《迴歸分析》

試題評析

本年度迴歸分析只考四題，第一題爲簡單迴歸，應屬送分題。第二題爲複迴歸；缺截距項，求OLSE與E($\hat{\beta}_1$)、V($\hat{\beta}_1$)的結果，只要將迴歸模型含截距項的結果中，$\overline{X}$、$\overline{Y}$ 代0即可。第三題是比較複迴歸之ANOVA F-test與偏F-test之差異性。中上程度同學，二、三題應該都可以拿到分數。至於第四題爲確認離群值之後，檢測離群值是否對迴歸模型具有影響力，屬於比較冷僻題型，公式難記、不易拿分。

一、今對一組樣本資料(x_i, y_i)，$i=1,\cdots\cdots 20$ 適配一簡單線性迴歸模型 $Y_i=\beta_0+\beta_1 X_i+\varepsilon_i$，其中 ε_i 為 i.i.d.$N(0,\sigma^2)$。已知 $F_{1,18,0.05}=3.01$，$\overline{x}=6$，$\overline{y}=15$，$\sum(x_i-\overline{x})^2=25$，$\sum(y_i-\overline{y})^2=208$，$\sum(x_i-\overline{x})(y_i-\overline{y})=40$。

(一)試寫此模型之變異數分析（Analysis of Variance）表。（10分）

(二)試求此一迴歸線之斜率與截距。（10分）

(三)試求此一迴歸線斜率β_1之90%信賴區間。（5分）

答：

(一)

$$SSTO=S_{YY}=208$$

$$SSR=\frac{S_{XY}^2}{S_{XX}}=\frac{40^2}{25}=64$$

$$SSE=SSTO-SSR=144$$

ANOVA table

變異來源	平方和	自由度	均方	F值
迴歸	64	1	64	8
誤差	144	18	8	
總和	208	19		

(二)

(1) 斜率$\hat{\beta}_1=\frac{S_{XY}}{S_{XX}}=\frac{40}{25}=1.6$

(2) 截距$\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}=15-1.6\times 6=5.4$

(三)

β_1 之 90% C.I.

$$=\hat{\beta}_1\pm t_{\frac{0.1}{2}}(20-2)\cdot\sqrt{\frac{MSE}{S_{XX}}}$$

$$=1.6\pm\sqrt{3.01}\times\sqrt{\frac{8}{25}}=1.6\pm 0.9814=(0.6186, 2.5814)$$

二、若(x_{1i}, x_{2i}, y_i)，$i=1,\cdots\cdots,n$ 彼此獨立且來自截距項為零的線性迴歸模型 $Y_i=\beta_1 X_{1i}+\beta_2 X_{2i}+\varepsilon_i$，其中 ε_i 為 i.i.d.$N(0,\sigma^2)$。

（一）試求 β_1 與 β_2 之最小平方估計量 $\hat{\beta}_1$ 與 $\hat{\beta}_2$。（15分）

（二）試求 $E(\hat{\beta}_1)$ 與 $Var(\hat{\beta}_2)$。（10分）

答：

(一)

令 $Q=\sum_{i=1}^{n}(Y_i-\hat{\beta}_1X_{1i}-\hat{\beta}_2X_{2i})^2$

$$\begin{cases}\dfrac{\partial Q}{\partial\hat{\beta}_1}=-2\sum_{i=1}^{n}\left[(Y_i-\hat{\beta}_1X_{1i}-\hat{\beta}_2X_{2i})\cdot X_{1i}\right]=0\\ \dfrac{\partial Q}{\partial\hat{\beta}_2}=-2\sum_{i=1}^{n}\left[(Y_i-\hat{\beta}_1X_{1i}-\hat{\beta}_2X_{2i})\cdot X_{2i}\right]=0\end{cases}$$

