《迴歸分析》

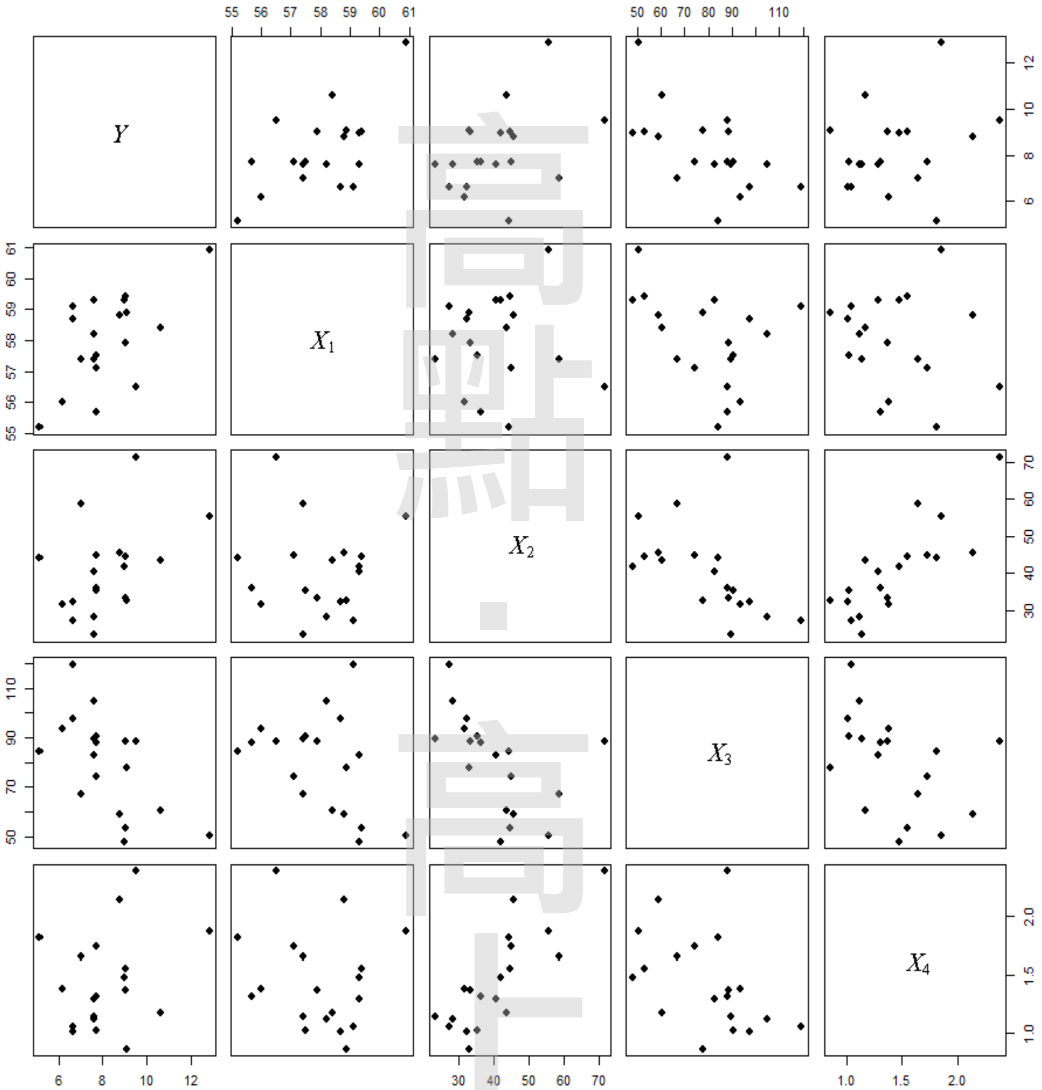
下表為2012年時19個縣市的資料，若以粗出生率為反應變數Y，四個解釋變數分別為勞動力參與率（X_1）、就業者之教育程度結構-大專及以上（X_2）、老化指數（X_3）、平均每人環保經費（X_4）。

