TALL SOM FORTELLER

Kurs i dataeditering: Validere og kontrollere

ASLAUG HURLEN FOSS OG ANE SEIERSTAD, 2024

Statistisk sentralbyrå
Statistics Norway

Plan for kurset

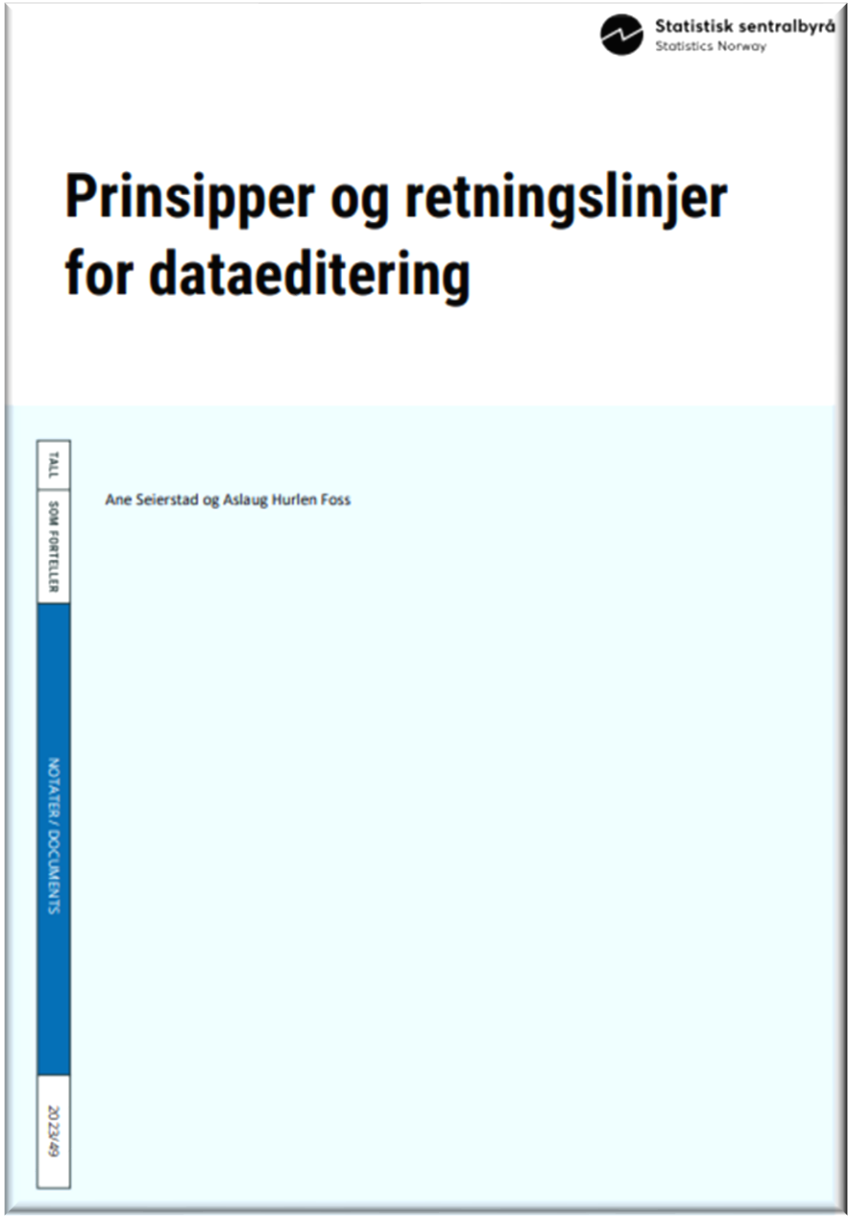
- 10:00-10:30 Introduksjon
- 10:30-11:00 Validering av data – øvelse med validate-pakken i R
- 11.00-11:30 Pandera pakken i Python med Jonatan Husø
- 11:30-12:15 Lunsj
- 12:15-14:00 Selektiv editering – Kostra pakken i R

Læringsmålet

- Kjenne til **SSB** sine prinsipper og retningslinjer for dataeditering
- Kjenne til **Eurostats** metodikk for validering
- Kjenne til **UNECE** sin modell for dataeditering
- Vite hvilke **logiske kontroller** som er anbefalt og kunne sette de opp
- Forstå **selektiv editering** basert på mistenkelige og innflytelse av verdier
- Forstå og bruke **metodene**: HB, kvartilmetode, regresjon og forskjellige mål på innflytelse

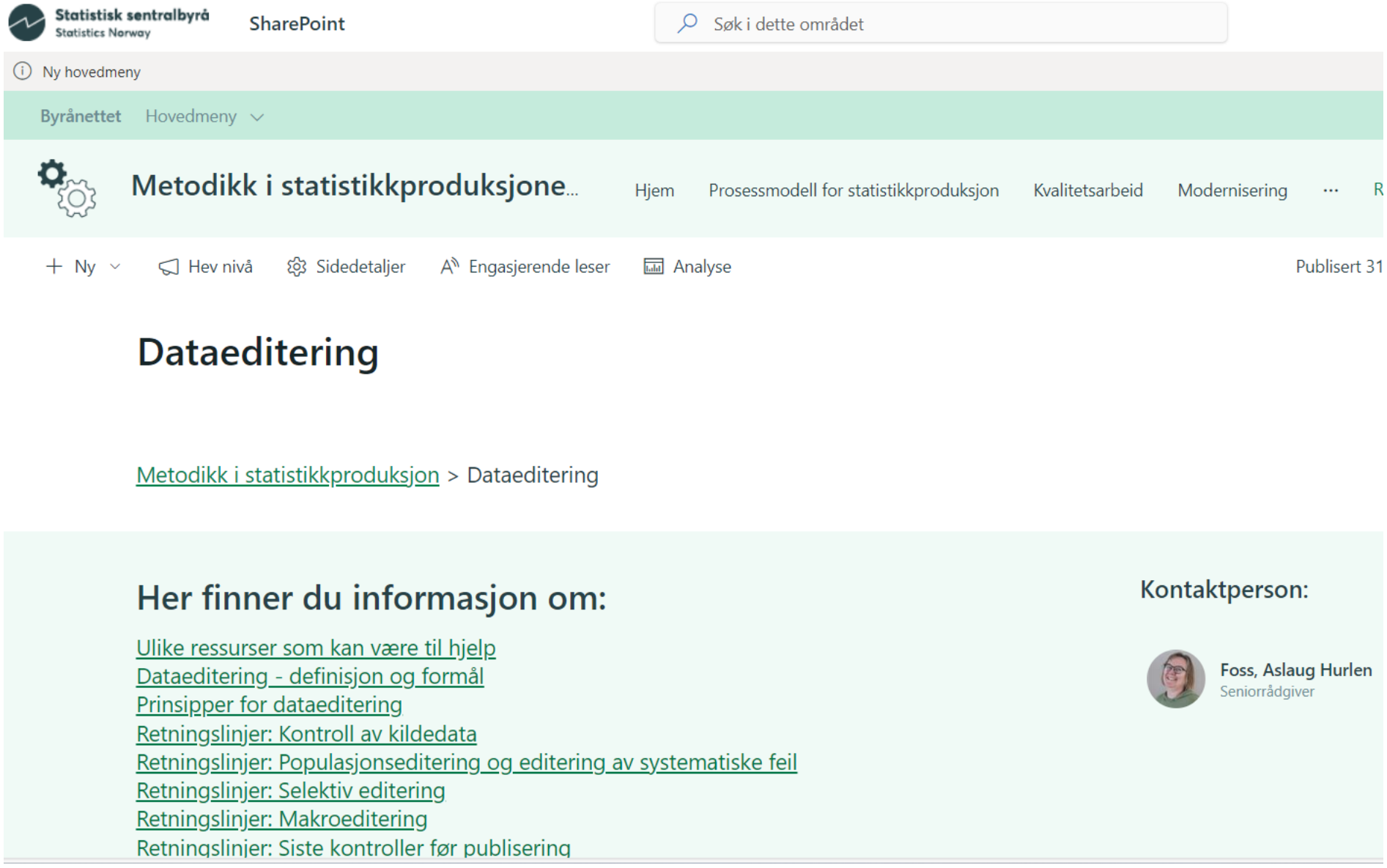
Notat

- Utarbeidet 2023
- Godkjent av alle direktørene
- Oversatt til engelsk

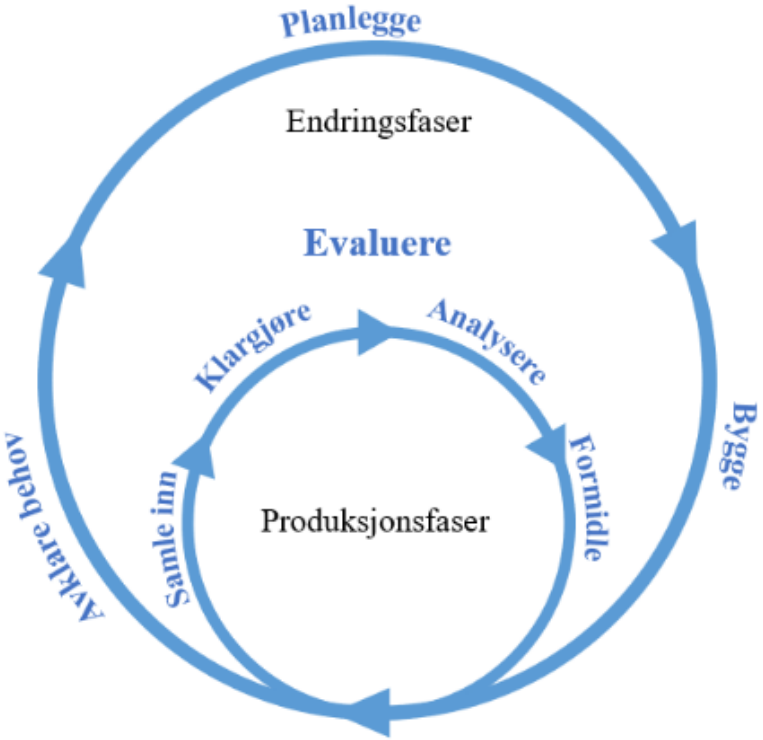


Statistisk sentralbyrå
Statistics Norway

Byrånettet: metodikk i statistikkproduksjon



Prosessmodell for statistikkproduksjon



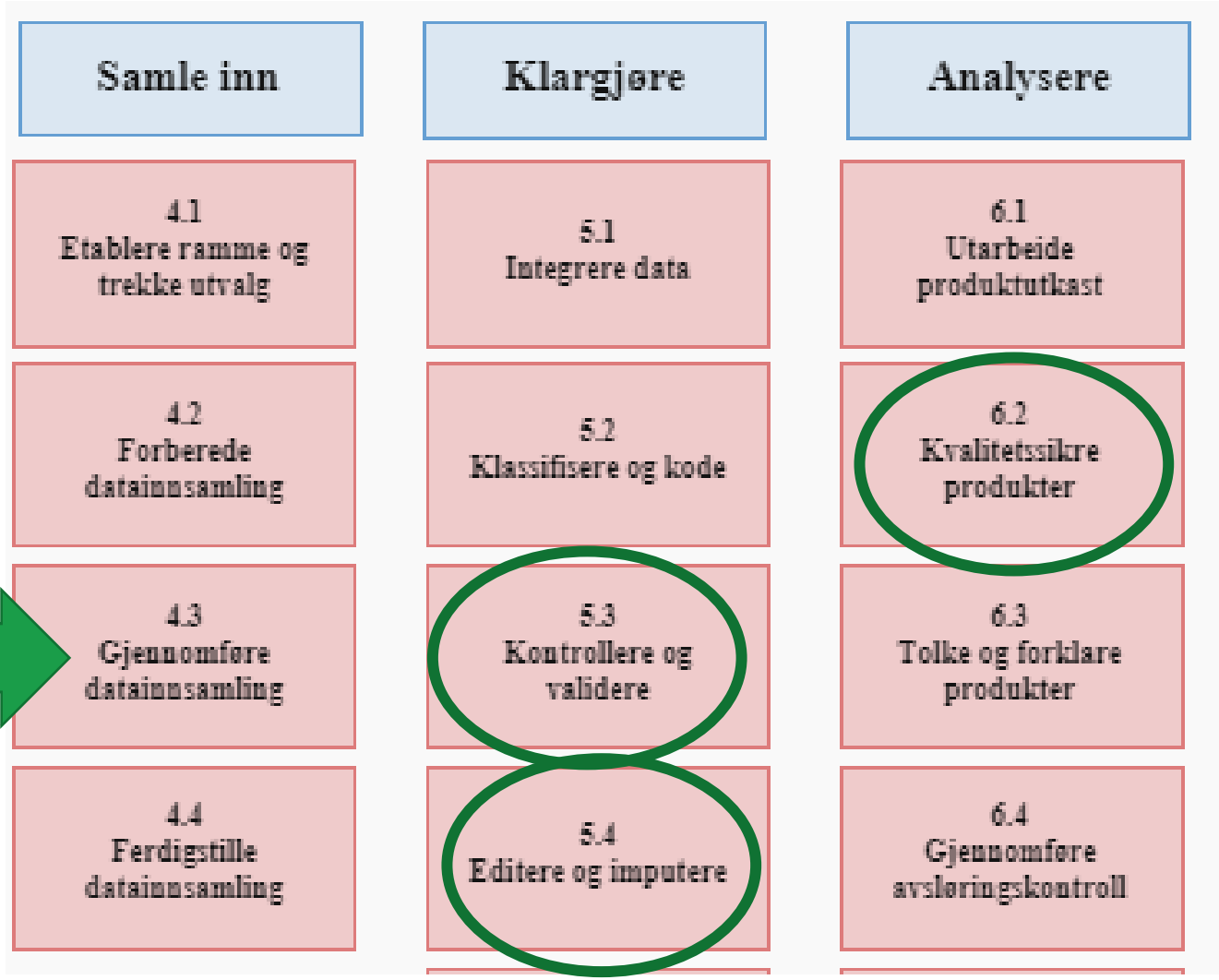
Produksjonsfaser og endringsfaser i prosessmodellen

Overordnede prosesser

Avklare behov	Planlegge	Bygge	Samle inn	Klargjøre	Analysere	Formidle	Evaluere
1.1 Identifisere behov	2.1 Planlegge produkt	3.1 Gjenbruke eller bygge innsamlings-instrumenter	4.1 Etablere ramme og trekke utvalg	5.1 Integrere data	6.1 Utarbeide produktutkast	7.1 Oppdatere outputsystemer	8.1 Samle input til evalueringen
1.2 Undersøke og bekrefte behov	2.2 Planlegge variabler	3.2 Gjenbruke eller bygge prosess- og analysekomponenter	4.2 Forberede datainnsamling	5.2 Klassifisere og kode	6.2 Kvalitetssikre produkter	7.2 Produsere formidlings-produkter	8.2 Utføre evalueringen
1.3 Etablere produktmål	2.3 Planlegge innsamling	3.3 Gjenbruke eller bygge formidlings-komponenter	4.3 Gjennomføre datainnsamling	5.3 Kontrollere og validere	6.3 Tolke og forklare produkter	7.3 Håndtere formidling av produkter	8.3 Bli enige om tiltaksplan
1.4 Identifisere begreper	2.4 Planlegge ramme og utvalg	3.4 Sette sammen arbeidsflyt	4.4 Ferdigstille datainnsamling	5.4 Editere og imputere	6.4 Gjennomføre avsløringskontroll	7.4 Markedsføre formidlings-produkter	
1.5 Kontrollere data-tilgjengelighet	2.5 Planlegge klargjøring og analyse	3.5 Teste produksjons-systemer		5.5 Avlede nye variabler og enheter	6.5 Ferdigstille produkter	7.5 Håndtere brukerstøtte	
1.6 Forberede og levere forretnings-begrunnelse	2.6 Planlegge produksjonssystemer og arbeidsflyt	3.6 Teste statistisk forretningsprosess		5.6 Beregne vekter			
		3.7 Ferdigstille produksjonssystem		5.7 Beregne aggregater			
				5.8 Ferdigstille datafiler			

