# 《迴歸分析》

參考之查表值: F分布α=0.05 臨界值$F_{0.05}$(df1, df2)。

| | df1 | |
|---|---|---|
| | 1 | 2 |
| df2 17 | 4.4513 | 3.5915 |
| 43 | 4.0670 | 3.2145 |
| 44 | 4.0617 | 3.2093 |
| 87 | 3.9506 | 3.1013 |
| 88 | 3.9493 | 3.1001 |

一、(一)何謂多重共線性（multicollinearity）？多重共線性對估計結果有何影響？如何偵測複迴歸模型中存在多重共線性？請詳述所需要的判斷準則。（12分）

(二)一位分析師進行迴歸分析資料並配適複迴歸模型如(1)。

$$Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_7 X_{7i} + \varepsilon_i, i = 1,...,100 \qquad (1)$$

所獲得初步結果如圖 1。請用圖 1 部分統計電腦套裝軟體輸出結果，說明這位分析師所配適的模型是否合適？如果模型(1)不合適，請說明原因並提供所有可以解決問題的方法。（10 分）

圖 1

| Analysis of Variance | | | | | |
|---|---|---|---|---|---|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 7 | 3019.404 | 431.343 | 78.26 | <0.0001 |
| Error | 92 | 507.065 | 5.512 | | |
| Corrected Total | 99 | 3526.470 | | | |

| Parameter Estimates | | | | | |
|---|---|---|---|---|---|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr >\| t \| |
| Intercept | 1 | -0.087 | 0.249 | -0.35 | 0.727 |
| X1 | 1 | 4.874 | 3.855 | 1.26 | 0.209 |
| X2 | 1 | 3.790 | 5.895 | 0.64 | 0.522 |
| X3 | 1 | 2.972 | 5.855 | 0.51 | 0.613 |
| X4 | 1 | 4.964 | 5.723 | 0.87 | 0.388 |
| X5 | 1 | 0.930 | 6.714 | 0.14 | 0.890 |
| X6 | 1 | -12.532 | 6.136 | -2.04 | 0.044 |
| X7 | 1 | 13.154 | 4.479 | 2.94 | 0.004 |

| | |
|---|---|
| **試題評析** | 本題有關線性重合之問題，與課本第六章第一節至第三節中所論述之內容完全一模一樣。 |
| **考點命中** | 《迴歸分析熱門題庫》，高點文化出版，趙治勳編著，第六章。 |

**答**：

(一)多重共線性指解釋變數(自變數)間存在著線性關係(或近似線性關係)

$$\lambda_1 X_{1i} + \lambda_2 X_{2i} + \cdots + \lambda_k X_{ki} + \delta_i = 0$$

多重共線性對估計如果之影響：

1.估計迴歸係數之標準誤較沒有多重共線性的時候大。

2.估計迴歸係數間之共變異數趨近無窮大。

多重共線性之偵測方法：

變異數膨脹因子(VIF)：當$VIF_j=\frac{1}{1-R_j^2}>10$表示模型可能有多重共線性之問題。

(二)該分析師所配適之複迴歸模型並不合適，原因是模型之$R^2=\frac{3019.404}{3526.470}=0.8562$很高且F檢定顯著，但大部份迴歸係數之t檢定卻不顯著，這就很有可能存在多重共線性之問題。可能之解決方法有：

1.增加樣本數。

2.多變量分析中主成份分析。

3.刪除產生多重共線性之自變數。

二、醫院分析師希望研究患者滿意度（$Y$）與患者年齡（$X_1$，以年為單位），疾病嚴重程度指數（$X_2$）以及焦慮指數（$X_3$）之間的關係。分析師隨機選擇了46名患者並收集了數據。請使用圖 2 部分統計電腦套裝軟體輸出結果來回答以下問題。

圖 2

**Dependent Variable:Y**

**Adjusted R-Square Selection Method**

| Number in Model | Adjusted R-Square | R-Square | SSE | Variables in Model |
|---|---|---|---|---|
| 2 | 0.6610 | 0.6761 | 4330.4997 | x1 x3 |
| 3 | 0.6595 | 0.6822 | 4248.8407 | x1 x2 x3 |
| 2 | 0.6389 | 0.6550 | 4613.0002 | x1 x2 |
| 1 | 0.6103 | 0.6190 | 5093.9155 | x1 |
| 2 | 0.4437 | 0.4685 | 7106.3941 | x2 x3 |
| 1 | 0.4022 | 0.4155 | 7814.3912 | x3 |
| 1 | 0.3491 | 0.3635 | 8509.0444 | x2 |
| SS Total=13369 | | | | |

