02402 Statistik (Polyteknisk grundlag)

Uge 9: Multipel lineær regression

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Dagsorden

- Praktiske informationer
- Opsummering om simpel lineær regression
- Motiverende eksempel

1. Multipel lineær regression
    - Mindste kvadraters metode (Least squares)
    - Hypotesetest og konfidensintervaller for β_i'erne
    - Konfidens- og prædiktionsinterval for "linjen"
2. Modelkontrol
3. Modelopbygning
    - Kurvelinearitet
    - Kollinearitet
    - Modeludvælgelse (Model selection)
    - Et avanceret eksempel

02402 Statistik (Polyteknisk grundlag)

- Praktiske informationer
- Opsummering om simpel lineær regression
- Motiverende eksempel

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Øvelser

I dag forsøger vi med "klasseundervisning" i to af lokalerne:

I bygning 324:

Lokale 40: Thea (taler dansk)

Lokale 60: Drin (taler engelsk)

Fra kl 10.15 og frem gennemgås øvelse 6.1 og 6.2 på tavlen.
Er der tid til overs arbejder man videre med Projekt2.

(Kyril, som normalt er i lokale 60, vil i stedet gå rundt i foyer-områderne)

Eksamen

Hvordan forbereder man sig til eksamen?

Til formelsamlingen

I kap 5 og 6 er der en del formler og tilhørende Python kode.

Tilføj selv noter til formelsamlingen!

Fx:

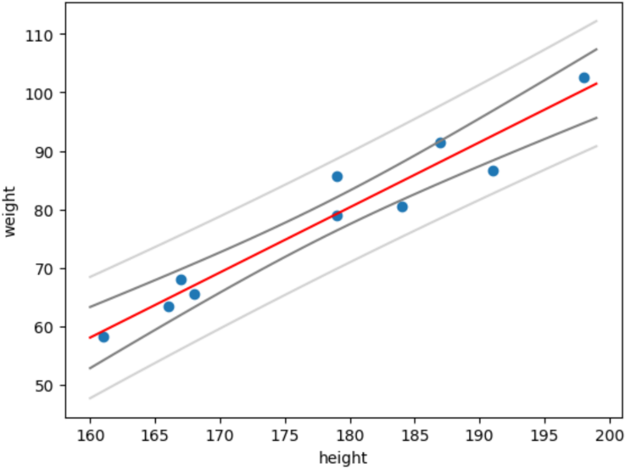
- $RSS = \dots$
- `print(my_fit.scale)`
- `print(my_fit.pvalues)`
- ...

02402 Statistik (Polyteknisk grundlag)

- Praktiske informationer
- Opsummering om simpel lineær regression
- Motiverende eksempel

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Opsummering: Simpel lineær regression



02402 Statistik (Polyteknisk grundlag)

- Praktiske informationer
- Opsummering om simpel lineær regression
- Motiverende eksempel

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Eksempel: Ozonkoncentration

- Gå til Python notebook:
  "multiple_linear_regression.ipynb" i VS Code
  Example: Ozon concentration
  (+ KAHOOT x4)

Visual Studio Code

02402 Statistik (Polyteknisk grundlag)

Multipel lineær regression

- Mindste kvadraters metode (Least squares)
- Hypotesetest og konfidensintervaller for β_i'erne
- Konfidens- og prædiktionsinterval for "linjen"

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Den lineære regressionsmodel

- Den *lineære regressionsmodel* (general linear model, GLM):

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \ldots + \beta_p x_{p,i} + \varepsilon_i \quad (i = 1, \ldots, n).$$

- Y_i er den *afhængige variabel*.
- $x_{1i}, x_{2i}, \ldots, x_{pi}$ er *forklarende variable*.
- Der er i alt p forklarende variable i modellen.
- ε_i er afvigelsen (residualen).
- Vi antager $\varepsilon_i \sim N(0, \sigma^2)$ (og i.i.d.).

