106 年公務人員高等考試三級考試試題

類　科：統計
科　目：迴歸分析

甲、申論題部份
參考之查表值：F分佈 α=0.05，臨界值$F_{0.05}(df1,df2)$，$t_{0.05}(28)=1.701$，$t_{0.025}(28)=2.048$。

	df1		
df2		1	2
	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：

㈠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)

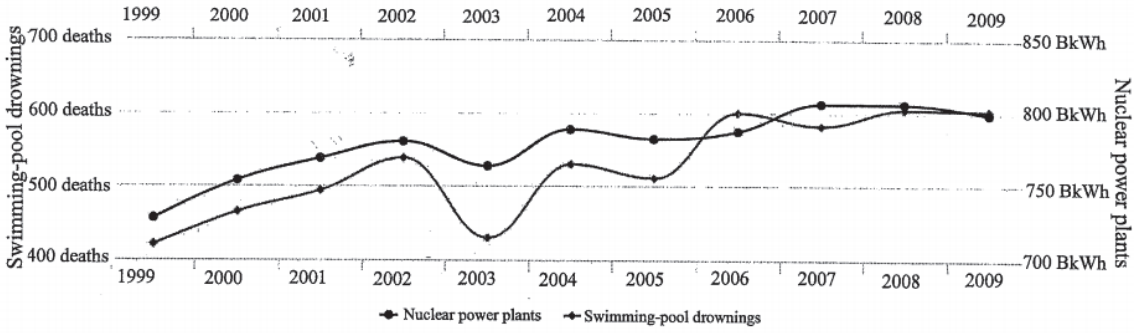

圖 1

㈡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，有時會遇到高維度的解釋變數的情況，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方法是否仍適用?如果你的答案為不適用，請說明理由。(10 分)

【擬答】：

㈠相關指的是直線關係，雖然在此兩變數之間具有高度的相關性(r=0.9012)，並且散布圖亦呈現高度的直線關聯，僅能說明兩變項間可能會有很強的直線關係，但不代表兩者之間具有因果關係。因果關係其中一個條件是必須具有生物贊同性(biological plausibility)，所以沒有理由當發電量越大，就要越多人溺水。

㈡(a) *1.* 選擇調整後判定係數 R_a^2 大者。因為當自變數個數增加時，R_a^2 不一定會變大。

2. 選擇 C_p 值越接近 P 者[C_p 值最小者]，模型越佳。

其中 $C_P=\dfrac{SSE(\text{所選模型})}{MSE(\text{全選})}-(n-2P)$

3. 選擇 PRESS 值越小者模型越佳。

先將第一筆樣本觀測值移除，並利用其他的樣本資料點作迴歸來預測第一筆樣本點，記為 $y_{(1)}^*$，並且計算此樣本點的殘差 $y_{(1)}-y_{(1)}^*$。用同樣的方式再找出其他 $n-1$ 筆資料的殘差值，並計算其平方和。此時的平方和即為所求：$PRESS=\sum(y_{(i)}-y_{(i)}^*)^2$

4.選擇 AIC 較小者模型越佳。

$AIC = n\ln SSE + n\ln n + 2p$

㈥當變數個數(p)大於樣本個數(n)時，會造成共線性問題，導致參數無法估計，無法確認個別自變數對依變數有多大影響。從數學求解的角度來說，有 p 個未知數就需要有 p 個方程式才能求解(full rank)，所以當 p 大於 n 時，稱為過度配適，無法得到唯一解。再者，從直觀的角度來說，若樣本數是極端的情況等於 1 筆資料，我可以選擇任何斜率的直線來解釋這筆樣本點，可知當 p 大於 n 時，我們無法有效地採用常見的選模方式來選出最佳的模型。

二、一位分析師隨機抽取55位大學生並蒐集到五個變數。該分析師希望研究界身高(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	

㈠假設該分析師採用模型1。在顯著水準α=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α=0.05檢定H_0：β_3=β_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)。

㈡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α=0.05下，請檢定H_0：β_1=β_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?

