02402 Statistik (Polyteknisk grundlag)

Uge 8: Simpel lineær regression

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Dagsorden

1. Introduktion til lineær regression
2. Mindste kvadraters metode (Least squares)
   - Standardfejl på parameter estimaterne
   - Matrice notation
3. Hypotesetest og konfidensintervaller for β_0 og β_1
4. Konfidensinterval og prædiktionsinterval for linjen
5. Korrelation og forklaret varians

02402 Statistik (Polyteknisk grundlag)

Introduktion til lineær regression

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Statistisk analyse med **forklarende variable**

For to grupper er den forklarende variabel "bare" hvilken gruppe du tilhører.

Nu skal vi snakke om **kontinuærte** forklarende variable.

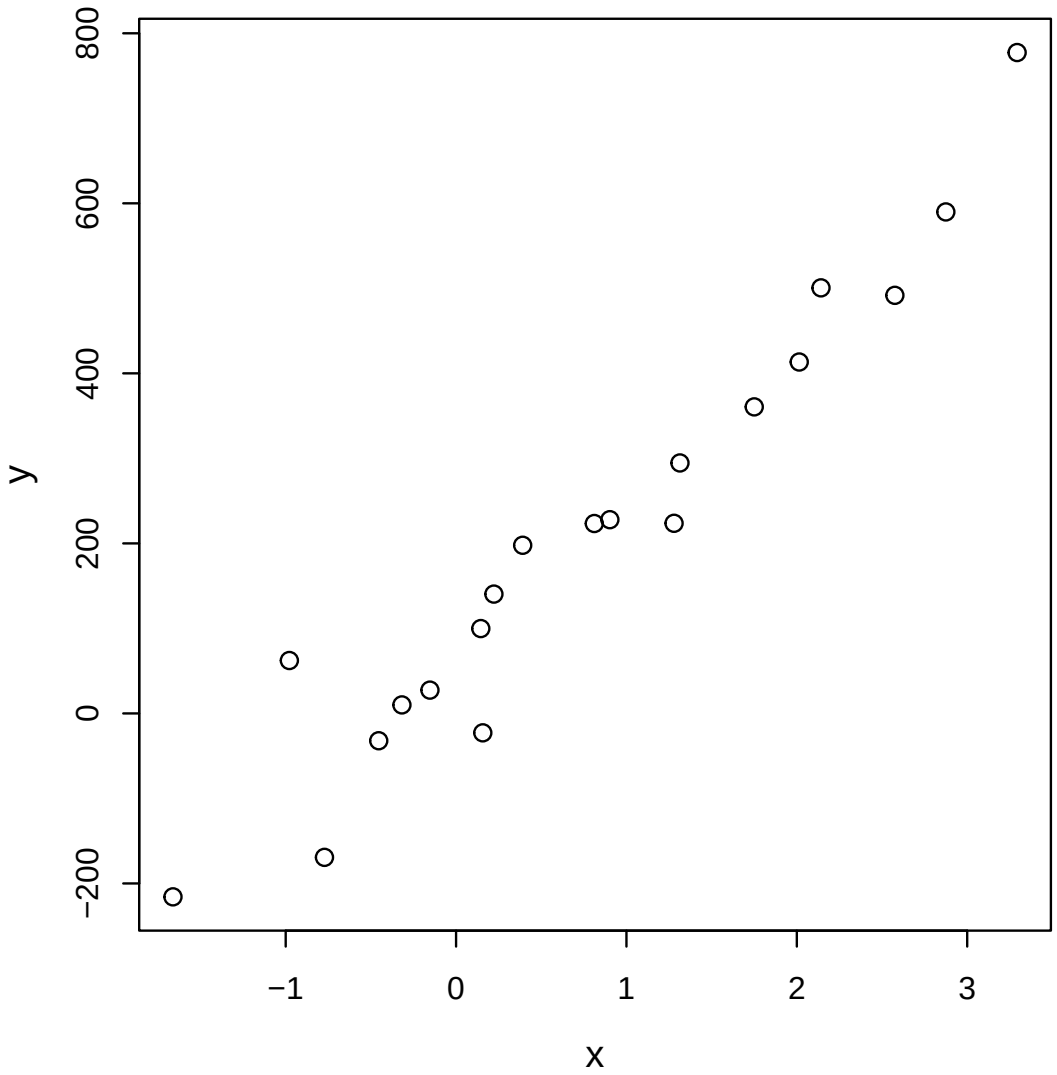
Motiverende eksempel

Hvad skal vi kunne efter i dag?