$$\begin{cases}\sum_{i=1}^{n}X_{1i}^2\cdot\hat{\beta}_1+\sum_{i=1}^{n}X_{1i}X_{2i}\cdot\hat{\beta}_2=\sum_{i=1}^{n}X_{1i}Y_i\\ \sum_{i=1}^{n}X_{1i}X_{2i}\cdot\hat{\beta}_1+\sum_{i=1}^{n}X_{2i}^2\cdot\hat{\beta}_2=\sum_{i=1}^{n}X_{2i}Y_i\end{cases}$$

$$\hat{\beta}_1=\frac{\begin{vmatrix}\sum_{i=1}^{n}X_{1i}Y_i & \sum_{i=1}^{n}X_{1i}X_{2i}\\ \sum_{i=1}^{n}X_{2i}Y_i & \sum_{i=1}^{n}X_{2i}^2\end{vmatrix}}{\begin{vmatrix}\sum_{i=1}^{n}X_{1i}^2 & \sum_{i=1}^{n}X_{1i}X_{2i}\\ \sum_{i=1}^{n}X_{1i}X_{2i} & \sum_{i=1}^{n}X_{2i}^2\end{vmatrix}}$$

$$=\frac{(\sum_{i=1}^{n}X_{2i}^2)(\sum_{i=1}^{n}X_{1i}Y_i)-(\sum_{i=1}^{n}X_{1i}X_{2i})(\sum_{i=1}^{n}X_{2i}Y_i)}{(\sum_{i=1}^{n}X_{1i}^2)(\sum_{i=1}^{n}X_{2i}^2)-(\sum_{i=1}^{n}X_{1i}X_{2i})^2}$$

同理

$$\hat{\beta}_2=\frac{(\sum_{i=1}^{n}X_{1i}^2)(\sum_{i=1}^{n}X_{2i}Y_i)-(\sum_{i=1}^{n}X_{1i}X_{2i})(\sum_{i=1}^{n}X_{1i}Y_i)}{(\sum_{i=1}^{n}X_{1i}^2)(\sum_{i=1}^{n}X_{2i}^2)-(\sum_{i=1}^{n}X_{1i}X_{2i})^2}$$

(二)

$$(1)\quad E\hat{\beta}_1=E\left[\frac{(\sum_{i=1}^{n}X_{2i}^2)(\sum_{i=1}^{n}X_{1i}Y_i)-(\sum_{i=1}^{n}X_{1i}X_{2i})(\sum_{i=1}^{n}X_{2i}Y_i)}{(\sum_{i=1}^{n}X_{1i}^2)(\sum_{i=1}^{n}X_{2i}^2)-(\sum_{i=1}^{n}X_{1i}X_{2i})^2}\right]$$

$$= \frac{(\sum_{i=1}^{n} X_{2i}^2)(\sum_{i=1}^{n} X_{1i}EY_i) - (\sum_{i=1}^{n} X_{1i}X_{2i})(\sum_{i=1}^{n} X_{2i}EY_i)}{(\sum_{i=1}^{n} X_{1i}^2)(\sum_{i=1}^{n} X_{2i}^2) - (\sum_{i=1}^{n} X_{1i}X_{2i})^2}$$

$$\frac{(\sum_{i=1}^{n} X_{2i}^2)(\beta_1\sum_{i=1}^{n} X_{1i}^2 - \beta_2\sum_{i=1}^{n} X_{1i}X_{2i}) - (\sum_{i=1}^{n} X_{1i}X_{2i})(\beta_1\sum_{i=1}^{n} X_{1i}X_{2i} - \beta_2\sum_{i=1}^{n} X_{2i}^2)}{(\sum_{i=1}^{n} X_{1i}^2)(\sum_{i=1}^{n} X_{2i}^2) - (\sum_{i=1}^{n} X_{1i}X_{2i})^2}$$

