《迴歸分析》

試題評析	今年地特迴歸分析可以說是歷屆試題中(含高考)最難的一份，完全沒有計算， 除了第四題外，都著重在模型之轉換與證明，且需要考生把一些常用的證明技巧應用在其他類題中，若考生沒有把迴歸分析中之數理推導加強訓練的話，很容易考取極低分數。本卷基本分20分，中等程度者會有50分。
考點命中	1.《高點・高上迴歸分析熱門題庫書籍》趙治勳編著，頁2-43 2.《高點・高上迴歸分析熱門題庫書籍》趙治勳編著，頁2-19 3.《高點・高上迴歸分析熱門題庫書籍》趙治勳編著，頁2-23(P.I.之證明)、(C.I.之證明) 4.《高點・高上迴歸分析熱門題庫書籍》趙治勳編著，頁2-46

一、假定 y 表示公司產品一年銷售量。與 y 相關的變數（如廣告費用、人事成本等）有 $x_1,x_2,\ldots,x_p$ ，我們考慮線性迴歸（Regression）模型： $y=\beta_0+\beta_1x_1+\cdots+\beta_px_p+\varepsilon$

其中 ε 服從常態分配。公司一資深經理認為，變數 x_2 為 x_1 的兩倍效力而 x_3 又為 x_2 的 $1/3$ 的效力。請問當我們有 n 個觀察值($y_i,x_{i1},\ldots,x_{ip}$)， $i=1,\ldots,n$ 時，您要如何驗證這個經理的說法是否正確(請說明您的假設)。(25 分)

答：

題中「變數 x_2 爲 x_1 的兩倍效力」意謂著自變數 x_2 對應變數 Y 之邊際影響是 x_1 的兩倍, 即 $\beta_2=2\beta_1$

題中「 x_3 又爲 x_2 的 $1/3$ 倍效力」意謂著自變數 x_3 對應變數 Y 之邊際影響是 x_2 的 $1/3$ 倍，即 $\beta_3=\frac{1}{3}\beta_2 \overset{\because \beta_2=2\beta_1}{=} \frac{2}{3}\beta_1$

若該資深經理之想法正確的話，模型修改如下：

$$\begin{aligned}Y&=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\cdots+\beta_px_p+\varepsilon\\&=\beta_0+\beta_1x_1+2\beta_1x_2+\frac{2}{3}\beta_1x_3+\beta_4x_4+\cdots+\beta_px_p+\varepsilon\\&=\beta_0+\beta_1(x_1+2x_2+\frac{2}{3}x_3)+\beta_4x_4+\cdots+\beta_px_p+\varepsilon\\&=\beta_0+\beta_1x^*+\beta_4x_4+\cdots+\beta_px_p+\varepsilon \quad \text{其中 } x^*=x_1+2x_2+\frac{2}{3}x_3\end{aligned}$$

再檢定以上模型中 β_1 是否爲零, 即可驗證該資深經理之想法是否正確

假設： $Y_i=\beta_0+\beta_1x_i^*+\beta_4x_{4i}+\cdots+\beta_px_{pi}+\varepsilon_i$, 其中 $x_i^*=x_{1i}+2x_{2i}+\frac{2}{3}x_{3i}$

$$\varepsilon_i \overset{iid}{\sim} N(0,\sigma^2),\ i=1,2,\cdots,n$$

$$H_0:\beta_1=0 \quad \text{vs} \quad H_1:\beta_1\neq 0$$

T.S.: $T=\frac{\hat{\beta}_1-0}{S(\hat{\beta}_1)}\sim t_{((n-1)-(p-2)=n-p+1)}$

R.R.: Reject H_0 at α if $|T^*|>t_{\frac{\alpha}{2}(n-p+1)}$

若以上檢定拒絕 H_0 ，即我們有足夠證據去推論該資深經理之想法是正確的。

二、假定我們有一線性迴歸模型：$y_i = x_i'\beta + \varepsilon_i$， $i = 1,\ldots,n$

其中 $\varepsilon_1,\ldots,\varepsilon_n$ 為 idd 隨機變數其平均數為 0 及變異數為 σ^2。令 $\hat{y}_i$ 為 y_i 之預測值。

