0.002	120.7799	243.0359
x3	-.7730071	1.489935	-0.52	0.640	-5.514646	3.968632
_cons	16262.03	33881.48	0.48	0.664	-91563.97	124088

. predict hat

Regression [2]：

regression line (線性)[3]：$\hat{y}$ = a0 + a1*x1 + a2*x2 + a3*x3

= 16262.03 + 689.15*x1 + 181.91*x2 – 0.77*x3

(d) t statistic

x1(t statistic) = b / SE_b =689.1547 / 560.8114 = 1.228853

x2(t statistic) = b / SE_b =181.9079 / 19.20788 = 9.470484

x3(t statistic) = b / SE_b =-0.7730071 / 1.489935 = -0.51882

(e) P-value

R-squared = 0.9981 = 99.81% > 80% ⇒ 雖然很好，但

x1(P-value)：0.307 = 30.7% > 5%

x2(P-value)：0.002 = 0.2% < 5% ⇒ 保留

x3(P-value)：0.640 = 64.0% > x1(P-value) > 5% ⇒ 太大，因此刪除 x3

(f)

```
. regress y x1 x2
```

Source	SS	df	MS		
Model	2.0665e+11	2	1.0332e+11	Number of obs =	7
Residual	434934599	4	108733650	F(2, 4) =	950.24
				Prob > F =	0.0000
				R-squared =	0.9979
				Adj R-squared =	0.9968
Total	2.0708e+11	6	3.4514e+10	Root MSE =	10428

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x1	822.5774	450.55	1.83	0.142	-428.35	2073.505
x2	173.9798	10.5209	16.54	0.000	144.7691	203.1905
_cons	13157.63	30148.86	0.44	0.685	-70549.04	96864.29

. predict hat

Regression [2]：

regression line (線性)[3]：$\hat{y} = a0 + a1*x1 + a2*x2$

= 13157.63 + 822.58*x1 + 173.98*x2

R-squared = 0.9979 = 99.79%，與之前的 99.81%差異很少，且

x1(P-value)：0.142 = 14.2% < 之前的 30.67%

x2(P-value)：0.000 = 0.0% < 之前的 0.2%

所以刪除 x3 是 O.K 的。

(g)

冪級數展開 $\Rightarrow \sum(a1*x1+a2*x2)^2$ [4]

= a1*x1 + a2*x2 + a3*x1^2 + a4*x1*x2 + a5*x2^2

= a1*x1 + a2*x2 + a3*x4 + a4*x5 + a5*x6

(h)

刪除 x3 的 data，並導入二次項 x4(= x1^2)、x5(= x1*x2)、x6(= x2^2)。

. edit

- preserve

- drop x3

(3 vars, 7 obs pasted into editor)

```
. summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
y	7	660572.9	185778.3	409705	894528
x1	7	125.2857	22.61794	84	145
x2	7	3128.857	968.5973	1842	4406
x4	7	16135	5215.426	7056	21025
x5	7	409062.1	180149.8	154728	638870
x6	7	1.06e+07	6093417	3392964	1.94e+07

(i)

```
. correlate
(obs=7)
```

	y	x1	x2	x4	x5	x6
y	1.0000					
x1	0.9243	1.0000				
x2	0.9981	0.9086	1.0000			
x4	0.9453	0.9978	0.9306	1.0000		
x5	0.9988	0.9346	0.9976	0.9534	1.0000	
x6	0.9873	0.8621	0.9942	0.8890	0.9855	1.0000

(j)

