Statistik och Dataanalys I

Föreläsning 20 - Inferens i linjär regression

Mattias Villani

Statistiska institutionen
Stockholms universitet

mattiasvillani.com @matvil @matvil mattiasvillani

Översikt

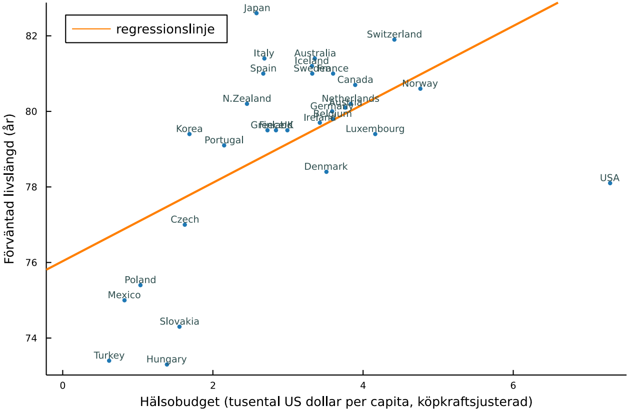
- **Inferens i enkel linjär regression**
- **Regression som sannolikhetsmodell**
- **Prediktionsintervall**
- **Inferens i multipel linjär regression**

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)

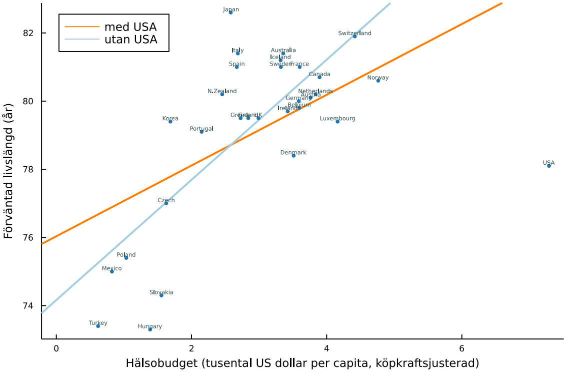
$$\text{lifespan} = 76.035 + 1.03757 \cdot \text{spending}$$

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

- 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

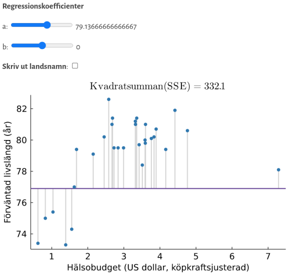
$$\hat{y}_i = b_0 + b_1 x_i$$

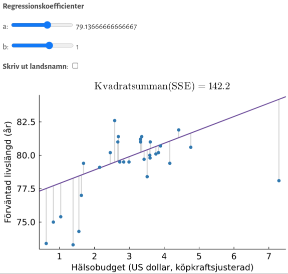
- **Residual**

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

- **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:

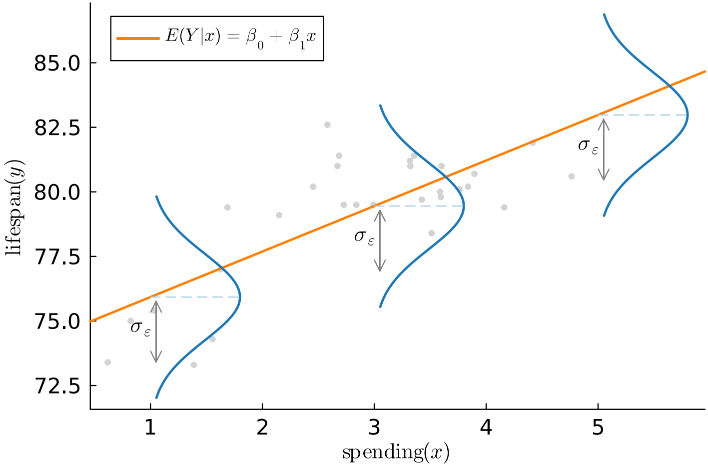
$$Y = \beta_0 + \beta_1 x + \varepsilon, \quad \varepsilon \sim N(0, \sigma_\varepsilon^2)$$

- 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**.

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?