㈢假設分析師採用模型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α=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。

【擬答】：

㈠(a) $SSR(X_1, X_2, X_3, X_4) = 590.21 + 224.35 + 1.4 + 0.43 = 816.39$

$SSR(X_1, X_2) = 590.21 + 224.35 = 814.56$

$H_0 : \beta_3 = \beta_4 = 0$　　$H_1 : \beta_3, \beta_4$ 不全為 0

$\alpha = 0.05$

$$F^* = \frac{SSR(X_3, X_4 \mid X_1, X_2)/2}{SSE(X_1, X_2, X_3, X_4)/55-5}$$

$$= \frac{[SSR(X_1, X_2, X_3, X_4) - SSR(X_1, X_2)]/2}{4.77}$$

$$= \frac{(816.39 - 814.56)/2}{4.77} = 0.192 \notin C$$

$C : \{F^* > F_{0.05}(2,50) = 3.183\}$

不拒絕 H_0，沒有顯著證據說在有 X_1, X_2 下，需要加入 X_3 與 X_4

⑹ $R^2_{Y,X_3,X_4|X_1,X_2} = \frac{SSR(X_3, X_4 \mid X_1, X_2)}{SSE(X_1, X_2)} = \frac{1.83}{240.18} = 0.0076$

㈡⑷ $H_0 : \beta_1 = \beta_2 = 0$　　$H_1 : \beta_1, \beta_2$ 不全為 0

$\alpha = 0.05$

$$F^* = \frac{SSR(X_1, X_2)/2}{SSE(X_1, X_2)/55-3}$$

$$= \frac{(590.21 + 224.35)/2}{4.62} = 88.156 \in C$$

$C : \{F^* > F_{0.05}(2,52) = 3.175\}$

拒絕 H_0，有顯著證據說 β_1, β_2 不全為 0，

左前臂長度(X_1)與左腳長度(X_2)同時解釋解釋身高顯著。

⑹ $adj\ R^2 = 1 - \frac{n-1}{n-p}\frac{SSE}{SSTO} = 1 - \frac{55-1}{55-3} \times \frac{240.18}{814.56 + 240.18} = 0.7635$

代表以左前臂長度與左腳長度對於身高解釋力有 76.35%

⑹僅是 Y_i 單位轉換，並不會改變調整後判定係數大小。

㈢⑷ $H_0 : \beta_2 = 0$　　$H_1 : \beta_2 \neq 0$

$$F^* = \frac{SSR(X_2 \mid X_1)/1}{SSE(X_1, X_2)/55-5}$$

$$= \frac{224.35/1}{4.62} = 48.561 \in C$$ (或藉由表 2 中的 F value)

$C : \{F^* > F_{0.05}(1,52) = 4.027\}$

拒絕 H_0，有顯著證據說 $\beta_2 \neq 0$，在有 X_1 情況下，需要加入 X_2

⑹前述表中的 F 檢定中，皆須有常態分配的假設才能執行。

三、㈠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)

㈡圖2A是一組數據的散佈圖，圖2B提供兩條估計線，實現估計式 $\hat{Y}_i = 2.8 + 4.97X_i$ 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)

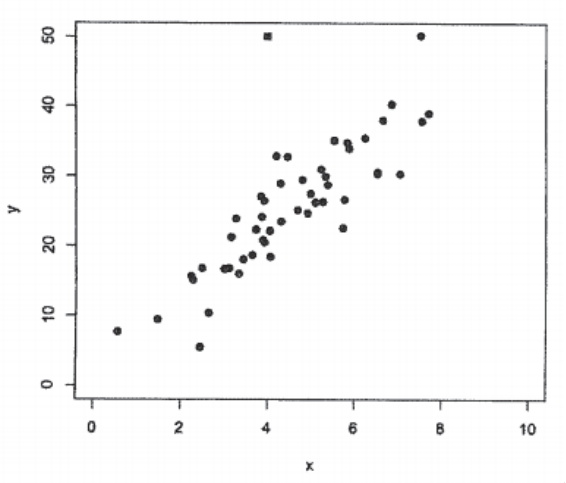

圖 2A

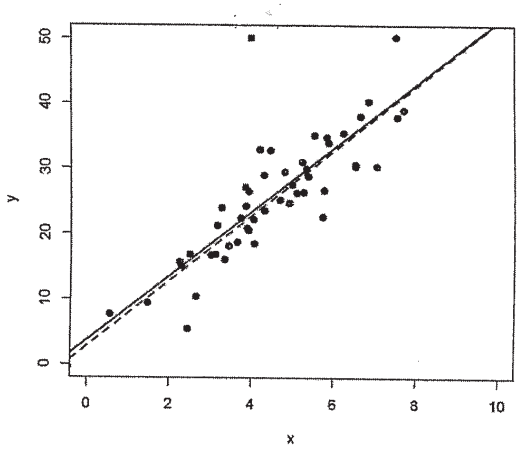

圖 2B

㈢圖3A是另一組數據的散佈圖，圖3B提供兩條估計線，實現估計式$\hat{Y}_i = 6.95 + 4.08X_i$包含第41點觀察值$\left((X_{41}, Y_{41}) = (10,16)\right)$，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)