- Gå til Python notebook "simple_linear_regression.ipynb" i VS Code
  Example: Height and Weight

Start med et scatterplot

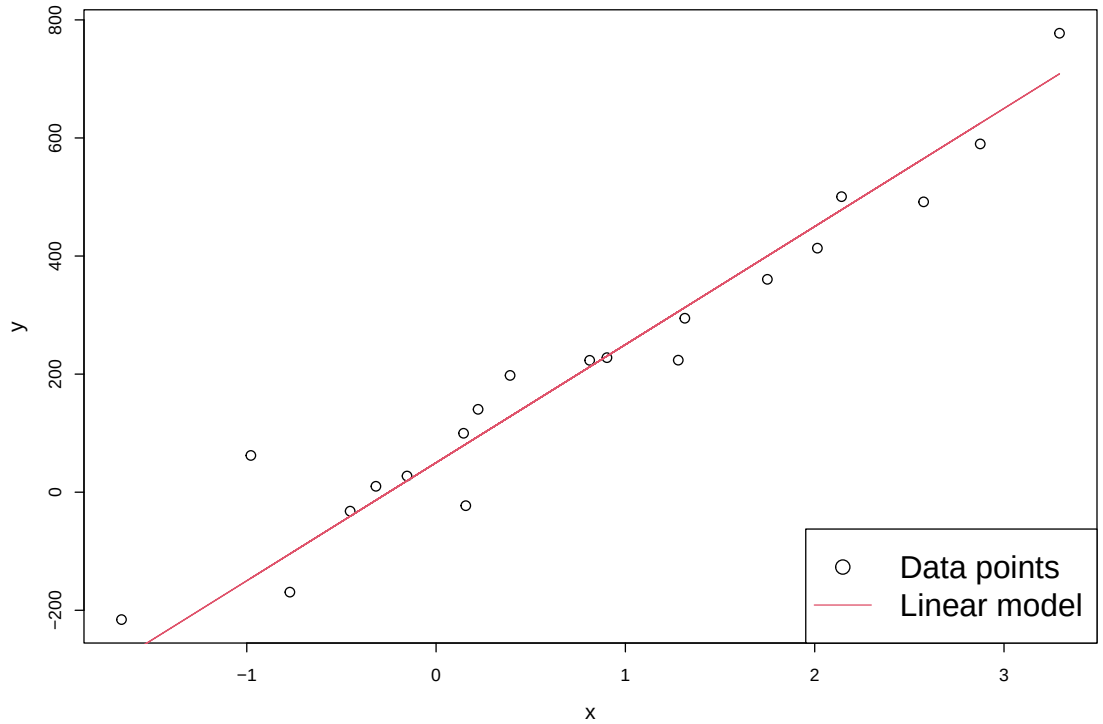
- Vi har n par datapunkter (x_i, y_i).



En lineær model

Hvis datapunkterne ligger på en ret linje, kan sammenhængen mellem x- og y-værdierne beskrives ved ligningen:

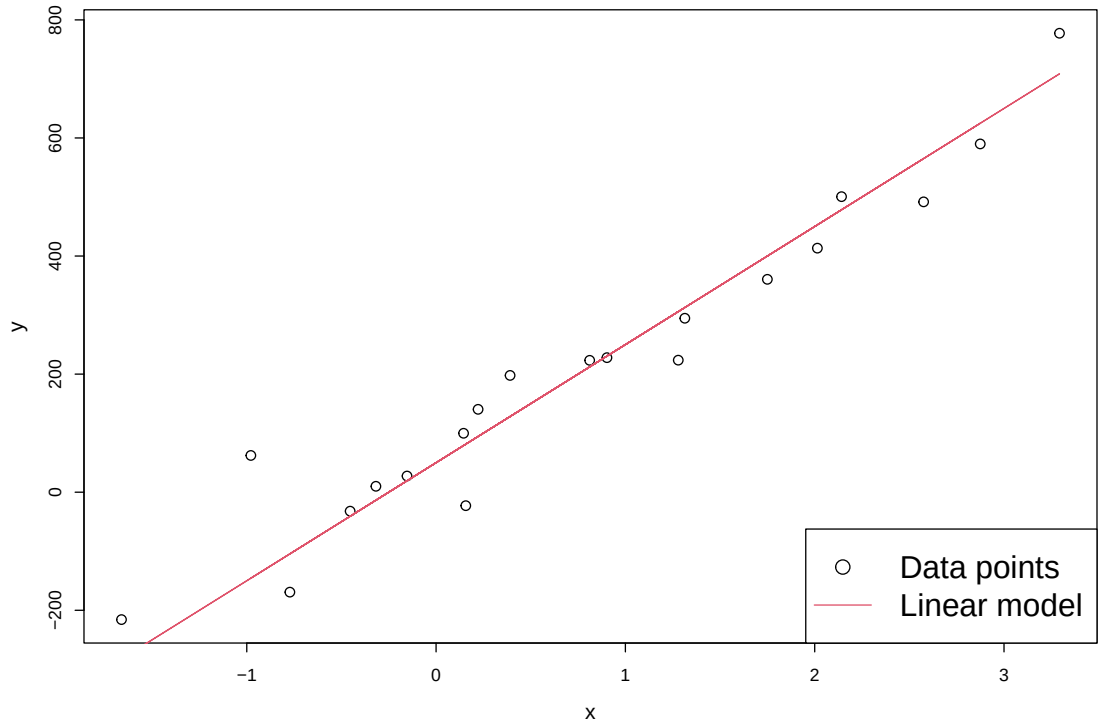
$$y_i = \beta_0 + \beta_1 x_i.$$



En lineær model

Hvis datapunkterne ligger på en ret linje, kan sammenhængen mellem x- og y-værdierne beskrives ved ligningen:

$$y_i = \beta_0 + \beta_1 x_i.$$



- Vi mangler en beskrivelse af den *tilfældige variation*.

