R-Prøve STV 4020A

Instruksjoner:

- Prøven skal besvares med et fungerende R-script som sendes til erlend.langorgen@stv.uio.no innen kl. 12.00.
- 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 kommentarer i scriptet.
- 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. Dersom dere har lyst til å bruke en pakke, men ikke får lastet den inn, kan får dere uttelling for å skrive riktig kode basert på pakken.
- Kommentarer trenger ikke være lengre enn en setning. Kunnskaper om statistikk vektlegges mindre enn kjennskap til R.
- Prøven er laget slik at det ikke vil oppstå følgefeil i senere oppgaver dersom det er en oppgave du ikke klarer, med unntak av oppgave 1 (innlasting av data). Du skal bruke dette datasettet som utgangspunkt for alle oppgavene. Du står derfor fritt til å løse oppgavene i den rekkefølgen du vil.
- En rød tråd i oppgaven, er å undersøke effekten av utdanning på lønn. Denne røde tråden kan gjøre det lettere å løse oppgavene i den oppsatte rekkefølgen.
- Det er mulig å få følgefeil innad i en oppgave, skulle det skje, skriv kode som du mener ville løst oppgaven dersom du ikke hadde følgefeil. Lykke til!

Kodebok for «prove2.RData»

Variabelnavn	Variabelbeskrivelse
Exp	years of full-time work experience
Wks	weeks worked
bluecol	blue collar ?
ind	works in a manufacturing industry ?
south	resides in the south ?
smsa	resides in a standard metropolitan statistical area?
married	married ?
Sex	Male or female?
union	individual's wage set by a union contract ?
ed	years of education
black	is the individual black ?
lwage	logarithm of monthly wage (in USD)

Datasettet er basert på intervjuer med 4165 amerikanere på starten av 80-tallet.

Oppgaver:

Oppgave 1:

Last inn datasettet «data.RData». Du finner datasettet på https://github.com/martigso/stv4020aR ved å klikke deg inn på mappen for «gruppe 1» og deretter gruppen for «data».

Eventuelt kan du skrive inn urlen under for å gå direkte til mappen der datasettet ligger:

https://github.com/martigso/stv4020aR/tree/master/Gruppe%201/data

Oppgave 2:

Bruk R-funksjoner til å undersøke følgende

a) Er det noen missingverdier i datasettet?
b) Hvilken «klasse» har variabelen «bluecol»
c) Hva er median, gjennomsnitt, standardavvik, skjevhet (skewness) og kurtose (kurtosis) for variabelen «lwage»?

Oppgave 3:

lag en ny dummy-variabel i datasettet, «bluecol.d», slik at de som er blue collar workers har verdien 1, og de som ikke er blue collar workers har verdien 0. Vis hvordan du kan sjekke at den nye variabelen er omkodet til riktige verdier, og at den nye variabelen er numerisk

Oppgave 4:

Lag et histogram for «lwage». Lag deretter et scatterplot med «ed» på x-aksen og «lwage» på y-aksen. Finn til slutt korrelasjonen mellom «lwage» og «ed».

Oppgave 5:

a) Kjør en lineær regresjon med «lwage» som avhengig variabel, og «ed» som uavhengig variabel.
b) Lag et nytt datasett, «**data2**», bestående av dem som er ugift. Bruk deretter det nye datasettet, «**data2**», til å kjøre en lineær regresjon med «lwage» som avhengig og «ed» som uavhengig variabel
c) Lag et nytt datasett «**data3**», bestående av observasjoner i «**data2**» som ikke får lønnen sin bestemt av en fagforeningskontrakt. Bruk deretter det nye datasettet, «**data3**» til å kjøre en lineær regresjon med «lwage» som avhengig og «ed» som uavhengig variabel.
d) Kontroller at «**data3**» er konstruert riktig, ved å sjekke om datasettet kun består av personer som er ugift og som ikke får lønnen sin bestemt av fagforeningskontrakter.
e) Sammenlign regresjonskoeffisientene for «ed» fra de 3 modellene over. For hvilken gruppe ser det ut som om et års ekstra utdanning fører til størst økning i lønn? For hvilken gruppe ser det ut som om års ekstra utdanning fører til lavest økning i lønn? Kommenter.

Oppgave 6:

Opprett et nytt datasett med utgangspunkt i datasettet du lastet inn i oppgave 1, «**data4**», som kun består av variablene «lwage» og «ed». Vis deretter hvordan du fjerner «**data4**».

Oppgave 7:

a) Kjør en lineær regresjonsanalyse med «lwage» som avhengig variabel og med «ed», «south», «bluecol», «married», «union», «exp», «black» og «wks» som uavhengige variabler.
b) Lag en ny variabel i datasettet du lastet inn i oppgave 1, som du kaller «wage», ved å ta eksponentialfunksjonen av «lwage». Kjør en ny lineær regresjonsanalyse som er lik regresjonen i oppgave a, bortsett fra at du bruker wage som avhengig variabel i stedet for «lwage».
c) I oppgave b) måles den forventede effekten av et år ekstra utdanning på lønn i dollar. I oppgave a) er den forventede effekten av et års ekstra utdanning på lønn tilnærmet lik *den prosentvise endringen i lønn,* dersom vi ganger koeffisienten for utdanning med 100.

Vurder (subjektivt) om den estimerte effekten av utdanning på månedslønnen gjør det verdt å ta et ekstra års utdanning, dersom utdanningen koster 300 $. Skriv en kort kommentar

d) Lag histogram med residualene fra regresjonsmodellene i oppgave a) og b). Beregn også kurtose og skjevhet (skewness) for residualene fra de to modellene. Er residualene for modellen fra oppgave a) noenlunde normalfordelt? Er residualene fra oppgave b) noenlunde normalfordelt? Hvilken modell gir residualer som er mest normalfordelte?

Oppgave 8:

Kjør en lineær regresjonsanalyse med «lwage» som avhengig variabel og med «ed», «south», «bluecol», «married», «union», «exp», «black» og «wks» som uavhengige variabler. Legg også inn samspill mellom «married» og «ed» og «union» og «ed» som uavhengige variabler. Gir et års ekstra utdanning en sterkere effekt for dem som er gift enn dem som er ugift? Hvor stor forskjell er det i effekten av et års ekstra utdanning på «lwage» for en ugift person som ikke får lønnen sin fastsatt gjennom fagforeningskontrakter, og en gift person som får lønnen sin fastsatt gjennom fagforeningskontrakter?

Oppgave 9:

Kjør en lineær regresjonsanalyse med «lwage» som avhengig variabel og med «ed», «south», «bluecol», «married», «union», «exp», «black» og «wks» som uavhengige variabler. Legg også inn andregradsledd for effekten av utdanning (med andre ord: «ed» kvadrert/opphøyd i andre).

Tilsier modellen at effekten av utdanning er ikke-lineær?