Simple Linear Regression

~ Case study ~

簡單迴歸分析

簡單迴歸分析
（線性迴歸）

simple regression analysis

樣本迴歸線之係數

假設 $y = \hat{\beta}_0 + \hat{\beta}_1 x$ 為樣本迴歸線

$$\Rightarrow \begin{cases} \hat{\beta}_1 = \dfrac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \dfrac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \dfrac{S_{xy}}{S_x^2} \\ \\ \hat{\beta}_0 = \dfrac{\sum y_i \sum x_i^2 - \sum x_i \sum x_i y_i}{n\sum x_i^2 - (\sum x_i)^2} \end{cases}$$

其中 S_{xy} 為X與Y的共變異數
S_x^2 為樣本資料X所求得的變異數

實務上，$\hat{\beta}_1$比較好求與好記，所以會先計算出，從式(1)中可推得：$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$ ，所以計算出$\bar{x}$、$\bar{y}$ 後，即可求得$\hat{\beta}_0$。

$\Rightarrow$ $\hat{\beta}_0$，$\hat{\beta}_1$ 滿足常態分配：

$$\hat{\beta}_0 \sim N\left(\beta_0, \frac{\sum x_i^2}{n\sum(x_i - \bar{x})^2}\sigma^2\right);\quad \hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum(x_i - \bar{x})^2}\right)。$$

案例1

隨機抽樣5個人，得其身高與體重資料如下所示：

身高（y）	162	170	158	172	166
體重（x）	62	75	50	56	70

請問此身高與體重組合的最佳線性迴歸方程式為何？

「應用統計學 二版」p391，李德治、童惠玲，博碩文化

案例1 解說

先求斜率，根據五組數據，可得：

身高 (y)	162	170	158	172	166	$\bar{y}$ = 165.6
體重 (x)	62	75	50	56	70	$\bar{x}$ = 62.6
$x_i y_i$	10044	12750	7900	9632	11620	Sum= 51946
x_i^2	3844	5625	2500	3136	4900	Sum= 20005

➔ $$\hat{\beta}_1 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{51946 - 5 \times 62.6 \times 165.6}{20005 - 5 \times 62.6^2} = 0.2753$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 165.6 - 0.2753 \times 62.6 = 148.366$$

➔ 線性迴歸方程式為 $\mathrm{y} = 148.366 + 0.2753\,\mathrm{x}$

案例2

假設下列資料為某公司近五年的投資金額：

年度(x)	1	2	3	4	5
投資金額(y)	1	1	3	4	6

(1)試求迴歸方程式？

(2)請預測此公司第7年度的投資金額大約是多少？

「應用統計學 二版」p392，李德治、童惠玲，博碩文化

案例2 解說

先求斜率，根據五組數據，可得：

年度(x)	1	2	3	4	5	$\bar{x}$ =	3
投資金額(y)	1	1	3	4	6	$\bar{y}$ =	3
$x_i y_i$	1	2	9	16	30	Sum=	58
x_i^2	1	4	9	16	25	Sum=	55

(1) $$\hat{\beta}_1 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{58 - 5 \times 3 \times 3}{55 - 5 \times 3^2} = 1.3$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 3 - 1.3 \times 3 = -0.9$$

