Linear regresjon (enkel og multippel)

ISTx1003 Statistisk læring og Data Science

Stefanie Muff, Institutt for matematiske fag

November 1 og 5, 2021

Plan for i dag

- Hvem er vi?
- Statistisk læring og data science
- De tre temaene i modulen:
  - regresjon
  - klassifikasjon og
  - klyngeananlyse
- Læringsressurser og pensum
- Prosjektoppgaven og Blackboard-informasjon
- Tema: regresjon - med enkel lineær regresjon

Læringsmål (av modulen)

Etter du har gjennomført denne modulen skal du kunne:

- forstå når du kan bruke regresjon, klassifikasjon og klyngeananlyse til å løse et ingeniørproblem
- kunne gjennomføre multippel lineær regresjon på et datasett
- bruke logistisk regresjon og nærmeste nabo for å utføre en klassifikasjonsoppgave
- bruke hierarkisk og k-means klyngeanalyse på et datasett, forstå begrepet avstandsmål
- og kunne kommunisere resultatene fra regresjon/ klassifikasjon/klyngeanalyse til medstudenter og ingeniører
- bli en kritisk leser av resultater fra statistikk/maskinlæring/ statistisk læring/data science/kunstig intelligens når disse rapporteres i media, og forstå om resultatene er realistiske ut fra informasjonen som gis
- kunne besvare prosjektoppgaven på en god måte!

Hva er statistisk læring og data science?

Todo

Prosjektoppgaven

- Vi ser hvor informasjonen ligger på Blackboard og hvordan melde seg på gruppe.

- Vi ser på prosjektoppgaven på https://s.ntnu.no/isthub.

Læringsmål (i dag)

- Du kan lage en modell for å forstå sammenhengen mellom en respons og en eller flere forklaringsvariabler.

- Du kan lage en modell for å predikere en respons fra en eller flere forklaringsvariabler.

Læringsressurser

Alle ressurser er tilgjengelig her:

https://wiki.math.ntnu.no/istx1003/2021h/start

Tema Regresjon:

- **Kompendium**: Regresjon (pdf og html, by Mette Langaas)
- **Korte videoer**: (by Mette Langaas)
  - Multippel lineær regresjon: introduksjon (14:07 min)
  - Multippel lineær regresjon: analyse av et datasett (15:20 min)
- Denne forelesningen
- **Disse slides** med notater

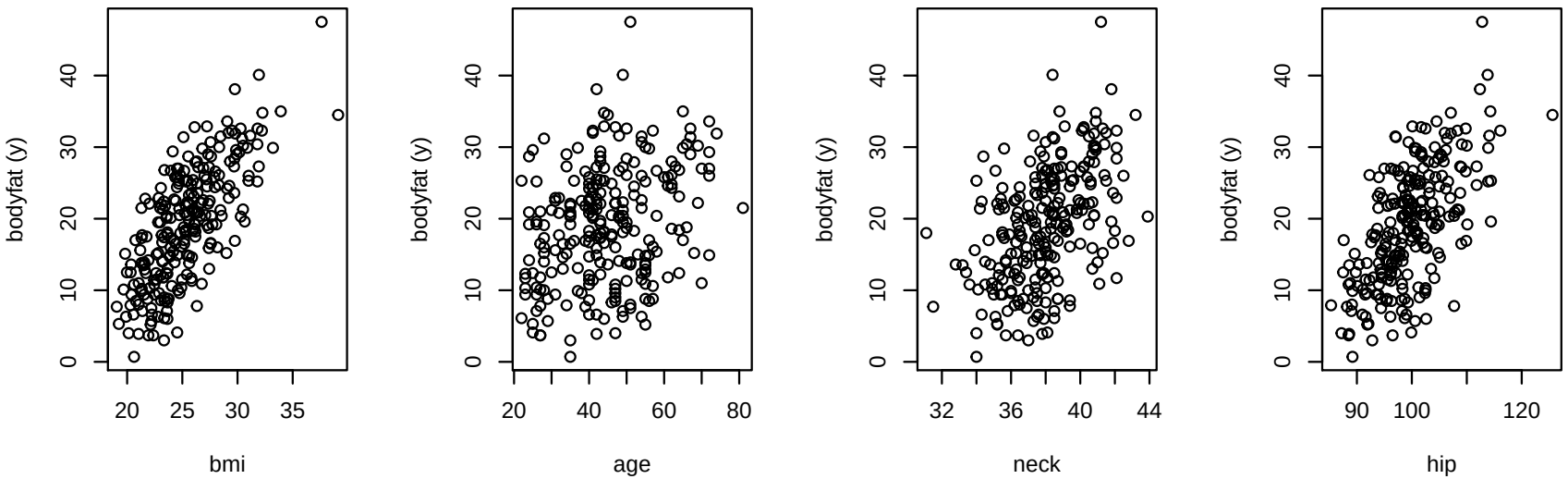
Regresjon – motiverende eksempel

(Veiledet læring - vi kjenner responsen)