Den lineære regressionsmodel

- Den *lineære regressionsmodel*:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (i = 1, \ldots, n).$$

- Y_i er den *afhængige variabel*.
- x_i er den *forklarende variabel*.
- ε_i er afvigelsen (residualen).
- Vi antager $\varepsilon_i \sim N(0, \sigma^2)$ (og i.i.d.).

Den lineære regressionsmodel

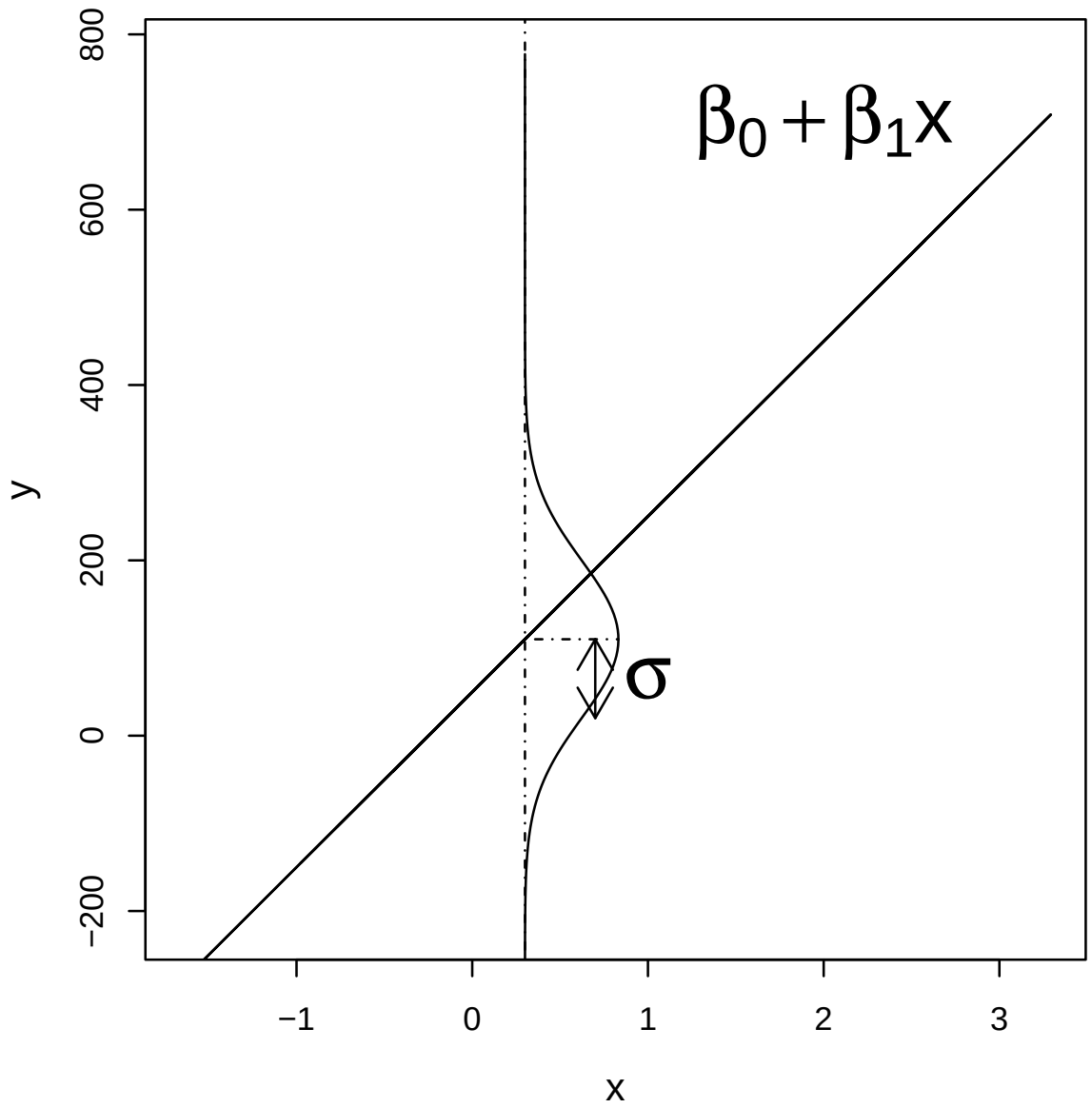
- Den *lineære regressionsmodel*:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (i = 1, \ldots, n).$$