➔ 線性迴歸方程式為 $\mathrm{y} = -0.9 + 1.3\ \mathrm{x}$

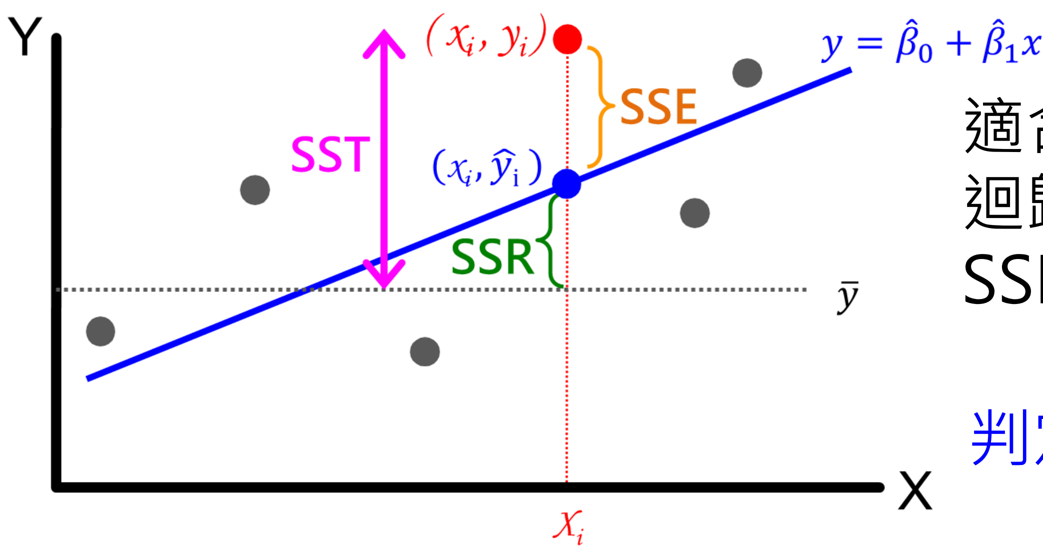
(2) 當 x=7時（第7年），預測的投資金額為：

$$\mathrm{y} = -0.9 + 1.3\ \mathrm{x} = -0.9 + 1.3 \times 7 = 8.2$$（千元）

迴歸模型之配適度檢定1

衡量迴歸方程式的解釋能力(判定係數)
也客觀檢定其適合度(F檢定)

判定係數 R^2



適合度：希望觀測的值都落在迴歸線上。所以SSE越小越好；SSR越大越好。

判定係數 $R^2 = \dfrac{SSR}{SST} = 1 - \dfrac{SSE}{SST}$

$$\Rightarrow \quad R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{\hat{\beta}_1^2 \sum(\hat{x}_i - \bar{x})^2}{\sum(y_i - \bar{y})^2} = \frac{\hat{\beta}_1^2\left(\sum x_i^2 - n\bar{x}^2\right)}{\sum y_i^2 - n\bar{y}^2}$$

$$= \hat{\beta}_1^2 \frac{S_x^2}{S_y^2} = \frac{S_{\hat{y}}^2}{S_y^2}$$

R^2越大，表示迴歸模型的解釋能力越強，配適度越大

※課本的Coefficient of Determination是以「r」表示。

迴歸模型的 F 檢定

迴歸模型的檢定，可以利用F檢定來做：

$$\begin{cases} H_0: \beta_1 = 0 \text{ (迴歸方程式不具有解釋力)} \\ H_1: \beta_1 \neq 0 \text{ (迴歸方程式具有解釋力)} \end{cases}$$

（或x不可解釋y）

變異來源	平方和	自由度	平均平方和	F值
迴歸	SSR	1	MSR=SSR/1	
隨機	SSE	n-2	MSE=SSE/(n-2)	$F^*=\frac{MSR}{MSE}$
總和	SST	n-1		

決策法則：

$F^* > F_{\alpha,1,n-2}$，則拒絕H_0

$$F^* = \frac{MSR}{MSE} = \frac{R^2/1}{(1-R^2)/(n-2)}$$

案例3

某飲料公司欲知各商店所裝設的自動販賣機數 X 與每個月所販賣的罐裝飲料數 Y 間的關係，隨機選取8家商店，其資料如下：

販賣機數(x)	1	1	1	2	4	4	5	6
飲料罐數(y)	568	577	652	657	755	759	840	832

(1)試求迴歸方程式？

(2)在顯著水準0.05下，請檢定此迴歸線是否適合（是否具有代表性）？

「應用統計學 二版」p400，李德治、童惠玲，博碩文化

案例3 解說(a)

先求斜率，根據8組數據，可得：

販賣機數(x)	1	1	1	2	4	4	5	6	$\bar{x}$ =	3
飲料罐數(y)	568	577	652	657	755	759	840	832	$\bar{y}$ =	705
$x_i y_i$	568	577	652	1314	3020	3036	4200	4992	Sum=	18359
x_i^2	1	1	1	4	16	16	25	36	Sum=	100
y_j^2	322624	332929	425104	431649	570025	576081	705600	692224	Sum=	4056236