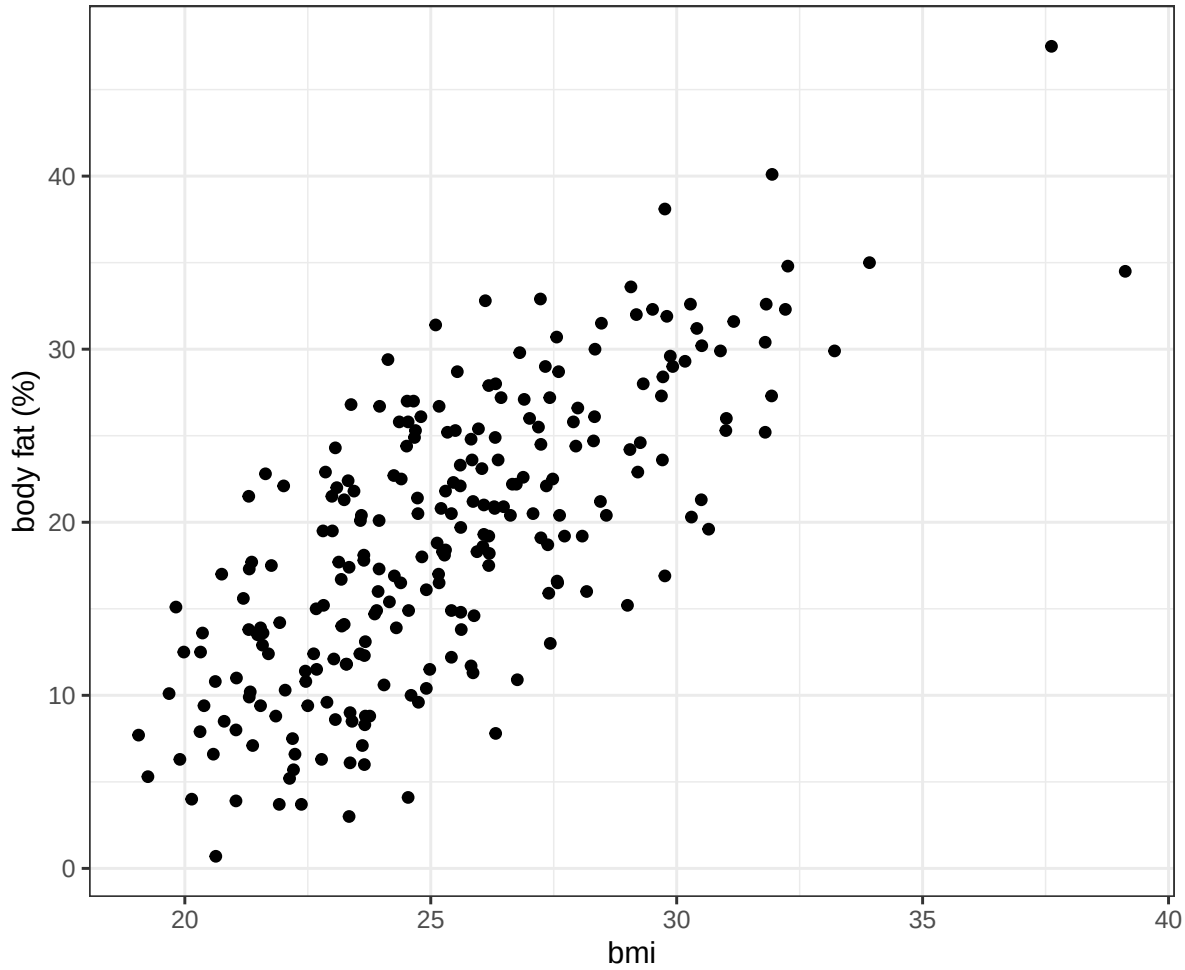
- Kropssfett er en viktig indikator for overvekt, men vanskelig å måle.

Spørsmål: Hvilke faktorer tillater præsis estimering av kroppsfettet?

