# 《迴歸分析》

| | |
|---|---|
| **試題評析** | 本年考題需要相當熟悉理論架構，計算量較去年少，但申論問答的部分又較去年多，考題難度約略中上，程度好的同學應可拿80分以上，但較難拿到90分以上。<br>第一題：考判定係數的評述，需就簡單迴歸與複迴歸加以分述才可以完整拿分，完整表達絕對是是否能全拿分的關鍵。屬於須考量全部觀點的申論題。<br>第二題：考複迴歸分析，模型設定及解釋係數的能力，且要熟悉殘差分析才有辦法設定正確模型。<br>第三題：考迴歸模型診斷，是正課與總複習再三強調的重點，故熟讀就能拿分。<br>第四題：考報表綜合題型，須熟悉額外平方和的概念輔以報表判讀解釋能力，才有辦法完整回答。 |
| **考點命中** | 1.《高點・高上迴歸分析講義》第一回，許勝雄編撰，頁36-37實例說明。<br>《高點・高上迴歸分析講義》第二回，許勝雄編撰，頁58-59定理說明。<br>2.《高點・高上迴歸分析講義》第三回，許勝雄編撰，與頁52例題7相同。<br>3.《高點・高上迴歸分析講義》第三回，許勝雄編撰，頁1-2、81；《第二回》頁33相同。<br>4.《高點・高上迴歸分析講義》第三回，許勝雄編撰，與頁48例題5相同。 |

一、迴歸分析中分別對下列陳述做一評述：

（一）如$R^2$（coefficient of determination判定係數）大（譬如0.95），則此模型良好，應採用。（6分）

（二）如$R^2$（coefficient of determination判定係數）小（譬如0.35），則此模型不佳，不應採用。（6分）

**答**：

(一)若在簡單迴歸分析中，當 $R^2=\frac{SSR}{SSTO}=\frac{\text{可解釋變異}}{\text{總變異}}$，若$R^2$(*判定係數*)*大*，

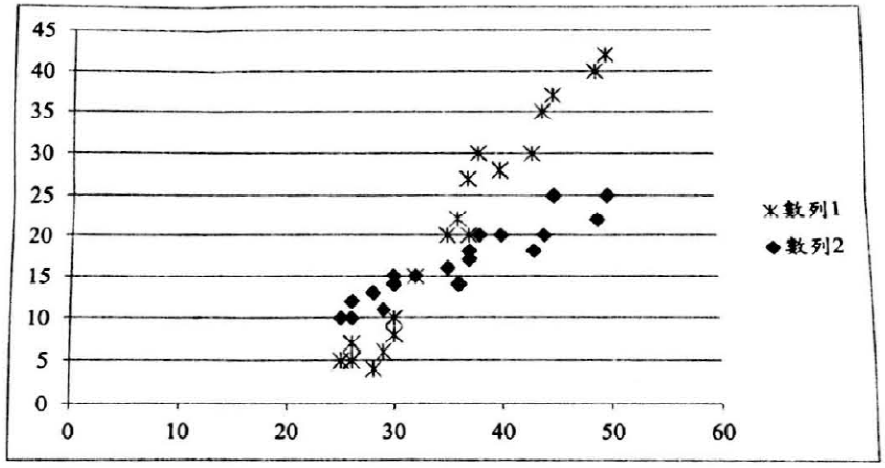
則F統計量$=\frac{(n-2)R^2}{1-R^2}$，F越顯著，則代表ANOVA-F檢定顯著，意味模型配適良好；而在複迴歸中，由於自變數越多，會使得複判定係數 $R^2=\frac{SSR}{SSTO}=\frac{\text{可解釋變異}}{\text{總變異}}$ 越大，故自變數個數增加，複判定係數就上升，但無法確認引進的變數是否皆為有效解釋變數，而在複迴歸中，由F統計量$=\frac{(n-k)R^2}{(k-1)(1-R^2)}$可知，當若要F檢定量夠大而拒絕虛無假設時，仍須考量樣本數、自變數個數與其他的因素方可判定，故複迴歸分析則須參考修正判定係數$R_a^2=\bar{R}^2=1-\frac{n-1}{n-k}(1-R^2)$為判定複迴歸的模型良好與否指標，避免引進過多自變數而令複判定係數變大造成無法辨別引進自變數是否正確。

