

# Kurs i dataeditering: Validere og kontrollere

ASLAUG HURLEN FOSS 2023



**Statistisk sentralbyrå**  
Statistics Norway

# Plan for kurset

- 10:00-10:40 Validering av numeriske og kategoriske verdier
- 10:40-11:30 Øvelse i R
- 11:30-12:00 Lunsj
- 12:00-12:40 Kvartilmetode, HB-metode, regresjon og innflytelse
- 12:40-13:30 Øvelse i R
- 13:30-14:00 Gjennomgang av øvelser og oppsummering



# Læringsmålet

- Målet er at alle skal lære de mest kjente metodene for å kontrollere data
- Kunne bruke R til å sette opp kontrollene.



# Materiale for kurset

- **Github:** <https://github.com/statisticsnorway/kurs-metode-validere>
- **Byrånettesiden «dataeditering»:** lenke til materialet og lenker til all bakgrunnsliteratur

<https://ssbno.sharepoint.com/sites/Metodikkistatistikkproduksjonen/SitePages/Dataeditering.aspx>



# Dataeditering

Dataeditering er kontroll, granskning og retting av data. All statistikk som publiseres er basert på data som har blitt kontrollert og i de fleste tilfeller også korrigert. Dataeditering er dermed en av de viktigste prosesser i en statistikkproduksjon for å sikre god kvalitet. Her får du tips til hvordan gjøre dataediteringen effektivt.

> 10 tips til dataeditering

---

> Prosesser i dataeditering

---

> Modernisering av dataeditering

---

> Programvare

---

> Kurs

---



Datarevisjon. Kontroll, granskning og retting av data.  
Anbefalt praksis - SSB



ESS Handbook - Methodology for data validation  
v1.1 - Rev2018 | CROS



Generic statistical data editing models -GSDem -  
UNECE Statswiki



Analysér og presentasjoner - wiki



Metoder brukerdokumentasjon - wiki

## Kontaktpersoner



Foss, Aslaug Hurlen  
Seniorrådgiver



Jentoft, Susie  
Seniorrådgiver



Seierstad, Ane  
Seniorrådgiver



Statistisk sentralbyrå  
Statistics Norway

# Følg med på Yammer....

Foss, Aslaug Hurlen

Hjemme-feed

Fellesskap

Innboks 10

Favoritter

Forbedre Yammer-opplevelsen din ved å lage en liste over favorittfellesskapene dine.

Mine samfunn

Dataeditering

Infografikk og datavisualise...

Hele SSB

R i SSB 1

SSB Kompetanse 1

Dapla

Hack4ssb

ISEE - Integrert System for ...

Innovasjonsgruppa A400

Sesongjustering

Konfidensialitet (SDC) i SSB

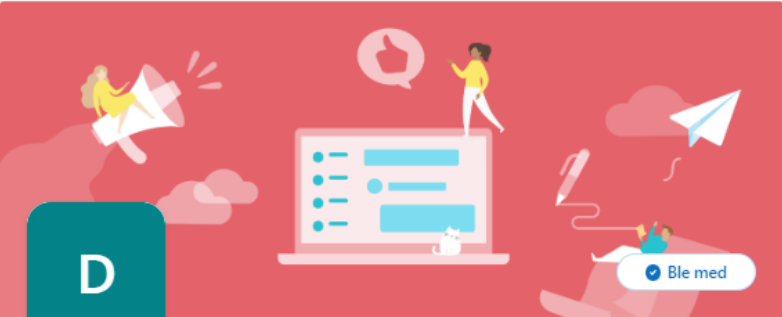
Gruppelederutvikling i SSB

VTL - Validation and Transf...

GDPR i SSB

Opprett et fellesskap

D



Ble med

Dataeditering

Samtaler Om Filer Arrangementer

Del tanker, ideer eller oppdateringer

Diskusjon Spørsmål Ros Avstemning

Alle samtaler

Nye innlegg

Foss, Aslaug Hurlen

1. jul.

Sett av 208

Svensk strategi

Denne artikkelen fra Sverige om strategi for dataeditering synes jeg var fin og stemmer godt overens med de prinsipper vi har!

Topic 3 SAB 6 May 2022 (003).pdf

Dataeditering • Filer

Lik

Kommenter

Del

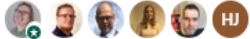
Thindberg, Christian og 3 andre

Thindberg, Christian

14. jul.

Veldig spennende [Foss, Aslaug Hurlen](#)!

Medlemmer • 18



Dette er en gruppe for nyheter og spørsmål om dataeditering - kontroll, granskning og retting av data.  
Edit description

Informasjon

Dataeditering er kontroll, granskning og retting av data. All statistikk som publiseres er basert på data som har blitt kontrollert og i de fleste tilfeller også korrigert. Dataeditering er dermed en av de viktigste prosesser i en statistikkproduksjon for å sikre god kvalitet.

Festet

Hjemmesiden til edite...

Fellesskapsressurser

SharePoint-bibliotek

SharePoint-område

OneNote

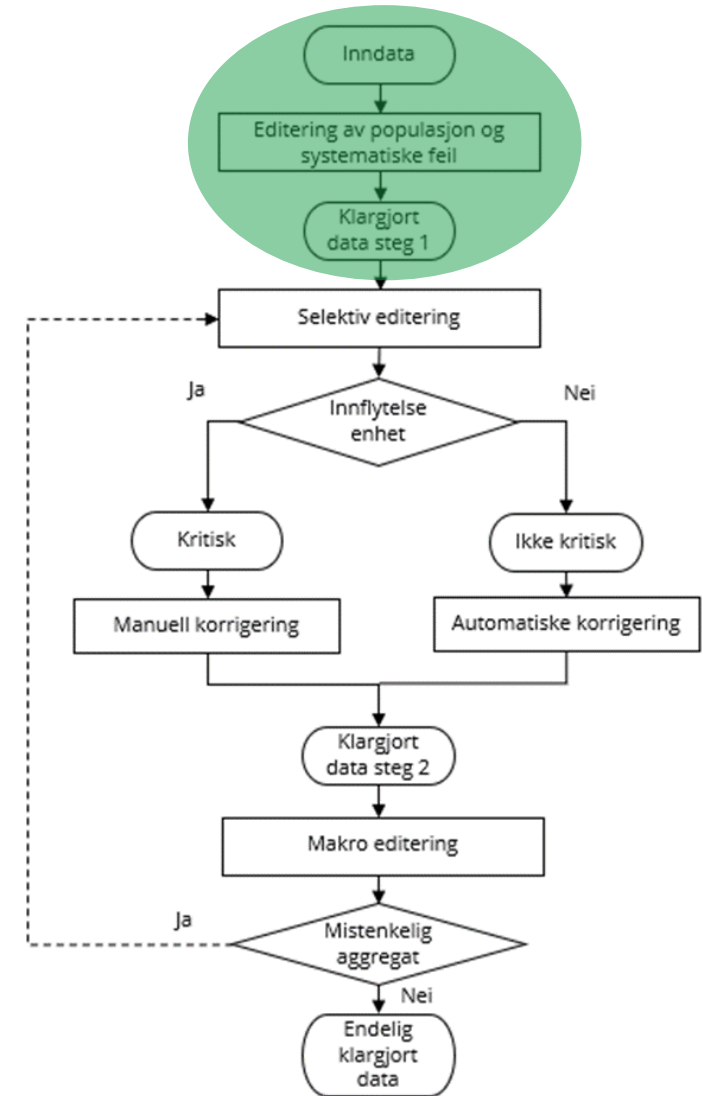
Planner

Opprett direktesendt arrangement

Statistisk sentralbyrå  
Statistics Norway

# Validering og prosessmodell

- Datavalidering er å verifisere om verdien er akseptabel
- Første prosess i prosessen GSDEM – Generic Statistical Data Editing Model
- <https://statswiki.unece.org/display/sde/GSDEM>



# Eksempler

- Ulovlige verdier - f.eks negative verdier
- Ulovlig kodeverdi - f.eks utgått kommunenummer
- Logiske feil – f.eks summen av alle underposter er forskjellig fra totalen





# Håndbok i datavalidering

## Methodology for data validation 2.0

Revised edition 2018

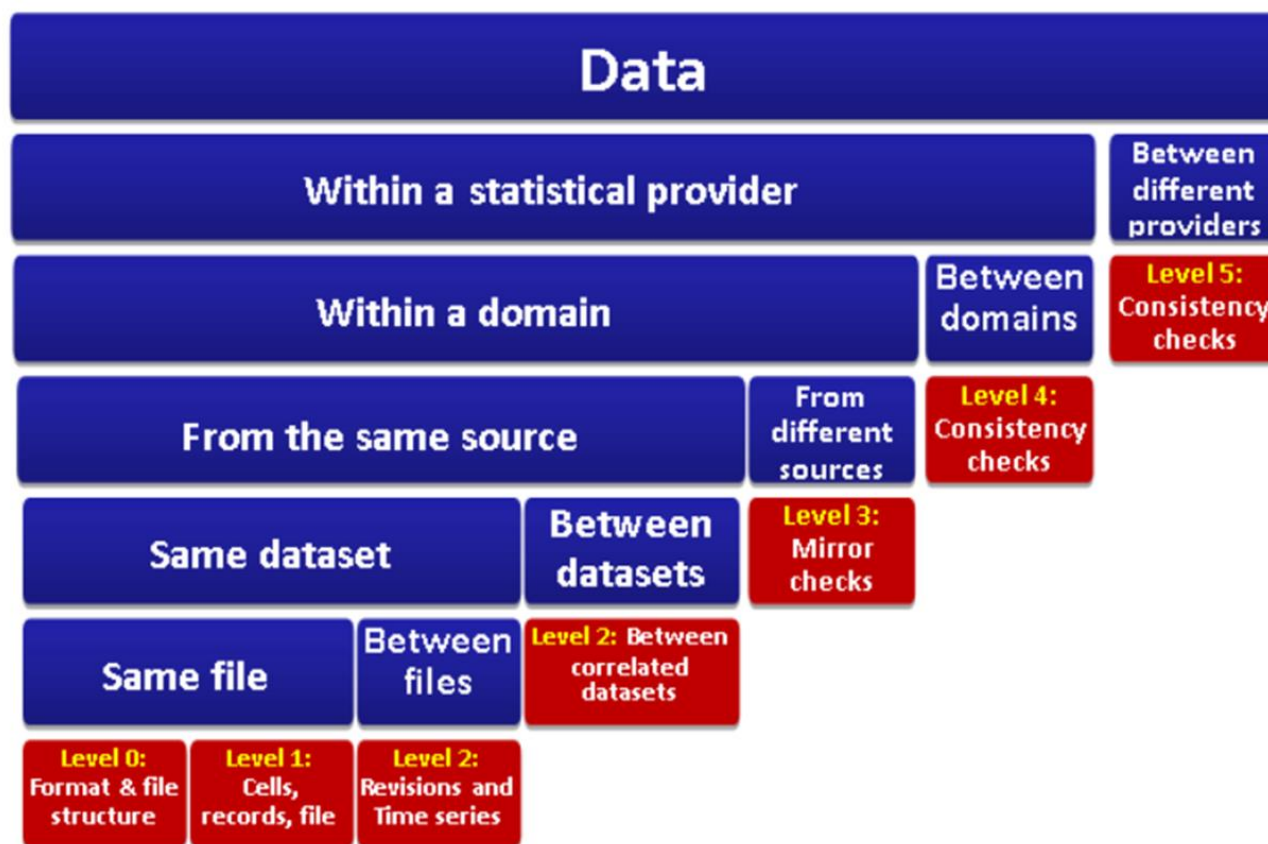
- Methodology for data validation, EUROSTAT
- [https://ec.europa.eu/eurostat/cros/system/files/ess\\_handbook\\_-\\_methodology\\_for\\_data\\_validation\\_v2.0\\_-\\_rev2018\\_0.pdf](https://ec.europa.eu/eurostat/cros/system/files/ess_handbook_-_methodology_for_data_validation_v2.0_-_rev2018_0.pdf)



**Statistisk sentralbyrå**  
Statistics Norway

# Nivåer for kontroller

Figure 1. Graphical representation of validation levels



- Innen en enhet (record)
- Innen et datasett
- Mellom datasett
- Konsistenssjekker mellom separate domener tilgjengelig i samme institusjon

# Type funksjoner

**Table 4: Overview of the classes and examples of numerical data**

Class ( $U\tau uX$ )	Description of input	Example function	Description of example
$ssss$	Single data point	$x > 0$	Univariate comparison with constant
$sssm$	Multivariate (in-record)	$x + y = z$	Linear restriction
$ssms$	Multi-element (single variable)	$\sum_{u \in S} x_u > 0$	Condition on aggregate of single variable
$ssmm$	Multi-element multivariate	$\frac{\sum_{u \in S} x_u}{\sum_{u \in S} y_u} < \epsilon$	Condition on ratio of aggregates of two

--



# Status for kontroller (regler)

	Respondent Records			Validation Rule Status							Overall Status
	x1	x2	x3	(1)	(2)	(3)	(4)	(5)	(6)	(7)	
record 1	4	3	2	P	P	P	P	P	P	P	P
record 2	4	3	missing	P	P	M	P	P	M	M	M
record 3	6	3	2	P	P	P	P	F	P	F	F
record 4	6	3	missing	P	P	M	P	F	M	M	F



# Enkel analyse av kontrollene (regler)

TABLE 7. COUNTS OF RECORDS THAT PASSED, MISSED AND FAILED FOR EACH VALIDATION RULE

VALIDATION RULE	RECORDS PASSED	RECORDS MISSED	RECORDS FAILED
(1)	4	0	0
(2)	4	0	0
(3)	2	2	0
(4)	4	0	0
(5)	2	0	2
(6)	2	2	0
(7)	1	2	1



# Kvalitetsindikatorer – kontrollere og validere



Indikator	Kommentarer
<b>Utslagsrate</b> - indikator uttrykkes som forholdet mellom antall verdier slått ut i kontroll og totalt antall verdier for en gitt variabel	Identifikasjon av feilaktige data i klargjøring - manglende, ugyldige eller uoverensstemmende oppføringer eller utpeking av dataposter som er feil

Kilde: Quality Indicators for the Generic Statistical Business Process Model (GSBPM) - For Statistics derived from Surveys and Administrative Data Sources. Version 2.0, October 2017



**Statistisk sentralbyrå**  
Statistics Norway

# «Edith» - laget av innovasjonsgruppa på næringsstatistikk - 400

- Prototype generell applikasjon laget i Python og Dash
- Hovedsakelig for numeriske verdier

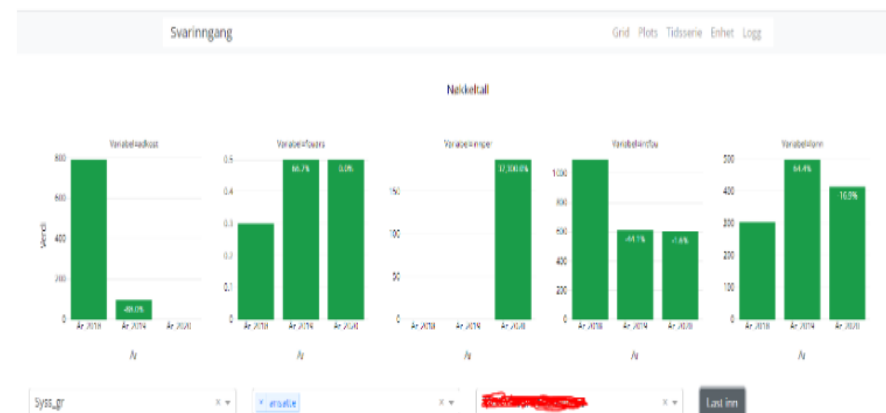
## 5. Enhet

Opprettet av Lyle, Christina, sist endret den 17.02.22

I Enhet velger man gruppering, variabel og foretak for å få oversikt over data fra de siste tre årene. Man editrer og godtar endringene.