Matrice notation

Den *lineære regressionsmodel* (general linear model, GLM):

$$Y_i = \beta_0 + \beta_1 \cdot x_{1i} + \beta_2 \cdot x_{2i} + \ldots + \varepsilon_i$$

På matriceform:

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \ldots \\ 1 & x_{12} & x_{22} & \ldots \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} & \ldots \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \end{bmatrix} \qquad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

02402 Statistik (Polyteknisk grundlag)

- Mindste kvadraters metode (Least squares)
- Hypotesetest og konfidensintervaller for β_i'erne
- Konfidens- og prædiktionsinterval for "linjen"

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Mindste kvadraters metode (*Least Squares*)

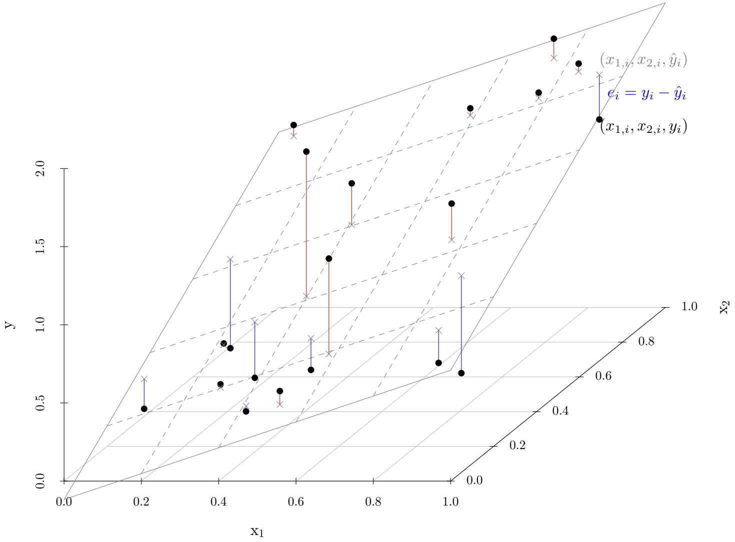
- Vi ønsker at **estimere** parametrene β_0, β_1, ... ,β_p.
- God ide: Lad os minimere variansen af residualerne (σ^2).
- Vi minimerer summen af de kvadrerede residualer ("Residual Sum of Squares", RSS):

$$RSS = \sum_{i=1}^{n} \varepsilon_i^2 = \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon}$$

- $\varepsilon_i^2 = (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1,i} - \hat{\beta}_2 x_{2,i} - \ldots - \hat{\beta}_p x_{p,i})^2 = (y_i - \hat{y}_i)^2$
- $\boldsymbol{\varepsilon}$ er en *vektor* med alle residualerne

Vi minimerer RSS mht. alle $(p+1)$ β'er.

Mindste kvadraters metode



Python

Brug Python til at estimere en lineær regressions model:

- `my_fit = smf.ols(formula = 'y ~ x1 + x2 + x3', data=...).fit()`
- Bemærk: "`y ~ x1 + x2 + x3`"

Print regressions-tabellen:

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

- Gå til Python notebook:
  "multiple_linear_regression.ipynb" i VS Code
  Example: Ozon concentration
  with multiple linear regression

Least Squares estimater

Theorem 6.17

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

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}, \tag{6-44}$$

and the covariance matrix of the estimates is

$$V[\hat{\boldsymbol{\beta}}] = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}, \tag{6-45}$$

and central estimate for the residual variance is

$$\hat{\sigma}^2 = \frac{RSS}{n-(p+1)}. \tag{6-46}$$

Bemærkning: For *multipel* lineær regression gives formlerne for $\hat{\beta}_i$ og $\hat{\sigma}_{\beta_i}$ *kun* på matriceform (Python udfører beregningerne for os).

