ISTx1002 Usikkerhet og støy i målinger

Usikkerhet og feil i variabler i regresjon

Stefanie Muff, Institutt for matematiske fag, NTNU Trondheim

Oktober 31 2023

Plan for i dag

Pensum og læringsressurser

Husk lenken til den eksterne modulsiden:

https://wiki.math.ntnu.no/istx1002/2023h/start

Pensum del 3:

- Denne forelesningen
- **Disse slides** med alle notater og beregninger som er vist fram
- Shiny app: https://stefaniemuff.shinyapps.io/MEC_ChooseL/

Bruk igjen (kanskje) målingene fra del 1?

https://docs.google.com/spreadsheets/d/1jswtNl8zmIdWwjuuRKSnKzaN-M5DLFISrnCIXl17w10/edit#gid=0

Oversikt

- Usikkerhet og målefeil i en variable (x) og i responsen (y) i en enkel linear regresjon.
- Effekt av usikkerhet og målefeil i regresjon.
- Når bør man være bekymret?
- Enkle metoder for å korrigere for usikkerhet og feil i en regresjonsvariabel.

Kilder til usikkerheit og målefeil i regresjonsvariabler.

- **Upresise målinger** i feltarbeid eller labor (lengde, vekt, blodtrykk...)
- Feil grunnet **ufullstendige eller skjeve observasjoner** (for eksempel selvrapporterte kostvaner, helsehistorie).
- Skjeve observasjoner grunnet **preferansebasert utvalg** eller gjentatte observasjoner.
- Avrundingsfeil, sifferpreferanse.
- **Feilklassifisering** (for eksempel feil i eksponerings- eller sykdomsklassifisering).

Merk: "Feil" og "usikkerhet" er ofte brukt synonymt.

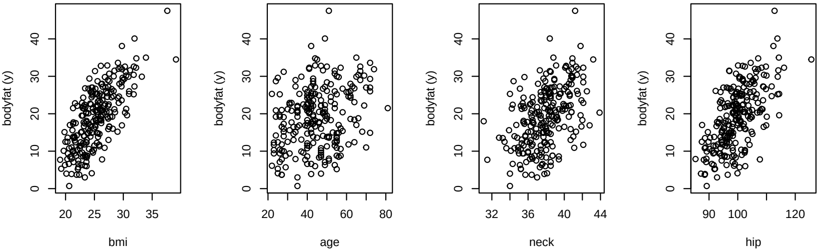
Linear regresjon - hva var det igjen for noe?

Motiverende eksempel

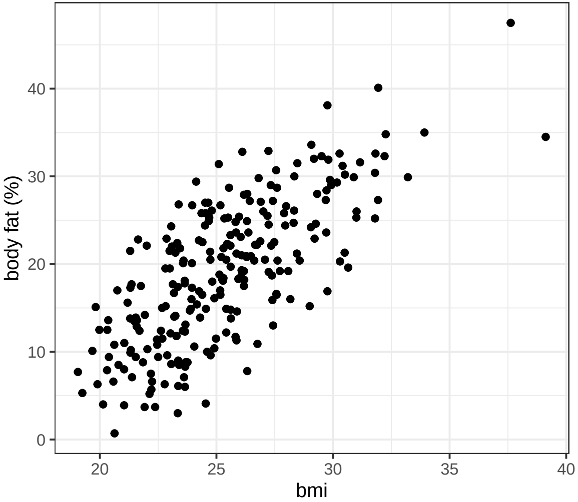
- Kropssfett er en viktig indikator for overvekt, men vanskelig å måle.

Spørsmål: Hvilke faktorer tillater præsis estimering av kroppsfettet?

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



Her ser vi bare på *enkel linear regresjon* (regresjon med bare en forklaringsvariabel):



Enkel linear regresjon

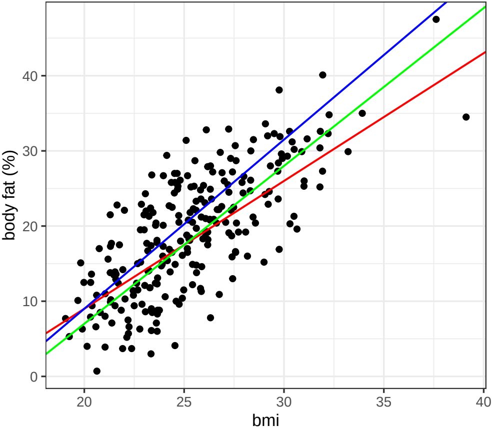
- En kontinuerlig responsvariabel Y
- Bare *en forklaringsvariabel* x
- Relasjon mellom Y og x er antatt å være *linear*.