Statistisk sentralbyrå
Statistics Norway

Dataeditering er kontroll, granskning og retting av data



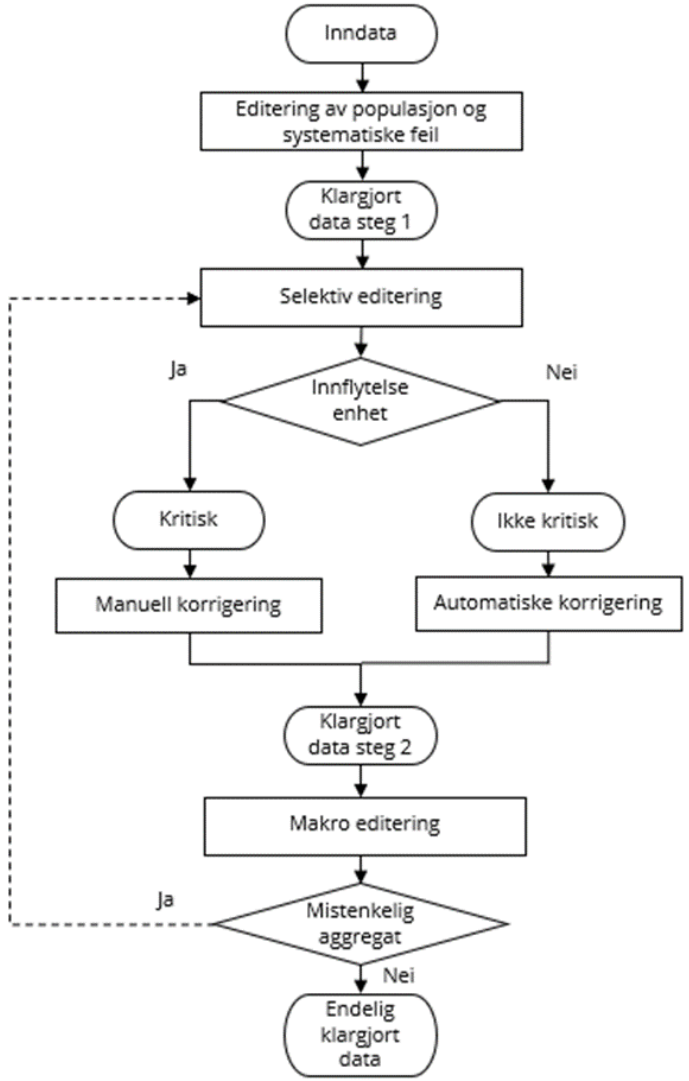
Kontroll

Statistisk sentralbyrå
Statistics Norway

Prosessmodell for dataeditering

Generic Statistical Data Editing Model – GSDEM:

- Populasjon
- Systematiske feil
- Selektiv editering
- Manuell og automatisk korrigering
- Makro editering



Programvare for kontrollere data

- Dynarev og Driller
- R – validate- pakken og kostra-pakken
- Python - Pandera

- Under utvikling – næringsstatistikk

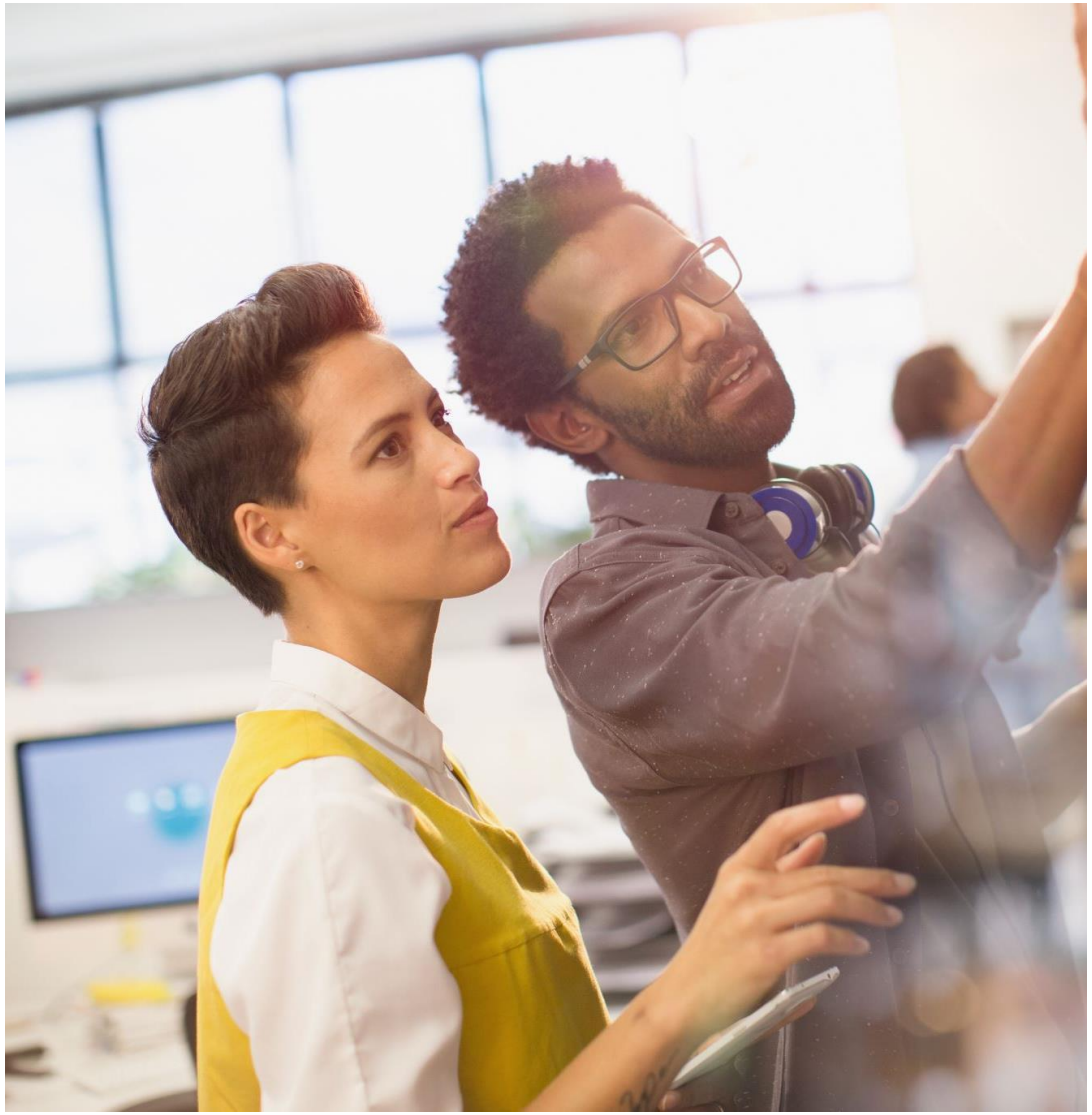
TALL

SOM FORTELLER

Prinsipper

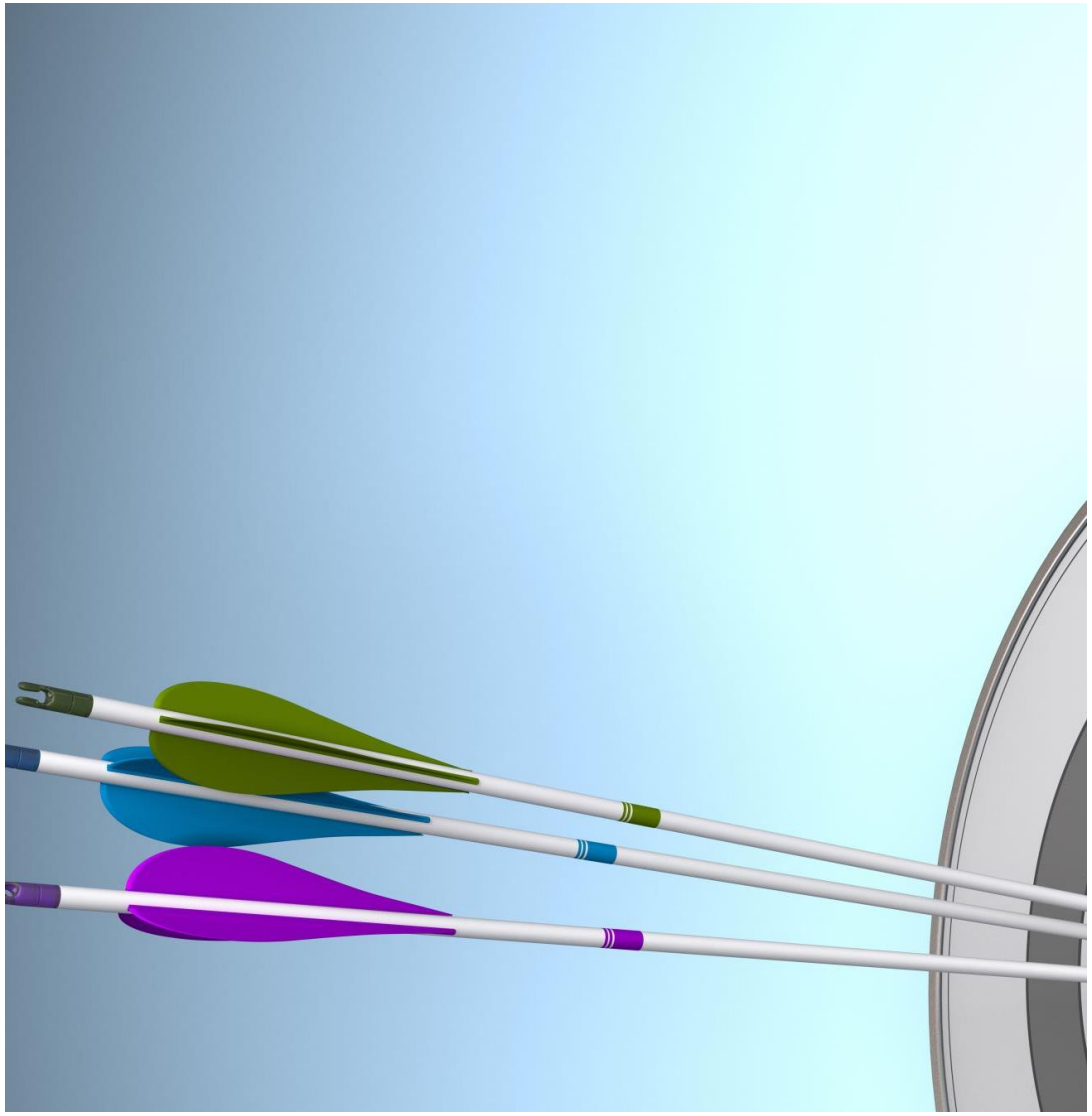
Kunnskap

1. God kunnskap om fagfeltet for statistikken og bakgrunnen til kildedata, er grunnlaget for en god editeringsprosess



Formål

2. Formålet med dataediteringen skal være klart formulert

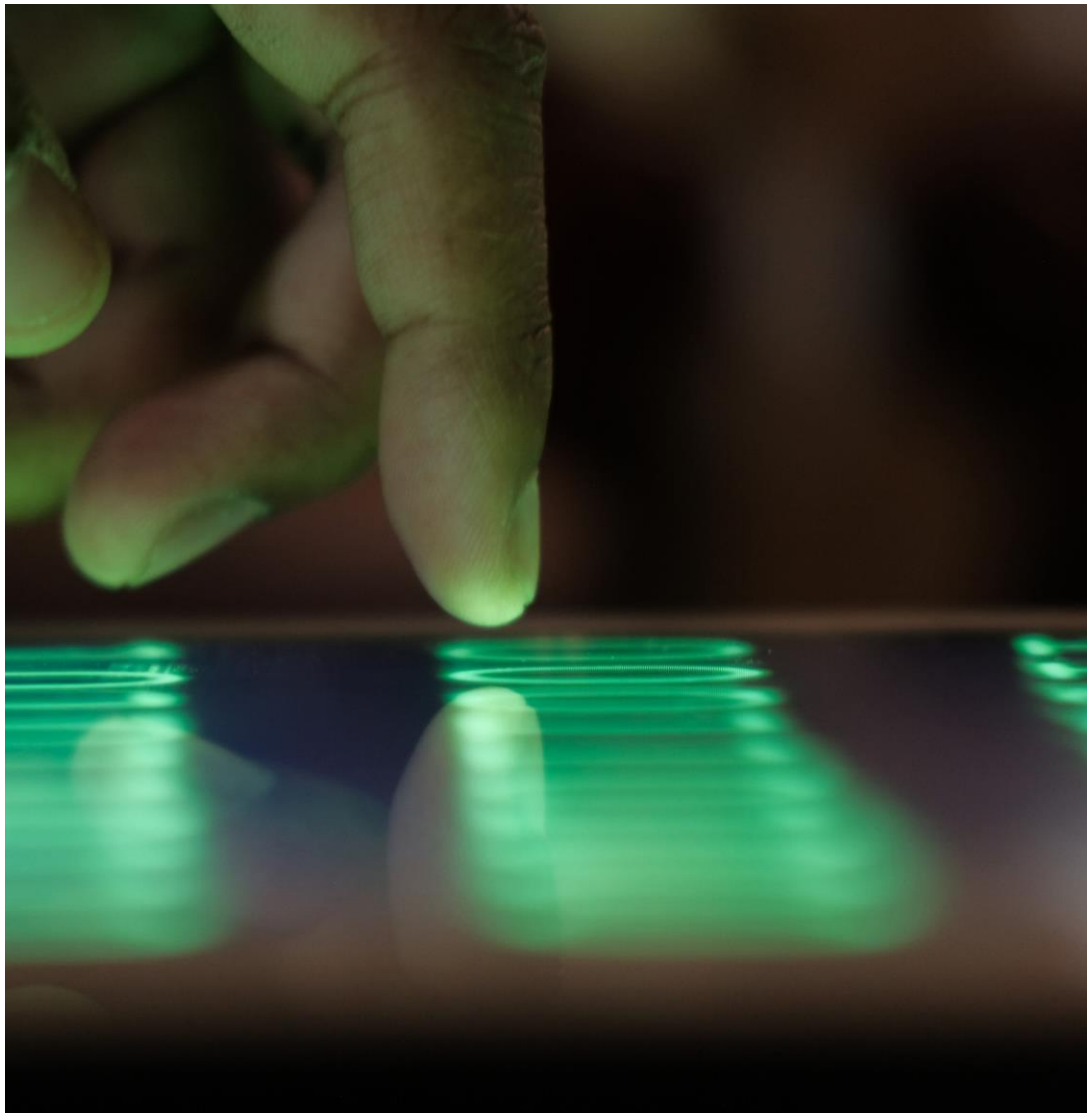


Statistisk sentralbyrå
Statistics Norway

Gode data

3. Gode data inn er best

- Gode spørreskjemaer
- God dialog med datalevrandør



Statistisk sentralbyrå
Statistics Norway

Kontroller

4. Kontroller alltid data

Den som mottar data, er ansvarlig for å kontrollere data ut ifra behovene til statistikken som skal produseres

Plasser

5. Jo tidligere jo bedre

- 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.

Dokumenter

6. Kontrollene, kontrollutslagene og endringene skal være veldokumenterte

Automatiser

7. Automatiser editeringsprosessen så mye som mulig

Kan nye funksjoner brukes?

