R-prøve 2 STV 1020 V18, fredag 11. mai 12.15-14.00

Kodebok:

Datasettet består av 7425 observasjoner fordelt på 5 variabler. Observasjonsenhetene er individer, og kommer fra en kanadisk spørreundersøkelse. Det er en del missing i datasettet

Variabler:

wages	Average hourly wage rate (from all sources of employment) in Canadian dollar.
education	Number of years of schooling.
Age	in years
Sex	Female, Male.
language	A factor with levels: English, French, Other.

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 fredag» 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 wages.csv eller wages.Rdata fra data-mappen på github (https://github.com/langoergen/stv1020R/tree/master/data) som et objekt i R-Studio
2) Finn ut hvor mange kvinner og menn som snakker fransk.
3) Hvor mange observasjoner har missing på variabelen wages? Hvor mange observasjoner har missing på minst en variabel totalt?
4) Lag et histogram som viser fordelingen til wages. Opprett deretter en ny variabel ved å ta den naturlige logaritmen til wages, wages_log, i datasettet ditt, og lag et nytt histogram. Hvilket histogram ser mest normalfordelt ut?
5) Lag to nye datasett, ett med menn over 50, og ett med kvinner under 50. Hva er medianlønnen for menn over 50, og hva er medianlønnen for kvinner under 50?
6) Lag et spredningsplot (scatterplot) med alder på x-aksen, og timelønn på y-aksen. Tegn deretter en lineær regresjonslinje oppå plottet.
7) Lag en korrelasjonsmatrise mellom alder, timelønn og år med utdanning. Tolk en av sammenhengene substansielt, og gjør en signifikanstest av denne sammenhengen. Fjern alle observasjoner som har missing på en eller flere av variablene som inngår i matrisen.
8) Kjør en lineær regresjonsanalyse med timelønn som avhengig variabel, og utdanning, alder, kjønn og språk som uavhengige variabler. Tolk effekten av ett år ekstra med utdanning substansielt. Er effekten statistisk signifikant?
9) Kjør en lineær regresjonsanalyse med timelønn som avhengig variabel, og utdanning, alder, kjønn, språk og et samspillsledd mellom alder og kjønn som uavhengige variabler. Hva er den forventede timelønnsforskjellen mellom en mann på 20 og en kvinne på 20? Hva er den forventede timelønnsforskjellen mellom en mann på 60 og en kvinne på 60? Oppgi også hvor mange observasjoner som fjernes i regresjonsmodellen på grunn av missing-verdier. Gjør til slutt en test for multikolinearitet (du kan bruke en funksjon fra pakken «car»). Hvilke to variabler har høyest multikolinearitet?
10) Opprett en ny variabel i datasettet ditt, cohort, slik at alle som har alder under 30 får verdien 1, alle som har alder 30-39 får verdien 2, alle som har alder 40-49 får verdien 3, alle som har alder 50-59 får verdien 4, mens alle som er eldre enn 59 får verdien 5. Aggreger deretter data på kjønn og den nye cohort-variabelen – beregn enten gj.snitt eller median for education og wages i aggregeringen. Lag til slutt et spredningsplot,

med utdanning på x-aksen, timelønn på y-aksen. Fargelegg punktene ut fra verdien på kjønn, og bruk facet_wrap() til å lage separate paneler i plottet for de forskjellige verdiene på cohort-variabelen.