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

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

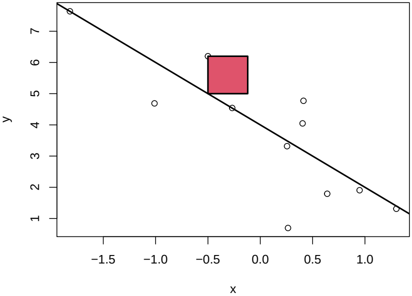
for alle i. Men..

Hvilken linje er best?



Enkel linear regresjon

a) Kan vi tilpasse den “rette” eller “beste” linjen til dataene?



- $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$.
- $\widehat{e}_i = \widehat{y}_i - y$
- $\widehat{\beta}_0$ og $\widehat{\beta}_1$ velges slik at

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

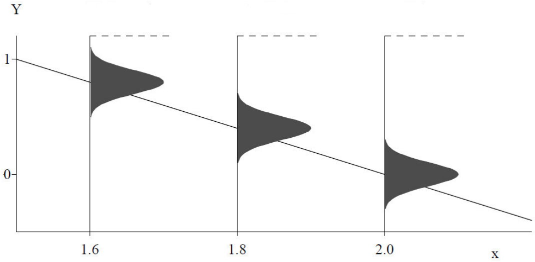
minimeres.

Linear regresjon – fundamentale antakelser

$$Y_i = \underbrace{\beta_0 + \beta_1 x_i}_{\hat{y}_i} + \varepsilon_i$$

med

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



Todo: Say (in class) that this implies four things: mean 0, var constant, Normal distribution and independence.

Gjør vi andre antagelser for en linear regresjonsmodell?

Ja! Men hvilke?

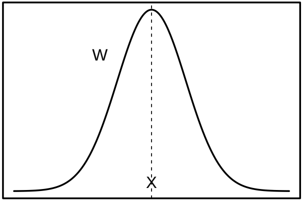
www.menti.com

“Klassisk” målefeil

Kanskje det mest vanlige tilfelle av feil og usikkerhet er så-kalt klassisk målefeil.

Vi vil måle størrelsen X, men vi kan bare måle W med feil U:

$$\begin{aligned} W &= X + U \\ U &\sim \mathrm{N}(0, \sigma_u^2) \end{aligned} \quad .$$



Eksempel: Måling av vår reaksjonstid, upresise målinger av en konsentrajon, en vekt etc.

Men hva skjer hvis variabel x har målefeil/usikkerhet, og inngår som forklaringsvariable i en regresjon?

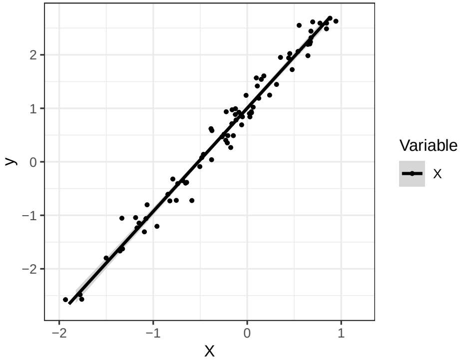
Du kan se selv:

https://stefaniemuff.shinyapps.io/MEC_ChooseL/

Illustrasjon

Vi genererer data som samsvarer med modellen nedenfor, og så vil vi estimere β_0 og β_x.

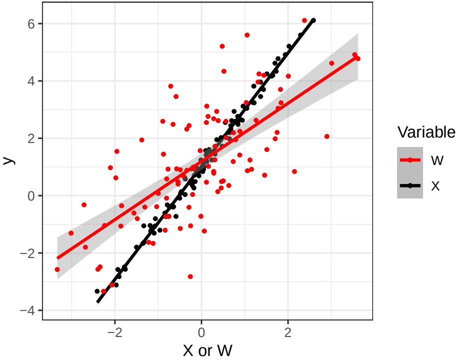
$$Y = \underbrace{1}_{\beta_0} + \underbrace{2}_{\beta_x} \cdot X + \varepsilon, \quad \varepsilon \sim \mathrm{N}(0, \sigma^2) \quad \text{med} \quad \sigma^2 = 1\ .$$



Illustrasjon, del II

Det dumme er at vi ikke kjenner X, men bare en upresis versjon W som vi kan bruke som forklaringsvariabel:

$$W = X + U \ , \qquad U \sim \mathrm{N}(0, \sigma_u^2) \quad \text{med } \sigma_u^2 = 0.8 \ .$$



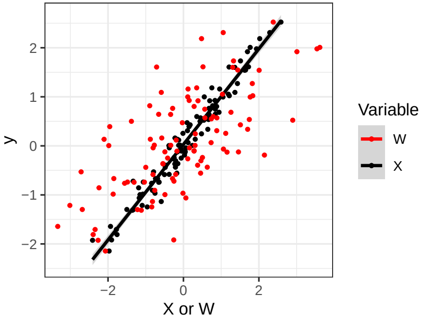
Den “tredoble utfordringen med målefeil’ ’

(Carroll et al., 2006}

1. Bias (skjevhet): Inkludering av feilaktige variabler i etterfølgende analyser kan føre til skjeve estimater av parameter.