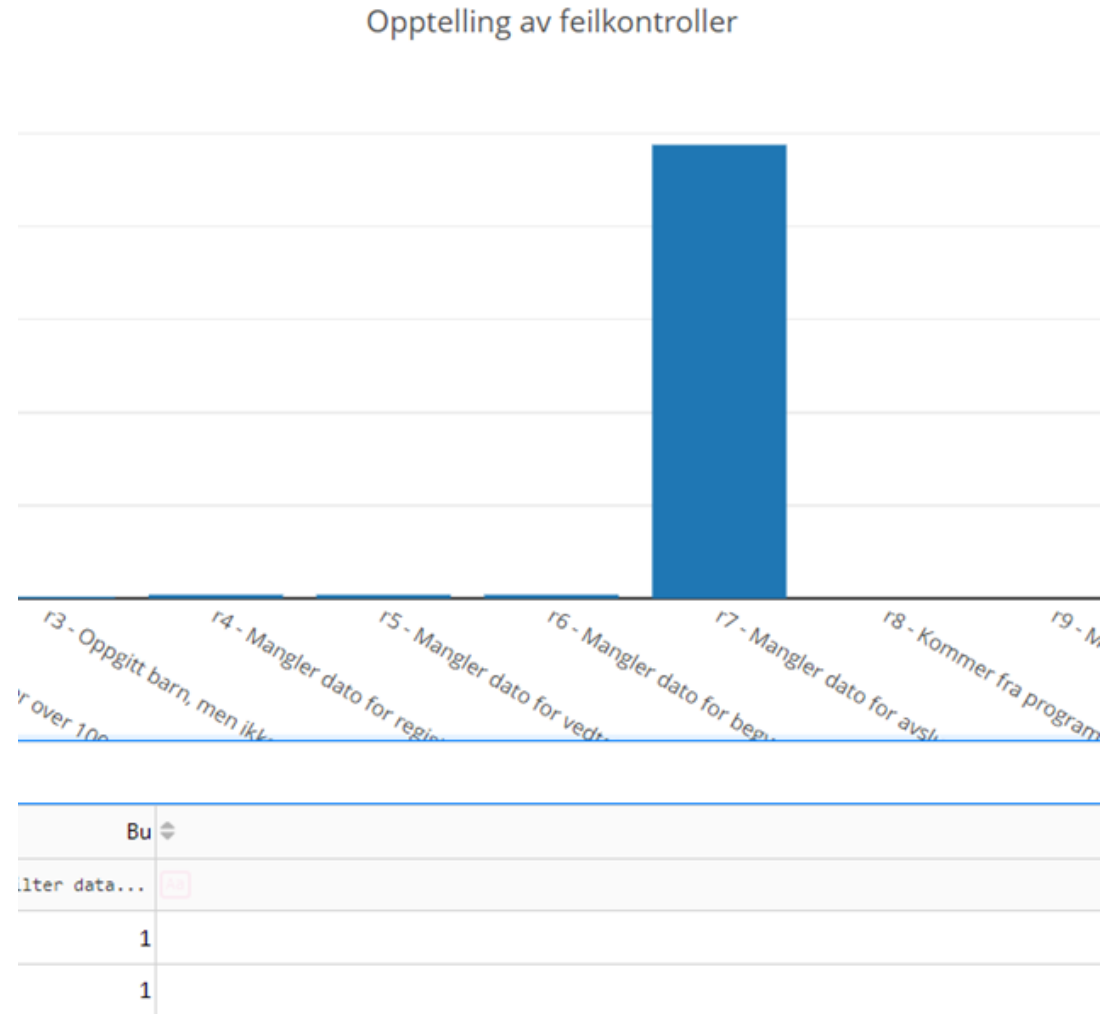
Figur 13



**Statistisk sentralbyrå**  
Statistics Norway

# Prototype fra IT-700

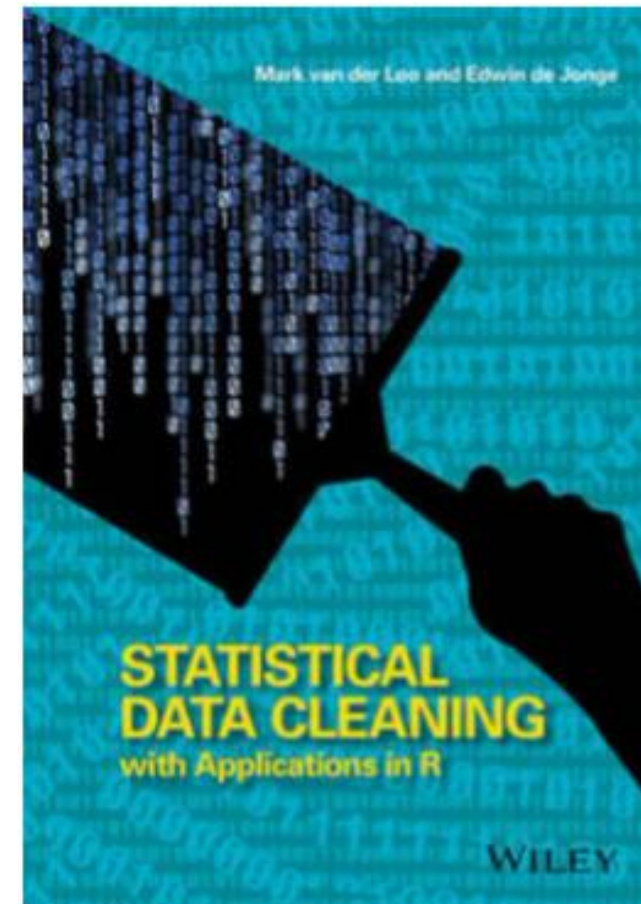
- Laget i Python og Dash
- Sette opp kontroller for numeriske og kategoriske variabler
- Sette opp regelretting
- Endring manuelt
- Logging





# Validate pakke i R

- Valideringspakken er ment å lage:
  - Sjekke data lett
  - Vedlikeholde kontrollene enkelt
  - Mulig å reprodusere resultatene
- Bygget av Mark van der Loo and Edwin de Jonge, Statistics Netherlands
- <https://cran.r-project.org/web/packages/validate/vignettes/cookbook.html>



## The Data Validation Cookbook

Mark P.J. van der Loo

2022-03-24

## Preface

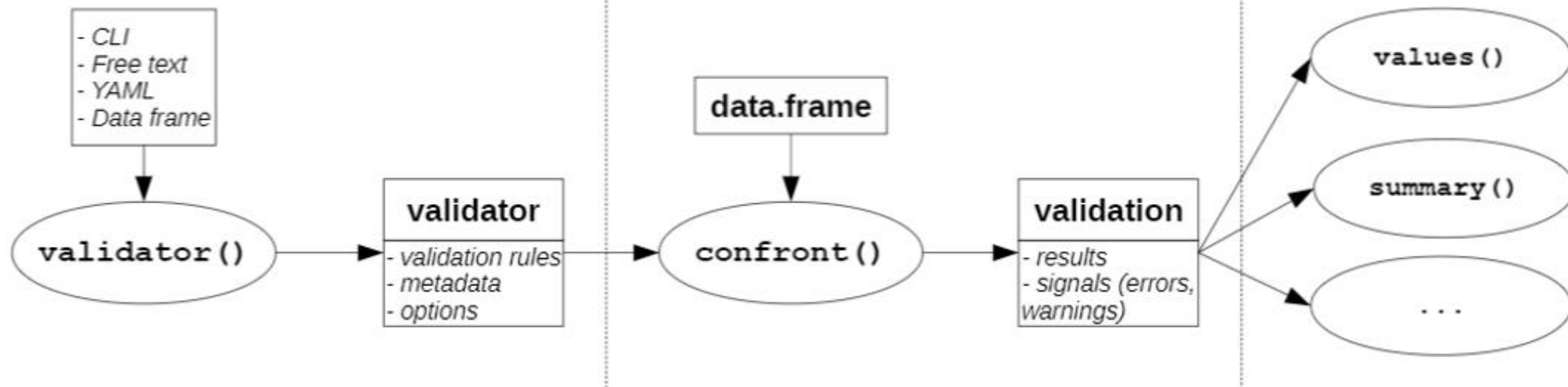
This book is about checking data with the `validate` package for R.

# Validate-pakken

Sette opp og  
vedlikeholde  
kontroller

Kjøre kontroller på  
data

Analyse av  
kontroller



# Datasett

- Lager eksempeldatasett i R
  - `ID<-c("1","2","3","4")`
  - `var1<-c(2,9,-1,7)`
  - `var2<-c(9,1,4,8)`
  - `mydata <- data.frame(ID, var1, var2)`

	▲	ID	▼	var1	▼	var2	▼
1		1		2		9	
2		2		9		1	
3		3		-1		4	
4		4		7		8	



# Validator

Objekt

Funksjon

Kontroll

`v <- validator( var1 > 0, var1 <= var2, mean(var1) < 10)`

```
> v
Object of class 'validator' with 3 elements:
v1: var1 > 0
v2: var1 <= var2
v3: mean(var1) < 10
|
```



# Validation syntax

- Enhver funksjon som begynner med "is." .
- Binær sammenligning: `<`, `<=`, `=`, `!=`, `>=`, `>` and `%in%`.
- Logiske operatorer: `!`, `all()`, `any()`.
- Binære logiske operatorer: `&`, `&&`, `|`, `||`
- Logisk implikasjon e.g. `if (staff > 0) staff.costs > 0`.



# Spesialfunksjoner

- `is_complete` – kontroll om variabel er komplett
- `is_unique` – kontroll om det er dubletter
- `in_range` – kontroll som setter min og maks verdi (eller dato)
- `in_linear_sequence` – kontroll om det er en komplett sekvens av tall (2,4,6) eller datoer (mars, april, mai)



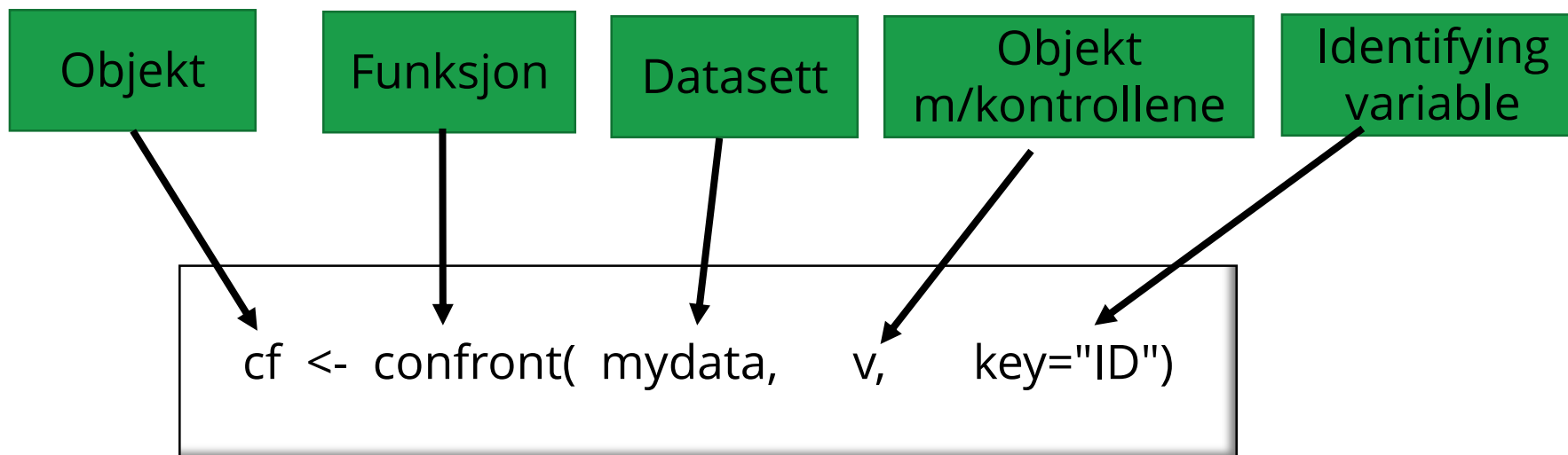
# Klassifiseringer og kodelister, Klass

- Bruk Klass til å holde orden kodelister og klassifiseringer
- Hent informasjonen fra Klass til å kontrollere kodelister ved pakken KlassR
  - `sn <- GetKlass(klass = 131, date = "2019-01-01")`
  - `komliste <- as.vector(sn[,c("code")])`
  - `regler <- validator( region %in% komliste)`



The screenshot shows the Statistics Norway website. At the top, there is a header with the logo and name 'Statistisk sentralbyrå Statistics Norway'. To the right of the logo are links for 'AA A', 'ENGLISH', and 'COOKIES'. Below the header is a navigation bar with three main categories: 'STATISTIKK', 'FORSKNING', and 'INNRAPPORTERING'. Below this, there is a breadcrumb trail: 'Forsiden > Metadata > Klassifikasjoner og kodelister'. The main heading is 'Klassifikasjoner og kodelister'. Below the heading is a paragraph explaining that classifications are 'official' code systems where categories should be mutually exclusive and exhaustive, while code lists are not 'official' and can be tailored for specific statistics. It mentions that users can include searches in code lists by checking a box and being careful of the number of results. Below this is a link 'Mer om SSBs system for klassifikasjoner og kodelister' and another link 'KLASS API guide. REST-API med formatene XML, JSON og CSV.' At the bottom, there is a search bar with the placeholder text 'Søk etter kodeverk' and a search button labeled 'Søk'.

# Kjøre kontrollene



```
> cf
Object of class 'validation'
Call:
confront(dat = mydata, x = v, key = "ID")

Confrontations: 3
With fails      : 2
Warnings        : 0
Errors          : 0
```





# Resultater av kjøring av kontroller på datasettet

- Mulig å hente ut informasjon med:
  - **summary:** Oppsummert resultat som returnerer som data.frame
  - **aggregate:** Aggregert validering - indikatorer
  - **values:** Få verdiene i en matrise, eller en liste over matriser hvis regler har annen dimensjonsstruktur for utdata
  - **errors:** Få feilmeldinger når kontrollene blir kjørt på data
  - **warnings:** Få advarsler når kontrollene blir kjørt på data
  - **sort :** Aggregere og sortere på forskjellige måter



# Metadata for kontrollene

- Følgende funksjoner kan bli brukt for å få eller sette **metadata**:
  - **origin** : Hvor var kontrollen laget?
  - **names** : Navnet til kontrollen
  - **created** : Når er kontrollen laget
  - **label** : Kort beskrivelse av kontrollen
  - **description**: Lang beskrivelse av kontrollen
  - **meta**: Sette eller gi generisk metadata



# Summary

summary(cf)

	name	items	passes	fails	nNA	error	warning	expression
1	v1	4	3	1	0	FALSE	FALSE	var1 > 0
2	v2	4	3	1	0	FALSE	FALSE	(var1 - var2) <= 1e-08
3	v3	1	1	0	0	FALSE	FALSE	mean(var1) < 10

- Hvor mange dataelementer som ble sjekket mot hver regel
- Hvor mange dataelementer som passerte, mislyktes eller resulterte i NA
- Hvorvidt kontrollen resulterte i en feil (kunne ikke utføres) eller ga en feil
- Uttrykket som faktisk ble evaluert for å utføre kontrollen



# Aggregate

aggregate(cf)

```
> aggregate(cf)
      npass nfail nNA rel.pass rel.fail rel.NA
v1         3     1   0    0.75    0.25     0
v2         3     1   0    0.75    0.25     0
v3         1     0   0    1.00    0.00     0
```

keys	If confront was called with key=
npass	Number of items passed
nfail	Number of items failing
nNA	Number of items resulting in NA
rel.pass	Relative number of items passed
rel.fail	Relative number of items failing
rel.NA	Relative number of items resulting in NA



# Values

```
values(cf)
```

```
> values(cf)
[[1]]
      v1      v2
1  TRUE   TRUE
2  TRUE  FALSE
3 FALSE   TRUE
4  TRUE   TRUE

[[2]]
      v3
[1,] TRUE
```