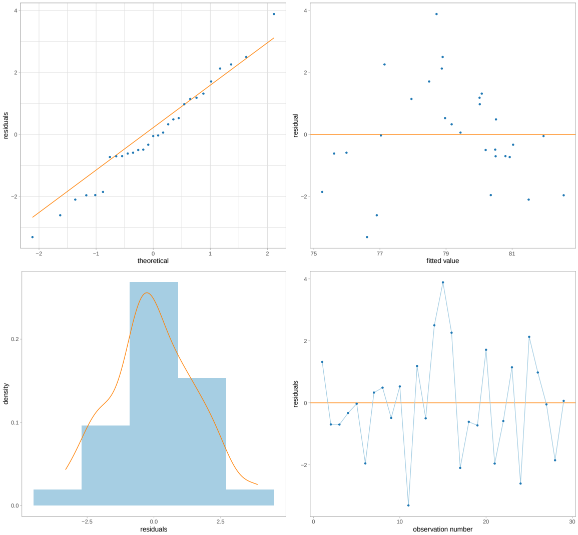
$$SD(\varepsilon_i) = \sigma_\varepsilon$$

4. **Normalfördelade ε?**

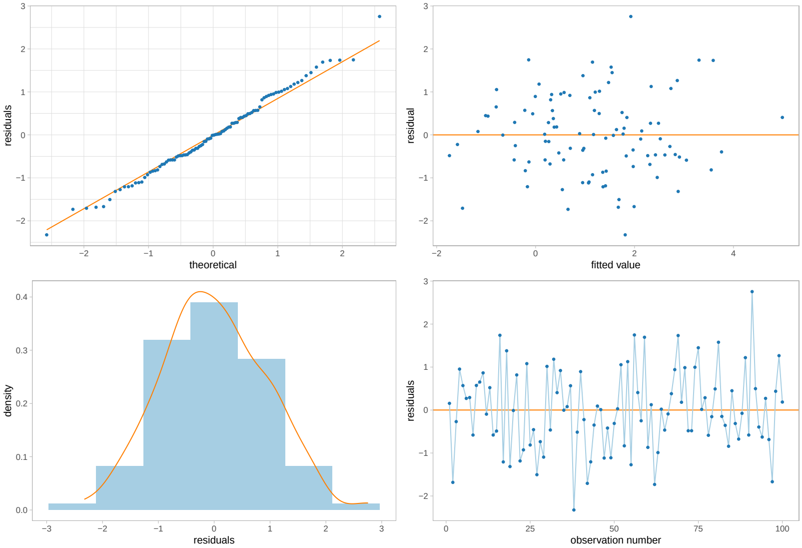
   Histogram, boxplot, QQ-plot för residualer e_i.

Residualanalys lifespan - sda1-paketet

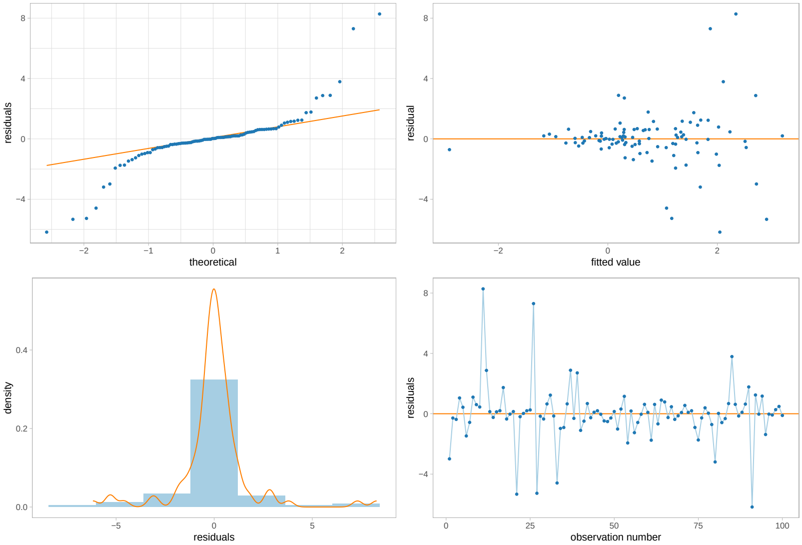
```
> model = lm(lifespan ~ spending, data = lifespan_no_usa)
> reg_residuals(model)
```



