《迴歸分析》

試題評析

今年考題著重在公式的熟練與推導，再加上題目繁瑣，恐怕時間內不易解完，若能考到60分以上，應該上榜有望。第一題(一)在第二回P37、P38；(二)在總複習P17，(三)在第二回P53；第二題在第三回P46（總複習P28、P29）、第三題利用第一回P20、P21即可推導；第四題在第二回P36、P37、P73《例題8》；第五題在第二回P52、P53、P64、總複習P16。

一、下列是某校120個學生三次測驗成績 X_1, X_2, Y 的資料：$\overline{x}_1 = 6.8, \overline{x}_2 = 7, \overline{y} = 74; s_1 = 1, s_2 = 0.8, s_Y = 9;$ $r_{12} = 0.6,\ r_{Y1} = 0.7,\ r_{Y2} = 0.5$（$s$ =標準差；r =相關係數）。若考慮迴歸模式為 $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$，回答以下問題：

(一)推導最佳迴歸線($\hat{y} = b_0 + b_1 x_1 + b_2 x_2$)（10分）

(二)求偏相關係數 $r_{Y1.2}$ 及 $r_{Y2.1}$。（5分）

(三)解釋(二)中 $r_{Y1.2}$ 及 $r_{Y2.1}$ 的意義。（5分）

答：

(一) $$b_1 = \frac{(S_{22}S_{1Y} - S_{12}S_{2Y})/S_{11}S_{22}}{(S_{11}S_{22} - S_{12}^2)/S_{11}S_{22}} = \frac{r_{Y1}\cdot\sqrt{\dfrac{S_{YY}}{S_{11}}} - r_{12}r_{Y2}\cdot\sqrt{\dfrac{S_{YY}}{S_{11}}}}{1 - r_{12}^2} = \frac{S_Y}{S_1}\cdot\frac{r_{Y1} - r_{12}r_{Y2}}{1 - r_{12}^2}$$

$$= \frac{9}{1}\cdot\frac{0.7 - 0.6\times 0.5}{1 - 0.6^2} = 5.625$$

$$b_2 = \frac{S_Y}{S_2}\cdot\frac{r_{Y2} - r_{12}r_{Y1}}{1 - r_{12}^2} = \frac{9}{0.8}\cdot\frac{0.5 - 0.6\times 0.7}{1 - 0.6^2} = 1.40625$$

$$b_0 = \overline{y} - b_1\overline{x}_1 - b_2\overline{x}_2 = 25.90625$$

$\therefore$ 最佳迴歸線 $\hat{y} = 25.90625 + 5.625x_1 + 1.40625x_2$

(二)1. $$r_{Y1\bullet 2}^2 = \frac{(r_{Y1} - r_{12}\cdot r_{Y2})^2}{(1 - r_{12}^2)(1 - r_{Y2}^2)} = \frac{(0.7 - 0.6\times 0.5)^2}{(1 - 0.6^2)(1 - 0.5^2)} = 0.3333$$

$\therefore\ r_{Y1\bullet 2} = \sqrt{0.3333} = 0.5773$

2. $$r_{Y2\bullet 1}^2 = \frac{(r_{Y2} - r_{12}\cdot r_{Y1})^2}{(1 - r_{12}^2)(1 - r_{Y1}^2)} = \frac{(0.5 - 0.6\times 0.7)^2}{(1 - 0.6^2)(1 - 0.7^2)} = 0.0196$$

$\therefore\ r_{Y2\bullet 1} = \sqrt{0.0196} = 0.14$

(三) $r_{Y1\bullet 2}$ 表示 x_2 固定之下，Y 與 x_1 的直線相關係數

$r_{Y2\bullet 1}$ 表示 x_1 固定之下，Y 與 x_2 的直線相關係數

二、甲生將一組包含Y及四個自變數，X_1,X_2,X_3,X_4 的資料做以下所有可能的模式的分析，其目的在於選取可能的最佳模式（$p=$ 模式參數個數；MSE=Mean square error; df=自由度）。