(1) $$\hat{\beta}_1 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{18359 - 8 \times 3 \times 705}{100 - 8 \times 3^2} = 51.393$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 705 - 51.393 \times 3 = 550.82$$

➔ 線性迴歸方程式為 $\mathrm{y} = 550.82 + 51.393\ \mathrm{x}$

案例3 解說(b)

(2) 建立假設檢定：$\begin{cases} H_0: \beta_1 = 0 \text{ (迴歸方程式不具有解釋力)} \\ H_1: \beta_1 \neq 0 \text{ (迴歸方程式具有解釋力)} \end{cases}$

$$SST = \sum y_i^2 - n\bar{y}^2 = 4056236 - 8 \times 705^2 = 80036$$

$$SSR = \hat{\beta}_1^2 \left[\sum x_i^2 - n\bar{x}^2\right] = 51.393^2\left(100 - 5 \times 3^2\right) = 73954$$

SSE＝SST – SSR=6082

變異來源	平方和	自由度	平均平方和	F值
迴歸	73954	1	MSR=SSR/1=73954	$F^* = \frac{MSR}{MSE} = \frac{73954}{1014} = 72.96$
隨機	6082	6	MSE=SSE/(n-2)=1014	
總和	80036	7		

因為 $F^* = 72.96 > F_{0.05, 1, 6} = 5.99$，所以拒絕 H_0，表示此迴歸線適合，X 對 Y 具有解釋力。

迴歸模型之
配適度檢定2

檢定斜率與截距的適合度

斜率項 β_1 的檢定

雙尾檢定

$$\begin{cases} H_0：\beta_1 = 0 \text{ (迴歸方程式不具有解釋力)} \\ H_1：\beta_1 \neq 0 \text{ (迴歸方程式具有解釋力)} \end{cases}$$

已知 $\hat{\beta}_1 \sim N\left(\beta_1, \dfrac{\sigma^2}{\sum(x_i - \bar{x})^2}\right)$ ，因母體變異數σ^2未知，依照

假設檢定方法，可用 t分配做檢定：

$$t^* = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{\sqrt{\dfrac{MSE}{\sum(x_i - \bar{x})^2}}}$$

決策法則： $|t^*| > t_{\alpha/2,\, n-2}$ 時，則拒絕虛無假設 H_0；表示迴歸方程式具適配度，或自變數X可以解釋變數Y

斜率項 β_1 的信賴區間

也因為 $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum(x_i - \bar{x})^2}\right)$ ，且母體變異數σ^{2}未知，可用

t分配做區間估計（信賴水準1-α）：

$$t^* = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{\sqrt{\frac{MSE}{\sum(x_i - \bar{x})^2}}}$$

以t分配做區間估計，所以1-α的信賴區間為：

$$\hat{\beta}_1 - t_{\frac{\alpha}{2},n-2}\sqrt{\frac{MSE}{\sum(x_i - \bar{x})^2}} \le \beta_1 \le \hat{\beta}_1 + t_{\frac{\alpha}{2},n-2}\sqrt{\frac{MSE}{\sum(x_i - \bar{x})^2}}$$

斜率項 β_1 的檢定

右尾檢定

$$\begin{cases} H_0: \beta_1 \le 0 \text{ (自變數對依變數不具正向影響力)} \\ H_1: \beta_1 > 0 \text{ (自變數對依變數具正向影響力)} \end{cases}$$

已知 $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum(x_i - \bar{x})^2}\right)$ ，母體變異數σ^2未知，依照假設

檢定方法，可用 t分配做檢定：

$$t^* = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{\sqrt{\frac{MSE}{\sum(x_i - \bar{x})^2}}}$$

決策法則： $t^* > t_{\alpha, n-2}$ 時，則拒絕虛無假設 H_0；表示迴歸方程式具適配度，或自變數X可以解釋變數Y

斜率項 β_1 的檢定

左尾檢定

$$\begin{cases} \mathrm{H}_0：\beta_1 \geq 0 \text{ (自變數對依變數不具負向影響力)} \\ \mathrm{H}_1：\beta_1 < 0 \text{ (自變數對依變數具負向影響力)} \end{cases}$$

