

ISTx1002 Usikkerhet og støy i målinger

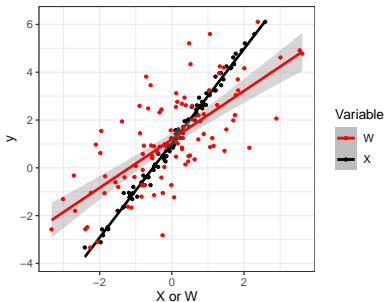
Usikkerhet og feil i variabler i regresjon

Stefanie Muff, Institutt for matematiske fag, NTNU Trondheim

Oktober 31 2023

Plan for i dag

- Lineær regresjon
- Klassisk målefeil i en forklaringsvariable
- To strategier for å handtere målefeil:
 - Med en analytisk formel \rightarrow Attenueringsfaktor
 - Med en heuristisk idé \rightarrow SIMEX



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>

Vi skal se hvordan måleusikkerheten påvirker en regressjon vi vil gjøre (detaljene kommer).

Oversikt

- Usikkerhet og målefeil i en variable (x) i en enkel lineær 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.

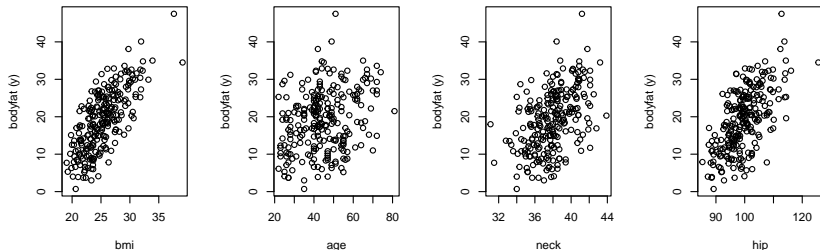
Lineær regresjon - hva var det igjen for noe?

Motiverende eksempel

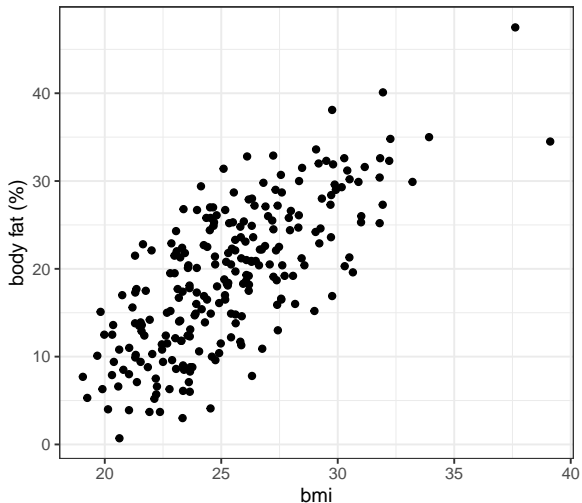
- Kroppsfett er en viktig indikator for overvekt, men vanskelig å måle.

Spørsmål: Hvilke faktorer tillater præsis estimering av kroppsfettet?

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



Vi begynner med *enkel lineær regresjon* (regresjon med bare en forklaringsvariabel):



Enkel lineær regresjon

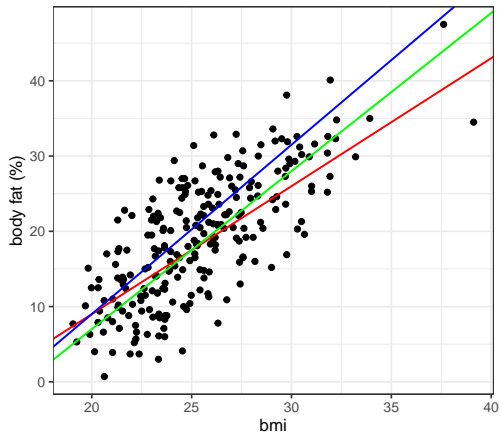
- En kontinuerlig responsvariabel Y
- Bare *en forklaringsvariabel* x
- Relasjon mellom Y og x er antatt å være *lineær*.

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

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

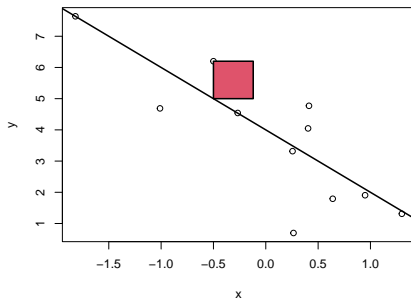
for alle i . Men..