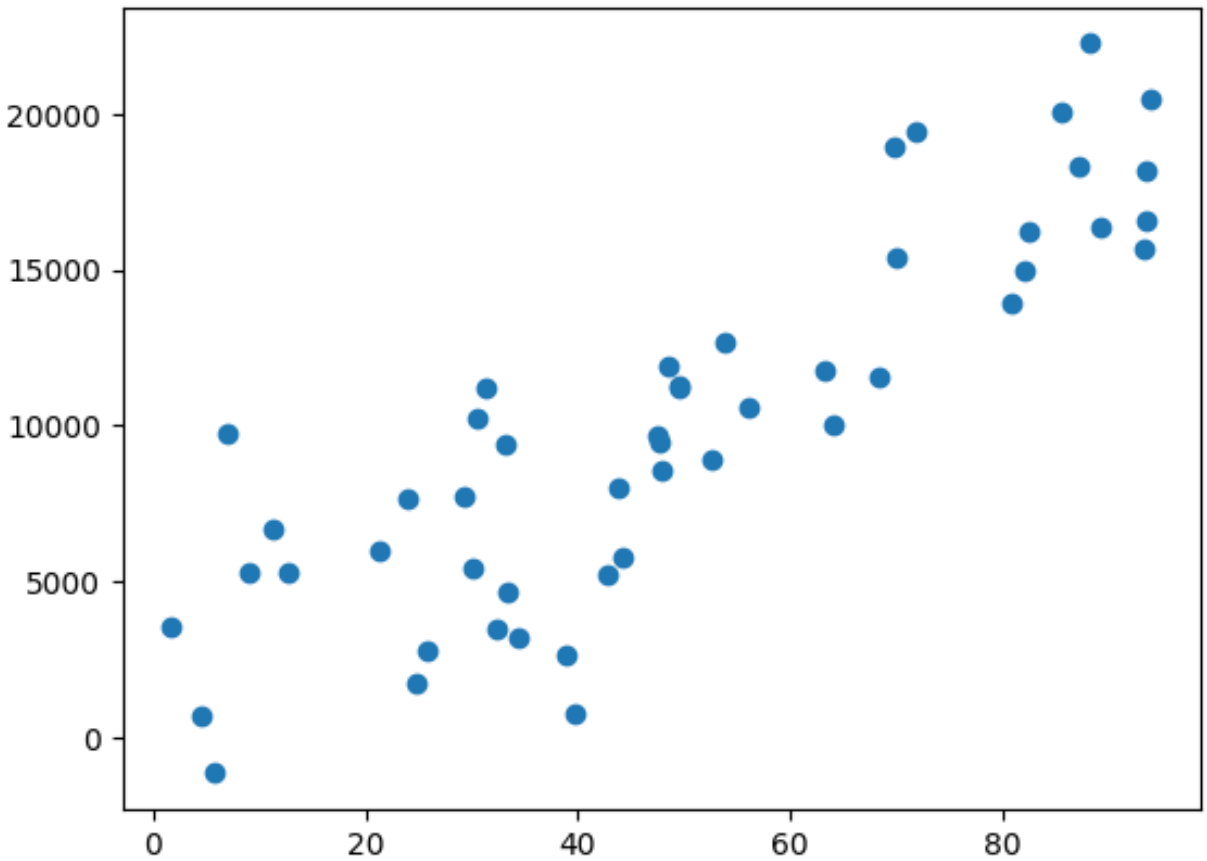
- Y_i er den *afhængige variabel*.
- x_i er den *forklarende variabel*.
- ε_i er afvigelsen (residualen).
- Vi antager $\varepsilon_i \sim N(0, \sigma^2)$ (og i.i.d.).

Overvej: *Hvilken slags fordeling følger Y_i? Er Y_i'erne ensfordelte?*

Illustration af den statistiske model



Kahoot



Kahoot!

(x7)

Mindste kvadraters metode (Least squares)

- Standardfejl på parameter estimaterne
- Matrice notation

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Mindste kvadraters metode

- Vi ønsker at **estimere** parametrene β_0 og β_1.

Mindste kvadraters metode

- Vi ønsker at **estimere** parametrene β_0 og β_1.
- God ide: Lad os minimere variansen af residualerne (σ^2).