```
. regress y x1 x2 x4 x5 x6
```

Source	SS	df	MS
Model	2.0707e+11	5	4.1413e+10
Residual	15541163.3	1	15541163.3
Total	2.0708e+11	6	3.4514e+10

Number of obs	7
F(5, 1)	2664.74
Prob > F	0.0147
R-squared	0.9999
Adj R-squared	0.9995
Root MSE	3942.2

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x1	25153.73	15542.54	1.62	0.352	-172333	222640.5
x2	-2131.418	1113.771	-1.91	0.307	-16283.22	12020.38
x4	-349.1905	182.3364	-1.92	0.306	-2665.995	1967.614
x5	25.81376	12.12574	2.13	0.280	-128.2583	179.8859
x6	-.1999492	.0884826	-2.26	0.265	-1.324227	.9243282
_cons	1371073	476699.7	2.88	0.213	-4685971	7428118

此時的 R-squared = 0.9999 ≅ 1。

試著導入三次方項觀察結果。

(k)

冪級數展開 $\Rightarrow \sum (a1*x1 + a2*x2)^3$ [4]

= a1*x1 + a2*x2 + a3*x1^2 + a4*x1*x2 + a5*x2^2 + a6*x1^3 +
a7*x1^2*x2 + a8*x1*x2^2 + a9*x2^3

= a1*x1 + a2*x2 + a3*x4 + a4*x5 + a5*x6 + a6*x7 + a7*x8 + a8*x9 +
a9*x10

(l)

導入三次項 x7(= x1^3)、x8(= x1^2*x2)、x9(= x1*x2^2)、x10(= x2^3)。

. edit

- preserve

(4 vars, 7 obs pasted into editor)

```
. summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
y	7	660572.9	185778.3	409705	894528
x1	7	125.2857	22.61794	84	145
x2	7	3128.857	968.5973	1842	4406
x4	7	16135	5215.426	7056	21025
x5	7	409062.1	180149.8	154728	638870
x6	7	1.06e+07	6093417	3392964	1.94e+07
x7	7	2122374	923263.2	592704	3048625
x8	7	5.45e+07	2.97e+07	1.30e+07	9.26e+07
x9	7	1.43e+09	9.55e+08	2.85e+08	2.81e+09
x10	7	3.82e+10	3.03e+10	6.25e+09	8.55e+10

(m)

```
. correlate
(obs=7)
```

	y	x1	x2	x4	x5	x6	x7	x8
y	1.0000							
x1	0.9243	1.0000						
x2	0.9981	0.9086	1.0000					
x4	0.9453	0.9978	0.9306	1.0000				
x5	0.9988	0.9346	0.9976	0.9534	1.0000			
x6	0.9873	0.8621	0.9942	0.8890	0.9855	1.0000		
x7	0.9623	0.9917	0.9490	0.9980	0.9683	0.9126	1.0000	
x8	0.9988	0.9348	0.9970	0.9540	0.9999	0.9851	0.9692	1.0000
x9	0.9891	0.8694	0.9952	0.8959	0.9877	0.9998	0.9188	0.9875
x10	0.9677	0.8144	0.9792	0.8450	0.9654	0.9953	0.8725	0.9651

	x9	x10
x9	1.0000	
x10	0.9942	1.0000

(n)

```
. regress y x1 x2 x4 x5 x6 x7 x8 x9 x10
```

Source	SS	df	MS		
Model	2.0708e+11	6	3.4514e+10	Number of obs	7
				F(6, 0)	.
				Prob > F	.
Residual	0	0	.	R-squared	1.0000
				Adj R-squared	.
Total	2.0708e+11	6	3.4514e+10	Root MSE	0

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x1	(dropped)					
x2	(dropped)					
x4	(dropped)					
x5	30.09948	.	.	.	.	.
x6	-1.809916	.	.	.	.	.
x7	.991401	.	.	.	.	.
x8	-.4552444	.	.	.	.	.
x9	.027047	.	.	.	.	.
x10	-.0002484	.	.	.	.	.
_cons	1066831	.	.	.	.	.

