

## Seminar 5 (Datasett: ess.rda)

Disse oppgavene er ment som repetisjon av basics, og forberedelse til prøven. På nett vil dere finne et R-script med ufullstendig kode, som dere kan bruke som hjelp til å løse oppgavene, alle funksjoner dere trenger er lagt inn her. Dersom dere klarer en oppgave uten dette scriptet er det supert. Jobb i grupper. Fasiten blir lagt opp i pausen, men ikke kikk på den med mindre dere står helt fast. Dersom det er noe som ikke fungerer, sjekk om parenteser, skrivefeil, eller hermetegn kan være feilkilden. Spør om hjelp dersom dere står fast.

1. Vis hvordan du importerer datasettet `ess.rda` til R med `load()`. Inspiser data ved hjelp av `str()`, `View()` og `summary()`.
2. Vis hvordan du oppretter et nytt datasett, «data», bestående kun av observasjoner fra Tyskland. Sjekk hvor mange observasjoner du vil få på forhånd.
3. Vis hvordan du lager et nytt datasett, «cordata» basert på «data», bestående av variablene «gender», «age», «income\_decile» og «trust\_police». Omkod gender til en dummyvariabel med verdien 1 for «female» og 0 for «male».
4. Vis hvordan du lager en korrelasjonsmatrise med «cordata» og `cor()`, bruk først «pairwise.complete.obs», bruk deretter «complete.obs». Lag deretter en ny versjon av cordata, der alle missing observasjoner er fjernet. Lag korrelasjonsmatrisen på nytt, og forklar forskjellen på «pairwise.complete.obs» og «complete.obs» (du kan også se på hjelpefilene for `cor`). Hvilke korrelasjoner vil ligge til grunn for en regresjon med de 4 variablene?
5. Omkod `cordata$trust_police` til numerisk. Vis hvordan du lager et scatterplot mellom «age» og «trust\_police» fra cordata, ved hjelp av `ggplot` og `geom_point`. Lag også en tabell med de to samme variablene.
6. Vis hvordan du legger til en lineær regresjonslinje til plottet ved hjelp av `geom_smooth` (husk på at dette bare fungerer for regresjon med to variabler), bruk `method = «lm»`.
7. Vis hvordan du kan finne gjennomsnitt, standardavvik, skewness og kurtose for `cordata$age`. Bruk «moments» pakken til skewness og kurtose. Lag et histogram for variabelen, og forklar til sidemannen hva de ulike tallene betyr.
8. Kjør en lineær regresjonsanalyse med `trust_police` som avhengig variabel, og de resterende variablene fra cordata som uavhengige variabler, lagre som objekt. Tolk resultatene
9. Bruk `str()` og eller `names()` til å finne ut hvordan du kan hente ut residualer fra modellen ved hjelp av indeksering. Lag et nytt objekt av disse residualene, som du kaller `m1resid` Kombiner residualene med «data», og lag et scatterplot av residualer mot faktiske y-verdier.
10. Omkod `cordata$trust_police` til en dummy variabel, «trust\_police\_d» med ifelse, slik at alle som har verdier lavere enn 5 får verdien 0 på den nye variabelen, mens resten får verdien 1. Kjør en logistisk regresjon lik modellen i oppgave 8, men med den nye dummyvariabelen som avhengig variabel. Finn ut hvordan du kan hente ut koeffisientene fra modellobjektet ditt ved hjelp av `names()` eller `str()`, og lag oddsrater av logitkoeffisientene. Kan du beregne effekten av gender når alle de to andre variablene er satt til sin gjennomsnittsverdi (bruk enten formelen for omregning, eller `predict` med `newdata` og `type = «response»`)? Er effekten av kjønn substansiell?