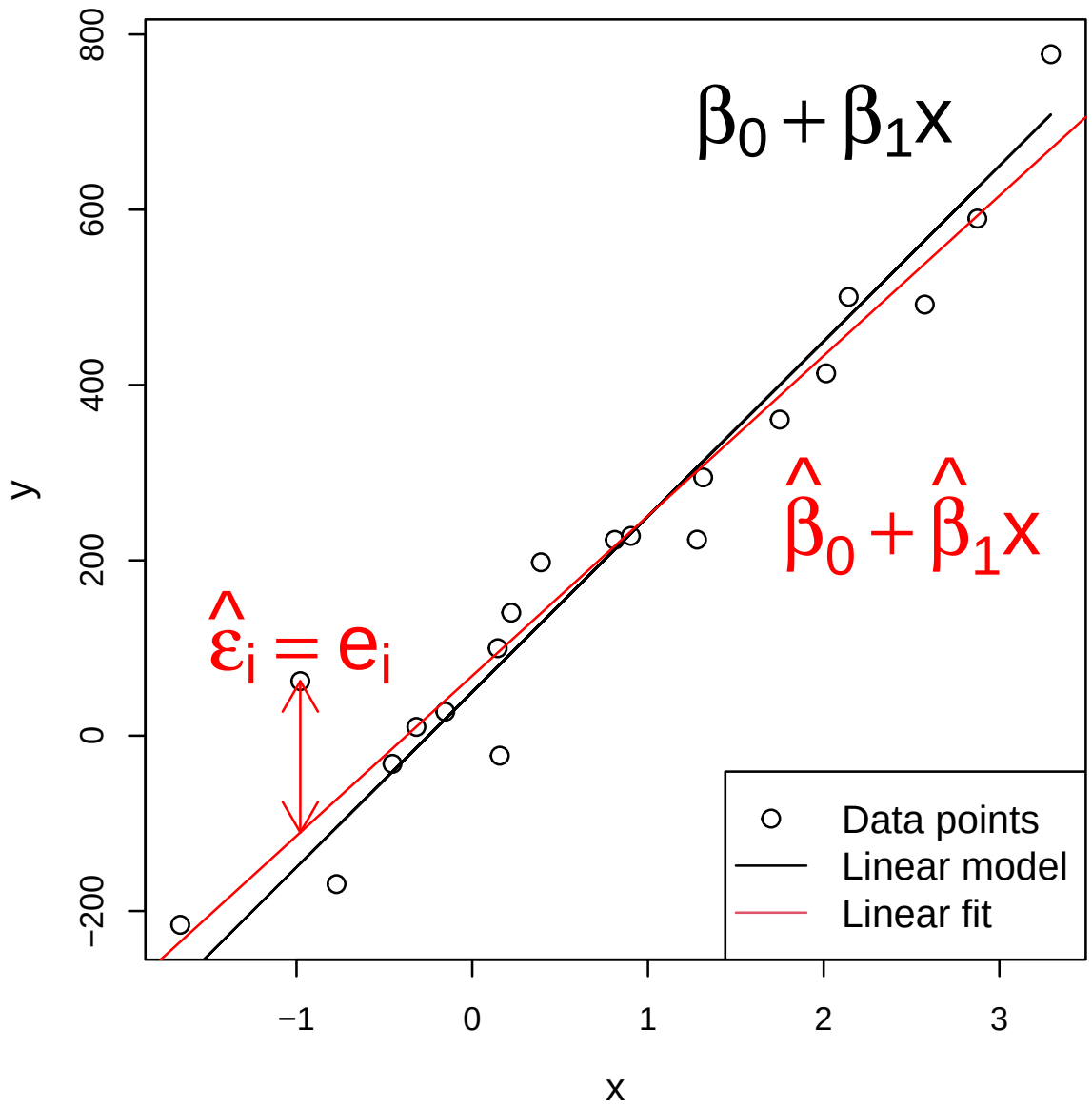
Mindste kvadraters metode

- Vi ønsker at **estimere** parametrene β_0 og β_1.
- God ide: Lad os minimere variansen af residualerne (σ^2).
- Vi minimerer summen af de kvadrerede residualer ("Residual Sum of Squares", RSS):

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 .$$

Dvs. at vi vælger $\hat{\beta}_0$ og $\hat{\beta}_1$, sådan at RSS bliver så lille som muligt.

Illustration af model, data og fit



'Least squares'-estimatorer

Theorem 5.4 Least squares estimators

The least squares estimators of β_0 and β_1 are given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(x_i - \bar{x})}{S_{xx}}, \tag{5-9}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}, \tag{5-10}$$

where $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$.

(Se udledning i "Proof" side 218-219)

Eksempel i Python: Estimér β_0 og β_1

Nu prøver vi at simulere noget data fra en kendt underliggende model.

Ud fra data skal vi estimere parametrene.

- Gå til Python notebook
  "simple_linear_regression.ipynb" i VS Code
  Example: Estimating Parameters

02402 Statistik (Polyteknisk grundlag)

- Standardfejl på parameter estimaterne
- Matrice notation

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Simulering

Vi udtager en ny stikprøve

Vil estimaterne af $\hat{\beta}_0$ and $\hat{\beta}_1$ så blive de samme?

- Gå til Python notebook
  "simple_linear_regression.ipynb" i VS Code
  Example: Variation of Parameters

Tænkepause

Hvad lærte vi fra simuleringen?

Hvad er fordelingerne af $\hat{\beta}_0$ og $\hat{\beta}_1$?

Hvad (tror du) har indflydelse på variansen af $\hat{\beta}_0$ og $\hat{\beta}_1$?

(Skriv dine egne noter her)

Variansen (og co-varians) af $\hat{\beta}_0$ og $\hat{\beta}_1$

Theorem 5.8 Variance of estimators

The variance and covariance of the estimators in Theorem 5.4 are given by

$$V[\hat{\beta}_0] = \frac{\sigma^2}{n} + \frac{\bar{x}^2\sigma^2}{S_{xx}}, \tag{5-27}$$

$$V[\hat{\beta}_1] = \frac{\sigma^2}{S_{xx}}, \tag{5-28}$$

$$\mathrm{Cov}[\hat{\beta}_0, \hat{\beta}_1] = -\frac{\bar{x}\sigma^2}{S_{xx}}, \tag{5-29}$$

where σ^2 is usually replaced by its estimate ($\hat{\sigma}^2$). The central estimator for σ^2 is

$$\hat{\sigma}^2 = \frac{RSS(\hat{\beta}_0, \hat{\beta}_1)}{n-2}. \tag{5-30}$$

When the estimate of σ^2 is used the variances also become estimates and we'll refer to them as $\hat{\sigma}^2_{\beta_0}$ and $\hat{\sigma}^2_{\beta_1}$.