(o)

雖然 R-squared = 1.0000，但由於 x1、x2、x4 的係數 dropped，可能是因爲三次方項的值太大而把較小值的係數給吸收掉，因爲試著先刪除 x10(= x2^3)並觀察結果。

. edit

- preserve

- drop x10

```
. regress y x1 x2 x4 x5 x6 x7 x8 x9
```

Source	SS	df	MS		
Model	2.0708e+11	6	3.4514e+10	Number of obs	7
				F(6, 0)	.
				Prob > F	.
Residual	0	0	.	R-squared	1.0000
				Adj R-squared	.
Total	2.0708e+11	6	3.4514e+10	Root MSE	0

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x1	(dropped)					
x2	(dropped)					
x4	1256.676	.	.	.	.	.
x5	-89.44979	.	.	.	.	.
x6	1.06113	.	.	.	.	.
x7	-11.01301	.	.	.	.	.
x8	.7976221	.	.	.	.	.
x9	-.0107663	.	.	.	.	.
_cons	1011747	.	.	.	.	.

R-squared = 1.0000，且仍有二項(x1、x2)的係數 dropped，因爲再試著刪除 x8(= x1^2*x2)與 x9(= x1*x2^2)並觀察結果。

. edit

- preserve

- drop x9

- drop x8

```
. regress y x1 x2 x4 x5 x6 x7
```

Source	SS	df	MS		
Model	2.0708e+11	6	3.4514e+10	Number of obs =	7
Residual	0	0	.	F(6, 0) =	.
				Prob > F =	.
				R-squared =	1.0000
				Adj R-squared =	.
Total	2.0708e+11	6	3.4514e+10	Root MSE =	0

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x1	1353788	.	.	.	.	.
x2	58214.78	.	.	.	.	.
x4	-11080.95	.	.	.	.	.
x5	-628.4204	.	.	.	.	.
x6	4.296064	.	.	.	.	.
x7	64.68168	.	.	.	.	.
_cons	-9.80e+07	.	.	.	.	.

此時 R-squared = 1.0000，且每一項的係數都有了，接著就可以寫出 regression line (非線性)。

(p)

. predict hat

(option xb assumed; fitted values)

Regression [2]：

regression line (非線性)[3]：

$$\hat{y} = -98000000 + 1353788*x1 + 58214.78*x2 - 11080.95*x4 - 628.42*x5 + 4.30*x6 + 64.68*x7$$

(q)

```
. correlate
(obs=7)
```

	y	x1	x2	x4	x5	x6	x7
y	1.0000						
x1	0.9243	1.0000					
x2	0.9981	0.9086	1.0000				
x4	0.9453	0.9978	0.9306	1.0000			
x5	0.9988	0.9346	0.9976	0.9534	1.0000		
x6	0.9873	0.8621	0.9942	0.8890	0.9855	1.0000	
x7	0.9623	0.9917	0.9490	0.9980	0.9683	0.9126	1.0000

(r)

圖 11 至圖 16 爲 y 與各項(x1、x2、x4、x5、x6、x7)之迴歸線。

. twoway (qfit y x1) (scatter y x1)