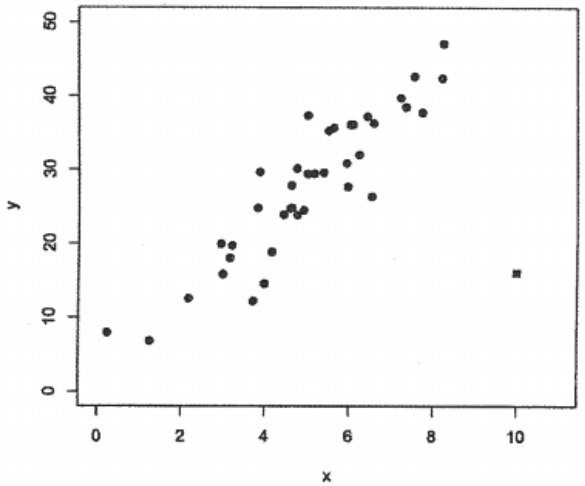

圖 3A

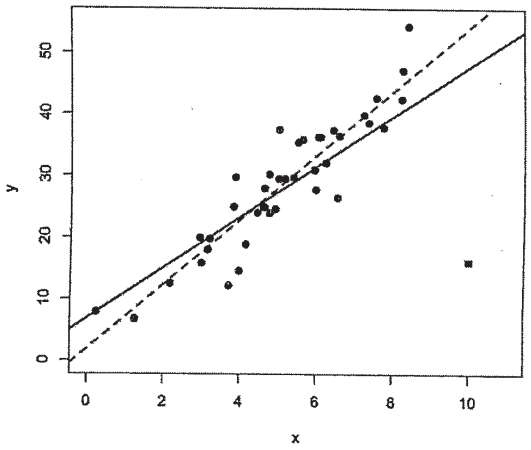

圖 3B

【擬答】：

㈠(a)離群值：某些觀察個案與其他資料間有明顯的區隔，離群值可能是對 Y 離群或是對 X 離群，亦可能同時對 X 與 Y 均離群。

可採用標準化殘差 $d_i = \frac{e_i}{\sqrt{MSE}}$ 的絕對值 $|d_i|$ 是否大於 2 來判斷；亦可利用槓桿值是否超過平均槓桿值的兩倍時，$h_{ii} > 2\frac{p}{n}$ 即被歸類為離群值。

(b)影響點：若觀察個案對於迴歸估計函數有較大的改變，即斜率受到較大的影響，此時稱為影響點。可採用 DFFITS，透過計算 $(DFFITS)_i = \frac{\hat{Y}_i - \hat{Y}_{(i)}}{\sqrt{MSE_{(i)}h_{ii}}}$ 來衡量，當 $(DFFITS)_i > 2\sqrt{p/n}$ 則該樣本可以視為具有影響力；亦可利用 DFBETAS，透過將所有的 n 個樣本配適出之迴歸係數 $\hat{\beta}_k$，與剔除第 i 個樣本後之配適迴歸係數 $\hat{\beta}_{k(i)}$ 差異作比較：$(DFBETAS)_{k(i)} = \frac{\hat{\beta}_k - \hat{\beta}_{k(i)}}{\sqrt{MSE_{(i)}(X'X)^{-1}_{kk}}}$，當 $|DFBETAS_{ik}| > 2$，判斷該樣本是否具有影響力。

㈡第 51 點觀察值為離群值，因為其對 Y 有所偏離；但此觀察值並未對整體迴歸直線斜率有太大影響，所以並非影響點。

㈢第 51 點觀察值為離群值，因為其對 X 有所偏離；且此觀察值對整體迴歸直線斜率影響頗大，所以第 51 點觀察值亦為影響點。

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：

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

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

㈠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）

㈡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)			

【擬答】：

㈠ $E(Y\mid X=x)=\hat{\beta}_0+\hat{\beta}_1x=(\overline{Y}-\hat{\beta}_1\overline{X})+\hat{\beta}_1x=\overline{Y}+\hat{\beta}_1(x-\overline{X})$

$\hat{\beta}_1=\frac{SXY}{SXX}=\frac{360.2128}{556.0186}=0.6478$ ， $\hat{\alpha}=\overline{Y}=34.7065$