X Variables in Model	p	df	SSE_p	R_P^2	MSE_P	C_P	$PRESS_P$
None	1	53	4	0	0.075	1721	4.12
X_1	2	52	3.5	0.12	0.067	1511	3.81
X_2	2	52	2.58	0.35	0.05	1100	2.86
X_3	2	52	2.22	0.44	0.043	939	2.43
X_4	2	52	1.88	0.53	0.036	788	2.03
X_1X_2	3	51	2.23	0.44	0.04	949	2.64
X_1X_3	3	51	1.41	0.65	0.03	580	1.61
X_1X_4	3	51	1.88	0.53	0.036	789	2.12
X_2X_3	3	51	0.74	0.81	0.015	284	0.84
X_2X_4	3	51	1.39	0.65	0.027	574	1.58
X_3X_4	3	51	1.25	0.69	0.024	508	1.43
$X_1X_2X_3$	4	50	0.11	0.97	0.002	3.1	0.145
$X_1X_2X_4$	4	50	1.39	0.65	0.028	575	1.65
$X_1X_3X_4$	4	50	1.12	0.72	0.022	452	1.33
$X_2X_3X_4$	4	50	0.47	0.88	0.009	162	0.55
$X_1X_2X_3X_4$	5	49	0.11	0.97	0.0022	5	0.15

從如何判定模式中該包括那些自變數的方向上，回答下列問題：

(一)請說明 R_P^2 判定準則的內容並依此決定最適模式。（5分）

(二)請說明 MSE_P 判定準則的內容並依此決定最適模式。（5分）

(三)請說明 C_P 判定準則的內容並依此決定最適模式。（5分）

(四)請說明PRESSp判定準則的內容並依此決定最適模式。（5分）

答：

(一) R_p^2 準則：$R_p^2=\dfrac{SSR_p}{SSTO}$

R_p^2 缺點在於自變數($X_i\neq 0$)愈多個時，R_p^2 會增大，因此用 R_p^2 準則時，不要求取 R_p^2 最大者為最佳自變數組合，只要找到適當的自變數時(R_p^2 夠大)即可，對model而言，即再增加多餘的自變數，R_p^2 增加量只有一點點，就不需要篩選進來。

1.一個自變數以 X_4 之 $R_p^2=0.5$ max

2.二個自變數以 X_2X_3 之 $R_p^2=0.81$ max

3.三個自變數以 $X_1X_2X_3$ 之 $R_p^2=0.97$ max

4.四個自變數 $R_p^2=0.97$ 並沒有增加

∴ 最適當模式：選用 $X_1X_2X_3$

(二)MSE$_P$準則：

$$R_a^2 = 1 - \frac{SSE_P/n-p}{SSTO/n-1} \uparrow \quad \Leftrightarrow \quad MSE_p \downarrow$$

當自變數個數 $p-1$ 增加時，R_a^2 並不一定變大，因爲若 SSE_P 減少量，不足以彌補自由度減小時，則 MSE 將會遞增(R_a^2 將會遞減)，因此採用 R_a^2 準則時，不要求取 MSE_P 最小者(R_a^2 最大者)，只要在取min(MSE_P)過程中，由遞減開始遞增就停止。

1.一個自變數以 X_4 之 $MSE_P = 0.036$ min
2.二個自變數以 X_2X_3 之 $MSE_P = 0.015$ min
3.三個自變數以 $X_1X_2X_3$ 之 $MSE_P = 0.002$ min
4.四個自變數 $X_1X_2X_3X_4$ 之 $MSE_P = 0.0022$ 增加
$\therefore$ 最適當模式選用 $X_1X_2X_3$

(三) C_P 準則：

$$C_p = \frac{SSE_p}{MSE_F} - (n-2p)$$

取 C_p 最小者，即爲最佳自變數組合

1.一個自變數：$n=54$，$p=2$，以 X_4 之 C_p 最小
$C_2 = 788$
2.二個自變數：$n=54$，$p=3$，以 X_2X_3 之 C_p 最小
$C_3 = 284$
3.三個自變數：$n=54$，$p=4$，以 $X_1X_2X_3$ 之 C_p 最小
$C_4 = 3.1$
4.四個自變數：$n=54$，$p=5$
$C_5 = 5$
$\because C_4$ 最小 $\therefore$ 最適當模式選用 $X_1X_2X_3$

