

# 統計學講義

## 第六部分：迴歸分析

### 1 簡單線性迴歸

#### 1.1 模型與估計

**定義 1.1** (簡單線性迴歸模型).

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \dots, n$$

其中  $\beta_0$  為截距， $\beta_1$  為斜率， $\varepsilon_i \sim N(0, \sigma^2)$  i.i.d.

**定理 1.1** (OLS 估計量). 最小平方法 (OLS) 估計量為：

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{S_{XY}}{S_{XX}}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

其中： $S_{XX} = \sum (X_i - \bar{X})^2$ ， $S_{XY} = \sum (X_i - \bar{X})(Y_i - \bar{Y})$ 。

**證明.** OLS 最小化殘差平方和  $Q(\beta_0, \beta_1) = \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2$ 。

**Step 1 :** 對  $\beta_0$  偏微分並令其為零：

$$\frac{\partial Q}{\partial \beta_0} = -2 \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i) = 0$$

展開得  $\sum Y_i - n\beta_0 - \beta_1 \sum X_i = 0$ ，故

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

**Step 2 :** 對  $\beta_1$  偏微分並令其為零：

$$\frac{\partial Q}{\partial \beta_1} = -2 \sum_{i=1}^n X_i (Y_i - \beta_0 - \beta_1 X_i) = 0$$

展開得  $\sum X_i Y_i - \beta_0 \sum X_i - \beta_1 \sum X_i^2 = 0$ 。

**Step 3 :** 將  $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$  代入：

$$\begin{aligned} \sum X_i Y_i - (\bar{Y} - \hat{\beta}_1 \bar{X}) \sum X_i - \hat{\beta}_1 \sum X_i^2 &= 0 \\ \sum X_i Y_i - n\bar{X}\bar{Y} + \hat{\beta}_1 n\bar{X}^2 - \hat{\beta}_1 \sum X_i^2 &= 0 \\ \sum X_i Y_i - n\bar{X}\bar{Y} &= \hat{\beta}_1 (\sum X_i^2 - n\bar{X}^2) \end{aligned}$$

由於  $S_{XY} = \sum X_i Y_i - n\bar{X}\bar{Y}$  且  $S_{XX} = \sum X_i^2 - n\bar{X}^2$ ，故

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

**定理 1.2** (OLS 估計量的不偏性).  $\hat{\beta}_0$  與  $\hat{\beta}_1$  是  $\beta_0$  與  $\beta_1$  的不偏估計量。

**證明.** 證明  $E[\hat{\beta}_1] = \beta_1$  :

將  $\hat{\beta}_1$  改寫為  $Y_i$  的線性組合。令  $w_i = \frac{X_i - \bar{X}}{S_{XX}}$ ，則  $\sum w_i = 0$  且  $\sum w_i X_i = 1$ 。

$$\hat{\beta}_1 = \frac{\sum (X_i - \bar{X}) Y_i}{S_{XX}} = \sum w_i Y_i$$

由模型  $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$  :

$$\begin{aligned} E[\hat{\beta}_1] &= E\left[\sum w_i Y_i\right] = \sum w_i E[Y_i] = \sum w_i (\beta_0 + \beta_1 X_i) \\ &= \beta_0 \sum w_i + \beta_1 \sum w_i X_i = \beta_0 \cdot 0 + \beta_1 \cdot 1 = \beta_1 \end{aligned}$$

**證明**  $E[\hat{\beta}_0] = \beta_0$  :

$$E[\hat{\beta}_0] = E[\bar{Y} - \hat{\beta}_1 \bar{X}] = E[\bar{Y}] - \bar{X} E[\hat{\beta}_1] = (\beta_0 + \beta_1 \bar{X}) - \bar{X} \beta_1 = \beta_0$$

**定理 1.3** (OLS 估計量的變異數).

$$\text{var}(\hat{\beta}_1) = \frac{\sigma^2}{S_{XX}}$$

$$\text{var}(\hat{\beta}_0) = \sigma^2 \left( \frac{1}{n} + \frac{\bar{X}^2}{S_{XX}} \right)$$

**證明.** 證明  $\text{var}(\hat{\beta}_1) = \sigma^2 / S_{XX}$  :

由  $\hat{\beta}_1 = \sum w_i Y_i$ ，其中  $w_i = (X_i - \bar{X}) / S_{XX}$ ，且  $Y_i$  獨立：

$$\begin{aligned} \text{var}(\hat{\beta}_1) &= \text{var}\left(\sum w_i Y_i\right) = \sum w_i^2 \text{var}(Y_i) = \sigma^2 \sum w_i^2 \\ &= \sigma^2 \sum \frac{(X_i - \bar{X})^2}{S_{XX}^2} = \sigma^2 \cdot \frac{S_{XX}}{S_{XX}^2} = \frac{\sigma^2}{S_{XX}} \end{aligned}$$

