《迴歸分析》

試題評析	今年考卷難度屬於中上，除了第二與第四大題屬於年年必考的計算題型外，其他題目都考的很活，今年如：離群值、影響值、維度災難與 Gauss-Markov 定理都入題了，雖然這些主題都有在課堂中介紹過，但要以實務角度進行分析，對於沒有實務分析經驗的人來說作答時可能會無從下筆。 今年考卷平均分數 50 分，想考取的話應該要有 70 分以上。由於今年考卷有不少申論題，故能得到 90 分以上應該只有少數考生而已。
考點命中	一、《迴歸分析熱門題庫》，高點文化出版，趙治勳編著，頁 2-29。 二、《迴歸分析熱門題庫》，高點文化出版，趙治勳編著，頁 2-43。 三、《迴歸分析熱門題庫》，高點文化出版，趙治勳編著，頁 2-48。 四、《迴歸分析熱門題庫》，高點文化出版，趙治勳編著，A-47 例 5，A-50 例 2。 五、《迴歸分析熱門題庫》，高點文化出版，趙治勳編著，頁 2-13。

參考之查表值：F 分佈 $\alpha=0.05$，臨界值$F_{0.05}(df1,df2)$，$t_{0.05}(28)=1.701$，$t_{0.025}(28)=2.048$。

	df1		
		1	2
df2	28	4.196	3.340
	29	4.183	3.328
	50	4.034	3.183
	52	4.027	3.175

一、請回答下列問題：

(一)圖 1 是探討美國在游泳池溺斃（Swimming-pool drownings）的人數和美國核能發電廠發電（Nuclear power plants）數量數之間的關係，這兩個變數的相關係數為 90.12%。請試述以簡單線性迴歸分析是否具有因果關係或意義？請說明理由。（5 分）

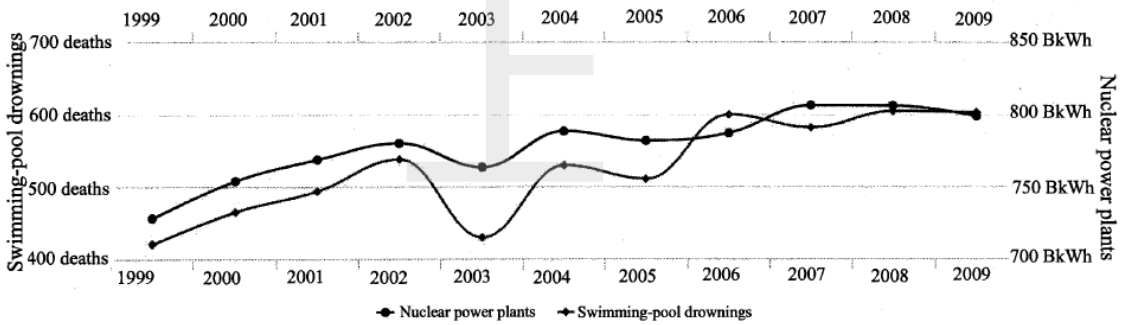

圖 1

(二)一位數據分析師擬研究滷肉飯銷售量受到那些因素所影響。所蒐集的可能解釋變數有價格、店內坪數、客流量、附近店家數、店內位置數、營業時間、店齡、配菜種類、選取肉的部位、米的種類等十個可能的解釋變數。該分析師計畫作複迴歸分析，要選擇重要解釋變數來描述反應變數（滷肉飯銷售量），請試述四種選擇重要變數的方法。又大數據的時代來臨，我們應用迴歸分析，有時會遇到高維度解釋變數的情況，解釋變數的個數（p）大到超過於樣本數（n）的情況，在高維度的解釋變數情況，請試述上述四種選擇重要變數之方法是否仍適用？如果你的答案為不適用，請說明理由。（10 分）