"Standard Errors" for $\hat{\beta}_0$ og $\hat{\beta}_1$

$$\hat{\sigma}_{\beta_0} = \sqrt{\frac{\hat{\sigma}^2}{n} + \frac{\bar{x}^2\hat{\sigma}^2}{S_{xx}}} = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}, \tag{5-43}$$

$$\hat{\sigma}_{\beta_1} = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} = \hat{\sigma}\sqrt{\frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}}, \tag{5-44}$$

$\hat{\sigma}_{\beta_0}$ og $\hat{\sigma}_{\beta_1}$ er "**Standard Errors**" for $\hat{\beta}_0$ og $\hat{\beta}_1$!

Desuden har vi:

$$\hat{\sigma} = \sqrt{\frac{RSS(\hat{\beta}_0,\hat{\beta}_1)}{n-2}} = \sqrt{\frac{\sum_i^n \hat{\varepsilon}_i^2}{n-2}}$$

(kend forskel på de forskellige σ'er!)

Kahoot!

(x2)

Python

Python funktion til at estimere lineær regressions model:

- `my_fit = smf.ols(formula = 'y  x', data=...).fit()`

Print modellen i en tabel:

- `print(my_fit.summary(slim=True))`

- Gå til Python notebook
  "simple_linear_regression.ipynb" i VS Code
  Example: Estimate standard errors of parameters

02402 Statistik (Polyteknisk grundlag)

- Standardfejl på parameter estimaterne
- Matrice notation

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Den Generelle Linære Model for simpel lineær regression

Model for "simpel lineær regression":

$$Y_i = \beta_0 + \beta_1 \cdot x_i + \varepsilon_i$$

Kan også skrives på matriceform:

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon, \quad \varepsilon_i \sim N(0, \sigma^2)$$

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \qquad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Her er $\{x_1, x_2, ..., x_n\}$ én kontinuær forklarende variabel

Theorem 5.23

The estimators of the parameters in the simple linear regression model are given by

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}, \tag{5-67}$$

and the covariance matrix of the estimates is

$$\mathrm{V}[\hat{\boldsymbol{\beta}}] = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}, \tag{5-68}$$

and central estimate for the error variance is

$$\hat{\sigma}^2 = \frac{RSS}{n-2}. \tag{5-69}$$

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} \qquad \mathbf{V}[\hat{\beta}] = \begin{bmatrix} \mathbf{V}[\hat{\beta}_0] & \mathbf{Cov}[\hat{\beta}_0, \hat{\beta}_1] \\ \mathbf{Cov}[\hat{\beta}_0, \hat{\beta}_1] & \mathbf{V}[\hat{\beta}_1] \end{bmatrix}$$

Hypotesetest og konfidensintervaller for β_0 og β_1

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Hypotesetest for β_0 og β_1

Vi kan udføre hypotesetest for parametrene i en lineær regressionsmodel:

$$H_{0,i}: \quad \beta_i = \beta_{0,i},$$
$$H_{1,i}: \quad \beta_i \neq \beta_{0,i}.$$

Theorem 5.12 Test statistics

Under the null hypothesis ($\beta_0 = \beta_{0,0}$ and $\beta_1 = \beta_{0,1}$) the statistics

$$T_{\beta_0} = \frac{\hat{\beta}_0 - \beta_{0,0}}{\hat{\sigma}_{\beta_0}}, \tag{5-45}$$

$$T_{\beta_1} = \frac{\hat{\beta}_1 - \beta_{0,1}}{\hat{\sigma}_{\beta_1}}, \tag{5-46}$$

are t-distributed with $n-2$ degrees of freedom, and inference should be based on this distribution.

Hypotetest for β_0 og β_1

- Test om parametrene er signifikant forskellige fra 0:

$$H_{0,i} : \beta_i = 0\,, \quad H_{1,i} : \beta_i \neq 0.$$

- (Se eksempel 5.13 for et eksempel på en hypotesetest)

- Gå til Python notebook
  "simple_linear_regression.ipynb" i VS Code
  Example: Hypothesis test for parameters

Konfidensintervaller for β_0 og β_1

Method 5.15 Parameter confidence intervals

$(1-\alpha)$ confidence intervals for β_0 and β_1 are given by

$$\hat{\beta}_0 \pm t_{1-\alpha/2} \cdot \hat{\sigma}_{\beta_0}, \tag{5-52}$$

$$\hat{\beta}_1 \pm t_{1-\alpha/2} \cdot \hat{\sigma}_{\beta_1}, \tag{5-53}$$

where $t_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of a t-distribution with $n-2$ degrees of freedom. Where $\hat{\sigma}_{\beta_0}$ and $\hat{\sigma}_{\beta_1}$ are calculated from the results in Theorem 5.8, and Equations (5-43) and (5-44).

Konfidensintervaller for β_0 og β_1

- I output tabellen (regressions-tabellen) kan $\hat{\sigma}_{\beta_0}$ og $\hat{\sigma}_{\beta_1}$ aflæses under "std err".

