第 14 章　簡單線性迴歸與相關

Simple Linear Regression and Correlation

用一個自變數 x 預測一個應變數 y。模型 $y=\beta_0+\beta_1 x+\varepsilon$，以**最小平方法**由樣本求 b_0, b_1。再用三項工具評估模型：估計標準誤 S_e、判定係數 R^2、對斜率 β_1 的 t 檢定（檢定「是否有線性關係」就是檢定 $\beta_1=0$）。

目錄

1　名詞速查（五欄詞表）

中文術語	English	符號	一句定義	用途／易混
自變數／應變數	independent / dependent	x / y	預測用 / 被預測	x 解釋、y 反應
迴歸模型	regression model	$y=\beta_0+\beta_1 x+\varepsilon$	母體真實關係＋隨機誤差	$\varepsilon\sim N(0,\sigma^2)$
最小平方估計	least squares	b_1, b_0	使 $\sum(y-\hat{y})^2$ 最小的係數	$\hat{y}=b_0+b_1x$
斜率／截距	slope / intercept	b_1 / b_0	x 增一單位 y 變動量 / 截距	$b_1=S_{xy}/S_{xx}$
誤差平方和	sum of sq. error	SSE	$\sum(y-\hat{y})^2$，$df=n-2$	不可解釋變異
迴歸平方和	sum of sq. regression	SSR	$\sum(\hat{y}-\bar{y})^2$，$df=1$	可解釋變異
估計標準誤	standard error of est.	S_e	$\sqrt{\text{SSE}/(n-2)}$	越小配適越好
判定係數	coefficient of determ.	R^2	SSR/SST，可解釋變異比例	$R^2=r^2$
相關係數	correlation	r	線性關係強度與方向	$r=S_{xy}/(S_xS_y)$

2 核心概念：模型與最小平方

簡單線性迴歸模型 $y=\beta_0+\beta_1 x+\varepsilon$，其中 β_0 截距、β_1 斜率（x 對 y 的邊際影響）、ε 隨機誤差。

誤差項四條件：(1) ε 常態；(2) $E(\varepsilon)=0$；(3) 標準差 σ 固定（不隨 x 變）；(4) 各誤差獨立。於是每個 x 下 $y\sim N(\beta_0+\beta_1 x,\ \sigma^2)$。

推導：最小平方法求 b_0, b_1

要找 $\hat{y}=b_0+b_1x$ 使誤差平方和 $D=\sum(y_i-b_0-b_1x_i)^2$ 最小。對 b_0, b_1 偏微分並令為 0（正規方程）：

$$\sum y_i = nb_0 + b_1\sum x_i, \qquad \sum x_i y_i = b_0\sum x_i + b_1\sum x_i^2.$$

解得

$$b_1=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}=\frac{S_{xy}}{S_{xx}}, \qquad b_0=\bar{y}-b_1\bar{x}.$$

（取 $\sum(y-\hat{y})^2$ 而非 $\sum(y-\hat{y})$，因後者恆為 0。）

係數估計與計算簡式

$$b_1=\frac{S_{xy}}{S_{xx}}, \qquad b_0=\bar{y}-b_1\bar{x}, \qquad \hat{y}=b_0+b_1x.$$

$$S_{xy}=\sum x_iy_i-\frac{\sum x_i\sum y_i}{n}, \quad S_{xx}=\sum x_i^2-\frac{(\sum x_i)^2}{n}, \quad S_{yy}=\sum y_i^2-\frac{(\sum y_i)^2}{n}.$$

3 評估模型：SSE、S_e、R^2、ANOVA

變異分解與三平方和

$$\underbrace{\sum(y_i-\bar{y})^2}_{\text{SST}=S_{yy}}=\underbrace{\sum(\hat{y}_i-\bar{y})^2}_{\text{SSR}=\frac{S_{xy}^2}{S_{xx}}}+\underbrace{\sum(y_i-\hat{y}_i)^2}_{\text{SSE}=S_{yy}-\frac{S_{xy}^2}{S_{xx}}}.$$

自由度：$\text{SST}=n-1$，$\text{SSR}=1$，$\text{SSE}=n-2$。

估計標準誤與判定係數

$$S_e=\sqrt{\text{MSE}}=\sqrt{\frac{\text{SSE}}{n-2}} \quad (\sigma \text{ 的不偏估計}), \qquad R^2=\frac{\text{SSR}}{\text{SST}}=\frac{S_{xy}^2}{S_{xx}S_{yy}}=1-\frac{\text{SSE}}{\text{SST}}.$$