- Logisk korrigering
- Modellbasert imputering
- Maskinlæring og KI

Effektiviser

8. Effektiviser editeringsarbeidet

Gjør den menneskelige interaksjonen bedre:

- Gode grensesnitt
- Visualisering
- Selektiv editering
- Drilling i data mellom makro og mikro

Evaluer

9. Dataediteringen bør evalueres

- Lag kvalitetsindikatorer
- Analyser indikatorene
- Sett inn tiltak

Statistisk sentralbyrå
Statistics Norway

TALL SOM FORTELLER

Metodikk for validering av data Eurostats

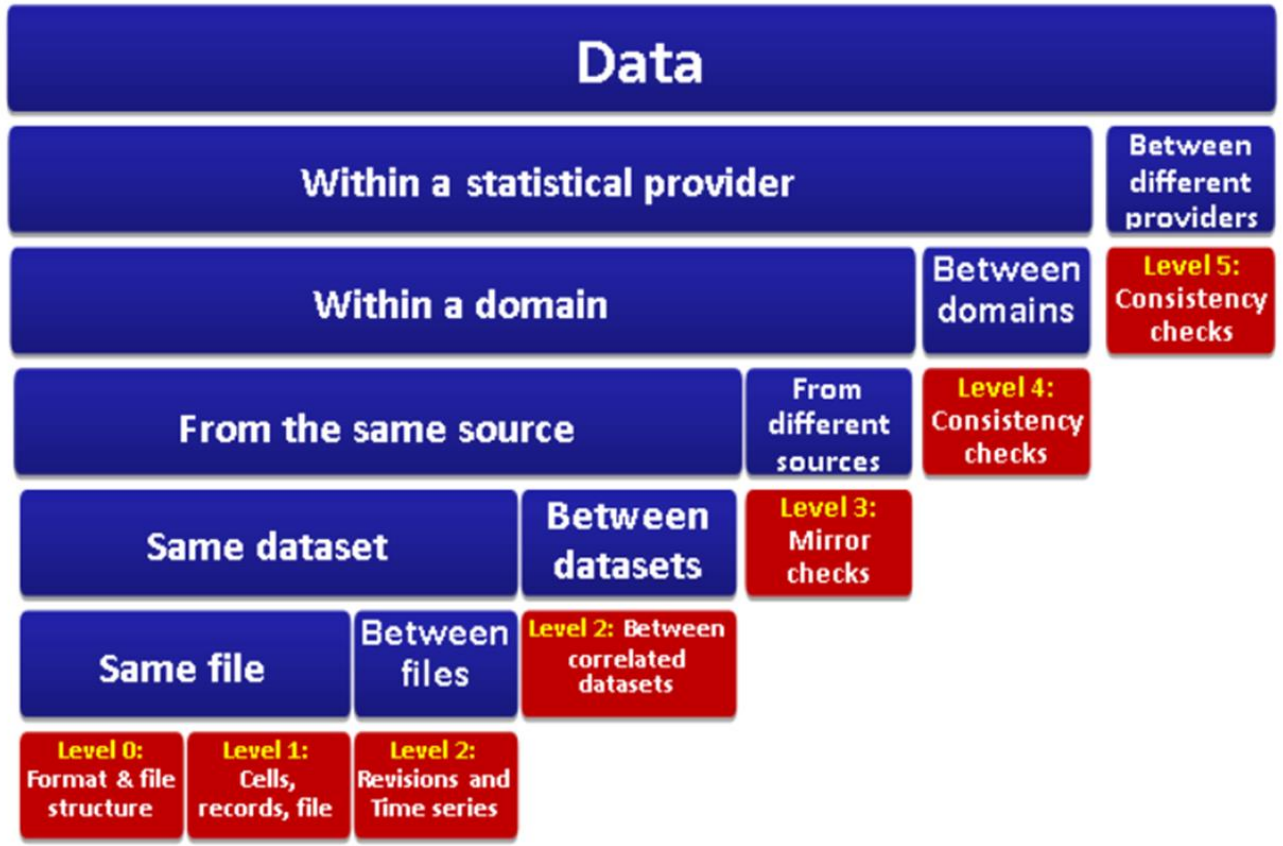
Håndbok i datavalidering

- Methodology for data validation, EUROSTAT

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
Kontrollutslagsrate - 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

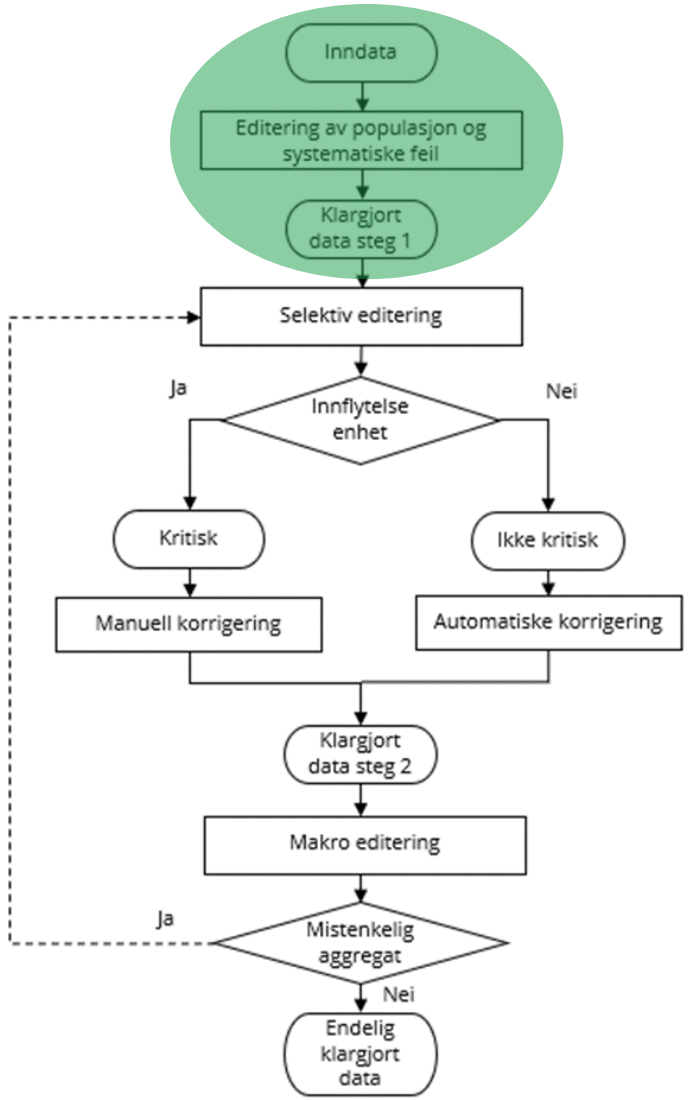
Gruppeøvelse

- Er kontrollene, kontrollutslagene og endringene **veldokumentert** i din statistikk?

TALL SOM FORTELLER

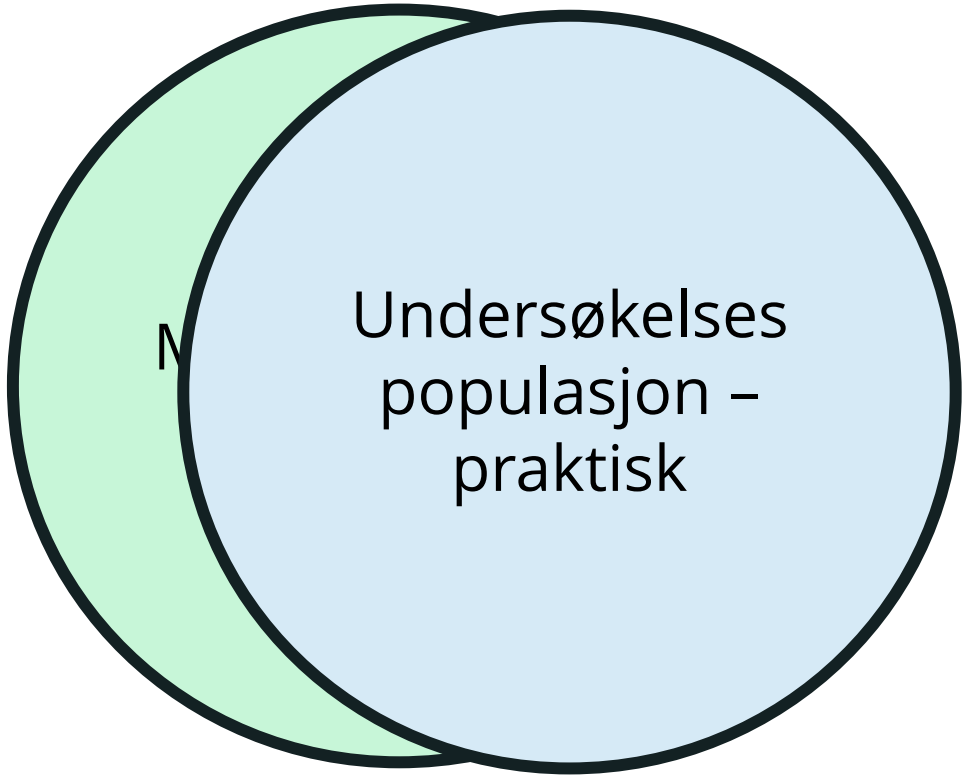
Editering av populasjon

Editering av populasjon og systematiske feil



Statistisk sentralbyrå
Statistics Norway

Editering av populasjon



Editering av populasjon –er verifisering og valg av kvalifiserte enheter med tilhørende klassifiseringsvariabler (f.eks. næring, juridisk status).

lbyrå

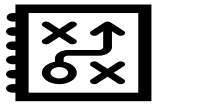
Editering av populasjon - personstatistikk

Retningslinjer personstatistikk:

- Størrelsen på populasjonen må kontrolleres for å se om den er rimelig i forhold til tidligere perioder eller andre datakilder.
- Det skal sjekkes for dubletter, og hvis de finnes så skal de ryddes opp i.

Editering av populasjon – næringsstatistikk

Hvordan editere populasjonen::



Kombinasjoner av variabler – høyeste verdi i binæring?

Fagkunnskap om emne og enheter - juridisk enhet ikke lik statistisk enhet

Følge med på nyheter – hva skjer med de viktige enhetene?

Maskinlæring - klassifisering

Statistisk sentralbyrå
Statistics Norway

TALL SOM FORTELLER

Editering av systematiske feil

Editering av systematisk feil

Systematiske feil – er gjennomgående feil som trekker i en retning og forekommer for mange enheter i undersøkelsen

- Kan ha stor påvirkning på statistikken - skjevhet
- Store utvalg reduserer ikke skjevhet fra systematiske feil
- Oppdages ofte med sammenligning med annen statistikk

Editering av åpenbare feil

Åpenbare feil - er observasjoner som har klart uriktige verdier

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

- Omfatter kontrollfunksjoner som:
  - Validere – logiske kontroller numeriske og kategoriske variabler

Innledende kontroller - retningslinjer

- Kontroller at datasettet inneholder alle enheter og variabler
- Kontroller at variabler har riktig format.
- Kontroller at datacellene har verdier.
- Kontroller at verdiene er i gyldig verdiområde.

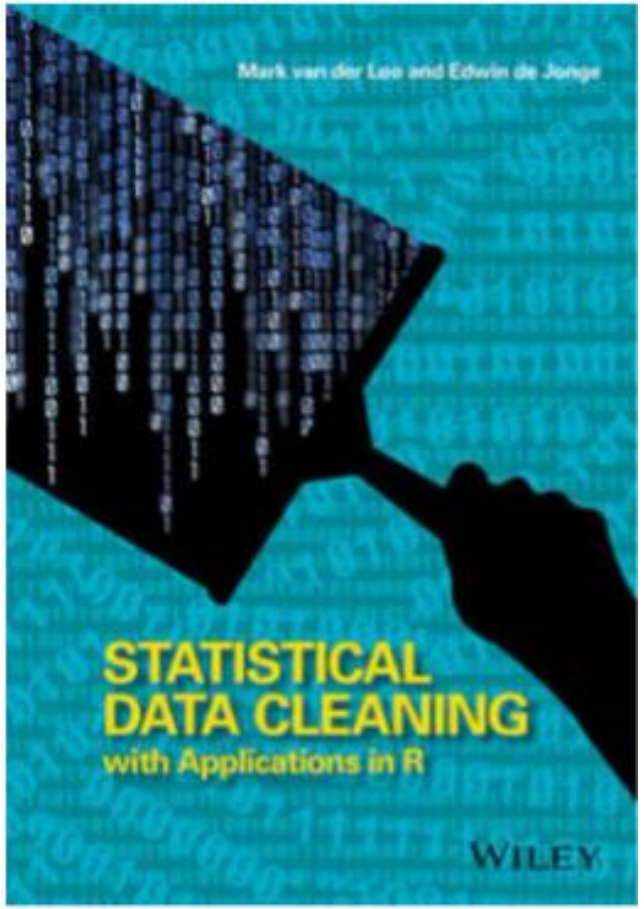
Tusenfeil - Verdi oppgitt i feil enhet

- Eksempel:
  - Svar i kroner når det skal oppgis i 1000 kroner.
  - Eller i årsverk når de skal oppgi svaret i antall timer per uke.
- Automatisk oppretting – ut fra regler

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

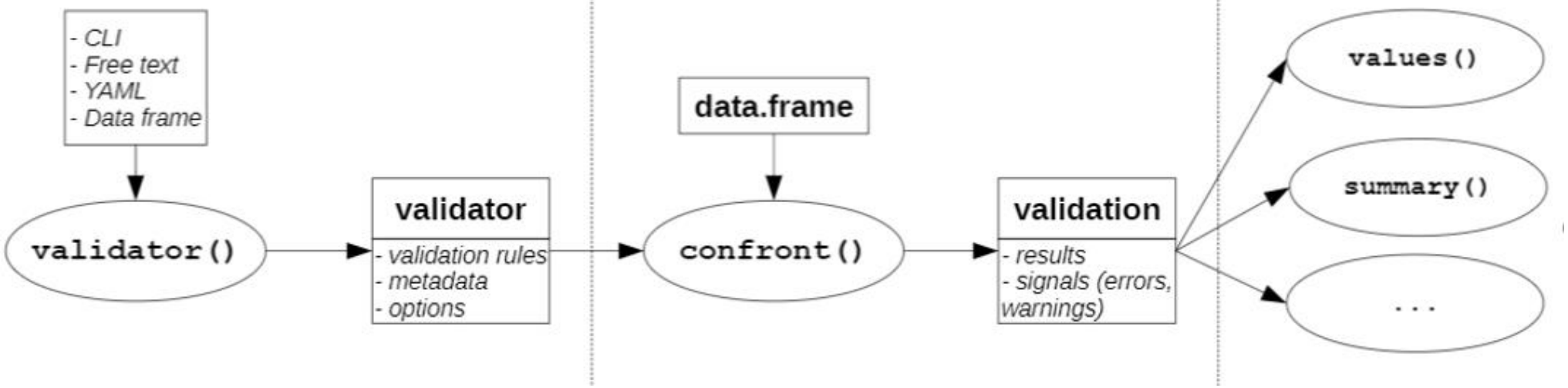
This book is about checking data with the validate package for R.

Validate-pakken

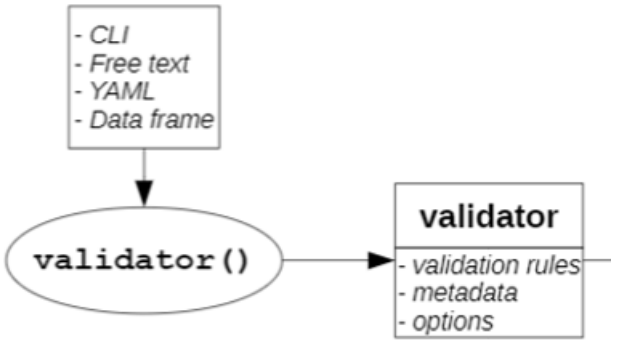
Sette opp og vedlikeholde kontroller

Kjøre kontroller på data

Analyse av kontroller



Validator – sette opp kontroller



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.

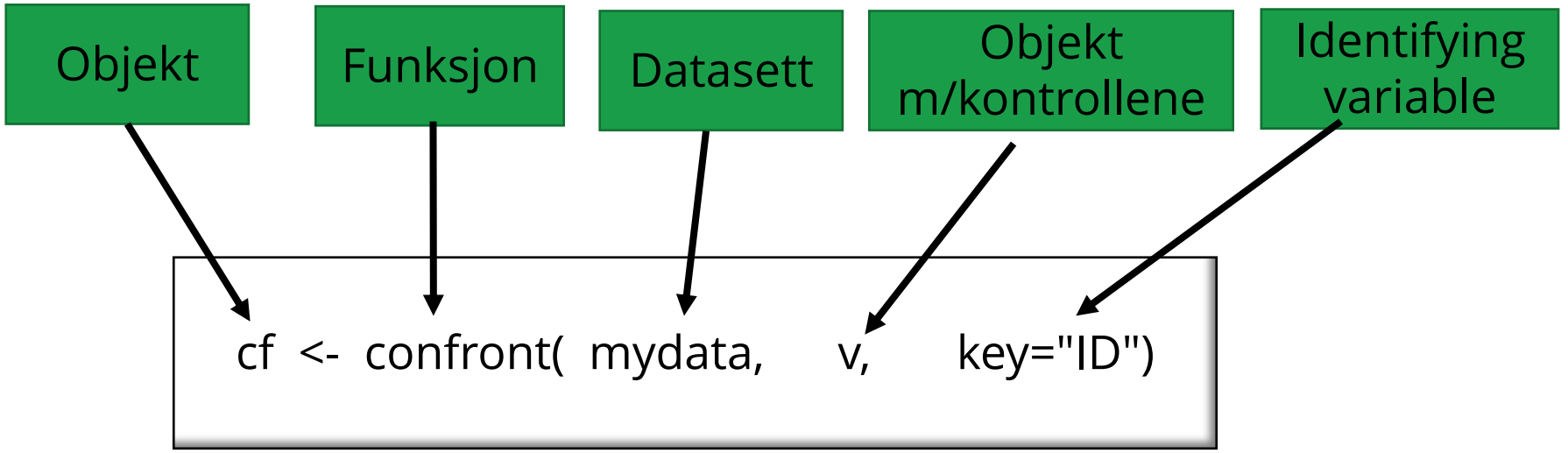
Spesialfunksjoner

- 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<-sn %>% select("code"))
  - regler<-validator(region %in% komliste)

Confront - 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

- summary: Oppsummert resultat
- aggregate: Forholdstall - indikatorer
- values: Får resulatene i en matrise True, False, Missing
- 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

Summary – summerer opp resultatet

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 – gir relative tall

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 – attributt variabel for utslag

