R-prøve 1 STV 1020 V18, tirsdag 8. mai 12.15-14.00

Kodebok:

Datasettet labour består av 753 observasjoner og 8 variabler. Observasjonene er fra en amerikansk panelundersøkelse av gifte kvinner

Variabler:

lfp	labor-force participation; a factor with levels: `no`; `yes`
k5	number of children 5 years old or younger.
k618	number of children 6 to 18 years old.
age	in years.
wc	wife's college attendance; a factor with levels: `no`; `yes`.
hc	husband's college attendance; a factor with levels: `no`; `yes`.
lwg	log expected wage rate; for women in the labor force, the actual wage rate; for women not in the labor force, an imputed value based on the regression of `lwg` on the other variables.
inc	family income exclusive of wife's income.

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 1 tirsdag» 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 labour.csv eller labour.Rdata fra data-mappen på github (https://github.com/langoergen/stv1020R/tree/master/data) som et objekt i R-Studio
2) Hvor mange av kvinnene i utvalget deltar i arbeidsstyrken?
3) Opprett en ny variabel, lfp.d ved å omkode lfp slik de som deltar i arbeidsstyrken får verdien 1, mens de som ikke deltar i arbeidsstyrken får verdien 0.
4) Opprett to nye datasett: a) ett datasett bestående av kvinner som gikk på college og som har færre enn 3 barn mellom 6 og 18 år og b) ett datasett bestående av kvinner som ikke gikk på college som har 3 eller flere barn mellom 6 og 18 år. Hva er medianinntekten til familien til kvinnene i de to nye datasettene?
5) Beregn den bivariate korrelasjonen mellom alder og lwg (pearsons r) og test om korrelasjonen er signifikant forskjellig fra 0. Uavhengig av signifikans: tilsier korrelasjonen at eldre eller yngre kvinner tjener mest?
6) Lag et boxplot med hc på x-akse og forventet inntekt på y-akse. Ser det ut som om de som har menn som gikk på college har høyest eller lavest forventet medianinntekt?
7) Estimer en regresjonsmodell med forventet inntekt (lwg) som avhengig variabel, og wc, k5, k618 og age som uavhengige variabler. Lagre modellen som et objekt, og tolk effekten av at kvinner har utdannelse fra college på forventet inntekt substansielt.
8) Lag et spredningsplot (scatterplot) med alder på x-aksen og forventet inntekt på y-aksen. Fargelegg punktene ut fra verdi på variabelen wc.
9) Estimer en regresjonsmodell med forventet inntekt som avhengig variabel, og wc, k5, k618, age og age kvadrert (andregradsledd) som uavhengige variabler. Lagre modellen som et objekt, og tolk effekten av å gå fra å være 30 til 50 år på forventet inntekt substansielt. Gjør til slutt en vif-test for å sjekke om det er mye multikolinearitet mellom variablene i modellen (du kan bruke en funksjon fra pakken «car» til dette). Hvilke to variabler har høyest multikolinearitet?