答：

(一)雖然溺斃人數與核能發電數量之相關係數為 90.12%，呈現高度正線性相關，然而未能表示有強烈之因果關係存在，因為以溺斃人數作為應變數及核能發電數量作為自變數下之判定係數 R^2_{YX} ，與以核能發電數量作為應變數及溺斃人數作為自變數下之判定係數 R^2_{XY} 是相同的，因此迴歸分析是不能驗證因果關係，因果關係是研究者於設立模型時之先驗假設。

又影響溺斃人數之因素可能不只核能發電數量，或許存在其他顯著影響溺斃人數之因素沒有被研究者考慮進來模型中，單以核能發電數量預測游泳池溺斃人數，除非研究者經過詳細的文獻探討，否則也不應該就此下定結論。

(二)選擇重要變數之方法：

1.向前/向後逐步迴歸

2.主成分分析法 PCA

3.線性判別分析 LDA

4.核主成分分析法 Kernal PCA

當發生「維度災難」時，以上四種方法都不適用以選擇重要變數，因為模型容易產生「高度擬合 overfitting」之問題，即模型過度配適資料，使得當有新資料進行預測時反而模型之預測能力不佳。又樣本數少於待估計參數個數時，就沒有足夠的自由度以估計模型中之所有參數，使得估計工作無法進行。

二、一位分析師隨機抽取 55 位大學生並蒐集到五個變數。該分析師希望研究身高（Y，英吋）與受測者左前臂長度（X_1，公分）、左腳長度（X_2，公分）、頭圍（X_3，公分）和鼻長（X_4，公分）之間的關係。該分析師考慮配適下列三個迴歸模型：

模型 1：$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \varepsilon_i$

模型 2：$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$

模型 3：$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$

請使用表 1 和表 2 中部分 R 統計軟體輸出之變異數分析表（ANOVA，Analysis of Variance）報表來回答以下問題：（每小題 10 分，共 30 分）

表 1　模型 1 ANOVA 表

Response：Y	DF	Sum of squares	Mean square	F value
X_1	1	590.21	590.21	123.8106
$X_2 \mid X_1$	1	224.35	224.35	47.0621
$X_3 \mid X_1, X_2$	1	1.4	1.4	0.294
$X_4 \mid X_1, X_2, X_3$	1	0.43	0.43	0.0896
Error	50	238.35	4.77	

表 2　模型 2 ANOVA 表

Response：Y	DF	Sum of squares	Mean square	F value
X_1	1	590.21	590.21	127.782
$X_2 \mid X_1$	1	224.35	224.35	48.572
Error	52	240.18	4.62	

(一)假設該分析師採用模型 1。在顯著水準 $\alpha = 0.05$ 之下，請檢定 X_3 和 X_4 兩個解釋變數是否

可以從給定模型 1 中刪除。也就是用 $\alpha=0.05$ 檢定H_0：$\beta_3=\beta_4=0$，並試述對立假設，檢定統計量之值、決策法則和結論。並請計算偏相關係數$R^2_{Y,X_3,X_4|X_1,X_2}$（partial R^2）。

(二)假設該分析師採用模型 2。也就是在模型中僅考慮了兩個解釋變數，這兩個解釋變數是學生的左前臂長度（X_1）和左腳長度（X_2）。該分析師想知道這兩個解釋變數是否與身高（Y）有線性關係。在顯著水準 $\alpha=0.05$ 之下，請檢定H_0：$\beta_1=\beta_2=0$。並請試述檢定統計量之值、決策法則和結論。另請計算模型 2 的調整的複判定係數R^2（adj R^2，the adjusted R-squared）並試述其意義。又該分析師要把身高的單位英吋轉公分（英吋乘以 2.54），試述模型 2 的adj R^2是否改變？

(三)假設該分析師採用模型 3。只考慮模型中具有一個解釋變數，為學生的左前臂長度（X_1）。在顯著水準 $\alpha=0.05$ 下，該分析師想知道一個額外的解釋變數 X_2 是否在解釋身高上具有顯著的貢獻。也就是說，該分析師想知道 X_2 對模型 3 的貢獻。請協助回答此問題並說明對立假設、檢定統計量之值、決策法則和結論。在表 1 和表 2 的F檢定中，請試述需要做何假設，才能執行這些F檢定。

