《迴歸分析》

試題評析	今年考題屬於綜合觀念題型，題目有點瑣碎，作答時間內想要拿到80分以上必須很有耐心循序漸進的解完。
考點命中	第一題：《高點迴歸分析總複習講義第一回》，秦大成編撰，頁59《例題4》相似。 第二題：《高點迴歸分析講義第二回》，秦大成編撰，頁92-93。 第三題： (一)《高點迴歸分析講義第三回》，秦大成編撰，頁8《例題4》。 (二)《高點迴歸分析講義第二回》，秦大成編撰，頁43《例題1》。 (三)《高點迴歸分析講義第三回》，秦大成編撰，頁12。 (四)(六)(七)《高點迴歸分析講義第三回》，秦大成編撰，頁93《例題4》。 (五)《高點迴歸分析講義第二回》，秦大成編撰，頁64《例題4》。 (八)《高點迴歸分析講義第二回》，秦大成編撰，頁43《例題1》。 (九)《高點迴歸分析講義第二回》，秦大成編撰，頁59。 (十)《高點迴歸分析講義第二回》，秦大成編撰，頁94《例題1》。 (十一)《高點迴歸分析講義第二回》，秦大成編撰，頁59。

一、考慮簡單線性迴歸模型如下：

$$E(Y|X=x)=\beta_0+\beta_1 x \qquad (1)$$

若解釋變數X的值替代為Z=aX+b，a≠0且b為常數，則模型（1）改寫為：

$$E(Y|Z=z)=\gamma_0+\gamma_1 z \qquad (2)$$

(一)請比較β_0與γ_0、β_1與γ_1的關係。（10分）

(二)請問模型(1)與模型(2)的判定係數是否改變？（回答是或否即可）（2分）

答：

(一) $E(Y|Z=z)=r_0+r_1 z=r_0+r_1(ax+b)=(r_0+r_1 b)+(r_1 a)x$

$\therefore \beta_0=r_0+r_1 b \ , \ \beta_1=r_1 a$

(二) 否

二、若反應變數為Y，解釋變數為X_j，j=1,2,..,p，及n個觀測值。考慮線性迴歸模型如下：

$$Y_i=\beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}+\ldots+\beta_p X_{ip}+\varepsilon_i，i=1,2,\ldots,n \qquad (3)$$

其中ε_i為均數是0，變異數是σ^2的隨機誤差項。若將模型(3)以向量及矩陣方式表達如下：

$$Y=X\beta+\varepsilon \qquad (4)$$

(一)請分別定義Y、X、β及ε之向量及矩陣之表達式，並標示其行與列的大小。（8分）

(二)試求模型(4)中，β的最小平方估計式。（10分）

(三)證明題(二)所得的最小平方估計式為不偏的。（5分）

(四)若欲求得β的最大概似估計式，需對誤差ε有如何的假設？（2分）

答：

(一) $Y=\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{n\times 1}$, $X=\begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}_{n\times(p+1)}$, $\beta=\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}_{(p+1)\times 1}$,

$$\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n\times 1}$$

(二) 令 $Q(\hat{\beta}) = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = (Y-\hat{Y})^T(Y-\hat{Y}) = (Y - X\hat{\beta})^T(Y - X\hat{\beta})$

$= (Y^T - \hat{\beta}^T X^T)(Y - X\hat{\beta}) = Y^TY - Y^TX\hat{\beta} - \hat{\beta}^TX^TY + \hat{\beta}^TX^TX\hat{\beta}$

where $(Y^TX\hat{\beta})_{1\times 1} = (\hat{\beta}^TX^TY)^T_{1\times 1}$ ：純量

$= Y^TY - 2\hat{\beta}^TX^TY + \hat{\beta}^T(X^TX)\hat{\beta}$

$\Rightarrow \frac{dQ}{d\hat{\beta}} = -2X^TY + 2(X^TX)\hat{\beta} = 0 \Rightarrow$ OLSE $\hat{\beta} = (X^TX)^{-1}X^TY$

(三) $E\hat{\beta} = E[(X^TX)^{-1}X^TY] = (X^TX)^{-1}X^TEY$

$= (X^TX)^{-1}X^T \cdot (X\beta) = \beta$ (unbiased)

(四) $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n \overset{iid}{\sim} N(0, \sigma^2)$

三、表一為民國101年縣市有關教育的資料（最後兩列分別為各變數值的加總與平方後之加總），圖一為其對應之兩兩變數散布圖矩陣（Scatter matrix），表二為這些變數之變異共變異矩陣（Variance-covariance matrix）。若考慮X_3及X_4放入模型中，表三為其估計結果。表四為僅考慮X_4放入模型中的估計結果。