Hvilken linje er best?



Enkel lineær regresjon

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



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

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

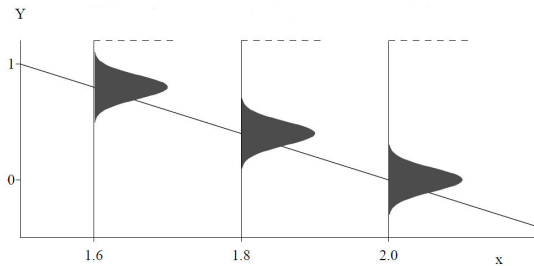
minimeres.

Lineær regresjon – fundamentale antakelser

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

med

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



Gjør vi andre antagelser for en lineær regresjonsmodell?

Ja! Men hvilke?

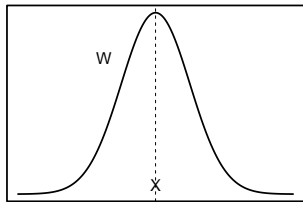
www.menti.com

“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 :

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



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

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

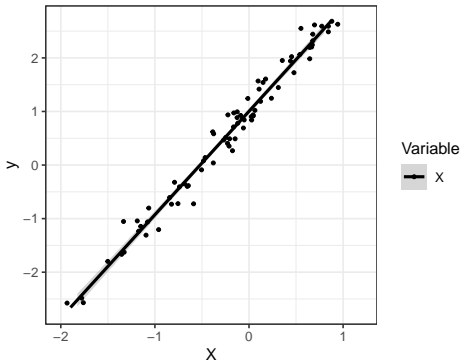
Du kan se selv:

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 . Vi vet at $X \sim N(0, \sigma_x^2)$ med $\sigma_x^2 = 1$, og

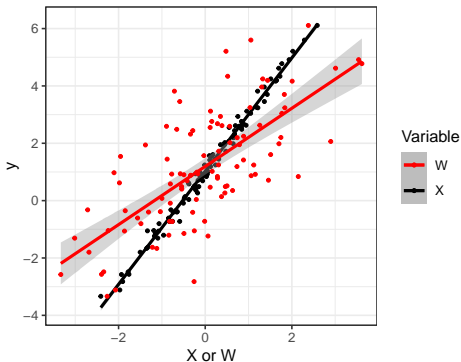
$$Y = \underset{\beta_0}{1} + \underset{\beta_x}{2} \cdot X + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2) \quad \text{med} \quad \sigma^2 = 1.$$



Illustrasjon, del II

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

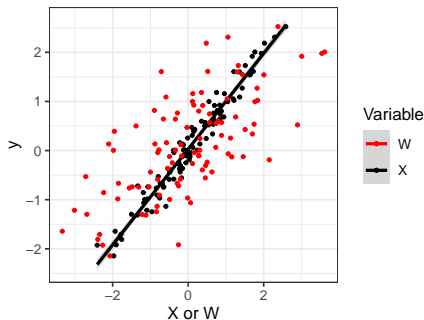
$$W = X + U, \quad U \sim N(0, \sigma_u^2) \quad \text{med } \sigma_u^2 = 0.8 .$$



Den “tredoble utfordringen med målefeil”

(Carroll et al., 2006)

1. **Skjevhet (bias)**: 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 idé hvordan feil og usikkerhet oppstår.
- Konkret trenger vi en **feilmodell** og kunnskap om **parametrene** i feilmodellen.

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

Mulige Strategier:

- 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 lineær 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

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

Er det en god idé?

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

$$E[\beta_x^*] = \underbrace{\frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2}}_{\lambda} \beta_x , \quad (1)$$

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 ?

Utgivning:

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

En heuristisk idé: Simulation Extrapolation (SIMEX)

Foreslatt av Cook og Stefanski (1994).