縣市	粗出生率（Y）	勞動力參與率（X_1）	就業者之教育程度結構-大專及以上（X_2）	老化指數（X_3）	平均每人環保經費（X_4）
新北市	8.79	58.8	45.59	59.00	2.14
臺北市	9.54	56.5	71.40	88.31	2.38
臺中市	9.04	59.4	44.49	53.28	1.55
臺南市	7.58	59.3	40.46	82.69	1.29
高雄市	7.72	57.1	44.88	74.13	1.73
宜蘭縣	7.71	57.5	35.30	90.68	1.02
桃園縣	8.99	59.3	41.89	48.07	1.48
新竹縣	10.64	58.4	43.61	60.40	1.17
苗栗縣	9.05	57.9	33.24	88.50	1.37
彰化縣	9.07	58.9	32.76	77.93	0.86
南投縣	6.63	58.7	32.12	97.38	1.01
雲林縣	7.60	58.2	28.24	104.76	1.12
嘉義縣	6.62	59.1	27.02	119.34	1.05
屏東縣	6.16	56.0	31.56	93.39	1.38
臺東縣	7.61	57.4	23.28	89.59	1.14
花蓮縣	7.71	55.7	36.20	88.16	1.31
基隆市	5.17	55.2	44.16	84.23	1.81
新竹市	12.85	60.9	55.35	50.40	1.86
嘉義市	7.00	57.4	58.54	66.89	1.65

變數間的相關係數矩陣如下：

	Y	X_1	X_2	X_3	X_4
Y	1	0.605	0.413	-0.616	0.248
X_1	0.605	1	-0.014	-0.386	-0.154
X_2	0.413	-0.014	1	-0.522	0.819
X_3	-0.616	-0.386	-0.522	1	-0.429
X_4	0.248	-0.154	0.819	-0.429	1

兩兩變數間的散布圖如下：



下列為六個不同迴歸模型的估計結果：

	估計值	標準誤
截距項	5.7314	1.3622
X_2	0.0605	0.0323

	估計值	標準誤
截距項	12.5929	1.4054
X_3	-0.0552	0.0171

	估計值	標準誤
截距項	11.3715	2.6355
X_2	0.0185	0.0335

X_3	-0.0493	0.0205
	估計值	標準誤
截距項	-36.3909	11.7069
X_1	0.7256	0.2009
X_2	0.0617	0.0247
	估計值	標準誤
截距項	-25.9329	14.1083
X_1	0.5948	0.2221
X_2	0.0402	0.0296
X_3	-0.0251	0.0196
	估計值	標準誤
截距項	-24.9931	15.3071
X_1	0.5811	0.2394
X_2	0.0472	0.0463
X_3	-0.0256	0.0205
X_4	-0.2671	1.3228

下表為配適線性迴歸模型，不同變數所得之模型選取準則的結果。

模型	模型中的變數	p	SSE_p	R_p^2	$R_{a,p}^2$	C_p
A	X_1	2	34.310	0.365	0.328	6.625
B	X_2	2	44.836	0.171	0.122	13.260
C	X_3	2	33.557	0.379	0.343	6.150
D	X_4	2	50.735	0.062	0.006	16.977
E	X_1、X_2	3	24.695	0.543	0.486	2.565
F	X_1、X_3	3	25.019	0.537	0.479	2.769
G	X_1、X_4	3	27.841	0.485	0.421	4.548
H	X_2、X_3	3	32.928	0.391	0.315	7.754
I	X_2、X_4	3	43.507	0.195	0.095	14.422
J	X_3、X_4	3	33.540	0.380	0.302	8.139
K	X_1、X_2、X_3	4	22.277	0.588	0.506	3.041
L	X_1、X_2、X_4	4	24.694	0.543	0.452	4.564
M	X_1、X_3、X_4	4	23.862	0.559	0.470	4.040
N	X_2、X_3、X_4	4	31.562	0.416	0.300	8.893
O	X_1、X_2、X_3、X_4	5	22.212	0.589	0.472	5.000

表中p為各模型中迴歸係數的個數，SSE_p 為該模型下所得的誤差平方和（error sum of squares），R_p^2 為其判定係數（coefficient of determination），$R_{a,p}^2$ 為調整的判定係數（adjusted coefficient of determination），C_p 為Mallows' C_p criterion。