**證明**  $\text{var}(\hat{\beta}_0)$  :

$$\text{var}(\hat{\beta}_0) = \text{var}(\bar{Y} - \hat{\beta}_1 \bar{X}) = \text{var}(\bar{Y}) + \bar{X}^2 \text{var}(\hat{\beta}_1) - 2\bar{X} \text{cov}(\bar{Y}, \hat{\beta}_1)$$

由於  $\text{cov}(\bar{Y}, \hat{\beta}_1) = \text{cov}\left(\frac{1}{n} \sum Y_i, \sum w_i Y_i\right) = \frac{\sigma^2}{n} \sum w_i = 0$ ，故

$$\text{var}(\hat{\beta}_0) = \frac{\sigma^2}{n} + \bar{X}^2 \cdot \frac{\sigma^2}{S_{XX}} = \sigma^2 \left( \frac{1}{n} + \frac{\bar{X}^2}{S_{XX}} \right)$$

## 1.2 變異數分解

**定理 1.4** (變異數分解——ANOVA).

$$\text{SST} = \text{SSR} + \text{SSE}$$

其中：

- $\text{SST} = \sum (Y_i - \bar{Y})^2$  (總變異)
- $\text{SSR} = \sum (\hat{Y}_i - \bar{Y})^2$  (迴歸變異)
- $\text{SSE} = \sum (Y_i - \hat{Y}_i)^2$  (殘差變異)

**證明.** 將  $Y_i - \bar{Y}$  分解為  $(Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})$  :

$$\begin{aligned} \text{SST} &= \sum (Y_i - \bar{Y})^2 = \sum [(Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})]^2 \\ &= \sum (Y_i - \hat{Y}_i)^2 + \sum (\hat{Y}_i - \bar{Y})^2 + 2 \sum (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) \end{aligned}$$

**關鍵：**證明交叉項為零。令  $e_i = Y_i - \hat{Y}_i$  (殘差)，則：

$$\sum e_i(\hat{Y}_i - \bar{Y}) = \sum e_i \hat{Y}_i - \bar{Y} \sum e_i$$

由 OLS 正規方程式， $\sum e_i = 0$  且  $\sum e_i X_i = 0$ 。

又  $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ ，故

$$\sum e_i \hat{Y}_i = \hat{\beta}_0 \sum e_i + \hat{\beta}_1 \sum e_i X_i = 0$$

因此交叉項為零， $\text{SST} = \text{SSE} + \text{SSR}$ 。

**推論 1.1** (SSR 的另一形式).  $\text{SSR} = \hat{\beta}_1^2 S_{XX} = \hat{\beta}_1 S_{XY}$ 。

**證明.**

$$\begin{aligned} \text{SSR} &= \sum (\hat{Y}_i - \bar{Y})^2 = \sum (\hat{\beta}_0 + \hat{\beta}_1 X_i - \bar{Y})^2 \\ &= \sum [(\bar{Y} - \hat{\beta}_1 \bar{X}) + \hat{\beta}_1 X_i - \bar{Y}]^2 = \sum [\hat{\beta}_1 (X_i - \bar{X})]^2 \\ &= \hat{\beta}_1^2 \sum (X_i - \bar{X})^2 = \hat{\beta}_1^2 S_{XX} \end{aligned}$$

又  $\hat{\beta}_1 = S_{XY}/S_{XX}$ ，故  $\text{SSR} = \hat{\beta}_1^2 S_{XX} = \hat{\beta}_1 \cdot S_{XY}$ 。

**定理 1.5** ( $R^2$  與相關係數的關係).  $R^2 = r_{XY}^2$ ，其中  $r_{XY}$  是  $X$  與  $Y$  的樣本相關係數。

**證明.** 樣本相關係數  $r_{XY} = \frac{S_{XY}}{\sqrt{S_{XX} \cdot S_{YY}}}$ 。

由  $R^2 = \text{SSR}/\text{SST} = \hat{\beta}_1^2 S_{XX}/S_{YY}$  且  $\hat{\beta}_1 = S_{XY}/S_{XX}$ ：

$$R^2 = \frac{(S_{XY}/S_{XX})^2 \cdot S_{XX}}{S_{YY}} = \frac{S_{XY}^2}{S_{XX} \cdot S_{YY}} = r_{XY}^2$$

