《迴歸分析》

一、考慮一多元線性迴歸模型，其反應變數為 Y，解釋變數為 $X_1, X_2,\ldots,X_k$，有 n 個觀測值，線性迴歸模型為 $Y_i=\beta_0+\beta_1X_{i1}+\beta_2X_{i2}+\ldots+\beta_kX_{ik}+\varepsilon_i$，$i=1,\ldots,n$，其中誤差項 ε_i 之期望值為0，變異數為 σ^2，且兩兩獨立，此模型以向量及矩陣方式表示為 $Y=X\beta+\varepsilon(*)$，其中

$$Y=\begin{bmatrix}Y_1\\Y_2\\\vdots\\Y_n\end{bmatrix}_{n\times1},\quad X=\begin{bmatrix}1 & X_{11}\cdots X_{1k}\\1 & X_{21}\cdots X_{2k}\\ & \vdots\ \cdots\ \vdots\\1 & X_{n1}\cdots X_{nk}\end{bmatrix}_{n\times(k+1)},\quad \beta=\begin{bmatrix}\beta_0\\\beta_1\\\vdots\\\beta_k\end{bmatrix}_{(k+1)\times1},\quad \varepsilon=\begin{bmatrix}\varepsilon_1\\\varepsilon_2\\\vdots\\\varepsilon_n\end{bmatrix}_{n\times1}$$

請回答下列問題：（每小題5分，共30分）

(一) 以向量及矩陣方式，試求出參數向量 β 之最小平方估計量向量 b。

(二) 承題(一)，令 A 為一個 $2\times(k+1)$ 的矩陣，求 Ab 之變異數－共變異數矩陣。

(三) 配適值向量表為 $\hat{Y}=HY$，寫出矩陣 H。

(四) 求出殘差向量 $e=Y-\hat{Y}$ 之變異數－共變異數矩陣。

(五) 令 A 為對稱矩陣，則 $Y'AY$ 稱為 Y 之二次式，將此模型之SSE（error sum of square）$=e'e$ 表成二次式，其中 Y' 和 e' 分別是 Y 和 e 之轉置矩陣。

(六) 求出 β 之最大概似估計量，對誤差項向量需要什麼假設。

試題評析	本題是要考驗考生是否具有矩陣法迴歸分析之能力，但沒有過多之計算，只要熟讀課本內容就可以獲得高分。
考點命中	《迴歸分析熱門題庫》，高點文化出版，趙治勳編著，第二章。

答：

(一) $\min \underset{\sim}{e}^T \underset{\sim}{e}=\min(Y-X\underset{\sim}{\hat{\beta}})^T(Y-X\underset{\sim}{\hat{\beta}})$

令 $$\frac{d\,\underset{\sim}{e}^T\underset{\sim}{e}}{d\,\underset{\sim}{\hat{\beta}}}=\frac{d}{d\,\underset{\sim}{\hat{\beta}}}(Y^TY-Y^TX\underset{\sim}{\hat{\beta}}-\underset{\sim}{\hat{\beta}}^TX^TY+\underset{\sim}{\hat{\beta}}^TX^TX\underset{\sim}{\hat{\beta}})$$

$$=-X^TY-X^TY+2X^TX\underset{\sim}{\hat{\beta}}=0$$

$$\Rightarrow X^TX\underset{\sim}{\hat{\beta}}=X^TY$$

$$\Rightarrow \underset{\sim}{\hat{\beta}}=b=(X^TX)^{-1}X^TY$$

(二)已知 $V(b)=(X^TX)^{-1}\sigma^2$

$V(Ab)=AV(b)A^T=A(X^TX)^{-1}A^T\sigma^2$

(三) $H=X(X^TX)^{-1}X^T$

(四) $e=Y-\hat{Y}=Y-HY=(I-H)Y$

$V(e)=(I-H)\sigma^2$

(五) $SSE=e^Te=[(I-H)Y]^T[(I-H)Y]=Y^T(I-H)Y$ 表成二次式

(六)除了題意中之假設外，仍然需要有 n 元常態分配之假設

二、某無人車研發公司欲預測它的行車時間 Y，考慮了三個高度相關的解釋變數分別是行駛里數 X_1，車種 X_2，車輪 X_3，收集過去20個月資料得到

$SSR(X_1, X_2, X_3) = 4000$，

$SSR(X_1) = 1000$，

$SSR(X_1|X_2) = 600$，

$SSE(X_1, X_2, X_3) = 800$，

請回答下列問題：（每小題5分，共15分）

(一) 求偏判定係數 $r^2_{YX_3.X_1X_2}$ 。

(二) 檢定偏相關係數 $\rho_{YX_3.X_1X_2}$ 是否為0，請求出 F 檢定統計量的值。

(三) 求偏判定係數 $r^2_{YX_1X_2.X_3}$（以最簡分數表示）。

試題評析	本題型於以往考古題中都常出現，不過由於資訊不足無法計算，考生只需要列出算式即可。
考點命中	《迴歸分析熱門題庫》，高點文化出版，趙治勳編著，第三章。

答：

題目有誤，$SSR(X_1 | X_2) = 600$ 應該改為 $SSR(X_2 | X_1) = 600$ 才能夠回答(二)，且因為顯見未給 $SSR(X_3)$ 故無法計算(三)