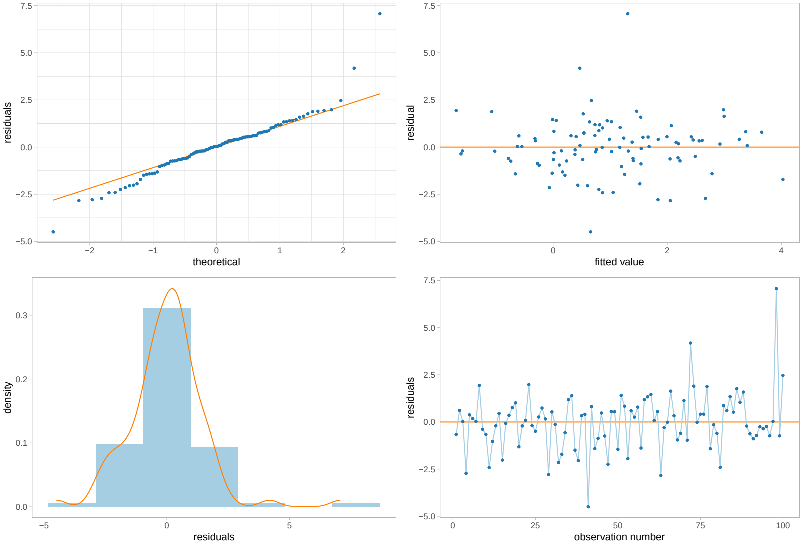
Residualer simulerade data - alla antaganden OK



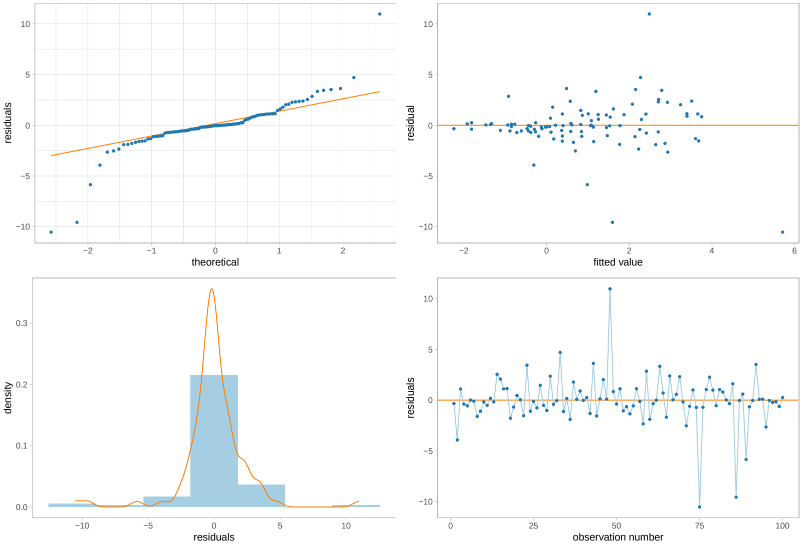
Trouble in paradise 1 - heteroscedastisk varians



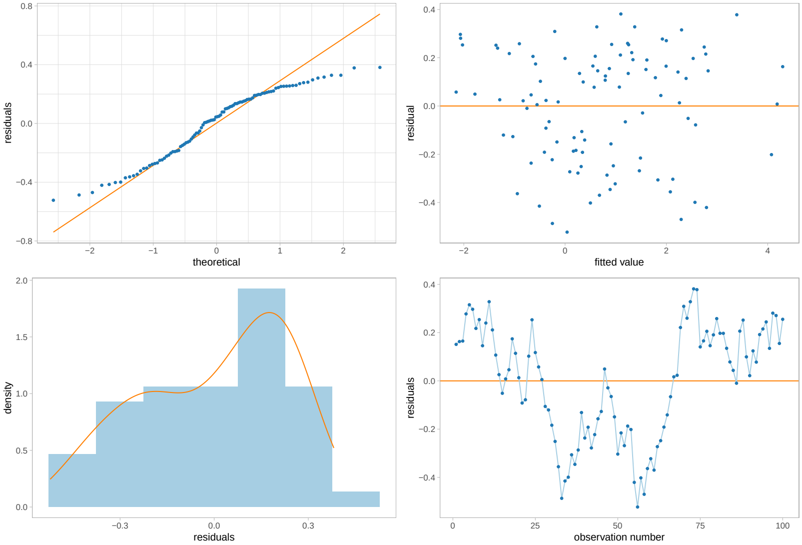
Trouble in paradise 2 - icke-normala ε (outliers)