(一)請問若 $\varepsilon_1,\ldots,\varepsilon_n$ 不具有常態分配，那 $\sum_{i=1}^{n}(y_i - \hat{y}_i) = 0$ 是否仍是對的。請驗證。(13 分)

(二)我們假設 $\varepsilon_1,\ldots,\varepsilon_n$ 為 idd $N(0,1)$ 的隨機變數。令 $\bar{\varepsilon}$ 為 $\varepsilon_1,\ldots,\varepsilon_n$ 的平均，請找出 $\bar{\varepsilon}$ 的 $100(1-\alpha)\%$ 信賴區間。(12 分)

答：

(一)仍然是對了，因爲從正規方程式中，可得知

$$\sum Y_i = n\hat{\beta}_0 + \sum X_i\hat{\beta}_1 = \sum(\hat{\beta}_0 + \hat{\beta}_1 X_i) = \sum \hat{Y}_i$$

故 $\sum(Y_i - \hat{Y}_i) = \sum Y_i - \sum \hat{Y}_i = 0$

由以上證明可得知 $\sum(Y_i - \hat{Y}_i) = 0$ 並未用到 ε_i 服從常態分配之假設

(二)

$$\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n \overset{iid}{\sim} N(0,1),\ \bar{\varepsilon} = \frac{\sum \varepsilon_i}{n} \sim N(0, \frac{1}{n})$$

$$P(-z_{\frac{\alpha}{2}} \le \frac{\bar{\varepsilon} - 0}{1/\sqrt{n}} \le z_{\frac{\alpha}{2}}) = 1-\alpha \Rightarrow P(-z_{\frac{\alpha}{2}}\frac{1}{\sqrt{n}} \le \bar{\varepsilon} \le z_{\frac{\alpha}{2}}\frac{1}{\sqrt{n}}) = 1-\alpha$$

$\bar{\varepsilon}$ 之 $100(1-\alpha)\%$ 信賴區間爲 $(-z_{\frac{\alpha}{2}}\frac{1}{\sqrt{n}}, z_{\frac{\alpha}{2}}\frac{1}{\sqrt{n}})$

三、我們考慮一個工業產品某一品質變數 y，它滿足下面迴歸模型：$y_i = x_i'\beta + \varepsilon_i$， $i = 1,\ldots,n$

其中 x_i 包含截距項的 $(p+1)$ 向量且 $\varepsilon_1,\ldots,\varepsilon_n$ 為 idd $N(0,\sigma^2)$ 的變數。若有一向量 x_a 想知道這時對應變數 y_a 的表現。我們重覆觀察 m 次產生下面樣本模型：$y_{aj} = x_{aj}'\beta + \varepsilon_{aj}$， $j = 1,\ldots,m$

其中 $\varepsilon_{aj}'s$ 與 $\varepsilon_i's$ 獨立且具相同分配，令 $y_a = \frac{1}{m}\sum_{j=1}^{m} y_{aj}$。

(一)找出 y_a 的 $100(1-\alpha)\%$ 信賴區間。(13 分)

(二)令 $E(y_a \mid x_a) = x_a'\beta$，考慮假設 H_0：$E(y_a \mid x_a) = 3$ vs. H_1：$E(y_a \mid x_a) > 3$，請導出顯著水準為 α 的檢定。(12 分)

答：

(一)由題意，$Y_i = x_i'\beta + \varepsilon_i, \varepsilon_i \overset{iid}{\sim} N(0,\sigma^2), i = 1,2,\cdots,n$

在向量 x_a 下之樣本迴歸線 $\hat{Y}_a = x_a'\hat{\beta} \sim N(x_a'\beta, x_a'(x_a x_a')^{-1}x_a\sigma^2)$