(一) $r^2_{YX_3\cdot X_1X_2} = \dfrac{SSR(X_3 | X_1, X_2)}{SSE(X_1, X_2)}$

其中 $SSR(X_3 | X_1, X_2) = SSR(X_1, X_2, X_3) - SSR(X_1, X_2)$

$= SSR(X_1, X_2, X_3) - [SSR(X_1) + SSR(X_2 | X_1)]$

$= 4000 - [1000 + 600] = 2400$

(二) $H_0 : \rho_{YX_3\cdot X_1X_2} = 0$ vs $H_1 : \rho_{YX_3\cdot X_1X_2} \neq 0$

T.S.: $F = \dfrac{SSR(X_3 | X_1, X_2)/1}{SSE(X_1, X_2, X_3)/20-3-1} = \dfrac{2400/1}{800/20-3-1} = 48$

(三) $r^2_{YX_1X_2\cdot X_3} = \dfrac{SSR(X_1, X_2 | X_3)}{SSE(X_3)} = \dfrac{SSR(X_1, X_2, X_3) - SSR(X_3)}{SSE(X_3)}$

$= \dfrac{SSR(X_1, X_2, X_3) - SSR(X_3)}{SSE(X_3)}$

三、一國內規模最大的律師事務所專門辦理職災案件，總經理想了解他們在捷運上的廣告有沒有增加他們的業務量，根據過去隨機抽取11月的資料，利用兩個解釋變數：一為單月廣告費用X_1（單位為百萬元，平均值為1單位，標準差為2單位），另一個為主要競爭對手單月廣告費用X_2（單位為百萬元，平均值為1單位，標準差為2單位）來預測職災案件的每月增加件數Y（單位為件，平均值為3件，標準差為2件），下表是以不同解釋變數配適每月增加件數Y之迴歸模型的參數最小平方估計量和誤差平方和SSE。

迴歸模型代號	迴歸模型中的解釋變數	參數最小平方估計量	SSE
LM1	X_1	$b_1=0.3$	5
LM2	X_2	$b_2=-0.1$	8
LM3	X_1, X_2	$b_1=0.2, b_2=-0.2$	4

下面所有小題的計算若除不盡，一律四捨五入到小數第二位，否則不給分。

(一)分別求出三個迴歸模型LM1～LM3之判定係數。（6分）

(二)分別求出三個迴歸模型LM1～LM3之修正判定係數。（6分）

(三) 使用題(二)的結果求出 Y 和 X_1 之相關係數以及 Y 和 X_2 之相關係數。（4分）

(四) 針對迴歸模型 LM3 於試卷上依序填入下列ANOVA表中(1)～(8)之8個空格內容。（8分）

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	(1)	(3)	(6)	(8)
Error	(2)	(4)	(7)	
Corrected Total	10	(5)		

(五) 針對迴歸模型LM3欲檢定 X_2 的係數是否為0，求出偏 F 檢定的計算值。（6分）

(六) 針對迴歸模型LM3，當迴歸係數在什麼條件下，MSR的值望值為 σ^2 ？（5分）

(七) 若三個變數之變異數－共變異數矩陣為：

$$V=\begin{bmatrix}1 & 1 & 0\\ 1 & 4 & 1\\ 0 & 1 & 16\end{bmatrix}$$

為了求其相關係數矩陣，需在 V 前後乘一個對角矩陣，寫出此對角矩陣。（5分）

試題評析	本題大多數題型於以往考古題中都常出現，跟上一題一樣考驗考生迴歸之矩陣法。
考點命中	《迴歸分析熱門題庫》，高點文化出版，趙治勳編著，第三章。

答：

由 $S_Y=2$ 得知 $SS_Y=SST=4\times(11-1)=40$

(一) LM1 之判定係數為 $R^2_{Y1}=1-\dfrac{SSE(X_1)}{SST}=1-\dfrac{5}{40}=\dfrac{7}{8}$

LM2 之判定係數為 $R^2_{Y2}=1-\dfrac{SSE(X_2)}{SST}=1-\dfrac{8}{40}=0.8$

LM3 之判定係數為 $R^2_{Y12}=1-\dfrac{SSE(X_1,X_2)}{SST}=1-\dfrac{4}{40}=0.9$

(二) LM1 之修正判定係數為 $\bar{R}^2_{Y1}=1-\dfrac{SSE(X_1)/(11-1-1)}{SST/(11-1)}=1-\dfrac{5/9}{40/10}=\dfrac{31}{36}$