## #Dataset with indicators

```
ind<-as.data.frame(values(cf))
```

# add indicators to datasett

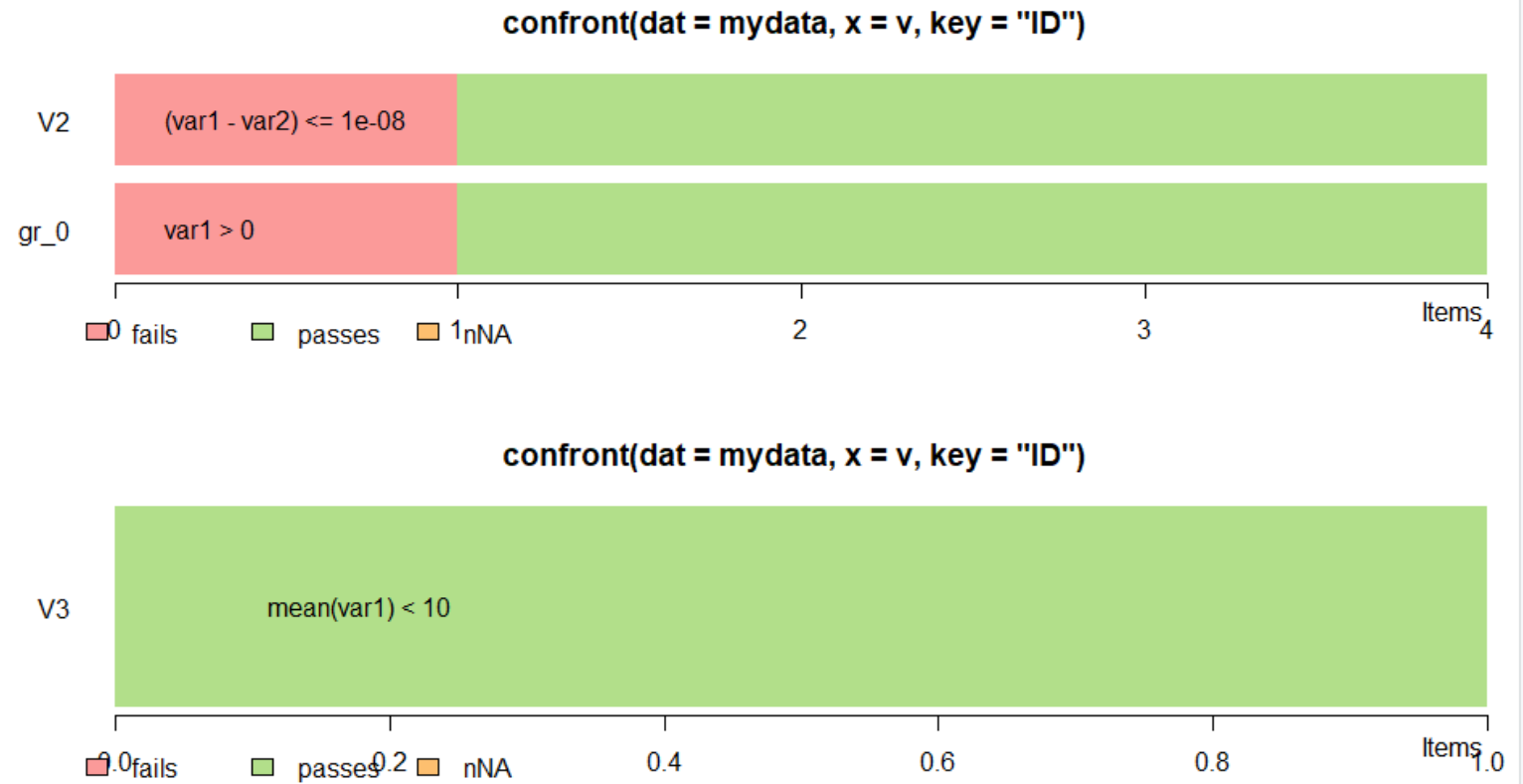
```
mydata2 <- mydata %>%
  mutate(greater_0 = pull(ind, V1),
         V2= pull(ind, V2))
```

	ID	var1	var2	V1	V2
1	1	2	9	TRUE	TRUE
2	2	9	1	TRUE	FALSE
3	3	-1	4	FALSE	TRUE
4	4	7	8	TRUE	TRUE



# Grafikk

plot(cf)



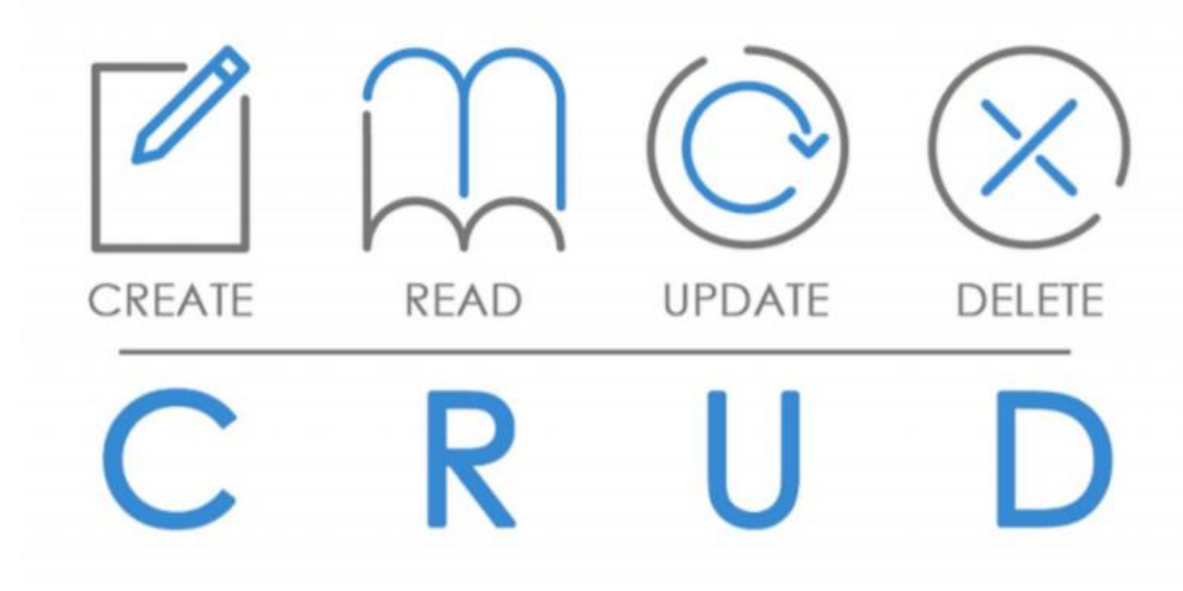
# Nederland lager dashboard

## Example: Validation Report Standard



# Reglene er data

- To sett med regler kan bli slått sammen: `rules<-rules1+rules2`
- Reglene bør versjoneres
- Reglene må vedlikeholdes
- Dokumentasjon!





# Eksempel i R på Jupyter



# Kursmaterialet

[https://github.com/statisticsnorway/R\\_kontrollfunksjoner](https://github.com/statisticsnorway/R_kontrollfunksjoner)

← → ↻ 🏠 [github.com/statisticsnorway/R\\_kontrollfunksjoner](https://github.com/statisticsnorway/R_kontrollfunksjoner)

Search or jump to... / Pull requests Issues Marketplace Explore

[statisticsnorway](#) / [R\\_kontrollfunksjoner](#) Public Edit Pins

[Code](#) [Issues](#) [Pull requests](#) [Actions](#) [Projects](#) [Wiki](#) [Security](#) [Insights](#) [Settings](#)

[main](#) 2 branches 0 tags [Go to file](#) [Add file](#) [Code](#)

**Your main branch isn't protected**  
Protect this branch from force pushing, deletion, or require status checks before merging

**aslaugfoss** Add files via upload

Eksempler.R	Add files via upload
Losninger.R	Add files via upload
Oppgaver.R	Add files via upload
Presentasjon.pdf	Add files via upload

**Clone** ⓘ

HTTPS SSH GitHub CLI

[https://github.com/statisticsnorway/R\\_kontrollfunksjoner](https://github.com/statisticsnorway/R_kontrollfunksjoner)

Use Git or checkout with SVN using the web URL.

Open with GitHub Desktop

Download ZIP

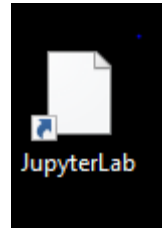
1 / months ago



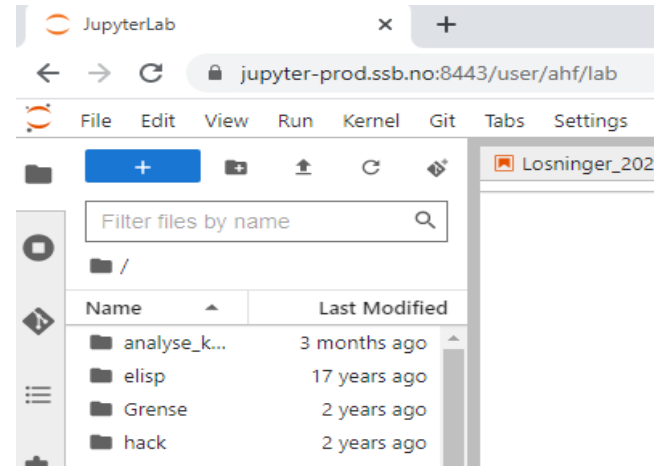
**Statistisk sentralbyrå**  
Statistics Norway

# Starte opp Jupyter i produksjonssonen

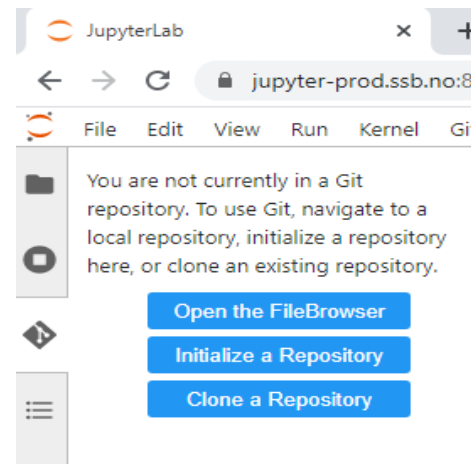
- Trykk på ikonet:



- Stå i «filutforsker»



- Trykk på Github-ikonet:



Clone a repo

Enter the Clone URI of the repository

<https://host.com/org/repo.git>

Cancel

CLONE

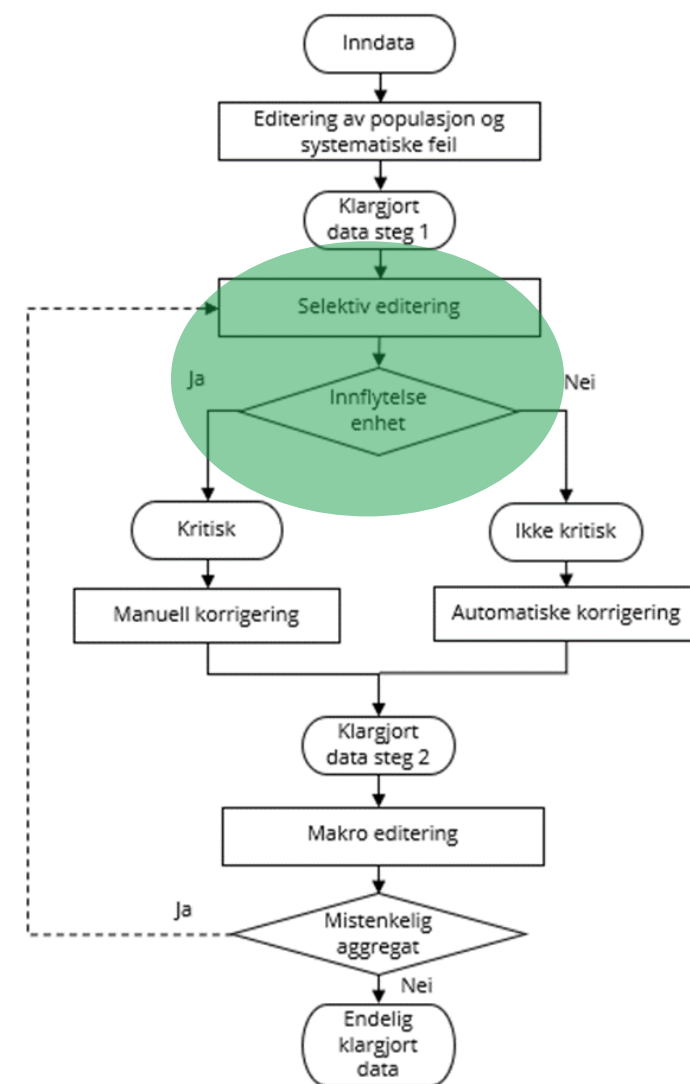
# Øvelse

- Oppgave 1-4 pluss ekstraoppgave
  - Kirkedata med variabelen døpte
  - «Oppgaver\_jupyter»
- Oppgavene
  - Github: <https://github.com/statisticsnorway/kurs-metode-validere>
  - Hvis det er utfordrende å kode; kjør «losninger», varier parametere og vurder resultatene
  - Ekstra kokeboken: <https://cran.r-project.org/web/packages/validate/vignettes/cookbook.html>
- Diskuter funksjonene og metodene med andre!



# Selektiv editering og prosessmodell

- **Selektiv editering** - er en generell tilnærming for å oppdage innflytelsesrike feil med hensyn til hovedresultatene.
- **Sannsynlig feil** – er observasjoner med verdier som ligger utenfor det som er forventet



# Kontrollmetoder

- Tusenfeil
- HB-metoden
- Kvartilmetoden
- Robust regresjon
- Enhetens andel av totalen
- Enhetens andel av endringstall
- Analyse av aggregat



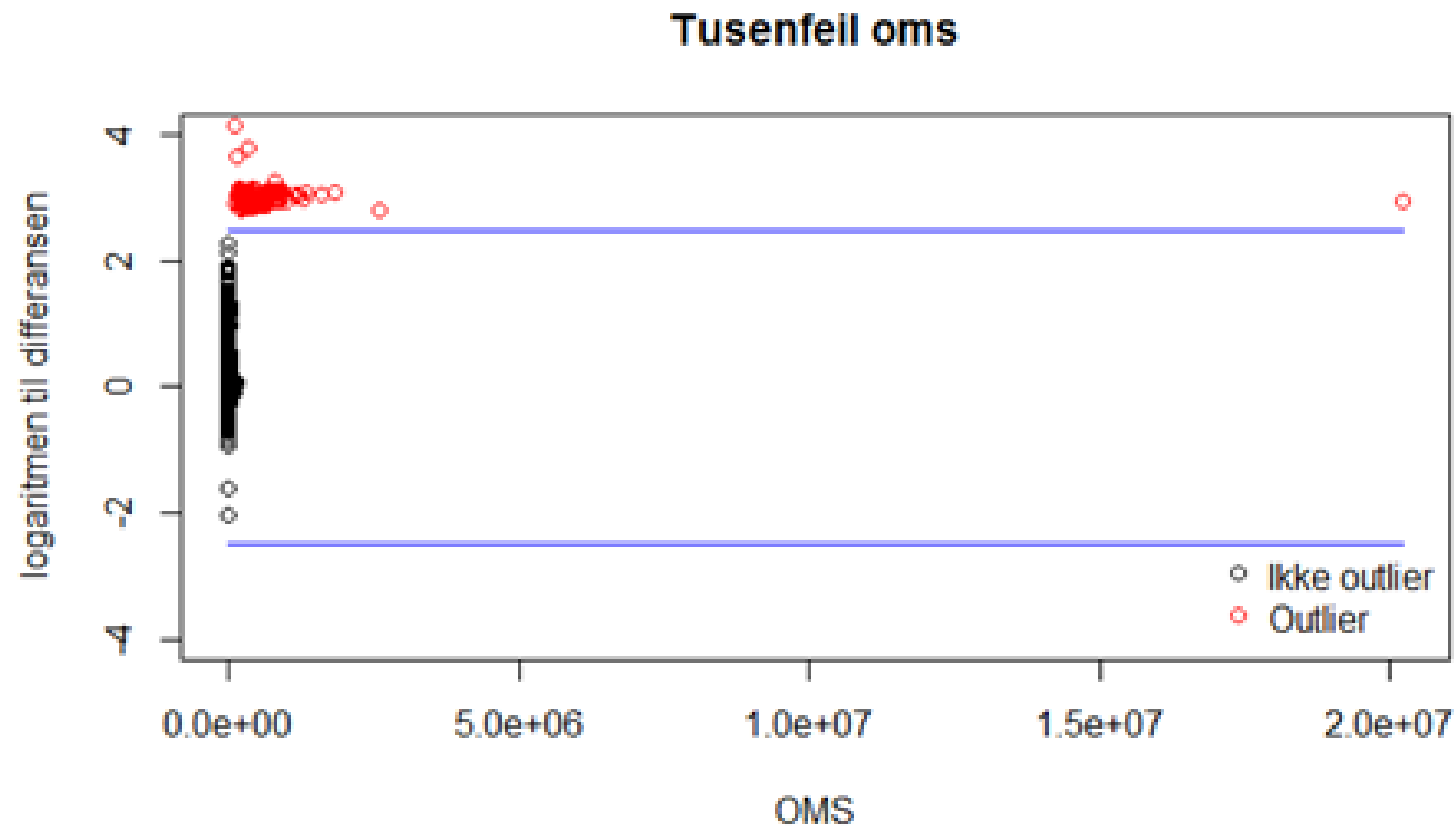
# Tusenfeil

- Målet for funksjonen: å oppdage at noen har oppgitt svaret i feil enhet
- Eksempel
  - Svar i kroner når det skal oppgis i 1000 kroner.
  - Eller i årsverk når de skal oppgi svaret i antall timer per uke.