$SSR=\hat{\beta}_1^2\cdot SXX=233.3614$

$SSE=SSTO-SSR=353.0085-233.3614=119.6471$

$MSE=\frac{119.6471}{30-2}=4.2731$

$S(\hat{\beta}_1)=\sqrt{Var(\frac{SXY}{SXX})}=\sqrt{\frac{MSE}{SXX}}=\sqrt{\frac{4.2731}{556.0186}}=0.0877$

$S(\hat{\alpha})=\sqrt{Var(\overline{Y})}=\sqrt{\frac{MSE}{n}}=\sqrt{\frac{4.2731}{30}}=0.3774$

$Cov(\hat{\alpha},\hat{\beta}_1)=Cov(\overline{Y},\hat{\beta}_1)=Cov(\frac{1}{n}\sum Y_i,\sum\frac{(X_i-\overline{X})Y_i}{SS_X})$

$=\frac{1}{n\cdot SS_X}\sum(X_i-\overline{X})V(Y_i)$

$=\frac{1}{n\cdot SS_X}\sum(X_i-\overline{X})\sigma^2=0$

㈡ ANOVA表

Source	$S.S$	$d.f$	$M.S$	F^*
Regression	233.3614	1	233.3614	54.6117
Error	119.6471	28	4.2731	
Total	353.0085	29		

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

$\alpha=0.05$

$C:\{F^*>F_{0.05}(1,28)=4.196\}$

$F^*=54.6117\in C$

拒絕H_0，有顯著證據說$\beta_1\neq 0$，每日最高溫顯著解釋冰品銷售。

$$\widetilde{\beta_1}=\frac{1}{n-1}\sum_{i=2}^{n}\left[\frac{Y_i-Y_{i-1}}{X_i-X_{i-1}}\right]$$

五、一位分析師擬以　　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$\beta_1$。他可以證明$\tilde{\beta}_1$是一個不偏估計式。請寫出$\beta_1$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）。

【擬答】：

根據Gauss-Markov定理：

若$E(\varepsilon_i)=0$，$\text{var}(\varepsilon_i)=\sigma^2$，且對所有的$i\neq j$，$Cov(\varepsilon_i,\varepsilon_j)=0$條件下

利用最小平方法所求得之β_1估計式$\hat{\beta}_1=\frac{SS_{XY}}{SS_X}=\frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sum(X_i-\bar{X})^2}$

具有最小變異，$\hat{\beta}_1$具有BLUE特性。所以$\hat{\beta}_1$相較$\tilde{\beta}_1$為最佳的估計式。

由$\hat{\beta}_1=\frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sum(X_i-\bar{X})^2}=\sum k_iY_i$，其中$k_i=\frac{X_i-\bar{X}}{SS_X}$

在此考慮$\tilde{\beta}_1=\sum a_iY_i$為$\beta_1$之任意線性不偏估計式

假設$a_i=k_i+c_i$，c_i為任意常數

$$E(\tilde{\beta}_1)=E\left(\sum a_iY_i\right)=\sum a_iE(Y_i)=\sum a_i(\beta_0+\beta_1x_i)$$
$$=\beta_0\sum a_i+\beta_1\sum a_ix_i$$

可得$\sum a_i=0$，$\sum a_ix_i=1$

$$Var(\tilde{\beta}_1)=Var\left(\sum a_iY_i\right)=\sum a_i^2Var(Y_i)=\sum a_i^2\sigma^2$$
$$=\sum(k_i+c_i)^2\sigma^2=\sigma^2(\sum k_i^2+\sum c_i^2+2\sum k_ic_i)$$

此處$\sum k_ic_i=\sum k_i(a_i-k_i)=\sum k_ia_i-\sum k_i^2$

$$=\frac{\sum a_i(X_i-\bar{X})}{SS_X}-\frac{1}{SS_X}$$
$$=\frac{\sum a_iX_i-\bar{X}\sum a_i}{SS_X}-\frac{1}{SS_X}=\frac{1}{SS_X}-\frac{1}{SS_X}=0$$

得$Var(\tilde{\beta}_1)=\sigma^2(\sum k_i^2+\sum c_i^2)$

所以最小值發生於$\sum c_i^2=0$，即$\tilde{\beta}_1=\sum a_iY_i=\sum k_iY_i$才會有最小變異數

$$\tilde{\beta}_1=\frac{1}{n-1}\sum_{i=2}^{n}\left[\frac{Y_i-Y_{i-1}}{X_i-X_{i-1}}\right]\neq\sum k_iY_i=\hat{\beta}_1$$

故$\hat{\beta}_1$相較$\tilde{\beta}_1$為最佳的估計式。