$$= \frac{\beta_1\left[(\sum_{i=1}^{n} X_{2i}^2)(\sum_{i=1}^{n} X_{1i}^2) - (\sum_{i=1}^{n} X_{1i}X_{2i})^2\right]}{(\sum_{i=1}^{n} X_{1i}^2)(\sum_{i=1}^{n} X_{2i}^2) - (\sum_{i=1}^{n} X_{1i}X_{2i})^2}$$

$$= \beta_1$$

(2) $V(\hat{\beta}_1) = V\left[\frac{(\sum_{i=1}^{n} X_{2i}^2)(\sum_{i=1}^{n} X_{1i}Y_i) - (\sum_{i=1}^{n} X_{1i}X_{2i})(\sum_{i=1}^{n} X_{2i}Y_i)}{(\sum_{i=1}^{n} X_{1i}^2)(\sum_{i=1}^{n} X_{2i}^2) - (\sum_{i=1}^{n} X_{1i}X_{2i})^2}\right]$

$$= \frac{(\sum_{i=1}^{n} X_{2i}^2)^2(\sigma^2\sum_{i=1}^{n} X_{1i}^2) - (\sum_{i=1}^{n} X_{1i}X_{2i})^2(\sigma^2\sum_{i=1}^{n} X_{2i}^2)}{\left[(\sum_{i=1}^{n} X_{1i}^2)(\sum_{i=1}^{n} X_{2i}^2) - (\sum_{i=1}^{n} X_{1i}X_{2i})^2\right]^2}$$

$$= \frac{\sigma^2 \cdot \sum_{i=1}^{n} X_{2i}^2\left[(\sum_{i=1}^{n} X_{2i}^2)(\sum_{i=1}^{n} X_{1i}^2) - (\sum_{i=1}^{n} X_{1i}X_{2i})^2\right]}{\left[(\sum_{i=1}^{n} X_{1i}^2)(\sum_{i=1}^{n} X_{2i}^2) - (\sum_{i=1}^{n} X_{1i}X_{2i})^2\right]^2}$$

$$= \frac{\sum_{i=1}^{n} X_{2i}^2}{(\sum_{i=1}^{n} X_{1i}^2)(\sum_{i=1}^{n} X_{2i}^2) - (\sum_{i=1}^{n} X_{1i}X_{2i})^2}\ \sigma^2$$

三、某一研究想要知道房屋的價格Y與房屋的坪數X_1，屋齡X_2，房間數X_3與空屋率X_4的關係。今收集30間房屋的資料並對此資料配適一迴歸模型 $Y_i=\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+\beta_3X_{3i}+\beta_4X_{4i}+\varepsilon_i$，其中$\varepsilon_i$為 i.i.d.$N(0,\sigma^2)$。已知 $F_{2,27,0.05}=3.35$，$F_{2,25,0.05}=3.39$。

Source of Variation(變異來源)	SS	df	MS
$SSR(X_1,X_2,X_3,X_4)$	300	4	75
$SSR(X_3,X_4,\|X_1,X_2)$	40	2	20
$SSR(X_3,X_4)$	150	2	75
$SSE(X_1,X_2,X_3,X_4)$	200	25	8

(一)假設迴歸模型中僅考慮房間數X_3與空屋率X_4。試就此模型在$\alpha=0.05$下檢定$H_0:\beta_3=\beta_4=0$。（請務必將完整之檢定寫出，包括H_0，H_1，檢定量，拒絕區域，結論等）（10分）

(二)假設迴歸模型中已考慮坪數X_1與屋齡X_2。試就此模型在$\alpha=0.05$下檢定$H_0:\beta_3=\beta_4=0$。（請務必將完整之檢定寫出，包括H_0，H_1，檢定量，拒絕區域，結論等）（10分）

