

Sannolikhet och statistik XVI

Johan Jonasson

May 2022

Linjär regression. Mycket vanligt med linjära samband, t.ex.

- Ström som funktion av spänning.
- $\text{Log}(C14)$ som funktion av ålder.
- Rödförskjutning som funktion av avstånd till galax.
- Längd hos dotter som funktion av mammans längd.

Sällan perfekta samband: data kommer inte riktigt att ligga på en linje. Modell:

$$Y_k = a + bx_k + \epsilon_k, \quad k = 1, 2, \dots, n$$

där $\epsilon_1, \epsilon_2, \dots, \epsilon_n$ är oberoende och $N(0, \sigma^2)$. a , b , σ okända parametrar. Talen $x_1, x_2, \dots, x_n$ kan ses som givna fixa tal.

ML-skattning av a och b : Det gäller att $Y_k \sim N(a + bx_k, \sigma^2)$, så

$$f_{Y_k}(y_k) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(y_k - a - bx_k)^2}.$$

Därför blir

$$\begin{aligned} L(a, b, \sigma; y_1, \dots, y_n) &= f_{Y_1, \dots, Y_n}(y_1, \dots, y_n) \\ &= \frac{1}{\sigma^n (2\pi)^{n/2}} e^{-\frac{1}{2\sigma^2} \sum_{k=1}^n (y_k - a - bx_k)^2}. \end{aligned}$$

Att maximera m.a.p. a och b är att minimera

$$\sum_{k=1}^n (y_k - a - bx_k)^2 =: \mathcal{L}(a, b).$$

Alltså: minsta-kvadrat-metoden.

Vi ska lösa

$$\frac{\partial \mathcal{L}}{\partial a} = -2 \sum_{k=1}^n (y_k - a - bx_k) = 0$$

$$\frac{\partial \mathcal{L}}{\partial b} = -2 \sum_{k=1}^n x_k (y_k - a - bx_k) = 0.$$

Detta ger

$$\hat{b} = \frac{\sum_{k=1}^n (y_k - \bar{y})(x_k - \bar{x})}{\sum_{k=1}^n (x_k - \bar{x})^2}$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x}.$$

Man brukar skriva

$$S_{xx} = \sum_{k=1}^n (x_k - \bar{x})^2, \quad S_{xy} = \sum_{k=1}^n (y_k - \bar{y})(x_k - \bar{x}), \quad S_{yy} = \sum_{k=1}^n (y_k - \bar{y})^2.$$

Då blir

$$\hat{b} = \frac{S_{xy}}{S_{xx}}.$$

Man kan beräkna att

$$\hat{b} \sim N\left(b, \frac{\sigma^2}{S_{xx}}\right), \hat{a} \sim N\left(a, \frac{\sigma^2 \sum_{k=1}^n x_k^2}{nS_{xx}}\right).$$

Om σ^2 känd, gör konfidensintervall/test för a resp b med dessa.
T.ex.

$$\frac{\hat{b} - b}{\sigma / \sqrt{S_{xx}}} \sim N(0, 1)$$

så

$$b \in \hat{b} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{S_{xx}}} \quad (1 - \alpha).$$

Om σ^2 är okänd, ersätt med s^2 och få

$$\frac{\hat{b} - b}{s/\sqrt{S_{xx}}} \sim t_{n-2}$$

och

$$b \in \hat{b} \pm F_{t_{n-2}}^{-1}(1 - \alpha/2) \frac{s}{\sqrt{S_{xx}}} \quad (1 - \alpha).$$

Analogt för a (men inte lika intressant).

Ok, så vad är s^2 ? Jo,

$$s^2 = \frac{1}{n-2} \sum_{k=1}^n (y_k - \hat{a} - \hat{b}x_k)^2.$$

Den är vvr för σ^2 . Formel:

$$s^2 = \frac{1}{n-2} \left(S_{yy} - \frac{S_{xy}^2}{S_{xx}} \right).$$

Formler som gör det lättare att räkna:

$$S_{xy} = \sum (x_k - \bar{x})(y_k - \bar{y}) = \sum x_k y_k - \frac{1}{n} \left(\sum x_k \right) \left(\sum y_k \right).$$

Då får vi också

$$S_{xx} = \sum (x_k - \bar{x})^2 = \sum x_k^2 - \frac{1}{n} \left(\sum x_k \right)^2$$

och analogt för y .

