Reproduserbarhet i sammenheng med Stordata

Data Science Assignment HVL University

Anne Grethe Lilleland, Arian Steen

17/09/20

## *Introduksjon*

I denne oppgaven tar vi for oss viktigheten av reproduserbarhet når man arbeider med store data.Vi vil se på hva *R* og *R-studio* er, og hvordan dette kan hjelpe oss med reproduserbarhet og replikerbarhet. Vi vil også forklare hva *R-markdown* og *R-Notebook* er.

## *Reproduserbarhet og replikerbarhet*

### **Hvorfor bryr vi oss om reproduserbarhet og replikerbarhet?**

Forskere prøver ofte å reprodusere eksprimenter for å øke troverdigheten til forsøket og resultatene i eksprimentet. Hvis eksprimentet kan reproduseres og få de samme svarene øker tiltroen/tillten til resultatet.Reproduserbarhet går ut på å bruke de samme dataene som allerede er brukt i forsåket, utføre da samme analysene, og ende opp med de samme svarene.

Replikerbarhet blir sett på som “gull” standarden innenfor forskning. Replikerbarhet går ut på at det skal være mulig å gjøre forsøket på nytt, med nye data, og ende opp med samme svar. Reproduserbarhet er nødvendig for å få replikerbarhet, men ikke tilstrekkelig nok til å gi replikerbarhet.

Reproduserbarhet og replikerbarhet har vist seg og ikke være noen selvfølge innen forskning. Det er flere meninger og tanker rundt dette. Noen hevder forskningen burde være reproduserbar for at den skal kunne publiseres. Andre går så lang og mener den også bør kunne replikeres.Vi vil ikke gå nærmere inn på dette, men vi vil se mer på grunnene til at studier ikke er reproduserbare og replikerbare, og ulike løsninger på dette senere i oppgaven.

### **Hvorfor R og R studio?**

Både R og R studio er åpen kildekode programvarer og er gratis. Det brukes i sammenheng med forskning og datavisualisering. R er ett kodespråk for statistisk databehandling og grafikk, mens R-studio er ett integrert utviklingsmiljø for R. R og R studio gjør sammarbeid enklere via feks. Git og Github. Man kan importere og manipulere data og kjøre dette i ett maskinlæringsprogram via R studio. Skal man kjøre linerar regrisjon eller lignende kan man bruke R og R studio. Analyser gjort i R og R studio er reproduserbare, da alt som er gjort ligger inne i “history” og kan bli hentet frem igjen.

### **R-markdown og R-notebooks**

R-Markdown gjør det mulig å skape rike, fullt dokumenterete, reproduserbare analyser (kilde). R koder, kommentarer, metadata ligger da inni det samme dokumentet. Dette gjør at man i etterkant kan se hvilke data som har blitt brukt. Hvilke formler som er brukt for å få svarene i analysen. Alt som har blitt gjort i analysen kan bli funnet igjen. Det er ikke gjort analyser separert hvor svarene har blitt plottet inn i dokumentet i etterkant.Dette dokumentet er skrevet med R notebook, som er et r-markdown dokument med kode “chunks” som kan utføres uavhengig og interaktivt, hvor resultatet er synlig med en gang. Dokumentet blir en blanding av vanlig skrift og koder, noe som gjør det svært effektivt iforhold til andre skriveprogrammer. Kodene som brukes er ikke synlige i det endelige resultatet som blir publisert, kun resultatene av disse kodene. Ønsker man derrimot å se hvilke koder som er blitt brukt er det mulig i etterkant. Dette gjør at dokumentet blir reproduserbart.

## *Hvorfor er ikke flere studier reproduserbare og replikerbare?*

Det kan være ulike grunner til at studier ikke er reproduserbare og/eller replikerbare. En årsak kan være feilaktig forkastning av null hypotesen. Dette kan være resultat av feil i analysen eller ett dårlig utvalg. Det kan også være ett resultat av feilaktig innplotting som har gått uanmerket. Når analysen blir gjort på ny igjen vil resultatet bli annerledes. Innen psykologifaget ble det gjort en undersøkelse hvor de prøvde å replikere 100 undersøkelser fra de siste 3 årene som var blitt publisert i de tre mest respekterte journalene innenfor faget. De fant ut att effektene fra funnene i disse studiene ble halvert etter at de ble gjennomført på ny (replikert).De fant også at i 95% av de orginale undersøkelsen hadde de fått ett signifikans nivå på 0,05 eller under, mens etter at undersøkelsen ble utført på ny var det kun 36% som hadde ett signifikant resultat.