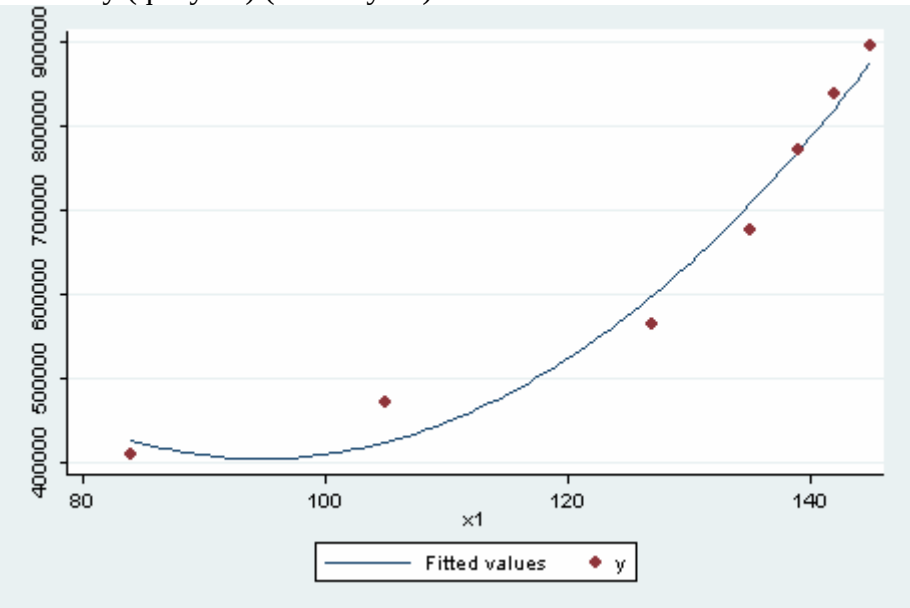

圖 11

. twoway (qfit y x2) (scatter y x2)

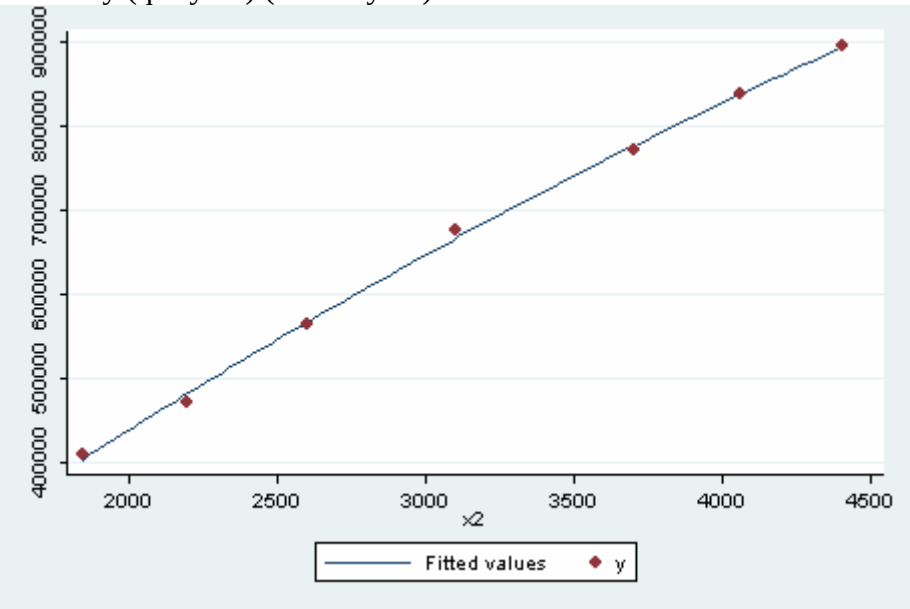

圖 12

. twoway (qfit y x4) (scatter y x4)