(一)請計算額外平方和（extra sum of squares）$SSR\left(X_2\middle|X_1,X_3\right)=$？（4分）

(二)假設這位分析師採用模型是$Y_i=\beta_0+\beta_1X_{1i}+\beta_3X_{3i}+\varepsilon_i$（2）

該分析師想知道在模型(2)之下，增加疾病嚴重程度指數（$X_2$）此額外變數，解釋其在顯著水準$\alpha$=5%下是否有顯著的貢獻，並敘述對立假設、檢定統計量之值、決策法則和結論。（8分）

(三)假設這位分析師採用模型是$Y_i=\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+\beta_3X_{3i}+\varepsilon_i$（3）

請檢定疾病嚴重程度指數（$X_2$）和焦慮指數（$X_3$）兩個解釋變數是否可以從模型(3)中刪除，也就是在已經有患者年齡（$X_1$）解釋變數之下，解釋變數$X_2$和$X_3$可否從模型中移除？請在顯著水準$\alpha$=5% 檢定，並協助敘述對立假設、檢定統計量之值、決策法則和結論。在本小題的檢定問題中，請試述需要作何假設，才能執行這些統計檢定。（10分）

| 試題評析 | 本題為迴歸分析問題中之邊際檢定計算題，這種題型在國考中常出現，考生想必容易獲得高分。 |
|---|---|
| 考點命中 | 《迴歸分析熱門題庫》，高點文化出版，趙治勳編著，第三章第四節。 |

**答**：

(一)$SSR(X_2\mid X_1,X_3)=SSE(X_1,X_3)-SSE(X_1,X_2,X_3)$

$= 4330.4997 - 4248.8407 = 81.659$

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

T.S.: $F = \dfrac{SSR(X_2 \mid X_1, X_3)/1}{SSE(X_1, X_2, X_3)/42} \sim F_{(1,42)}$

R.R.: Reject $H_0$ at $\alpha = 0.05$ if $F^* > F_{(1,42)0.05} = 4.0727$

$\because F^* = \dfrac{81.659/1}{4248.8407/42} = 0.8072$ $\therefore$ don't reject $H_0$

我們沒有足夠證據去推論 $X_2$ 對模型之貢獻是顯著的

(三)假設 $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

$H_0 : \beta_2 = \beta_3 = 0$ vs $H_1$ : 至少一個$\beta_i \neq 0$

T.S.: $F = \dfrac{SSR(X_2, X_3 \mid X_1)/2}{SSE(X_1, X_2, X_3)/42} \sim F_{(2,42)}$

R.R.: Reject $H_0$ at $\alpha = 0.05$ if $F^* > F_{(2,42)0.05} = 3.2199$

$\because F^* = \dfrac{791.0748/2}{4248.8407/42} = 3.91$ $\therefore$ reject $H_0$

我們有足夠證據去推論 $X_1$ 對模型之貢獻是顯著的。

其中 $SSR(X_2, X_3 \mid X_1) = SSE(X_1) - SSE(X_1, X_2, X_3)$

$= 5039.9155 - 4248.8407 = 791.0748$

三、一位分析師考慮對三組數據配適一個簡單迴歸模型 $Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$ ，其中 $\beta_0$ 、 $\beta_1$ 為參數， $\varepsilon$ 為隨機誤差，且假設其為具均數0，標準差 $\sigma$ 之常態分配。

(一)配適模型後，三組數據之殘差分析圖分別為3(a)、3(b)、3(c)，請分別說明配適迴歸模型是否恰當？若模型不合適或偏離模型假設時，請指出不恰當之處並請提出修正的方法。（21分）

