《迴歸分析》

試題評析

本卷共四大題，考試範圍都沒有超過課程範圍。第一大題是迴歸係數之估計與推論題型，比較困難是第三小題，考生要有推導預測區間之能力才能獲得滿分。第二大題是迴歸之變異數分析，比較困難是第二小題缺適度檢定。第三大題要用到偏相關係數與複判定係數之觀念作答，此題型不難。第四大題是有關選模與選自變數之題目，不用計算，只要有熟記各指標之判斷準則就可以很快地作答。綜言之，今年考卷以75分作爲基本分。

一、若 $(y_i, x_i), i=1,2,...,n$ ，彼此獨立並滿足線性迴歸模型 $y_i=\beta_0+\beta_1 x_i+\varepsilon_i$ ，其中截距項 β_0 是已知的，ε_i 為i.i.d. $N(0,\sigma^2)$ 。

(一)試求 β_1 之最小平方估計量（least squares estimator）$\hat{\beta}_1$ 。（10分）

(二)試求 $E(\hat{\beta}_1)$ 和 $Var(\hat{\beta}_1)$ 。（10分）

(三)試求當另一獨立解釋變數之觀測值 $x=x_0$ 時，其反應變數 y_0 之 $(1-\alpha)100\%$ 的預測區間。（10分）

答：

(一) 目的: $\min\sum_{i=0}^{n} e_i^2 = \min\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2 = \min_{\hat{\beta}_1}\sum_{i=1}^{n}(Y_i-\beta_0-\hat{\beta}_1 X_i)^2$

$$\text{令}\ \frac{\partial\sum_{i=0}^{n} e_i^2}{\partial\hat{\beta}_1} = 2\sum_{i=1}^{n}(Y_i-\beta_0-\hat{\beta}_1 X_i)(-X_i)=0$$

$$\sum X_i^2\hat{\beta}_1 = \sum X_i Y_i - \sum X_i\beta_0 \quad \text{(正規方程式normal equation)}$$

$$\Rightarrow \hat{\beta}_1 = \frac{\sum X_i Y_i - \sum X_i\beta_0}{\sum X_i^2}$$

(二) $E(\hat{\beta}_1) = E(\frac{\sum X_i Y_i - \sum X_i\beta_0}{\sum X_i^2}) = \frac{\sum X_i E(Y_i) - \sum X_i\beta_0}{\sum X_i^2}$

$$= \frac{\sum X_i(\beta_0+\beta_1 X_i) - \sum X_i\beta_0}{\sum X_i^2} = \frac{\sum X_i\beta_0 + \sum X_i^2\beta_1 - \sum X_i\beta_0}{\sum X_i^2}$$

$$= \frac{\sum X_i^2\beta_1}{\sum X_i^2} = \beta_1$$

$$V(\hat{\beta}_1) = V(\frac{\sum X_i Y_i - \sum X_i\beta_0}{\sum X_i^2}) = \frac{\sum X_i^2 V(Y_i)}{[\sum X_i^2]^2} = \frac{\sum X_i^2\sigma^2}{[\sum X_i^2]^2} = \frac{\sigma^2}{\sum X_i^2}$$

(三) $E(\hat{Y}\mid x_0 - Y\mid x_0) = \beta_0+\beta_1 x_0 - (\beta_0+\beta_1 x_0) = 0$

$$V(\hat{Y}\mid x_0 - Y\mid x_0) = V(\hat{Y}\mid x_0) + V(Y\mid x_0)$$

$$= \frac{x_0^{\ 2}}{\sum X_i^{\ 2}}\sigma^2 + \sigma^2 = (1+\frac{x_0^{\ 2}}{\sum X_i^{\ 2}})\sigma^2$$