答：

(一) $H_0:\beta_3=\beta_4=0$ vs $H_1:$至少一個$\beta_j\neq 0, j=3,4$

T.S.: $F=\dfrac{SSR(X_3,X_4\mid X_1,X_2)/2}{SSE(X_1,X_2,X_3,X_4)/(55-4-1)}\sim F_{(2,50)}$

R.R.: Reject H_0 at $\alpha=0.05$ if $F^*>F_{(2,50)0.05}=3.183$

$\because SSR(X_3,X_4\mid X_1,X_2)=SSR(X_3\mid X_1,X_2)+SSR(X_4\mid X_1,X_2,X_3)$

$=1.4+0.43=1.83$

$$F^*=\frac{SSR(X_3,X_4\mid X_1,X_2)/2}{SSE(X_1,X_2,X_3,X_4)/(55-4-1)}=\frac{1.83/2}{238.35/50}=0.1919$$

$\therefore$ don't reject H_0

我們沒有足夠證據去推論 X_3,X_4 對應變數 Y 具有顯著的影響。

$$R^2_{YX_3X_4|X_1X_2}=\frac{SSR(X_3,X_4\mid X_1,X_2)}{SSE(X_1,X_2)}=\frac{1.83}{240.18}=0.007619$$

(二) $H_0:\beta_1=\beta_2=0$ vs $H_1:$至少一個$\beta_j\neq 0, j=1,2$

T.S.: $F=\dfrac{SSR(X_1,X_2)/2}{SSE(X_1,X_2)/(55-2-1)}\sim F_{(2,52)}$

R.R.: Reject H_0 at $\alpha=0.05$ if $F^*>F_{(2,52)0.05}=3.175$

$$\because F^*=\frac{SSR(X_1,X_2)/2}{SSE(X_1,X_2)/(55-2-1)}=\frac{[SSR(X_1)+SSR(X_2\mid X_1)]/2}{SSE(X_1,X_2)/52}$$

$$=\frac{[590.21+224.35]/2}{240.18/52}=88.1779$$

∴ reject H_0

我們有足夠證據去推論 X_1, X_2 對應變數 Y 具有顯著的影響。

$$adjR^2 = 1 - \frac{SSE(X_1, X_2)/55-2-1}{SST(X_1, X_2)/55-1} = 1 - \frac{240.18/52}{(590.21+224.35+240.18)/55-1} = 0.7635$$

$$R^2 = 1 - \frac{SSE(X_1, X_2)}{SST(X_1, X_2)} = 1 - \frac{240.18}{590.21+224.35+240.18} = 0.7723$$

由 $adjR^2$ 與 R^2 之結果相差不遠，得知模型之解釋能力為 77.23%，且不是因為自變數個數之影響而有中高度之解釋能力。

應變數 Y 之衡量單位改變，不影響模型之解釋能力，故 $adjR^2$ 不會改變。

(三) $H_0 : \beta_2 = 0$ vs $H_1 : \beta_2 \neq 0$

T.S.: $F = \frac{SSR(X_2 \mid X_1)/1}{SSE(X_1, X_2)/55-2-1} \sim F_{(1,52)}$

R.R.: Reject H_0 at $\alpha = 0.05$ if $F^* > F_{(1,52)0.05} = 4.027$

$$\because F^* = \frac{SSR(X_2 \mid X_1)/1}{SSE(X_1, X_2)/55-2-1} = \frac{224.35/1}{240.18/52} = 48.5727$$

