TALL SOM FORTELLER

Kurs i dataeditering: Validere og kontrollere

ASLAUG HURLEN FOSS 20.04.2021

Statistisk sentralbyrå
Statistics Norway

Plan for kurset

- 09:00-09:40 Validering av numeriske og kategoriske verdier
- 09:40-10:10 Øvelse i R
- 10:10-10:20 Kaffepause
- 10:20-11:00 Kvartilmetode, HB-metode, regresjon og innflytelse
- 11:00-11:30 Øvelse i R
- 11:30-12:00 Gjennomgang av øvelse og oppsummering

Praktisk gjennomføring

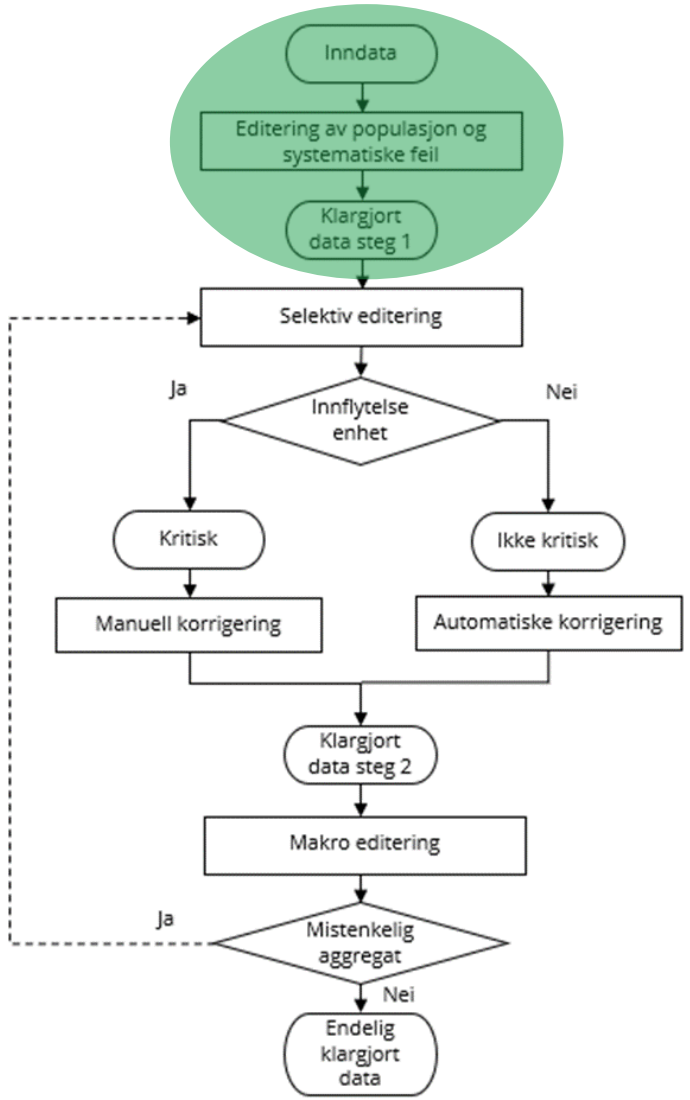
- Er det greit med opptak av kurset på video?
- Hvis du vil stille spørsmål eller gi en kommentar, si fra i chatten!
- Sett mikrofonen på lydløs når du ikke snakker

Læringsmålet

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

Validering og prosessmodell

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

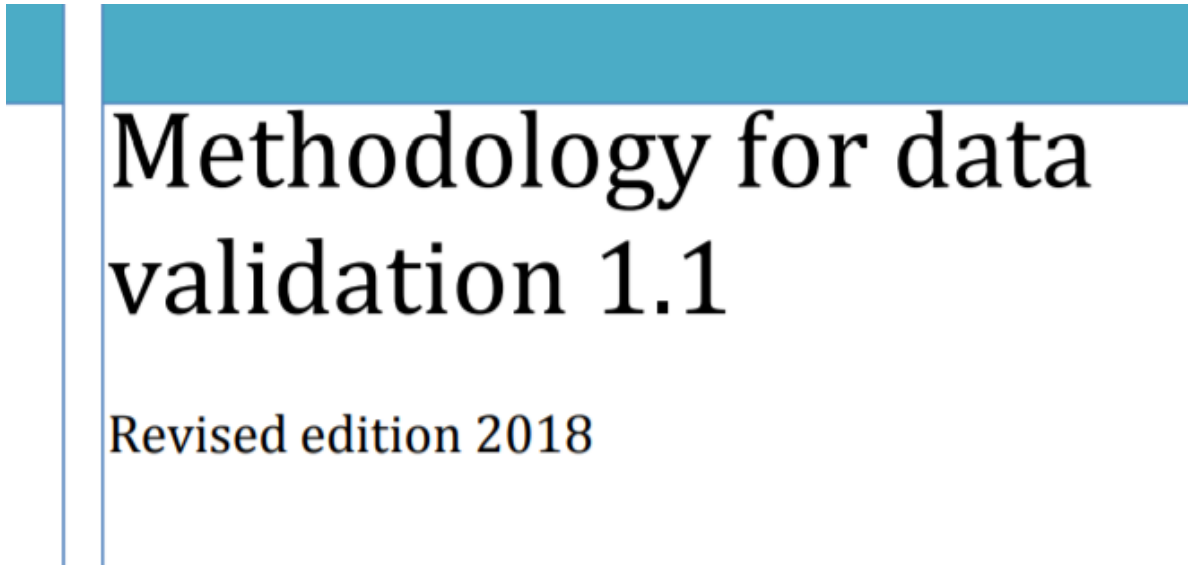


Statistisk sentralbyrå
Statistics Norway

Eksempler

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

Håndbok i datavalidering



- Methodology for data validation, EUROSTAT https://ec.europa.eu/eurostat/cros/content/ess-handbook-methodology-data-validation-v11-rev2018-0_en

Nivåer for kontroller

- Innen en enhet (record)
- Innen et datasett
- Mellom datasett
- Konsistens sjekker 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

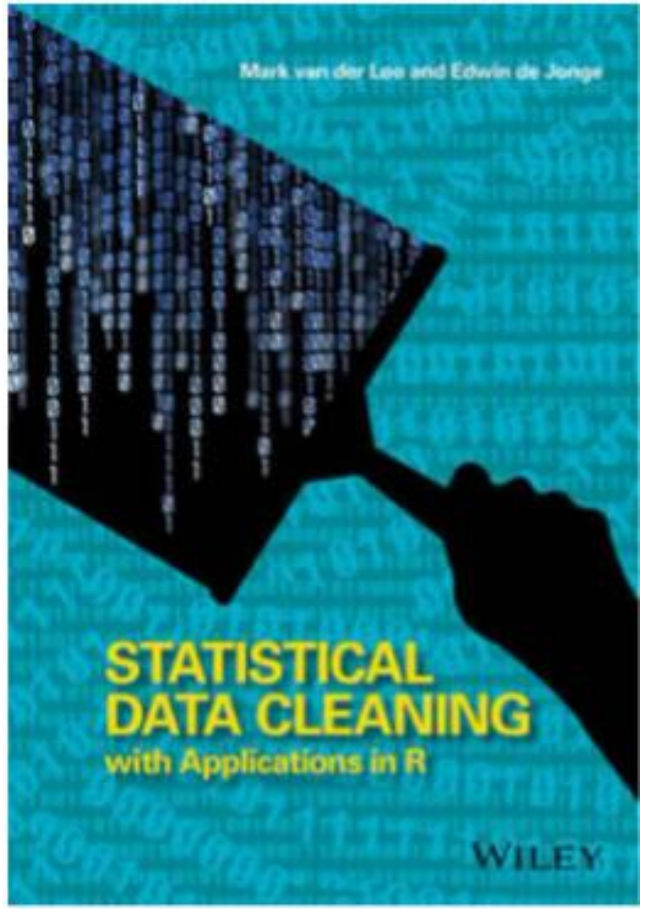
Kvalitetsindikatorer – kontrollere og validere

Indikator	Kommentarer
Utslagsrate - 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

Validate pakke i R

- Valideringspakken er ment å lage:
  - Sjekke data lett
  - Vedlikeholde kontrollene enkelt
  - Mulig å reprodusere resultatene



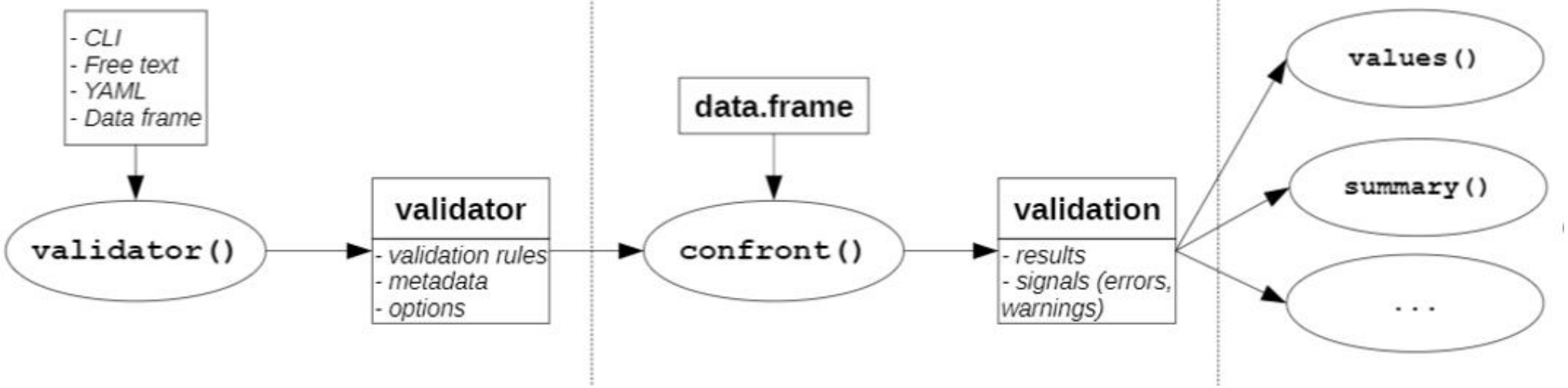
- Bygget av Mark van der Loo and Edwin de Jonge, Statistics Netherlands
- https://cran.r-project.org/web/packages/validate/vignettes/introduction.html