其中 $V(\hat{Y}_a) = V(x_a'\hat{\beta}) = x_a'V(\hat{\beta})x_a = x_a'(x_a x_a')^{-1}\sigma^2 x_a = x_a'(x_a x_a')^{-1}x_a\sigma^2$

再由題意，$Y_{aj} = x_a'\beta + \varepsilon_{aj}, \varepsilon_{aj} \overset{iid}{\sim} N(0,\sigma^2), j = 1,2,\cdots,m$

可得 $E(Y_{aj}) = x_a'\beta$， $V(Y_{aj}) = V(\varepsilon_{aj}) = \sigma^2$

故 $Y_a = \frac{\sum Y_{aj}}{m} \sim N(x_a\beta, \frac{\sigma^2}{m})$

其中 $E(Y_a)=E(\frac{\sum Y_{aj}}{m})=x_a^{'}\beta$ ，$V(Y_a)=V(\frac{\sum Y_{aj}}{m})=\frac{\sigma^2}{m}$

經由以上討論後可以得到，$\hat{Y}_a-Y_a\sim N(0,(x_a^{'}(x_ax_a^{'})^{-1}x_a+\frac{1}{m})\sigma^2)$ （$\because \varepsilon_{aj}\perp\varepsilon_i$）

標準化後，$\frac{(\hat{Y}_a-Y_a)-0}{\sqrt{(x_a^{'}(x_ax_a^{'})^{-1}x_a+\frac{1}{m})\sigma^2}}\sim N(0,1)$

樞紐量: $\frac{\hat{Y}_a-Y_a}{\sqrt{(x_a^{'}(x_ax_a^{'})^{-1}x_a+\frac{1}{m})MSE}}\sim t_{(n-p-1)}$ 其中 $MSE=\frac{\sum(Y_i-\hat{Y}_i)^2}{n-p-1}$

機率區間: $P(-t_{\frac{\alpha}{2}(n-p-1)}\le\frac{\hat{Y}_a-Y_a}{\sqrt{(x_a^{'}(x_ax_a^{'})^{-1}x_a+\frac{1}{m})MSE}}\le t_{\frac{\alpha}{2}(n-p-1)})=1-\alpha$

結論: Y_a 之 $100(1-\alpha)\%$ 信賴區間爲

$$(\hat{Y}_a\mp t_{\frac{\alpha}{2}(n-p-1)}\sqrt{(x_a^{'}(x_ax_a^{'})^{-1}x_a+\frac{1}{m})MSE})$$

(二)

由題意，$Y_i=x_i^{'}\beta+\varepsilon_i,\varepsilon_i\overset{iid}{\sim}N(0,\sigma^2),i=1,2,\cdots,n$

在向量 x_a 下之樣本迴歸線 $\hat{Y}_a=x_a^{'}\hat{\beta}\sim N(x_a^{'}\beta,x_a^{'}(x_ax_a^{'})^{-1}x_a\sigma^2)$

故 $\frac{\hat{Y}_a-x_a^{'}\beta}{\sqrt{(x_a^{'}(x_ax_a^{'})^{-1}x_a)\sigma^2}}=\frac{\hat{Y}_a-E(Y_a\mid x_a)}{\sqrt{(x_a^{'}(x_ax_a^{'})^{-1}x_a)\sigma^2}}\sim N(0,1)$

樞紐量: $\frac{\hat{Y}_a-E(Y_a\mid x_a)}{\sqrt{(x_a^{'}(x_ax_a^{'})^{-1}x_a)MSE}}\sim t_{(n-p-1)}$ 其中 $MSE=\frac{\sum(Y_i-\hat{Y}_i)^2}{n-p-1}$

機率區間: $P(\frac{\hat{Y}_a-E(Y_a\mid x_a)}{\sqrt{(x_a^{'}(x_ax_a^{'})^{-1}x_a)MSE}}\le t_{\alpha(n-p-1)})=1-\alpha$