∴ reject H_0

我們有足夠證據去推論 X_2 對應變數 Y 具有顯著的影響。

假設 $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

1.隨機性且獨立性
2.平均數為零
3.變異數齊一性
4.常態分配
5.模型正確性假設

三、(一)在作迴歸分析時，經常會遇到離群值和有影響力觀察值（influential data point）的問題。請試述何謂離群值和有影響力觀察值。並請分別試述兩種判斷準則偵測迴歸分析中的離群值和有影響力觀察值。（12 分）

(二)圖 2A 是一組數據的散佈圖，圖 2B 提供兩條估計線，實線估計式 $\hat{Y}_i = 2.8 + 4.97X_i$ 包括第 51 點觀察值 $((X_{51}, Y_{51}) = (4,50))$，虛線估計式 $\hat{Y}_i = 3.68 + 4.98X_i$ 不包括第 51 點觀察值。請試述這組數據集是否包含任何離群值？並請試述這組數據是否包含任何有影響力觀察值？另請說明理由。（4 分）

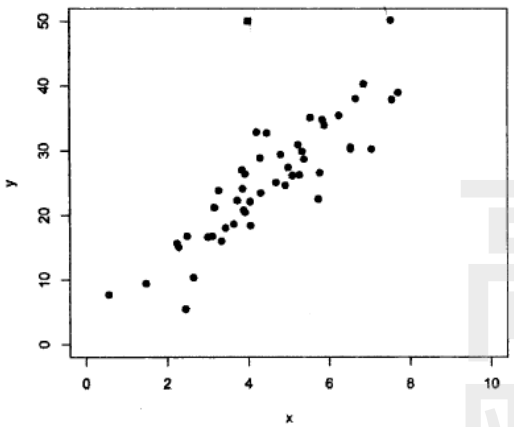

圖 2A

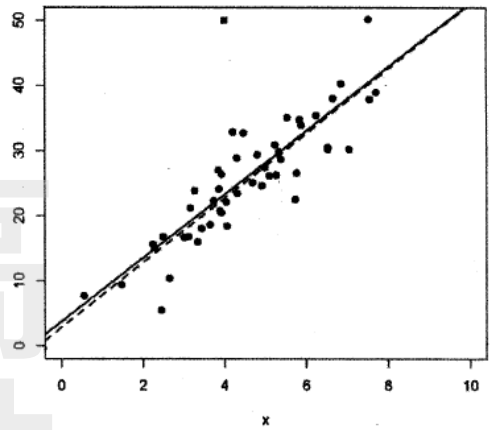

圖 2B

(三)圖 3A 是另一組數據的散佈圖，圖 3B 提供兩條估計線，實線估計式 $\hat{Y}_i = 6.95 + 4.08X_i$ 包括第 41 點觀察值 $((X_{41}, Y_{41}) = (10,16))$，虛線估計式 $\hat{Y}_i = 1.93 + 5.21X_i$ 不包括第 41 點觀察值。請試述這組數據集是否包含任何離群值？並請試述這組數據集是否包含任何有影響力觀察值？另請說明理由。（4 分）