Validate pakken

Sette opp og vedlikeholde kontroller

Kjøre kontroller på data

Analyse av kontroller

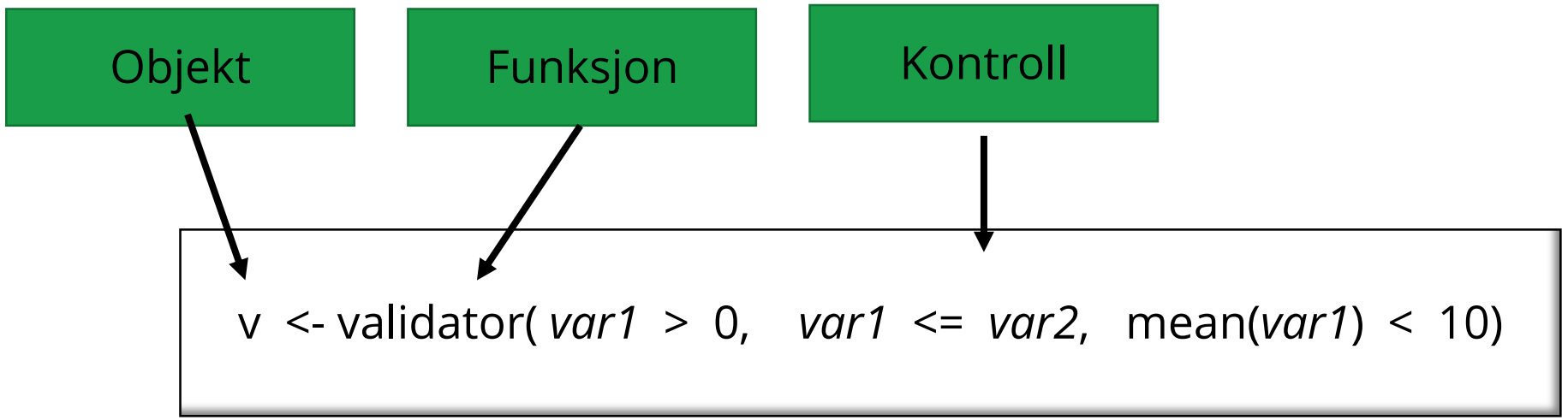


Datasett

- Lager eksempel datasett 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




```
> 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.

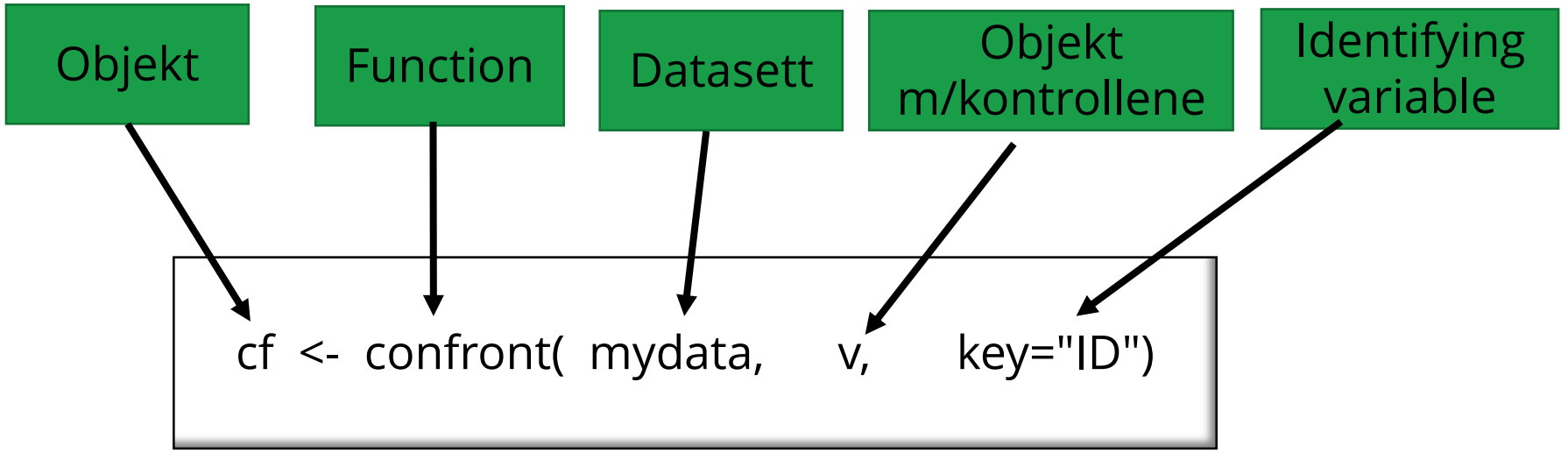
Spesialfunksjoner

- 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)

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 å 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 indikators

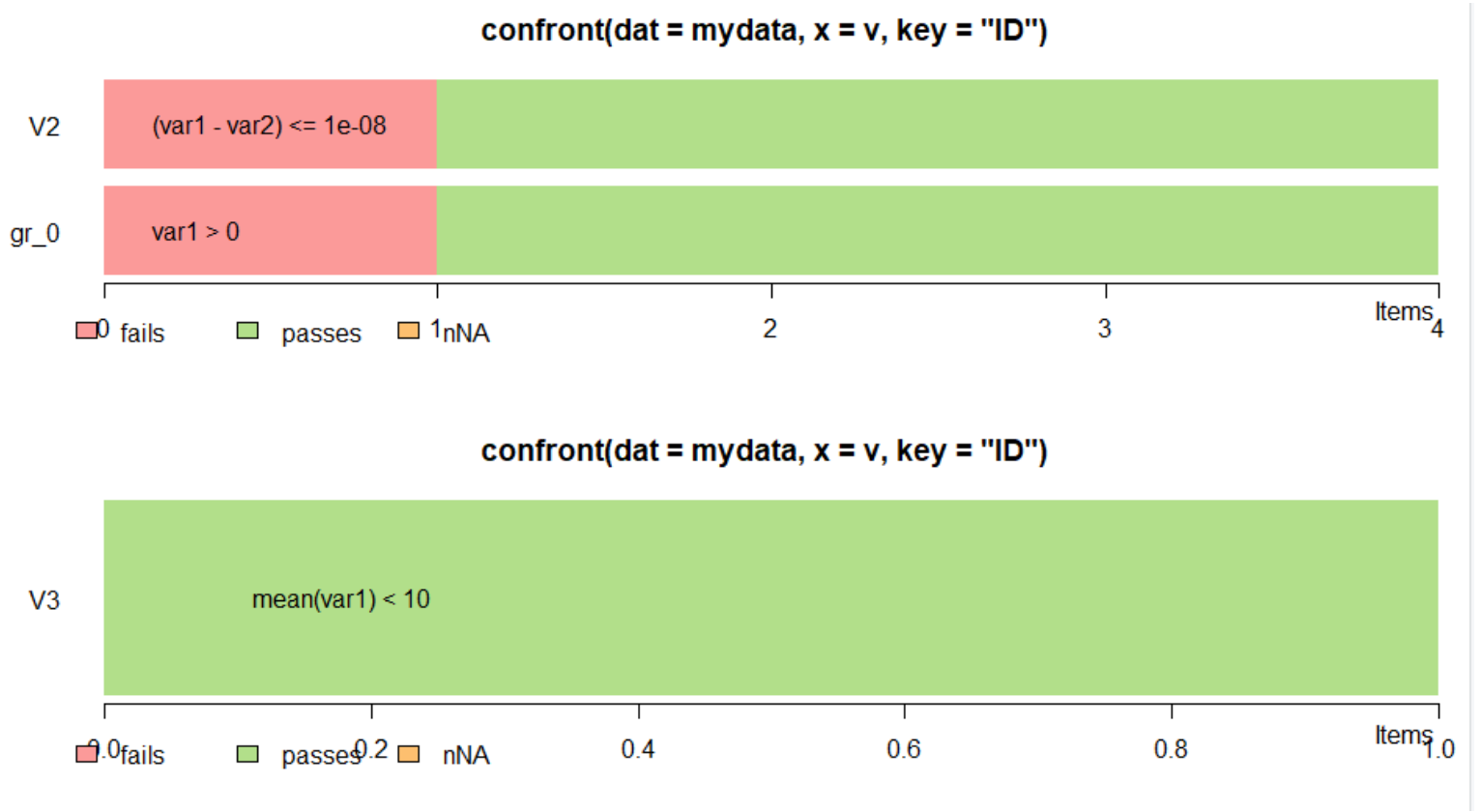
```
ind<-as.data.frame(values(cf))