已知 $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum(x_i - \bar{x})^2}\right)$ ，母體變異數σ²未知，依照假設

檢定方法，可用 t分配做檢定：

$$\mathrm{t}^* = \frac{\hat{\beta}_1 - \beta_1}{\mathrm{S}_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{\sqrt{\frac{\mathrm{MSE}}{\sum(\mathrm{x_i} - \bar{\mathrm{x}})^2}}}$$

決策法則： $t^* < -t_{\alpha, n-2}$ 時，則拒絕虛無假設 H_0；表示自變數X對依變數Y有顯著的負面影響。

截距項 β_0 的檢定

雙尾檢定

$$\begin{cases} \mathrm{H_0}：\beta_0 = 0 \text{ (迴歸方程式沒有通過原點)} \\ \mathrm{H_1}：\beta_0 \neq 0 \text{ (迴歸方程式沒有通過原點)} \end{cases}$$

已知 $\hat{\beta}_0 \sim N\left(\beta_0, \dfrac{\sum x_i^2}{n\sum(x_i-\bar{x})^2}\,\sigma^2\right)$，母體變異數$\sigma^2$未知，依照假設檢定方法，可用 t分配做檢定：

$$\mathrm{t}^* = \frac{\hat{\beta}_0 - \beta_0}{\mathrm{S}_{\hat{\beta}_0}} = \frac{\hat{\beta}_0}{\sqrt{\dfrac{\sum x_i^2}{n} \times \dfrac{MSE}{\sum(x_i-\bar{x})^2}}}$$

決策法則： $|t^*| > t_{\alpha/2,\, n-2}$ 時，則拒絕虛無假設 H_0；表示迴歸方程式沒有通過原點。

斜率項 β_0的信賴區間

也因為 $\hat{\beta}_0 \sim N\left(\beta_0, \frac{\sum x_i^2}{n\sum(x_i-\bar{x})^2}\sigma^2\right)$，母體變異數σ²未知，可用

t分配做區間估計（信賴水準1-α）：

$$\mathrm{t}^* = \frac{\hat{\beta}_0 - \beta_0}{\mathrm{S}_{\hat{\beta}_0}} = \frac{\hat{\beta}_0}{\sqrt{\frac{\sum x_i^2}{n} \times \frac{MSE}{\sum(x_i-\bar{x})^2}}}$$

以t分配做區間估計，所以1-α的信賴區間為：

$$\hat{\beta}_0 - \mathrm{t}_{\frac{\alpha}{2},\mathrm{n}-2}\sqrt{\frac{\sum x_i^2}{n} \times \frac{MSE}{\sum(x_i-\bar{x})^2}} \leq \beta_0 \leq \hat{\beta}_0 + \mathrm{t}_{\frac{\alpha}{2},\mathrm{n}-2}\sqrt{\frac{\sum x_i^2}{n} \times \frac{MSE}{\sum(x_i-\bar{x})^2}}$$

案例4

某飲料公司想瞭解廣告費用（x）與飲料的銷售量（y）之間的關係，於是進行實驗，每個月打一次廣告，總共進行十個月的實驗，並記錄每個月的飲料銷售量，經過整理資料如下（n=10）：

$$\sum x_i = 28，\quad \sum x_i^2 = 303.4，$$

$$\sum y_i = 75，\quad \sum y_i^2 = 598.5，\quad \sum x_i y_i = 237$$

(1)試求迴歸方程式。

(2)試求判定係數。

(3)在顯著水準0.05下，廣告的花費對飲料的銷售量成正向影響？

(4)試求 β_1 的95%信賴區間。

「應用統計學 二版」p403，李德治、童惠玲，博碩文化

案例4 解說(a)

(1) $\bar{x} = \frac{\sum x}{n} = \frac{28}{10} = 2.8$ ， $\bar{y} = \frac{\sum y}{n} = \frac{75}{10} = 7.5$

$$\hat{\beta}_1 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{237 - 10 \times 2.8 \times 7.5}{303.4 - 10 \times 2.8^2} = 0.12$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 7.5 - 0.12 \times 2.8 = 7.164$$