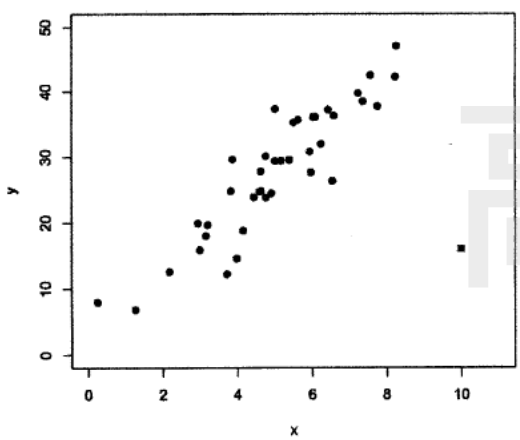

圖 3A

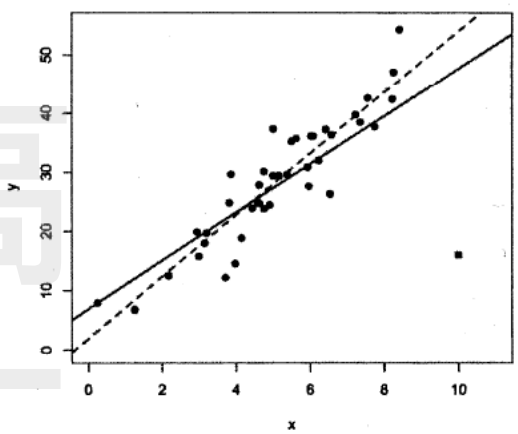

圖 3B

答：

(一)離群值：比較大部份的觀察值而言，該筆資料較大或較小。

偵測方法：

1.殘差絕對值大於 3，可視為離群值。

2.殘差對應變數估計值 $\hat{Y}$ 之散佈圖，當某資料遠離大部份資料點時，可視為離群值。

影響值：該筆資料雖然沒有比較大或小，但它對於配適模型之結果影響很嚴重。

偵測方法：

1.DFFITS：$|(DFFITS)_i| = \left|\frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)}h_{ii}}}\right| > 1$，可視為離群值。

2.DFBETAS：$|(DFBETAS)_{ki}| = \left|\frac{\hat{\beta}_k - \hat{\beta}_{k(i)}}{\sqrt{MSE_{(i)}c_{ii}}}\right| > 1$，可視為離群值。

(二) $(X_{51},Y_{51})=(4,50)$

$\hat{Y}_{51}=2.8+4.97\times4=22.68$ ， $\hat{Y}_{51(51)}=3.68+4.98\times4=23.6$

$|e_{51}|=|Y_{51}-\hat{Y}_{51}|=|50-22.68|=27.32>3$，故第 51 筆觀察值判斷為離群值。

$|\frac{\hat{Y}_{51}-\hat{Y}_{51(51)}}{\hat{Y}_{51}}|=|\frac{22.68-23.6}{22.68}|=4.056\%$ ，故第 51 筆觀察值判斷不為影響值。

註：由於題目沒有給 $MSE_{(51)}$，DFFITS 與 DFBETAS 無法求得，只利用兩個模型之改變百分比以衡量第 51 筆觀察值是否有可能為影響值。

(三) $(X_{41},Y_{41})=(10,16)$

$\hat{Y}_{41}=6.95+4.08\times10=47.75$ ， $\hat{Y}_{41(41)}=1.93\times5.21\times10=54.03$

$|e_{41}|=|Y_{41}-\hat{Y}_{41}|=|16-47.75|=31.75>3$，故第 41 筆觀察值判斷為離群值。

$|\frac{\hat{Y}_{41}-\hat{Y}_{41(41)}}{\hat{Y}_{41}}|=|\frac{47.75-54.03}{22.68}|=27.69\%$ ，故第 41 筆觀察值判斷為影響值。

註：由於題目沒有給 $MSE_{(41)}$，DFFITS 與 DFBETAS 無法求得，只利用兩個模型之改變百分比以衡量第 41 筆觀察值是否有可能為影響值。

四、一位數據分析師受冰飲企業老闆的委託，欲知道每日最高溫和該公司冰品銷售是否有線性關係，以作為未來商品促銷的依據。他蒐集了每日最高溫（X，以攝氏為單位）和冰品銷售（Y），共 30 個樣本點。下列是這些數據的統計量：

$n=30$ ， $\overline{X}=28.9892$ ， $\overline{Y}=34.7065$ ， $SXY=\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})=360.2128$

$SXY=\sum_{i=1}^{n}(X_i-\overline{X})^2=556.0186$ ， $SYY=\sum_{i=1}^{n}(Y_i-\overline{Y})^2=353.0085$

(一)在配適 $E(Y\mid X=x)=\alpha+\beta_1(x-\overline{X})$ 的簡單線性迴歸方程式下，請利用最小平方法計算參數估計值（$\hat{\alpha}$ 和 $\hat{\beta}_1$）與分別之標準誤。並請試述 $\hat{\alpha}$ 和 $\hat{\beta}_1$ 的共變異數，也就是 Cov($\hat{\alpha}$, $\hat{\beta}_1$)。（15 分）