Exempel: För att observera hur en kvinnas längd beror av hennes mors längd, mätte vi sex mödra-dotterpar. Vi tror på ett linjärt samband. Data:

(170.1, 173.4), (161.3, 162.6), (177.1, 170.5)

(167.0, 162.8), (168.6, 171.2), (162.2, 165.5)

Vi får

$$S_{xy} = \sum x_k y_k - \frac{1}{n} \left(\sum x_k \right) \left(\sum y_k \right) = 91.323$$

$$S_{xx} = \sum x_k^2 - \frac{1}{n} \left(\sum x_k \right)^2 = 166.628.$$

Detta ger

$$\hat{b} = \frac{S_{xy}}{S_{xx}} = 0.548.$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x} = 75.74.$$

Den skattade regressionslinjen blir alltså

$$y = 75.74 + 0.548x.$$

Vi får också

$$s^2 = \frac{1}{4} \sum_{k=1}^6 (y_k - \hat{a} - \hat{b}x_k)^2 = 14.34.$$

Symmetriskt 95% konfintervall för b :

$$b \in \hat{b} \pm F_{t_4}^{-1}(0.975) \frac{s}{\sqrt{S_{xx}}} = 0.548 \pm 2.776 \frac{\sqrt{14.34}}{\sqrt{166.6}} = 0.55 \pm 0.81.$$

Om vi gör ett test av $H_0 : b = 0$ mot $H_A : b \neq 0$ kan vi alltså inte förkasta H_0 på 5% signnivå. □

Exempel: Vid Virginia Polythèque mätte man mängden syre, y , som krävs för nedbrytning av partiklar som funktion av mängden utsläppta partiklar, x , vid garvning. Linjärt samband. Siffror:

$$n = 33, \sum x_k y_k = 41355, \sum x_k^2 = 41086, \sum y_k^2 = 43117$$

$$\sum x_k = 1104, \sum y_k = 1124.$$

Skatta reglinjen $y = a + bx$ och ge 95% konfintervall för b .

$$S_{xy} = \sum x_k y_k - \frac{1}{n} \left(\sum x_k \right) \left(\sum y_k \right) = 41355 - \frac{1}{33} \cdot 1104 \cdot 1124 = 3572.$$

$$S_{xx} = \sum x_k^2 - \frac{1}{n} \left(\sum x_k \right)^2 = 41086 - \frac{1}{33} \cdot 1104^2 = 4152.$$

Detta ger:

$$\hat{b} = \frac{S_{xy}}{S_{xx}} = \frac{3572}{4152} = 0.86.$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x} = \frac{1}{33}(1124 - 0.86 \cdot 1104) = 5.29.$$

Regressionslinjen:

$$y = 5.29 + 0.86x.$$

Konfintervall: Vi har $F_{t_{31}}^{-1}(0.975) = 2.04$. Vi behöver s .

$$S_{yy} = \sum y_k^2 - \frac{1}{n} \left(\sum y_k \right)^2 = 43117 - \frac{1}{33} 1124^2 = 4833.$$

$$s^2 = \frac{1}{n-2} \left(S_{yy} - \frac{S_{xy}^2}{S_{xx}} \right) = \frac{1}{31} \left(4833 - \frac{3572^2}{4152} \right) = 56.74.$$

Alltså

$$b = 0.86 \pm 2.04 \cdot \frac{\sqrt{56.74}}{\sqrt{4152}} = 0.86 \pm 0.24 \quad (95\%).$$



Prediktion i linjär regression. Antag att vi vill ge ett prediktionsintervall för en ny observation

$$Y = a + bx + \epsilon$$

för ett givet värde x . Rimlig punktguessning: $\hat{a} + \hat{b}x$. Skriv

$$D = Y - (\hat{a} + \hat{b}x).$$

Vi har $\mathbb{E}[D] = 0$. Vad är $\mathbb{V}\text{ar}(D)$? Eftersom $\hat{a} = \bar{Y} - \hat{b}\bar{x}$, får vi

$$D = Y - \bar{Y} - \hat{b}(x - \bar{x}).$$

Obs att Y och $\bar{Y}$ är oberoende, ty Y är en ny obs. Därför är också Y och $\hat{b}$ oberoende. Det gäller också att $\mathbb{C}\text{ov}(\hat{b}, \bar{Y}) = 0$, ty

$$\begin{aligned} \mathbb{C}\text{ov}(\bar{Y}, S_{xy}) &= \mathbb{C}\text{ov}\left(\bar{Y}, \sum x_k Y_k - \frac{1}{n} \left(\sum x_k\right) \left(\sum Y_k\right)\right) \\ &= \sum x_k \mathbb{C}\text{ov}(\bar{Y}, Y_k) - \bar{x} \sum \mathbb{C}\text{ov}(\bar{Y}, Y_k) \\ &= \frac{1}{n} \sigma^2 \sum x_k - \bar{x} \sigma^2 = 0. \end{aligned}$$