(四)預測平方和(prediction sum of squares)

$$PRESS_P = \sum_{i=1}^{n} (Y_i - \hat{Y}_{i(i)})^2$$

其中 $\hat{Y}_{i(i)}$ 第一個 i 爲第 i 個預測值
第二個 i 爲剔除第 i 個後所fit的regression function
$\therefore$ 好的模式應具有較小的 $PRESS_P$ 值
$\because X_1X_2X_3$ 之 $PRESS_P$ 最小
$\therefore$ 最適當模型選用 $X_1X_2X_3$。

三、在簡單線性迴歸模式下，請回答下列問題：
(一)請解釋自變數值大小的分散程度如何影響 b_1（β_1 的估計式）的變異數大小。（5分）
(二)請解釋為何殘差不是獨立的隨機變數。（5分）

答：

(一) $Var(b_1) = \dfrac{\sigma^2}{\sum (x_i - \bar{x})^2} = \dfrac{\sigma^2}{(n-1)S_X^2}$

$\therefore$ 當自變數 X 的分散度愈大，則 $Var(b_1)$ 愈小。

(二)當 $i \neq j$

1. $\text{cov}(Y_i, Y_j)=0$

2. $\text{cov}(Y_i, \hat{Y}_j)=\text{cov}(Y_i, \bar{Y}+\hat{\beta}(x_j-\bar{x}))$

$$=\text{cov}(Y_i, \bar{Y})+(x_j-\bar{x})\text{cov}(Y_i, \frac{\sum_{k=1}^{n}(x_k-\bar{x})\,Y_k}{S_{XX}})$$

$$=\frac{\sigma^2}{n}+\frac{(x_j-\bar{x})\cdot(x_i-\bar{x})\sigma^2}{S_{XX}}$$

3.同理 $\text{cov}(\hat{Y}_i, Y_j)=\frac{\sigma^2}{n}+\frac{(x_j-\bar{x})\cdot(x_i-\bar{x})\sigma^2}{S_{XX}}=\text{cov}(Y_i, \hat{Y}_j)$

4. $\text{cov}(\hat{Y}_i, \hat{Y}_j)=\text{cov}(\hat{\alpha}+\hat{\beta}x_i, \hat{\alpha}+\hat{\beta}x_j)$

$$=(\frac{1}{n}+\frac{\bar{x}^2}{S_{XX}})\sigma^2-\frac{\bar{x}x_j}{S_{XX}}\sigma^2-\frac{\bar{x}x_i}{S_{XX}}\sigma^2+\frac{x_ix_j}{S_{XX}}\sigma^2$$

$$=\frac{\sigma^2}{n}+\frac{(x_i-\bar{x})\cdot(x_j-\bar{x})\sigma^2}{S_{XX}}=\text{cov}(Y_i, \hat{Y}_j)$$

$\therefore \text{cov}(e_i, e_j)=\text{cov}(Y_i-\hat{Y}_i, Y_j-\hat{Y}_j)$

$$=\text{cov}(Y_i, Y_j)-2\text{cov}(Y_i, \hat{Y}_j)+\text{cov}(\hat{Y}_i, \hat{Y}_j)$$

$$=0-\text{cov}(Y_i, \hat{Y}_j)$$

$$=-\left[\frac{\sigma^2}{n}+\frac{(x_i-\bar{x})\cdot(x_j-\bar{x})}{S_{XX}}\sigma^2\right]\neq 0$$

則e_i與e_j不是indep.

$\therefore e_1, e_2, \cdots, e_n$不是獨立隨機變數

四、經濟學家想了解一個新的保險方案被接受的速度(Y)與保險公司大小(X_1)及公司種類(X_2)的相關性，在考慮$Y_i=\beta_0+\beta_1X_{i1}+\beta_2X_{i2}+\varepsilon_i$的模式下，$Y_1$=接受新的保險方案所需時間（月），$X_{11}$=保險公司大小，$X_{12}=1$（證券公司），$X_{12}=0$（基金公司），收集樣本資料共20家公司。統計分析結果如下：