# 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

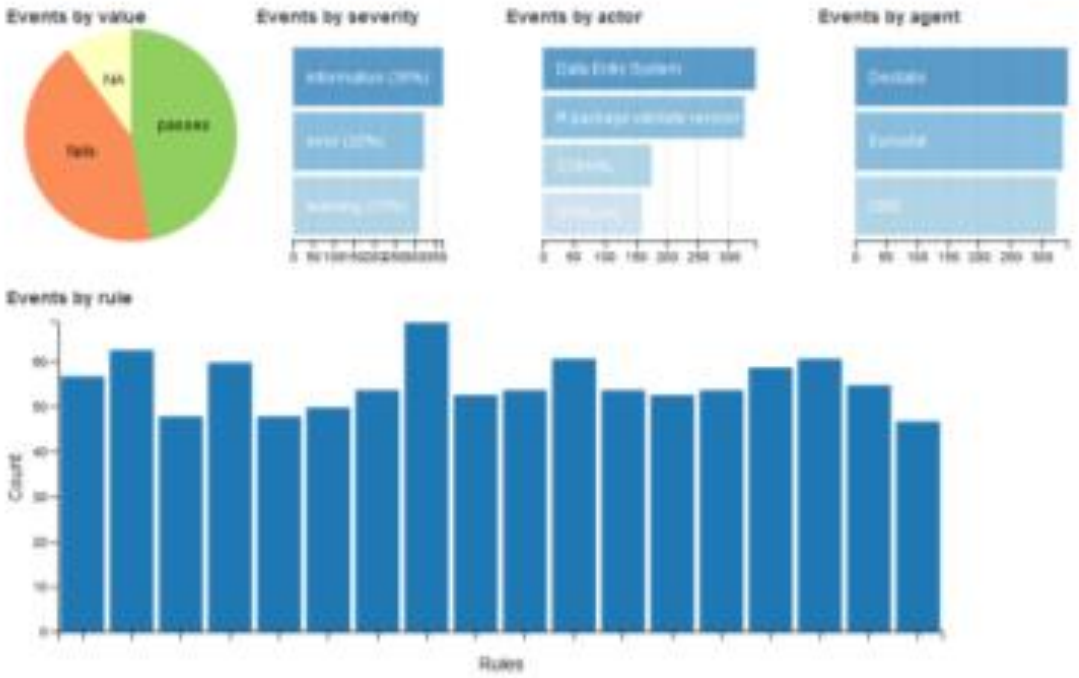
Grafikk

plot(cf)



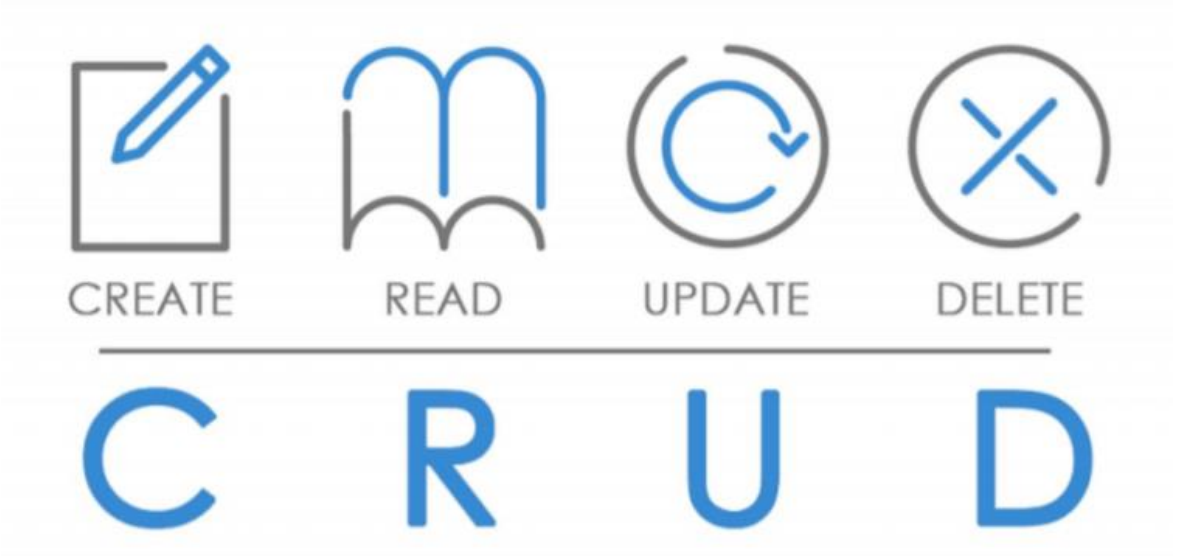
Nederland lager dashboard for

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!



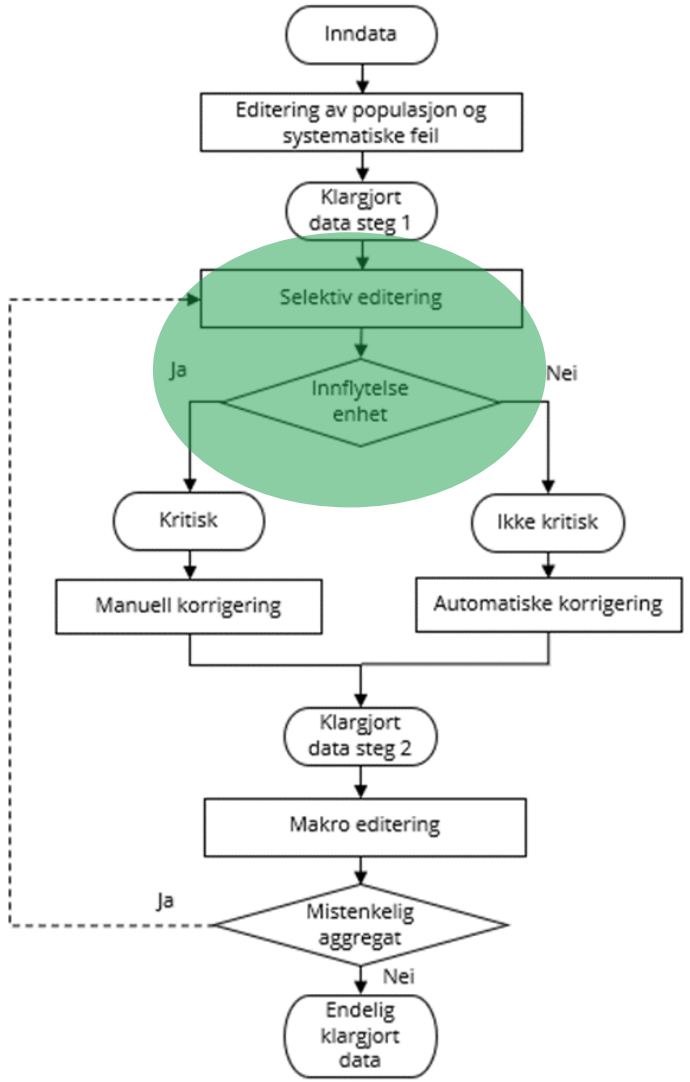
Eksempel og øvelse

- Oppgave 1-4 pluss ekstra oppgave
  - Kirkedata med variabelen døpte
  - Last inn r-pakken validate
- Oppgavene ligger på
  - teams <>
  - Github https://github.com/statisticsnorway/R_kontrollfunksjoner
- Neste leksjon starer 10.20

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



Statistisk sentralbyrå
Statistics Norway

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.