Vi får

$$\mathbb{V}\text{ar}(D) = \sigma^2 + \frac{1}{n}\sigma^2 + (x - \bar{x})^2 \mathbb{V}\text{ar}(\hat{b}) = \left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}\right) \sigma^2.$$

Alltså

$$\frac{D}{\sigma \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}}} \sim N(0, 1).$$

Man kan ana att

$$\frac{D}{s \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}}} \sim t_{n-2}.$$

Detta ger ett prediktionsintervall

$$Y \in \hat{a} + \hat{b}x \pm F_{t_{n-2}}^{-1}(1 - \alpha/2) s \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}}$$

med prediktionsgrad $1 - \alpha$.

Exempel: Gör ett 90% prediktionsintervall för en ny observation Y vid $x = 40$ i Virginia Polythèque-exemplet. Vi har $F_{t_{31}}^{-1}(0.95) = 1.696$, $\bar{x} = 1104/33 = 33.45$, så $x - \bar{x} = 6.55$ så

$$Y \in 5.29 + 0.86 \cdot 40 \pm 1.696 \sqrt{56.74} \sqrt{\frac{34}{33} + \frac{6.55^2}{4152}} = 36.7 \pm 13.0.$$



Exempel: Gör en linjär regression med data $(1, 1)$, $(2, 2)$, $(3, 5)$, ett 90 % konfidensintervall för b , ett 90% prediktionsintervall för $x = 4$ och ett 90% prediktionsintervall för $x = 2$.

Vi har

$$n = 3, \sum x_k = 6, \sum y_k = 8, \sum x_k y_k = 20, \sum x_k^2 = 14, \sum y_k^2 = 30$$

Vi får

$$S_{xy} = 20 - \frac{1}{3} \cdot 8 \cdot 6 = 4, S_{xx} = 14 - \frac{1}{3} 6^2 = 2, S_{yy} = 30 - \frac{1}{3} 8^2 = \frac{26}{3}.$$

Alltså

$$\hat{b} = \frac{4}{2} = 2$$

och

$$\hat{a} = \bar{y} - \hat{b}\bar{x} = \frac{1}{3}(8 - 2 \cdot 6) = -\frac{4}{3}.$$

Regressionslinjen blir

$$y = -\frac{4}{3} + 2x.$$

Vidare är

$$s^2 = \frac{1}{n-2} \left(S_{yy} - \frac{S_{xy}^2}{S_{xx}} \right) = \frac{26}{3} - \frac{4^2}{2} = \frac{2}{3}.$$

Konfintervallet är

$$b \in \hat{b} \pm F_{t_1}^{-1}(0.95) \frac{s}{\sqrt{S_{xx}}} = 2 \pm 6.31 \frac{\sqrt{2/3}}{\sqrt{2}} = 2 \pm 3.64 \quad (90\%).$$

Prediktionsintervall för $x = 4$: Det gäller att $x - \bar{x} = 2$, så det 90%-iga predintervallet blir

$$\begin{aligned} Y \in \hat{a} + \hat{b}x \pm F_{t_{n-2}}^{-1}(0.95)s \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}} &= -\frac{4}{3} + 2 \cdot 4 \pm 6.31 \sqrt{\frac{2}{3}} \sqrt{\frac{4}{3} + \frac{2^2}{2}} \\ &= 6.7 \pm 9.4. \end{aligned}$$

Prediktionsintervall för $x = 2$: Här är $x - \bar{x} = 0$, i övrigt samma siffror, så vi får

$$\begin{aligned} Y \in \hat{a} + \hat{b}x \pm F_{t_{n-2}}^{-1}(0.95)s\sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}} &= -\frac{4}{3} + 2 \cdot 2 \pm 6.31\sqrt{\frac{2}{3}}\sqrt{\frac{4}{3}} \\ &= 2.7 \pm 6.0. \end{aligned}$$

Vi ser att det är svårare att prediktera med extrapolation än med interpolation.