- Gå til Python notebook "simple_linear_regression.ipynb" i VS Code
  - "Example: Confidence interval for parameters"
  - + KAHOOT (x1)

Konfidensinterval og prædiktionsinterval for linjen

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Konfidensinterval for regressions**linjen**

En model for Y_i-værdierne (interessevariablen) kan skrives som

$$Y \sim N(\mu(x), \sigma^2),$$

hvor $\mu(x) = \beta_0 + \beta_1 x$ (gennemsnit som funktion af x)

For en bestemt x-værdi ($x = x_0$) kan vi finde et $(1-\alpha)$-**konfidensinterval for gennemsnittet** (for $\hat{\mu}(x_0)$):

$$(\hat{\beta}_0 + \hat{\beta}_1 x_0) \pm t_{1-\alpha/2} \cdot \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

(i bogen kaldes $\hat{\mu}(x_0) = \hat{Y}(x_{new})$)

Python

Python kode til at beregne $\hat{\mu}(x_0)$ for nye x-værdier:
Definér ønskede x-værdier:

- `x_new = pd.DataFrame('x': ... )`

Beregn $\hat{\mu}(x_0)$ for hver x-værdi:

- `my_fit.get_prediction(x_new).summary_frame(alpha=...)`

(Konfidensinterval fås med `mean_ci_lower` og `mean_ci_upper`)

- Gå til Python notebook
  "simple_linear_regression.ipynb" i VS Code
  Example: Confidence interval for the line

Prædiktionsinterval for en ny observation

En model for Y_i-værdierne (interessevariablen) kan skrives som

$$Y \sim N(\mu(x), \sigma^2),$$

dvs. observationer ved samme x-værdi (x_0) vil være normalfordelte omkring $\mu(x_0)$ og med spredning σ.

For en bestemt x-værdi $(x = x_0)$ kan vi finde et $(1-\alpha)$-**Prædiktionsinterval for observationen** (for $Y(x_0)$):

$$(\hat{\beta}_0 + \hat{\beta}_1 x_0) \pm t_{1-\alpha/2} \cdot \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

(i bogen kaldes $Y(x_0) = Y_{new}$)

Python

Python kode til at beregne prædiktionsinterval for observationer (ved nye x-værdier):
Definér ønskede x-værdier:

- `x_new = pd.DataFrame('x': ... )`
- `my_fit.get_prediction(x_new).summary_frame(alpha=...)`

(denne gang bruges `obs_ci_lower` og `obs_ci_upper`)

- Gå til Python notebook
  "simple_linear_regression.ipynb" i VS Code
  Example: Prediction interval for observations

Konfidens- og prædiktionsintervaller i bogen

Method 5.18 Intervals for the line

The (1-α) **confidence interval** for the line $\hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}}$ is

$$\hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}} \pm t_{1-\alpha/2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_{\text{new}} - \bar{x})^2}{S_{xx}}}, \tag{5-59}$$

and the (1-α) **prediction interval** is

$$\hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}} \pm t_{1-\alpha/2}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_{\text{new}} - \bar{x})^2}{S_{xx}}}, \tag{5-60}$$

where $t_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of the t-distribution with $n-2$ degrees of freedom.

Konfidens- og prædiktionsintervaller

- Konfidensintervallet angiver usikkerheden for *regressionslinjen*.
- Prædiktionsintervallet angiver usikkerheden for en *ny observation*.
- For fastholdt α er prædiktionsintervallet større end konfidensintervallet.
- Prædiktionsintervallet kan aldrig blive mindre end den tilfældige variation i data (altså den fra fejlleddet ε).

Kahoot!

(x2)

Korrelation og forklaret varians

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Forklaret varians og korrelation

- Den forklarede varians i en model er R^2 (R-squared).
- Beregnes med

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2},$$

  hvor $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.
- Andelen af den totale varians, der er forklaret med modellen.

Kahoot!

(x2)

Forklaret varians og korrelation

- Korrelationen ρ er et mål for *lineær sammenhæng* mellem to stokastiske variable.
- Den estimerede (dvs. empiriske) korrelation opfylder

$$\hat{\rho} = R = \sqrt{R^2}\, sgn(\hat{\beta}_1),$$

hvor $sgn(\hat{\beta}_1)$ er -1 for $\hat{\beta}_1 \leq 0$ og 1 for $\hat{\beta}_1 > 0$

Forklaret varians og korrelation

- Korrelationen ρ er et mål for *lineær sammenhæng* mellem to stokastiske variable.
- Den estimerede (dvs. empiriske) korrelation opfylder

$$\hat{\rho} = R = \sqrt{R^2}\, sgn(\hat{\beta}_1),$$

  hvor $sgn(\hat{\beta}_1)$ er -1 for $\hat{\beta}_1 \leq 0$ og 1 for $\hat{\beta}_1 > 0$
- Altså:
  - Positiv korrelation ved positiv hældning.
  - Negativ korrelation ved negativ hældning.

