Statistik och Dataanalys I

Föreläsning 20 - Inferens i linjär regression - populationsmodell och samplingfördelning

Mattias Villani

Statistiska institutionen
Stockholms universitet

mattiasvillani.com @matvil @matvil mattiasvillani

Översikt

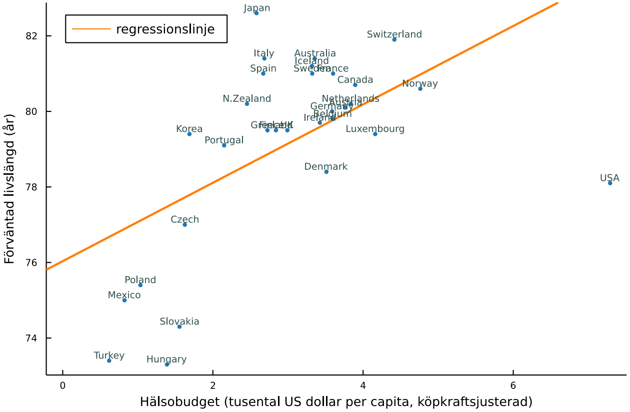
- **Inferens i enkel linjär regression**
- **Regression som sannolikhetsmodell**
- **Samplingfördelning regression**

Samband - hälsovårdsbudget och livslängd

sda1 lifespan

Källa: boken 'Regression and other stories' och OECD.

Regression - hälsovårdsbudget och livslängd



Anpassad regressionslinje och tolkning

- Skattad regressionslinje hälsobudget (x) $\rightarrow$ livslängd (y)

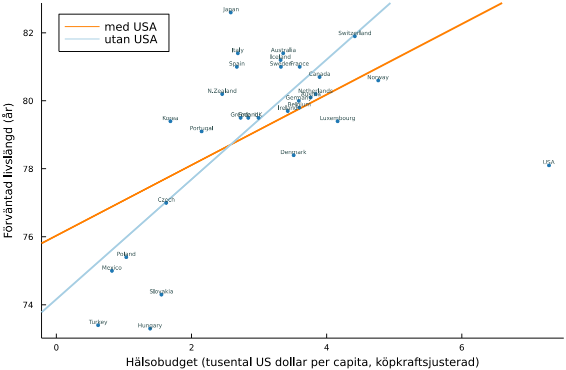
$$\text{lifespan} = 76.035 + 1.03757 \cdot \text{spending}$$

$$\hat{y} = \underbrace{76.035}_{b_0} + \underbrace{1.038}_{b_1} \cdot x$$

- Tolkning **intercept** b_0: **genomsnittlig** livslängd är ca 76 år om spending $= 0$.

- Tolkning **lutning** b_1: **genomsnittlig** livslängd ökar med 1.038 år om `spending` ökar med 1 (tusen US dollar per capita).

Inflytelserika observationer



- Med USA

$$\text{lifespan} = 76.035 + 1.038 \cdot \text{spending}$$

- Utan USA

$$\text{lifespan} = 74.164 + 1.763 \cdot \text{spending}$$

Minsta-kvadrat-metoden

- **Anpassat värde/prediktion** för i:te observationen

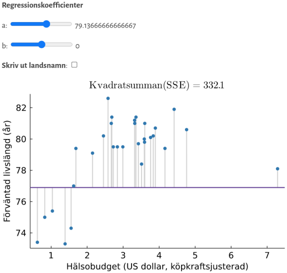
$$\hat{y}_i = b_0 + b_1 x_i$$

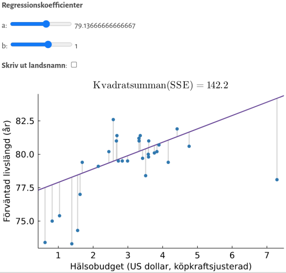
- **Residual**

$$e_i = y_i - \hat{y}_i$$

- **Minsta-kvadrat-skattning**: välj b_0 och b_1 som minimerar

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$





Regression i R

```
> library(sda1)
> lifespan_no_usa = lifespan[1:29,] # ta bort outliern USA
> model = lm(lifespan ~ spending, data = lifespan_no_usa)
> summary(model)

Call:
lm(formula = lifespan ~ spending, data = lifespan_no_usa)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3108 -0.7016 -0.0507  1.1458  3.8860

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  74.1639     0.8782   84.45  < 2e-16 ***
spending      1.7629     0.2890    6.10 1.63e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.678 on 27 degrees of freedom
Multiple R-squared:  0.5795,	Adjusted R-squared:  0.5639
F-statistic: 37.21 on 1 and 27 DF,  p-value: 1.626e-06
```

