

# Regressions- och tidsserieanalys

## Föreläsning 3 - Regression som sannolikhetsmodell

**Mattias Villani**

Statistiska institutionen  
Stockholms universitet

Institutionen för datavetenskap  
Linköpings universitet



[mattiasvillani.com](http://mattiasvillani.com)



@matvil



[mattiasvillani](https://github.com/mattiasvillani)

- Regression som sannolikhetsmodell
- Konfidensintervall
- Hypotestest
- Prediktionsintervall

# Repetition sannolikhetsmodeller

- Underliggande **populationsmodell**:

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

- Medelvärdet

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

är en **estimator** för  $\mu$ .

- **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**

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

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

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

- $\alpha$  är interceptet i den underliggande populationen.
- $\beta$  är lutningen på regressionslinjen i den underliggande populationen.

# Regression som sannolikhetsmodell

- Stickprov/datamaterial med  $n$  observationspar

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

- Vanligt att anta oberoende feltermen  $\varepsilon$  för alla observationer:

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

- Antar oftast också samma varians
- Modell för hela stickprovet

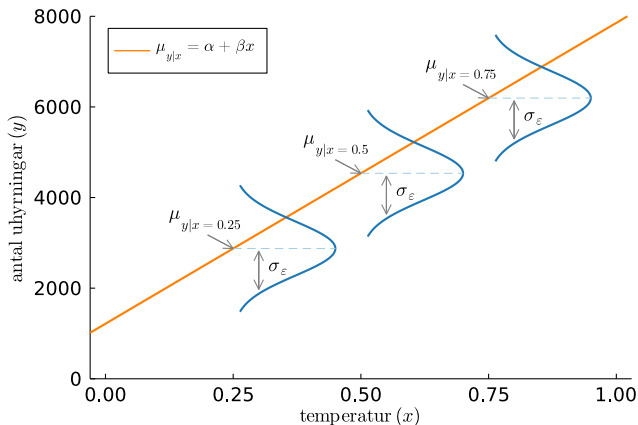
$$y_i = \alpha + \beta x_i + \varepsilon_i, \quad \varepsilon_i \overset{\text{ober}}{\sim} N(0, \sigma_\varepsilon^2)$$

# Regression som sannolikhetsmodell

- Regression som modell för betingad fördelning

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

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



- Simulera regressionsdata med stickprovstorlek  $n$ :
  - ▶ Bestäm populationens parametrar  $\alpha$ ,  $\beta$  och  $\sigma_\varepsilon^2$ .
  - ▶ Bestäm  $x_1, \dots, x_n$  (som antas vara icke-slumpmässiga)
  - ▶ Simulera feltermerna  $\varepsilon_1, \dots, \varepsilon_n$  från  $N(0, \sigma_\varepsilon^2)$ .
  - ▶ Beräkna  $y_i = \alpha + \beta x_i + \varepsilon_i$  för varje observation.

# Samplingfördelning - minstakvadratskattningen

## ■ Minstakvadratetestimatorerna

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

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

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

## ■ 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  $\sigma_b^2$  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

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

# Approximativt konfidensintervall för $b$

- **Approximativt 95% konfidensintervall för  $\beta$  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  $\beta$ .
- Hälsobudgetdata

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

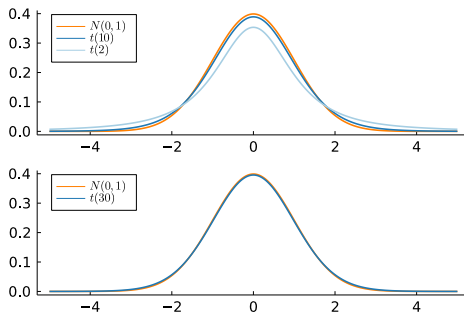
- Intervallet  $[0.468, 1.608]$  täcker eller täcker inte det sanna värdet  $\beta$ . Vi vet inte vilket.

# Exakt konfidensintervall för $b$ - student $t$

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

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

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

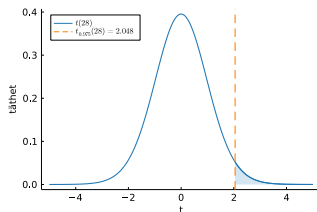


# Exakt konfidensintervall för $b$ - student $t$

## ■ Exakt 95% konfidensintervall för $\beta$

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

- $t$ -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 $b$

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

# Hypotesttest för $\beta$

- Hypotesttest 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  $t_{\text{crit}}$  hämtas från tabell:

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

- **P-värde** = sannolikheten att observera  $t_{\text{obs}}$  eller något ännu mer extremt **givet att  $H_0$  är sann**.

- Under  $H_0$  har vi att  $t \sim t(n - 2)$ .

# Hypotesttest för $\beta$ - 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 $\beta$ - hälsobudgetdata

## Hälsobudget - regression

The REG Procedure

Model: MODEL1

Dependent Variable: lifespan

Number of Observations Read	30
-----------------------------	----

Number of Observations Used	30
-----------------------------	----

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	56.90723	56.90723	12.74	0.0013
Error	28	125.08244	4.46723		
Corrected Total	29	181.98967			

Root MSE	2.11358	R-Square	0.3127
Dependent Mean	79.13667	Adj R-Sq	0.2881
Coeff Var	2.67080		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	76.03502	0.95084	79.97	<.0001
spending	1	1.03757	0.29071	3.57	0.0013

# 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.
- $Root\ MSE = \sqrt{MSE}$ , dvs  $s_e$ . Residualstandardavvikelse.
- Sum of Squares Regression (SSR) heter **Sum of Squares Model** (SSM) i SAS.



# Konfidensintervall för regressionslinjen

- **Regressionslinjen i populationen** är

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

som skattas med minsta kvadratmetoden genom formeln

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

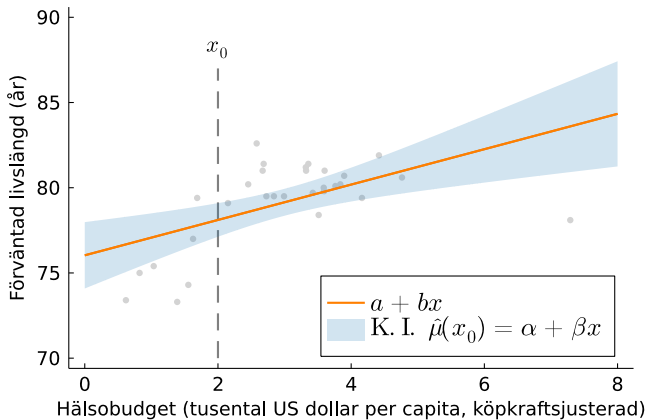
- **Standardavvikelsen för skattningen av regressionslinjen** vid ett givet  $x$ -värde  $x = x_0$  kan skattas med

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

- **95% konfidensintervall för regressionslinjen**  $\hat{\mu}_{y|x} = a + bx$

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

# 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**  $\alpha$  och  $\beta$ , dvs osäkerhet om  $\mu_{y|x}$ .
  - ▶ **Variationen i de enskilda  $y$ -värdena kring regressionlinjen**  $\mu_{y|x}$ . Alla observationer "träffas av ett  $\varepsilon$ " som har varians  $\sigma_\varepsilon^2$ .
- Prognosvariansen:

$$\sigma_{\hat{y}(x_0)}^2 = \sigma_{\hat{\mu}_{y|x_0}}^2 + \sigma_\varepsilon^2$$

- 95%-igt prognosintervall för en enskild observation vid  $x = x_0$

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

# Prediktionsintervall