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 indikators

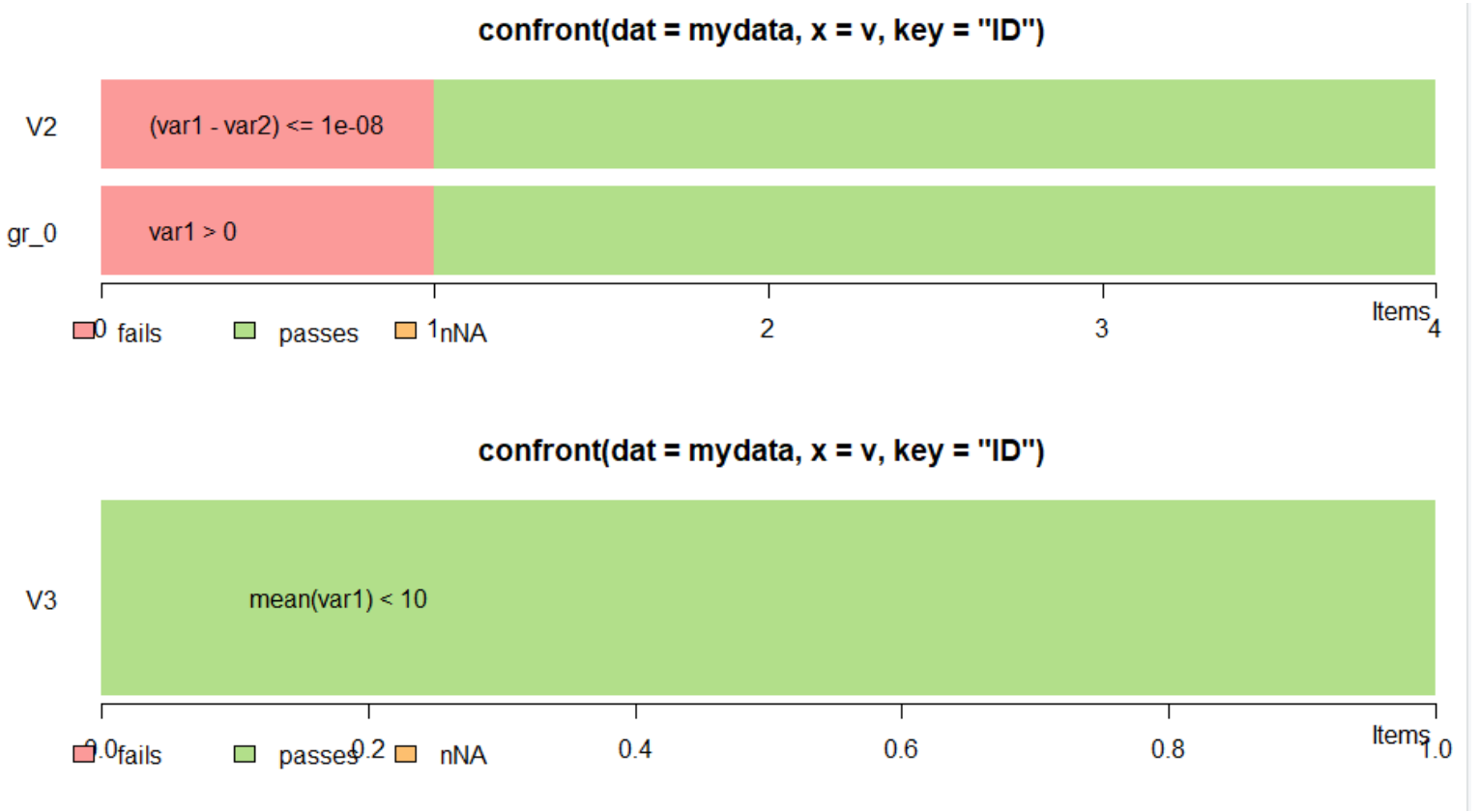
```
ind<-as.data.frame(values(cf))

# add indikators 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)



Metadata for kontrollene

- 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

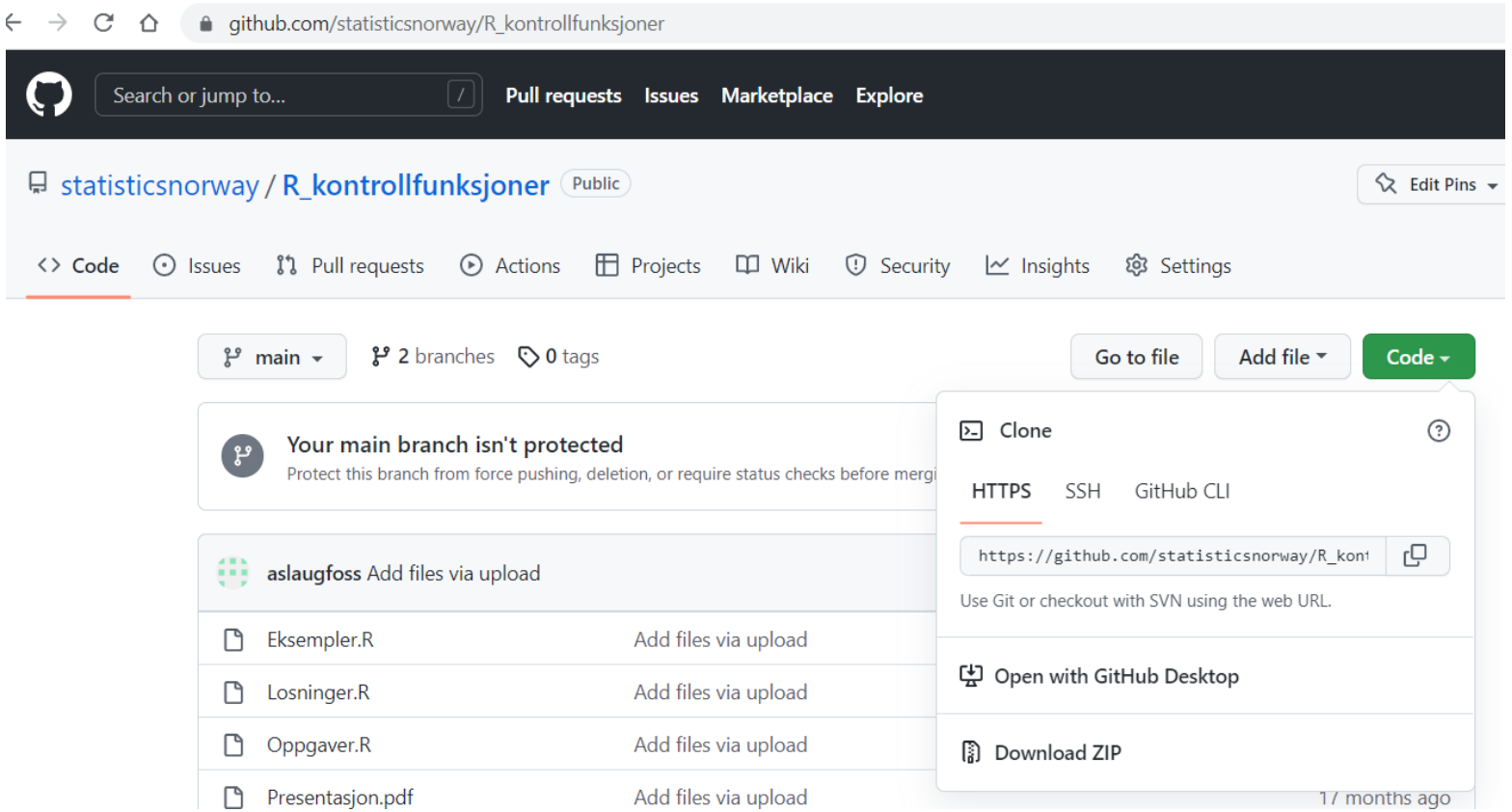
Innledende kontroller

- Kontroller at datasettet inneholder alle enheter og variabler
  - Funksjoner i R: nrow() og ncol()
- Kontroller at variabler har riktig format.
  - Funksjoner i R: is.numeric() og is.character()
- Kontroller at datacellene har verdier.
  - Funksjoner i R: all_complete(), is_complete()
- Kontroller at verdiene er i gyldig verdiområde.
  - Numerisk: in_range(variabel, min=0, max=1)
  - Kategorisk: variabel %in% kodeliste

Eksempel i R på Jupyter

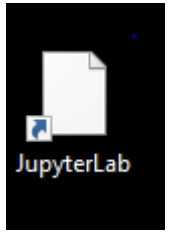
Kursmaterialet

https://github.com/statisticsnorway/kurs-metode-validere

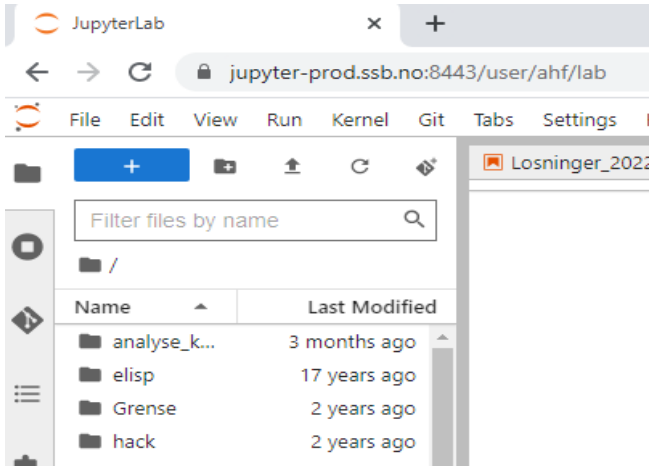


Starte opp Jupyter i produksjonssonen

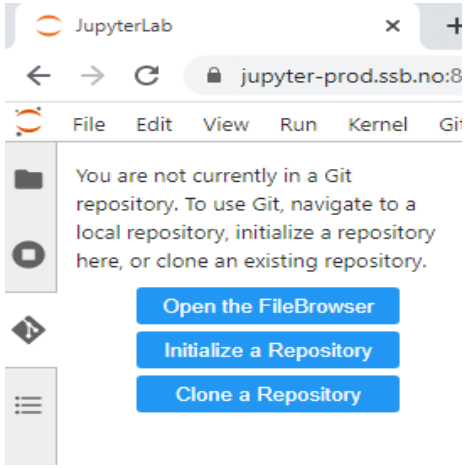
- Trykk på ikonet:

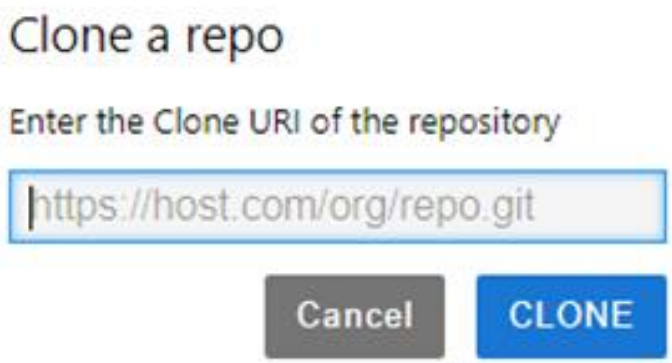


- Stå i «filutforsker»



- Trykk på Github-ikonet:





Øvelse

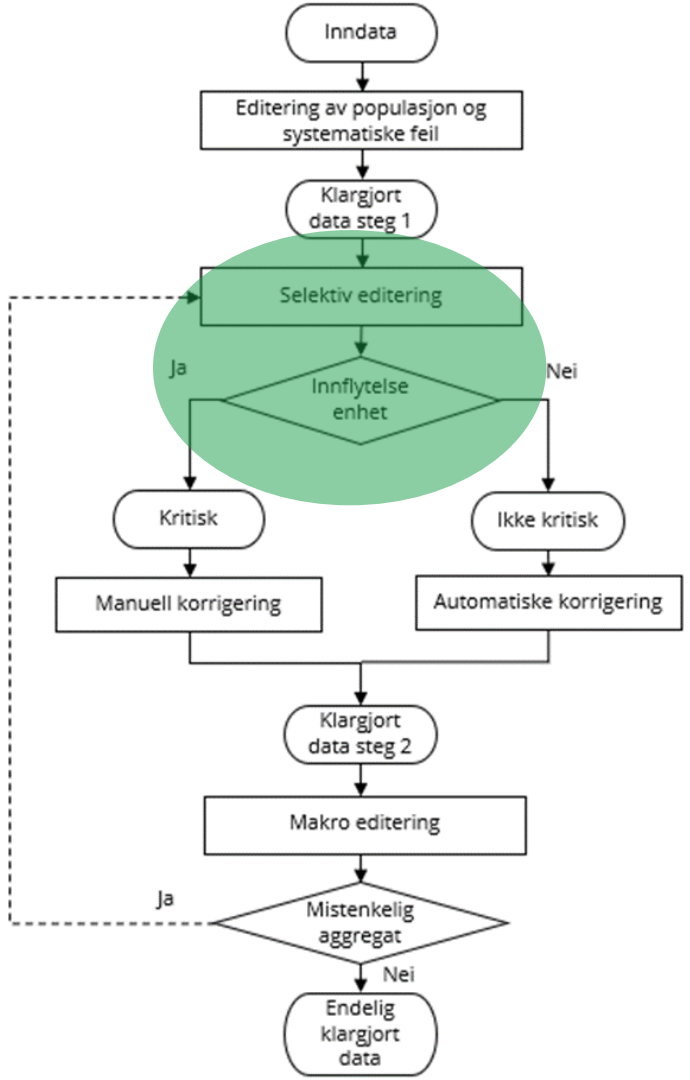
- Oppgave 1-4 pluss ekstraoppgave
  - Kirkedata med variabelen døpte
  - «Oppgaver_jupyter»
- Oppgavene
  - Github: https://github.com/statisticsnorway/kurs-metode-validere
  - Hvis det er utfordrende å kode; kjør «losninger», varier parametere og vurder resultatene
  - Ekstra kokeboken: https://cran.r-project.org/web/packages/validate/vignettes/cookbook.html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

TALL SOM FORTELLER

Selektiv editering

Selektiv editering

- Selektiv editering - er en generell tilnærming for å oppdage innflytelsesrike feil med hensyn til hovedresultatene.
- Mistenkelig – er observasjoner med verdier som ligger utenfor det som er forventet
- Innflytelse – er verdi som har stor påvirkning på resultatet



Statistisk sentralbyrå
Statistics Norway

Selektiv editering - retningslinjer

- De enhetene og verdiene som har høy innflytelse og er mistenkelige bør plukkes ut til manuell inspeksjon.
- De enhetene og verdiene som er mistenkelige og har lav innflytelse bør korrigeres automatisk.

Mistenkelige verdier: Kontrollmetoder

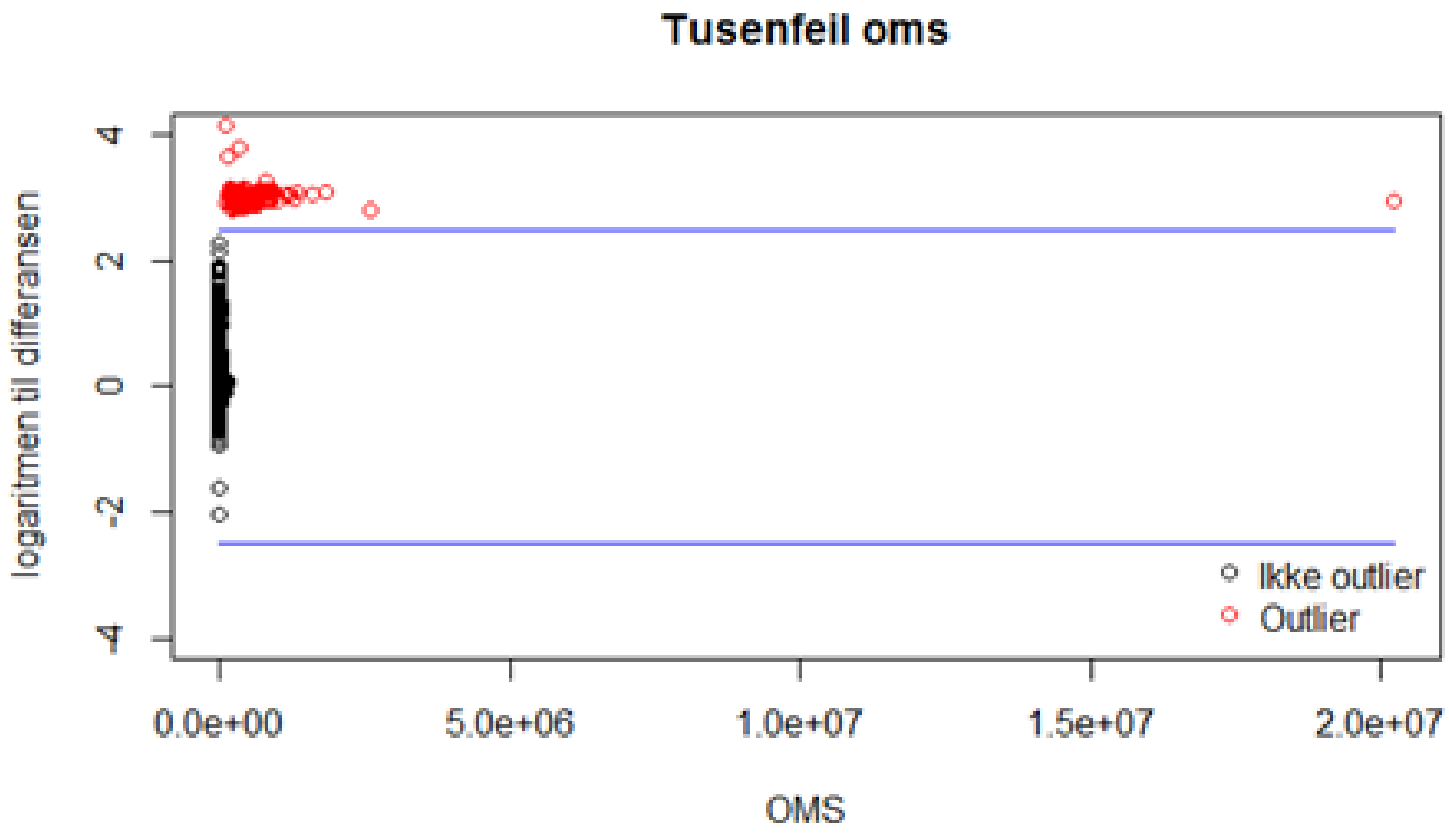
- Tusenfeil
- HB-metoden
- Kvartilmetoden
- Robust regresjon

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

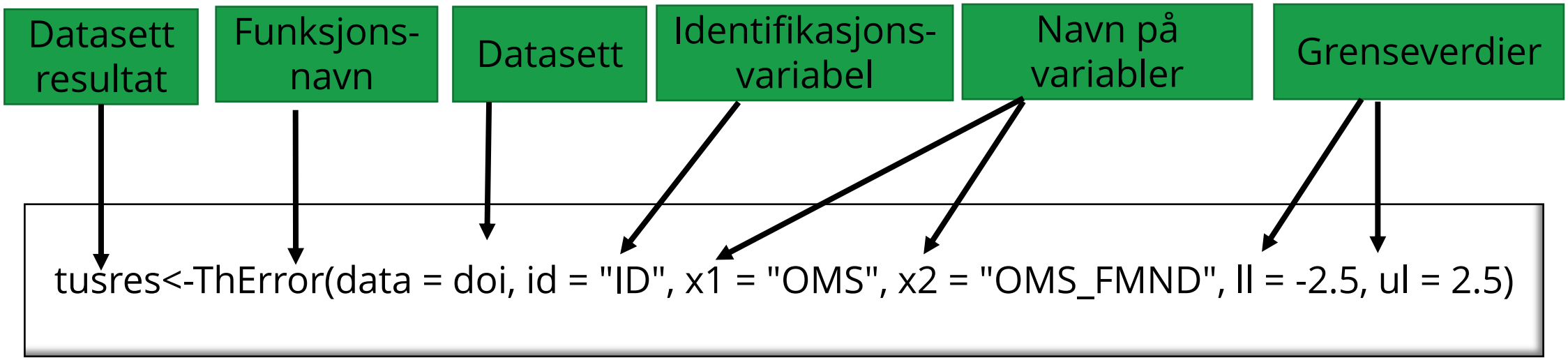


Statistisk sentralbyrå
Statistics Norway

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 log10 (x1 / x2) = log10 (x1) - log10 (x2). Standard -2,5
- ul Øvre grense for log10 (x1 / x2) = log10 (x1) - log10 (x2). Standard +2,5

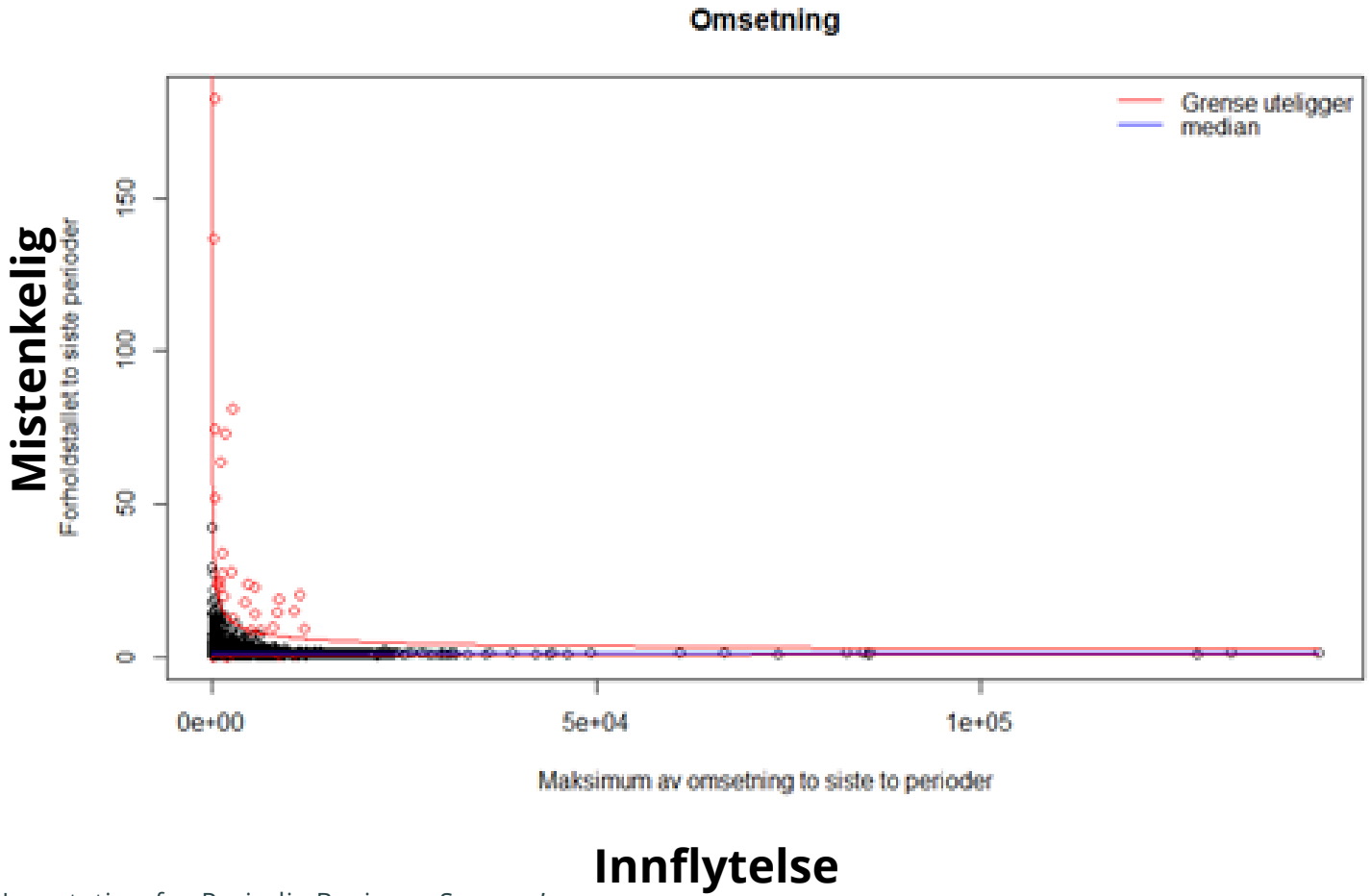
Output

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 log10 (x1) - log10 (x2)
- lowereLimit Inngangsparameteren ll