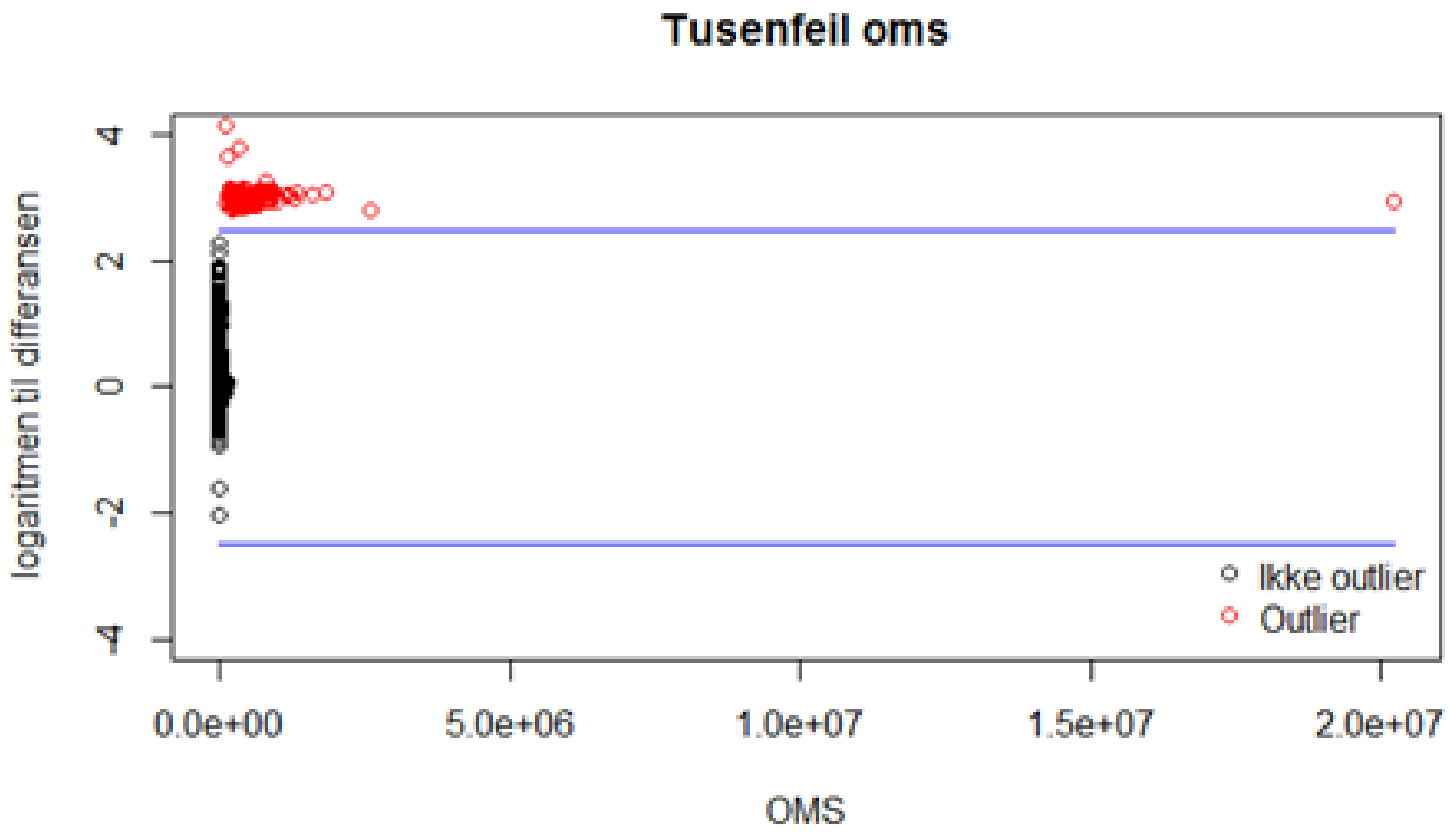
Tusenfeil

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

Tusenfeil: Siffermetoden - logaritmen

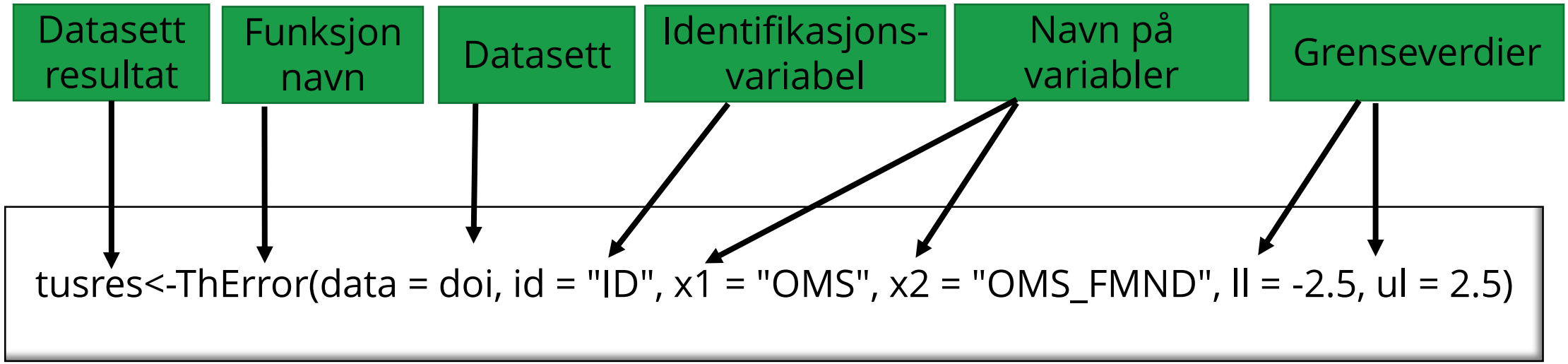
- I siffermetoden teller vi forskjell i antall siffer mellom årets og forrige års verdi, dette gjør vi ved hjelp av den matematiske funksjonen logaritmen.
- Hvis det er en forskjell på 3 siffer er det en tusenfeil, 6 siffer millionfeil osv.
- Metoden fungerer bare for verdier som er større enn null og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

Eksempel: DOI- detaljomsetningsindeksen



Statistisk sentralbyrå
Statistics Norway

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

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

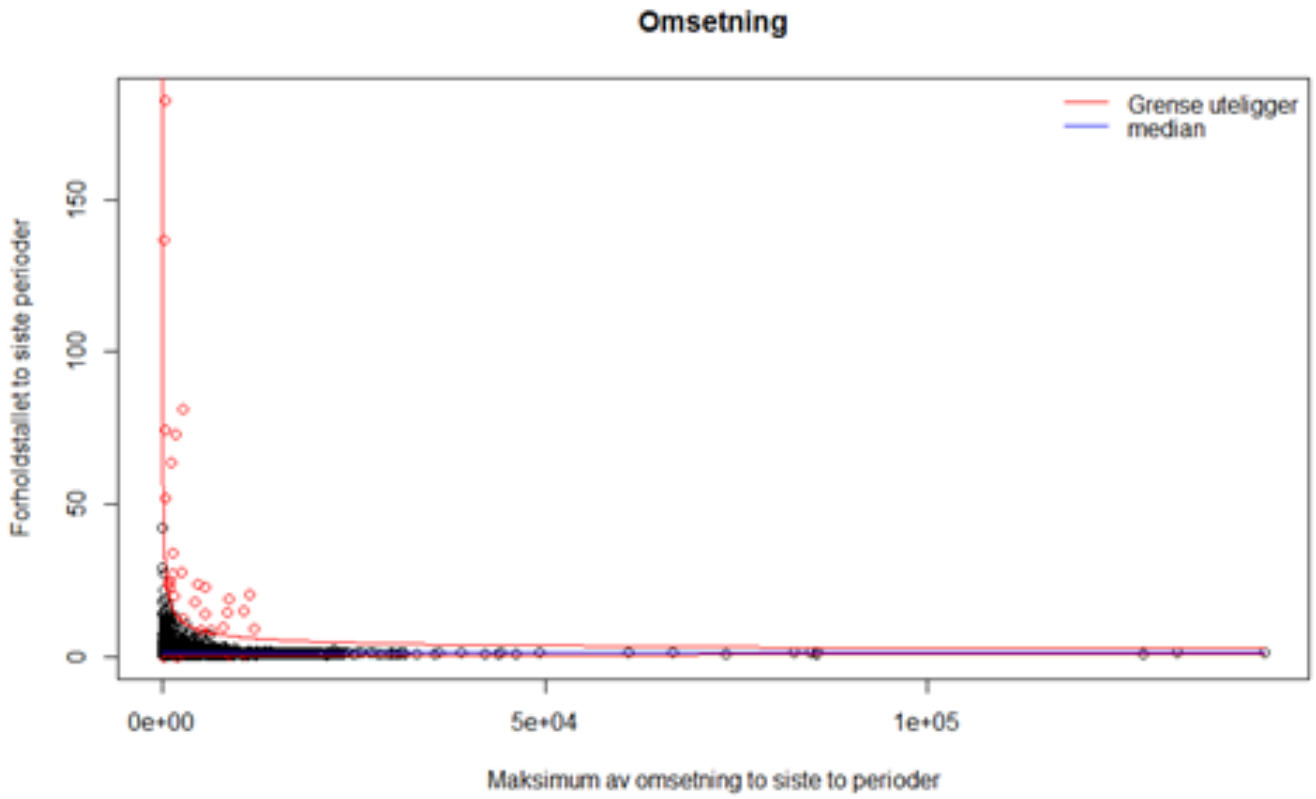
$$E_i = S_i * \left(MAX\left(X_i(t-1), X_i(t)\right)\right)^U, \quad 0 \leq U \leq 1$$

Akseptgrenser

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

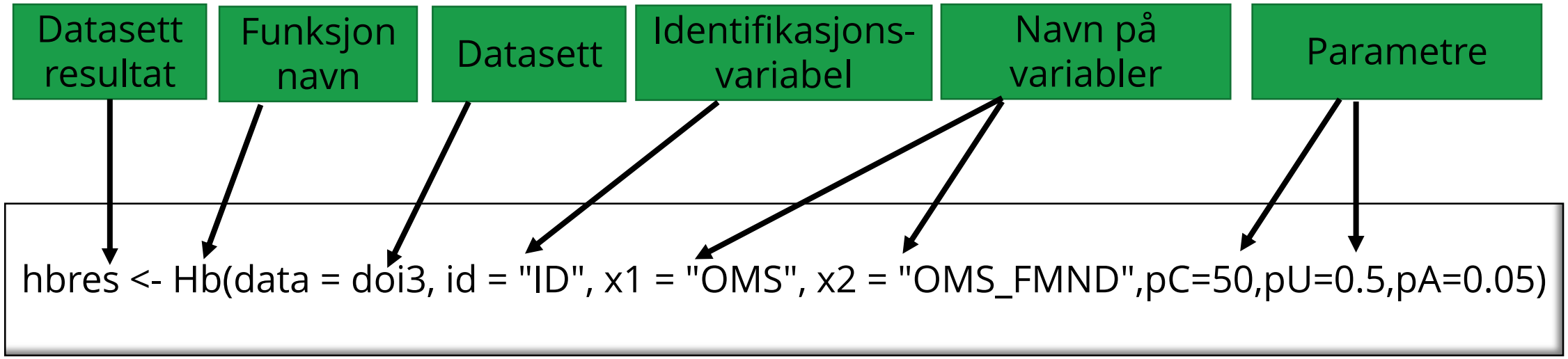
$$\left\{ E_{median} - C * D_{Q1} \quad , \quad E_{median} + C * D_{Q3} \right\}$$