$R^2\in[0,1]$，越接近 1 線性關係越強；且 $R^2=r^2$（r ＝樣本相關係數）。

變異來源	SS	df	MS	F
迴歸	SSR	1	MSR=SSR	$F=\dfrac{\text{MSR}}{\text{MSE}}$
誤差	SSE	$n-2$	MSE=$\dfrac{\text{SSE}}{n-2}$	
總和	SST	$n-1$		

4 對斜率 β_1 的推論（檢定線性關係）

b_1 的抽樣分配、檢定與信賴區間

$E(b_1)=\beta_1$，$\text{se}(b_1)=\dfrac{S_e}{\sqrt{S_{xx}}}$，故

$$T=\frac{b_1-\beta_{10}}{S_e/\sqrt{S_{xx}}}\ \sim\ t_{n-2}.$$

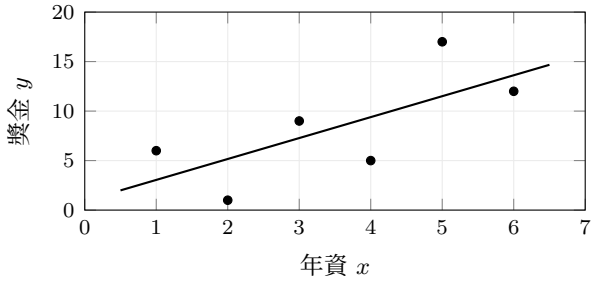
β_1 的 $100(1-\alpha)\%$ 信賴區間：$b_1 \pm t_{\alpha/2,\,n-2}\dfrac{S_e}{\sqrt{S_{xx}}}$。

「有沒有線性關係」＝檢定 $\beta_1=0$　若 $\beta_1=0$，迴歸線水平，y 不隨 x 變，即無線性關係。故雙尾檢定 $H_0:\beta_1=0$ vs $H_1:\beta_1\neq 0$；拒絕 H_0 表示有顯著線性關係。（左／右尾用於檢定特定方向。）

5 例題

例題 14.1：年資 x 與績效獎金 y（配適迴歸線＋檢定）

六位員工：$x=1,2,3,4,5,6$；$y=6,1,9,5,17,12$。算得 $\sum x=21, \sum y=50, \sum xy=212, \sum x^2=91, \sum y^2=576,\ n=6$。



係數：$S_{xy}=212-\frac{21\cdot 50}{6}=37$，$S_{xx}=91-\frac{21^2}{6}=17.5$，$S_{yy}=576-\frac{50^2}{6}=159.33$。
$b_1=\frac{37}{17.5}=2.114$，$b_0=\frac{50}{6}-2.114\cdot\frac{21}{6}=0.934$，故 $\hat{y}=0.934+2.114x$。
評估：$\text{SSE}=159.33-\frac{37^2}{17.5}=81.10$，$S_e=\sqrt{81.10/4}=4.50$，$R^2=\frac{37^2}{17.5\cdot 159.33}=0.49$。
檢定 $\beta_1=0$（$\alpha=0.05$）：$T=\frac{2.114-0}{4.50/\sqrt{17.5}}=\frac{2.114}{1.076}=1.96$；$df=n-2=4$；$R=\{|T|\geq t_{0.025,4}=2.776\}$；$1.96<2.776\Rightarrow$ **Do not reject** H_0：此樣本**無足夠證據**說年資與獎金有線性關係（樣本小、變異大）。

例題 14.2 ／ 14.3 ／ 14.4：二手車里程 x 與售價 y（$n=100$）

$\sum x=3601.1, \sum y=1484.1, \sum xy=53155.93, \sum x^2=133986.59, \sum y^2=22055.23$。
$S_{xy}=53155.93-\frac{3601.1\cdot 1484.1}{100}=-287.995$，$S_{xx}=4307.378$，$S_{yy}=29.702$。
(14.2 迴歸線)：$b_1=\frac{-287.995}{4307.378}=-0.0669$，$b_0=\frac{1484.1}{100}-(-0.0669)\frac{3601.1}{100}=17.25$，$\hat{y}=17.25-0.0669x$（里程每增千英里，售價降約 66.9 美元）。
(14.3 標準誤)：$\text{SSE}=29.702-\frac{(-287.995)^2}{4307.378}=10.446$，$S_e=\sqrt{10.446/98}=0.3265$。
(14.4 檢定 $\beta_1=0$，$\alpha=0.05$**)**：$T=\frac{-0.0669-0}{0.3265/\sqrt{4307.378}}=-13.45$；$df=98$；$R=\{|T|\geq t_{0.025,98}\approx 1.984\}$；$|-13.45|\geq 1.984\Rightarrow$ **Reject** H_0：售價與里程有顯著線性（負）關係。

