Regressions- och tidsserieanalys

Föreläsning 3 - Regression som sannolikhetsmodell

Mattias Villani

Statistiska institutionen
Stockholms universitet

Institutionen för datavetenskap
Linköpings universitet


mattiasvillani.com


@matvil


mattiasvillani

Översikt

- Regression som sannolikhetsmodell
- Konfidensintervall
- Hypotestest
- Prediktionsintervall

Repetition sannolikhetsmodeller

- Underliggande **populationsmodell**:

$$X_1, \ldots, X_n \overset{\text{ober}}{\sim} N(\mu, \sigma^2), \quad \sigma^2 \text{ känd}$$

- Medelvärdet

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

är en **estimator** för μ.

- **Väntevärdesriktig** (rätt i genomsnitt över alla möjliga stickprov)

$$\mathbb{E}(\bar{X}) = \mu$$

- **Samplingfördelningen** (hur medelvärdet varierar från stickprov till stickprov):

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Regression som sannolikhetsmodell

- Underliggande **populationsmodell** för regression:

$$y = \alpha + \beta x + \varepsilon, \quad \varepsilon \sim N(0, \sigma_\varepsilon^2)$$

- Regression är en modell för den **betingade fördelningen**

$$y|x \sim N\left(\mu_{y|x}, \sigma_\varepsilon^2\right)$$

där det betingade väntevärdet för y nu beror på x genom regressionen

$$\mu_{y|x} = \alpha + \beta x$$

- α är interceptet i den underliggande populationen.
- β är lutningen på regressionslinjen i den underliggande populationen.

Regression som sannolikhetsmodell

- Stickprov/datamaterial med n observationspar $(y_1, x_1), \ldots (y_n, x_n)$.
- Vanligt att anta oberoende feltermer ε för alla observationer:

$$\varepsilon_1, \ldots, \varepsilon_n \overset{\text{ober}}{\sim} N(0, \sigma_\varepsilon^2)$$

- Modell för hela stickprovet

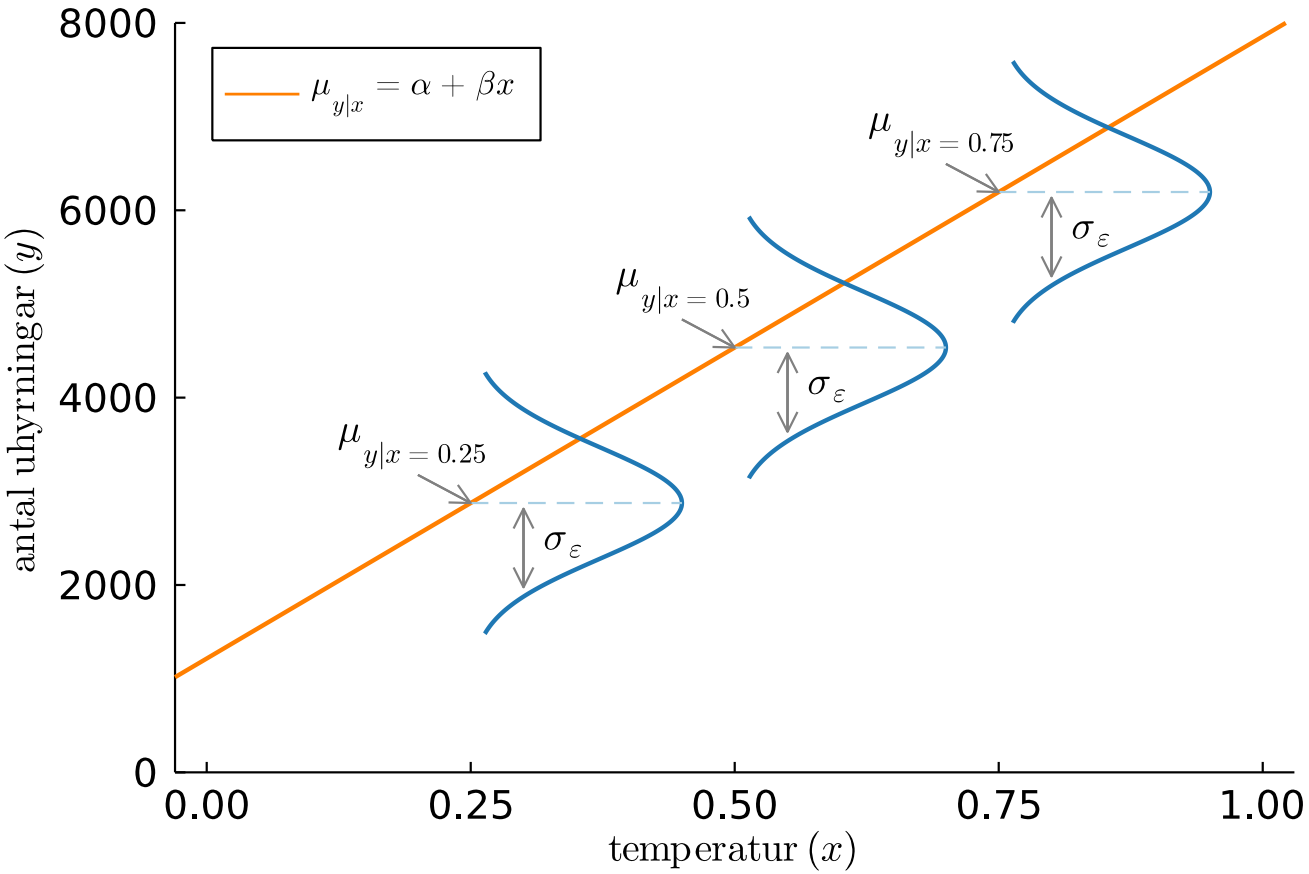
$$y_i = \alpha + \beta x_i + \varepsilon_i, \quad \varepsilon_i \overset{\text{ober}}{\sim} N(0, \sigma_\varepsilon^2)$$