(二)在何種情況下，需要採用加權最小平方法（Weighted least squares）估計未知的參數？請協助提供散佈圖和殘差圖說明。（7分）

圖3(a) 標準化殘差時間序列圖（Standardized residuals vs. time）

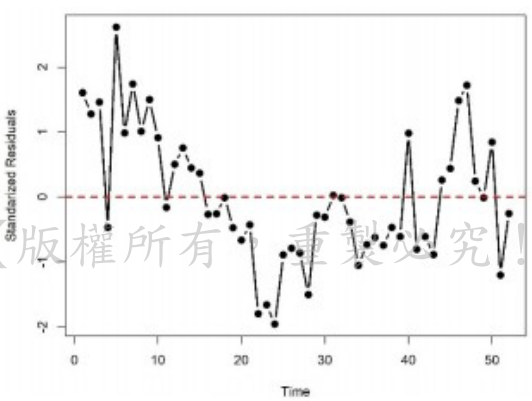


圖3(b) 標準化殘差對預測值圖（Standardized residuals vs. $\hat{y}_i$）

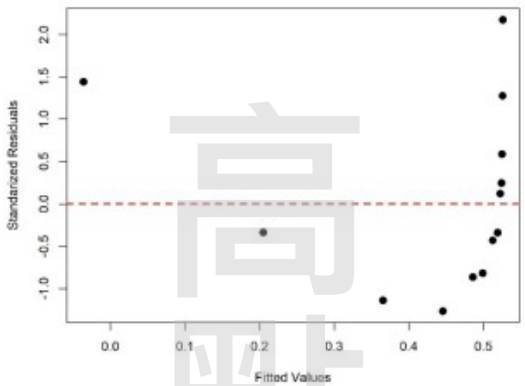
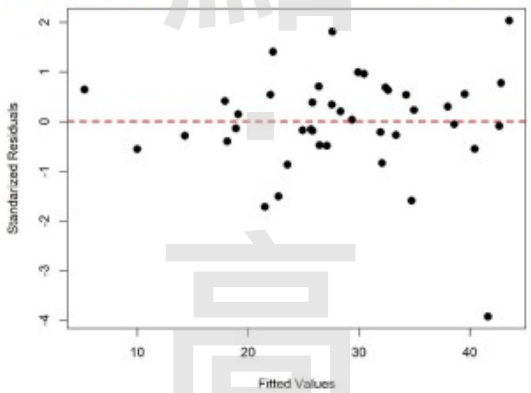



圖 3(c) 標準化殘差對預測值圖（Standardized residuals vs. $\hat{y}_i$）




| | |
|---|---|
| **試題評析** | 本題為殘差分析之相關問題，考古題已經出現過類似的題目。 |
| **考點命中** | 《迴歸分析熱門題庫》，高點文化出版，趙治勳編著，第四章。 |

**答**：

(一)圖3（ａ）中殘差在時間上具有規則性波動，可判斷模型違反獨立性假設，修正方法則是利用時間序列分析，建立ＡＲＩＭＡ模型。

圖3（ｂ）中殘差非隨機散佈在０之上下範圍內，可判斷Ｘ與Ｙ之關係為二次式，故該分析師之線性模型並不合適，修正方法則是建立模型為$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i$。

圖3（ｃ）中有一個點較遠，判斷這筆觀察值是離群值，修正方法則是刪除該筆樣本重新建模。

(二)加權最小平方法是在模型違反了變異數同質性假設時，其中一種修正模型之方法，例如：當殘差圖呈現喇叭狀時，模型有可能是$V(\varepsilon_i) = kX_i$。

四、一位資料分析師受託分析一組數據，想要了解一個特定基因，稱之GT基因，是否有影響老鼠斷奶時的重量。該分析師預計配適模型1和模型2。

Y=斷奶時的重量（公克為單位）

X1=年齡（以日為單位）

X2=品種（品種 A=1，B=0）

X3=GT 基因（有此基因=1，無此基因=0）

X4=性別（公老鼠=1，母老鼠=0）

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：$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{1i} X_{2i} + \varepsilon_i$,

請使用圖 4 和圖 5 中部分統計電腦套裝軟體輸出變異數分析（ANOVA, Analysis of Variance）回答下列問題：

(一)請計算模型1和模型2的調整的複判定係數$R^2$（the adjusted R-squared）。試述其意義，並判斷何種模型為佳。（8分）

(二)在顯著水準5%下，請檢定「GT 基因」在模型1中是否影響老鼠的重量？（4分）

(三)請解釋在考慮模型1下，請說明如何檢定老鼠的性別之兩條迴歸線是相同的迴歸線。並請列出虛無假設、對立假設、檢定統計量及決策法則。（4分）

(四)在顯著水準5%下，請檢定$X_{1i}$ $X_{2i}$相乘項在模型2中是否對解釋反應變數Y有顯著貢獻？請試述虛無假設、檢定統計量之值、決策法則和結論，以及所需要之假設。請解釋$X_{1i}$ $X_{2i}$該項在迴歸模型的意義。（12分）