(三)試解釋上面兩小題結果不盡相同之原因。（5分）

答：

(一)model：$Y_i=\beta_0+\beta_3X_{3i}+\beta_4X_{4i}+\varepsilon_i$

$$SSTO=SSR(X_1X_2X_3X_4)+SSE(X_1X_2X_3X_4)$$
$$=300+200=500$$
$$SSE(X_1X_2)=SSTO-SSR(X_1X_2)$$
$$=500-150=350$$

1.檢定假設

$$\begin{cases}H_0:\beta_3=\beta_4=0\\ H_1:\text{不全爲}0\end{cases}$$

2.拒絕區域

$$C\{F|F>F_{0.05}(2,30-3)=3.35\}$$

3.檢定統計值

$$F=\frac{SSR(X_1X_2)/2}{SSE(X_1X_2)/(30-3)}=\frac{75}{350/27}=5.78\in C$$

∴ reject H_0, 有充分證據顯示β_3,β_4不全爲0

(二)model：$Y_i=\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+\beta_3X_{3i}+\beta_4X_{4i}+\varepsilon_i$

1.檢定假設

$$\begin{cases}H_0:\beta_3=\beta_4=0\\ H_1:\text{不全爲}0\end{cases}\Leftrightarrow\begin{cases}H_0:\text{不值得引進 }X_3,X_4\\ H_1:\text{值得引進 }X_3,X_4\end{cases}$$

2.拒絕區域

$$C\{F|F>F_{0.05}(2,30-5)=3.39\}$$

3.檢定統計值

$$F=\frac{SSR(X_3X_4|X_1X_2)/2}{SSE(X_1X_2X_3X_4)/(30-5)}=\frac{20}{8}=2.5\notin C$$

∴ Not reject H_0, 無充分證據顯示β_3,β_4不全爲0

(三)

1.在(一)中，表示房屋的價格可以用房間數(X_3)與空屋率(X_4)來預測

2.在(二)中，表示當model已含有 X_1 (坪數)， X_2 (屋齡)兩個自變數時，不值得引進 X_3 (房屋數)與 X_4 (空屋率)當Y之預測變數。

∴(一)採用迴歸ANOVA F-test與(二)偏F檢定結果未必相同

四、在模型診斷時，我們常用DFFITS, Cook's Distance，DFBETAS方法辨認具有影響力的個案（Influential cases）。

(一)試比較DFFITS, Cook's Distance，DFBETAS此三種方法之差異。（15分）

(二)試說明此三種方法辨認具有影響力的個案之判定原則。（10分）

答：

(一)

1. DFFITS：剔除ith個離群值後，對單一適配值（*fitted value*）的影響。

其中， $(DFFITS)_i = \dfrac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)}h_{ii}}}$

2.Cook's Distance：剔除ith個離群值後，對所有適配值（*fitted value*）的影響。

其中， $D_i = \dfrac{\sum_{j=1}^{n}\left(\hat{Y}_j - \hat{Y}_{j(i)}\right)^2}{pMSE}$

3.DFBETAS：剔除ith個離群值後，對於每一個迴歸係數（*regression coefficients*）的影響。

其中， $\left(DFBETAS\right)_{k(i)} = \dfrac{b_k - b_{k(i)}}{\sqrt{MSE_{(i)}c_{kk}}} \qquad k = 0,1,\cdots,p-1$

(二)

1.DFFITS判定原則：

對於中、小規模的資料集而言，當 $\left|DFFITS\right| > 1$ ；或者是，大規模資料集而言， $\left|DFFITS\right| > 2\sqrt{\dfrac{p}{n}}$ ，則可視爲離群值具有影響力。

2.Cook's Distance判定原則：

Cook's Distance中， D_i 與F(*p,n-p*)所對應的百分位數值來確認個別的影響。當百分位數值小於10%或是20%時，表示第*i*個個案對於適配值影響不大；當百分位數值大於50%時，表示第*i*個個案對於適配迴歸函數具有影響。

3.DFBETAS判定原則：

對於中、小規模的資料集而言，當 $DFBETAS > 1$ ；或者是，大規模資料集而言，

$DFBETAS > \dfrac{2}{\sqrt{n}}$ ，則可視爲離群值具有影響力。