Regression som sannolikhetsmodell

- Regression som modell för betingad fördelning

$$y|x \sim N\left(\mu_{y|x}, \sigma_\varepsilon^2\right)$$

$$\mu_{y|x} = \alpha + \beta x$$



Simulera data

- Simulera regressionsdata med stickprovstorlek n:
  - Bestäm populationens parametrar β_0, β_1 och σ^2.
  - Bestäm $x_1, \ldots, x_n$(som antas vara icke-slumpmässiga)
  - Simulera feltermer $\varepsilon_1, \ldots, \varepsilon_n$ från $N(0, \sigma^2)$.
  - Beräkna $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ för varje observation.

Samplingfördelning - minstakvadratskattningen

- Minstakvadratestimatorerna

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

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

$$s_e^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$$

- **Väntevärdesriktiga**

$$\mathbb{E}(b) = \beta$$
$$\mathbb{E}(a) = \alpha$$
$$\mathbb{E}(s_e^2) = \sigma_\varepsilon^2$$

Samplingfördelning för b

- Estimatorn för lutningskoefficienten

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

har **samplingvarians** (hur mycket varierar b över olika stickprov)

$$\sigma_b^2 = \frac{\sigma_\varepsilon^2}{\sum(x_i - \bar{x})^2}$$

- En estimator av den teoretiska samplingvariansen σ_b^2 är

$$s_b^2 = \frac{s_e^2}{\sum(x_i - \bar{x})^2}$$

- Se AJÅ för en motsvarande formel för att skatta samplingvariansen för a.
- Hälsobudgetdata

$$s_b^2 = \frac{4.467}{52.861} = 0.085 \qquad s_b \approx \sqrt{0.085} \approx 0.291$$

Approximativt konfidensintervall för b

- Approximativt 95% konfidensintervall för b för **stora stickprov** ($n \geq 30$)

$$[b - 1.96 \cdot s_b, b + 1.96 \cdot s_b]$$

- Hälsobudgetdata

$$[1.038 - 1.96 \cdot 0.291, 1.038 + 1.96 \cdot 0.291] = [0.468, 1.608]$$

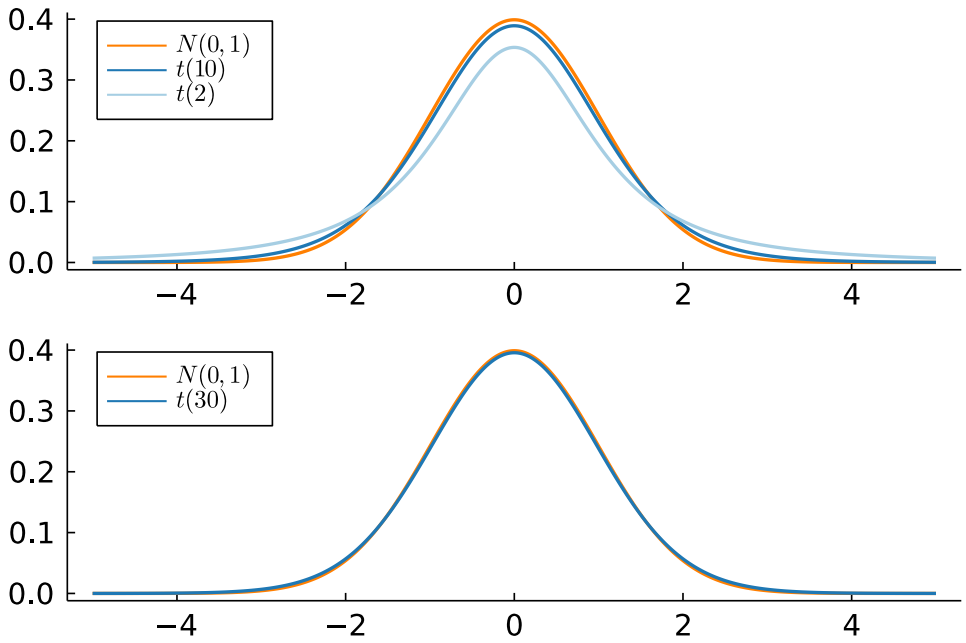
- **I 95% av alla stickprov från populationen** täcker intervallet $[0.468, 1.608]$ den sanna lutningen β.

Exakt konfidensintervall för b - student t

- För **små** n är normalapproximationen inte tillräckligt bra.
- Estimatorn b följer en **t-fördelning** med $n-2$ **frihetsgrader**.

$$\frac{b-\beta}{s_b} \sim t(n-2)$$

- För $n \to \infty$ blir t-fördelningen alltmer lik normalfördelningen.
- t-fördelningen konvergerar mot normalfördelningen när $n \to \infty$.



Exakt konfidensintervall för b - student t

- Exakt 95% konfidensintervall för b

$$[b - t_{0.025}(n-2) \cdot s_b, b + t_{0.025}(n-2) \cdot s_b]$$

- $t_{0.025}(n-2)$ är det värde som har 0.025 (2.5%) sannolikhetsmassa till vänster om sig i t-fördelningen med $n-2$ frihetsgrader.
- TO BE CONTINUED!