因此，機率區間爲

$$P(-t_{\frac{\alpha}{2},n-1} \le \frac{\hat{Y} \mid x_0 - Y \mid x_0 - 0}{\sqrt{\hat{V}(Y \mid x_0 - Y \mid x_0)}} \le t_{\frac{\alpha}{2},n-1}) = 1-\alpha$$

$$\Rightarrow P(-t_{\frac{\alpha}{2},n-1} \le \frac{\hat{\beta}_0 + \hat{\beta}_1 x_0 - Y \mid x_0}{\sqrt{(1+\frac{x_0^2}{\sum X_i^{\ 2}})MSE}} \le t_{\frac{\alpha}{2},n-1}) = 1-\alpha$$

因此，$(1-\alpha)100\%$ 之預測區間爲

$$(\hat{\beta}_0 + \hat{\beta}_1 x_0 \mp t_{\frac{\alpha}{2},n-1}\sqrt{(1+\frac{x_0^2}{\sum X_i^{\ 2}})MSE})$$

二、假設樣本資料$(y_i, x_i), i=1,2,...,n$，配適簡單線性迴歸模型$y_i = \beta_0 + \beta_1 x_1 + \varepsilon_i$，其中$\varepsilon_i$為i.i.d. $N(0,\sigma^2)$。若樣本數$n=20$，$\bar{x} = \sum_{i=1}^{n} x_i / n = 5$，$\bar{y} = 3$，$\sum_{i=1}^{n}(x_i - \bar{x})^2 = 160$，$\sum_{i=1}^{n}(y_i - \bar{y})^2 = 83.2$且$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = 80$。

(一)試寫出模型之ANOVA（analysis of variance）表。（10分）

(二)若解釋變數$x_i = 2, i = 1,2,3,4; x_i = 4, i = 5,...,8;\ x_i = 6, i = 9,...,12; x_i = 8, i = 13,...,16$及$x_i = 10, i = 17,...,20$，且其純誤差平方和（pure error sum of squares）為23.2。試問此時(一)中的模型是否仍恰當？請寫出檢定統計量之分布和自由度。（臨界值（critical value）=3.29。）。（15分）

答：

(一)根據題目已知 $SS_X = 160$, $SS_Y = 83.2$, $SS_{XY} = 80$

$$SST = SS_Y = 83.2,\ SSR = \hat{\beta}_1 SS_{XY} = \frac{(SS_{XY})^2}{SS_X} = 40,\ SSE = SST - SSR = 43.2$$

ANOVA TABLE				
source	SS	d.f.	MS	F
Reg	40	1	40	$F^* = 16.667$
Error	43.2	18	2.4	
Total	83.2	19		

(二)

ANOVA TABLE				
source	SS	d.f.	MS	F
Reg	40	1	40	
Error	43.2	18	2.4	
Lack of Fit	20	3	6.667	$F^* = 4.31$
Pure Error	23.2	15	1.547	
Total	83.2	19		

H_0：母體迴歸線爲直線 vs H_1：母體迴歸線不爲直線

T.S.: $F = \dfrac{MSLF}{MSPE} \sim F_{(3,15)}$

R.R.: Reject H_0 at α if $F^* > F_{(3,15)\alpha} = 3.29$

$\because F^* = 4.31 \quad \therefore reject H_0$

我們有足夠證據去推論(一)中之線性模型是適當的。

三、若 $(y_i, x_{i1}, x_{i2}, x_{i3}), i = 1,2,...20$ ，滿足迴歸模型

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i,$$

其中 ε_i 為i.i.d. $N(0,\sigma^2)$。且其 $SS_{\text{Re}g}$（sum of squares for regression）=19.0064，$SS_{\text{Res}} = 11.3792$， $R(\beta_1 \mid \beta_0) = 10.2070$ 和 $R(\beta_2 \mid \beta_0, \beta_1) = 6.2856$。試求：

（一）R^2 值。（5分）

（二）檢定上述迴歸模型是否顯著？（請寫出檢定統計量之分布和自由度。臨界值=3.24。）（10分）

（三）若檢定 $\beta_3 = 0$ 不顯著，試求此時 σ^2 之估計量。（5分）

答：

(一) $R^2 = \dfrac{SSR(X_1, X_2, X_3)}{SST(X_1, X_2, X_3)} = \dfrac{SSR(X_1, X_2, X_3)}{SSR(X_1, X_2, X_3) + SSE(X_1, X_2, X_3)}$