故 $E(Y_a\mid x_a)$ 之 $100(1-\alpha)\%$ 左界信賴區間爲 $(\hat{Y}_a-t_{\alpha(n-p-1)}\sqrt{(x_a^{'}(x_ax_a^{'})^{-1}x_a)MSE},\infty)$

因此，針對檢定 $H_0:E(Y_a\mid x_a)=3$ vs $H_1:E(Y_a\mid x_a)>3$

在顯著水準爲 α 下，當區間 $(\hat{Y}_a-t_{\alpha(n-p-1)}\sqrt{(x_a^{'}(x_ax_a^{'})^{-1}x_a)MSE},\infty)$ 包含 3 時，我們會拒絕 H_0。(此爲信賴區間檢定法)

四、令 y 表示嬰兒身高。一奶粉公司考慮如何選取對身高有益的元素（變數）。現有 x_1, x_2, x_3, x_4 四種變數供選取。若 A, B 為 x_1, x_2, x_3, x_4 的部分集合。我們令 $F(A|B)=\frac{MSR(A|B)}{MSE(A,B)}$，表示當 B 集合變數已在模型內時 A 集合變數的部分 F（partial F）檢定統計量。再令 $P(A|B)$ 表示 $F(A|B)$ 觀察值的 P 值。我們現在有下面 P 值：

$P(x_1)=0.03$ ， $P(x_2)=0.02$ ， $P(x_3)=0.01$ ， $P(x_4)=0.02$

$P(x_i|x_1)=0.04$ ， $i=2,3$ ， $P(x_4|x_1)=0.01$

$P(x_i|x_2)=0.02$ ， $i=1,3$ ， $P(x_4|x_2)=0.01$

$P(x_1|x_3)=0.04$ ， $P(x_2|x_3)=0.05$ ， $P(x_4|x_3)=0.035$

$P(x_i|x_4)=0.05$ ， $i=1,2,3$ ， $P(x_i|x_1,x_2)=0.06$ ， $i=3,4$

$P(x_i|x_1,x_3)=0.07$ ， $i=2,4$ ， $P(x_i|x_1,x_4)=0.07$ ， $i=2,3$

$P(x_i|x_2,x_3)=0.06$ ， $i=1,4$ ， $P(x_i|x_2,x_4)=0.06$ ， $i=1,3$

$P(x_i|x_3,x_4)=0.06$ ， $i=1,2$

我們考慮 Forward selection 選取變數。在下面問題請說明理由及每階段選取結果。

（一）當 $\alpha=0.05$ 時，請找出選取的模型。（15 分）

（二）當 $\alpha=0.03$ 時，請找出選取的模型。（10 分）

答：

(一)因為 $P(x_i), i=1,2,3,4$ 都小於 $\alpha=0.05$ ，且最小值為 $P(x_3)=0.01$，故先考慮 x_3 加入模型中。再由 $P(x_i|x_3), i=1,2,4$ 去看，只有 $P(x_i|x_3), i=1,4$ 小於 $\alpha=0.05$ ，且最小值為 $P(x_4|x_3)=0.035$ ，故再考慮 x_4 加入模型中。然後觀察 $P(x_i|x_3,x_4), i=1,2$ 都已經沒有小於 $\alpha=0.05$ 。因此，在 $\alpha=0.05$ 時，Forward selection 選取的模型應該包含 x_3, x_4 兩個自變數。

(二)因為 $P(x_i), i=1,2,3,4$ 中只有 $P(x_i), i=2,3,4$ 小於 $\alpha=0.03$，且最小值為 $P(x_3)=0.01$，故先考慮 x_3 加入模型中。然後觀察 $P(x_i|x_3), i=1,2,4$ 都已經沒有小於 $\alpha=0.03$。因此，在 $\alpha=0.03$ 時，Forward selection 選取的模型應該只包含 x_3 一個自變數。