Linjär regression

Betrakta följande modell:

$$Y(x) \text{ är } \mathrm{N}(m(x), \sigma)$$

det vill säga Y är en normalfördelad stokastisk variabel vars väntevärde beror på en underliggande parameter x. Vi kommer i huvudsak att betrakta funktioner

$$m(x) = \alpha' + \beta x,$$

det vill säga att väntevärdet för $Y(x)$ är linjär i x. Problemet nu är hur vi med observationer

$$y_1(x_1), y_2(x_2), \ldots, y_n(x_n)$$

skall skatta α' och β.

För vår observationsserie, låt oss byta parametrisering till

$$m(x) = \alpha' + \beta x = \alpha' + \beta x - \beta\overline{x} + \beta\overline{x} = \underbrace{\alpha' + \beta\overline{x}}_{=\alpha} + \beta(x - \overline{x}) = \alpha + \beta(x - \overline{x}),$$

det vill säga vi har infört α istället för α'. Med minsta kvadratmetoden får vi att skattningarna av α och β är de värden som minimerar

$$Q(\alpha, \beta) = \sum_{i=1}^{n}(y_i - m(x_i))^2 = \sum_{i=1}^{n}(y_i - (\alpha + \beta(x_i - \overline{x}))^2.$$

Derivering ger ekvationen

$$\begin{aligned}\frac{\partial}{\partial\alpha}Q(\alpha,\beta) &= -2\sum_{i=1}^{n}(y_i - (\alpha + \beta(x_i - \overline{x}))) = -2\left(\sum_{i=1}^{n} y_i\right) + 2n\alpha + \beta\underbrace{\sum_{i=1}^{n}(x_i - \overline{x})}_{=0} \\ &= -2n\overline{y} + 2n\alpha = 0,\end{aligned}$$

eller $\alpha^* = \overline{y}$. Vidare,

$$\begin{aligned}\frac{\partial}{\partial\beta}Q(\alpha,\beta) &= -2\sum_{i=1}^{n}(y_i - \overline{y} - \beta(x_i - \overline{x}))(x_i - \overline{x}) \\ &= -2\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x}) + 2\beta\sum_{i=1}^{n}(x_i - \overline{x})^2 = 0,\end{aligned}$$

eller

$$\beta^* = \frac{\displaystyle\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x})}{\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{S_{xy}}{S_{xx}}.$$

Skattningarna beskrivs av motsvarande stokastiska variabler:

$$\alpha^* = \overline{Y} \qquad \beta^* = \frac{S_{xY}}{S_{xx}}$$

som båda är linjärkombinationer av normalfördelade variabler och alltså är normalfördelade. Slutligen får vi en skattning av σ^2 som

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - (\alpha^* + \beta^*(x_i - \overline{x})))^2 = \frac{1}{n-2}Q(\alpha^*, \beta^*)$$

Sammanfattning

$$\begin{aligned}
\alpha^* &= \overline{Y} \text{ är N}\left(\alpha, \frac{\sigma}{\sqrt{n}}\right) \\
\beta^* &= \frac{S_{xY}}{S_{xx}} \text{ är N}\left(\beta, \frac{\sigma}{\sqrt{S_{xx}}}\right) \\
S^2 &= \frac{1}{n-2}(S_{YY} - \beta^* S_{xY}) = \frac{1}{n-2}Q(\alpha^*, \beta^*) \\
\frac{(n-2)S^2}{\sigma^2} &= \frac{Q(\alpha^*, \beta^*)}{\sigma^2} \text{ är } \chi^2(n-2)
\end{aligned}$$

Fixera en punkt x_0. Ett konfidensintervall för $m_0 = m(x_0) = \mathsf{E}[Y(x_0)] = \alpha + \beta(x_0 - \overline{x})$ erhålls som följer. Vi skattar $\alpha + \beta(x_0 - \overline{x})$ med den normalfördelade variabeln

$$m_0^* = \alpha^* + \beta^*(x_0 - \overline{x})$$

som har väntevärde

$$\mathsf{E}[m_0^*] = \mathsf{E}[\alpha^* + \beta^*(x_0 - \overline{x})] = \alpha + \beta(x_0 - \overline{x}) = m_0$$

och varians

$$\mathrm{V}(m_0^*) = \mathrm{V}(\alpha^*) + \mathrm{V}(\beta^*)(x_0 - \overline{x})^2 = \frac{\sigma^2}{n} + \sigma^2\frac{(x_0 - \overline{x})^2}{S_{xx}}.$$

(Här utnyttjade vi att α^* och β^* är oberoende!) Alltså är

$$\frac{m_0^* - m_0}{S\sqrt{\frac{1}{n} + \frac{(x_0-\overline{x})^2}{S_{xx}}}} \text{ är } t(n-2).$$

Ett $1-\gamma$ konfidensintervall ges av

$$m_0 \in m_0^* \pm t_{\gamma/2} S\sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}.$$