SIMEX består av to faser:

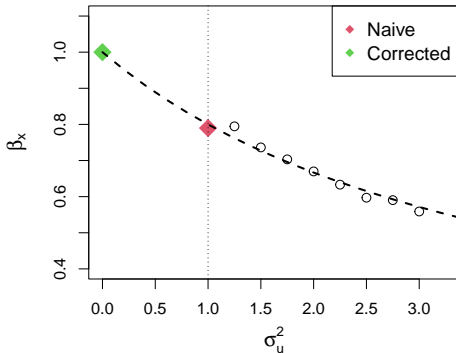
- **Fase 1: Simulasjon.** Usikkerheten i dataen (x variable) er økt for å bestemme hvordan størrelsen av interesse (vanligvis en regresjonsparameter/stigningstall β_x) påvirkes av feilen/usikkerheten.
- **Fase 2: Ekstrapolajon** Den observerte trenden ekstrapoleres deretter *tilbake* til en hypotetisk feilfri verdi.

Illustrasjon av SIMEX ideen

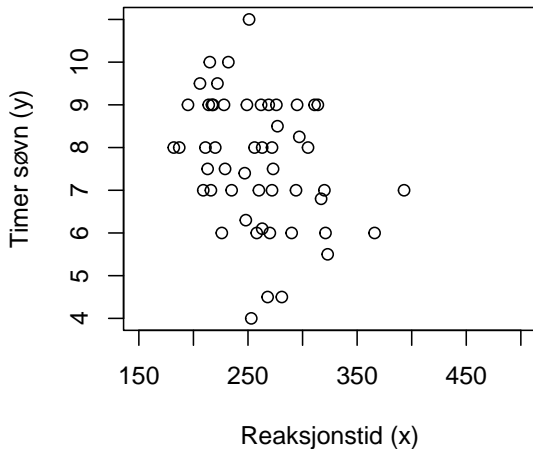
Størrelsen av interesse: β_x (stigningstall i en regresjon).

Problem: Forklaringsvariablen x er målt med usikkerhet:

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



Exsempel 1: Våre reaksjonstider



Modell:

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

med x_i =søvnstid og x_i =reaksjonstid til person i , målt i 3. omgang.

Problem: Vi vet jo at reaksjonstiden er ganske unøyaktig og har mye målefeil i seg. Vi kan estimere feilvariansen (det har jeg gjort for dere ved bruk av deres data) som

$$\sigma_u^2 = 565$$

For å bruke SIMEX proseduren, antar vi jo

$$w_i = x_i + u_i , \quad u_i \sim N(0, \sigma_u^2) .$$

Det er en ganske plausibel antakelse, og så bruker vi SIMEX proseduren!

Vi sammenligner

- en regresjon hvor vi bruker reaksjonstiden ved 3. målingen som forklaringsvariable, med
- en korrigert version som vi får etter vi anvender SIMEX proseduren.

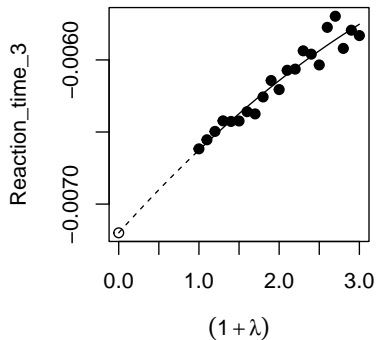
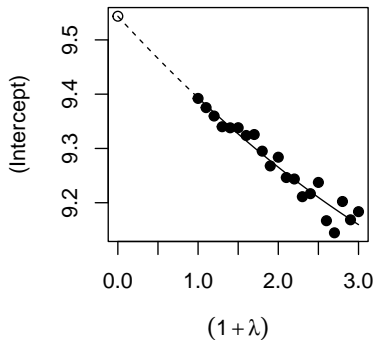
Naiv resultat med med feilaktig BMI:

##	Estimate	Std. Error	t value	Pr(> t)
## (Intercept)	9.392168756	0.669439617	14.029897	3.852253e-19
## Reaction_time_3	-0.006616798	0.002374351	-2.786782	7.458553e-03

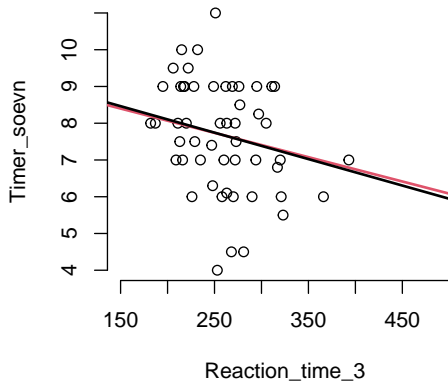
Estimater etter SIMEX-korrektoren ble anvendt:

##	Estimate	Std. Error	t value	Pr(> t)
## (Intercept)	9.543600821	0.559054066	17.070980	9.805241e-23
## Reaction_time_3	-0.007199182	0.001912446	-3.764384	4.328353e-04

Se på grafiske resultater med en kvadratisk ekstrapolasjonsfunksjon:



Uten (rødt) og med korrektur (svart):



Eksempel 2: Sammenheng av BMI og kropps fett

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

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

med

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

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

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

→ Bruk SIMEX proseduren!