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

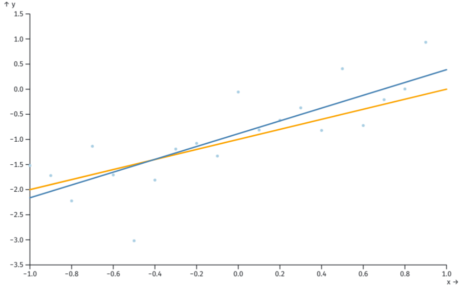
intercept, β_0 -1

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

Konfidensintervall för b_1

- Estimatorn b_1 följer en *t*-**fördelning** med $n-2$ **frihetsgrader**:

$$\frac{b_1 - \beta_1}{s_{b_1}} \sim t_{n-2}$$

- Varför $n-2$? Skattar två parametrar, β_0 och β_1. Förlorar två frihetsgrader.
- **95%-igt konfidensintervall för β_1**

$$b_1 \pm t_{0.025,n-2} \cdot s_{b_1}$$

- `lifespan` data: $n = 29$, och $t_{0.025,27} = 2.052$ från tabell.
- 95%-igt konfidensintervall för β_1

$$1.763 \pm 2.052 \cdot 0.289 = (1.170, 2.356)$$

Konfidensintervall i R

- R:

```
> model = lm(lifespan ~ spending, data = lifespan_no_usa) # utan USA
> confint((model))
```

- sda1-paketet:

```
> model = lm(lifespan ~ spending, data = lifespan_no_usa) # utan USA
> reg_summary(model, conf_intervals = TRUE, anova = FALSE)

Measures of model fit
-----------------------------------------------
Root MSE       R2   R2-adj
 1.67836  0.57952  0.56394

Parameter estimates
-----------------------------------------------
            Estimate Std. Error t value   Pr(>|t|)  2.5 %  97.5 %
(Intercept)  74.1639    0.87822 84.4482 2.9262e-34 72.362 75.9658
spending      1.7629    0.28900  6.1002 1.6256e-06  1.170  2.3559
```

Hypotesttest för β

- **Hypotestest för lutningen** i regressionen

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

- **Teststatistiska**

$$T = \frac{b_1 - 0}{s_{b_1}}$$

- Under H_0 har vi att $T \sim t_{n-2}$.
- Vi förkastar nollhypotesten på signifikansnivån $\alpha = 0.05$ om

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

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

$$t_{\text{crit}} = t_{0.025,n-2}$$

- **P-värde** räknas som tidigare, men från t_{n-2} fördelning.

Hypotesttest för β - lifespan data

- $n = 29$, så $n - 2 = 27$, och $t_{\text{crit}} = t_{0.025}(27) = 2.052$.

$$t_{\text{obs}} = \frac{1.763 - 0}{0.289} = 6.100$$

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

- Vi förkastar nollhypotesen att `spending` inte är korrelerad med `lifespan`.

- `spending` är en **signifikant förklarande variabel** för livslängd på signifikansnivå 5%.

- Testets p-värde visar att vi tokförkastar H_0!

$$p = 1.6256e - 06 = 0.0000016256$$

- $1.6256\text{e-}06$. Flytta punkten/kommat sex steg till vänster.

Hypotestest 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
```

Prediktionsintervall

- Antag att vi gör en prognos vid ett nytt $x = x_\star$

$$\hat{y}_\star = b_0 + b_1 x_\star$$

- **Prediktionsintervall** för $\hat{y}_\star$ - **två källor av osäkerhet**:
  - De **okända parametrarna** β_0 och β_1, dvs osäkerhet om regressionslinjen vid $x_\star$.
  - **Variationen i de enskilda y-värdena kring regressionlinjen**. Alla observationer "träffas av ett ε" med standardavvikelse σ_ε.

- **Prediktionsvariansen**:

$$\sigma^2_{\text{prediktion}} = \sigma^2_{\text{regressionslinjen vid } x_\star} + \sigma^2_\varepsilon$$

- **95%-igt prediktionsintervall** för enskild observation vid $x_\star$

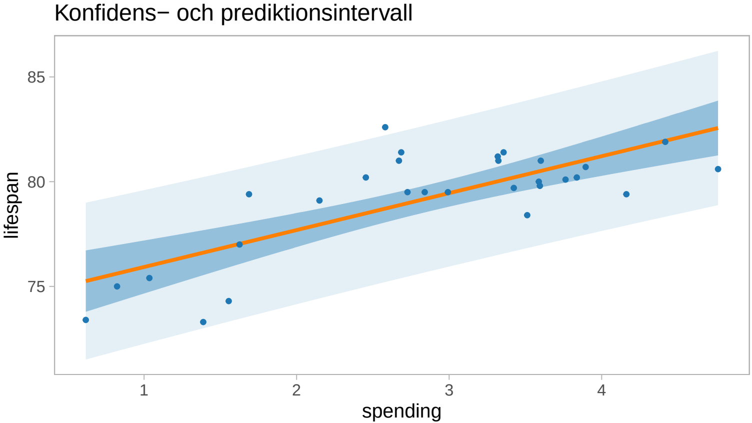
$$\hat{y}_\star \pm t_{0.025,n-2} \cdot \sqrt{\frac{s_e^2}{n} + s_{b_1}^2 (x_\star - \bar{x})^2 + s_e^2}$$

Prediktionsintervall