Residualvarians

- **Residualvariansen** - hur bra regressionslinjen passar data:

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

- Kom ihåg: stickprovsvariansen delar med $n-1$ eftersom vi måste beräkna $\bar{y}$ först:

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

- Residualvariansen delar med $n-2$ eftersom vi måste beräkna både b_0 och b_1 först. **Väntevärdesriktig**.

- **Residualstandardavvikelsen** (`residual standard error` i R)

$$s_e = \sqrt{s_e^2}$$

- Hälsobudgetdata

$$s_e^2 = \frac{76.056}{29-2} \approx 2.817 \qquad s_e = \sqrt{2.817} \approx 1.678 \text{ år}$$

Regression som sannolikhetsmodell

- **Populationsmodell** för enkel regression:

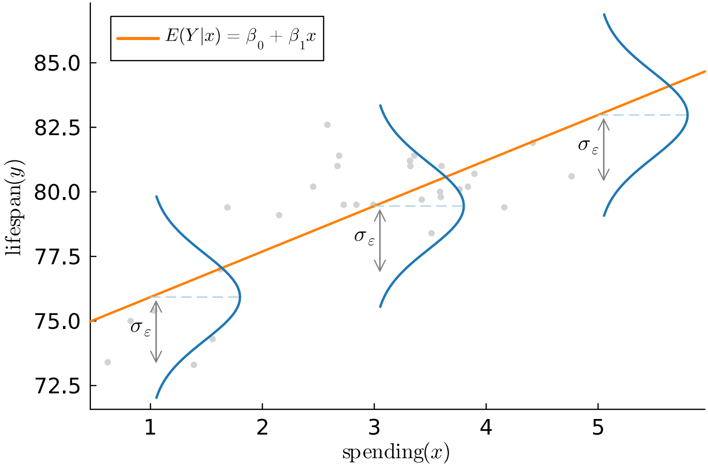
$$Y = \beta_0 + \beta_1 x + \varepsilon, \quad \varepsilon \sim N(0, \sigma_\varepsilon^2)$$

- β_0 är interceptet i populationen/modellen.
- β_1 är lutningen på regressionslinjen i populationen.
- **Regressionlinjen** i populationen är ett **betingat väntevärde**:

$$E(Y|x) = \beta_0 + \beta_1 x$$

- β_1 : hur Y förändras **i genomsnitt** när x ökar med en enhet.
- "i genomsnitt" = (betingat) väntevärde.
- Responsvariabeln y kommer avvika från populationens regressionslinje med en **slumpmässig "felterm"** ε.

Regression som modell för betingad fördelning



Regression som sannolikhetsmodell

- **Populationsmodell** för hela stickprovet:

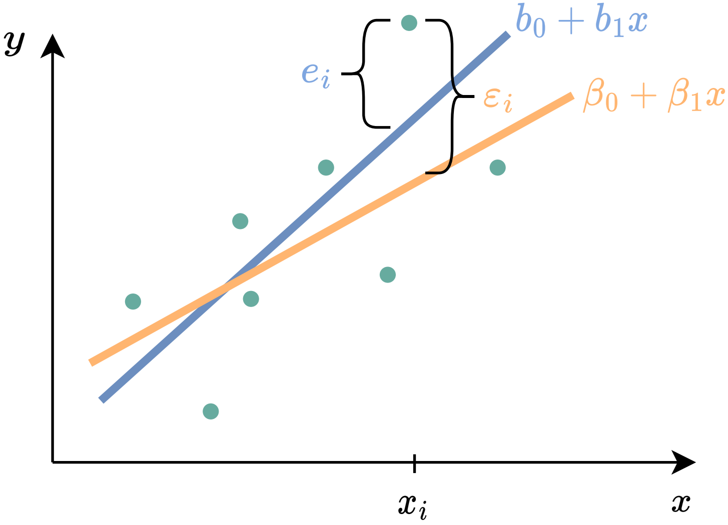
$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma_\varepsilon)$$

- **Stickprov/datamaterial** med n observationspar

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

- I regression antar vi att **x-variabeln inte är slumpmässig**.