LM2 之修正判定係數為 $\bar{R}^2_{Y2}=1-\dfrac{SSE(X_2)/(11-1-1)}{SST/(11-1)}=1-\dfrac{8/9}{40/10}=\dfrac{7}{9}$

LM3 之修正判定係數為 $\bar{R}^2_{Y12}=1-\dfrac{SSE(X_1,X_2)/(11-2-1)}{SST/(11-1)}=1-\dfrac{4/8}{40/10}=\dfrac{7}{8}$

(三) 已知 $R^2 = 1-(1-\bar{R}^2)\dfrac{n-k-1}{n-1}$

Y 與 X_1 之相關係數為 $r_{Y1} = sign(b_1)\sqrt{R_{Y1}^2} = sign(b_1)\sqrt{1-(1-\bar{R}_{Y1}^2)\dfrac{n-k-1}{n-1}}$

$$= +\sqrt{1-(1-\frac{31}{36})\frac{11-1-1}{11-1}} = 0.94$$

Y 與 X_2 之相關係數為 $r_{Y2} = sign(b_2)\sqrt{R_{Y2}^2} = sign(b_2)\sqrt{1-(1-\bar{R}_{Y2}^2)\dfrac{n-k-1}{n-1}}$

$$= -\sqrt{1-(1-\frac{7}{9})\frac{11-1-1}{11-1}} = -0.89$$

(四)

ANOVA TABLE

source	d.f.	SS	MS	F
Reg	2	36	18	$F^* = 36$
Error	8	4	0.5	
Total	10	40		

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

T.S.: $F = \dfrac{SSR(X_2 \mid X_1)/1}{SSE(X_1, X_2)/11-2-1} = \dfrac{1/1}{8/11-2-1} = 1$

其中 $SSR(X_2 \mid X_1) = SSR(X_1, X_2) - SSR(X_1) = (40-4)-(40-5) = 1$

(六) 在 $\beta_0 = \beta_1 = \beta_2 = \cdots = \beta_k = 0$ 下, $E(MSR) = \sigma^2$

(七) 令 $D = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 4 \end{bmatrix}$ 為 V 之對角線值開根號(即各變數之標準差)所組成之對角矩陣

在 V 之前後乘上對角矩陣 $D^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1/2 & 0 \\ 0 & 0 & 1/4 \end{bmatrix}$ 可以求出相關係數矩陣

四、某大數據資料分析公司以 Y 對三個變數 X_1,X_2,X_3 所做的複迴歸分析中，樣本大小 $n = 14$，得到複判定係數 $R^2 = 90\%$，又將 Y 改對變數 X_2 做迴歸分析時，得到判定係數 $R^2 = 70\%$，請回答下列問題：（每小題5分，共15分）

(一) 此公司欲檢定 X_1,X_3 的迴歸係數是否為0，請求出 F 檢定統計量的值。

(二) 設速食店營收為 Y（單位為百萬元），廣告費為 X_1，有得來速（drive-through）服務時 $X_2 = 1$，沒有則 $X_2 = 0$，做迴歸分析得到迴歸平面為 $Y = 1 + 1.5X_1 + 2X_2$，因有得來速服務多出來的平均營收為多少元？

(三) 承題(二)，令 R_j^2 代表解釋變數 X_j 對另一個解釋變數做迴歸分析得到的判定係數，$j = 1,2$，且 $R_1^2 = 0.65$，$R_2^2 = 0.95$，求出 X_2 之變異數膨脹因子（VIF），若VIF大於10代表模式有何問題？

試題評析	本題所有題型於以往考古題中都常出現，考生獲得滿分不難。
考點命中	《迴歸分析熱門題庫》，高點文化出版，趙治勳編著，第三章。

答：

(一)設 $Y=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\varepsilon$　　$R_F^2=0.9$

$Y=\beta_0+\beta_2X_2+\varepsilon$　　$R_R^2=0.7$

$H_0:\beta_1=\beta_3=0$ vs H_1:至少一個 $\beta_i\neq 0$, $i=1,3$

T.S.: $F=\dfrac{n-k-1}{k}\dfrac{R_F^2-R_R^2}{1-R_F^2}=\dfrac{14-3-1}{2}\times\dfrac{0.9-0.7}{1-0.9}=10$

(二)有得來速之平均營收： $E(Y\mid X_1,X_2=1)=(\beta_0+\beta_2)+\beta_1X_1$

沒有得來速之平均營收： $E(Y\mid X_1,X_2=0)=\beta_0+\beta_1X_1$

因此，有得來速多出來之平均營收為 $E(Y\mid X_1,X_2=1)-E(Y\mid X_1,X_2=0)=\beta_2$,估計值為 $\hat{\beta}_2=2$ (百萬元)

(三) X_2 之變異數膨脹因子為 $VIF_2=\dfrac{1}{1-R_2^2}=\dfrac{1}{1-0.95}=20$ ，而當 VIF 大於 10 表示模型可能有線性重合之問題