下列問題皆在顯著水準為0.05下，進行統計假設檢定：

一、若由X_2與X_3的散布圖判斷，該圖中可能有一個離群值。請將該離群值排除後，重新計算X_2與X_3相

關係數。（10分）

（下列的問題皆是在無離群值存在的狀況下作答。）

二、檢定模型A的X_1的迴歸係數是否為0？（10分）

三、針對模型0，寫出其變異數分析表。檢定其迴歸係數是否同時為0。（10分）

四、若 $SSR_{eg}(X_i | X_j)$ 代表給定X_j已在模型中，X_i加入模型中的額外平方和（extra sum of squares）。請分別計算 $SSR_{eg}(X_2 | X_1)$、$SSR_{eg}(X_2 | X_3)$、$SSR_{eg}(X_2, X_3 | X_1)$、$SSR_{eg}(X_1, X_2 | X_3, X_4)$。（12分）

五、藉由迴歸估計結果及報表，詳細說明「就業者之教育程度結構-大專及以上」（X_2）此一變數對於粗出生率的影響，是否具統計顯著意義？（10分）

六、請詳細說明前述各模型選取準則的定義，包括SSE_p、R_p^2、$R_{a,p}^2$及C_p。並說明他們在模型選取的判斷原則為何？（12分）

七、在不同的p下，請依各準則判斷其所得之最適模型。（12分）

八、請決定一個影響粗出生率的最適迴歸模型，並說明理由。（6分）

九、計算所得的最適迴歸模型的均方誤（MSE），並說明其意義。（6分）

十、寫出線性迴歸模型之誤差項的假設。並針對誤差項的各項假設，分別提出一種殘差分析的圖形，及說明在符合假設下各圖形應呈現的型態。（12分）

試題評析	今年考卷難度屬於中下，所有考題均於考古題中常常出現，且沒有太多計算。最後一題殘差分析圖形也在趙治勳編著的《迴歸分析熱門題庫》第二篇第四章都有詳細介紹。 今年考卷平均分數70分，考取的話應該要有80分以上，預估今年獲得高分之人數應該會不少。

答：

(一)

由X2與X3之散布圖判斷出臺北市可能為離群值，刪除該值後X2與X3之相關係數為

$$r_{23} = \frac{SS_{23}}{\sqrt{SS_2}\sqrt{SS_3}} = -0.7632$$

其中
$$SS_2 = \sum X_{2i}^2 - \frac{(\sum X_{2i})^2}{n} = 28638.2077 - \frac{(698.69)^2}{18} = 1517.779$$

$$SS_3 = \sum X_{3i}^2 - \frac{(\sum X_{3i})^2}{n} = 120068.662 - \frac{(1428.82)^2}{18} = 6650.518$$

$$SS_{23} = \sum X_{2i}X_{3i} - \frac{(\sum X_{2i})(\sum X_{3i})}{n} = 53036.5039 - \frac{(698.69)(1428.82)}{18} = -2424.732$$

(二)

假設模型A：$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_{ij}$，$\varepsilon_{ij} \overset{iid}{\sim} N(0,\sigma^2)$,$i = 1,2,\ldots,19$

$H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$

T.S.: $T = \dfrac{r_{Y1}\sqrt{n-2}}{\sqrt{1-r_{Y1}^2}} \sim t_{(19-2=17)}$

R.R.: Reject H_0 at $\alpha = 0.05$ if $|T^*| > t_{(17)0.025} = 2.11$

$\because |T^*| = \left|\dfrac{0.605\sqrt{19-2}}{\sqrt{1-0.605^2}}\right| = 3.133$　　$\therefore$ reject H_0

我們有足夠證據去推論 $\beta_1 \neq 0$

(三)

假設模型O：$Y_i=\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+\beta_3X_{3i}+\beta_4X_{4i}+\varepsilon_{ij}$，$\varepsilon_{ij}\overset{iid}{\sim}N(0,\sigma^2)$,$i=1,2,\cdots,19$