Eksempel: DOI- detaljomsetningsindeksen

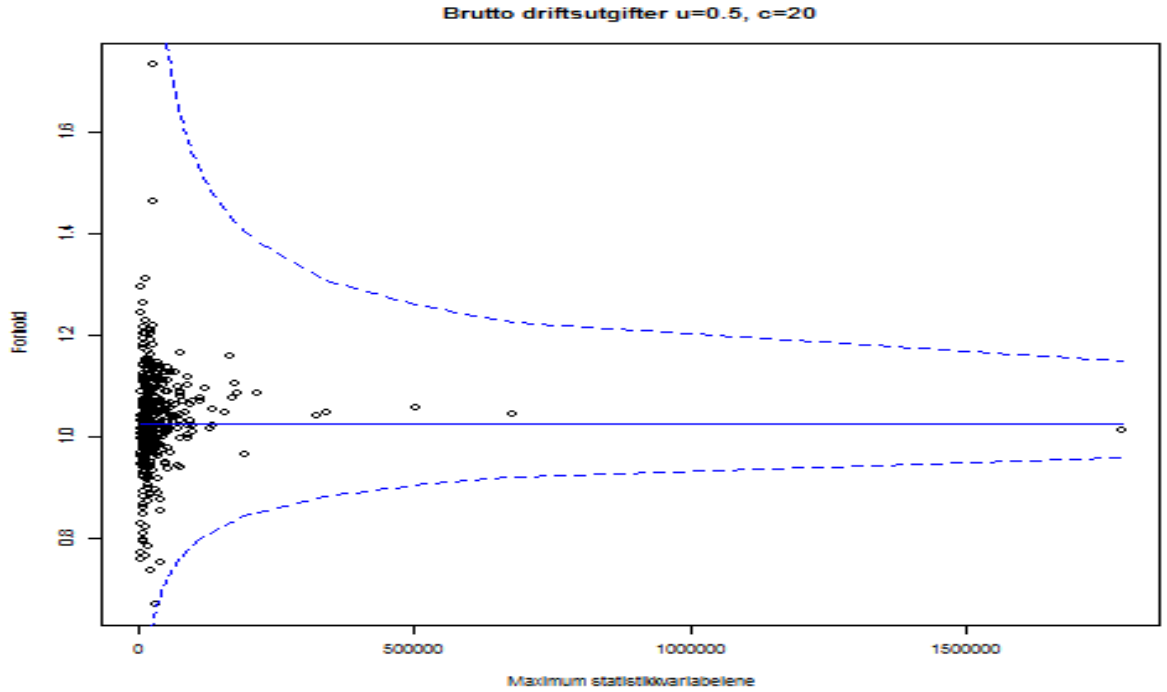


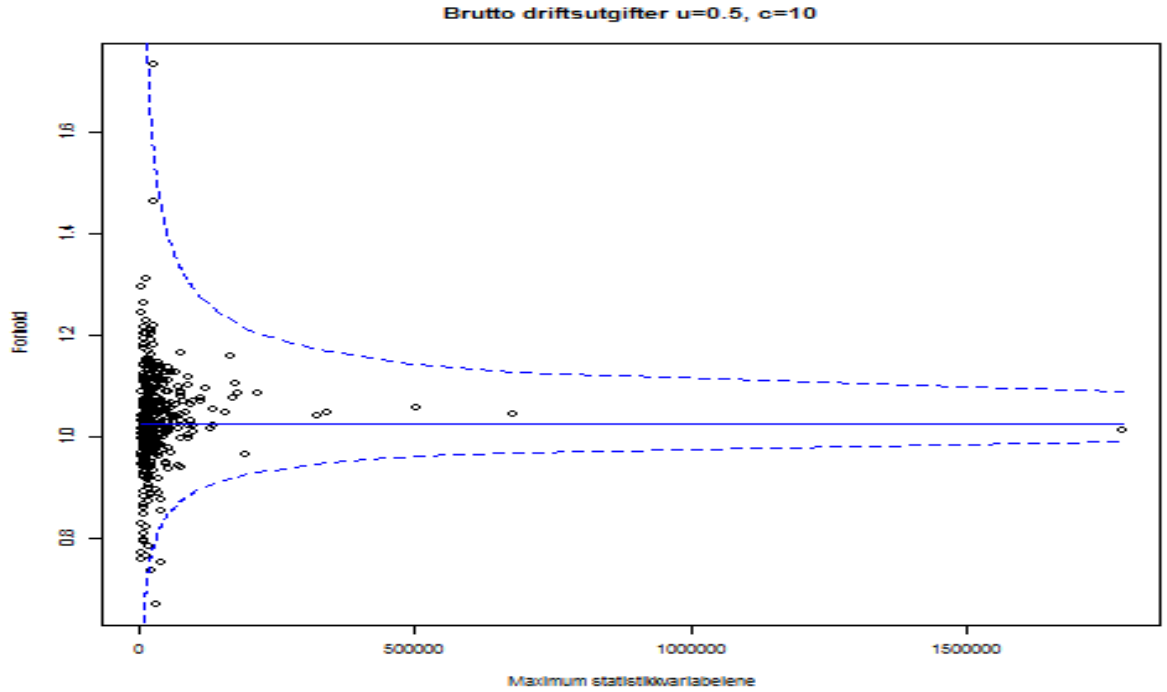
Statistisk sentralbyrå
Statistics Norway