圖 4 模型 1 的變異數分析

| Analysis of Variance | | | | | |
|---|---|---|---|---|---|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 4 | 2067.1901 | 516.798 | 74.40 | <.0001 |
| Error | 88 | 611.2830 | 6.946 | | |
| Corrected Total | 92 | 2678.4731 | | | |

| Parameter Estimates | | | | | |
|---|---|---|---|---|---|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
| Intercept | 1 | -3.5571 | 1.5904 | -2.24 | 0.0278 |
| X1 | 1 | 0.7971 | 0.0540 | 14.75 | <.0001 |
| X2 | 1 | 3.1411 | 0.5614 | 5.59 | <.0001 |
| X3 | 1 | -2.1286 | 0.5958 | -3.57 | 0.0006 |
| X4 | 1 | -3.1229 | 0.6234 | -5.01 | <.0001 |

圖 5 模型 2 的變異數分析

| Analysis of Variance | | | | | |
|---|---|---|---|---|---|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 5 | 2161.1224 | 432.224 | 72.68 | <.0001 |
| Error | 87 | 517.3508 | 5.947 | | |
| Corrected Total | 92 | 2678.4731 | | | |

| 試題評析 | 本題為虛擬變數之相關考題。 |
|---|---|
| 考點命中 | 《迴歸分析熱門題庫》，高點文化出版，趙治勳編著，第五章。 |

**答**：

(一)模型一：$R_1^2=1-\frac{SSE}{SST}=1-\frac{611.2830}{2678.4731}=0.7718$

模型二：$R_2^2=1-\frac{SSE}{SST}=1-\frac{517.3508}{2678.4731}=0.8068$

意義：經由自由度調查後，模型具有多少之解釋能力，其為越大越好。

判斷：以調整後判定係數為準則，模型二較為合適。

(二)

$H_0:\beta_3=0$ vs $H_1:\beta_3\neq 0$

T.S.: $T=\frac{\hat{\beta}_3-0}{S(\hat{\beta}_3)}\sim t_{(88)}$

R.R.: Reject $H_0$ at $\alpha=0.05$ if $\alpha>p-value$

Q $p-value=0.0006$ ∴ reject $H_0$

我們有足夠證據去推論ＧＴ基因顯著影響老鼠重量。

(三)

$H_0:\beta_4=0$ vs $H_1:\beta_4\neq 0$

T.S.: $T=\frac{\hat{\beta}_4-0}{S(\hat{\beta}_4)}\sim t_{(88)}$

R.R.: Reject $H_0$ at $\alpha=0.05$ if $\alpha>p-value$

Q $p-value<0.0001$ ∴ reject $H_0$

我們有足夠證據去推論老鼠性別之兩條迴歸線不為相同的。

(四)假設 $\varepsilon_i\overset{iid}{\sim}N(0,\sigma^2)$

令 $X_5=X_1X_2$

$H_0:\beta_5=0$ vs $H_1:\beta_5\neq 0$

T.S.: $F=\frac{SSR(X_5\mid X_1,X_2,X_3,X_4)/1}{SSE(X_1,X_2,X_3,X_4,X_5)/87}\sim F_{(1,87)}$

R.R.: Reject $H_0$ at $\alpha=0.05$ if $F^*>F_{(1,87)0.05}=3.9506$

Q $F^*=\frac{93.9323/1}{517.3508/87}=15.7961$ ∴ reject $H_0$

我們有足夠證據去推論 $X_1$ 與 $X_2$ 之交互作用項對應變數Ｙ具有顯著的貢獻。

其中 $SSR(X_5\mid X_1,X_2,X_3,X_4)=SSR(X_1,X_2,X_3,X_4,X_5)-SSR(X_1,X_2,X_3,X_4)$

$=2161.1224-2067.1901=93.9323$

$X_5=X_1X_2$ 之意義：

由於 $E(Y\mid X_1,X_2=0,X_3,X_4)=\beta_0+\beta_1X_1+\beta_3X_3+\beta_4X_4$

$E(Y\mid X_1,X_2=1,X_3,X_4)=(\beta_0+\beta_2)+(\beta_1+\beta_5)X_1+\beta_3X_3+\beta_4X_4$

可得知(1)當 $X_1=0, X_3=0, X_4=0$ 時品種A與品種B之老鼠重量相差 $\beta_2$ 。

(2)品種A下，老鼠年齡每增加一日對老鼠平均重量之影響為 $\beta_1+\beta_5$ 。

但在品種B，老鼠年齡每增加一日對老鼠平均重量之影響僅為 $\beta_1$ 。