➔ 線性迴歸方程式為 $\mathrm{y} = 7.164 + 0.12\ \mathrm{x}$

(2)判定係數 $R^2 = \frac{\widehat{\beta}_1^2\left(\sum x_i^2 - n\bar{x}^2\right)}{\sum y_i^2 - n\bar{y}^2} = \frac{0.12^2\left(303.4 - 10 \times 2.8^2\right)}{598.5 - 10 \times 7.5^2} = 0.3$

案例4 解說(b)

(3)依題意，相當檢定：
$$\begin{cases} H_0：\beta_1 \le 0 \text{ (x對y不具正向影響力)} \\ H_1：\beta_1 > 0 \text{ (x對y具正向影響力)} \end{cases}$$

$$SSE = \sum y_i^2 - \hat{\beta}_0^2 \sum y_i - \hat{\beta}_1^2 \sum x_i y_i$$

$$= 598.5 - 7.164 \times 7 - 0.12 \times 237 = 32.76$$

$$\mathrm{M}SE = \frac{\mathrm{SSE}}{\mathrm{n}-2} = \frac{32.76}{10-2} = 4.095$$

➔ $$t^* = \frac{\hat{\beta}_1}{\sqrt{\frac{\mathrm{MSE}}{\sum(\mathrm{x_i} - \bar{\mathrm{x}})^2}}} = \frac{\hat{\beta}_1}{\sqrt{\frac{\mathrm{MSE}}{\sum x_i^2 - n\bar{x}^2}}} = \frac{0.12}{\sqrt{\frac{4.095}{303.4 - 10 \times 2.8^2}}} = 0.984$$

因為 $t^*=0.984 < -t_{0.05,\,8} = 1.86$，接受虛無假設 H_0；表示廣告的花費對飲料的銷售量，無足夠證據證明具正向影響。

案例4 解說(c)

(4) β_1 在顯著水準0.05下的信賴區間為：

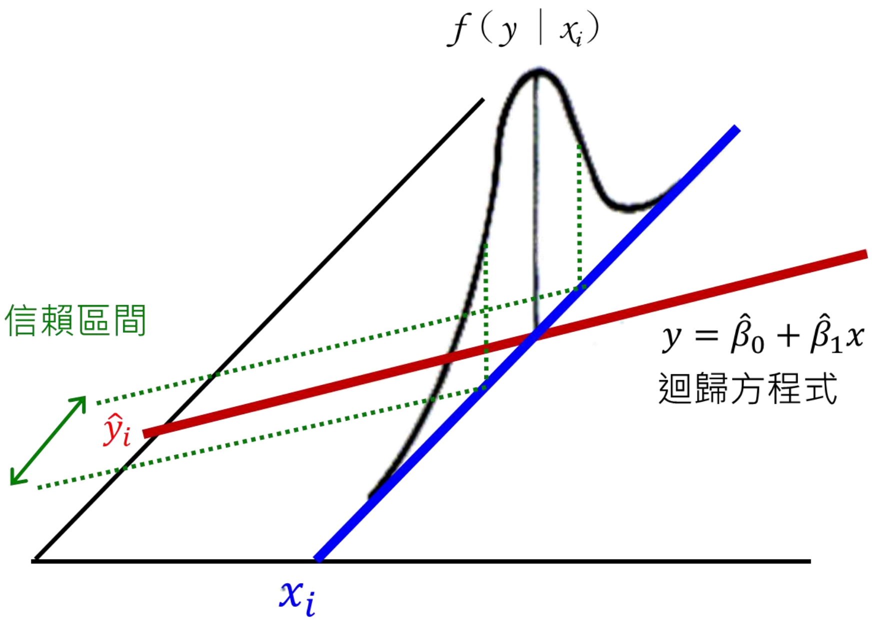
$$\hat{\beta}_1 - \mathrm{t}_{\frac{\alpha}{2},\mathrm{n}-2}\sqrt{\frac{\mathrm{MSE}}{\sum(\mathrm{x_i}-\bar{\mathrm{x}})^2}} \le \beta_1 \le \hat{\beta}_1 + \mathrm{t}_{\frac{\alpha}{2},\mathrm{n}-2}\sqrt{\frac{\mathrm{MSE}}{\sum(\mathrm{x_i}-\bar{\mathrm{x}})^2}}$$