02402 Statistik (Polyteknisk grundlag)

- Mindste kvadraters metode (Least squares)
- Hypotesetest og konfidensintervaller for β_i'erne
- Konfidens- og prædiktionsinterval for "linjen"

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Hypotesetest og konfidensinterval for β_i'erne

Theorem 6.2 Hypothesis tests and confidence intervals

Suppose the we are given parameter estimates $(\hat{\beta}_0, \ldots, \hat{\beta}_p)$ and their corresponding standard errors $(\hat{\sigma}_{\beta_0}, \ldots, \hat{\sigma}_{\beta_p})$, then under the null hypothesis

$$H_{0,i}: \quad \beta_i = \beta_{0,i}, \tag{6-15}$$

the t-statistic

$$T_i = \frac{\hat{\beta}_i - \beta_{0,i}}{\hat{\sigma}_{\beta_i}}, \tag{6-16}$$

will follow the t-distribution with $n-(p+1)$ degrees of freedom, and hypothesis testing and confidence intervals should be based on this distribution. Further, a central estimate for the residual variance is

$$\hat{\sigma}^2 = \frac{RSS(\hat{\beta}_0, \ldots, \hat{\beta}_p)}{n-(p+1)}. \tag{6-17}$$

Konfidensinterval for β_0 og β_1

Method 6.5 Parameter confidence intervals

$(1-\alpha)$ confidence interval for β_i is given by

$$\hat{\beta}_i \pm t_{1-\alpha/2}\, \hat{\sigma}_{\beta_i}, \tag{6-20}$$

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

Husk: $\hat{\beta}_i$ og $\hat{\sigma}_{\beta_i}$ er beregnet med formlerne på matriceform *eller* med Python

Kahoot!

(x2)

Tænkepause

$$\hat{\beta}_{wind} \pm t_{1-\alpha/2} \cdot \hat{\sigma}_{\beta_{wind}}$$
$$-0.0693 \pm t_{1-\alpha/2} \cdot 0.015 \quad \text{(fra eksemplet i Python)}$$

(For 95% konfidensintervallet bruges $\alpha = 0.05$)

Hvad er $t_{1-\alpha/2}$ (fra hvilken fordeling)?

____

____

________________________________ (tegn også fordeling og markér $t_{1-\alpha/2}$)

Hvordan ville vi finde $t_{1-\alpha/2} = t_{0.975}$ i Python (skriv koden)?

____

Hvad er så 90% konfidens interval for β_{wind}?

____

Hypotesetest for for β_i'erne

Method 6.4 Level α t-tests for parameters

1. Formulate the *null hypothesis*: $H_{0,i} : \beta_i = \beta_{0,i}$, and the alternative hypothesis $H_{1,i} : \beta_i \neq \beta_{0,i}$
2. Compute the test statistic $t_{\text{obs},\beta_i} = \frac{\hat{\beta}_i - \beta_{0,i}}{\hat{\sigma}_{\beta_i}}$
3. Compute the evidence against the *null hypothesis*

$$p\text{-value}_i = 2P(T > |t_{\text{obs},\beta_i}|) \tag{6-19}$$

4. If the p-value$_i < \alpha$ reject $H_{0,i}$, otherwise accept $H_{0,i}$

Kahoot!

(x2)

Fortolkning af parametre

Hvad er $\hat{\beta}_i$ udtryk for?

Fortolkning af parametre

Hvad er $\hat{\beta}_i$ udtryk for?

- Den forventede ændring i y når x_i ændres én enhed.
- Effekten af x_i givet de øvrige variable.
- Effekten af x_i korrigeret for de øvrige variables effekt.
- Effekten af x_i "når de andre variable er uændret".

Fortolkning af parametre

Hvad er $\hat{\beta}_i$ udtryk for?

- Den forventede ændring i y når x_i ændres én enhed.
- Effekten af x_i givet de øvrige variable.
- Effekten af x_i korrigeret for de øvrige variables effekt.
- Effekten af x_i "når de andre variable er uændret".

- Afhænger af hvad der ellers er i modellen!
- Generelt: IKKE en kausal effekt/interventionseffekt!

