R-prøve 1 STV 1020 V18, fredag 11. mai 10.15-12.00

Kodebok:

Datasettet består av 45 observasjoner fordelt på 5 variabler. Observasjonene er av yrker, og er basert på en kanadisk spørreundersøkelse.

Variabler:

occupation	Name of occupation
type	Type of occupation. A factor with the following levels: prof, professional and managerial; wc, white-collar; bc, blue-collar.
income	Percentage of occupational incumbents in the 1950 US Census who earned $3,500 or more per year (about $36,000 in 2017 US dollars).
education	Percentage of occupational incumbents in 1950 who were high school graduates (which, were we cynical, we would say is roughly equivalent to a PhD in 2017)
prestige	Percentage of respondents in a social survey who rated the occupation as “good” or better in prestige

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 work.csv eller work.Rdata fra data-mappen på github (https://github.com/langoergen/stv1020R/tree/master/data) som et objekt i R-Studio
2) Hvilken klasse har variablene i datasettet?
3) Hva er maksimums og minimumsverdien til prestige? Vis hvordan du kan indeksere deg frem til observasjonene som har disse verdiene i datasettet ved hjelp av en logisk test.
4) Lag et spredningsplot (scatterplot) mellom inntekt og utdanning. Fargelegg punktene i plottet ut fra verdi på variabelen «type».
5) Lag en korrelasjonsmatrise mellom inntekt, utdanning og prestisje. Hvilken variabel korrelerer sterkest med prestisje?
6) Opprett en ny variabel i datasettet, type2, ved å omkode de observasjonene som jobber som politimenn og kokker til «prof», la andre observasjoner på den nye variabelen type2 beholde verdien de har på variabelen «type».
7) Lag et boxplot med type på x-aksen og prestisje på y-aksen. Hvilken gruppe ser ut til å ha minst variasjon i inntekt?
8) Estimer en regresjonsanalyse med prestisje som avhengig variabel og inntekt, utdanning og type yrke som uavhengige variabler. Lagre modellen som et objekt. Tolk effekten av en enhets økning i inntekt og en enhets økning i utdanning substansielt. Hvor høy er forklart varians (R i annen)?
9) Opprett to nye variabler i datasettet, en med forventet verdi (fitted.values) og en med residualer fra modellen i oppgave 8. Lag deretter et plot for å se på innflytelsesrike observasjoner (du kan bruke en funksjon fra pakken car). Indekser deretter følgende observasjoner: De to observasjonene med de høyeste studentiserte residualene, og de to observasjonene med høyest innflytelse (hat-value). Hva kjennetegner variabelverdiene og residualene til de to observasjonene med størst residualer? Hva kjennetegner variabelverdiene og residualene til de to observasjonene med høyest hat-value? Estimer til slutt to nye regresjonsmodeller med samme variabler som i oppgave 8, en uten de to observasjonene med høyest hatte-verdi, og en uten de to observasjonene som har de høyeste studentiserte residualene. Endres resultatene fra oppgave 8?