- upperLimit Inngangsparameteren ul

Kontroll mot forrige periode: HB- metoden

- Tar hensyn til nivåendring mellom perioder
- Tar hensyn til at små verdier har større variasjon enn store verdier



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

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

Akseptgrenser

$$D_{Q1} = MAX\left(E_{median} - E_{Q1}, \left|A * E_{median}\right|\right)$$
$$D_{Q3} = MAX\left(E_{Q3} - E_{median,}, \left|A * E_{median}\right|\right)$$

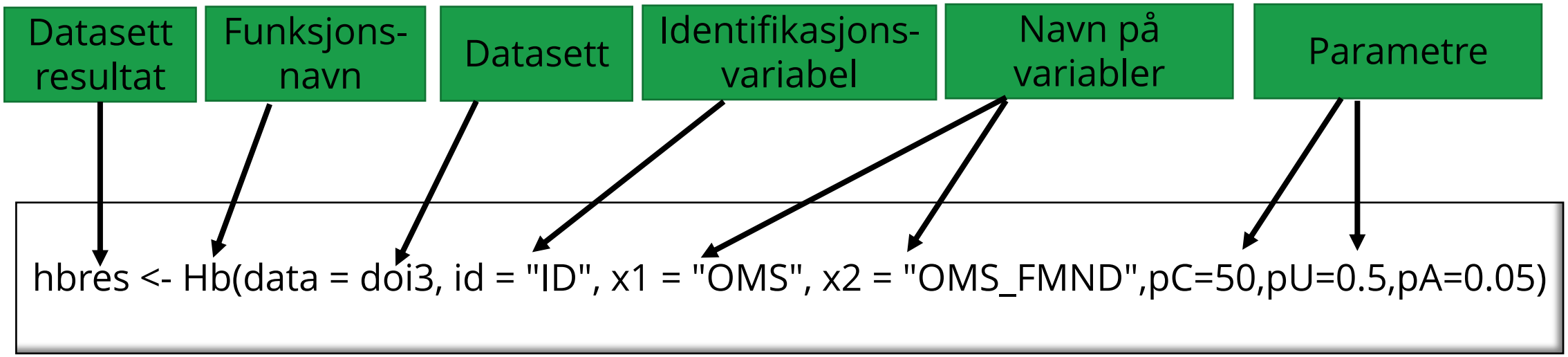
$$\left\{ E_{median} - C * D_{Q1} \quad , \quad E_{median} + C * D_{Q3} \right\}$$

Statistisk sentralbyrå
Statistics Norway

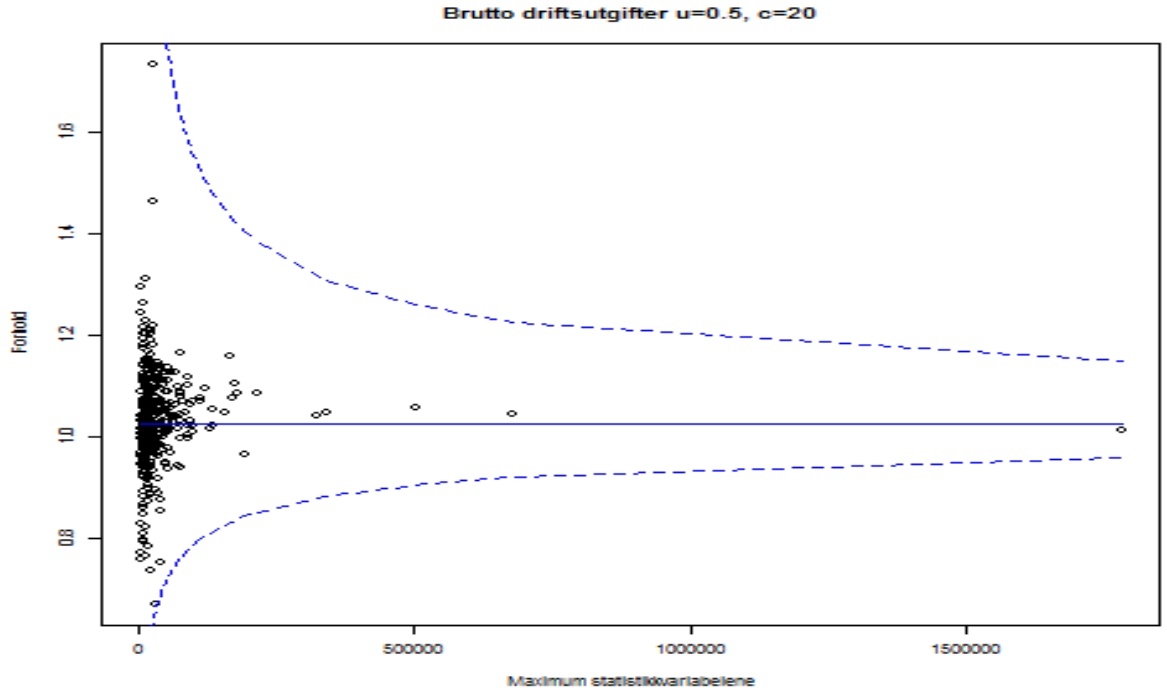
Parametre

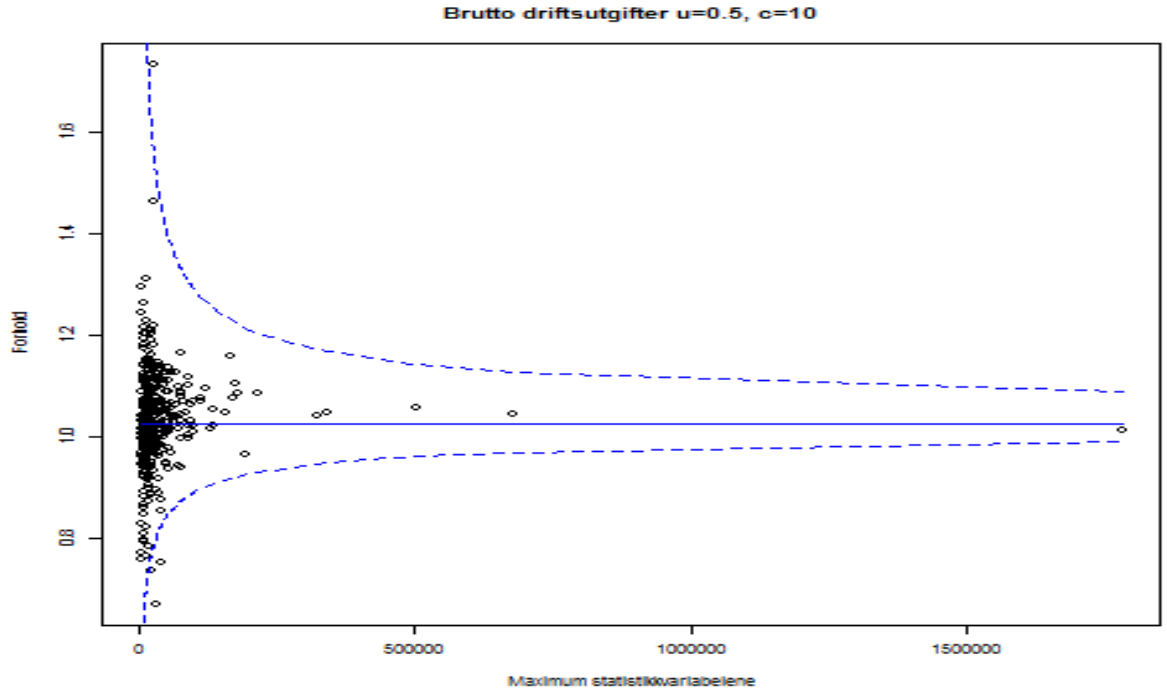
- C parameter for bredden på intervallet. Default verdi 20.
- U parameter justerer for krumning av grensen. Default verdi 0.5.
- A parameter justerer for liten forskjell mellom median og 1. og 3. kvartil. Default Verdi 0.05.

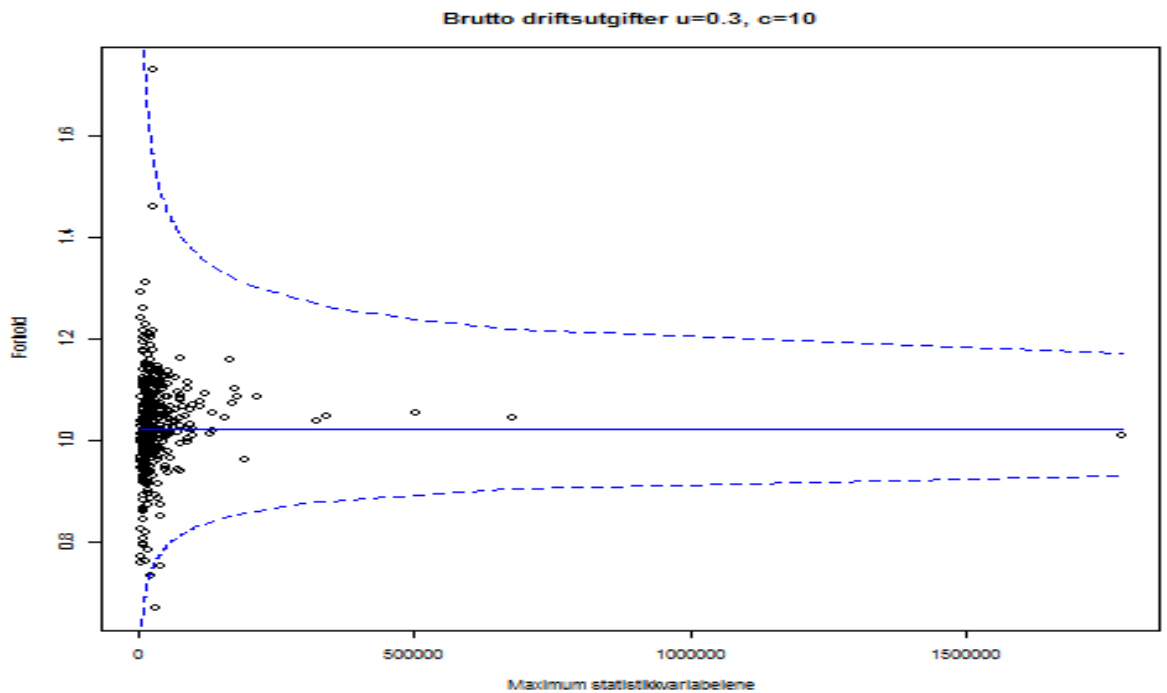
HB-funksjon

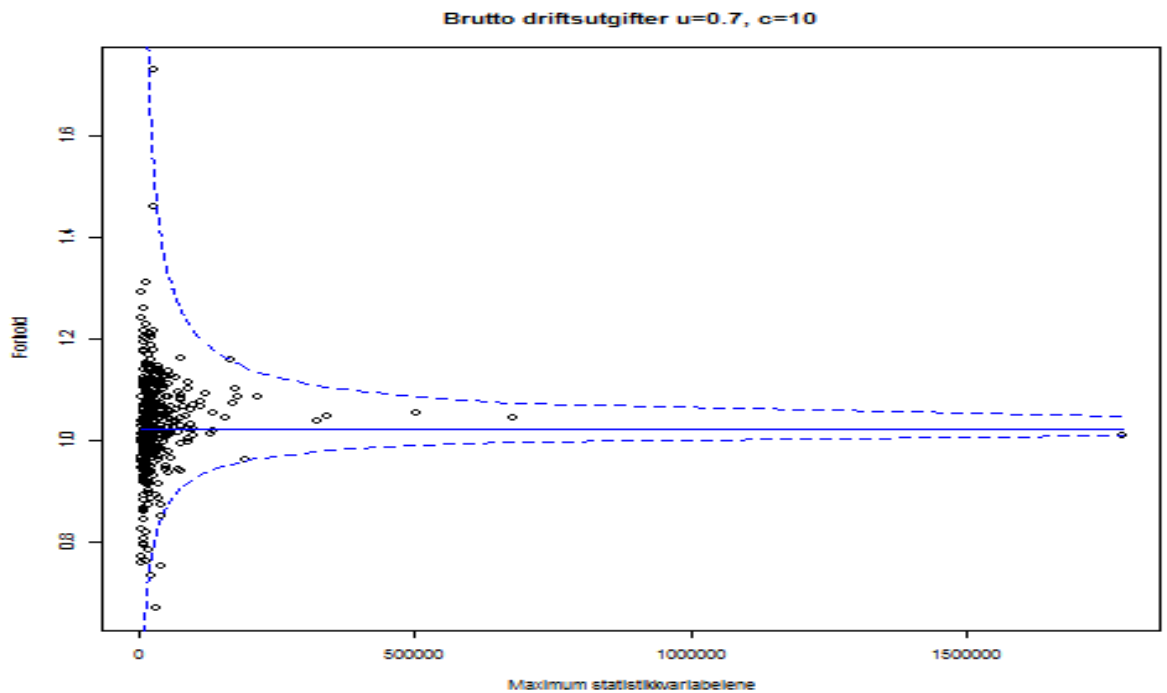


Input parametre er:
- data Input of Hb er et datasett av type dataframe.
- id Navn på en identifikasjonsvariabel.
- x1 Navn på variabel i periode t.
- x2 Navn på variabel i periode t-1.
- pU Parameter som justerer for forskjellige nivåer av variablene. Default verdi 0,5. pU<1
- pA Parameter som justerer 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
Forhold
0.8
1.0
1.2
1.4
1.6
0
500000
1000000
1500000
Maximum statistikkvariabelene


Brutto driftsutgifter u=0.5, c=10
Forhold
0.8
1.0
1.2
1.4
1.6
0
500000
1000000
1500000
Maximum statistikkvariabelene


Brutto driftsutgifter u=0.3, c=10
Forhold
0.8
1.0
1.2
1.4
1.6
0
500000
1000000
1500000
Maximum statistikkvariabelene


Brutto driftsutgifter u=0.7, c=10
Forhold
0.8
1.0
1.2
1.4
1.6
0
500000
1000000
1500000
Maximum statistikkvariabelene

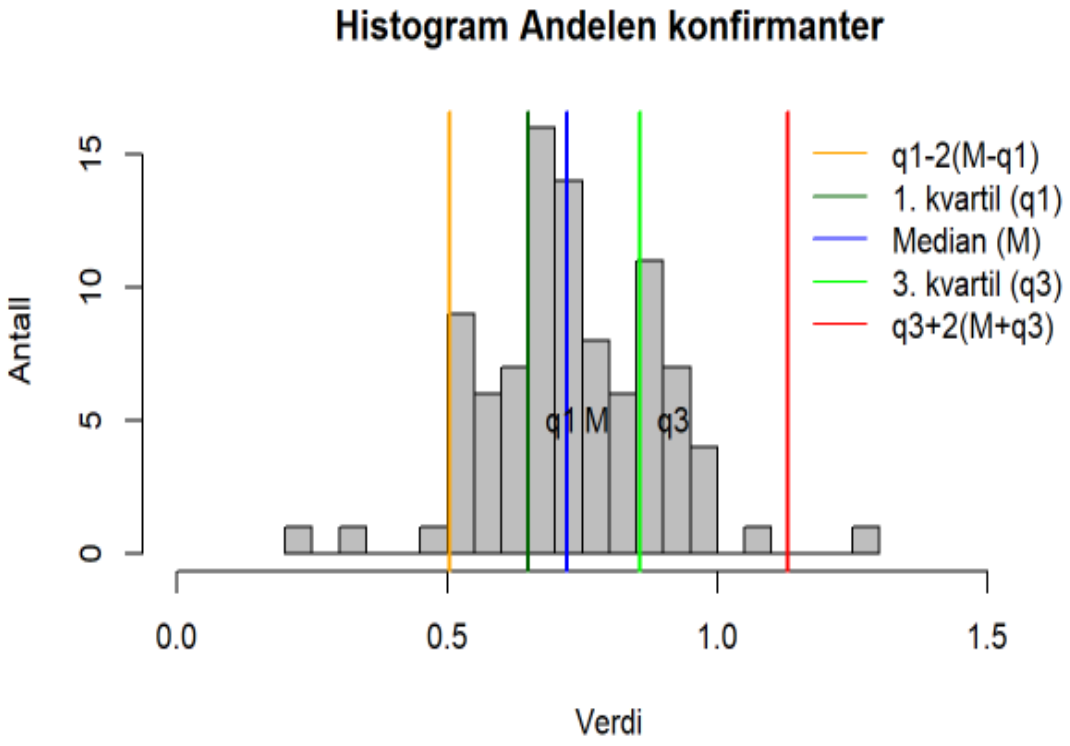
Øvelser tusenfeil og HB metoden

- Oppgave 5, 6 og 7
  - Bruk av variabelen døpte
  - Kirkedata_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!

Kvartilmetode

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

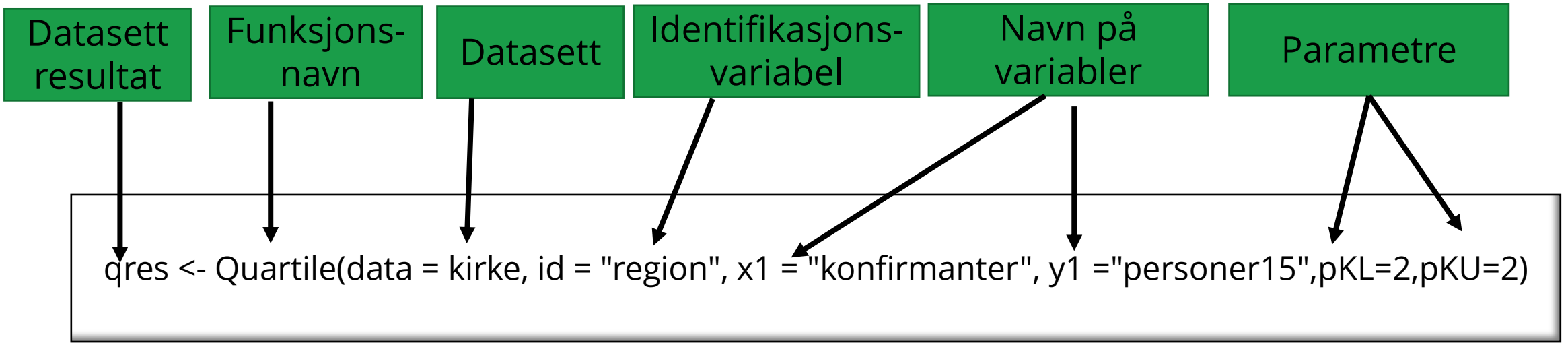
$(q_{0.25}-k_1(q_{0.5}-q_{0.25}),\ q_{0.75}+k_2(q_{0.75}-q_{0.5}))$



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



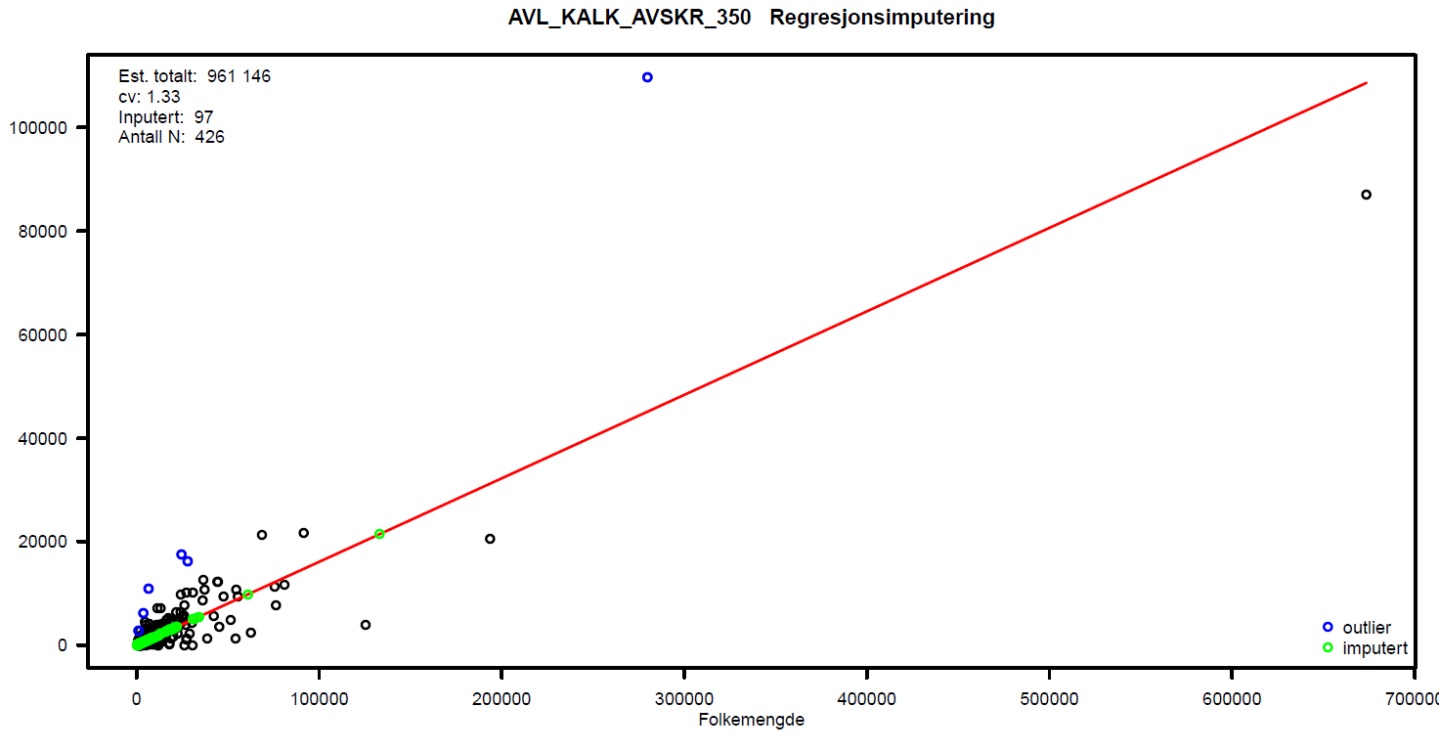
Input parametre er:
- data Input til Quartile er et datasett med klassedata.frame.
- id Navn på identifikasjonsvariabel.
- x1 Navn på x-variabel i periode t.
- y1 Navn på y-variabel i periode t.
- x2 Navn på x-variabel i periode t-1. Valgfri
- y2 Navn på y-variabel i periode t-1. Valgfri
- strataName 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 -pKL * (M - q1), q3 + pKU * (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 - OutlierRegression

- 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.



Statistisk sentralbyrå
Statistics Norway

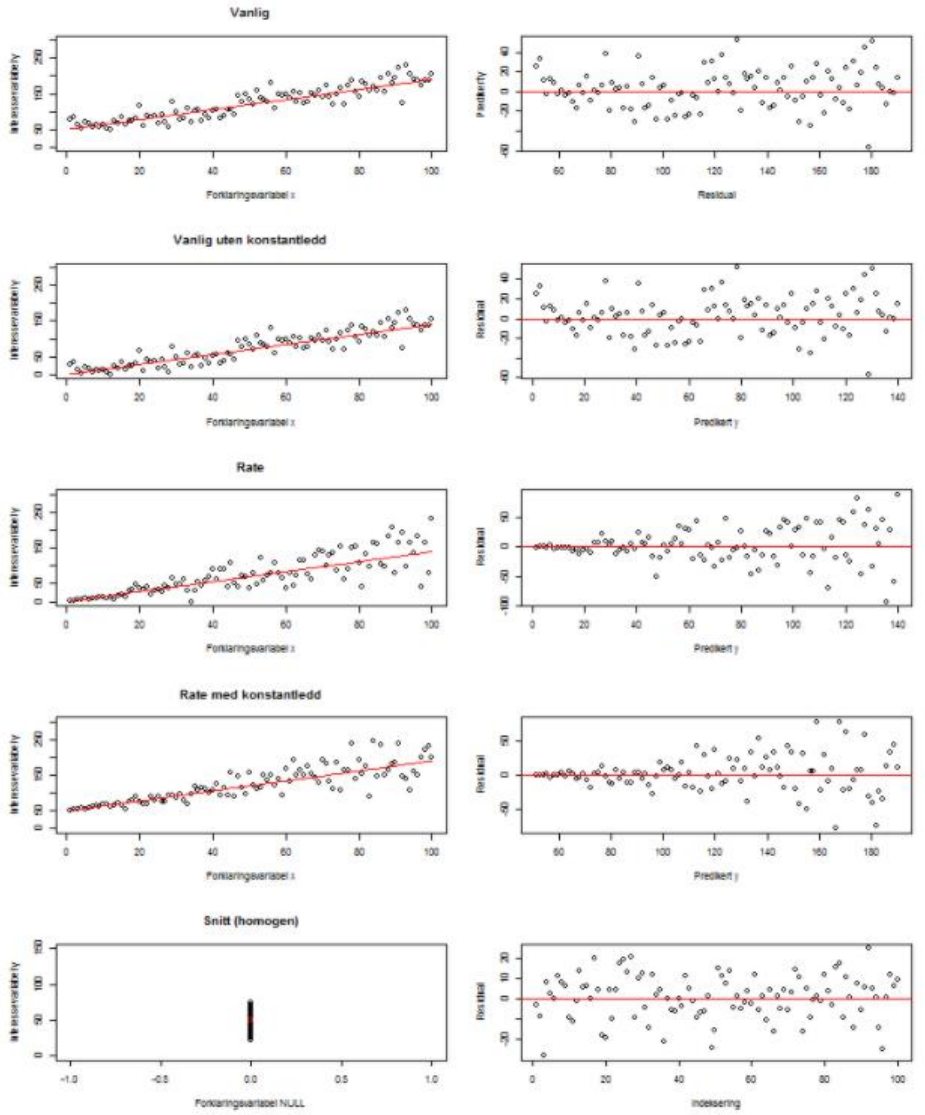