### **wale of London**

Historien om JP Morgan Chase tabben som kostet selskapet 2 Milliarder dollar, aksjeinvestorer som jobbet for JP Morgan hadde slått inn en feil i excel spreadsheet, som underestimerte feilen i den syntetiske modellen de investoren brukte for å kalkulere risikoen i handelen. Videre undersøkelser viste hvordan tabben en enkel tabbe i excel spreadsheet remodelerte den finansielle risikomodellen som ble brukt i sammenheng med investeringen.

## *Historisk data arkivering*

## *Løser R notebooks problemet med reproduserbarhet?*

## *Code chunks og text chunks*

## *Reproduserbarhet i ulike sektorer*

## *Konklusjon*

# ***Litteraturliste***

## **Innledning og Historien om “R”**

Utviklingen av diverse dataprogrammer (software) har styrket analytikernes evne til å gjennomføre flerdimensjonal data-analyse. En slik type analyse, gir muligheter til å gjennomføre komplekse beregninger med svært raske resultater. Blant de ulike dataprogrammer som brukes til dette finner man, for eksempel. SPSS,STATA,EXCEL og R. Den sistnevnte “R” tilbyr unike muligheter som skiller det fra andre tilgjengelige statistiske verktøyene. Faktisk har R blitt hovedverktøyet til data-analyse. De R ekspertene som bruker R-programmet, kalles Data-Analyst eller Data Scietist, på grunn av det brede omfanget som “R” kan tilby.

I 1991 ble “R” opprettet av Ross Ihaka og Robert Gentleman i “Department of Statistics ved University of Auckland”. I 1993 ble den første kunngjøringen av “R” gjort for publikum. Ross og Roberts erfaring med å utvikle “R” er dokumentert i en artikkel i det vitenskapelige journal (Ihaka,R & Gentleman,R 1996)

Populæriteten til “R” kommer fra diverse kilder. For det første er den basert på multiplatform kode, som vil gjøre det mulig å bruke den på tvers av ulike platformer (Dette fører til enklere sammarbeid mellom flere parter) og åpen kildekode (Det sikrer reproduserbarheten av analysen), det blir brukt av profesjonelle i arbeidslivet, store selskaper og studenter som foreksempel (Økonomisk Data-analyse) brukt av HVL studenter innen økonomifeltet. “R” gir oss muligheten til å jobbe raskt og iretaivt med store datamengder.

For det andre, er brukervennligheten til “R” og det faktum at det finnes store mengder læremateriale åpent,fritt og lett tilgjengelig på nettet, gjør det mulig å reprodusere data, gjennomføre analyser og løse diverse problemstillinger. “R” gjør det mulig for forskere å fange opp og kommunisere detaljene i arbeidsflyten med større effektivitet. I følgende paragrafer forteller vi nærmere om hvordan disse ulike elementene øker vår mulighet til å gjennomføre en analyse ut ifra et datasett ved bruk av “R”.

## **Reproduserbarhet og R-Notebooks**

Repetering og intern reproduserbarhet et aspekt innen Data Science feltet, og alle som håndterer og jobber/deler store mengder data. Det handler om overensstemmelse mellom resultatene av målinger av samme målstørrelse utført under samme målebetingelser (VIM, 1993). Reproduserbarhet avspeiler i hvilken grad en har overenstemmelse i gjentatte målinger på samme prøvemateriale, forutsatt at de utføres med samme betingelser.

Dette er en viktig aspekt for å få sikre metodikken i en analyse. For å måle graden av reproduserbarhet kan en benytte to eller flere kontrollmaterialer (Som for eksempel resultat av en prøve) som dekker måleområdet, og analysere det for eksempel 10 ganger, og se hvordan resultatet endrer seg.

På denne måten kan man validere den gjennomførte analysen. Verdien av hva forskjellen mellom to testresultater oppnådd under reproduserbarhetsbetingelser kan forventes å oppstå i utgangpunk i forskjellige konfidensintervall (sannsynlighet) som et intervallestimat. For eksempel på omtrent 99%, 95% eller 90%. Hvis vi bruker 95% konfidensintervall, vil det si at det er 95 prosent av tilfellene det man utfører eksperimentet, vil populasjonens gjennomsnitt ligge innenfor dette intervallet(Bergsaker, 2019).

## **Eksempler på "Code cunk,Text chunk og Sessioninfo innenfor R og R Notebook**

### Text Chunk

# Her kan man putte tekst i en såkalt "kode-chunk", gjerne for å forklare kodeplotten nedenfor som for eksempel.For å sette inn en code cunk kan man bruke snarveien ctrl+alt+i på windows og cmd+alt+i på OSX.