### 1.3 $\sigma^2$ 的估計

**定理 1.6** ( $\sigma^2$  的不偏估計).

$$s^2 = \text{MSE} = \frac{\text{SSE}}{n-2}$$

是  $\sigma^2$  的不偏估計量。

**證明.** 需證明  $E[\text{SSE}] = (n-2)\sigma^2$ 。

由  $\text{SSE} = \text{SST} - \text{SSR} = \sum (Y_i - \bar{Y})^2 - \hat{\beta}_1^2 S_{XX}$ ：

$$E[\text{SSE}] = E[\text{SST}] - S_{XX} E[\hat{\beta}_1^2]$$

**計算**  $E[\text{SST}]$ ：

$$E[\text{SST}] = E\left[\sum Y_i^2 - n\bar{Y}^2\right] = \sum E[Y_i^2] - n E[\bar{Y}^2]$$

由  $E[Y_i^2] = \text{var}(Y_i) + (E[Y_i])^2 = \sigma^2 + (\beta_0 + \beta_1 X_i)^2$ ，且  $E[\bar{Y}^2] = \text{var}(\bar{Y}) + (E[\bar{Y}])^2 = \sigma^2/n + (\beta_0 + \beta_1 \bar{X})^2$ ，  
經計算：

$$E[\text{SST}] = (n-1)\sigma^2 + \beta_1^2 S_{XX}$$

計算  $E[\hat{\beta}_1^2]$ ：

$$E[\hat{\beta}_1^2] = \text{var}(\hat{\beta}_1) + (E[\hat{\beta}_1])^2 = \frac{\sigma^2}{S_{XX}} + \beta_1^2$$

合併：

$$\begin{aligned} E[\text{SSE}] &= (n-1)\sigma^2 + \beta_1^2 S_{XX} - S_{XX} \left( \frac{\sigma^2}{S_{XX}} + \beta_1^2 \right) \\ &= (n-1)\sigma^2 + \beta_1^2 S_{XX} - \sigma^2 - \beta_1^2 S_{XX} = (n-2)\sigma^2 \end{aligned}$$

故  $E[\text{MSE}] = E[\text{SSE}/(n-2)] = \sigma^2$ 。

## 1.4 假設檢定

**定理 1.7** (斜率的  $t$  檢定). 在  $H_0: \beta_1 = 0$  下，檢定統計量

$$t = \frac{\hat{\beta}_1}{\text{SE}(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{s/\sqrt{S_{XX}}} \sim t_{n-2}$$

**證明. Step 1：**標準化  $\hat{\beta}_1$ 。

由定理 1.2 和 1.3， $\hat{\beta}_1 \sim N(\beta_1, \sigma^2/S_{XX})$ ，故在  $H_0: \beta_1 = 0$  下：

$$Z = \frac{\hat{\beta}_1 - 0}{\sigma/\sqrt{S_{XX}}} \sim N(0, 1)$$

**Step 2：** $\text{SSE}/\sigma^2$  的分配。

可證明  $\text{SSE}/\sigma^2 \sim \chi_{n-2}^2$  (殘差有  $n$  個觀測值減去 2 個估計參數)。

**Step 3：** $\hat{\beta}_1$  與  $\text{SSE}$  獨立。

由常態分配的性質， $\hat{\beta}_1$  (依賴  $\bar{Y}$  的方向) 與  $\text{SSE}$  (垂直於該方向) 獨立。

**Step 4：**構造  $t$  統計量。

由  $t$  分配定義：

$$t = \frac{Z}{\sqrt{\chi_{n-2}^2/(n-2)}} = \frac{\hat{\beta}_1/(\sigma/\sqrt{S_{XX}})}{\sqrt{\text{SSE}/\sigma^2/(n-2)}} = \frac{\hat{\beta}_1}{s/\sqrt{S_{XX}}} \sim t_{n-2}$$

**定理 1.8** (整體  $F$  檢定). 在  $H_0: \beta_1 = 0$  下，

$$F = \frac{\text{MSR}}{\text{MSE}} = \frac{\text{SSR}/1}{\text{SSE}/(n-2)} \sim F_{1, n-2}$$

且  $F = t^2$ 。

**證明.** 證明  $F \sim F_{1,n-2}$  :

在  $H_0: \beta_1 = 0$  下,  $SSR/\sigma^2 \sim \chi_1^2$  (因為  $SSR$  是一個參數的函數)。

由於  $SSR$  與  $SSE$  獨立 (可由正交分解證明), 且  $SSE/\sigma^2 \sim \chi_{n-2}^2$  :

$$F = \frac{SSR/\sigma^2/1}{SSE/\sigma^2/(n-2)} = \frac{SSR/1}{SSE/(n-2)} \sim F_{1,n-2}$$