以題二中迴歸模型(4)的表達方式，表五為僅考慮X_3在模型中$(X^TX)^{-1}$與(X^TY)的結果（上標T代表矩陣的轉置）。

請回答下列問題：

(一)在Y與X_1的散布圖中，可看到一個明顯的離群值（Outlier），請說明為那一個縣市？（2分）

(二)請計算所有變數之兩兩變數間的相關係數矩陣（Correlation matrix）。（10分）

(三)若將題(一)中所發現的離群值排除後，再計算Y與X_1的相關係數。另外，若將該離群值排除，已知不會影響X_3及X_4的相關係數。請建議後續統計分析（包含迴歸分析）該如何處理此一離群值。（6分）

(四)請說明表三中三個「*t statistic*」的意義，及其值與所對應之*p value*所代表之結論。（6分）

(五)請說明表三中「Residual standard error」的意義。（5分）

(六)請說明表三中「F-statistic」的意義，及其值與所對應之*p value*所代表之結論。（5分）

(七)請比較題(四)及題(六)的結論是否一致？無論一致與否，皆請說明為何能有這樣的結果。（5分）

(八)表三與表四中所得到X_4的迴歸係數估計皆為正的，是否可說明「國中生視力不良率愈高，大專以上學歷所占比例愈高；高視力不良率可提升國民的教育程度，因此視力不良率很高不是一件不好的事。」請評論引號中的話。（5分）

(九)請說明為何表四的「Multiple R-squared」比表三的值小，但表四的「Adjusted R-squared」卻比表三的值大。（5分）

(十)僅考慮X_3在模型中的簡單線性迴歸模型，請計算其截距與斜率的估計值。（6分）

(十一)若考慮下列三個模型：

$Y=\beta_{01}+\beta_{31}X_3+\beta_{41}X_4+\varepsilon$

$Y=\beta_{02}+\beta_{32}X_3+\varepsilon$

$Y=\beta_{03}+\beta_{43}X_4+\varepsilon$

那一個模型為最適模型？請寫出理由及所根據的準則。（8分）

表一

	15歲以上民間人口之教育程度結構－大專及以上（Y，%）	平均每一教師教導學生數－國小（X_1）	平均每一教師教導學生數－國中（X_2）	視力不良率－國小（X_3，%）	視力不良率－國中（X_4，%）
新北市	37.9	15.0	14.4	54.6	77.6
臺北市	63.3	12.8	12.7	52.4	78.1
臺中市	39.0	16.3	14.0	53.4	78.1
臺南市	34.9	15.9	14.6	49.6	75.9
高雄市	37.6	16.0	14.0	51.2	74.1
宜蘭縣	29.2	13.6	13.4	43.3	68.2
桃園縣	35.1	17.1	14.3	49.5	74.1
新竹縣	34.8	15.2	12.8	47.0	70.1
苗栗縣	27.1	12.9	12.1	43.9	67.2
彰化縣	28.1	16.0	14.5	52.7	79.6
南投縣	27.7	11.8	13.1	41.4	66.8
雲林縣	24.7	13.5	13.8	44.9	65.8
嘉義縣	22.9	12.1	12.8	41.5	67.4
屏東縣	27.0	13.5	14.6	38.9	61.2
臺東縣	19.4	9.5	11.8	31.1	55.1
花蓮縣	29.4	10.9	12.3	36.6	59.5
基隆市	35.1	14.8	12.5	53.0	74.4
新竹市	46.1	16.7	13.2	49.7	73.3
嘉義市	49.8	17.5	14.7	54.0	78.7
總和	**649.1**	**271.1**	**255.6**	**888.7**	**1345.2**
平方和	**24124.55**	**3957.55**	**3454.52**	**42366.01**	**96138.94**

表二

	Y	X_1	X_2	X_3	X_4
Y	108.29	11.24	1.91	47.15	49.93
X_1	11.24	4.97	1.50	12.37	12.70
X_2	1.91	1.50	0.89	3.36	3.65
X_3	47.15	12.37	3.36	44.35	45.79
X_4	49.93	12.70	3.65	45.79	49.93

表三

	Estimate	Std Err	t statistic	p value
Intercept	-26.12	29.4966	-0.886	0.389
X_3	0.58	1.2380	0.468	0.646
X_4	0.47	1.1667	0.402	0.693

Residual standard error: 8.048 on 16 degrees of freedom

Multiple R-squared: 0.4683, Adjusted R-squared: 0.4018

F-statistic: 7.046 on 2 and 16DF, p-value: 0.006383

表四

	Estimate	Std Err	t statistic	p value
Intercept	-36.63	18.6525	-1.964	0.06611
X_4	1.00	0.2622	3.813	0.00139