ANOVA TABLE

source	SS	d. f.	MS	F
Reg	31.8318	4	7.95795	$F^*=1.58657$
Error	22.212	14	1.58657	
Total	54.0438	18		

$H_0:\beta_1=\beta_2=\beta_3=\beta_4=0$ vs H_1:至少一個$\beta_i\neq\beta_j, i\neq j$

T.S.: $F=\dfrac{MSR}{MSE}\sim F_{(4,14)}$

R.R.: Reject H_0 at $\alpha=0.05$ if $F^*>F_{(4,14)0.05}=3.1122$

$\because F^*=1.58657$ $\therefore$ don't reject H_0

我們沒有足夠證據去推論模型O之迴歸係數不為同時為0

(四)

$SS\,\mathrm{Reg}(X_2\mid X_1)=SSE(X_1)-SSE(X_1,X_2)=34.31-24.695=9.615$

$SS\,\mathrm{Reg}(X_2\mid X_3)=SSE(X_3)-SSE(X_2,X_3)=33.557-25.019=8.538$

$SS\,\mathrm{Reg}(X_2,X_3\mid X_1)=SSE(X_1)-SSE(X_1,X_2,X_3)=34.31-22.277=12.033$

$SS\,\mathrm{Reg}(X_1,X_2\mid X_3,X_4)=SSE(X_3,X_4)-SSE(X_1,X_2,X_3,X_4)=33.54-22.212=11.328$

(五)

假設模型：$Y_i=\beta_0+\beta_2X_{2i}+\varepsilon_{ij}$，$\varepsilon_{ij}\overset{iid}{\sim}N(0,\sigma^2)$,$i=1,2,\cdots,19$

$H_0:\beta_2=0$ vs $H_1:\beta_2\neq 0$

T.S.: $T=\dfrac{\hat{\beta}_2-0}{S(\hat{\beta}_2)}\sim t_{(19-2=17)}$

R.R.: Reject H_0 at $\alpha=0.05$ if $|T^*|>t_{(17)0.025}=2.11$

$\because |T^*|=\left|\dfrac{0.0605-0}{0.0323}\right|=1.873$ $\therefore$ don't reject H_0

我們沒有足夠證據去推論就業者之教育程度結構-大專及以上對於粗出生率具有統計顯著之意義

(六)

$\underline{SSE_p}$

$SSE_p=\sum_{i=1}^{19}(Y_i-\hat{Y}_i)^2$　選取準則：越小越好

$\underline{R_p^2}$

$R_p^2=\dfrac{SSR}{SST}=1-\dfrac{SSE}{SST}$　選取準則：越大越好

$\underline{R_{a,p}^2}$

$R_{a,p}^2=1-\dfrac{SSE/(n-p)}{SST/(n-1)}$　選取準則：越大越好

C_p

$$C_p=\frac{SSE}{MSE(X_1,X_2,\cdots,X_{p-1})}-(n-2p)$$ 選取準則：C_p 越接近 p 越好

(七)

在 $p=2$ 下，模型C為最適模型

在 $p=3$ 下，模型E為最適模型

在 $p=4$ 下，模型K為最適模型

(八)

模型K之均方誤為 $MSE=\frac{SSE}{n-p}=\frac{22.277}{19-4}=1.4851$

(九)

模型K之均方誤為 $MSE=\frac{SSE}{n-p}=\frac{22.277}{19-4}=1.4851$

(十)

1.線性之迴歸模型有五個基本假設:

(1) ε 之隨機性 $Cov(\varepsilon_i,\varepsilon_j)=0$，$i\neq j$

(2) $E(\varepsilon_i)=0$

(3) $V(\varepsilon_i)=\sigma^2$

(4) $\varepsilon_i\sim N$

(5)模型之正確性

2.殘差分析圖形：

(1)利用殘差與觀察順序繪製散布圖，在符合 ε 間不相關假設下，圖形應該呈現上下隨機跳動

(2)利用殘差與應變數估計值($\hat{Y}_i$)繪製散布圖，在符合變異數同質性假設下，圖形應該在某個範圍內隨機分布

(3)繪製殘差之常態機率圖，在符合 ε 服從常態分配假設下，圖形應該接近直線周圍