S.V.	SS	df	MS
Regression	1504.41	2	752.2
Error	176.39	17	10.38
Total	1680.8	19	

Regression Coefficient	Estimated Regression coefficient	Estimated Standard deviation
β_0	33.87407	1.81386
β_1	-0.10174	0.00889
β_2	8.05547	1.45911

(一)寫下估計的迴歸線並解釋迴歸係數估計值b_1與b_2的意義。(5分)

(二)求β_2的95%信賴區間,並解釋該區間之含義。(5分)

(三)檢定$H_0:\beta_1=0$ vs. $H_a:\beta_1\neq 0$並說明檢定結果。($\alpha=0.05$)(5分)

(四)假設經濟學家採用包含X_1及X_2交互項(interaction term)的模式,

$Y_i=\beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}+\beta_3 X_{i1}X_{i2}+\varepsilon_i$,說明模式中係數$\beta_1,\beta_2$與$\beta_3$的意義。(5分)

(五)利用Bonferroni procedure求β_0與β_1的90%聯合信賴區並解釋該區間的意義。

$(t_{0.975}(17)=2.11, t_{0.95}(17)=1.74)$(5分)

答:

(一)1. $\hat{Y}_i=33.87407-0.10174X_{i_1}+8.05547X_{i_2}$

2. $b_1=-0.10174$

表示當X_{i2}固定時,保險公司大小(X_{i1})每增加一單位時,則平均接受新方案的速度$\hat{Y}$將減小0.10174個月。

3. $b_2=8.05547$

表示當X_{i1}固定時,證券公司比基金公司平均接受新方案的速度將增加8.05547個月。

(二) β_2之95% C.I. $=b_2\pm t_{\frac{0.05}{2}}(20-3)\cdot S_{b_2}$

$=8.05547\pm 2.11\times 1.45911=(4.97675 , 11.13419)$

(三)1. $\begin{cases}H_0:\beta_1=0\\ H_1:\beta_1\neq 0\end{cases}$

2. $C=\left\{T \;\middle|\; |T|>t_{\frac{0.05}{2}}(20-3)=2.11\right\}$

3. $|T|=\left|\dfrac{b_1}{S_{b_1}}\right|=\left|\dfrac{-0.10174}{0.00889}\right|=11.444\in C$

$\therefore$ reject H_0,$\beta_1\neq 0$

(四) $EY_i=\beta_0+\beta_1X_{i1}+\beta_2X_{i2}+\beta_3X_{i1}X_{i2}$

β_1與β_2表示線型效果係數,β_3表示交互效果係數

1.當$X_2=0$時

$E(Y|X_2=0)=\beta_0+\beta_1X_1$表示基金公司平均接受新保險方案之迴歸線

β_1表示此迴歸線之斜率。

2.當$X_2=1$時

$E(Y|X_2=1)=(\beta_0+\beta_2)+(\beta_1+\beta_3)X_1$表示證券公司平均接受新保險方案之迴歸線

$(\beta_1+\beta_3)$表示此迴歸線之斜率,β_2表示二條迴歸線之截距差。

(五) $C_1^2=2$

1. β_0之90% Bonferroni C.I.

$=b_0\pm t_{1-\frac{0.1}{2\times 2}}(20-3)\cdot S_{b_0}=33.87407\pm 2.11\times 1.81386=(30.04683 , 37.70131)$

2. β_1之90% Bonferroni C.I.

$=b_1\pm t_{1-\frac{0.1}{2\times 2}}(20-3)\cdot S_{b_1}=-0.10174\pm 2.11\times 0.00889=(-0.12050 , -0.08298)$

表示此一成對區間包含β_0與β_1可信度達90%。

五、營養學家欲研究體脂量(Y)與三個可能預測變數，肌皮脂厚度(X_1)，大腿圍(X_2)及上臂圍(X_3)之間的關係。由年齡介於25至34歲健康女性族群中抽出20位，並收集體脂量，肌皮脂厚度，大腿圍及上臂圍等資料。營養學家藉由此份資料進行以下四種迴歸模式分析：