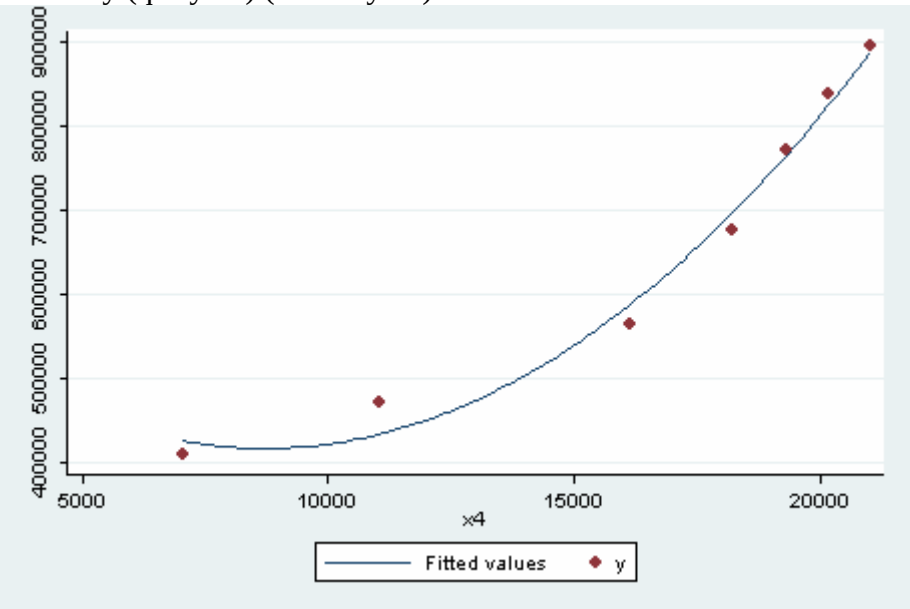

圖 13

. twoway (qfit y x5) (scatter y x5)

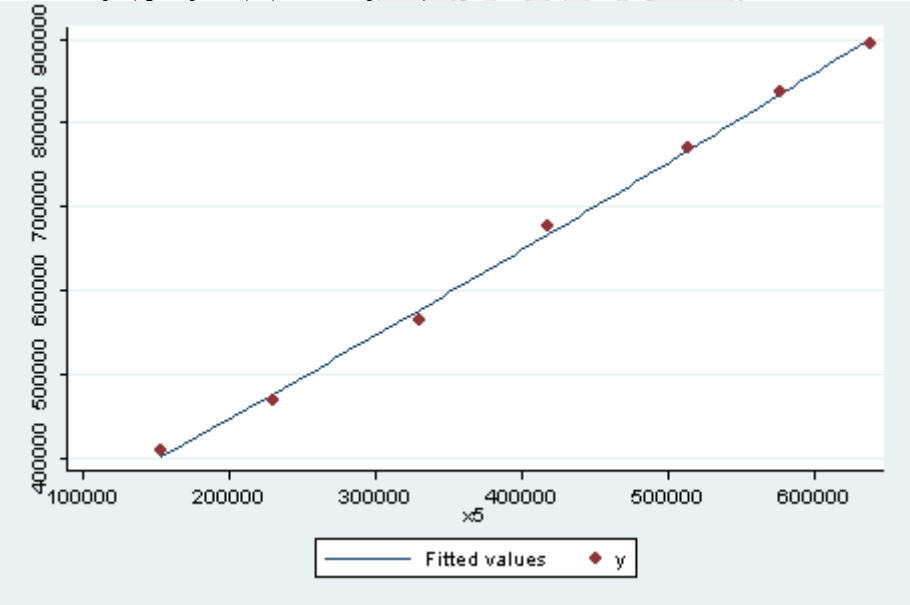

圖 14

. twoway (qfit y x6) (scatter y x6)

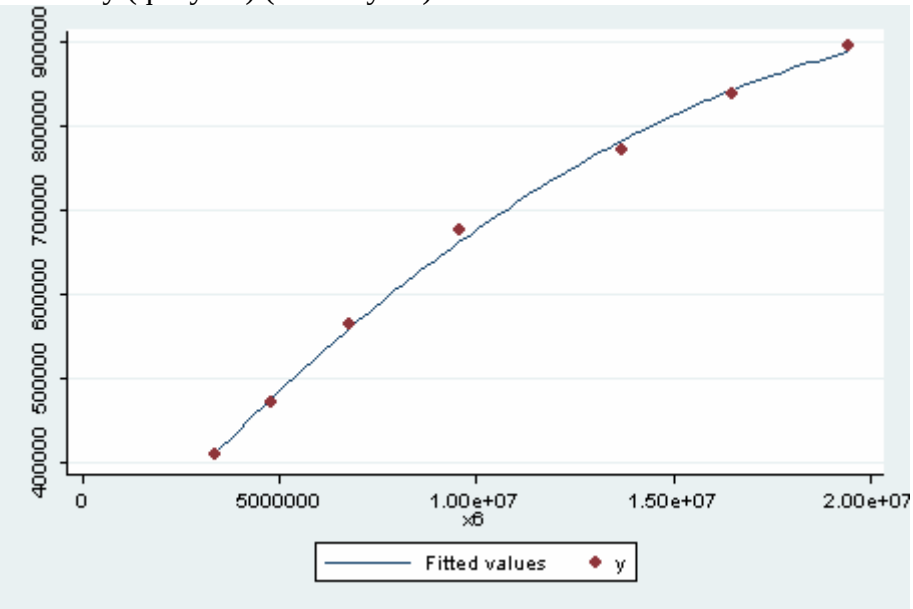

圖 15

. twoway (qfit y x7) (scatter y x7)