Robust regresjon

- Vi har en multivariat versjon liggende som kan brukes

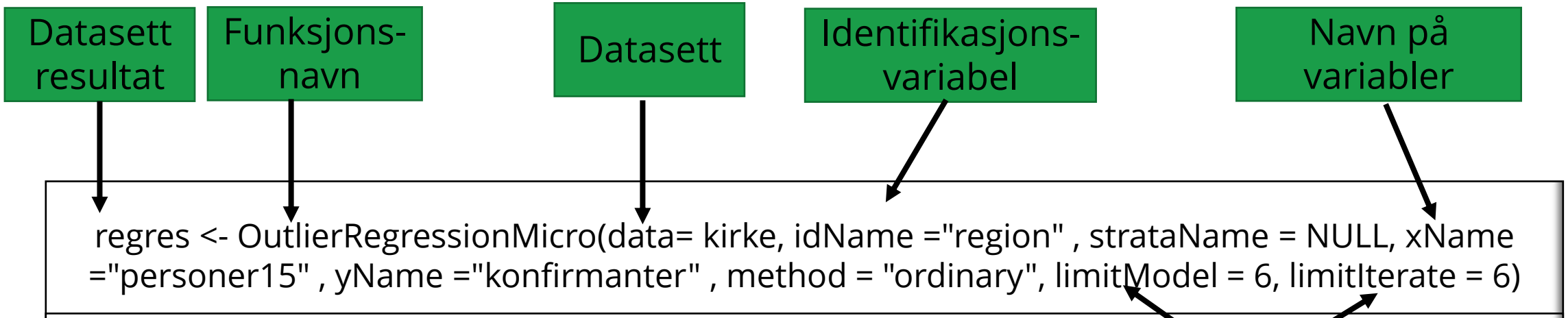
- Kan kjøres innen grupper
  - homogene grupper som ligner på hverandre
  - Må være nok observasjoner innen hver gruppe
  - Grupper kalles ofte strata av statistikere

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



Input parametre er:

- data - Input datasett m
- dName - Navn på ID-variabel (er)
- strataName - Navn på starta-variabel. Enkeltlag når NULL (standard)
- xName - Navn på x-variabel
- yName - Navn på y-variabel
- method Metoden (modell og vekt) kodet som en streng:
  - ordinary" (standard)
  - "ratio«
  - "noconstant«
  - "mean"
  - "ratioconstant".
- limitModel - Studentiserte residual grenser. Over grensen -> outlier.
- limitIterate - Studentiserte residual grenser for iterativ beregning av studentiserte residualer.

- 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

Øvelser kvartilmetode og regresjon

- Oppgave 8 og 9
  - Bruk av variabelen døpte
  - Kirkedata_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!

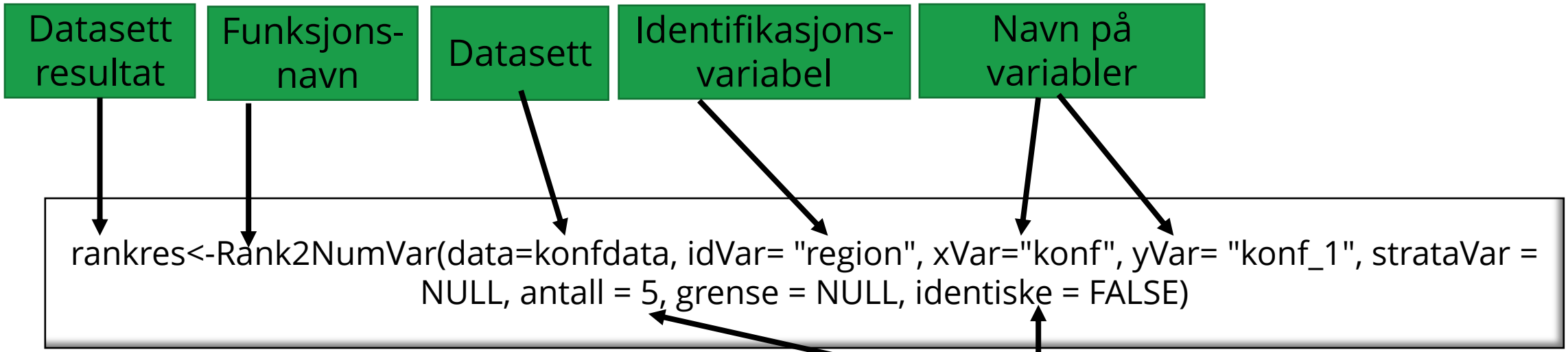
Innflytelse

- Kontroll metoder:
  - Enhetens andel av totalen - Rank2NumVar
  - Enhetens andel av endringstall – Diff2NumVar

Rank2NumVar - 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



Input parametre er:
- data Input datasett
- idVar Navn på en identifikasjonsvariabel.
- xVar Navn på x-variabelen som skal sammenlignes.
- yVar Navn på y-variabelen som skal sammenlignes.
- strataVar Navn på stratifiseringsvariabel. Valgfri. Hvis strataVar er gitt, blir beregningen og notering utføres innenfor hvert stratum.
- antall Parameter som spesifiserer hvor mange av de største enhetene som skal vises. Standard 5.
- grense Parameter som spesifiserer en terskel for enhetene som skal vises. Denne parameteren overstyrer ANTALL. Valgfri.
-identiske Når TRUE brukes bare enheter med verdi på både x og y i beregningene.
Standard FALSE.

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

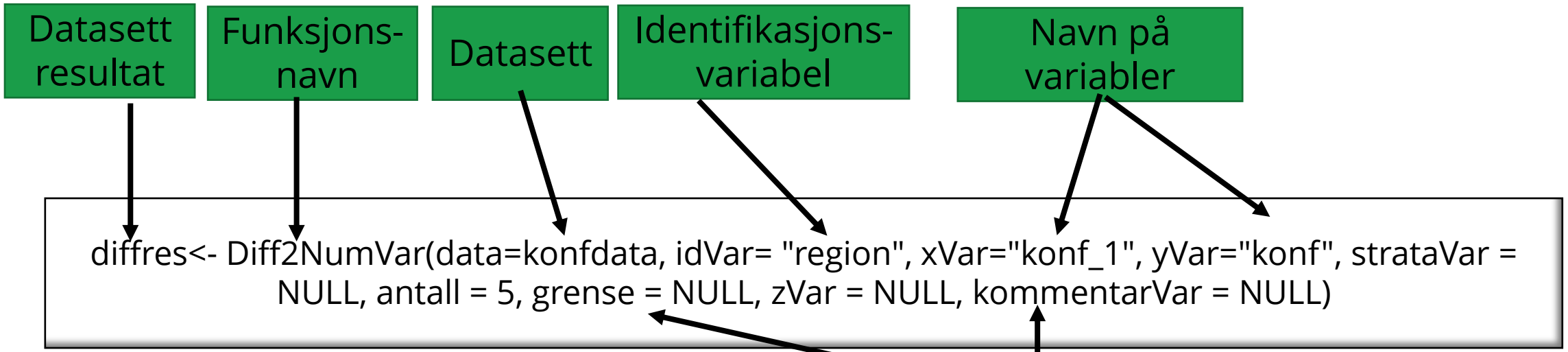
id <chr>	x <int>	y <int>	strata <chr>	forh <dbl>	xRank <int>	yRank <int>	›
1201	1668	1740	1	1.0431655	1	1	
0301	1645	1722	1	1.0468085	2	2	
5001	1046	1085	1	1.0372849	3	3	
1103	930	793	1	0.8526882	4	4	
0219	743	758	1	1.0201884	5	5	

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



Input parametre er:
- data Input datasett
- idVar Navn på en identifikasjonsvariabel.
- xVar Navn på x-variabelen som skal sammenlignes.
- yVar Navn på y-variabelen som skal sammenlignes.
- strataVar Navn på stratifiseringsvariabel. Valgfri. Hvis strataVar er gitt, blir beregningen og notering utføres innenfor hvert stratum.
- antall Parameter som spesifiserer hvor mange enheter med den største forskjellen som skal vises. Standard 5.
- grense Parameter som spesifiserer en terskel for enhetene som skal vises. Denne parameteren overstyrer antall. Valgfri.