(二)若在簡單迴歸分析中，當 $R^2=\frac{SSR}{SSTO}=\frac{\text{可解釋變異}}{\text{總變異}}$，若$R^2$(*判定係數*)*小*

代表此一模型SSE較高，配適較差，則F統計量$=\frac{(n-2)R^2}{1-R^2}$所計算的檢定統計量會越不顯著(越小)，模型**相對**比較不佳(容易無法拒絕虛無假設)，但實務上仍需端視所研究的主題而來判讀而定；就複迴歸的觀點而言，由於F統計量為由F統計量$=\frac{(n-k)R^2}{(k-1)(1-R^2)}$可知，則$R^2$(*判定係數*)*小*使F檢定量夠而拒絕虛無假設時，仍須考量樣本數、自變數個數其他的因素方可判定，故複迴歸分析仍需參考修正判定係數$R_a^2=\bar{R}^2=1-\frac{n-1}{n-k}(1-$

$R^2$)來判定複迴歸模型優劣為宜。

二、某公司兩業務單位（A,B）之員工業績（Y）及年齡（X）散佈圖如下，數列1為單位A，數列2為單位B。欲觀察年齡與所屬業務單位如何影響員工業績，請以員工業績（Y）為應變數設一個複迴歸模式，並解釋模式中每個迴歸係數之涵意。（24分）




**答**：

由題意，建立以下複迴歸模式：

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 D_i + \beta_3 X_i D_i + \varepsilon_i \ ， \quad 其中\ D_i = \begin{cases} 1 \ ， 單位A \\ 0 \ ， 單位B \end{cases}$$

當$D_i = 0$*時代表單位B*，模式變為$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$，$\beta_0$為員工基本業績，

$\beta_1$為單位B每單位年齡變動造成員工業績的影響數。

當$D_i = 1$*時代表單位A*，模式變為$Y_i = \beta_0 + \beta_1 x_i + \beta_2 D_i + \beta_3 X_i D_i + \varepsilon_i$ ，$\beta_0$為員工基本業績，

$\beta_1$為單位A每單位年齡變動造成員工業績的影響數。

$\beta_2$*代表每增加一單位年齡，所能增加單位A員工業績，*$\beta_3$*代表單位A與單位B平均員工基本業績差。*

三、在複迴歸模型診斷中，

（一）「某人對應變數（Y）做常態假設之檢定，發現非常態，故採Y之變數變換處理」，評論之。（3分）

（二）如何觀察各解釋變數與應變數之線性假設是否成立？（3分）

（三）VIF（變異數膨脹係數）值過大，表示為何？（3分）

（四）某$h_{ii} = 0.7$，表示為何？（3分）

**答**：

(一)若應變數不符合常態分配，則實務上會使用以下轉換：

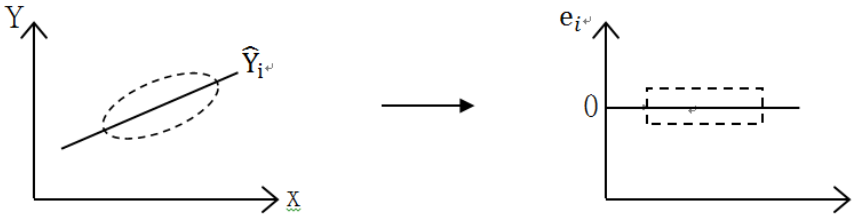
**Box − Cox Transformation**：

$$令\ T(\varepsilon) = \begin{cases} \frac{\varepsilon^{\lambda}-1}{\lambda} \ ， \text{if } \lambda \neq 0 \\ \ln\varepsilon \ ， \text{if } \lambda = 0 \end{cases} \quad \text{where：} \lim_{\lambda \to 0} \frac{\varepsilon^{\lambda}-1}{\lambda} = \ln\varepsilon \ ，求\lambda之MLE$$