Vi undersøer 243 mannlige deltakere. Kroppsfett (%), BMI og andre forklaringsvariabler ble målet. Spredningsplott:



For en model for funker god for prediksjon trenger vi *multippel linear regresjon*. Men vi begynner med *enkel linear regresjon* (bare en forklaringsvariabel):



Enkel linear regresjon

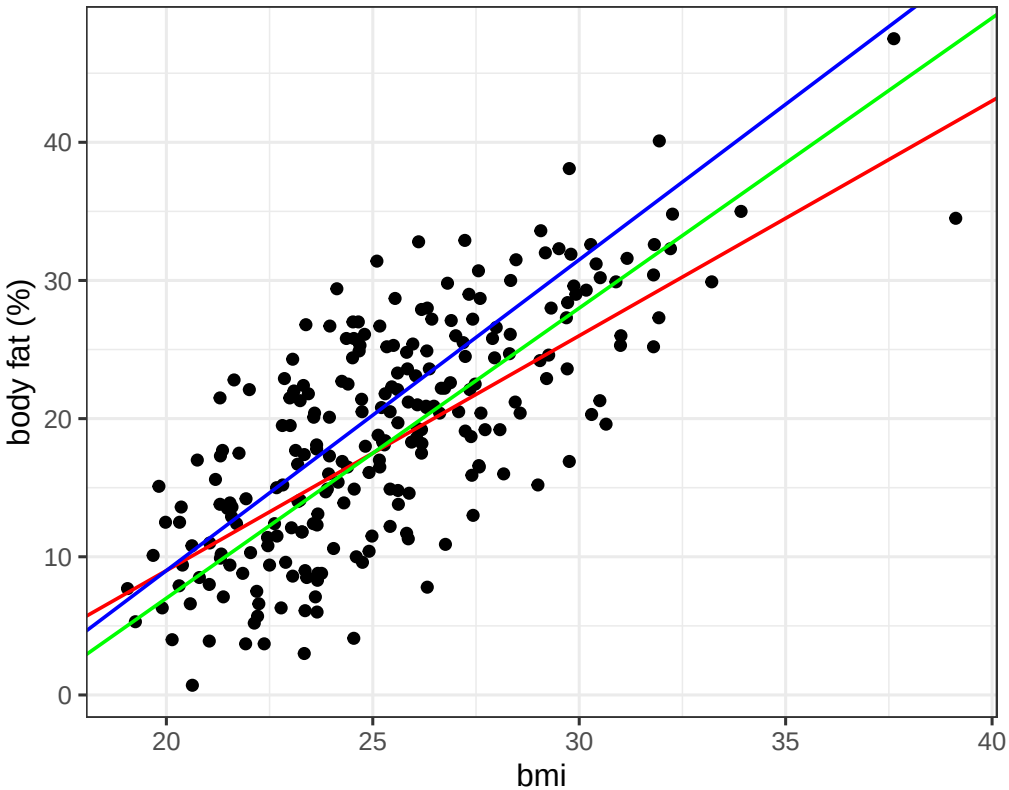
- En kontinuerlig respons variabel Y
- Bare *en forklaringsvariable* x_1
- Relasjon mellom Y og x er antatt å være *linear*.

Hvis den lineare relasjonen mellom Y og x er perfekt, så gjelder

$$y_i = \beta_0 + \beta_1 x_{1i}$$

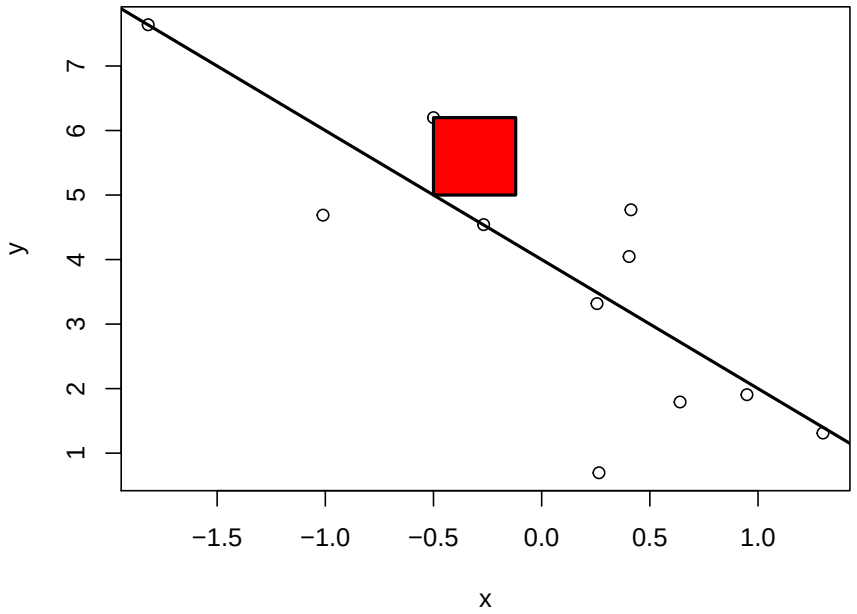
for alle i. Men..