- zVar Navn på den opprinnelige y-variabelen, før du redigerer. Valgfri.
- kommentarVar Navn på en variabel som gir informasjon om redigering. Valgfri

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 Inngangsidentifikasjonsvariabelen

- 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: | Diff |

- DiffProsAvx Forskjellen i prosent av x: (Diff / x) * 100

- DiffProsAvSumx Forskjellen i prosent av stratum totalt for x: (Diff / stratum x) * 100

- DiffProsAvTotx Forskjellen i prosent av totalen for x: (Diff / total x) * 100

- SumDiffProsAvSumx Stratumforskjellen i prosent av stratum totalt for x: ((stratum y - stratum x) / stratum x) * 100

- SumDiffProsAvTotx Stratumforskjellen i prosent av totalen for x: ((stratum y - stratum x) / totalt x) * 100

- z Variabelen input z

- Endring Forskjellen mellom z og y: y - z

- KommentarVar Input kommentar-variabelen

Øvelse

- Oppgave 10-11
  - Bruk av variabelen døpte
  - Kirkedata_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 statistkkbyrået.

(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.

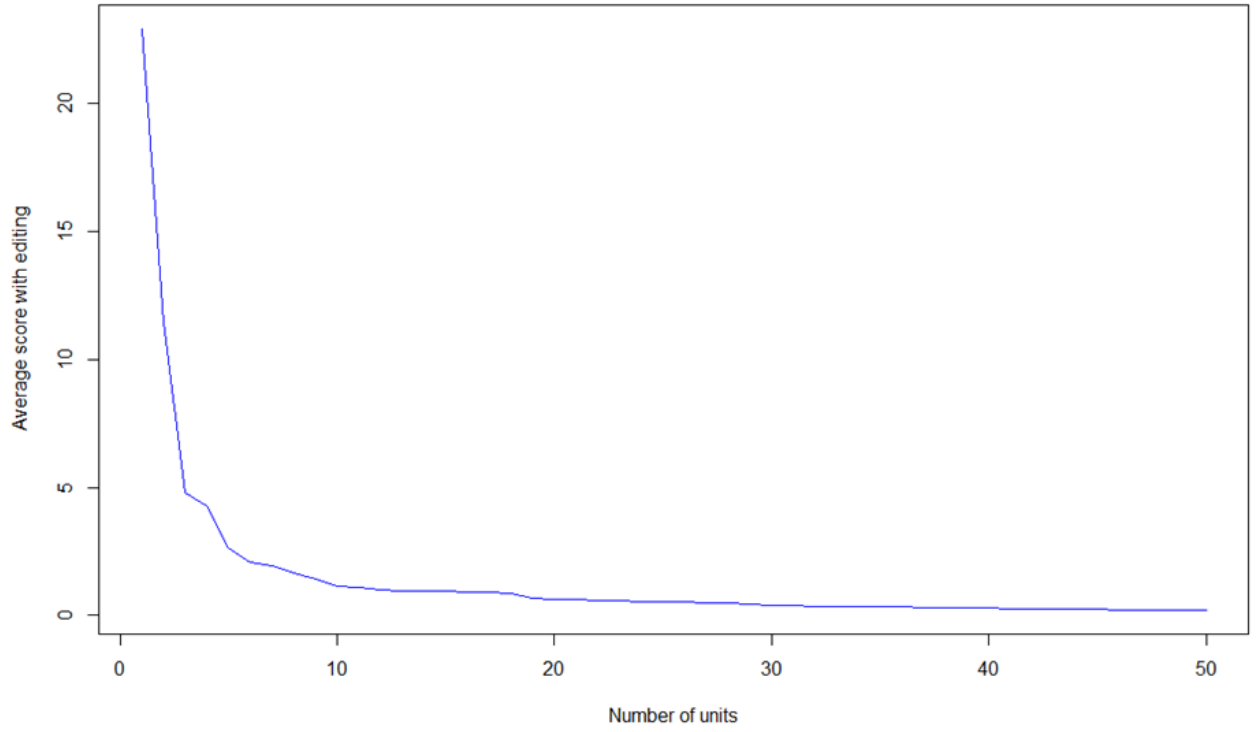
$$score(\mathrm{i}) = \sum_{v} \frac{|y_{i,v,t} - y_{i,v,t-1}|}{\sum_i y_{i,v,t-1}} \frac{1}{Antall\ v}$$

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

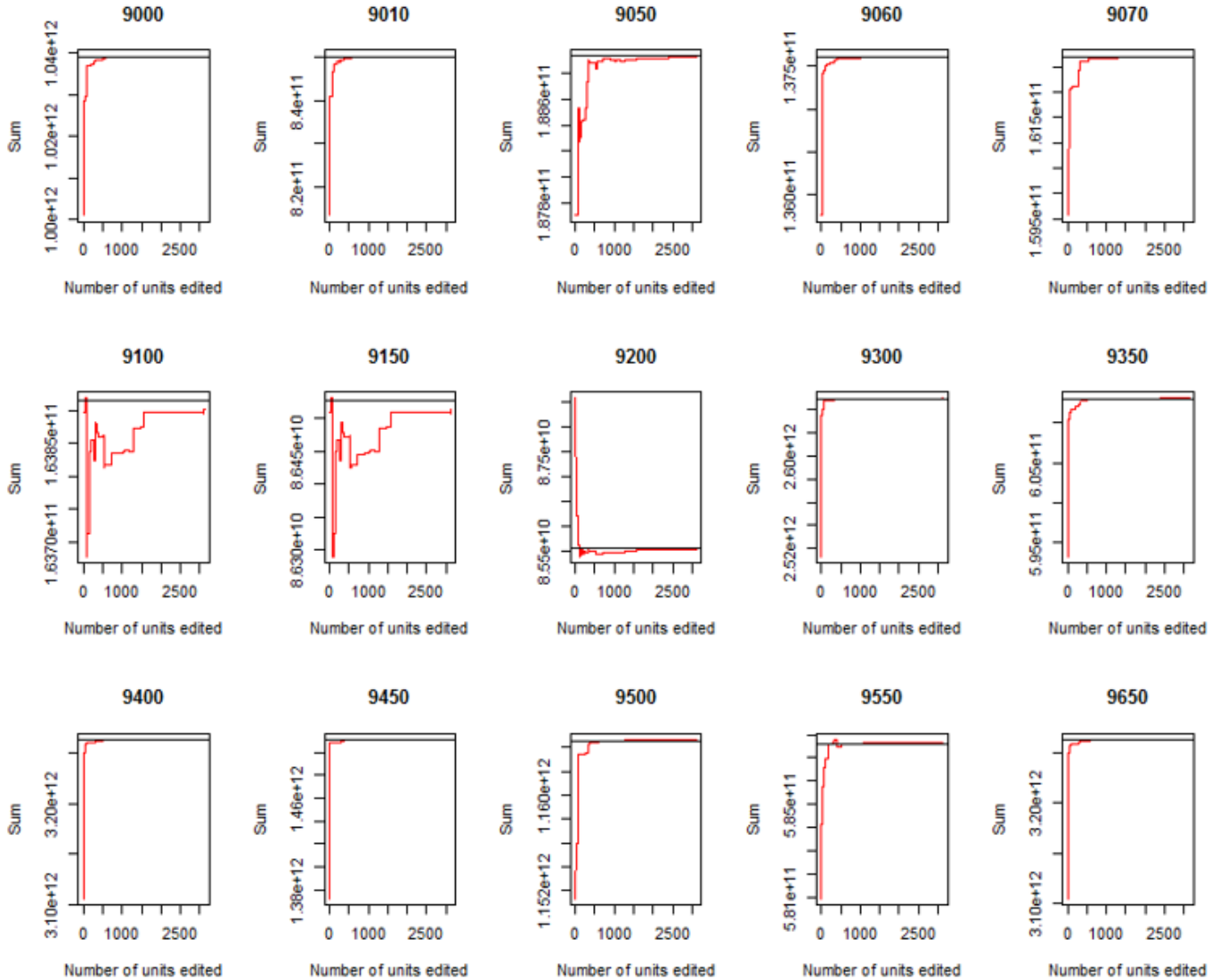
Offentlig eide foretak

- 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

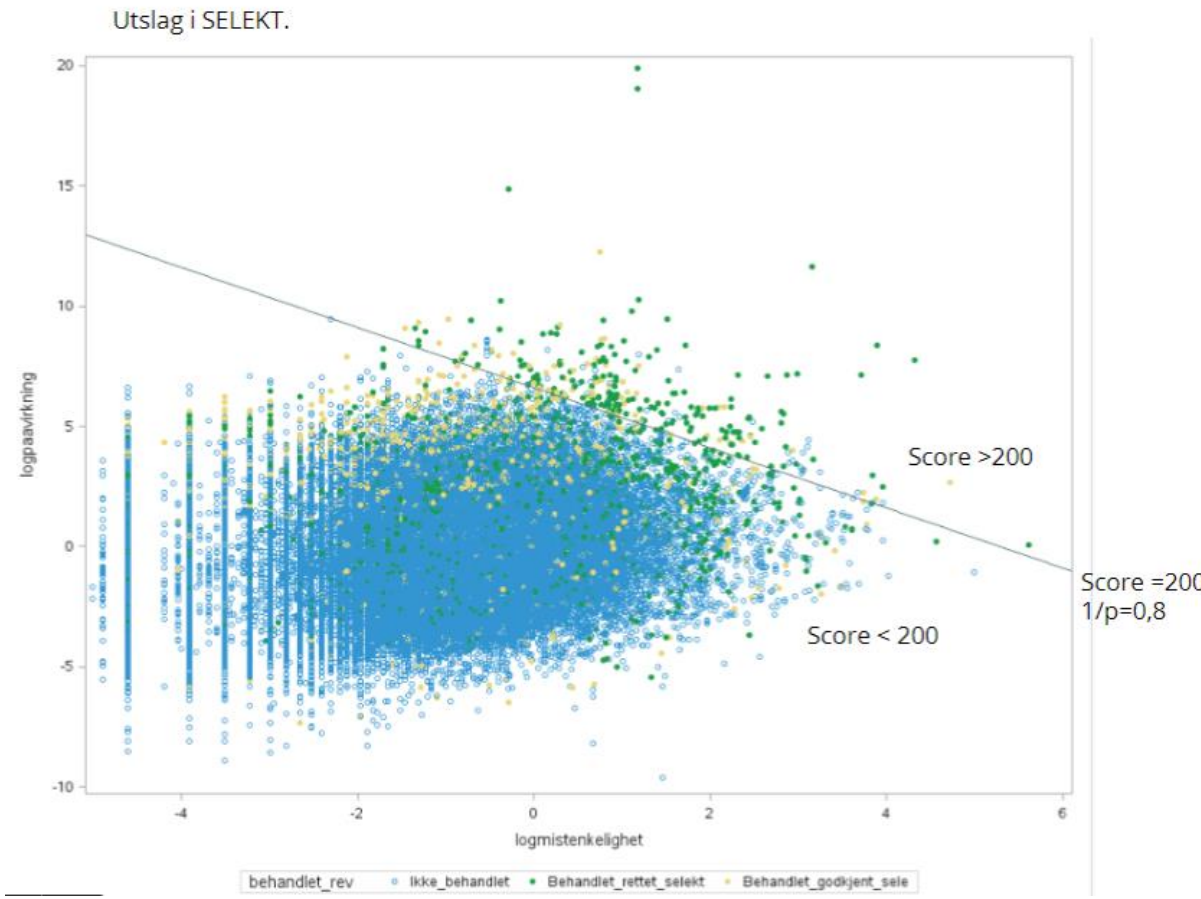


Selekt - poengfungsjon

- Poengfunksjonen beregnes som et produkt av:
  - Mål for mistenkelighet
  - Mål for potensiell påvirkning for flere grupper
  - Parameteren p vekter viktigheten av potensiell påvirkning

- Score=mistenkelig x (potensiell påvirkning)^p

- Scoreverdien brukes til å prioritere oppfølging av utslagene i synkende rekkefølge.



Statistisk sentralbyrå
Statistics Norway

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.

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

albyrå

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.

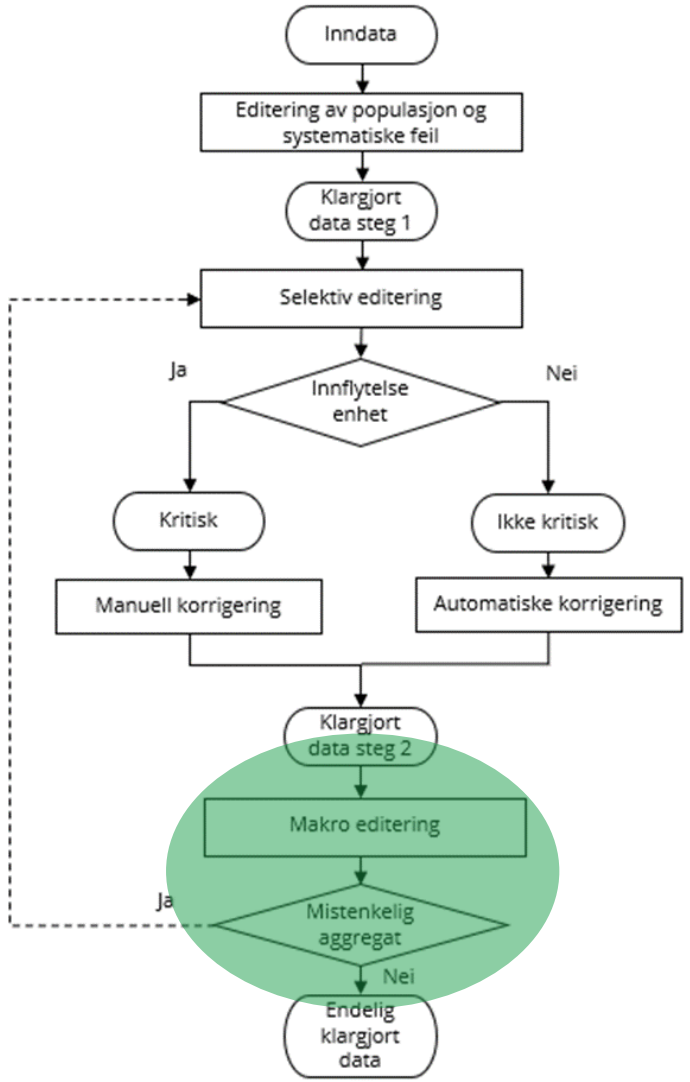
$$Potential\ impact_i = \underset{over\ v=1-5}{\text{maximum}} \left\{ \frac{\left| Invoiced\ value_i - Quantity_i \cdot UP_{Q2}(i) \right|}{\sum_{k \in g_v} Invoiced\ value_k^*} \cdot \frac{1}{O_v} \cdot f^{\,10\log\left(\sum_{k \in g_v} Invoiced\ value_k^* \right)} \right\}$$

TALL SOM FORTELLER

Makro editering

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.



Statistisk sentralbyrå
Statistics Norway

Makroeditering - Retningslinjer

• Det skal vurderes om aggregatene er plausible gitt historisk forløp

• Plausibiliteten av aggregatet skal vurderes i lys av utviklingen i avledede størrelser, for eksempel forholdstall.

• Ved endringer utenom det som anses som normalt skal datagrunnlaget bli undersøkt for å enten finne den potensielle feilen eller kunne forklare endringen.

• Aggregatene bør vurderes opp mot tilsvarende aggregater for sammenlignbare land.