Residualerna e_i skattar populationens ε_i



De fyra antaganden om populationen i regression

1. Sambandet mellan y och x är **linjärt**

$$E(Y|x) = \beta_0 + \beta_1 x$$

2. **Feltermerna ε_i är oberoende**

3. Feltermerna har **samma standardavvikelse** (homoskedastisk)

$$SD(\varepsilon_i) = \sigma_\varepsilon$$

4. **Feltermerna är normalfördelade**

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

Residualanalys för att undersöka de 4 antagandena

- **Residualer**:

$$e_i = y_i - \hat{y}_i$$

1. **Linjärt samband?**

   Plotta y_i mot x_i. Ser linjärt ut?
   Plotta e_i mot x_i. Konstant, eller mönster kvar?

2. **Oberoende ε?**

   Plotta residualer e_i mot anpassade värden $\hat{y}_i$.
   Tidsserier: plotta e_i mot tid (observationsnummer).

3. **Homoskedastiska ε?**

   Plotta residualer e_i mot x_i. Liknande spridning för alla x_i?

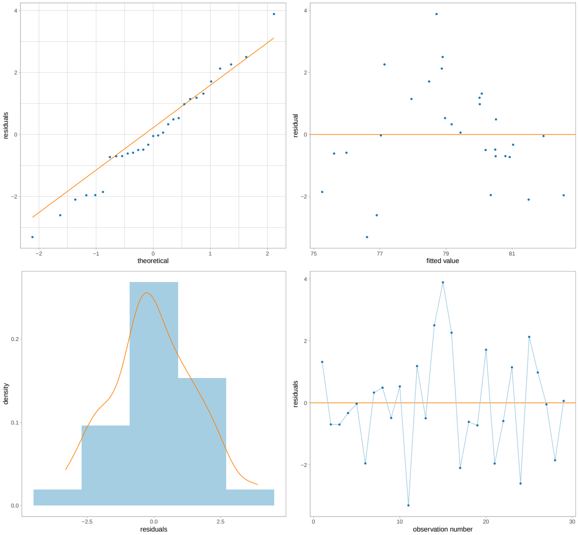
$$SD(\varepsilon_i) = \sigma_\varepsilon$$

4. **Normalfördelade ε?**

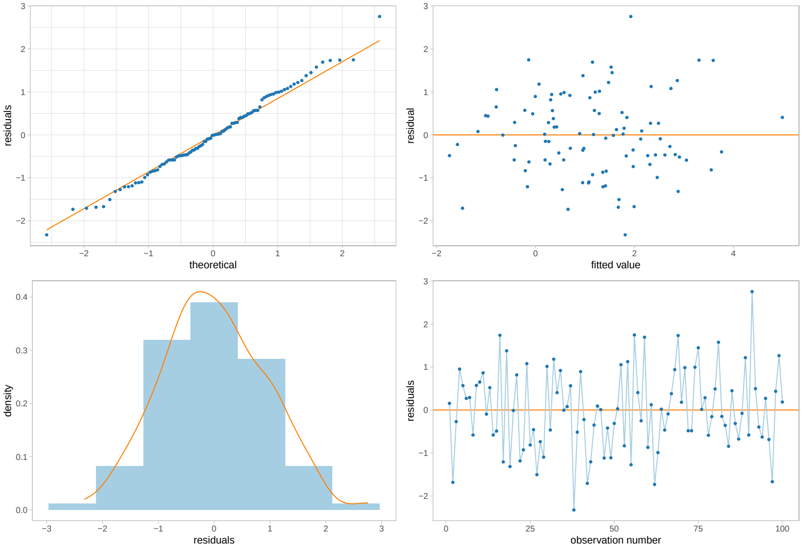
   Histogram, boxplot, QQ-plot för residualer e_i.