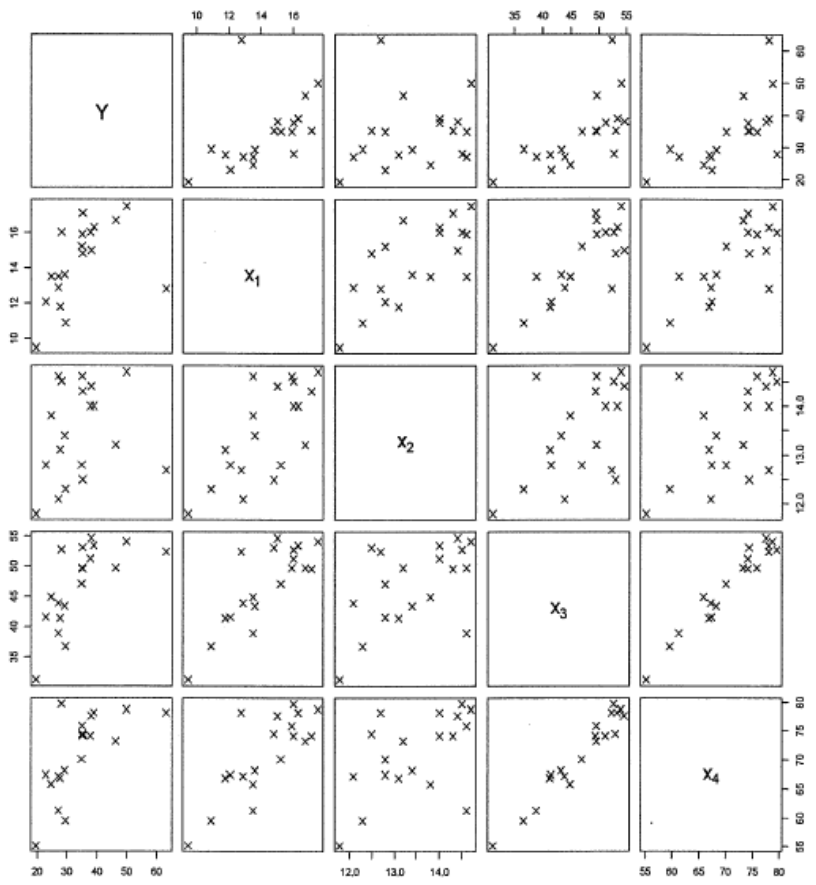
Residual standard error; 7.861 on 17 degrees of freedom
Multiple R-squared: 0.461, Adjusted R-squared: 0.4293
F-statistic: 14.54 on 1 and 17DF, p-value: 0.00139

表五

$(X^T X)^{-1}$

	Intercept	X_3
Intercept	2.7933940	-0.0585962
X_3	-0.0585962	0.0012528

$(X^T Y)^T = (649.1 \ 31209.5)$



圖一

答：

(一)台北市

(二) $r_{Y1}=\frac{11.24}{\sqrt{108.29\times4.97}}=0.4845$ ，$r_{Y2}=\frac{1.91}{\sqrt{108.29\times0.89}}=0.1946$

$r_{Y3}=\frac{47.15}{\sqrt{108.29\times44.85}}=0.6804$ ，$r_{Y2}=\frac{49.93}{\sqrt{108.29\times49.93}}=0.6790$

$r_{12}=\frac{1.6}{\sqrt{4.97\times0.89}}=0.7132$ ，$r_{13}=\frac{12.37}{\sqrt{4.97\times44.85}}=0.8332$，

$r_{14}=\frac{12.70}{\sqrt{4.97\times49.93}}=0.8062$

$r_{23}=\frac{3.86}{\sqrt{0.89\times44.85}}=0.5348$ ，$r_{24}=\frac{3.66}{\sqrt{0.89\times49.93}}=0.5475$

$r_{34}=\frac{45.79}{\sqrt{44.85\times49.93}}=0.9731$

$$\text{相關係數矩陣} = \begin{bmatrix} 1 & 0.4845 & 0.1946 & 0.6804 & 0.6790 \\ 0.4845 & 1 & 0.7132 & 0.8332 & 0.8062 \\ 0.1946 & 0.7132 & 1 & 0.5348 & 0.5475 \\ 0.6804 & 0.8332 & 0.5348 & 1 & 0.9731 \\ 0.6790 & 0.8062 & 0.5475 & 0.9731 & 1 \end{bmatrix}$$

(三)(1) $\text{cov}(Y, X_1) = \frac{\sum_{i=1}^{n} X_{i1} Y_i - n\bar{X}_1 \bar{Y}}{n-1}$

$n = 19$: $\bar{X}_1 = \frac{271.1}{19} = 14.2684$, $\bar{Y} = \frac{649.1}{19} = 34.1632$

$$\sum_{i=1}^{19} X_{i1} Y_i = (19-1) \times 11.24 + (19 \times 14.2684 \times 34.1632)$$
$$= 9463.9500$$