• Aggregater kan være gjenstand for editering når de inngår i et system der visse relasjoner må holde

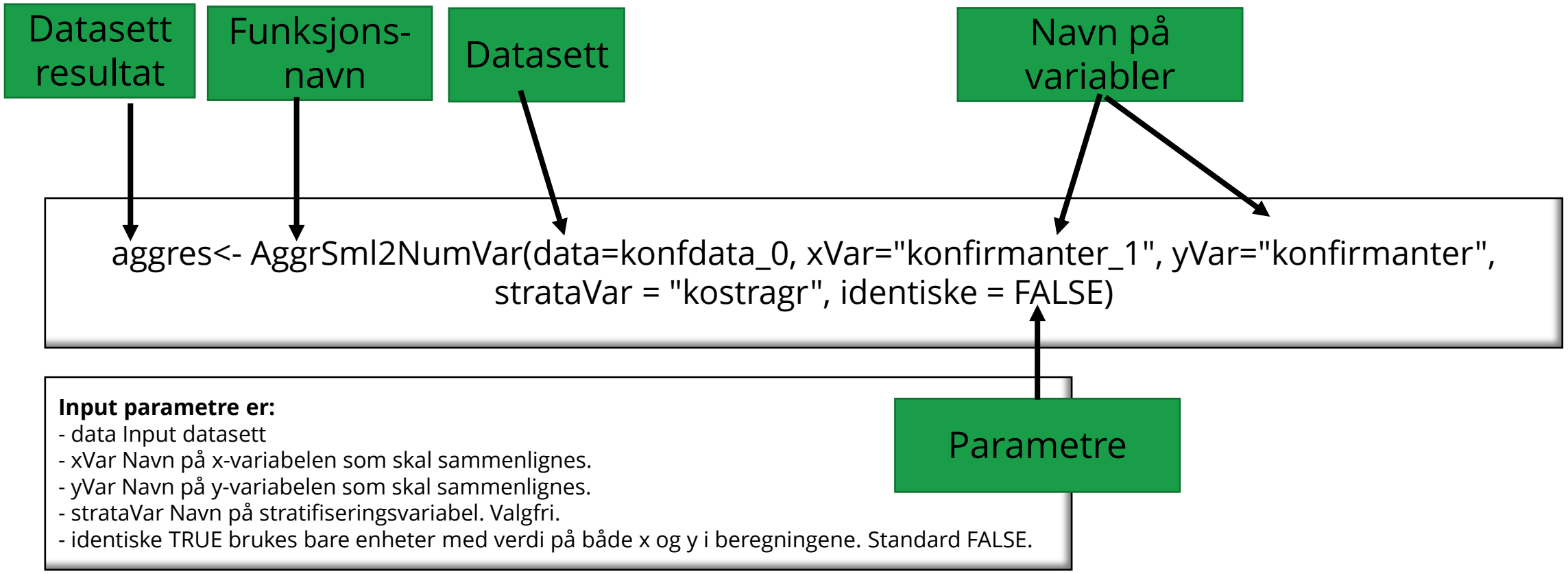
AggrSml2NumVa - Analyse av aggregat

strata <chr>	Antx <int>	Anty <int>	Sumx <int>	Sumy <int>	SumxProsAvTotx <dbl>	SumyProsAvToty <dbl>	Diff <int>	AbsDiff <int>	DiffProsAvSumx <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



Input parametre er:
- data Input datasett
- xVar Navn på x-variabelen som skal sammenlignes.
- yVar Navn på y-variabelen som skal sammenlignes.
- strataVar Navn på stratifiseringsvariabel. Valgfri.
- identiske TRUE brukes bare enheter med verdi på både x og y i beregningene. Standard FALSE.

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: (Sumx / Totx) *100

- SumyProsAvToty Stratumet totalt for y i prosent av befolkningen totalt for y: (Sumy / Toty) * 100

- Diff Forskjellen mellom stratum totalt av x og y: Sumy - Sumx

- AbsDiff Den absolutte forskjellen: | Diff |

- DiffProsAvSumx Forskjellen i prosent av stratum totalt for x: (Diff / Sumx) * 100

- AbsDiffProsAvSumx Den absolutte verdien av DiffProsAvSumx: | DiffProsAvSumx |

- DiffProsAvTotx Forskjellen i prosent av befolkningen totalt for x: (Diff / Totx) * 100

- AbsDiffProsAvTotx Den absolutte verdien av DiffProsAvTotx: | DiffProsAvTotx |

Dashboard makrokontroller

12293: Omsorgstjenester - supplerende nøkkeltall

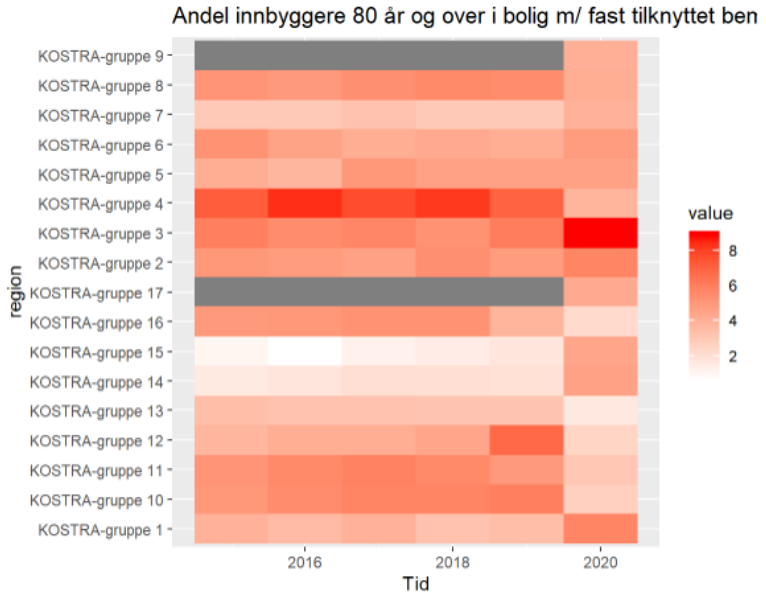
Landet | Fylke per variabel | Kostragruppe per variabel | Fylke samlet siste år | Robust multivariat oppdagelse av uteliggere

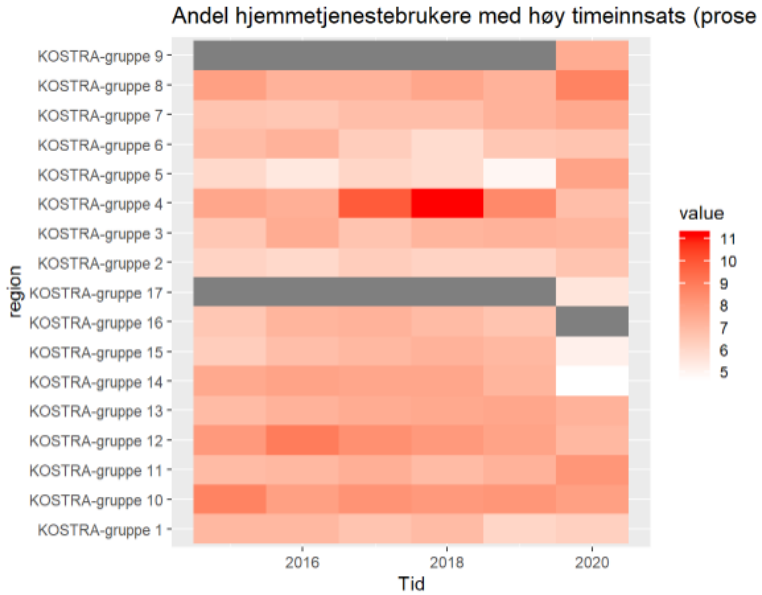
	statistikkvariabel	2015	2016	2017	2018	2019	2020	TrendSparkline
	All	/	/	/	/	/	/	All
1	Andel beboere 80 år og over i bolig m/ fast tilknyttet bemanning hele døgnet (prosent)	34.9	33.8	34.9	34.5	34.2	33.6	
2	Andel beboere i institusjon av antall plasser (belegg) (prosent)	98.8	98.8	98.1	0	0	0	
3	Andel 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	

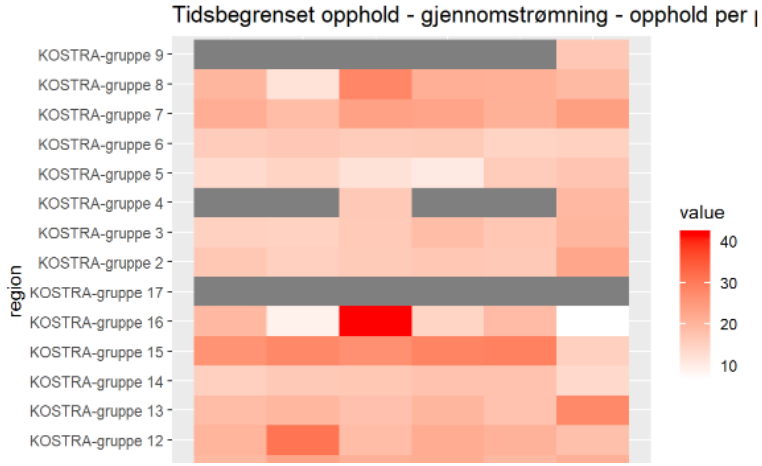
Statistisk sentralbyrå
Statistics Norway

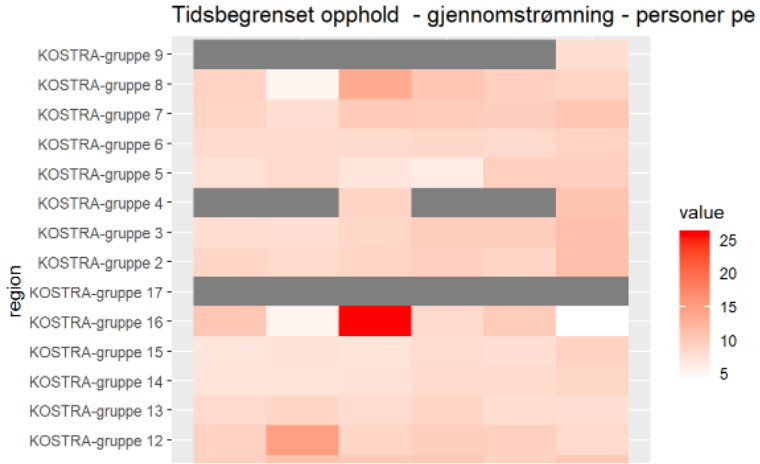
12293: Omsorgstjenester - supplerende nøkkeltall

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









Statistisk sentralbyrå
Statistics Norway

Eksempler i R

Øvelse

- Oppgave 12
  - Bruk av variabelen døpte
  - Kirkedata_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!

Læringsmålet

- Kjenne til **SSB** sine prinsipper og retningslinjer for dataeditering
- Kjenne til **Eurostats** metodikk for validering
- Kjenne til **UNECE** sin modell for dataeditering
- Vite hvilke **logiske kontroller** som er anbefalt og kunne sette de opp
- Forstå **selektiv editering** basert på mistenkelige og innflytelse av verdier
- Forstå og bruke **metodene**: HB, kvartilmetode, regresjon og forskjellige mål på innflytelse

TALL SOM FORTELLER

Takk!

Statistisk sentralbyrå
Statistics Norway