當這些假設成立時，才開始進行參數的估計與檢定。

(二)若要「觀察」解釋變數與應變數之線性假設是否成立，則通常會以殘差圖進行分析

直線假設：modal：$Y = \beta_0 + \beta_1 x + \varepsilon$，若以0為中心線水平帶，正負都沒有規則性傾向，則線性假設應成立。




$x_i(或\hat{Y}_i)$

散佈圖　　殘差圖

(三)變異數膨脹係數=$VIF_j = \frac{1}{1-R_j^2}$，$1 \le VIF < \infty$，$R_j^2$代表為共線性模型之複判定係數，VIF越大代表模型存在共線性問題，會造成OLSE估計值的變異數膨脹，進而使估計產生誤差。

(四)$h_{ii}$:稱為槓桿值(亦為帽子矩陣對角線元素)，係用來衡量第ith個觀測值與預測值的距離。

槓桿值有兩個主要的性質：$0 \le h_{ii} \le 1$，$\sum h_{ii} = k - 1 \; where \; k = 參數個數$

又 $\text{cov}(\mathbf{e}) = \sigma^2(\text{I} - \text{H})$，$\text{H} = \text{X}(\text{X}^\text{T}\text{X})^{-1}\text{X}^\text{T}$，$\text{hii} = \text{X}_i(\text{X}^\text{T}\text{X})^{-1}X_i^T$，若hii=0.7則

依照t化殘差圖定義：$\text{t}_\text{i} = \frac{\text{e}_\text{i} - \bar{\text{e}}}{\text{S}_{\text{e}_\text{i}}} = \frac{\text{e}_\text{i}}{\text{S}_{\text{e}_\text{i}}}$　$(\because \sum \text{e}_\text{i} = 0 \Rightarrow \bar{\text{e}} = 0)$

Where $\text{S}_{\text{e}_\text{i}} = \sqrt{[1 - \frac{1}{\text{n}} - \frac{(\text{x}_\text{i} - \bar{\text{x}})^2}{\text{S}_{\text{XX}}}] \cdot \text{MSE}} = \sqrt{(1 - \text{h}_{\text{ii}}) \cdot \text{MSE}}$

也可知，當hii=0.7使得殘差的變異數變小，使得t化殘差變大，較容易造成離群值的狀況。

四、針對某公司員工，得一迴歸分析如下：

模型 $\text{Y} = \beta_0 + \beta_1\text{X1} + \beta_2\text{X2} + \beta_3\text{X3} + \beta_\text{M}\text{M} + \beta_\text{R}\text{R} + \varepsilon$。

Y＝量化工作績效，X1＝年資（以年為單位），X2＝薪資（以萬元為單位），X3＝所屬小組之人數，$(\text{M},\text{R}) = (1,0)$ 為行銷部門，$(\text{M},\text{R}) = (0,1)$ 為研發部門，$(\text{M},\text{R}) = (0,0)$ 為行政部門。

應變數：Y

使用的觀測值數目：70

$\text{SSTO} = 36.304$

$\text{SSR}(\text{X1}) = 14.288$, $\text{SSR}(\text{X2} \mid \text{X1}) = 0.676$, $\text{SSR}(\text{X3} \mid \text{X1}, \text{X2}) = 5.766$,

$\text{SSR}(\text{M} \mid \text{X1}, \text{X2}, \text{X3}) = 1.212$, $\text{SSR}(\text{R} \mid \text{X1}, \text{X2}, \text{X3}, \text{M}) = 0.357$

參數估計值

| 變數 | 參數估計值 | 標準誤差 | t值 | Pr > \|t\| |
|---|---|---|---|---|
| Intercept | 0.535 | 0.491 | 1.09 | 0.2793 |
| X1 | 0.029 | 0.006 | 4.84 | <.0001 |
| X2 | −0.180 | 0.107 | −1.68 | 0.0951 |
| X3 | 0.044 | 0.011 | 4.00 | 0.0001 |
| M | 0.433 | 0.163 | 2.66 | 0.0099 |
| R | 0.175 | 0.137 | 1.28 | 0.2061 |