# Tusenfeil

## DOI- detaljomsetningsindeksen





# Tusenfeil

- Kan kjøre som automatisk oppretting tidlig i prosessen
- Bruk funksjonen til analysere
- Automatisk oppretting bør alltid kontrolleres

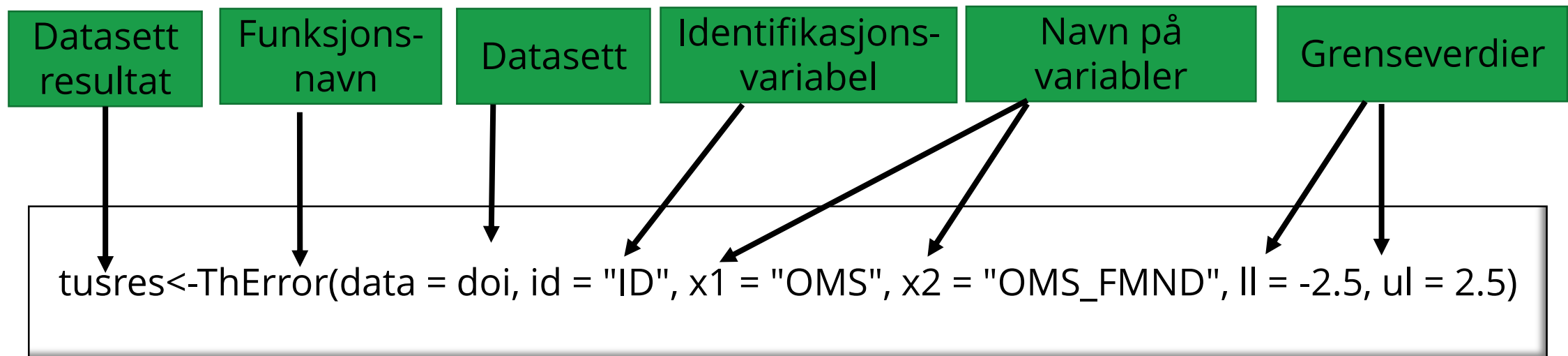


# Tusenfeil: Siffermetoden - logaritmen

- I siffermetoden teller vi forskjell i antall siffer mellom årets og forrige års verdi; dette gjør vi ved hjelp av den matematiske funksjonen logaritmen.
- Hvis det er en forskjell på 3 siffer er det en tusenfeil, 6 siffer millionfeil osv.
- Metoden fungerer bare for verdier som er større enn null og ikke missing



# Tusenfeil-funksjon



## Parametre:

- data Input datasett med klasse data.frame.
- id Navn på identifikasjonsvariabel.
- x1 Navn på variabel i periode t.
- x2 Navn på variabel i periode t-1.
- ll Nedre grense for  $\log_{10}(x1 / x2) = \log_{10}(x1) - \log_{10}(x2)$ . Standard -2,5
- ul Øvre grense for  $\log_{10}(x1 / x2) = \log_{10}(x1) - \log_{10}(x2)$ . Standard +2,5



# Output

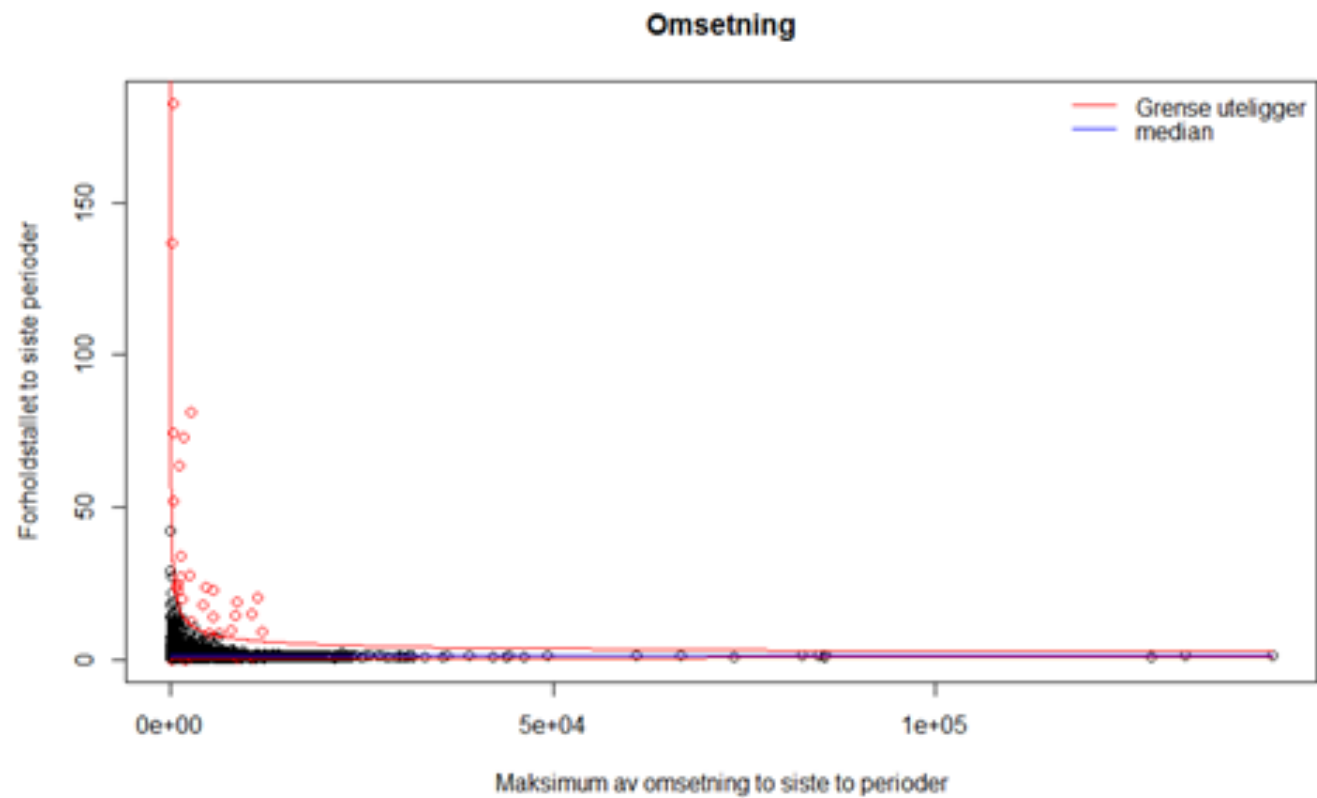
**Resultat:**

- id identifikasjonsvariabelen.
- x1-variabel
- x2-variabelen
- outlier En binær (1/0) variabel som indikerer om vi mistenker en 1000 feil eller ikke
- diffLog10 Forskjellen  $\log_{10}(x1) - \log_{10}(x2)$
- lowerLimit Inngangsparameteren ll
- upperLimit Inngangsparameteren ul



# HB- metoden

DOI- detaljomsetningsindeksen



# Hidiroglou-Berthelot (HB)

- Formålet med funksjonen er å finne avvikende verdier i forhold til forrige periode
- Egenskaper til funksjonen:
  - Funksjonen tar hensyn til nivåendring mellom perioder
  - Funksjonen tar hensyn til at små verdier har større variasjon enn store verdier
  - Funksjonen feiler hvis median og en av kvartilene er identiske.

Hidiroglou, M.A. and Berthelot, J.-M. (1986) 'Statistical editing and Imputation for Periodic Business Surveys'. Survey Methodology, Vol 12, pp. 73-83.

# Formlene

Endringskvoten  $R_i = \frac{X_i(t)}{X_i(t-1)}$

Symmetritransformasjonen  $S_i = \begin{cases} 1 - R_{median} / R_i, & 0 < R_i < R_{median} \\ R_i / R_{median} - 1, & R_i \geq R_{median} \end{cases}$

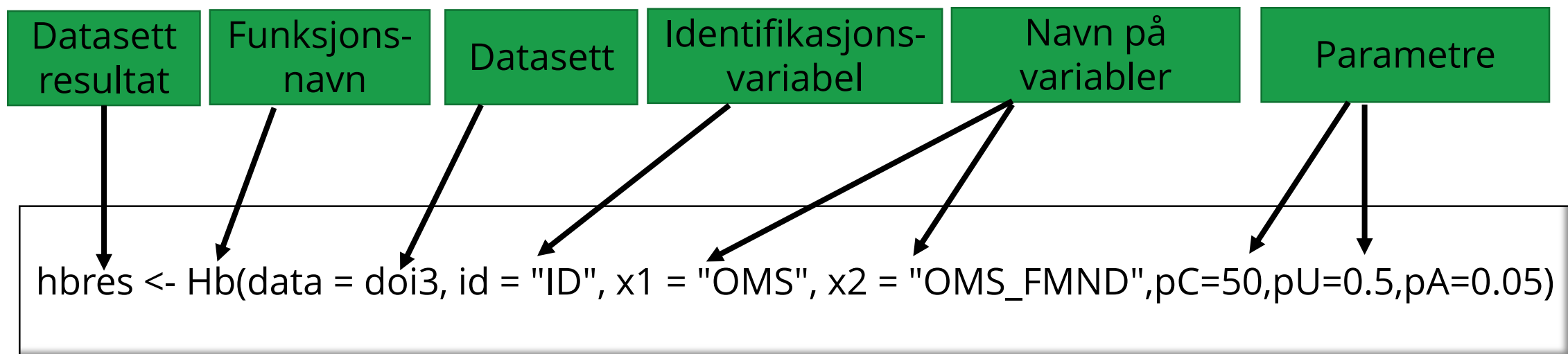
Størrelsestransformasjon  $E_i = S_i * (MAX(X_i(t-1), X_i(t)))^U, \quad 0 \leq U \leq 1$

Akseptgrenser 
$$\begin{aligned} D_{Q1} &= MAX(E_{median} - E_{Q1}, |A * E_{median}|) \\ D_{Q3} &= MAX(E_{Q3} - E_{median}, |A * E_{median}|) \end{aligned}$$

$$\{ E_{median} - C * D_{Q1}, \quad E_{median} + C * D_{Q3} \}$$



# HB-funksjon

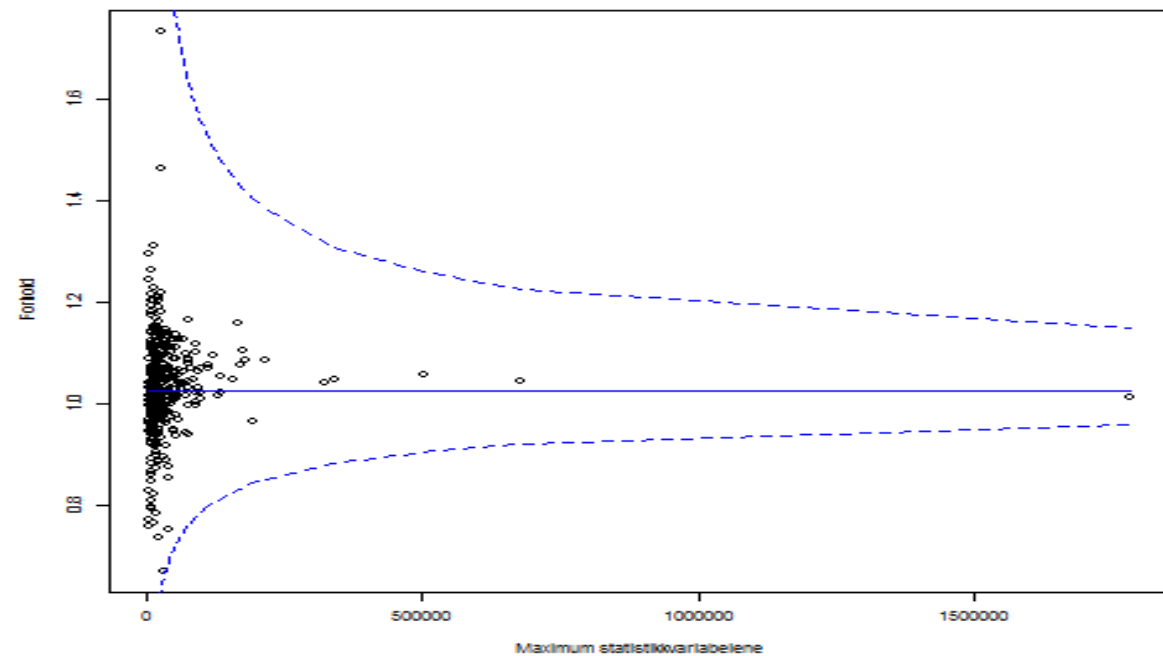


## Input parametre er:

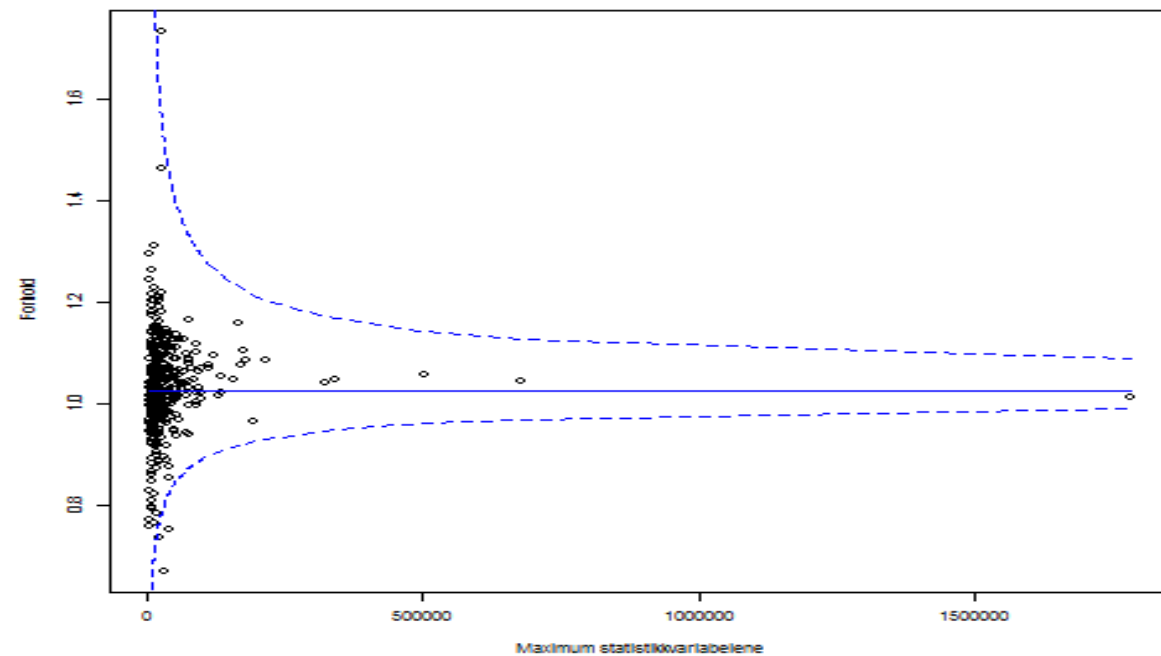
- data Input of Hb er et datasett av type dataframe.
- id Navn på en identifikasjonsvariabel.
- x1 Navn på variabel i periode t.
- x2 Navn på variabel i periode t-1.
- pU Parameter som justerer for forskjellige nivåer av variablene. Default verdi 0,5.  $pU < 1$
- pA Parameter som justerer for små forskjeller mellom median og 1. eller 3. kvartil. Standardverdi 0,05.
- pC Parameter som kontrollerer lengden på konfidensintervallet. Default verdi 20.



