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

- Underliggande **populationsmodell** för regression:

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

- Stickprov/datamaterial med n observationspar

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

- Vanligt att anta **oberoende feltermer** ε för alla observationer:

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

- Antar oftast också samma varians. **Homoskedastisk** modell.
- Modell för hela stickprovet

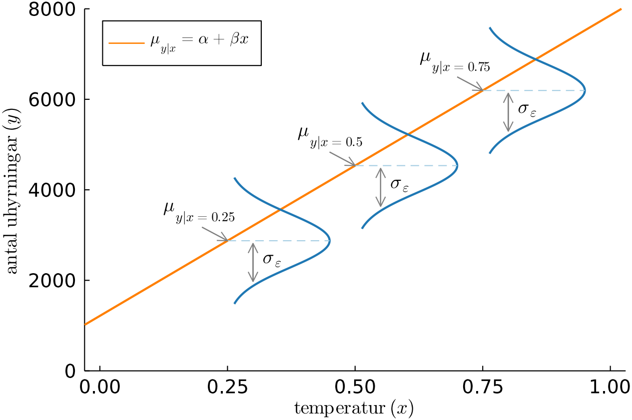
$$y_i = \alpha + \beta x_i + \varepsilon_i, \quad \varepsilon_i \overset{\text{ober}}{\sim} N(0, \sigma_\varepsilon^2)$$

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 α, β och σ_ε^2.
  - Bestäm $x_1, \ldots, x_n$ (som antas vara icke-slumpmässiga)
  - Simulera feltermer $\varepsilon_1, \ldots, \varepsilon_n$ från $N(0, \sigma_\varepsilon^2)$.
  - Beräkna $y_i = \alpha + \beta x_i + \varepsilon_i$ för varje observation.

```
library(regkurs)
simdata <- regsimulate(n = 500, betavect = c(1, -2, 1, 0), sigma_eps = 2)
lmfit <- lm(y ~ X1 + X2 + X3, data = simdata)
regsummary(lmfit, anova = F)
```

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 β i stora stickprov** $(n \geq 30)$

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

- **I 95% av alla stickprov från populationen** täcker intervallet $[b - 1.96 \cdot s_b, b + 1.96 \cdot s_b]$ den sanna lutningen β.
- Hälsobudgetdata

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

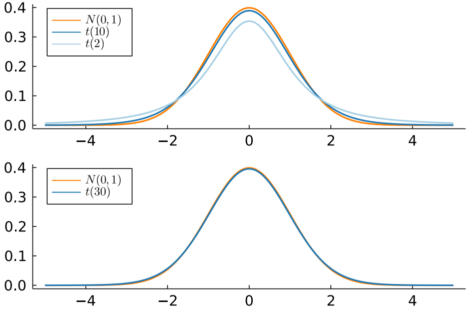
- Intervallet $[0.468, 1.608]$ täcker eller täcker inte det sanna värdet β. Vi vet inte vilket.

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 $n \to \infty$.

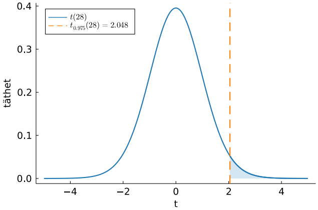


Exakt konfidensintervall för b - student t

- **Exakt 95% konfidensintervall för β**

$$\left[b - t_{0.975}(n-2) \cdot s_b, b + t_{0.975}(n-2) \cdot s_b\right]$$

- t-fördelningen med $n-2$ frihetsgrader har 0.975 (97.5%) sannolikhetsmassa till vänster om värdet $t_{0.975}(n-2)$.



- Hälsobudgetdata: $n = 28$, och $t_{0.975}(28) = 2.0484$ från tabell.
- Exakt 95% konfidensintervall för b

$$[1.038 - 2.0484 \cdot 0.291, 1.038 + 2.0484 \cdot 0.291] = [0.442, 1.634]$$

Exakt konfidensintervall - R

```
fit = lm(lifespan ~ spending, data = healthbudget) # skattar regression
regsummary(fit, conf_intervals = TRUE) # med konfidensintervall
```

```
 Parameter estimates
 ------------------------------------------------
             Estimate Std. Error t value   Pr(>|t|)    2.5 %  97.5 %
 (Intercept)  76.0350    0.95084 79.9663 1.3416e-34 74.08732 77.9827
 spending      1.0376    0.29071  3.5691 1.3164e-03  0.44209  1.6331
```

Hypotesttest för β

- **Hypotestest för lutningen** i regressionen

$$H_0 : \beta = 0$$
$$H_1 : \beta \neq 0$$

- **Teststatistiska**

$$t = \left| \frac{b - 0}{s_b} \right|$$

- Vi förkastar nollhypotesten på signifikansnivån $\alpha = 0.05$ om

$$t_{\text{obs}} > t_{\text{crit}}$$

där det kritiska värdet t_{crit} hämtas från tabell:

$$t_{\text{crit}} = t_{0.975}(n - 2)$$

- **P-värde** = sannolikheten att observera t_{obs} eller något ännu mer extremt **givet att H_0 är sann**.
- Under H_0 har vi att $t \sim t(n - 2)$.