Hvilken linje er best?



Enkel linear regresjon

a) Kan vi tilpasse den "rette" linje til dataene?



- $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i}$.
- $\hat{e}_i = \hat{y}_i - y$
- $\hat{\beta}_0$ og $\hat{\beta}_1$ velges slik at

$$SSE = \sum_i \hat{e}_i^2$$

minimeres.

b) Kan vi tolke linja? Hvor sikkert er jeg på $\hat{\beta}_1$ og linja? Vi trenger antakelser, KI og hypothesetest.

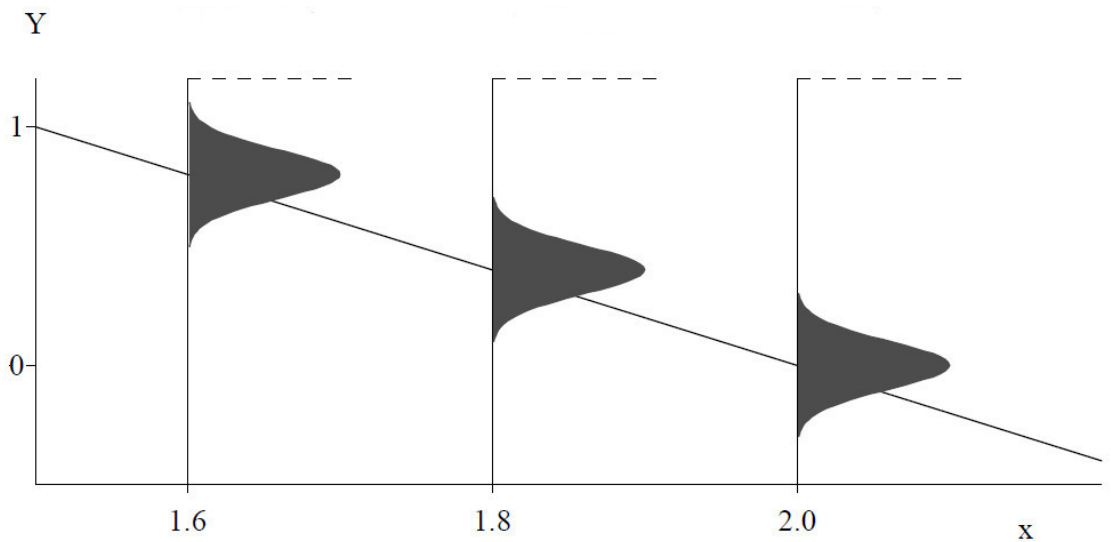
c) Fremtidige presisjoner av predikert y (kroppsfett)?

Linear regresjon – antakelser

$$Y_i = \underbrace{\beta_0 + \beta_1 x_{i1}}_{\hat{y}_i} + e_i$$

med

$$e_i \sim \mathsf{N}(0, \sigma^2) \ .$$



Do-it-yourself “by hand”

Her kan du finne de beste parametrene selv:

You can do this here:

https://gallery.shinyapps.io/simple_regression/

Multippel linear regresjon

Nesten det samme some enkel linear regresjon, we bare summerer flere forklaringsvariabler:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi} + e_i \ , \quad e_i \sim \mathsf{N}(0, \sigma^2) \ .$$

For eksempel:

$$\text{bodyfat}_i = \beta_0 + \beta_1 \text{BMI}_i + \beta_2 \text{age}_i + e_i \ .$$

Regresjonsanalyse i fem steg

Vi skal bruke statmodels.api og statmodels.formula.api for lineær regresjon:

Steg 1: Bli kjent med dataene ved å se på oppsummeringsmål og ulike typer plott

Steg 2: Spesifiser en matematisk modell

Steg 3: Initialiser og tilpass modellen

Steg 4: Presenter resultater fra den tilpassede modellen

Steg 5: Evaluer om modellen passer til dataene

Steg 1: Bli kjent med dataene

- Histogram og boksplott av forklaringsvariable(r) $(x_1, \ldots, x_p)$ og y.