Test for signifikant korrelation

- Test for signifikant korrelation (lineær sammenhæng) mellem to variable:

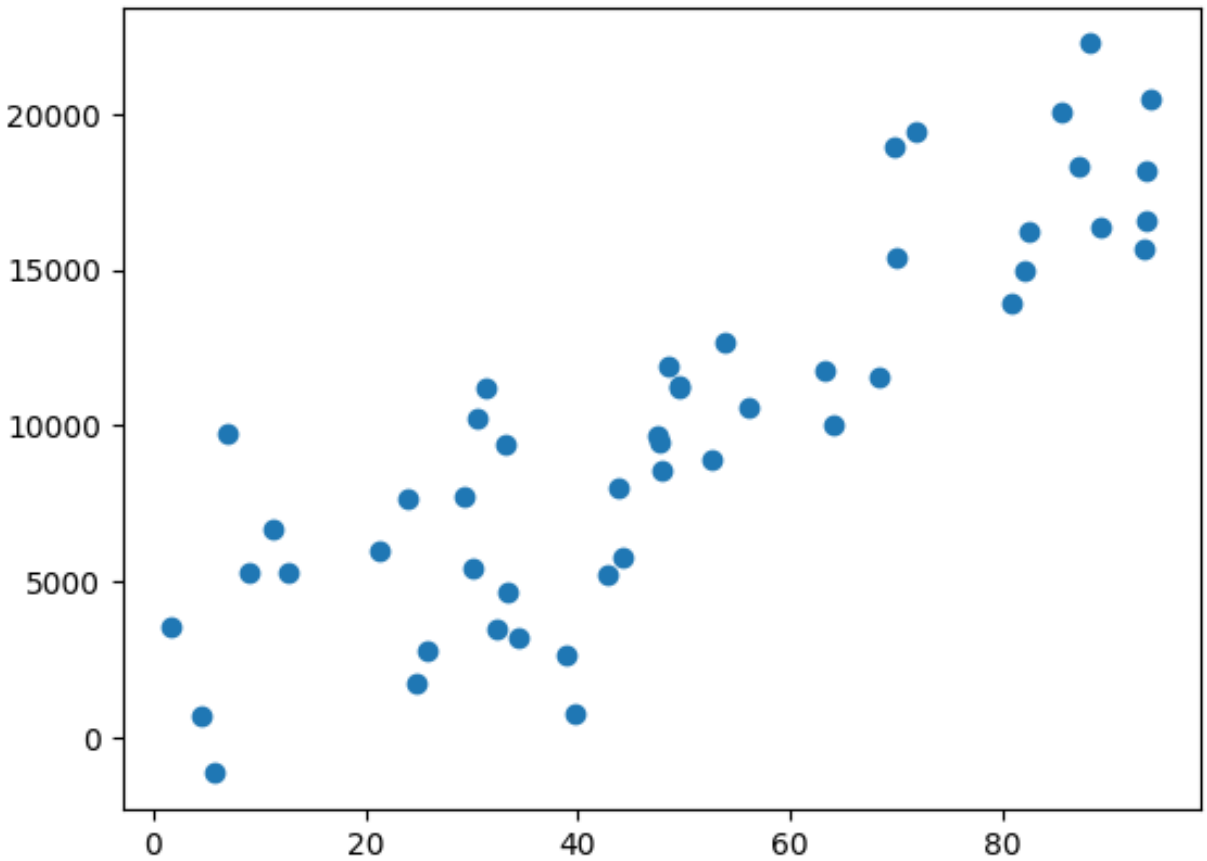
$$H_0 : \rho = 0,$$
$$H_1 : \rho \neq 0,$$

er ækvivalent med

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

hvor β_1 er hældningen i den simple lineære regressionsmodel.

Kahoot



Kahoot!

(x2)

Næste gang

Næste gang: Multipel lineær regression

Dagsorden

1. Introduktion til lineær regression
2. Mindste kvadraters metode (Least squares)
   - Standardfejl på parameter estimaterne
   - Matrice notation
3. Hypotesetest og konfidensintervaller for β_0 og β_1
4. Konfidensinterval og prædiktionsinterval for linjen
5. Korrelation og forklaret varians

Tjekliste

Efter i dag skal du kunne:

- Vurdere ud fra et scatterplot om det vil være relevant at lave en simpel lineær regressionsmodel for data
- Opstille en simpel lineær regressionsmodel og estimere dens parametre
- Beskrive princippet bag "Least Squares Estimation".
- Opskrive Design Matrice for en simpel lineær regresionsmodel
- Estimere usikkerheder (herunder konfidensintervaller) på parametrene i en simpel lineær regressions model
- Udføre hypotesetest for om en parameter er signifikant forskellig fra nul
- Beregne prædikterede værdier for en simpel lineær regressionsmodel, herunder også konfidensinterval samt prædiktionsinterval
- Beskrive sammenhængen mellem R^2 og correlations coefficienten (og hældningen af regressionslinjen)
- Udføre modelkontrol af en simpel lineær regressionsmodel (dette punkt dækkes hovedsageligt i projekterne).

Øvelser Uge 8

5.1 Kan besvares uden Python

5.2 Brug Python, men lav beregninger ud fra formler og brug først `smf.ols` til sidst (d).

5.6 Brug gerne Python

5.7 Brug gerne Python

Project Start på projekt 2