02402 Statistik (Polyteknisk grundlag)

- Mindste kvadraters metode (Least squares)
- Hypotesetest og konfidensintervaller for β_i'erne
- Konfidens- og prædiktionsinterval for "linjen"

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Konfidens- og prædiktionsintervaller

Method 6.9 Intervals for the line (by Python)

The (1-α) **confidence and prediction intervals** for the line $\hat{\beta}_0 + \hat{\beta}_1 x_{1,\text{new}} + \cdots + \hat{\beta}_p x_{p,\text{new}}$ are calculated in Python by

```
# Confidence and Prediction interval
fit.get_prediction(new_data).summary_frame(alpha=0.05)
```

Remark 6.10

Explicit formulas for confidence and prediction intervals are given in Section 6.6.

Konfidens- og prædiktionsintervaller med matrice notation

$$\boldsymbol{x}_{\text{new}} = [1, x_{1,\text{new}}, \ldots, x_{p,\text{new}}]$$

$$\begin{aligned} \mathrm{V}(\hat{Y}_{\text{new}}) &= \mathrm{V}(\boldsymbol{x}_{\text{new}}\hat{\boldsymbol{\beta}}) \\ &= \boldsymbol{x}_{\text{new}}\,\mathrm{V}(\hat{\boldsymbol{\beta}})\boldsymbol{x}_{\text{new}}^T \\ &= \sigma^2 \boldsymbol{x}_{\text{new}}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_{\text{new}}^T, \end{aligned} \tag{6-48}$$

$$\begin{aligned} \mathrm{V}(Y_{\text{new}}) &= \mathrm{V}(\boldsymbol{x}_{\text{new}}\hat{\boldsymbol{\beta}} + \varepsilon_{\text{new}}) \\ &= \boldsymbol{x}_{\text{new}}\,\mathrm{V}(\hat{\boldsymbol{\beta}})\boldsymbol{x}_{\text{new}}^T + \sigma^2 \\ &= \sigma^2(1 + \boldsymbol{x}_{\text{new}}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_{\text{new}}^T). \end{aligned} \tag{6-49}$$

Python

Beregn konfidensinterval of prædiktionsinterval med Python:

- `x_new = pd.DataFrame('x1':..., 'x2':..., 'x3':...)`
- `my_fit.get_prediction(x_new).summary_frame(alpha=...)`

- Gå til Python notebook "multiple_linear_regression.ipynb" i VS Code
  Example: Confidence and prediction intervals

02402 Statistik (Polyteknisk grundlag)

Modelkontrol

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Modelkontrol (Analyse af residualerne)

- Modelkontrol: Analysér residualerne for at tjekke om antagelserne er opfyldt.
- Samme antagelser som for den simple lineære model.

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \ldots + \beta_p x_{p,i} + \varepsilon_i \quad (i = 1, \ldots, n).$$

- Vi antager $\varepsilon_i \sim N(0, \sigma^2)$ og i.i.d.

Python

Man kan få residualer og fittede y-værdier direkte fra "fit" variablen:

- `my_fit = smf.ols(formula = 'y ~ x1 + x2 + x3', data=...).fit()`
- `residuals = my_fit.resid`
- `fittedvalues = my_fit.fittedvalues`

Herefter produceres diverse plot til visual inspektion af model antagelser.

- Gå til Python notebook "multiple_linear_regression.ipynb" i VS Code
  Example: Model control

Visual Studio Code

02402 Statistik (Polyteknisk grundlag)

Modelopbygning

- Kurvelinearitet
- Kollinearitet
- Modeludvælgelse (Model selection)
- Et avanceret eksempel

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

02402 Statistik (Polyteknisk grundlag)

- Kurvelinearitet
- Kollinearitet
- Modeludvælgelse (Model selection)
- Et avanceret eksempel

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

En kurvelineær model

Regressionsmodeller til ikke-lineær data baseret på Taylorudviklinger.

Hvis vi vil benytte en model af typen

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i,$$

så kan vi bruge en multipel lineær regressionsmodel

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \varepsilon_i,$$

hvor

$$x_{i,1} = x_i, \quad x_{i,2} = x_i^2$$

og bruge de samme metoder som for multipel lineær regression.