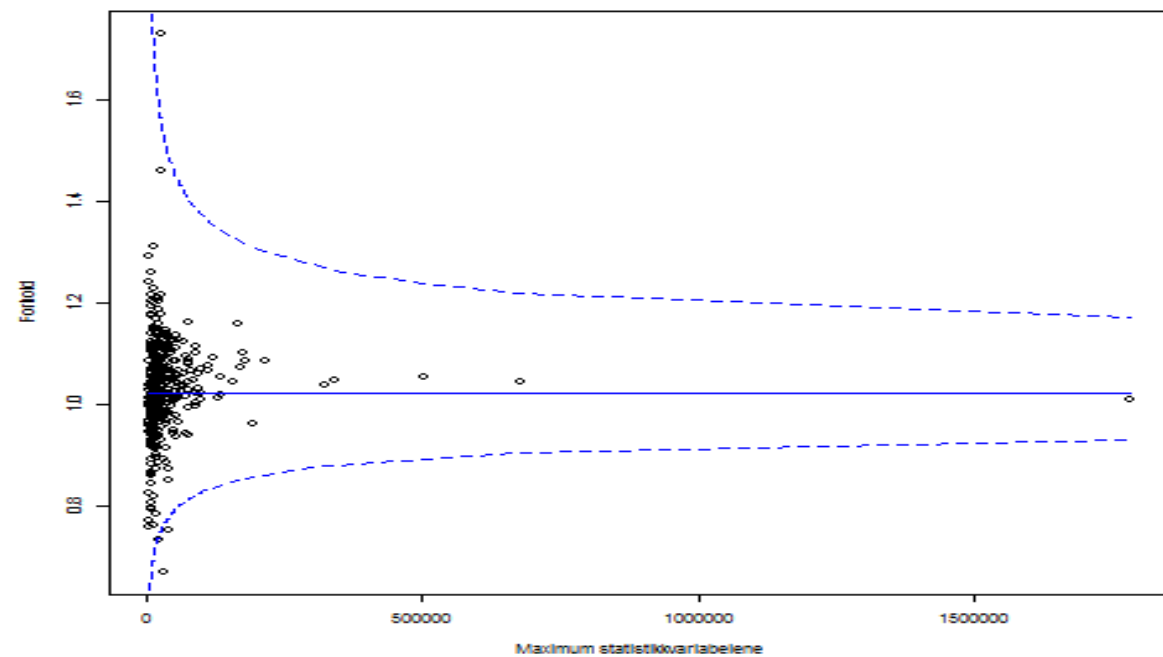
Brutto driftsutgifter  $u=0.5$ ,  $c=20$



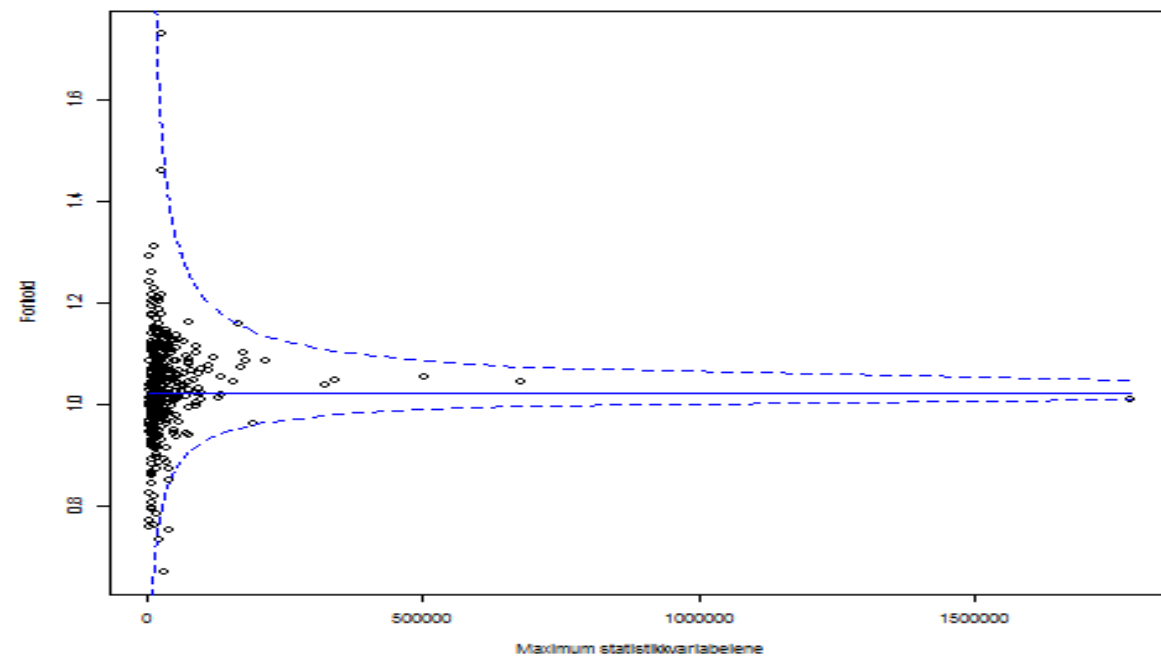
Brutto driftsutgifter  $u=0.5$ ,  $c=10$



Brutto driftsutgifter  $u=0.3$ ,  $c=10$

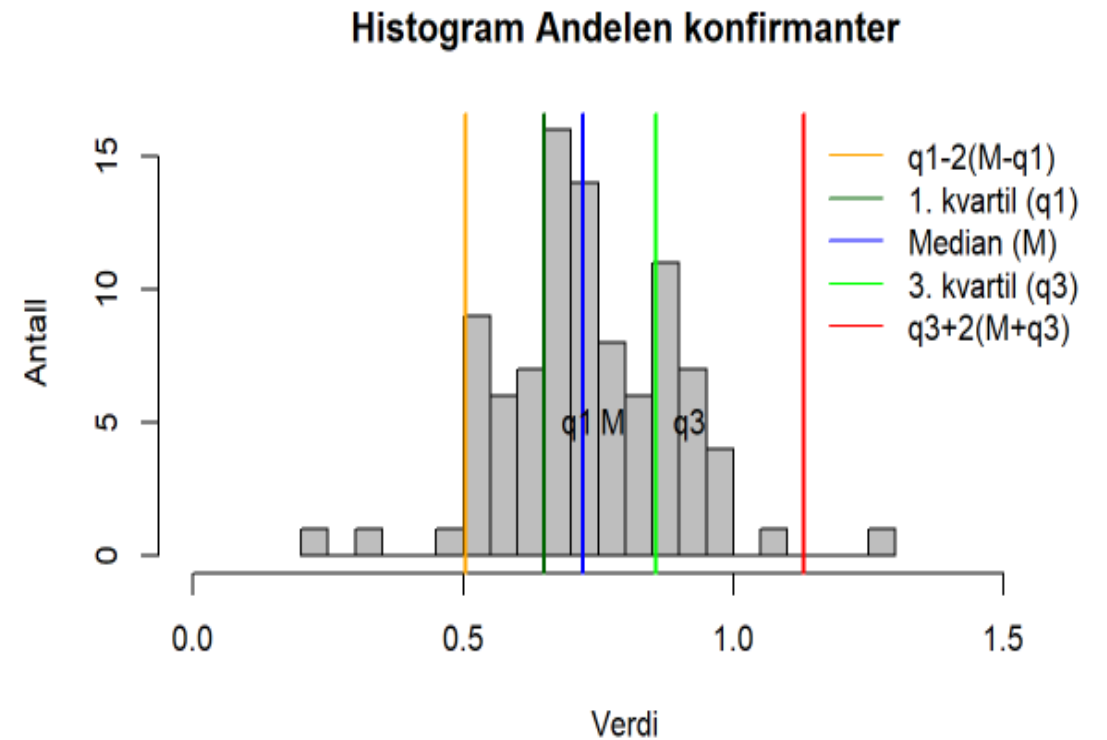


Brutto driftsutgifter  $u=0.7$ ,  $c=10$



# Kvartilmetode

- gjennomsnitt  $\bar{X} \pm k * std(X)$  følsomt for ekstremverdier
- $q_{0.5}$ ,  $q_{0.25}$  og  $q_{0.75}$  står for henholdsvis median, 1. og 3. kvartil
- Aksepterte verdier:  
 $(q_{0.25} - k_1(q_{0.5} - q_{0.25}), q_{0.75} + k_2(q_{0.75} - q_{0.5}))$

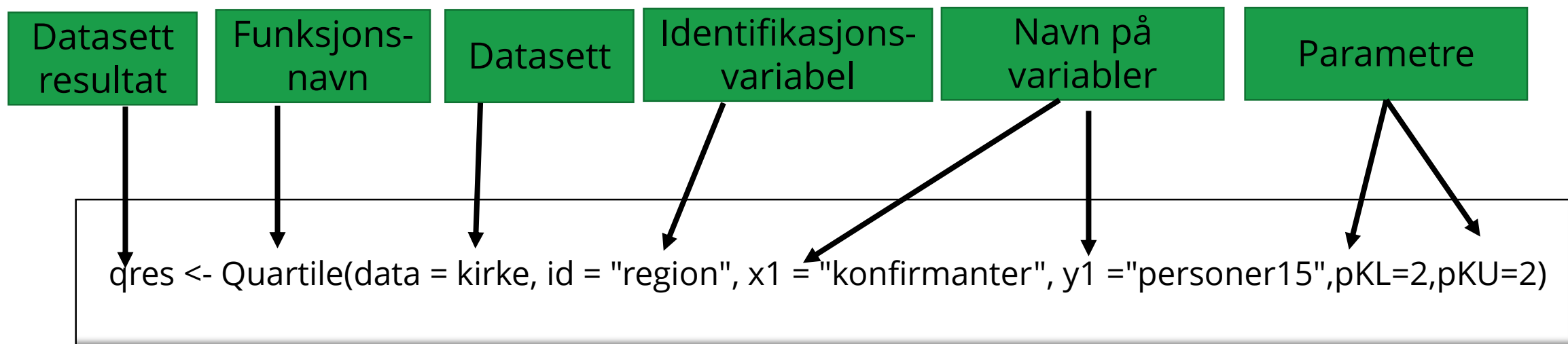


# Kvartilfunksjonen

- Formålet med funksjonen er å kontrollere en variabel mot en annen variabel av god kvalitet.
- Vær oppmerksom på:
  - Det er mulig å kjøre metoden innen grupper (stratum), men da med alle stratum med lik grenseverdi.
  - Det er mulig å sende inn informasjon om forrige periode for å bruke dette til vurdering av uteliggerne
  - I output fra metoden er enheter med verdier som mangler eller er null eller mindre ekskludert



# Kvartilfunksjon



## Input parametre er:

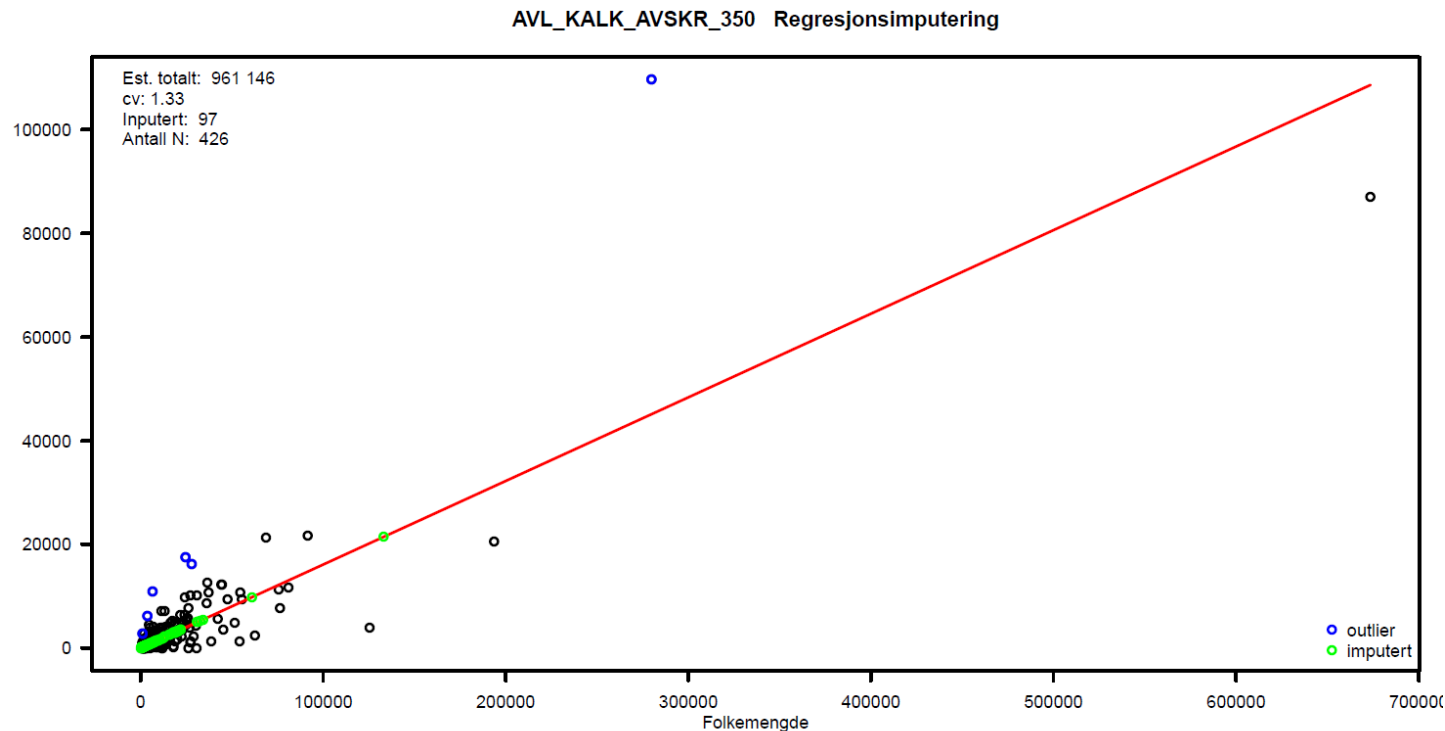
- data Input til Quartile er et datasett med klassesdata.frame.
- id Navn på identifikasjonsvariabel.
- x1 Navn på x-variabel i periode t.
- y1 Navn på y-variabel i periode t.
- x2 Navn på x-variabel i periode t-1. Valgfri
- y2 Navn på y-variabel i periode t-1. Valgfri
- strataName Navn på stratifiseringsvariabelen. Valgfri
- pKL Parameter for nedre grense.
- pKU-parameter for øvre grense.

### Output fra funksjonen:

- id identifikasjonsvariabelen
- x1-variabel
- y1-variabelen
- x2-variabelen - forrige periode - valgfri
- y2-variabelen - forrige periode - valgfri
- ratio Forholdet mellom x1 og y1; ratio2 Forholdet mellom x2 og y2, ratioAll Forholdet mellom summen av x1 og summen av y1 samlet over det hele datasett; ratioAll2 Forholdet mellom summen av x2 og summen av y2 samlet over det hele datasett; ratioStr Forholdet mellom summen av x1 og summen av y1 samlet over stratum; ratioStr2 Forholdet mellom summen av x2 og summen av y2 samlet over stratum
- lowerLimit Den nedre grensen for forholdet
- upperLimit Den øvre grensen for forholdet
- outlier En binær variabel som indikerer om observasjonen er utenfor grensene  $[q1 - p_{KL} * (M - q1), q3 + p_{KU} * (q3 - M)]$ , hvor M er henholdsvis median og q1 og q3, henholdsvis 1. og 3. kvartil.
- strata Strata navn eller nummer
- ranking Rangeringsgraden. For plotting

# Robust regresjon

- Metoden tar utgangspunkt i at det kan bli laget en lineær modell av sammenhengen mellom variablene og hvordan variasjonen er.
- Kjører iterativt for å finne en modell som ikke er påvirket av outliere.



