R-prøve 2 STV 1020 V18, tirsdag 8. mai 14.15-16.00

Kodebok

Datasettet pinochet består av 2700 observasjoner og 8 variabler. Observasjonene er personer fra en spørreundersøkelse før en folkeavstemning for/mot Pinochet. Det er noen missing-data.

Variabler:

region	A factor with levels: `C`, Central; `M`, Metropolitan Santiago area; `N`, North; `S`, South; `SA`, city of Santiago.
population	Population size of respondent's community.
sex	A factor with levels: `F`, female; `M`, male
age	in years.
education	A factor with levels (note: out of order): `P`, Primary; `PS`, Post-secondary; `S`, Secondary.
income	Monthly income, in Pesos.
statusquo	Scale of support for the status-quo.
vote	a factor with levels: `A`, will abstain; `N`, will vote no (against Pinochet); `U`, undecided; `Y`, will vote yes (for Pinochet).

Instruksjoner:

• Prøven skal besvares med et fungerende R-script (.R-fil, ikke sett noen punktum i filnavnet når du lagrer!) som lastes opp i innleveringsmappen «prøve 2» på Fronter. Innleveringsmappen finner dere i arkiv-mappen i Fronter-rommet for seminargruppen, jeg har åpnet mappen slik at dere får forsøkt dere på innlevering.

• Scriptet skal inneholde nødvendig kode for å besvare oppgavene samt kommentarer markert med # som forklarer fremgangsmåten dere har valgt. Der oppgavene ber dere oppgi bestemte verdier eller tolkninger skal disse også oppgis som korte kommentarer i scriptet (du trenger ikke skrive mer enn en linje eller to).

• Sørg for at koden er oversiktlig. For å skille oppgavene fra hverandre i scriptet, anbefales overskrifter av typen:

Oppgave 1
Oppgave 1:

• Flere av oppgavene kan løses på forskjellige måter, du står fritt til å velge fremgangsmåte selv. Det er lov å google og bruke alle hjelpemidler (som oversikten over funksjoner og feilsøkingsdokumentet på github). Det eneste som ikke er lov, er kommunikasjon med medstudenter. Dersom dere skriver kode som er riktig, men ikke klarer å løse en oppgave fullstendig kan dere likevel få god uttelling.

Lykke til!

Oppgaver:

1) Importer datasettet pinochet.csv eller pinochet.Rdata fra data-mappen på github (https://github.com/langoergen/stv1020R/tree/master/data) som et objekt i R-Studio
2) Hvor mange missing-verdier er det på variabelen vote?
3) Lag et histogram med variabelen income
4) Lag en ny variabel, decided_vote, der de som vil stemme for Pinochet får verdien 1, de som vil stemme mot Pinochet får verdien 0, mens alle andre blir satt til missing.
5) Lag en korrelasjonsmatrise mellom age, income og statusquo. Fjern missing ved listwise deletion. Tolk en av korrelasjonene substansielt, og utfør en signifikanstest for denne korrelasjonen
6) Estimer en regresjon med statusquo som avhengig variabel og alder, utdanning, kjønn og inntekt som uavhengige variabler. Lagre modellen som et objekt. Hva er den forventede effekten av alder? Hvor mange observasjoner har missing på en av variablene som inngår i modellen?
7) Lag et spredningsplot (scatterplot) med alder på x-aksen og statusquo på y-aksen. Legg deretter til en lineær regresjonslinje oppå plottet.
8) Opprett en ny variabel i datasettet ditt, income_log, ved å ta den naturlige logaritmen til income. Estimer en regresjon med statusquo som avhengig variabel og alder, utdanning, kjønn og log_income som uavhengige variabler. Lagre modellen som et objekt. Er effekten til alder forskjellig fra oppgave 6?
9) Opprett et nytt datasett uten missing på noen av variablene som inngikk i regresjonsmodellen i oppgave 8, og opprett en ny variabel i dette datasettet med restleddene (residualene) fra modellen du estimerte i oppgave 8. Lag et qqplot for å sjekke om restleddene er normalfordelt. Kommenter plottet kort.
10) Opprett en ny variabel, «no», med verdien 1 for dem som vil stemme nei til pinochet, 0 for alle andre. Aggreger pinochet basert på observasjonenes verdier på «region», «sex» og «education». Regn ut gjennomsnitt for variabelen «statusquo» og andelen nei-stemmer (av totalt antall observasjoner) for hver gruppe. Lagre som et nytt datasett.