**證明**  $F = t^2$  :

$$F = \frac{SSR}{MSE} = \frac{\hat{\beta}_1^2 S_{XX}}{s^2} = \frac{\hat{\beta}_1^2}{s^2/S_{XX}} = \left( \frac{\hat{\beta}_1}{s/\sqrt{S_{XX}}} \right)^2 = t^2$$

## 1.5 調整後 $R^2$

**定理 1.9** (調整後  $R^2$ ). 對於有  $k$  個自變數的多元迴歸 :

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1}(1-R^2) = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}$$

**證明.** 由  $R^2 = 1 - SSE/SST$  :

$$1 - R^2 = \frac{SSE}{SST}$$

調整後  $R^2$  用均方取代平方和 :

$$\bar{R}^2 = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)} = 1 - \frac{n-1}{n-k-1} \cdot \frac{SSE}{SST} = 1 - \frac{n-1}{n-k-1}(1-R^2)$$

這樣調整的原因：當增加無意義的自變數時， $SSE$  下降但自由度也下降， $\bar{R}^2$  可能下降，避免過度擬合。

## 2 信賴區間與預測

**定理 2.1** (迴歸係數的信賴區間).  $\beta_1$  的  $(1-\alpha)$  信賴區間為 :

$$\hat{\beta}_1 \pm t_{\alpha/2, n-2} \cdot \frac{s}{\sqrt{S_{XX}}}$$

**證明.** 由定理 1.7,  $\frac{\hat{\beta}_1 - \beta_1}{s/\sqrt{S_{XX}}} \sim t_{n-2}$ 。

$$\text{故 } P\left(-t_{\alpha/2, n-2} \leq \frac{\hat{\beta}_1 - \beta_1}{s/\sqrt{S_{XX}}} \leq t_{\alpha/2, n-2}\right) = 1 - \alpha$$

$$\text{移項得 } P\left(\hat{\beta}_1 - t_{\alpha/2, n-2} \cdot \frac{s}{\sqrt{S_{XX}}} \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2, n-2} \cdot \frac{s}{\sqrt{S_{XX}}}\right) = 1 - \alpha$$

**定理 2.2** (平均反應的信賴區間). 對於給定  $X = X_0$ ,  $E[Y|X_0] = \beta_0 + \beta_1 X_0$  的  $(1-\alpha)$  信賴區間 :

$$\hat{Y}_0 \pm t_{\alpha/2, n-2} \cdot s \sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{XX}}}$$

**證明.** 令  $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 X_0$  為  $E[Y|X_0]$  的估計。

**計算**  $\text{var}(\hat{Y}_0)$ ：

$$\hat{Y}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} + \hat{\beta}_1 X_0 = \bar{Y} + \hat{\beta}_1 (X_0 - \bar{X})$$

由於  $\text{cov}(\bar{Y}, \hat{\beta}_1) = 0$ ：

$$\begin{aligned} \text{var}(\hat{Y}_0) &= \text{var}(\bar{Y}) + (X_0 - \bar{X})^2 \text{var}(\hat{\beta}_1) \\ &= \frac{\sigma^2}{n} + (X_0 - \bar{X})^2 \cdot \frac{\sigma^2}{S_{XX}} = \sigma^2 \left( \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{XX}} \right) \end{aligned}$$

故  $\frac{\hat{Y}_0 - (\beta_0 + \beta_1 X_0)}{s \sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{XX}}}} \sim t_{n-2}$ ，由此得信賴區間。

**定理 2.3** (個別預測的預測區間). 對於給定  $X = X_0$ ，個別觀測值  $Y_{\text{new}}$  的  $(1 - \alpha)$  預測區間：

$$\hat{Y}_0 \pm t_{\alpha/2, n-2} \cdot s \sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{XX}}}$$

**證明.** 預測誤差為  $Y_{\text{new}} - \hat{Y}_0$ 。

由於  $Y_{\text{new}} = \beta_0 + \beta_1 X_0 + \varepsilon_{\text{new}}$  與  $\hat{Y}_0$  獨立：

$$\begin{aligned} \text{var}(Y_{\text{new}} - \hat{Y}_0) &= \text{var}(Y_{\text{new}}) + \text{var}(\hat{Y}_0) \\ &= \sigma^2 + \sigma^2 \left( \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{XX}} \right) = \sigma^2 \left( 1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{XX}} \right) \end{aligned}$$