$R^2 = \frac{(-287.995)^2}{4307.378 \cdot 29.702} = 0.648$（里程解釋了約 64.8% 的售價變異），$r = -\sqrt{0.648} = -0.805$。

6 公式整理（總表）

量	公式
斜率 / 截距	$b_1 = S_{xy}/S_{xx}$，$b_0 = \bar{y} - b_1\bar{x}$
三平方和	$S_{xy} = \sum xy - \frac{\sum x \sum y}{n}$，$S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$，$S_{yy} = \sum y^2 - \frac{(\sum y)^2}{n}$
SSR / SSE / SST	$\text{SSR} = \frac{S_{xy}^2}{S_{xx}}$，$\text{SSE} = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$，$\text{SST} = S_{yy}$
估計標準誤	$S_e = \sqrt{\text{SSE}/(n-2)}$
判定係數	$R^2 = \text{SSR}/\text{SST} = S_{xy}^2/(S_{xx}S_{yy}) = r^2$
斜率檢定	$T = \dfrac{b_1 - \beta_{10}}{S_e/\sqrt{S_{xx}}} \sim t_{n-2}$

7 易錯點總整理

簡單線性迴歸

- **自由度**：S_e 與斜率 t 檢定的自由度是 $n-2$（估了 b_0, b_1 兩個），不是 $n-1$。
- **「有線性關係」＝檢定** $\beta_1 = 0$（雙尾）；拒絕才表示有顯著線性關係，否則勿用迴歸線預測。
- $R^2 = r^2$：判定係數是相關係數的平方；R^2 大表示解釋力強，但**不等於因果**。
- **SSE 計算**：用 $S_{yy} - S_{xy}^2/S_{xx}$ 比逐點 $(y-\hat{y})^2$ 快；S_{xy} 可正可負（決定斜率正負與 r 正負）。
- **誤差四條件**：常態、均值 0、等變異、獨立；違反時模型推論不可靠。
- 外推（extrapolation）超出 x 觀測範圍的預測風險高。

8 計算跳板（數值代入演練）

量（例 14.1）	代入	結果
S_{xy}	$212 - \frac{21 \times 50}{6}$	$= 37$
S_{xx}	$91 - \frac{21^2}{6}$	$= 17.5$
$b_1,\ b_0$	$37/17.5$；$\frac{50}{6} - 2.114 \cdot \frac{21}{6}$	$b_1 = 2.114,\ b_0 = 0.934$
SSE，S_e	$159.33 - \frac{37^2}{17.5}$；$\sqrt{81.10/4}$	SSE$= 81.10$，$S_e = 4.50$
T（$\beta_1 = 0$）	$2.114/(4.50/\sqrt{17.5})$	$= 1.96$（$< t_{0.025,4} = 2.776$，不拒絕）

寫字區：自選資料，算 b_0, b_1、$\hat{y}$、R^2，並檢定 $\beta_1 = 0$

9 自我檢查

1. 最小平方法在最小化什麼？b_1、b_0 的公式為何？
2.「檢定 x, y 是否有線性關係」對應到檢定哪個參數？自由度多少？
3. R^2 的意義是什麼？它與相關係數 r 有何關係？
4. S_e 怎麼算？它大或小分別代表模型配適如何？
5. SST、SSR、SSE 的關係與各自的自由度？

寫字區：作答

參考答案：1. 最小化 $\sum(y-\hat{y})^2$；$b_1 = S_{xy}/S_{xx}$、$b_0 = \bar{y} - b_1\bar{x}$。　2. 檢定 $\beta_1 = 0$（斜率）；自由度 $n-2$。　3. R^2 ＝可解釋變異占總變異的比例 SSR/SST；$R^2 = r^2$。　4. $S_e = \sqrt{\text{SSE}/(n-2)}$；小表示配適好（誤差小）、大表示配適差。　5. SST=SSR+SSE，*df*：$n-1 = 1 + (n-2)$。