Hypotesttest för β - hälsobudgetdata

- $n = 30$, så $n - 2 = 28$, och $t_{\text{crit}} = t_{0.975}(28) = 2.0484$.

$$t_{\text{obs}} = \left| \frac{1.038 - 0}{0.291} \right| = 3.567$$

- Eftersom $t_{\text{obs}} > t_{\text{crit}}$ så förkastar vi nollhypotesen på 5% signifikansnivå.

- Vi förkastar nollhypotesen att hälsobudgetens storlek inte är korrelerad med livslängd.

- Testets p-värde

$$p = 0.0013237$$

vilket visar att vi t o m skulle ha förkastat på 1% nivån.

Hypotesttest för β - hälsobudgetdata

```
Analysis of variance - ANOVA
-----------------------------------------------
      df      SS      MS      F    Pr(>F)
Regr   1  56.907 56.9072 12.739 0.0013164
Error 28 125.082  4.4672
Total 29 181.990

Measures of model fit
-----------------------------------------------
Root MSE       R2   R2-adj
 2.11358  0.31269  0.28815

Parameter estimates
-----------------------------------------------
            Estimate Std. Error t value   Pr(>|t|)
(Intercept)  76.0350    0.95084 79.9663 1.3416e-34
spending      1.0376    0.29071  3.5691 1.3164e-03
```

Terminologi

- **Målvariabel** (y) kallas ofta **beroende variabel**. Även **responsvariabel**.
- **Förklarande variabel** (x) kallas ofta **kovariat**. Även **prediktor** eller **feature**.
- MSE $= s_e^2$. Residual varians.
- Root MSE $= \sqrt{\mathrm{MSE}}$, dvs s_e. Residualstandardavvikelse.

Konfidensintervall för regressionslinjen [1]

..

- **Regressionslinjen i populationen** är

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

som skattas med minsta kvadratmetoden genom formeln

$$\hat{\mu}_{y|x} = a + bx$$

- **Standardavvikelsen för skattningen av regressionslinjen** vid ett givet x-värde $x = x_0$ kan skattas med

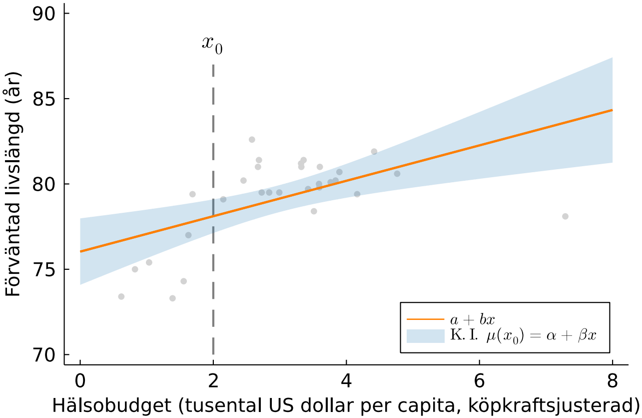
$$s_{\hat{\mu}_{y|x_0}} = s_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

- **95% konfidensintervall för regressionlinjen** $\mu_{y|x} = \alpha + \beta x$

$$\hat{\mu}_{y|x_0} \pm t_{0.975}(n-2) \cdot s_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

[1] Videon **Prediktionsintervall.mp4** finns i Videos mappen på Athena.

Konfidensintervall för regressionslinjen



Prediktionsintervall

- Antag att vi gjort en prognos vid punkten $x = x_0$.
- Prognosen är
$$\hat{y}(x_0) = \hat{\mu}_{y|x_0} = a + bx_0$$
- **Prognosintervall** för $\hat{y}(x_0)$ - **två källor av osäkerhet**:
  - De **okända parametrarna** α och β, dvs osäkerhet om $\mu_{y|x}$.
  - **Variationen i de enskilda y-värdena kring regressionlinjen** $\mu_{y|x}$. Alla observationer "träffas av ett ε" som har varians σ_ε^2.
- Prognosvariansen:
$$\sigma^2_{\hat{y}(x_0)} = \sigma^2_{\hat{\mu}_{y|x_0}} + \sigma^2_\varepsilon$$
- 95%-igt prognosintervall för en enskild observation vid $x = x_0$
$$\hat{y}(x_0) \pm t_{0.975}(n-2) \cdot s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

Prediktionsintervall