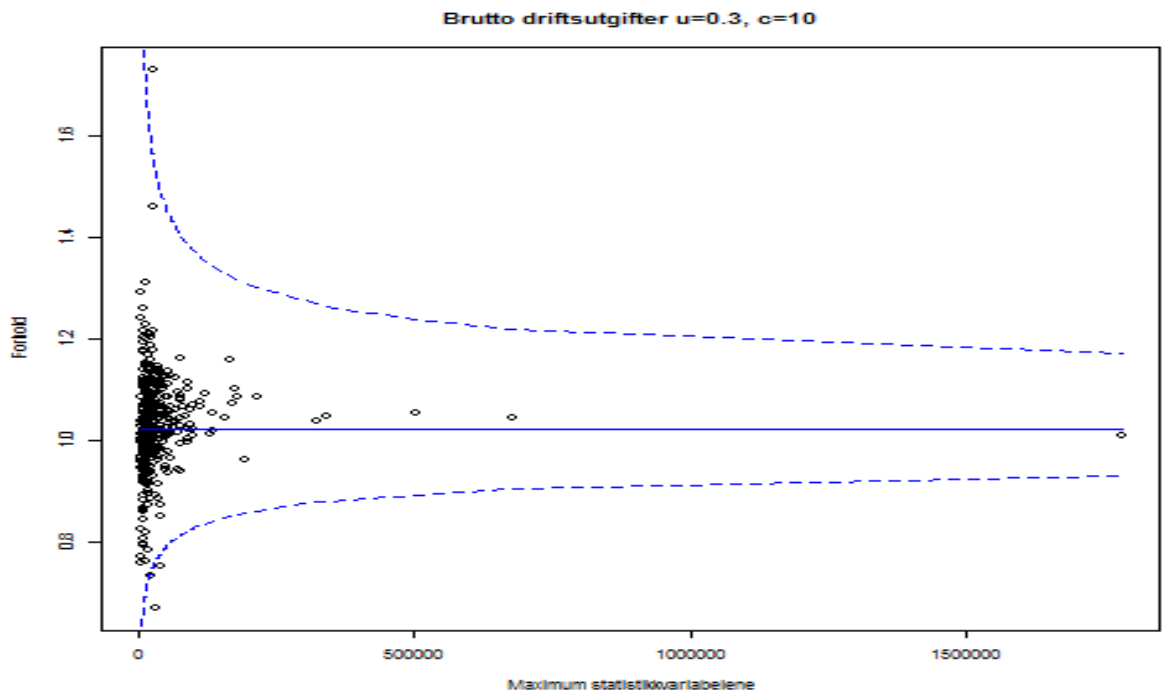
HB funksjon

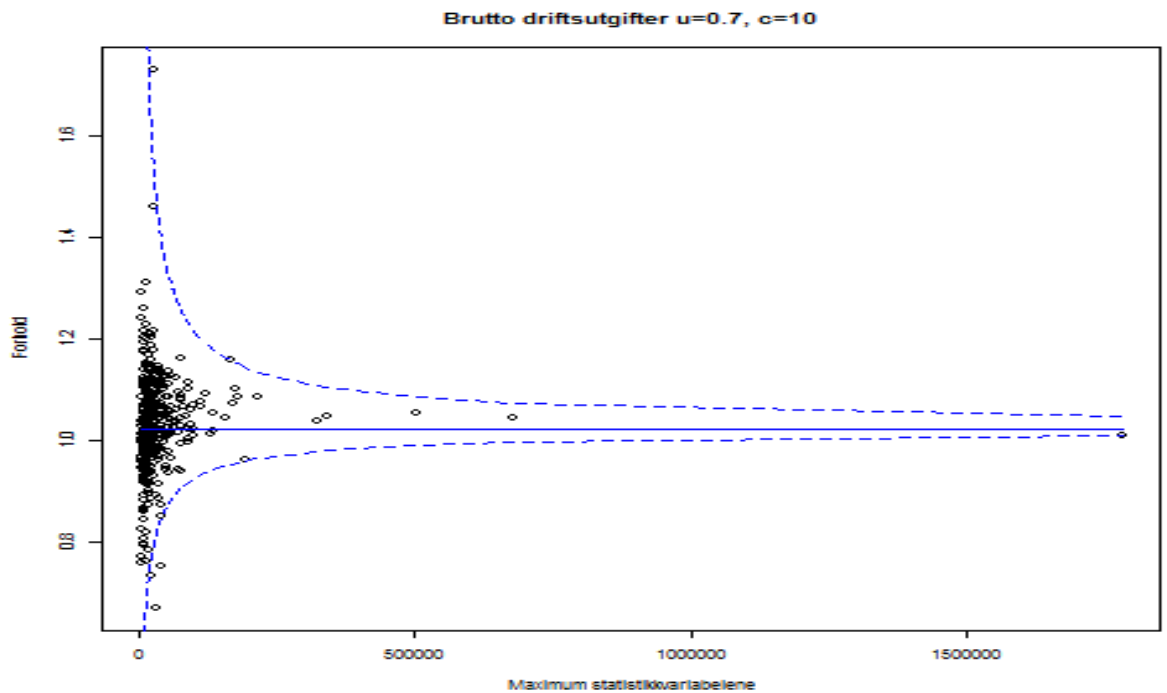


Input parametre er:
- data Input of Hb er et datasett av type dataframe.
- id Navn på en identifikasjonsvariabel.
- x1 Navn på variabel i periode t.
- x2 Navn på variabel i periode t-1.
- pU Parameter som justerer for forskjellige nivåer av variablene. Default verdi 0,5. pU<1
- pA Parameter som justerer 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

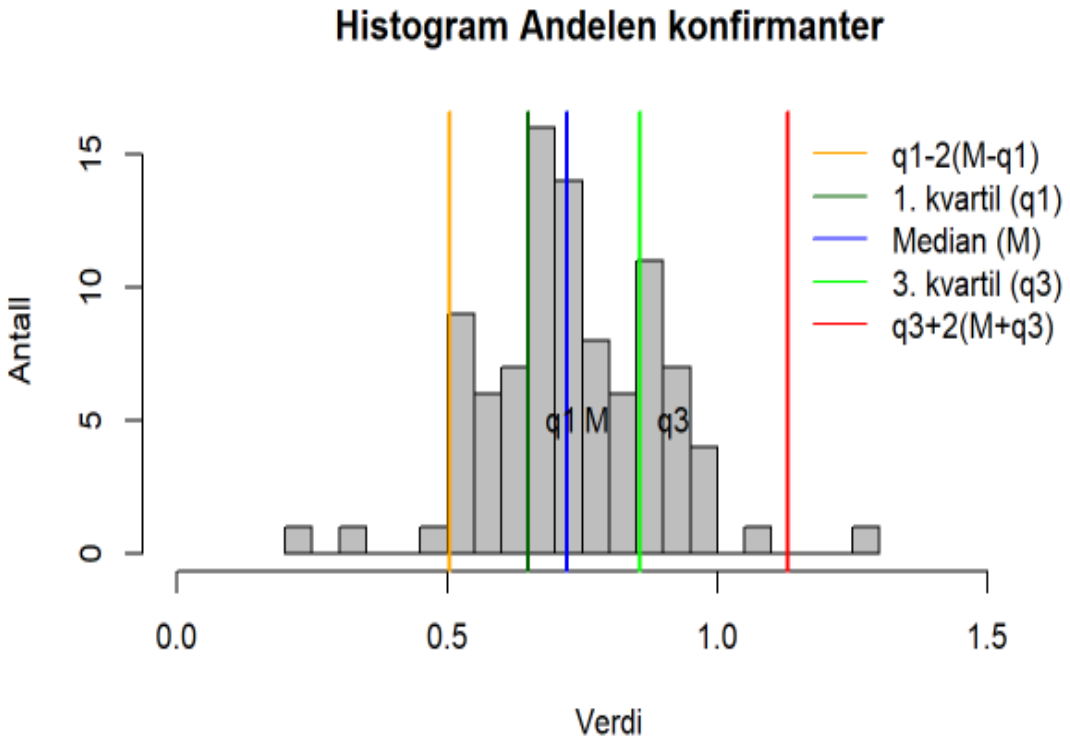

Brutto driftsutgifter u=0.3, c=10
Forhold
0.8
1.0
1.2
1.4
1.6
0
500000
1000000
1500000
Maximum statistikkvariabelene


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

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:

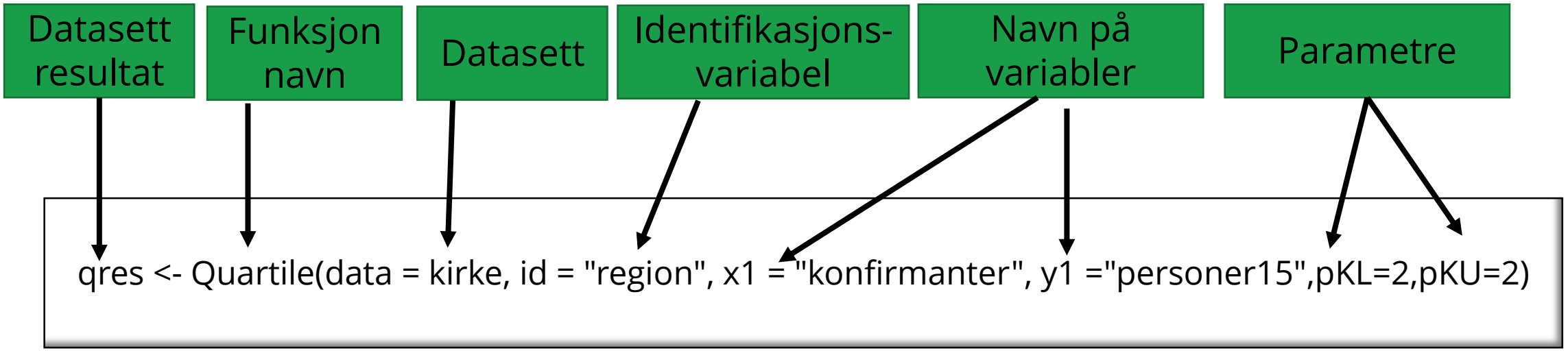
$(q_{0.25}-k_1(q_{0.5}-q_{0.25}),\ q_{0.75}+k_2(q_{0.75}-q_{0.5}))$



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

Kvartil funksjon



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

- Denne metodene er laget for å kontrollere to numeriske variabler mot hverandre.
- Metoden tar utgangspunkt at del kan bli laget en modell av sammenhengen mellom variablene og hvordan variasjonen er.
- Kjører iterativ for å finne en modell som ikke er påvirket av outliere