Residualanalys lifespan - sda1-paketet

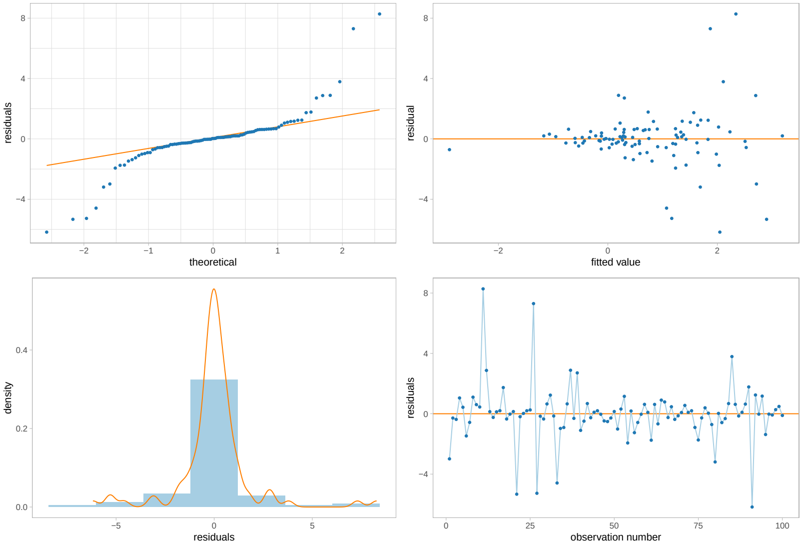
```
> model = lm(lifespan ~ spending, data = lifespan_no_usa)
> reg_residuals(model)
```



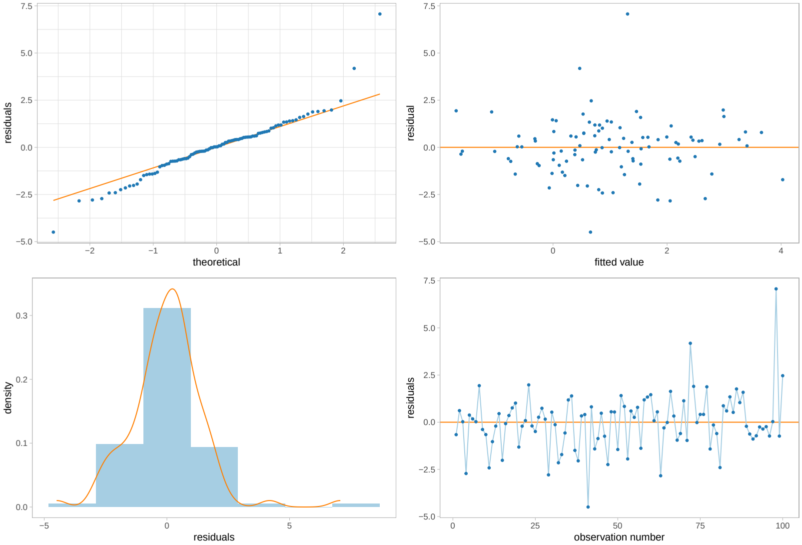
Residualer simulerade data - alla antaganden OK



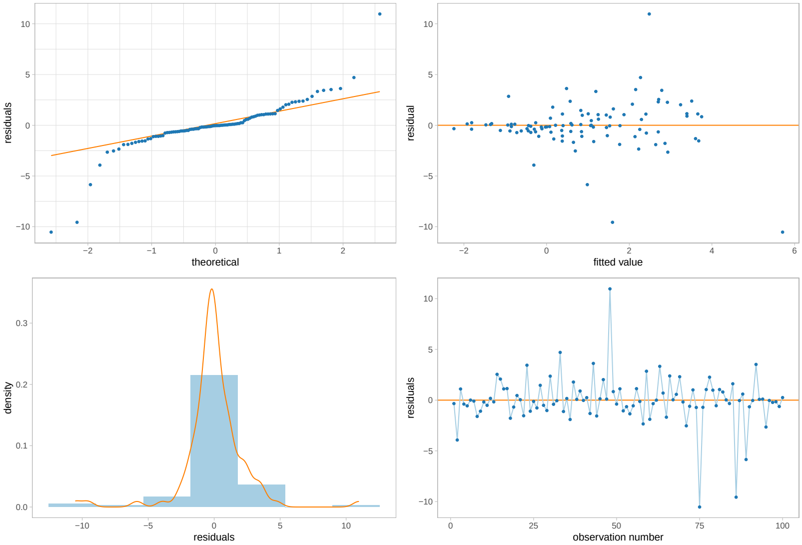
Trouble in paradise 1 - heteroscedastisk varians