```
> model = lm(lifespan ~ spending, data = lifespan_no_usa)
> predict(model, newdata = data.frame(spending = 3.323))
       1
80.02209
> predict(model, newdata = data.frame(spending = 4.323))
       1
81.78502
> predict(model, newdata = data.frame(spending = 4.323), interval = "prediction")
       fit      lwr      upr
1 81.78502 78.17388 85.39616
```

Plot av prediktionsintervall sda1-paketet

```
> reg_predict(lifespan ~ spending, data = lifespan_no_usa)
```



- Ljusblå band är prediktionsintervall (för ett x i taget).

Multipel regression - modell och samplingfördelning

- **Populationsmodell** för **multipel regression** med k förklarande variabler

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon, \quad \varepsilon \sim N(0, \sigma_\varepsilon)$$

- Varje β_j skattas med b_j med minsta-kvadrat-metoden.
- Estimatorn b_j följer en **t-fördelning** med $n-k-1$ **frihetsgrader**:

$$\frac{b_j - \beta_j}{s_{b_j}} \sim t_{n-k-1}$$

- Varför $n-k-1$? Skattar k lutningskoefficienter $(\beta_1, \beta_2, \ldots, \beta_k)$ och ett intercept (β_0).
- Formlerna för minsta-kvadratskattningar b_j och standardfelen s_{b_j} är komplicerade. Datorn får göra jobbet. 😅

Multipel regression - konfidensintervall och test

- **Populationsmodell multipel regression**

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon, \quad \varepsilon \sim N(0, \sigma_\varepsilon)$$

- **95%-igt konfidensintervall för β_1**

$$b_j \pm t_{0.025,n-k-1} \cdot s_{b_j}$$

- **Hypotestest för lutningen** i regressionen

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

- **Teststatistiska**

$$T = \frac{b_j - 0}{s_{b_j}}$$

- **Under H_0** har vi att $T \sim t_{n-k-1}$.
- Om vi **förkastar** H_0 så drar vi slutsatsen att $\beta_j \neq 0$ och säger att x_j **är en signifikant förklarande variabel**.

Multipel regression i R

```
> model = lm(lifespan ~ spending + gdp + doctorvisits, data = lifespan_no_usa)
> summary(model)

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

Residuals:
    Min      1Q  Median      3Q     Max 
-3.4860 -0.8975 -0.0762  1.1654  3.7609 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 74.07091    1.34241  55.178  < 2e-16 ***
spending     2.10379    0.55123   3.817 0.000792 ***
gdp         -0.02993    0.04230  -0.708 0.485723    
doctorvisits  0.02842    0.10867   0.262 0.795813    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.726 on 25 degrees of freedom
Multiple R-squared:  0.5884,	Adjusted R-squared:  0.5391 
F-statistic: 11.92 on 3 and 25 DF,  p-value: 4.894e-05
```

Simulera data med sda1 paketet

```
> library(sda1)
> simdata <- reg_simulate(n = 500, betavect = c(1, -2, 1, 0), sigma_eps = 2)
> head(simdata)
           y         X1         X2         X3
1 -0.8556858 -0.1638814 -1.2216823 -1.2885348
2  7.7482158 -0.4227647  0.8976398 -0.7506514
3  1.3913296  0.7428795 -0.3323370 -1.0195601
4 -1.2241392  1.3069117  0.2499938 -0.6624398
5  2.8649503 -0.9133748 -0.5069302  0.8182298
6 -0.0306757 -1.1655395 -2.3763845 -2.3674088
```

Skatta från simulerat data med sda1 paketet

```
> lmfit <- lm(y ~ X1 + X2 + X3, data = simdata)
> summary(lmfit)

Call:
lm(formula = y ~ X1 + X2 + X3, data = simdata)

Residuals:
    Min      1Q  Median      3Q     Max
-5.5087 -1.3133 -0.0259  1.3712  5.6610

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.98541    0.08979  10.974   <2e-16 ***
X1          -1.91591    0.08952 -21.403   <2e-16 ***
X2           0.94003    0.08944  10.510   <2e-16 ***
X3           0.06682    0.08574   0.779    0.436
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.99 on 496 degrees of freedom
Multiple R-squared:  0.549,	Adjusted R-squared:  0.5462
F-statistic: 201.2 on 3 and 496 DF,  p-value: < 2.2e-16
```