### Code Chunk

summary(iris)

## Sepal.Length Sepal.Width Petal.Length Petal.Width   
## Min. :4.300 Min. :2.000 Min. :1.000 Min. :0.100   
## 1st Qu.:5.100 1st Qu.:2.800 1st Qu.:1.600 1st Qu.:0.300   
## Median :5.800 Median :3.000 Median :4.350 Median :1.300   
## Mean :5.843 Mean :3.057 Mean :3.758 Mean :1.199   
## 3rd Qu.:6.400 3rd Qu.:3.300 3rd Qu.:5.100 3rd Qu.:1.800   
## Max. :7.900 Max. :4.400 Max. :6.900 Max. :2.500   
## Species   
## setosa :50   
## versicolor:50   
## virginica :50   
##   
##   
##

# Oppsumering av en plot som for eksempel

### Man kan også legge inn plot og visualisering som for eksempel



### Sessioninfo forklart og demonstrert

Sessioninfo, samler og viser informasjon om den nåværende økten som kjører i R. Informasjonen inneholder for eksempel, R versjonen som kjører, operativsystemet som er brukt og de ulike packages som er lastet. Dette er brukbart for å undersøke mulige feil og avvik, og bidrar til å gjøre prosjektet mer reproduserbart ved å gi oss mer informasjon slik at det for eksempel i fremtiden skal bli lett å reprodusere dette datasettet uavhengig av hva du kjører. Sessioninfo vil derfor hjelpe med å gjøre dette lettere ved å gi oss mer informasjon om økten. Nedenfor er dette demonstrert

sessionInfo()

## R version 4.0.2 (2020-06-22)  
## Platform: x86\_64-w64-mingw32/x64 (64-bit)  
## Running under: Windows 10 x64 (build 19041)  
##   
## Matrix products: default  
##   
## locale:  
## [1] LC\_COLLATE=Norwegian Bokmål\_Norway.1252   
## [2] LC\_CTYPE=Norwegian Bokmål\_Norway.1252   
## [3] LC\_MONETARY=Norwegian Bokmål\_Norway.1252  
## [4] LC\_NUMERIC=C   
## [5] LC\_TIME=Norwegian Bokmål\_Norway.1252   
##   
## attached base packages:  
## [1] stats graphics grDevices utils datasets methods base   
##   
## loaded via a namespace (and not attached):  
## [1] compiler\_4.0.2 magrittr\_1.5 tools\_4.0.2 htmltools\_0.5.0  
## [5] yaml\_2.2.1 stringi\_1.4.6 rmarkdown\_2.3 knitr\_1.29   
## [9] stringr\_1.4.0 xfun\_0.16 digest\_0.6.25 rlang\_0.4.7   
## [13] evaluate\_0.14

## Ulike overskrifter og skrivemåter

## Har forskerne incentiver til å være “Reproduserbare” eller må de tvinges?

Et av vilkårene for pålitelig vitenskapelig arbeid, er reproduserbarhet. Dette betyr at hvis en annen forsker gjentar undersøkelsen og målingen ved like betingelser og underliggende faktorer Er det betryggende å få tilsvarende (like) resultater. En høy reproduserbarhet uttrykkes som reliabilitet (Bartlett & Frost, 2008). Det er også viktig å være oppmerksom på at noen av målingene gir lik resultat, uavhengig av hvor, når og hvilken forsker som gjennomfører det. For eksempel, kjemiske analyser i et laboratorium må ha i utgangspunkt høy reproduserbarhet, og at målinger av studien kan etterprøves. Et annet eksempel er markedsundersøkelser som utforsker meninger om produkter, tjenester og kjøpemønstre.

Målinger i sammfunnsvitenskapelig perspektiv innebærer mange kontekstuelle betingelser, som gjør at resultatet ikke er reproduserbar. For eksempel måling av hvordan folk reagerer på nyheter i konkrete politiske atmosfære, eller hvordan velgerne reagerer på kandidater i politiske kampanjer og hvilke saker som er viktige for dem ved valg. Dette er noe som ikke er veldig reproduserbart, Forskjere som jobber med store datasett er nærmest nesten avhengig av å kunne reprodusere det de jobber med, dette er for å finne ulike avvik og feil i modellen som blir brukt. Økonomiske modeller som er forsket på må for eksempel kunne bli reprodusert, likt alle andre kvantitative felt innen forsking. Derfor er det Forskere har dermed incentiver til å ivareta datasett og kunne reprodusere modeller i framtiden.

## Er økt reproduserbarhet noe som vil tvinge seg fram eller er dagens økte interesse bare en blaff?

Økt kunnskapsnivå i sammfunnet, lett tilgjengelighet av datasett, materiale og utvikling av dataprogrammer og pakker innenfor ulike programm skreddersydd for hvert felt. Disse ulike faktorene gjør at den økte interessen vil øke. For eksempel finnes det pakker innenfor Rstudio som er skreddersydd for psykologer. (Psych) pakken i dette tilfellet. Alle disse faktorene gjør reproduserbarhet innen forsking og akademiske journaler viktigere framover.