Python

- Gå til Python notebook "multiple_linear_regression.ipynb" i VS Code
  Example: curvilinear regression
  (+ KAHOOT x1)

Visual Studio Code

Pas på med prædiktioner udenfor datagrundlag

Remark 6.12

In general one should be careful when extrapolation models into areas where there is no data, and this is in particular true when we use curvilinear regression.

02402 Statistik (Polyteknisk grundlag)

- Kurvelinearitet
- Kollinearitet
- Modeludvælgelse (Model selection)
- Et avanceret eksempel

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Kollinearitet

- Hvis to (eller flere) forklarende variable har en perfekt lineær sammenhæng, så kan vi ikke afgøre, hvilken som er forklarende.
- Også et problem hvis sammenhængen er *tæt på lineær*.
- Relateret til konceptet "confounders".
- Med to meget korrelerede x-variable:
  - *Sammen* kan det være at ingen af dem har en "unik" effekt.
  - *Separat* kan de have en stor effekt.

Python

- Gå til Python notebook "multiple_linear_regression.ipynb" i VS Code
  Example: Colinearity and confounding
  (+ KAHOOT x1)

Fortolkning af β_i'erne

Remark 6.14 Interpretation of parameters

In general we can interpret the parameters of a multiple linear regression model as the effect of the variable given the other variables. E.g. β_j is the effect of x_j when we have accounted for other effects (x_i, $i \neq j$). This interpretation is however problematic when we have strong collinearity, because the true effects are hidden by the correlation.

02402 Statistik (Polyteknisk grundlag)

- Kurvelinearitet
- Kollinearitet
- Modeludvælgelse (Model selection)
- Et avanceret eksempel

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Eksempel: Ozonkoncentration

Hvilke forklarende variable skal man inkludere i modellen?

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \ldots + \beta_p x_{p,i} + \varepsilon_i \quad (i = 1, \ldots, n).$$

Hvis β_k **ikke** er "signifikant" - så kan x_k udelades fra modellen?

- Gå til Python notebook:
  "multiple_linear_regression.ipynb" i VS Code
  Example: Model selection

Visual Studio Code

Modeludvidelse (forward selection)

- *Ikke inkluderet i bogen*
- Start med en *simpel lineær regressionsmodel* med én signifikant forklarende variabel
- *Udvid modellen* med andre forklarende variable én ad gangen
- *Stop* når der ikke er flere signifikante udvidelser

Modelreduktion (backward selection)

- *Beskrevet i bogen under sektion 6.5*
- Start med den fulde model
- Fjern den "mindst signifikante" variabel
- Stop når alle tilbageværende parametre er signifikante

Kahoot!

(x1)

Modeludvælgelse

- Der er ikke nogen sikker metode til at finde den bedste model!
- Det kræver subjektive beslutninger at udvælge en model.
- Forskellige procedurer, enten forward eller backward selection (eller begge), afhænger af forholdene.
- Der findes statistiske metoder og test til at sammenligne modeller.
- Her i kurset er kun backward selection beskrevet.

02402 Statistik (Polyteknisk grundlag)

- Kurvelinearitet
- Kollinearitet
- Modeludvælgelse (Model selection)
- Et avanceret eksempel

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Eksempel: Test af 3 forskellige patient behandlinger

I praksis er det vigtigt at man kan læse data ind fra en fil:

`pd.read_csv(...)`

Linear regression kan bruges på mange måder. Dette eksempel går lidt ud over kursets pensum (vi gennemgår det kun hvis der er tid).

- Gå til Python notebook:
  "multiple_linear_regression.ipynb" i VS Code
  Example: 3 treatment groups

Visual Studio Code

Dagsorden

- Praktiske informationer
- Opsummering om simpel lineær regression
- Motiverende eksempel

1. Multipel lineær regression
    - Mindste kvadraters metode (Least squares)
    - Hypotesetest og konfidensintervaller for β_i'erne
    - Konfidens- og prædiktionsinterval for "linjen"
2. Modelkontrol
3. Modelopbygning
    - Kurvelinearitet
    - Kollinearitet
    - Modeludvælgelse (Model selection)
    - Et avanceret eksempel