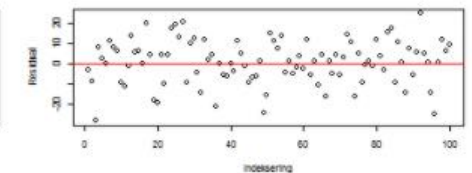
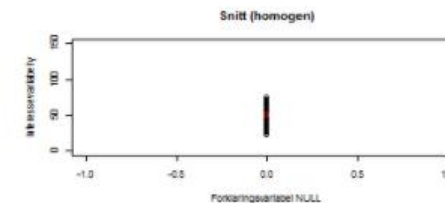
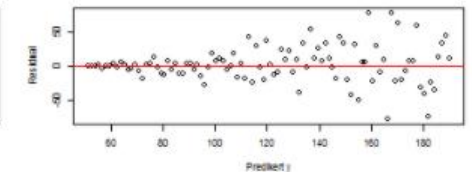
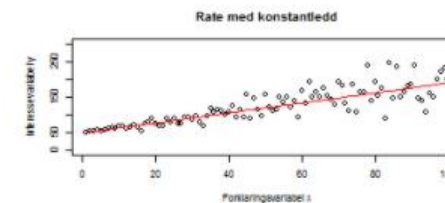
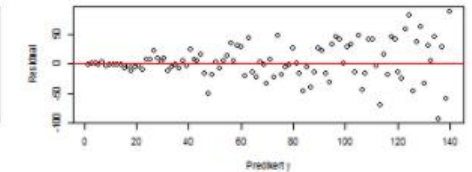
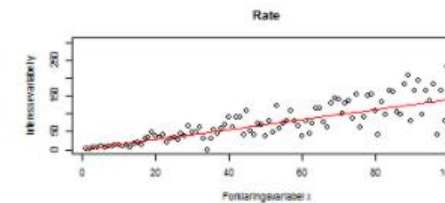
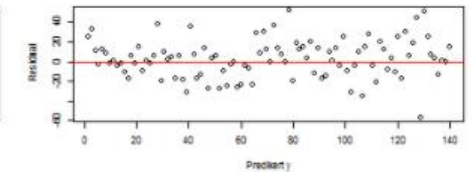
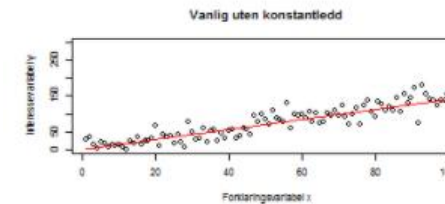
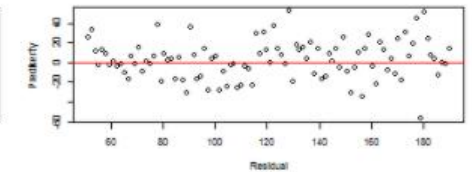
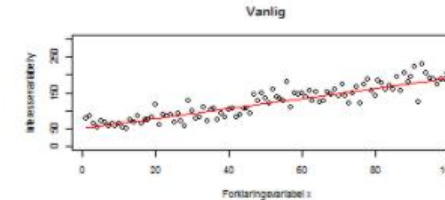
# Robust regresjon

- Vi har en multivariat versjon liggende som kan brukes
- Kan kjøres innen grupper
  - homogene grupper som ligner på hverandre
  - Må være nok observasjoner innen hver gruppe
  - Grupper kalles ofte strata av statistikere



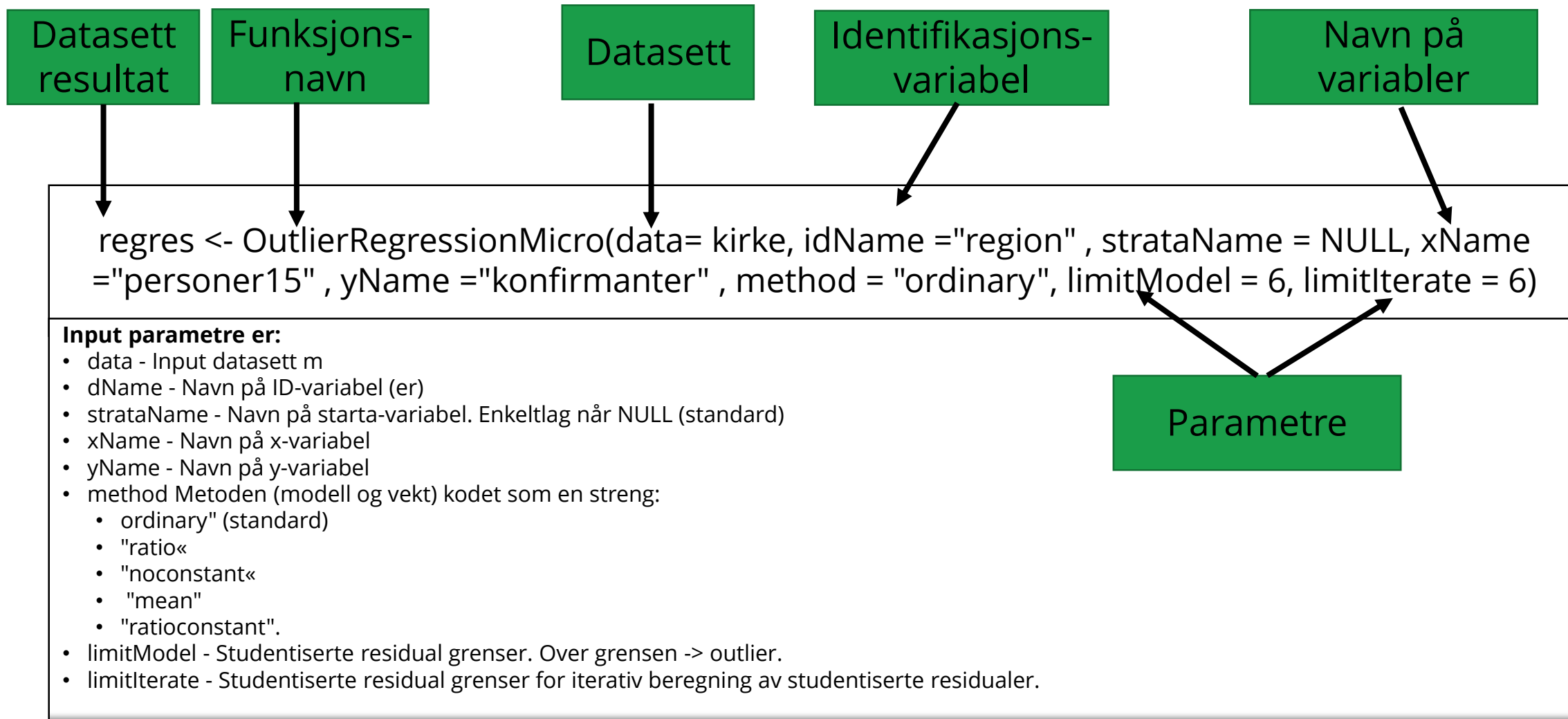
# Modeller

Navn	Modell	Varians
Vanlig med konstantledd	Rett linje	Lik for alle
Vanlig	Rett linje som går gjennom origo (punktet 0,0)	Lik for alle
Rate	Rett linje som går gjennom origo (punktet 0,0)	Økende med forklaringsvariabelen x
Rate med konstantledd	Rett linje	Økende med forklaringsvariabelen x
Snitt	Gjennomsnitt	Lik for alle





# Regresjonsfunksjon



- id - id fra input
- x - Variabelen input x
- y - Variabelen input y
- strata - Inndata-grupperingsvariabelen (kan være NULL)
- outlier - Dummy-variabel: outlier (1) eller ikke (0).
- kategori123 - De gruppene: representativ (1), riktig, men ikke representativ (2), galt (3).
- yHat - Tilpassede verdier
- rStud - De studentiserte residulaer fra siste iterasjon
- dffits - Diagonale elementer i hatmatrise fra siste iterasjon
- leaveOutResid Restmodellen utenfra fra siste iterasjon
- limLo - limitModel
- limUp - limitModel



# Rstud og dffits?

- Studenifiserte residualer - Avstand fra observasjon til regresjonslinjen ( $|rstudent| > 3$ )
- Dffits – en observasjons innflytelse på regresjonslinjen ( $|dffits| > 2[(m+1)/n]^{1/2}$ )
  - n antall observasjoner og m antall estimerte parametre

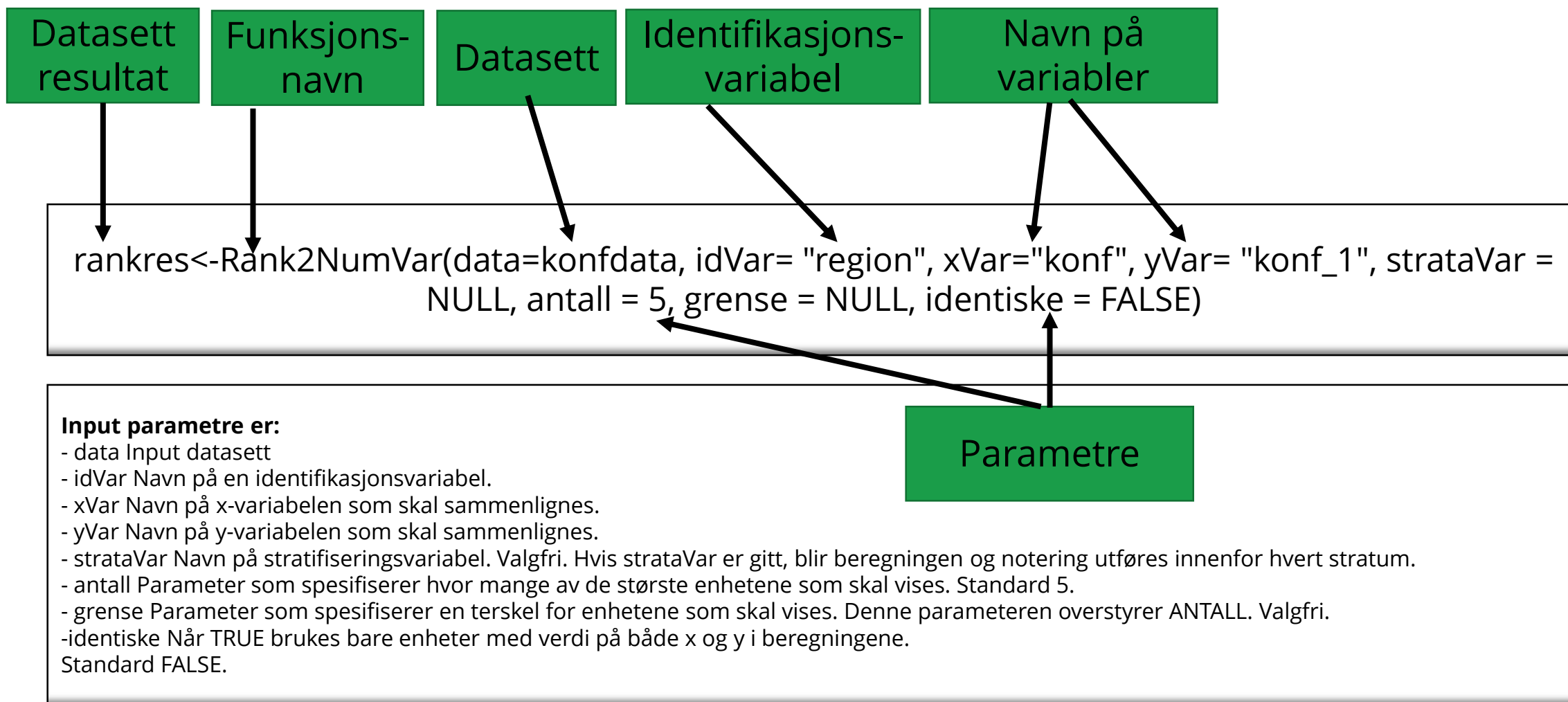


# Innflytelse på totalen

- Målet med metoden er å rangere de største verdiene og vise hvor stor andel de utgjør av totalen og eventuelt tilhørende stratum.
- Det kan også være nyttig å se på rangeringen forrige periode og hvor mye verdien utgjorde da.



# Rangering av variabler



# Output fra funksjonen

- - id identifikasjonsvariabelen
- - x Variabelen input x
- - y Variabelen input y
- - strata Inputstrata-variabelen hvis strataVar er gitt, "1" ellers
- - forh Forholdet mellom x og y:  $y / x$
- - xRank Rangeringen av x
- - yRank Rangeringen til y
- - xProsAvSumx x i prosent av total / stratum totalt for x
- - yProsAvSumy y i prosent av total / stratum totalt for y

id	x	y	strata	forh	xRank	yRank	
<chr>	<int>	<int>	<chr>	<dbl>	<int>	<int>	>
1201	1668	1740	1	1.0431655	1	1	
0301	1645	1722	1	1.0468085	2	2	
5001	1046	1085	1	1.0372849	3	3	
1103	930	793	1	0.8526882	4	4	
0219	743	758	1	1.0201884	5	5	

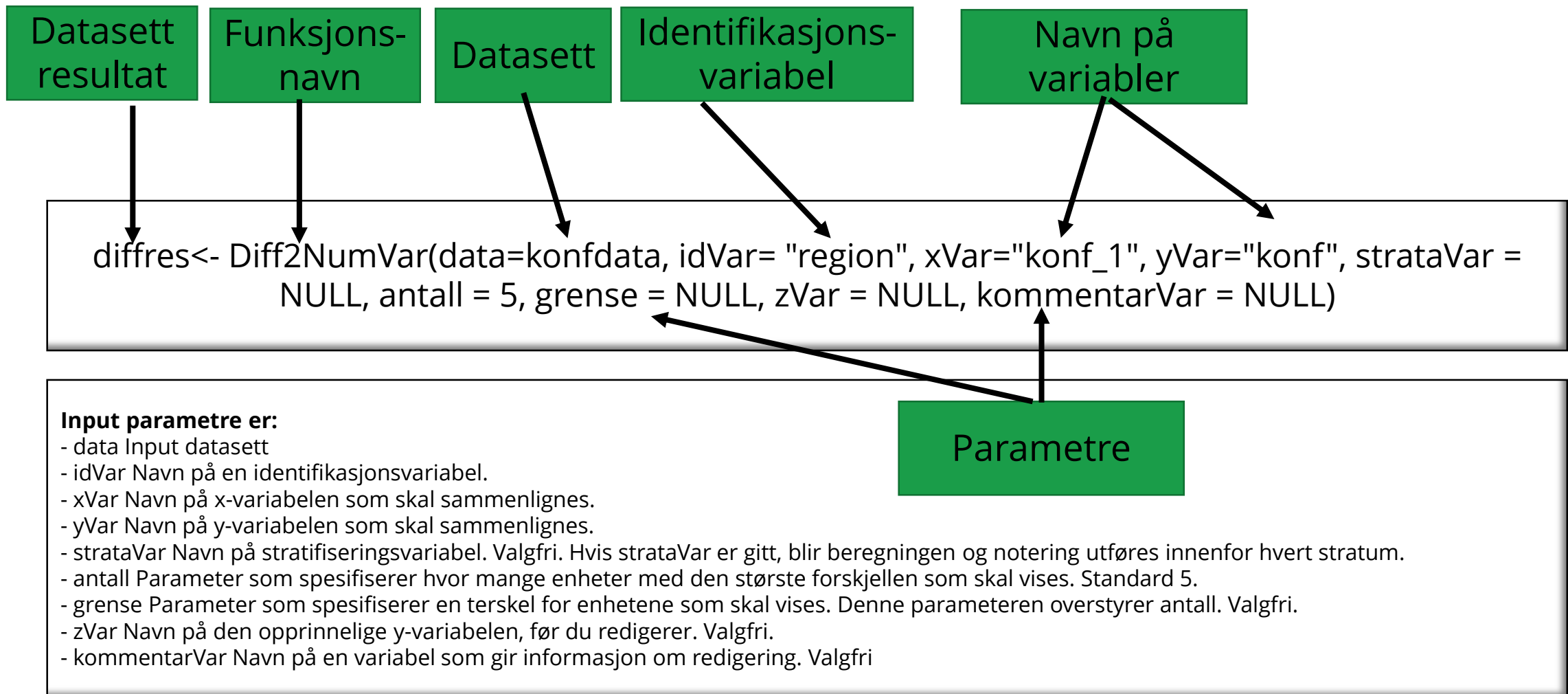


# Innflytelse på endringen

- Målet med funksjonen er å vise hvilke verdier som har størst innflytelse på endringstallet.
- Metoden gir også støtte til å forstå betydning av denne verdiendringen i forhold til både total og innen gruppe (stratum)