(M_1)Regression of Y on $X_1: \hat{Y} = -1.496 + 0.8572X_1$

S.V.	SS	df	MS
Regression	352.27	1	352.27
Error	143.12	18	7.95

(M_2)Regression of Y on $X_2: \hat{Y} = -23.364 + 0.8565X_2$

S.V.	SS	df	MS
Regression	381.97	1	381.97
Error	113.42	18	6.3

(M_3)Regression of Y on X_1 and $X_2: \hat{Y} = -19.174 + 0.2224X_1 + 0.6594X_2$

S.V.	SS	df	MS
Regression	385.44	2	192.72
Error	109.95	17	6.47

(M_4)Regression of Y on X_1, X_2 and $X_3: \hat{Y} = 117.08 + 4.334X_1 - 2.857X_2 - 2.186X_3$

S.V.	SS	df	MS
Regression	396.98	3	132.33
Error	98.41	16	6.15

Note：S.V.= Source of Variation; SS= Sum of squares; df=degree of freedoms; MS=Mean square

(一)求 $SSR(X_2|X_1), SSR(X_3|X_1, X_2), SSR(X_2, X_3|X_1), SSR(X_1, X_3|X_2)$ 及 $SSE(X_1, X_3|X_2)$。(5分)

(二)解釋(一)中 $SSR(X_3|X_1, X_2)$ 值及 $SSR(X_2, X_3|X_1)$ 值之意義。(5分)

(三)假設迴歸模式已包含 X_1 及 X_2，在 $\alpha = 0.01$，檢定 $\beta_3 = 0$ 是否成立並判定變數 X_3 是否該存在模式中？($F(0.99;1,16) = 8.53$) (5分)

(四)說明造成模式(M_1)與模式(M_3)中，X_1 的迴歸係數不相同之可能原因。(5分)

(五)計算並解釋下列部分判定係數(coefficients of partial determination)：$R^2_{Y2|1}$ 及 $R^2_{Y3|12}$。(5分)

答：

(一)1. $SSR(X_2 | X_1) = SSR(X_1X_2) - SSR(X_1) = 385.44 - 352.27 = 33.17$

2. $SSR(X_3 | X_1X_2) = SSR(X_1X_2X_3) - SSR(X_1X_2) = 396.98 - 385.44 = 11.54$

3. $SSR(X_2X_3 | X_1) = SSR(X_1X_2X_3) - SSR(X_1) = 396.98 - 352.27 = 44.71$

4. $SSR(X_1X_3 | X_2) = SSR(X_1X_2X_3) - SSR(X_2) = 396.98 - 381.97 = 15.01$

(二)1. $SSR(X_3 | X_1X_2)$

表示原模型 $Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \varepsilon$

再引進 X_3 之後，SSR 的增加量或 SSE 的減少量

2. $SSR(X_2X_3 | X_1)$

表示原模型 $Y = \beta_0 + \beta_1X_1 + \varepsilon$

再引進 X_2 與 X_3 之後，SSR 的增加量或 SSE 的減少量

(三)1. $\begin{cases} H_0 : \beta_3 = 0 \\ H_1 : \beta_3 \neq 0 \end{cases}$

2. $C = \left\{ F \;\middle|\; F > F_{0.01}(1\,,\ 20-4) = 8.53 \right\}$

3. $F = \dfrac{SSR(X_3 \mid X_1 X_2)/1}{SSE(X_1 X_2 X_3)/(n-k)} = \dfrac{11.54}{6.15} = 1.876 \notin C$

$\therefore$ not reject H_0，表示 $\beta_3 = 0$ 成立， X_3 不應該存在。

(四) $M_1 : \hat{Y} = \hat{\alpha} + \hat{\beta} X_1$

$M_3 : \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$

$\hat{\beta} = \hat{\beta}_1 + \hat{\beta}_2 \times \dfrac{S_{12}}{S_{11}}$

$\therefore$ 當 X_1 與 X_2 有共線性($r_{12} \neq 0 \iff S_{12} \neq 0$)時，兩模型 X_1 係數就不相同。

(五) $R^2_{Y2|1} = \dfrac{SSR(X_2 \mid X_1)}{SSE(X_1)} = \dfrac{33.17}{143.12} = 0.232$

$R^2_{Y3|12} = \dfrac{SSR(X_3 \mid X_1 X_2)}{SSE(X_1 X_2)} = \dfrac{11.54}{109.95} = 0.105$