2. ME fører til en tap av styrke (power) for å oppdage signaler.

3. ME skjuler viktige egenskaper ved dataene, noe som gjør det vanskelig å inspisere grafiske modeller.



Hvordan kan vi handtere målefeil i en regressjonsvariabel?

- Generelt sett hjelper det når man har en ide hvordan feil og usikkerhet oppstår.
- Konkret trenger vi en **feilmodell** og kunnskap om **parametrene** i feilmodellen.

Eksempel: For klassisk målefeil $W = X + U$ med $U \sim \mathrm{N}(0, \sigma_u^2)$ trenger vi kunnskap om variansen σ_u^2.

Strategi:

1) Ta gjentatte målinger (som for reaksjonstiden), så kan du estimere variansen (i.e., usikkerheten).
2) Bruk kunnskap om modellen og usikkerheten i X for å korrigere feilen i regresjonsparametrene (særlig i β_x).

En formel for å handtere et enkelt tilfelle

Vi fortsetter med klassisk målefeil i linear regresjon $y = \beta_0 + \beta_x x + \varepsilon_i$ hvor vi dessverre bare måler en feilaktig versjon $w = x + u$.

Siden vi ikke kjenner x, har vi i første omgang ikke noe annet valg enn å bruke w og tilpasse en modell gitt som

$$y = \beta_0^\star + \beta_x^\star w + \varepsilon_i \ .$$

Er det en god idé?

Vi har jo sett at parameteren $\beta_x^\star < \beta_x$, og det viser seg at den forventete verdien reduseres med en faktor $\lambda < 1$

$$\mathrm{E}[\beta_x^\star] = \underbrace{\frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2}}_{\lambda} \mathrm{E}[\beta_x] \ , \tag{1}$$

hvor λ heter attenueringsfaktor.

Oppgave:

- Se igjen på eksempelet vi brukte for illustrasjon. Hvor mye reduksjon forventer vi i estimaten til stigningstallet β_x?
- I praksis har vi jo *ikke* tilgang til de ekte verdiene x men bare kjenner de med usikkerhet w. Hvordan kan du bruke kunnskapet om formel (1) for å finne en korrigert versjon for estimaten av β_x?

Utregning:

Desssverre finnes det ikke mange tilfeller hvor vi har noe som formel (1).

En intuitiv ide: Simulation Extrapolation (SIMEX)

Foreslatt av Cook og Stefanski (1994).

SIMEX består av to faser:

- **Fase 1: Simulasjon.** Usikkerheten i dataen (x variable) er økt for å bestemme hvordan størrelsen av interesse (vanligvis en regresjonsparameter/stigningstall) påvirkes av feilen/usikkerheten.

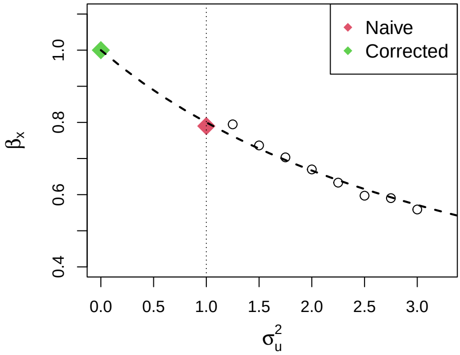
- **Fase 2: Ekstrapolajon** Den observerte trenden ekstrapoleres deretter *tilbake* til en hypotetisk feilfri verdi.

Illustrasjon av SIMEX ideen

Størrelsen av interesse: β_x (stigningstall i en regresjon).

Problem: Forklaringsvariable x er målt med usikkerhet:

$$w = x + u \ , \quad u \sim \mathrm{N}(0, \sigma_u^2) \ .$$



Exsempel: Vi ser igjen på en regresjonsmodell med kroppsfett som respons (y) og BMI som forklaringsvariabel x:

$$y_i = \beta_0 + \beta_x x_i + \varepsilon_i \ , \quad \varepsilon_i = \mathrm{N}(0, \sigma^2)$$

med

$\mathbf{y} = (y_1, \ldots, y_{100})^\top$: variabel med % kroppsvet målt ved 100 personer.

Problemet: BMI-en ble selvrapportert og har derfor målefeil! Ikke x_i blir observert, men heller

$$w_i = x_i + u_i \ , \quad u_i \sim \mathrm{N}(0, 4) \ .$$

$\rightarrow$ Bruk SIMEX proseduren!