# Differanse numeriske variabler





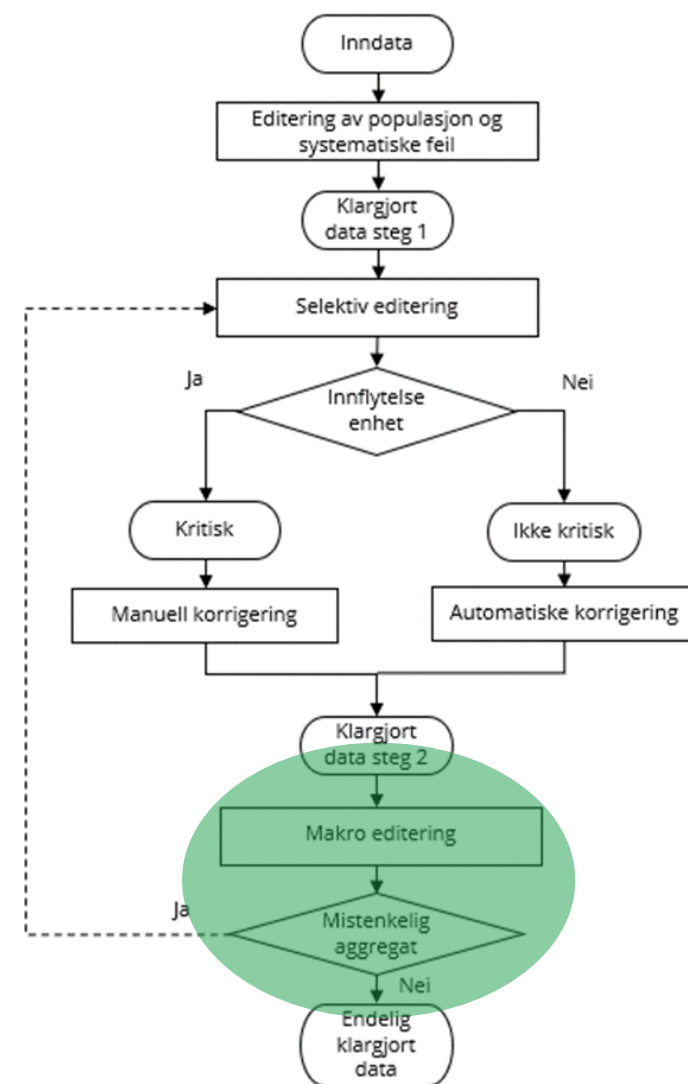
# Output fra funksjonen

id <chr>	x <int>	y <int>	Diff <int>
5037	136	336	200
1103	793	930	137
1523	25	136	111
0301	1722	1645	-77
5005	109	182	73

- strata Stratum (hvis strataVar er gitt, "1" ellers)
- id Inngangsideifikasjonsvariabelen
- x Variabelen input x
- y Variabelen input y
- Forh Forholdet mellom x og y:  $y / x$
- Diff Forskjellen mellom x og y:  $y - x$
- AbsDiff Den absolutte forskjellen:  $| \text{Diff} |$
- DiffProsAvx Forskjellen i prosent av x:  $(\text{Diff} / x) * 100$
- DiffProsAvSumx Forskjellen i prosent av stratum totalt for x:  $(\text{Diff} / \text{stratum } x) * 100$
- DiffProsAvTotx Forskjellen i prosent av totalen for x:  $(\text{Diff} / \text{total } x) * 100$
- SumDiffProsAvSumx Stratumforskjellen i prosent av stratum totalt for x:  $((\text{stratum } y - \text{stratum } x) / \text{stratum } x) * 100$
- SumDiffProsAvTotx Stratumforskjellen i prosent av totalen for x:  $((\text{stratum } y - \text{stratum } x) / \text{totalt } x) * 100$
- z Variabelen input z
- Endring Forskjellen mellom z og y:  $y - z$
- KommentarVar Input kommentar-variabelen

# Makroeditering og prosessmodell

- **Makroeditering** - er å analysere aggregater eller beregninger på hele populasjonen for å identifisere deler av datasett som kan inneholde potensielt innflytelsesrike feil.



# Analyse av aggregat

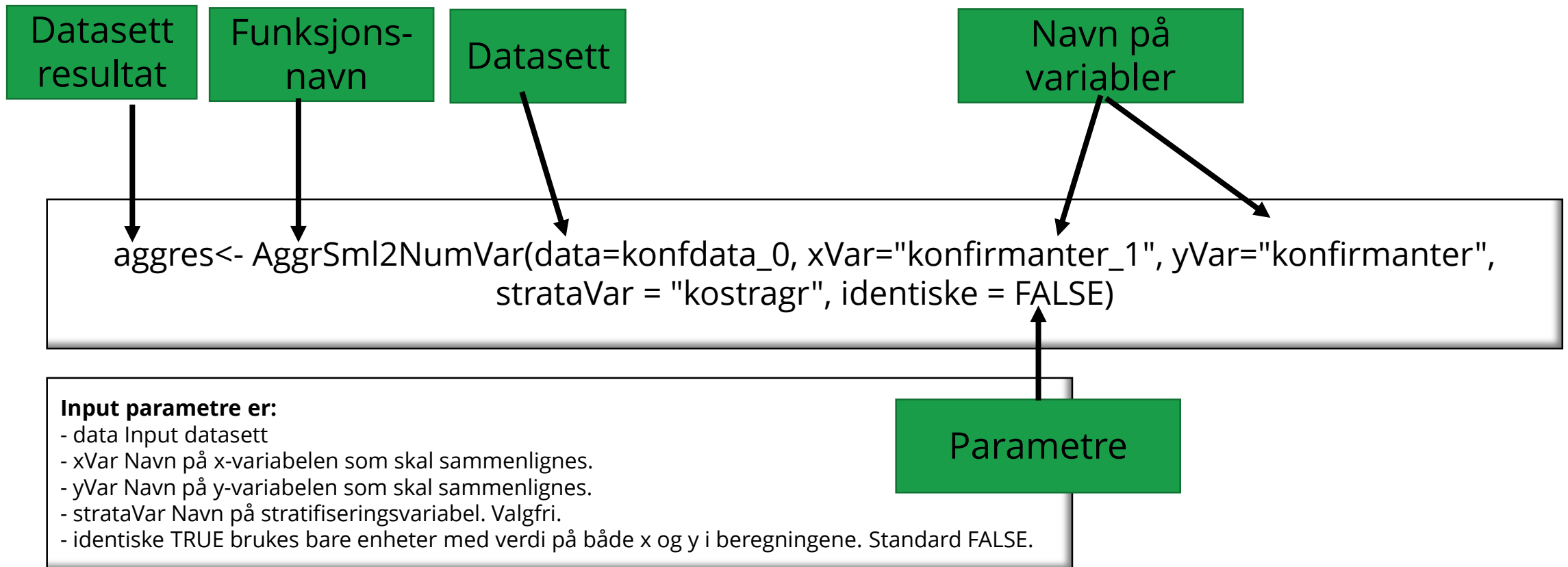
strata <chr>	Antx <int>	Anty <int>	Sumx <int>	Sumy <int>	SumxProsAvTotx <dbl>	SumyProsAvToty <dbl>	Diff <int>	AbsDiff <int>	DiffProsAv Sumx <dbl>
EKG01	13	13	406	403	1.1658291	1.1700491	-3	3	-0.7389163
EKG02	57	57	1695	1839	4.8671931	5.3392562	144	144	8.4955752
EKG03	34	34	957	901	2.7480258	2.6159162	-56	56	-5.8516196
EKG04	11	11	153	144	0.4393396	0.4180820	-9	9	-5.8823529
EKG05	31	31	478	461	1.3725772	1.3384432	-17	17	-3.5564854
EKG06	52	52	658	611	1.8894472	1.7739454	-47	47	-7.1428571
EKG07	30	30	2982	2854	8.5628141	8.2861539	-128	128	-4.2924212
EKG08	15	15	1427	1374	4.0976310	3.9891995	-53	53	-3.7140855
EKG10	24	24	1819	1770	5.2232592	5.1389252	-49	49	-2.6937878
EKG11	62	62	4710	4562	13.5247667	13.2450716	-148	148	-3.1422505

1-10 of 15 rows | 1-10 of 13 columns

Previous 1 2 Next



# Analyse av aggregat



# Output fra funksjonen

- strata Stratum (hvis strataVar er gitt, "1" ellers)
- Antx Antall enheter med x som ikke mangler brukt i samlingen
- Anty Antall enheter med y som ikke mangler brukt i aggregeringen
- Sumx Summen av x i stratum
- Sumy Summen av y i stratum
- SumxProsAvTotx Stratum totalt for x i prosent av befolkningen totalt for x:  $(\text{Sumx} / \text{Totx}) * 100$
- SumyProsAvToty Stratomet totalt for y i prosent av befolkningen totalt for y:  $(\text{Sumy} / \text{Toty}) * 100$
- Diff Forskjellen mellom stratum totalt av x og y:  $\text{Sumy} - \text{Sumx}$
- AbsDiff Den absolutte forskjellen:  $|\text{Diff}|$
- DiffProsAvSumx Forskjellen i prosent av stratum totalt for x:  $(\text{Diff} / \text{Sumx}) * 100$
- AbsDiffProsAvSumx Den absolutte verdien av DiffProsAvSumx:  $|\text{DiffProsAvSumx}|$
- DiffProsAvTotx Forskjellen i prosent av befolkningen totalt for x:  $(\text{Diff} / \text{Totx}) * 100$
- AbsDiffProsAvTotx Den absolutte verdien av DiffProsAvTotx:  $|\text{DiffProsAvTotx}|$



# Dashboard makrokontroller

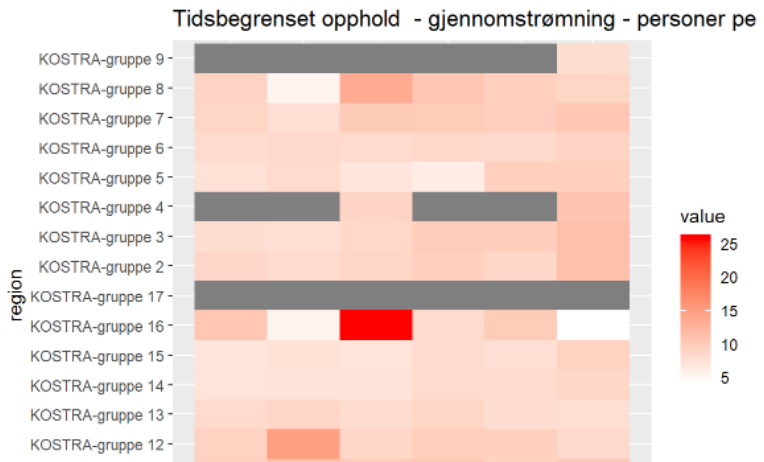
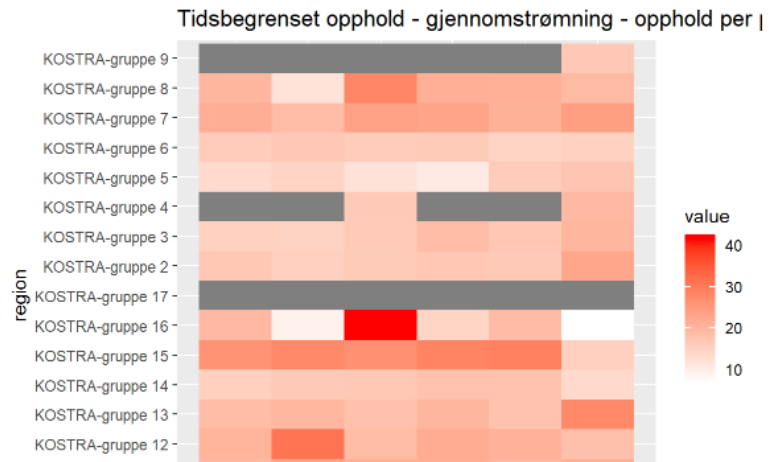
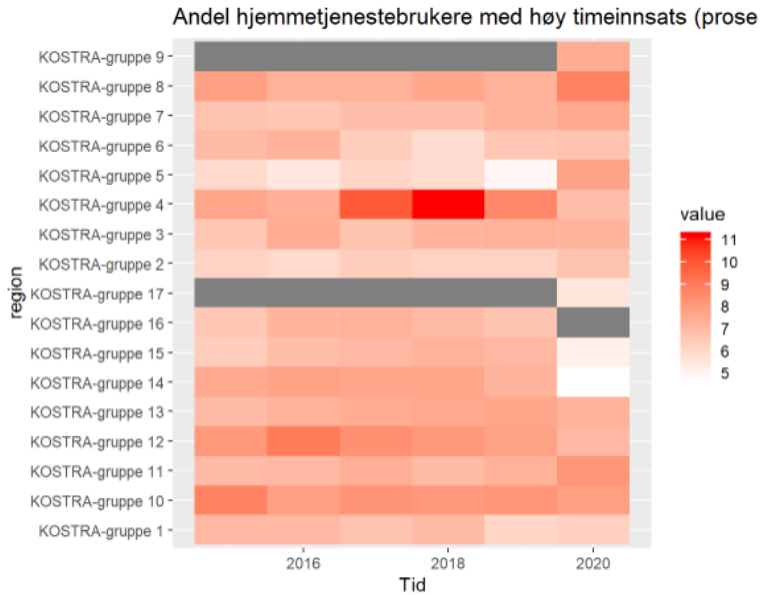
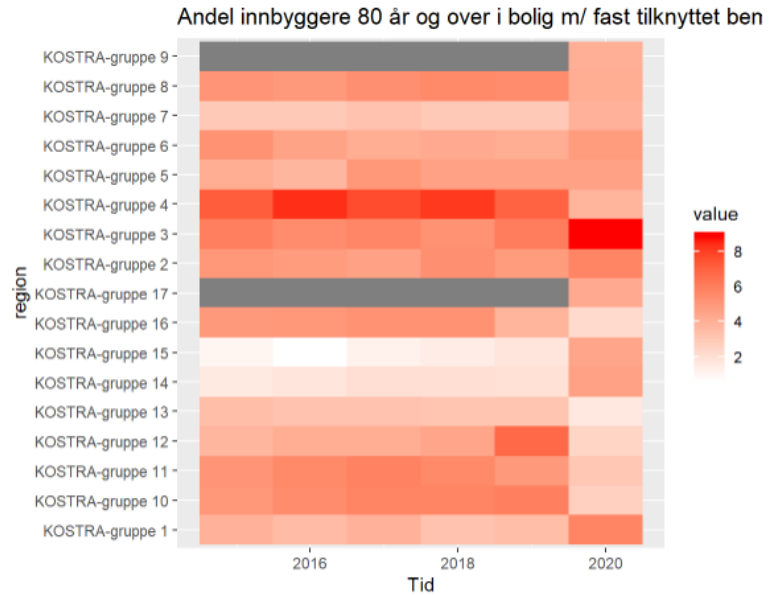
## 12293: Omsorgstjenester - supplerende nøkkeltall

Landet	Fylke per variabel	Kostragruppe per variabel	Fylke samlet siste år	Robust multivariat oppdagelse av uteliggere				
statistikkvariabel		2015	2016	2017	2018	2019	2020	TrendSparkline
All		/	/	/	/	/	/	All
1	Andel beboere 80 år og over i bolig m/ fast tilknyttet bemanning hele døgnet (prosent)	34.9	33.8	34.9	34.5	34.2	33.6	
2	Andel beboere i institusjon av antall plasser (belegg) (prosent)	98.8	98.8	98.1	0	0	0	
3	Andel hjemmetjenestebrukere med høy timeinnsats (prosent)	7.1	7.2	7.3	7.4	7.4	7.4	
4	Andel innbyggere 80 år og over i bolig m/ fast tilknyttet bemanning hele døgnet (prosent)	3.6	3.6	3.7	3.7	3.6	3.5	
5	Andel langtidsbeboere 31.12 vurdert av lege siste år (prosent)	49.9	54.3	54.5	65.5	66	65	
6	Brutto driftsutgifter, institusjon, per plass (kr)	1130391	1186616	1233936	1329501	1384651	1557795	
7	Fysioterapitimer pr. uke pr. beboer i sykehjem (antall)	0.41	0.43	0.42	0.42	0.43	0.45	
8	Korrigerte brutto driftsutgifter, institusjon, pr. kommunal plass (kr)	1077219	1112051	1148619	1222522	1282885	1406826	
9	Legetimer per uke per beboer i sykehjem (timer)	0.53	0.55	0.55	0.56	0.55	0.58	
10	Tidsbegrenset opphold - gjennomsnittlig antall døgn per opphold (antall)	19	19	19.3	18.4	19.5	17.8	
11	Tidsbegrenset opphold - gjennomstrømning - opphold per	19.2	19.2	19.6	20.6	19.4	20	



## 12293: Omsorgstjenester - supplerende nøkkeltall

Landet Fylke per variabel Kostragruppe per variabel Fylke samlet siste år Robust multivariat oppdagelse av uteliggere



# Eksempler i R i Jupyter





# Øvelse

- Oppgave 5-12
  - Bruk av variabelen døpte
  - Kirke\_data\_0 med fjernet null og missing for konfirmanter
  - Bruk av pakken Kostra og grafikkpakken plotly
- Hvis det er utfordrerne å kode, kjør «losninger», varier parametre og vurder resultatene
- Diskuter funksjonene og metodene med andre!