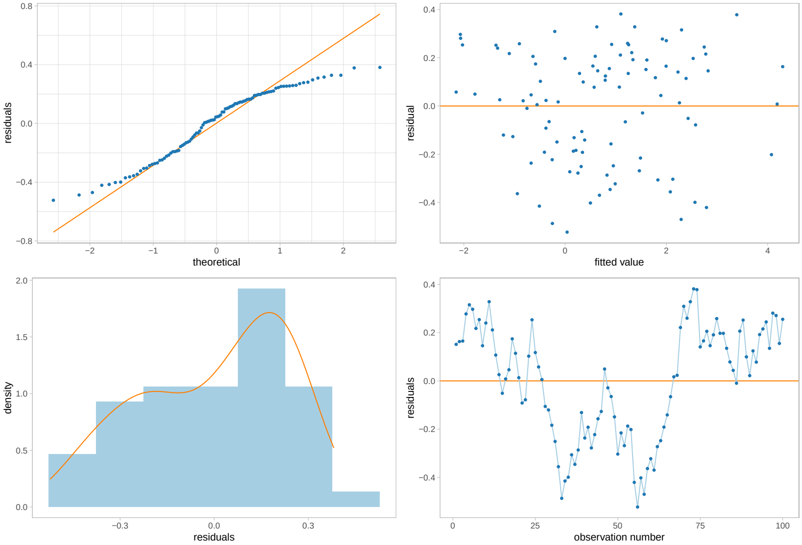
Trouble in paradise 2 - icke-normala ε (outliers)



Trouble in paradise 3 - icke-normala och hetero ε



Trouble in paradise 4 - ej oberoende ε



Minsta-kvadrat-skattningar är väntevärdesriktiga

- Minsta-kvadrat-estimatorerna:

$$b_1 = \frac{s_{xy}}{s_x^2}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

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

- **Väntevärdesriktiga**

$$E(b_0) = \beta_0$$
$$E(b_1) = \beta_1$$
$$E(s_e^2) = \sigma_\varepsilon^2$$

Standardfel för b_1

- Estimatorn för lutningskoefficienten

$$b_1 = \frac{s_{xy}}{s_x^2}$$

- Hur b_1 varierar mellan olika stickprov:

$$\sigma_{b_1} = SD(b_1) = \frac{\sigma_\varepsilon}{\sqrt{n-1}s_x}$$

- σ_{b_1} skattas med **standardfelet**

$$s_{b_1} = SE(b_1) = \frac{s_e}{\sqrt{n-1}s_x}$$

- Formel för $SE(b_0)$ slipper ni på SDA1.😅
- `lifespan data [sd(spending) = 1.097516]`

$$s_{b_1} = \frac{1.678}{\sqrt{29-1} \cdot 1.097516} \approx 0.289$$

Standardfel för b_1 i R

```
> library(sda1)
> lifespan_no_usa = lifespan[1:29,] # ta bort outliern USA
> model = lm(lifespan ~ spending, data = lifespan_no_usa)
> summary(model)

Call:
lm(formula = lifespan ~ spending, data = lifespan_no_usa)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.3108 -0.7016 -0.0507  1.1458  3.8860 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  74.1639     0.8782   84.45  < 2e-16 ***
spending      1.7629     0.2890    6.10 1.63e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.678 on 27 degrees of freedom
Multiple R-squared:  0.5795,	Adjusted R-squared:  0.5639 
F-statistic: 37.21 on 1 and 27 DF,  p-value: 1.626e-06
```

Samplingfördelning i regression - interaktivt

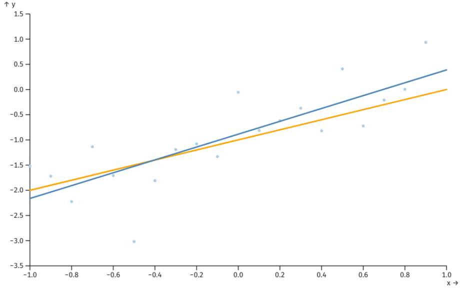
intercept, β_0 -1

lutning, β_1 1

sd felterm, σ_ε 0.5

Play dataset nummer: 207

Skattat intercept: $b_0 = -0.8866$
Skattad lutning: $b_1 = 1.275$



Populationsmodell
Skattad regressionslinje