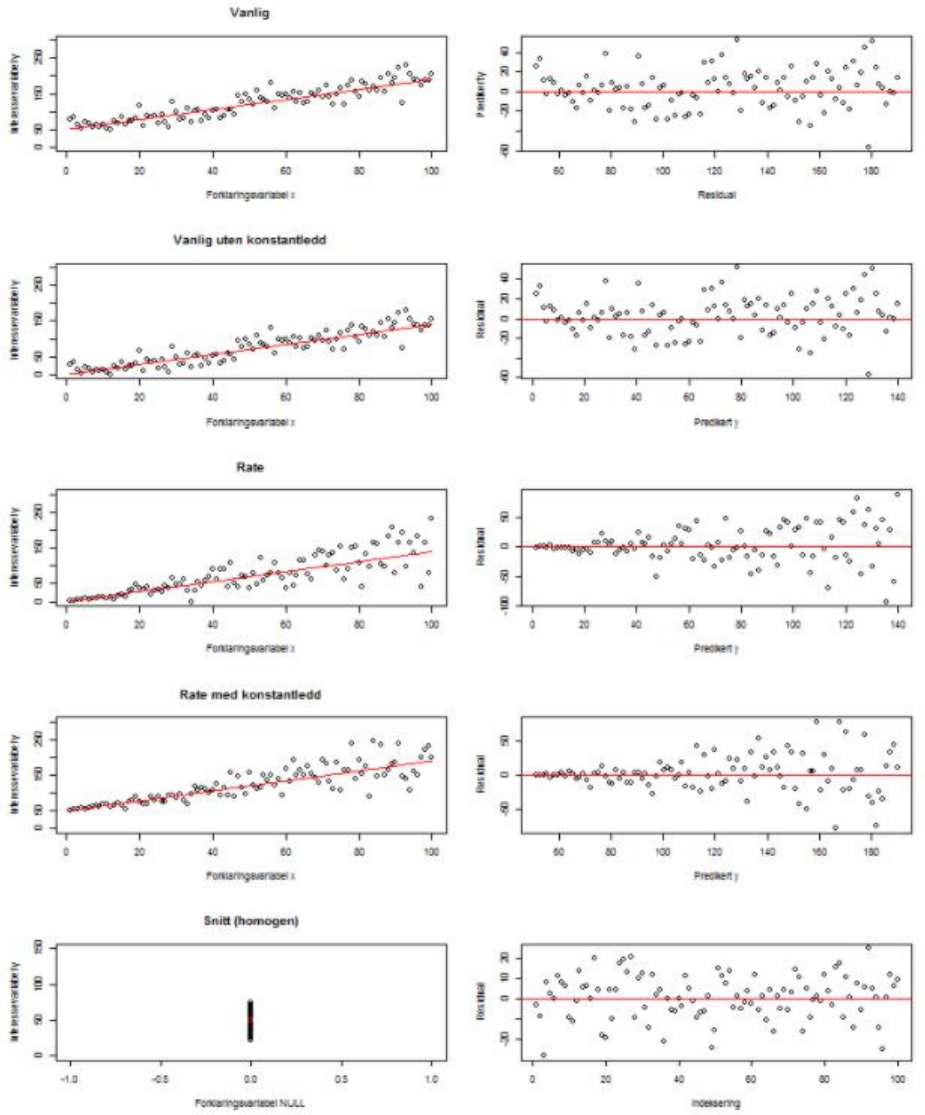
Robust regresjon

- Vi har multivariat versjon liggende som kan brukes

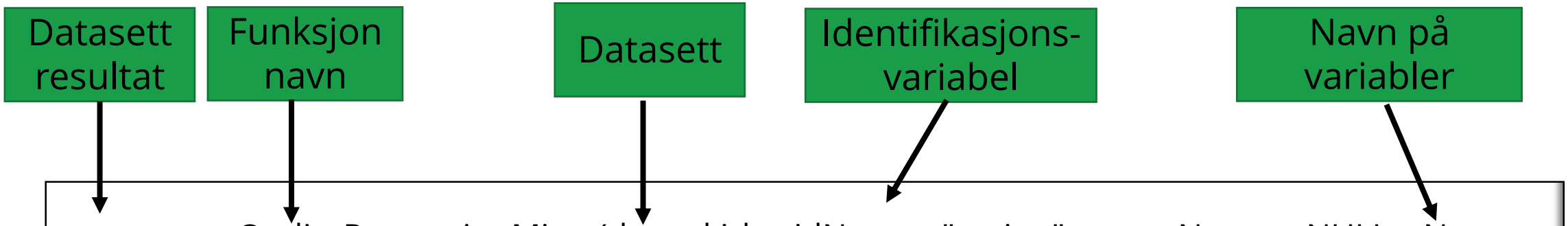
- Kan kjøres innen grupper
  - homogene grupper som ligner på hverandre
  - Må være nok observasjoner innen hver gruppe
  - Grupper kalles ofte strata av statistikere

Modeller

Navn	Modell	Varians
vanlig	Rett linje	Lik for alle
Vanlig med konstantledd	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



Regresjons funksjon




```
regres <- OutlierRegressionMicro(data= kirke, idName ="region" , strataName = NULL, xName
="personer15" , yName ="konfirmanter" , method = "ordinary", limitModel = 6, limitIterate = 6)
```

Parametre

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?

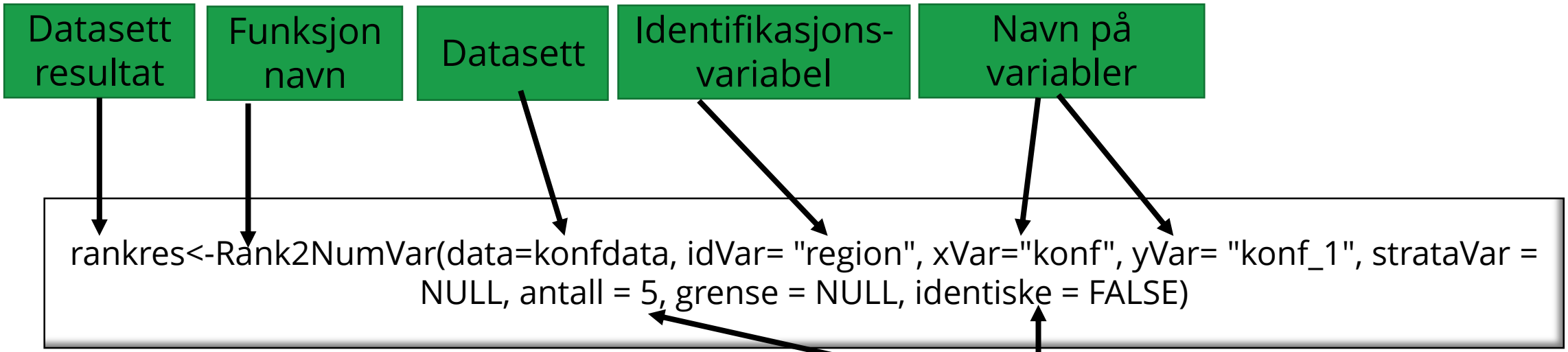
- Studenifiserte residualer - Avstand fra observasjon til regresjonslinjen (|rstudent| > 3)

- Dffits – en observasjons innflytelse på regresjonslinjen (|dffits| > 2[(m+1)/n]1/2)
  - n antall observasjoner og m antall estimerte parametre

Innflytelse på totalen

- Målet med metoden er å rangere de største verdiene og vise hvor stor andelen 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

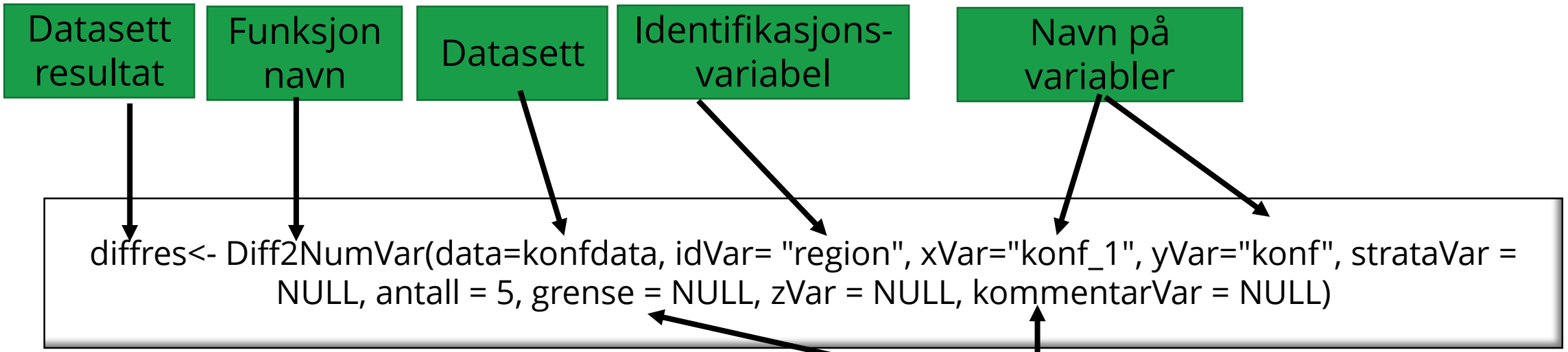
id <chr>	x <int>	y <int>	strata <chr>	forh <dbl>	xRank <int>	yRank <int>
1201	1668	1740	1	1.0431655	1	1
0301	1645	1722	1	1.0468085	2	2
5001	1046	1085	1	1.0372849	3	3
1103	930	793	1	0.8526882	4	4
0219	743	758	1	1.0201884	5	5