Vi sammenligner

- en regresjon hvor vi bruker den feilaktige BMI-en som forklaringsvariable, med
- en korrigert version som vi får etter vi anvender SIMEX proseduren.

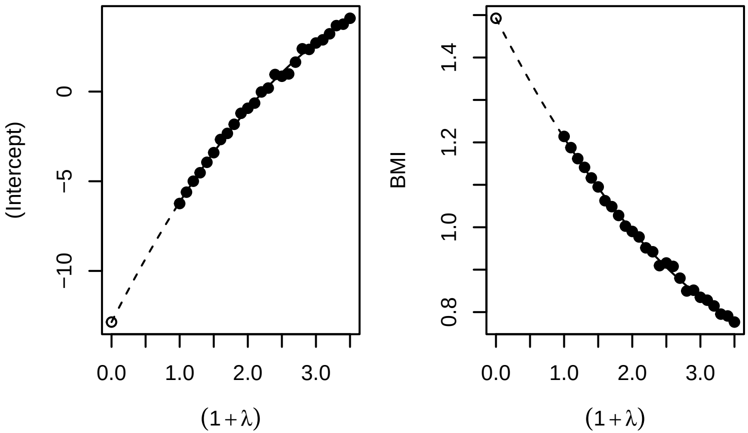
Naiv resultat med med feilaktig BMI:

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -6.240988 2.12868299 -2.931854 4.194014e-03
## BMI          1.214017 0.08860899 13.700829 1.673886e-24
```

Estimater etter SIMEX-korrekturen ble anvendt:

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -12.851018  3.1273875 -4.109186 8.247864e-05
## BMI           1.492441  0.1275759 11.698449 2.632916e-20
```

Se på grafiske resultater med en kvadratisk ekstrapolasjonsfunksjon:



Multippel linear regresjon

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

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \varepsilon_i \ , \quad \varepsilon_i \sim \mathsf{N}(0, \sigma^2) \ .$$

For eksempel:

$$\text{bodyfat}_i = \beta_0 + \beta_1 \text{bmi}_i + \beta_2 \text{age}_i + \varepsilon_i \ .$$

Hva skjer hvis en variabel i en *multippel regresjon* har usikkerhet?

Exsempel: Vi ser igjen på en regresjonsmodell med kroppsfett som respons (y) og BMI (x) som forklaringsvariabel, men nå har vi i tillegg også sex (z) som forklaringsvariabel:

$$y_i = \beta_0 + \beta_x x_i + \beta_z z_i + \epsilon_i \ , \quad \epsilon_i = \mathrm{N}(0, \sigma^2)$$

med $\mathbf{y}$ og $\mathbf{x}$ som før, med

$$w_i = x_i + u_i \ , \quad u_i \sim \mathrm{N}(0, 4) \ ,$$

og den binære variablen $\mathbf{z} = (z_1, \ldots, z_{100})^\top$ som indikerer om person i er en mann ($z_i = 1$) eller en kvinne ($z_i = 0$).

Vi sammenligner igjen

- en regresjon hvor vi bruker den feilaktige BMI-en som forklaringsvariable, med
- en korrigert version fra SIMEX.

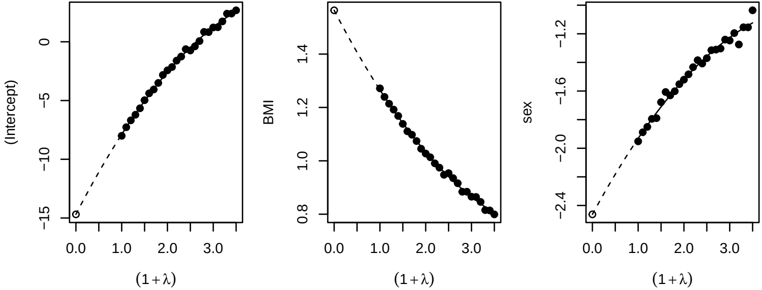
Naiv resultat med feilaktig BMI::

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -8.003714 2.07060335 -3.865402 2.005407e-04
## BMI          1.271558 0.08821382 14.414504 7.478782e-26
## sex         -1.951735 0.73625960 -2.650879 9.376840e-03
```

Estimater etter SIMEX-korrekturen ble anvendt:

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -14.689940  2.6954519 -5.449899 3.825138e-07
## BMI           1.564059  0.1159075 13.494022 5.467540e-24
## sex          -2.462127  0.7906688 -3.113980 2.426632e-03
```

Graphical results with quadratic extrapolation function:



Note: The `sex` variable has *not* been mismeasured, nevertheless it is affected by the error in BMI!

Reason: `sex` and BMI are correlated.