➔ 所以 β_1 的信賴區間：

$= 0.12 \pm 2.306 \times 0.1219 = [\ -0.161\ ,\ 0.101\ ]$

迴歸模型之信賴區間

全體依變數平均數的信賴區間

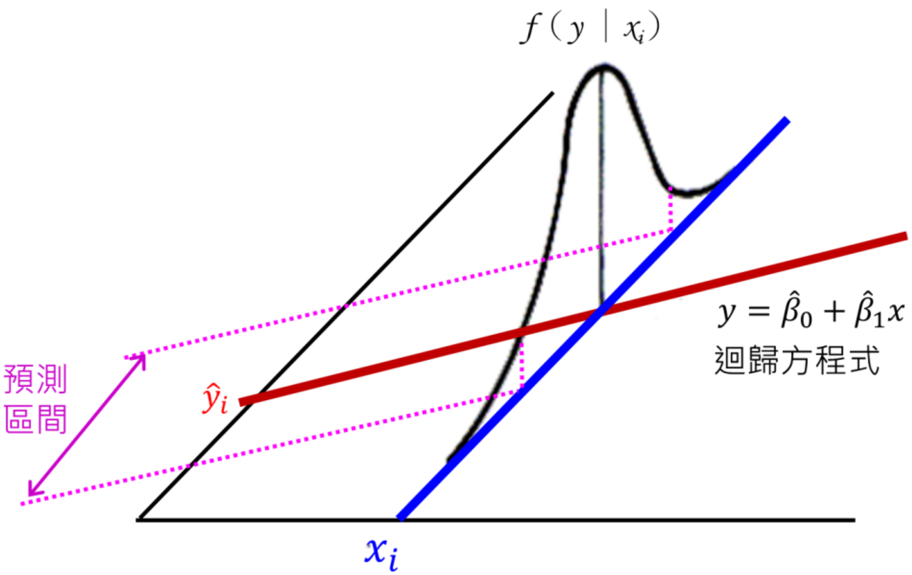


$$\hat{y}_i \sim N\left(E(y|x_i), \left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_k (x_k - \bar{x})^2}\right] \sigma^2\right)$$

$$E(y|x_i) = \hat{y}_i \pm \mathrm{t}_{\frac{\alpha}{2}, n-2} \sqrt{MSE\left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_k (x_k - \bar{x})^2}\right]}$$

其中 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ ， x_i為所給定的自變數值

個別依變數的信賴區間（預測區間）



$$(y_i - \hat{y}_i) \sim N\left(0\,,\left[1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_k (x_k - \bar{x})^2}\right]\sigma^2\right)$$

$$y_i = \hat{y}_i \pm \mathrm{t}_{\frac{\alpha}{2},\,n-2}\sqrt{MSE\left[1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_k (x_k - \bar{x})^2}\right]}$$

其中 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ ， x_i為所給定的自變數值

案例5

某人想瞭解書本的定價是否受書的頁數所影響，於是他隨機選取10本書，記錄其頁數與定價如下表：

頁數(x)	390	700	760	500	560	600	440	500	360	280
定價(y)	280	480	650	320	380	500	200	200	230	130

(1)試求迴歸方程式？

(2)在顯著水準0.05下，是否書本的頁數越多其定價越高？

(3)若某本書頁數是450頁，求同樣450頁的書籍平均定價的95%信賴區間。

(4)若此人欲購買一本新書，其頁數為450頁，求此本書定價的95%信賴區間。

「應用統計學 二版」p408，李德治、童惠玲，博碩文化

案例5 解說(a)