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

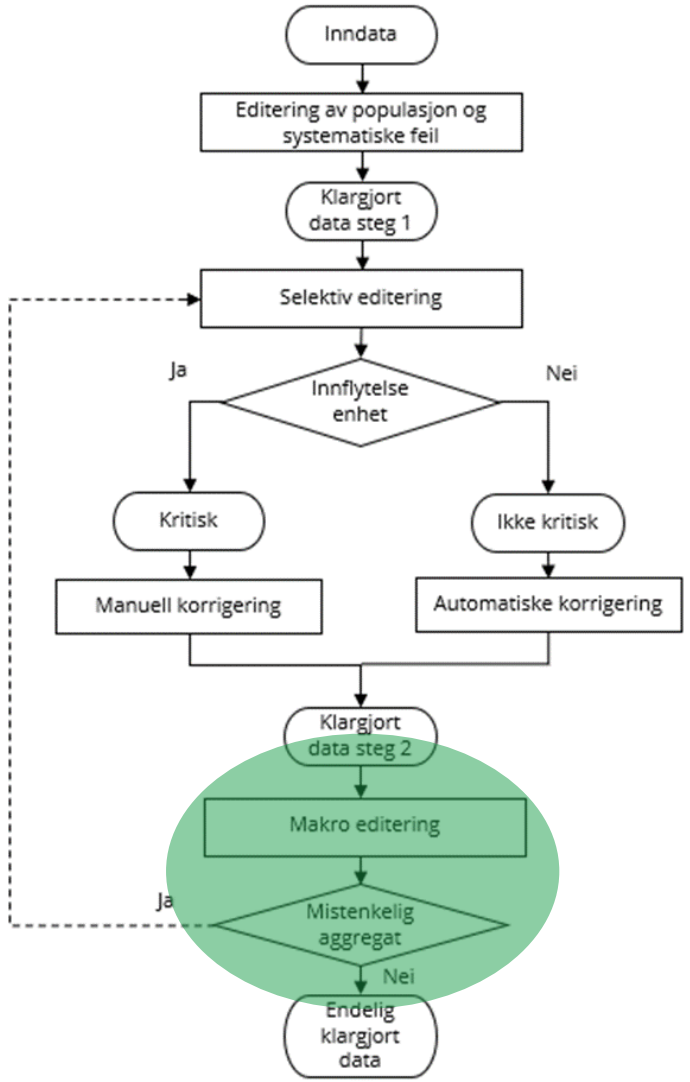
- z Variabelen input z

- Endring Forskjellen mellom z og y: y - z

- KommentarVar Input kommentar-variabelen

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

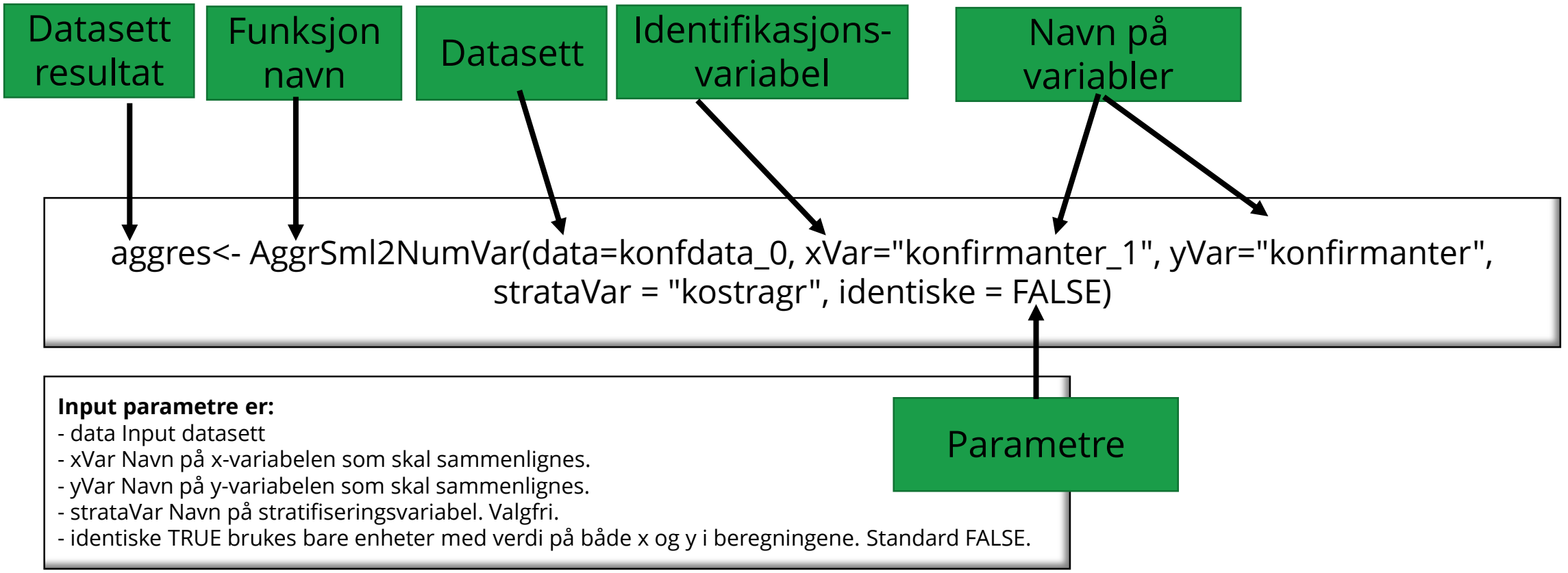
Analyse av aggregat

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



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 makro kontroller

12293: Omsorgstjenester - supplerende nøkkeltall

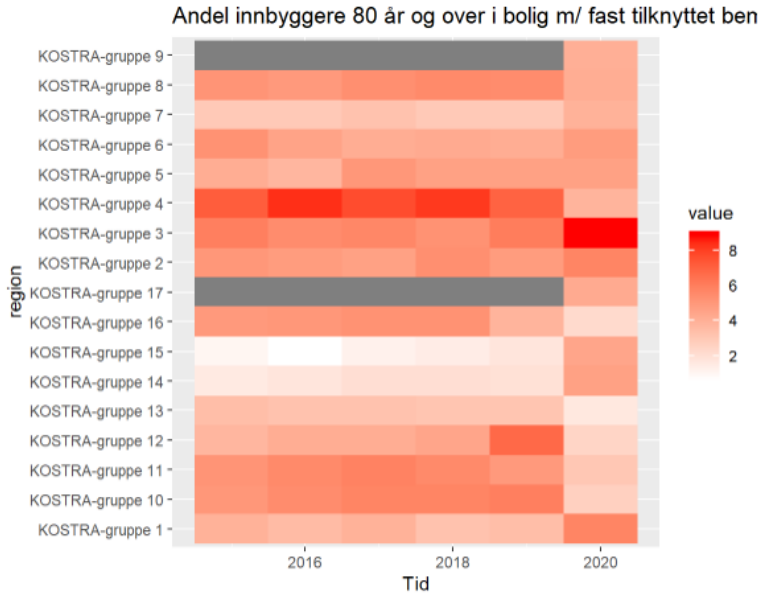
Landet | Fylke per variabel | Kostragruppe per variabel | Fylke samlet siste år | Robust multivariat oppdagelse av uteliggere

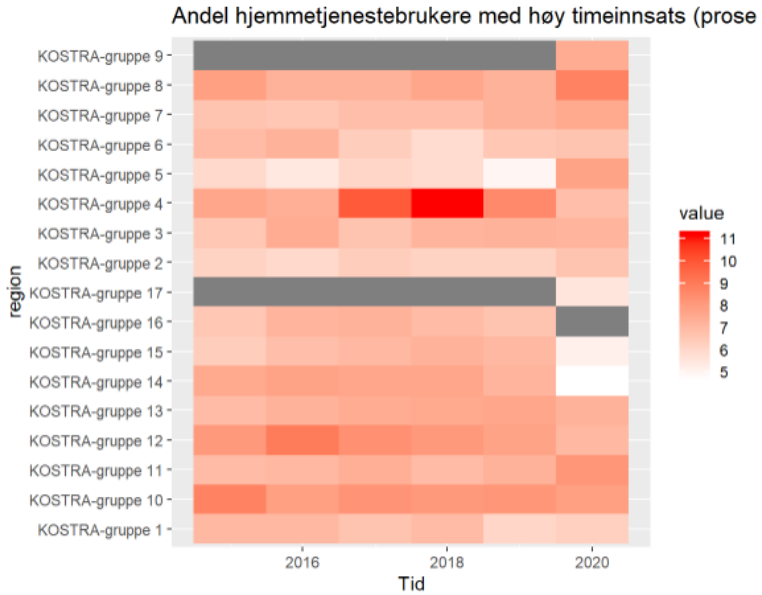
	statistikkvariabel	2015	2016	2017	2018	2019	2020	TrendSparkline
	All	/	/	/	/	/	/	All
1	Andel beboere 80 år og over i bolig m/ fast tilknyttet bemanning hele døgnet (prosent)	34.9	33.8	34.9	34.5	34.2	33.6	
2	Andel beboere i institusjon av antall plasser (belegg) (prosent)	98.8	98.8	98.1	0	0	0	
3	Andel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

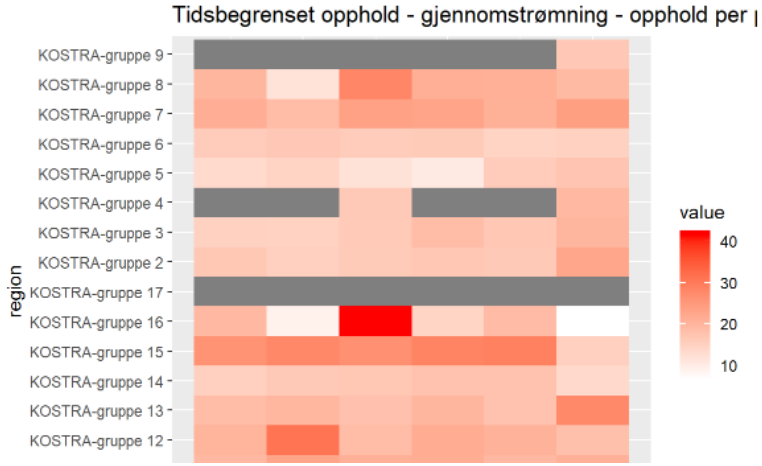
Statistisk sentralbyrå
Statistics Norway

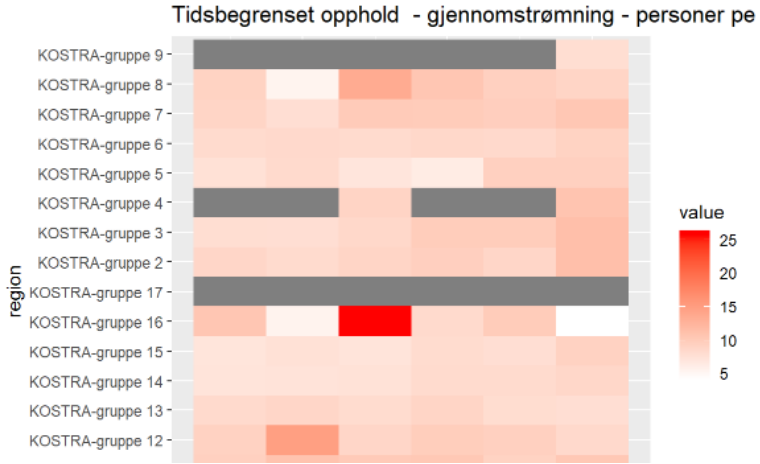
12293: Omsorgstjenester - supplerende nøkkeltall

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









Statistisk sentralbyrå
Statistics Norway

Eksempel og øvelse

- Oppgave 5-12
  - Kirkedata_0 med fjernet null og missing og bruk av variabelen døpte
  - Last inn r-pakken kostra og grafikkpakken plotly
- Oppgavene ligger på
  - teams <>
  - Github https://github.com/statisticsnorway/R_kontrollfunksjoner
- Gjennomgang av oppgaver starer 11.30

TALL SOM FORTELLER

Takk!

Statistisk sentralbyrå
Statistics Norway