$= \dfrac{19.0064}{19.0064 + 11.3792} = 62.55\%$

(二)

ANOVA TABLE				
source	SS	d.f.	MS	F
Reg	19.0064	3	6.3355	$F^* = 8.908$
Error	11.3792	16	0.7112	
Total	30.3856	19		

H_0：模型是不適當的 vs H_1：模型是適當的

T.S.: $F = \dfrac{MSR}{MSE} \sim F_{(3,16)}$

R.R.: Reject H_0 at α if $F^* > F_{(3,16)\alpha} = 3.24$

$\because F^* = 8.908 \quad \therefore reject H_0$

我們有足夠證據去推論模型是適當的。

(三) $SSR(X_1, X_2) = SSR(X_2 \mid X_1) + SSR(X_1) = 6.2856 + 10.2070 = 16.4926$

$SST(X_1, X_2) = SST(X_1, X_2, X_3) = 30.3856$

ANOVA TABLE				
source	SS	d.f.	MS	F
Reg	16.4926	2	8.2463	$F^* = 10.0909$
Error	13.8930	17	0.8172	
Total	30.3856	19		

故 $\hat{\sigma}^2 = MSE = 0.8172$

四、下表為一可能有4個解釋變數 x_1, x_2, x_3 和 x_4 之資料，分別配適具左方之解釋變數之迴歸模型時其餘變數之partial F 值。（假設模型誤差為獨立同分佈的 $N(0,\sigma^2)$ 。）

模型中之解釋變數	不在模型中之解釋變數的partial F 值			
	x_1	x_2	x_3	x_4
–	12.60	21.96	4.40	22.80
x_1	–	208.58	0.31	159.30
x_2	146.52	–	11.82	0.42
x_3	5.81	36.68	–	100.36
x_4	108.22	0.17	40.29	–
x_1x_2	–	–	1.83	1.86
x_1x_3	–	220.55	–	208.24
x_1x_4	–	5.03	4.24	–
x_2x_3	68.72	–	–	41.65
x_2x_4	154.01	–	96.94	–
x_3x_4	22.11	12.43	–	–
$x_1x_2x_3$	–	–	–	0.04
$x_1x_2x_4$	–	–	0.02	–
$x_1x_3x_4$	–	0.5	–	–
$x_2x_3x_4$	4.34	–	–	–

考慮以 $F_{IN} = F_{OUT} = 4$ 試分別寫出下列方法所選出的解釋變數：

(一)逐步迴歸法（Stepwise Regression）。（5分）

(二)向前選擇法（Forward Selection）。（5分）

(三)後退消去法（Backward Elimination）。（5分）

(四)在下述各模型與所提供的訊息中選擇最佳的模型。（10分）

解釋變數	R^2	R^2_{adj}（adjusted R^2）	C_p	s
x_2	.666	.636	142.5	9.07
x_4	.675	.645	138.7	8.96
x_1x_2	.979	.974	2.7	2.41
x_1x_4	.972	.967	5.5	2.73
$x_1x_2x_4$	.982	.976	3.0	2.31
$x_1x_3x_4$	.981	.975	3.5	2.37

答：

(一) 逐步迴歸法 (以4作爲準則,大於4考慮,小於4刪除)
選擇: x1,x2,x3,x4,x1x3,x1x4,x2x3,x2x4,x3x4,x2x3x4

(二) 向前選擇法 (找最大的加入)
選擇: x1,x2,x3,x4,x1x3,x2x4

(三) 後退消去法 (找最小的刪除)
選擇: x1,x2,x3,x4,x1x3,x2x3,x2x4

(四) R^2, R^2_{adj} 越大越好, C_p 越接近p越好(p:模型中考慮變數個數)，s越小越好
取x1x2x4