(二)請在試卷上，完成下列變異數分析表。在顯著水準 $\alpha=0.05$，請協助檢定 H_0：$\beta_1=0$。並請試述檢定統計量之值、決策法則、結論和所需要之假設。（10 分）

Source	Sum of Squares	DF	Mean square	F value
Regression	(1)	(4)		
Error	(2)	(5)	(6)	
Total	(3)			

答：

(一) $\hat{\alpha}=\overline{Y}=34.7065$ ， $\hat{\beta}_1=\frac{SS_{XY}}{SS_X}=\frac{360.2128}{556.0186}=0.6478$

$S(\hat{\alpha})=S(\overline{Y})=\sqrt{\frac{MSE}{n}}=\sqrt{\frac{4.2742}{30}}=0.3775$

$S(\hat{\beta}_1)=\sqrt{\frac{MSE}{SS_X}}=\sqrt{\frac{4.2742}{556.0186}}=0.08768$

其中 $MSE=\frac{SSE}{n-2}=\frac{SS_Y-\hat{\beta}_1^2 SS_X}{n-2}=\frac{353.0085-(0.6478)^2(556.0186)}{30-2}$=4.2742

$$Cov(\hat{\alpha},\hat{\beta}_1)=Cov(\frac{\sum Y_i}{n},\frac{SS_{XY}}{SS_X})=\frac{1}{nSS_X}Cov(\sum Y_i,\sum(X_i-\bar{X})Y_i)$$

$$=\frac{1}{nSS_X}\sum Cov(Y_i,(X_i-\bar{X})Y_i)=\frac{\sigma^2}{nSS_X}\sum(X_i-\bar{X})=\frac{\sigma^2}{nSS_X}(0)=0$$

(二)假設模型 $Y_i=\alpha+\beta_1(X_i-\bar{X})+\varepsilon_i$，$\varepsilon_i \overset{iid}{\sim} N(0,\sigma^2)$，$i=1,2,\cdots,30$

Source	SS	DF	MS	F value
Reg	233.3309	1	233.3309	54.5905
Error	119.6776	28	4.2742	
Total	353.0085	29		

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

T.S.: $F=\frac{SSR/1}{SSE/30-2}\sim F_{(1,28)}$

R.R.: Reject H_0 at $\alpha=0.05$ if $F^*>F_{(1,28)0.05}=4.196$

$\because F^*=54.5905$ $\therefore$ reject H_0

我們有足夠證據去推論 X 對應變數 Y 具有顯著的影響。

五、一位分析師擬以 $\tilde{\beta}_1=\frac{1}{n-1}\sum_{i=2}^{n}\left[\frac{Y_i-Y_{i-1}}{X_i-X_{i-1}}\right]$ 估計簡單線性迴歸模型 $Y_i=\beta_0+\beta_1X_i+\varepsilon_i$，$i$=1,...,$n$ 之斜率 β_1。他可以證明 $\tilde{\beta}_1$ 是一個不偏估計式。請寫出 β_1 的最小平方估計式 $\hat{\beta}_1$。在無須推導 $\tilde{\beta}_1$ 的變異數下，試述相較於最小平方估計式 $\hat{\beta}_1$，$\tilde{\beta}_1$ 和 $\hat{\beta}_1$ 何者為最佳之估計式？請詳細敘述所依據的理由或定理。（10 分）

答：

由 Gauss-Markov 定理得知最小平方估計式 $\hat{\beta}_1$ 為 β_1 之 BLUE，表示 $\hat{\beta}_1$ 為所有 β_1 之線性不偏估計式中變異數最小，由於 $\tilde{\beta}_1$ 為 β_1 之線性不偏估計式，故 $\tilde{\beta}_1$ 之變異數必定大於 $\hat{\beta}_1$ 之變異數。