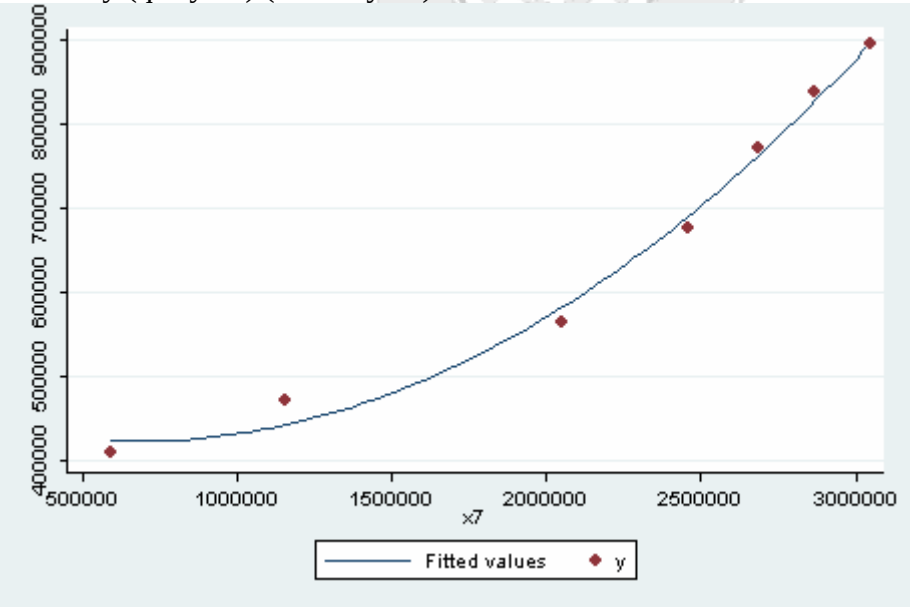

圖 16

(s)

只用線性項(x1、x2)分析 ⇒ R-squared = 0.9979

導入二次項(x4(= x1^2)、x5(= x1*x2)、x6(= x2^2))分析 ⇒ R-squared = 0.9999

導入三次項(x7(= x1^3))分析 ⇒ R-squared = 1.0000

因此二次方項與三次方項只影響了 0.0021，也就是 0.21%的 R-squared。

(t) 預測

由 regression line(非線性)[3]：

$$\hat{y} = -98000000 + 1353788*x1 + 58214.78*x2 - 11080.95*x4 - 628.42*x5 + 4.30*x6 + 64.68*x7$$

以 87 學年度(學校數、學系數)預測 86 學年度之學生人數爲 454304 人；

以 93 學年度(學校數、學系數)預測 94 學年度之學生人數爲 998181 人。

也可以預測未來幾年之學生人數。

Part IV. References

1. http://www.edu.tw/EDU_WEB/Web/STATISTICS/index.htm

2. Moore DS, McCabe GP, *Introduction to the Practice of Statistics*, W.H. Freeman and Company, New York (2005).

3. Masri SF, "A Hybrid Parametric/Nonparametric Approach for the Identification of Nonlinear Systems," *Probabilistic Engineering Mechanics*, Vol. 9, pp. 47-57 (1994).

4. Lin JW, Betti R, "On-line Identification and Damage Detection in Non-linear Structural Systems Using a Variable Forgetting Factor Approach," *Earthquake Engineering and Structural Dynamics*, Vol. 33(4), pp. 419-444 (2004).