頁數(x)	390	700	760	500	560	600	440	500	360	280	$\bar{x}=$	509
定價(y)	280	480	650	320	380	500	200	200	230	130	$\bar{y}=$	337
$x_i y_i$	109200	336000	494000	160000	212800	300000	88000	100000	82800	36400	Sum=	1919200
x_i^2	152100	490000	577600	250000	313600	360000	193600	250000	129600	78400	Sum=	2794900
y_j^2	78400	230400	422500	102400	144400	250000	40000	40000	52900	16900	Sum=	1377900

(1) $$\hat{\beta}_1 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{1919200 - 10 \times 509 \times 337}{2794900 - 10 \times 509^2} = 0.9989$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 337 - 0.9989 \times 509 = -171.45$$

➔ 線性迴歸方程式為 $\mathrm{y} = -171.45 + 0.9989\,\mathrm{x}$

案例5 解說(b)

(2)依題意，相當檢定： $\begin{cases} H_0：\beta_1 \le 0 \text{ (x對y不具正向影響力)} \\ H_1：\beta_1 > 0 \text{ (x對y具正向影響力)} \end{cases}$

$$SSE = \sum y_i^2 - \hat{\beta}_0^2 \sum y_i - \hat{\beta}_1^2 \sum x_i y_i$$

$$= 1377900 - (-171.45) \times 3370 - 0.9989 \times 1919200 = 38597.62$$

$$\mathrm{M}SE = \frac{\mathrm{SSE}}{\mathrm{n}-2} = \frac{38597.62}{10-2} = 4824.702$$

$$\Rightarrow \quad t^* = \frac{\hat{\beta}_1}{\sqrt{\dfrac{\mathrm{MSE}}{\sum x_i^2 - n\bar{x}^2}}} = \frac{0.9989}{\sqrt{\dfrac{4824.702}{2794900 - 10 \times 509^2}}} = 6.497$$

因為 $t^*=6.497 > -t_{0.05,8} = 1.86$ ，拒絕虛無假設 H_0；
表示有足夠證據顯示，書本的頁數越多，定價越高。

案例5 解說(c)

(3)依題意為求取迴歸方程式的信賴區間，已知 x_0=450，

$$\hat{y}_0 = -171.45 + 0.9989 \times 450 = 278.06$$

$$\Rightarrow E(y|x_0 = 450) = \hat{y}_0 \pm t_{\frac{\alpha}{2}, n-2} \sqrt{MSE\left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum \mathrm{x}_\mathrm{i}^2 - \mathrm{n}\bar{\mathrm{x}}^2}\right]}$$

$$= 278.06 \pm t_{\frac{0.05}{2}, 8} \sqrt{4824.702\left[\frac{1}{10} + \frac{(450 - 509)^2}{2794900 - 10 \times 509^2}\right]}$$

$$= 278.06 \pm 2.036 \times 23.765 = [229.674\,, 326.446]$$

案例5 解說(d)

(4)依題意為求取<u>一本有450頁書的定價預測區間</u>：

➔ $$(y|x_0 = 450) = \hat{y}_0 \pm \mathrm{t}_{\frac{\alpha}{2}, n-2} \sqrt{MSE\left[1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum \mathrm{x_i^2} - \mathrm{n\bar{x}^2}}\right]}$$

$$= 278.06 \pm \mathrm{t}_{\frac{0.05}{2}, 8} \sqrt{4824.702\left[1 + \frac{1}{10} + \frac{(450 - 509)^2}{2794900 - 10 \times 509^2}\right]}$$

$$= 278.06 \pm 2.036 \times 73.413 = [128.59 , 427.56]$$

相關分析(Correlation Analysis)

~迴歸模型之配適度檢定3

- 衡量兩隨機變數相關程度與變化的方向趨勢

相關分析(Correlation Analysis)

$$\rho_{xy} = \frac{\sum(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum(x_i - \mu_x)^2} \cdot \sqrt{\sum(y_i - \mu_y)^2}} = \frac{Cov(x, y)}{\sigma_x \sigma_y} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

其中， $-1 \le \rho_{xy} \le 1$

利用樣本資料來估算母體的相關係數：樣本相關係數

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2} \cdot \sqrt{\sum(y_i - \bar{y})^2}} = \frac{S_{xy}}{S_x S_y}$$