Vi sammenligner

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

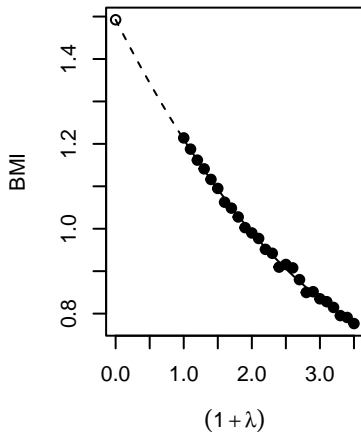
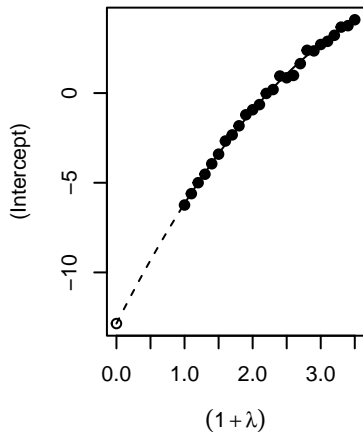
Naiv resultat med med feilaktig BMI:

	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

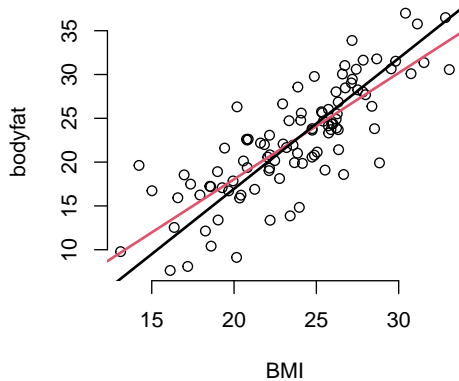
Estimater etter SIMEX-korrektoren ble anvendt:

	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:



Uten (rødt) og med korrektur (svart):



Multippel lineær regresjon

Nesten det samme som enkel lineær regresjon, vi bare summerer flere forklaringsvariabler:

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

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?

Eksempel: Vi ser igjen på en regresjonsmodell med kropps fett som respons (y) og BMI (x) som forklaringsvariabel, men nå har vi i tillegg også kjønn (z) som forklaringsvariabel:

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

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

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

og den binære variabelen $\mathbf{z} = (z_1, \dots, z_{100})^\top$ 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.

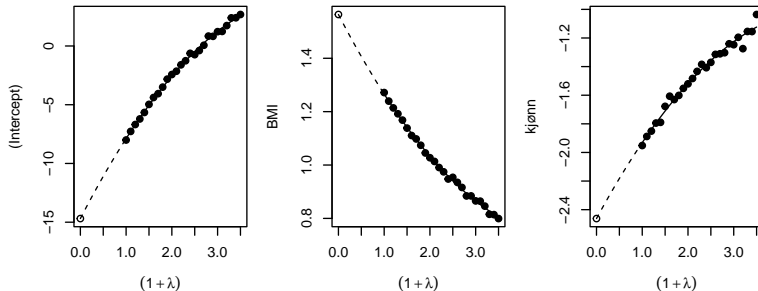
Naiv resultat med feilaktig BMI::

##	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
## kjønn	-1.951735	0.73625960	-2.650879	9.376840e-03

Estimater etter SIMEX-korrektoren 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
## kjønn	-2.462127	0.7906688	-3.113980	2.426632e-03

Grafiske resultater med kvadratisk ekstrapolasjonsfunksjon:



Merk: Variabelen `kjønn` har ikke blitt feilmålt, likevel er stigningstallet påvirket av feilen i BMI!

Grunn: `kjønn` og BMI er korrelert.