故預測區間比信賴區間多了  $\sigma^2$  這一項（個別觀測的隨機變異）。

### 3 多元線性迴歸

**定義 3.1** (多元迴歸模型).

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \varepsilon_i, \quad \varepsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$$

**定理 3.1** (整體  $F$  檢定 (多元迴歸)). 檢定  $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$ ：

$$F = \frac{\text{MSR}}{\text{MSE}} = \frac{\text{SSR}/k}{\text{SSE}/(n-k-1)} = \frac{R^2/k}{(1-R^2)/(n-k-1)} \sim F_{k, n-k-1}$$

**證明.** 在  $H_0$  下， $\text{SSR}/\sigma^2 \sim \chi_k^2$  ( $k$  個限制式)， $\text{SSE}/\sigma^2 \sim \chi_{n-k-1}^2$  ( $n$  個觀測減去  $k+1$  個參數)，且兩者獨立。

由  $F$  分配定義：

$$F = \frac{\text{SSR}/\sigma^2/k}{\text{SSE}/\sigma^2/(n-k-1)} = \frac{\text{SSR}/k}{\text{SSE}/(n-k-1)} \sim F_{k, n-k-1}$$

又  $R^2 = \text{SSR}/\text{SST}$ ，故  $\text{SSR} = R^2 \cdot \text{SST}$ ， $\text{SSE} = (1 - R^2)\text{SST}$ ：

$$F = \frac{R^2 \cdot \text{SST}/k}{(1 - R^2)\text{SST}/(n-k-1)} = \frac{R^2/k}{(1 - R^2)/(n-k-1)}$$

**定理 3.2** (個別係數  $t$  檢定). 對於  $H_0 : \beta_j = 0$  :

$$t = \frac{\hat{\beta}_j}{\text{SE}(\hat{\beta}_j)} \sim t_{n-k-1}$$

**證明.** 由多元迴歸理論， $\hat{\beta}_j \sim N(\beta_j, \sigma^2 c_{jj})$ ，其中  $c_{jj}$  是  $(\mathbf{X}'\mathbf{X})^{-1}$  的第  $j$  個對角元素。

以  $s^2$  估計  $\sigma^2$  後， $\frac{\hat{\beta}_j - \beta_j}{s\sqrt{c_{jj}}} \sim t_{n-k-1}$ 。

在  $H_0 : \beta_j = 0$  下，檢定統計量為  $t = \hat{\beta}_j / \text{SE}(\hat{\beta}_j)$ 。

## 4 計算範例

**例題 4.1.** 給定資料：(1, 2), (2, 4), (3, 5), (4, 4), (5, 5)，進行完整線性迴歸分析。

**解答. 1. 基本統計量：** $n = 5$ ， $\bar{X} = 3$ ， $\bar{Y} = 4$ ， $S_{XX} = 10$ ， $S_{XY} = 6$ ， $S_{YY} = 6$ 。

**2. 估計值：** $\hat{\beta}_1 = 6/10 = 0.6$ ， $\hat{\beta}_0 = 4 - 0.6(3) = 2.2$ 。

**3. ANOVA：** $\text{SSR} = (0.6)^2 \times 10 = 3.6$ ， $\text{SSE} = 6 - 3.6 = 2.4$ ， $R^2 = 0.6$ 。

**4. 假設檢定：** $\text{MSE} = 2.4/3 = 0.8$ ， $s = 0.894$ 。

$\text{SE}(\hat{\beta}_1) = 0.894/\sqrt{10} = 0.283$ ， $t = 0.6/0.283 = 2.12$ 。

$df = 3$ ， $t_{0.025,3} = 3.182$ 。因為  $|2.12| < 3.182$ ，不拒絕  $H_0$ 。

**5. 信賴區間：** $0.6 \pm 3.182 \times 0.283 = 0.6 \pm 0.90 = (-0.30, 1.50)$ 。

**例題 4.2.**  $n = 100$ ， $k = 4$ ， $R^2 = 0.36$ ，求  $F$  值和調整後  $R^2$ 。

**解答.  $F$  值：**

$$F = \frac{0.36/4}{(1 - 0.36)/(100 - 4 - 1)} = \frac{0.09}{0.64/95} = \frac{0.09}{0.00674} = 13.35$$

**調整後  $R^2$ ：**

$$\bar{R}^2 = 1 - \frac{99}{95}(1 - 0.36) = 1 - 1.042 \times 0.64 = 1 - 0.667 = 0.333$$