相關係數與判定係數間的關係

$$\Rightarrow r_{xy} = \hat{\beta}_1 \cdot \frac{S_\mathrm{x}}{S_\mathrm{y}}$$

$\Rightarrow \; r_{xy} = \pm\sqrt{R^2}$ 　其正負號與斜率項 $\hat{\beta}_1$ 相同

ρ_{xy}的統計推論

同樣地，我們利用樣本統計量 r_{xy} 來估算母體相關係數 ρ_{xy} ，也可以用來檢定兩變數母體間所抽取的樣本是否有顯著代表性。

假設檢定
(雙尾檢定)

$$\begin{cases} H_0 : \rho_{xy} = 0 \\ H_1 : \rho_{xy} \neq 0 \end{cases}$$

可利用斜率檢定推導而得

檢定統計量

$$t^* = \frac{r_{xy}}{\sqrt{\frac{1 - r_{xy}^2}{n - 2}}}$$

$$\mathrm{t}^* = \frac{\hat{\beta}_1}{\sqrt{\frac{\mathrm{MSE}}{\sum(\mathrm{x_i} - \bar{\mathrm{x}})^2}}}$$

決策法則

當 $|\mathrm{t}^*| > \mathrm{t}_{\frac{\alpha}{2},\mathrm{n}-2}$ 時，拒絕虛無假設H_0。

相同概念，也可以進行「右尾檢定」或「左尾檢定」。

案例6

某人想瞭解工作年資（x）與薪水（y）是否有關，隨機選取10個人分別記錄其年資與薪水，其資料如下：

年資(x)	8	6.2	7.1	7.55	8.75	8.15	10.25	9.6	11.3	7.7
薪資(y)	18	33	48	50	54	56	62	65	71	83

(1)試求迴歸方程式？

(2)求x、y的相關係數，並檢定在顯著水準0.05下，x、y是否具相關性？

「應用統計學 二版」p412，李德治、童惠玲，博碩文化

案例6 解說(a)

年資(x)	8	6.2	7.1	7.55	8.75	8.15	10.25	9.6	11.3	7.7	$\bar{x}$ =	8.46
薪資(y)	18	33	48	50	54	56	62	65	71	83	$\bar{y}$ =	54
$x_i y_i$	144	204.6	340.8	377.5	472.5	456.4	635.5	624	802.3	639.1	Sum=	4696.7
x_i^2	64	38.44	50.41	57.0025	76.5625	66.4225	105.0625	92.16	127.69	59.29	Sum=	737.04
y_j^2	324	1089	2304	2500	2916	3136	3844	4225	5041	6889	Sum=	32268

(1) $$\hat{\beta}_1 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{4696.7 \ - 10 \times 8.46 \times 54}{737.07 - 10 \times 8.46^2} = 6.017$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 54 - 6.017 \times 8.46 = 3.096$$

➔ 線性迴歸方程式為 $\mathrm{y} = 3.096 + 6.017\,\mathrm{x}$

案例6 解說(b)

(2)依題意，相當檢定：$\begin{cases} H_0: \rho_{xy} = 0 \\ H_1: \rho_{xy} \neq 0 \end{cases}$

$$r_{xy} = \hat{\beta}_1 \cdot \frac{S_x}{S_y} = 6.017 \times \frac{1.539}{18.583} = 0.498$$

$$\Rightarrow t^* = \frac{r_{xy}}{\sqrt{\frac{1-r_{xy}^2}{n-2}}} = \frac{0.498}{\sqrt{\frac{1-0.498^2}{10-2}}} = 1.624$$

因為 $t^*=1.624 < t_{0.025,\,8}=2.306$，不拒絕虛無假設 H_0；表示無足夠證據顯示，年資與薪水具相關性。

➔ 若以相關係數為0.498而言，應該要接受對立假設（兩者有關）才對，但檢定的結果卻相反，其原因在於樣本數不夠多。「檢定顯著」僅表示是統計上的顯著，不一定代表事實；欲求證事實，應該必須不斷地重複抽樣驗證或不斷地重複實驗，才可以更客觀驗證事實。

The End