$n = 18$ (**排除離群值**)

$$\bar{X}_1 = \frac{271.1-12.8}{18} = 14.35 \text{ , } \bar{Y} = \frac{649.1-63.3}{18} = 32.5444$$

$$\sum_{i=1}^{n} X_{i1}^2 = 3957.55 - 12.8^2 = 3793.71$$

$$\sum_{i=1}^{n} Y_i^2 = 24124.55 - 63.3^2 = 20117.66$$

$$S_{X_1}^2 = \frac{3793.71 - 18 \times 14.35^2}{18-1} = 5.1238 \text{ ,}$$

$$S_Y^2 = \frac{20117.66 - 18 \times 32.5444^2}{18-1} = 61.9791$$

$$\sum_{i=1}^{18} X_{i1} Y_i = 9463.95 - 63.3 \times 12.8 = 8653.71$$

$$\text{cov}(Y, X_1) = \frac{8653.71 - 18 \times 14.35 \times 32.5444}{18-1} = 14.5583$$

$$\hat{\rho}(Y, X_1) = \frac{\text{cov}(Y, X_1)}{S_{X_1} S_Y} = \frac{14.5583}{\sqrt{5.1238 \times 61.9791}} = 0.8169$$

(2)迴歸分析採成對抽樣，如果排除離群值，樣本只剩18對，後續統計分析；包括OLSE、迴歸係數檢定、相關係數檢定、….等等將隨之改變。若不排除離群值，可使用穩健迴歸來降低離群值的影響，或者用OLSE，讓離群值得權重很小得以降低離群值的影響。

(四) (1)

$|t_i|$, $i = 0, 1, 2$ **值越大** ，**則越有可能** reject $H_0 : \beta_i = 0$, $i = 0, 1, 2$

(2) Intercept : $p - value > \alpha$ ∴ **截距項** β_0 **不顯著**

X_3: $p - value > \alpha$ ∴ X_3 **不值得引進**

X_4: $p - value > \alpha$ ∴ X_4 **不值得引進**

(五) Residual standard error $\sqrt{MSE(X_3 X_4)} = 0.8048$

(六) (1) $F = 7.046 \begin{cases} \in C \Rightarrow \text{reject } H_0 : \beta_3 = \beta_4 = 0 \\ \notin C \Rightarrow \text{Do not reject } H_0 : \beta_3 = \beta_4 = 0 \end{cases}$

(2) $p - value = 0.006388 < \alpha$

$\therefore$ reject H_0，**有充分證據顯示**β_3，β_4**不全爲** 0

(七) (1) 第(四)小題檢定結果：迴歸係數全爲0，即 $\beta_3 = \beta_4 = 0$

第(六)小題檢定結果：迴歸係數不全爲0，即 $\beta_3 \neq 0$ **或** $\beta_4 \neq 0$

∴ 檢定結果不一致

(2)∵ X_3，X_4**存在高度共線性**

(八) 不可下此結論，原因如下：

(1)對表三而言：X_3，X_4**存在高度共線性**，將會嚴重影響OLSE的精確度($\text{Var}(\hat{\beta}_i)$**膨脹**)，無法用國中生視力不良率來預測大專以上所占人口比例。

(2)對表四而言：$r_{Y4} = \sqrt{0.461} = 0.68 > 0$ 表示 ”國中生視力不良率” 與 “大專以上所占人口比例” 具有中度正相關性，但因爲迴歸模型並未包括其他重要自變數，所以不能說高視力不良率可提升國民教育程度。

(九) 引進越多個自變數，則判定係數越大，所以表三的R^2會較大，但表三的調整判定係數卻較小，表示當X_4固定時，不值得引進X_3。

(十) $\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_3 \end{bmatrix} = (X^TX)^{-1}X^TY = \begin{bmatrix} 2.793394 & -0.0585962 \\ -0.0585962 & 0.0012528 \end{bmatrix} \begin{bmatrix} 649.1 \\ 31209.5 \end{bmatrix}$

$= \begin{bmatrix} -15.566 \\ 1.0645 \end{bmatrix}$

(十一) **模型** $Y = \beta_{01} + \beta_{31}X_3 + \beta_{41}X_4 + \varepsilon$ ，$R_a^2 = 0.4018$

模型 $Y = \beta_{02} + \beta_{32}X_3 + \varepsilon$ ， $R^2 = 0.6804^2 = 0.4689$

$$\Rightarrow R_a^2 = 1 - \frac{n-1}{n-k}(1 - R^2) = 0.43135$$

模型 $Y = \beta_{03}3 + \beta_{43}X_4 + \varepsilon$ ，$R_a^2 = 0.4293$

∵ **模型** $Y = \beta_{02} + \beta_{32}X_3 + \varepsilon$ ，R_a^2 **最大，∴ 爲最適當模型**