# Mer avanserte metoder

- Poengfunksjon DIFF fra Statistics Canada multivariat differanse
- Selekt-funksjon fra Svenske statistikkbyrået.

(<https://wiki.ssb.no/display/s880/Selektiv+editering+-+frokost+metodeseminar>)

- Multivariable metoder
- Maskinlæring



# Poengfunksjon Diff - multivariat

- Poengfunksjonen gir oss en prioritering på hvilke enheter som påvirker endringstallene mest.
- Hvilke variabler som skal inn i metoden må statistikkansvarlig selv velge.

$$score(i) = \sum_v \frac{|y_{i,v,t} - y_{i,v,t-1}|}{\sum_i y_{i,v,t-1}} \frac{1}{Antall\ v}$$

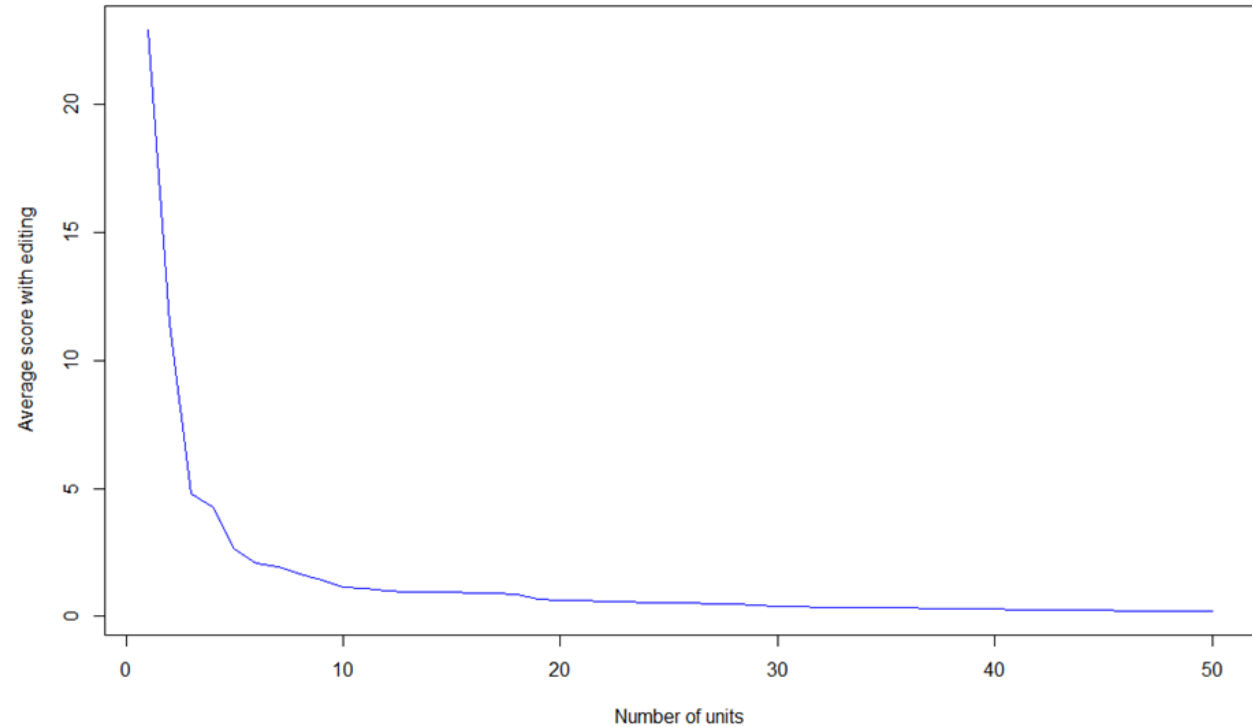
- Der i – er enhet, v- er variabel og t er periode.



# Offentlig eide foretak

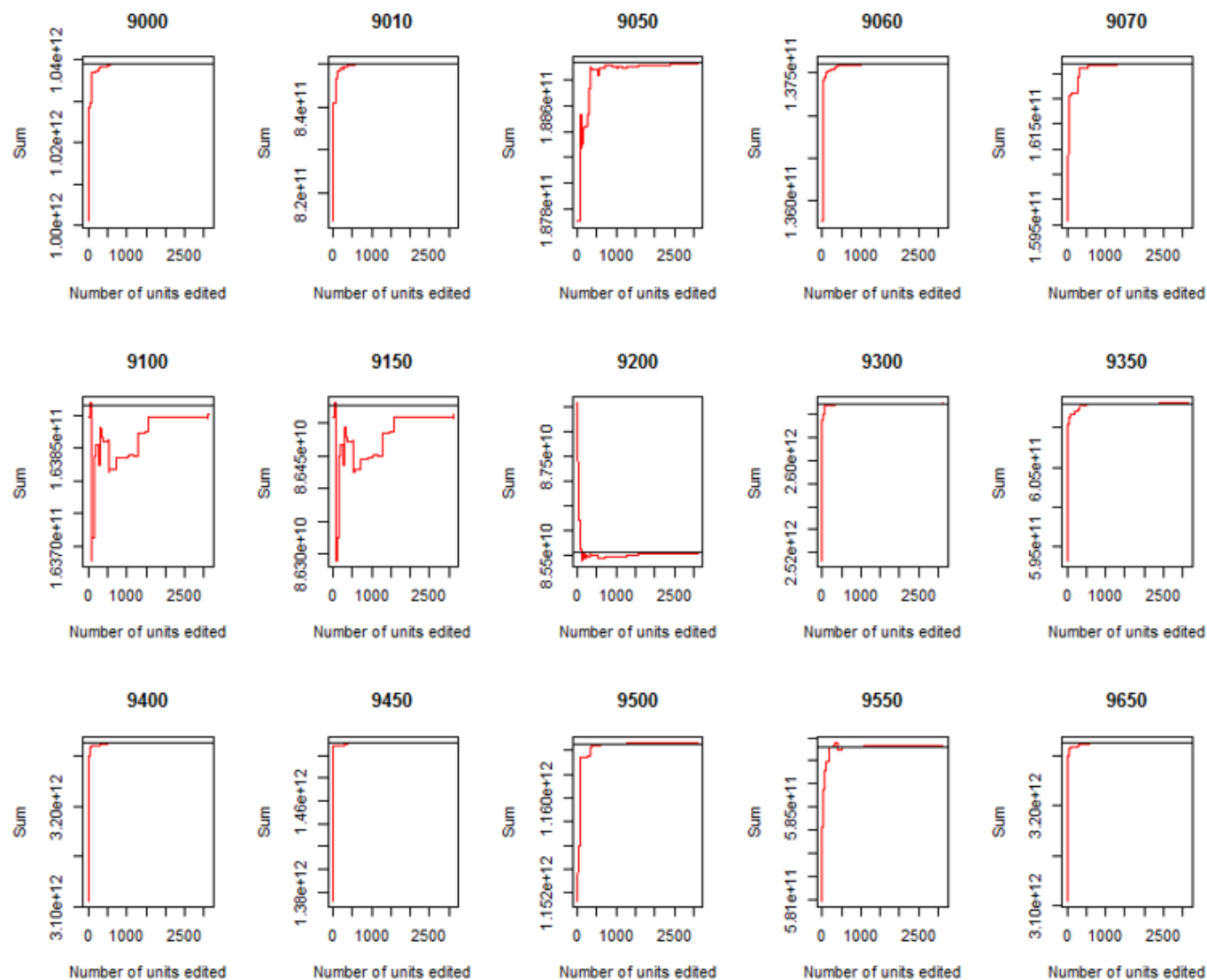
- Variabler brukt i formelen

p9000 - Sum driftsinntekter  
p9010 - Sum driftskostnader  
p9050 - Driftsresultat  
p9060 - Sum finansinntekter  
p9070 - Sum finanskostnader  
p9100 - Ordinært resultat før skattekostnad  
p9150 - Ordinært resultat  
p9200 – Årsresultat  
p9300 - Sum anleggsmidler  
p9350 - Sum omløpsmidler  
p9400 - Sum eiendeler  
p9450 - Sum egenkapital  
p9500 - Sum langsiktig gjeld  
p9550 - Sum kortsiktig gjeld  
p9650 - Sum egenkapital og gjeld



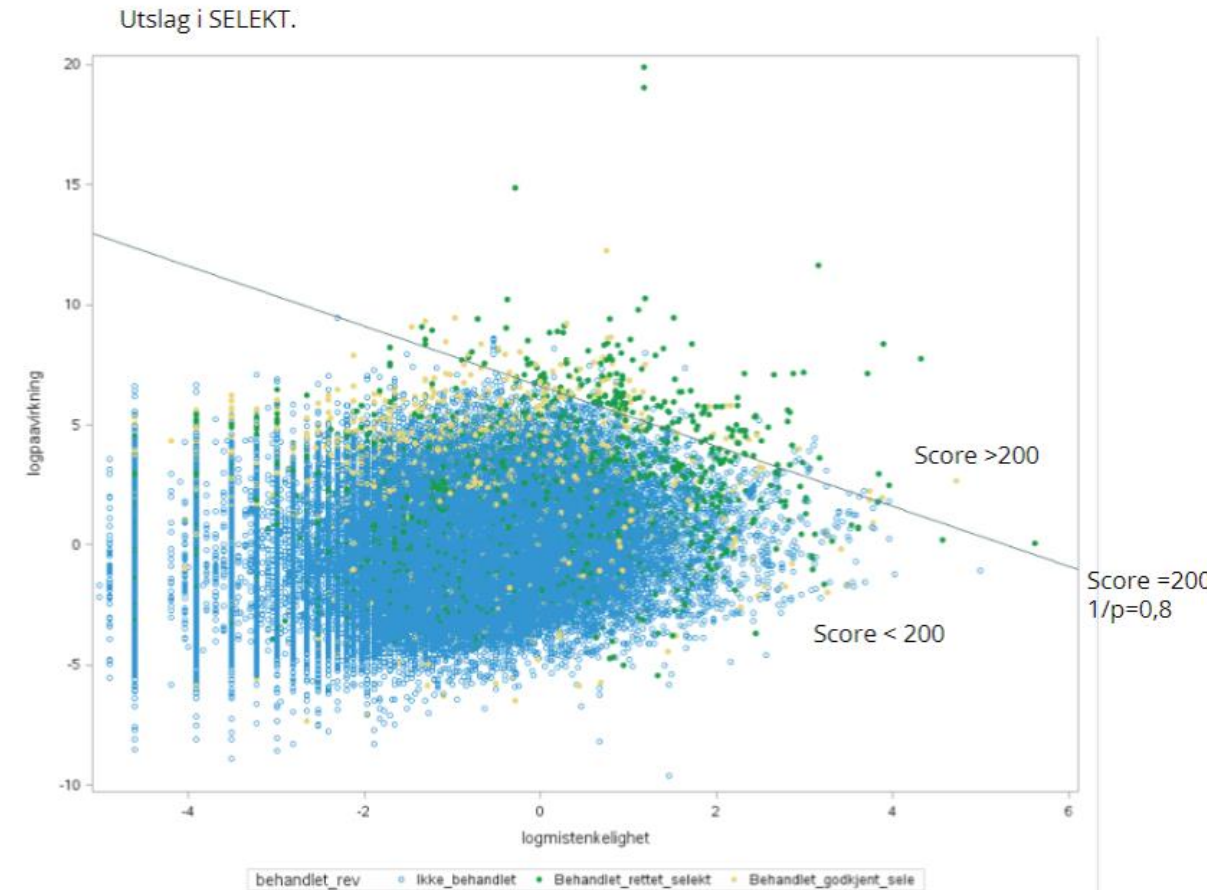
Figur 2. Effekt av å editere etter strategi med den høyeste enhet av poengfunksjon, antall enheter

3125



# Selekt - poengfunksjon

- Poengfunksjonen beregnes som et produkt av:
  - Mål for mistenkelighet
  - Mål for potensiell påvirkning for flere grupper
  - Parameteren  $p$  venter viktigheten av potensiell påvirkning
- Score = mistenkelig  $\times$  (potensiell påvirkning) $^p$
- Scoreverdien brukes til å prioritere oppfølging av utslagene i synkende rekkefølge.



# Mistenkelige verdier

- Lager grupperinger av data som er mest mulig **homogene**.
- Det settes krav til **tilstrekkelig records** i datagrunnlaget (12 records) og aktualitet (24 mnd tilbake i tid).
- Mistenkelighet beregnes som antall **interkvartile avstander** verdier avviker fra 1. eller 3. kvartil, og tas med videre inn i poengsummen.

$$Suspicion_i = \begin{cases} \frac{\log(UP_{Q1}(i)) - \log(UP_i)}{\log(UP_{Q3}(i)) - \log(UP_{Q1}(i))} & \text{if } UP_i < UP_{Q1}(i) \\ \frac{\log(UP_i) - \log(UP_{Q3}(i))}{\log(UP_{Q3}(i)) - \log(UP_{Q1}(i))} & \text{if } UP_i > UP_{Q3}(i) \end{cases}$$

# Potensiell påvirkning

- For hver mistenkelig record beregnes potensiell påvirkning som en estimert verdi feilrettingen vil ha på **ulike tabellsummer**, den høyeste verdien beholdes og tas med videre i poengfunksjonen.
- I telleren estimeres størrelsen på feilen som avviket mellom statistisk verdi og forventet verdi. Denne vektes mot summen av statistisk verdi for de 24 siste mnd.

$$Potential\ impact_i = \underset{over\ v=1-5}{\text{maximum}} \left\{ \frac{|Invoiced\ value_i - Quantity_i \cdot UP_{Q2}(i)|}{\sum_{k \in g_v} Invoiced\ value_k^*} \cdot \frac{1}{O_v} \cdot f^{10 \log \left( \sum_{k \in g_v} Invoiced\ value_k^* \right)} \right\}$$



# 10 tips for dataeditering

## **Tips 1. Sett deg godt inn i fagfeltet for statistikken og bakgrunnen til datasettet.**

Da vet du hvilke feil som kan forekomme og kan lettere vurdere om trendene statistikken viser er korrekte eller kanskje skyldes feil i datasettet.

## **Tips 2. Jobb for å få gode data inn**

Gode data inn er det mest effektive og er basisen for å lage statistikk av høy kvalitet.

## **Tips 3. Ha kontroller og korrigeringer så tidlig som mulig i prosessløpet**

Kontrollering og korrigering tidlig, fører til at de resterende prosesser ikke blir påvirket av feilen.

## **Tips 4. Kontroller at alle enheter er med i datasettet**

Kontroll av enheter kan være veldig vanskelig og krever mye kunnskap om fagområdet. Det inkluderer fjerning av enheter som ikke er kvalifiserte, og de enhetene som er relevante, bør oppdages og bli inkludert i undersøkelsen.

## **Tips 5. Sørg for at alle viktige variabler er dekket av en kontroll**

Ved mange variabler bør de viktigste variablene bli plukket ut og bli kontrollert grundigere enn mindre viktige variabler.



### **Tips 6. Automatiser korrigeringer så mye som mulig**

Økt bruk av automatisert imputering i produksjonsprosessen vil gjøre produksjonen mer effektiv.

### **Tips 7. Ha et makroperspektiv og prioriter det som påvirker statistikken mest**

Fokus på det som påvirker sluttproduktet mest gir overordnet perspektiv på statistikken og hjelp til å prioritere hva som er viktig, og med det sikre en effektiv ressursbruk.

### **Tips 8. Visualisering av datasettet kan gi en rask oversikt og hjelp til å avdekke feil**

Grafikk kan gi oversikt over trender og strukturer i data. Grafikk kan også gi oversikt over produksjonsløp og kvalitet.

### **Tips 9. Gransking effektiviseres ved hjelp av drilling i data og figurer**

Drilling i data gir mulighet til raskt å gå mellom nivåer i data, det effektiviserer leting etter årsak til mistenkelige verdier i statistikken

### **Tips 10. Dataediteringen bør evalueres for kontinuerlig forbedring av prosess og data**

Evaluering av produksjonsprosessen og datakvalitet er viktig for å kunne samle kunnskap slik at forbedringstiltak kan settes inn. Hvis tiltak blir satt inn, ved for eksempel at feil ikke oppstår igjen, vil prosessen bli mer effektiv og kvaliteten på statistikken bli høyere.



# Takk!



**Statistisk sentralbyrå**  
Statistics Norway