(一)如薪資（X2）之單位由萬元改為千元，上述提供之資料有那些不會改變？有那些會改變？變化為何？（8分）

(二)在其他變數固定下，檢定薪資（X2）對Y之效果是否顯著為負？（$\alpha = 0.05$）（8分）

(三)$\text{H}_0 : \beta_0 = 0$ vs. $\text{H}_1 : \beta_0 \ne 0$ 之檢定（$\alpha = 0.05$），有何結論？又，是否要去除截距項？（敘

明理由）（8分）

（四）就檢定 $H_0:\beta_M=\beta_R=0$ vs. $H_1:$ not $H_0$，請算出F檢定統計量之值及寫出決策法則。（$\alpha=0.05$）（12分）

（五）求 $R^2$（coefficient of determination判定係數），並解釋其涵意。（8分）

（六）經變數選取後，得估計迴歸式為 $\hat{Y}=0.01+0.025X1+0.045X3+0.342M$，請分別解釋X3及M的係數估計值之涵意。（8分）

**答**：

(一)由於統計量不具單位性，故 T 值不變進而 P-Value 也不變。

關於參數估計值與標準誤差，試先算出標準迴歸係數(Beta 係數)

$$\text{Beta}(\hat{\beta}_2)=\frac{S_2}{S_y}\hat{\beta}_2=\frac{0.1070}{0.7254}\times-0.180=-0.026651$$

$$\text{Where } S_y=\sqrt{S_y^2}=\sqrt{\frac{\sum(y_i-\bar{y})^2}{n-1}}=\frac{36.304}{70-1}=\sqrt{0.526145}=0.725358$$

$$\hat{\beta}_2^*=-0.026651*\frac{0.725358}{1.07}=-0.018$$

$$\text{Where}\hat{S}_2^*=\sqrt{\frac{\sum(10X_2-10\overline{X_2})^2}{70-1}}=\sqrt{100}*S_y=10*S_y$$

故單位變換後，參數估計值為原本 0.1 倍，而標準誤差為原本之 10 倍。

總結單位變換後，參數估計值與標準誤差會改變，檢定 T 值與 P-Value 不變。

(二)

Step1: $\begin{cases}H_0:\beta_2=0\\H_1:\beta_2\neq0\end{cases}$

Step2:p-value=0.0951>0.05 → *在顯著水準* $\alpha=0.05$ *之下，*Do not reject $H_0$

(三)

Step1: $\begin{cases}H_0:\beta_0=0\\H_1:\beta_0\neq0\end{cases}$

Step2:p-value=0.2793>0.05 → *在* $\alpha=0.05$ *之下，*Do not reject $H_0$

故截距項在檢定水準 $\alpha=0.05$ 不具統計顯著性，建議刪除。

(四) 採偏 F-Test

Step: $\begin{cases}H_0:\beta_M=\beta_R=0\\H_1:\beta_M=\beta_R\neq0\end{cases}$

Step2: 決策法則：Reject $F>F_{\alpha=0.05}(2,70-6)\leftrightarrow Reject\ H_0$

Step3:F 檢定統計量=

$$\frac{SSE(Reduced)-SSE(Full)/2}{SSE(Full)/(70-6)}=\frac{(15.574-14.005)/2}{14.005/64}=3.5850$$

(五)

$$R^2=\frac{SSR}{SSTO}=\frac{\text{可解釋變異}}{\text{總變異}}=\frac{22.299}{36.304}=0.6142$$代表此模型可以解釋總變異的 61.42%

(六)

$\widehat{\beta_3}=0.045$代表在固定其他因素下，所屬小組人數每增加一人，量化工作績效可以增加 0.045 個單位。

$\widehat{\beta_M}=0.342$代表在固定其他因素下，行銷部門與非行銷部門量化工作績效平均值差異。