R-prøve STV 4020A, seminar 1

Datasett: test4020a.RData eller test4020a.csv

Prøveresultater fra 420 amerikanske skoler fra 45 counties i 1999.

Variabler:

Distcod	district code
County	county
District	Name of school district
Grspan	grade span of district
Teachers	Number of teachers
mealpct	Percent of students qualifying for reduced-price lunch
testscr	average test score (read.scr+math.scr)/2
Compstu	computer per student
Expnstu	expenditure per student
Str	student teacher ratio
Avginc	district average income
Elpct	percent of English learners
Readscr	average reading score
mathscr	average math score

Instruksjoner:

• Prøven skal besvares med et fungerende R-script som lastes opp i innleveringsmappen på Fronter. Innleveringsmappen finner dere i arkiv-mappen i Fronter-rommet for seminargruppen.

• 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. Lykke til!

Oppgaver:

1) Importer data fra arkiv-mappen i Fronter-rommet for seminargruppen eller fra github (https://github.com/martigso/stv4020aR/tree/master/Gruppe%201/data), og last inn i R (velg mellom test4020a.csv og test4020a.RData, datasettene er like).
2) Opprett en ny variabel i datasettet, compstu.d, slik at observasjoner som har lavere verdi på compstu enn .1 blir satt til 0 på den nye variabelen. Resten av observasjonene skal du gi verdien 1 på compstu.d. Bruk kode for å sjekke om omkodingen fungerte, forklar hvordan du kontrollerer omkodingen.
3) Lag et nytt datasett, som du kaller la, bestående av observasjoner fra Los Angeles. Hvor mange observasjoner er det i det nye datasettet? Er medianen til avginc høyere eller lavere enn medianen i hele datasettet?
4) Lag et nytt datasett, «cordata», bestående av følgende variabler: testscr, compstu, expnstu, mealpct, str og elpct. Lag også en korrelasjonsmatrise som inneholder disse variablene. Hvilken variabel korrelerer sterkest positivt med testscr? Hvilken variabel korrelerer sterkest negativt med testscr
5) Lag et scatterplot mellom testscr (y-akse) og elpct (x-akse). Kommenter sammenhengen mellom variablene (du kan også bruke informasjon fra forrige oppgave).
6) Finn mean, sd, skewness og kurtose for testscr. Lag også et histogram for testscr. Det kan være nyttig å bruke en pakke til å løse deler av denne oppgaven, men det er også mulig å regne alt ut for hånd.
7) Kjør en lineær regresjon med testscr som avh. var, og compstu, expnstu, mealpct, str og elpct som uavhengige variabler. Kjør deretter en ny modell med samspill mellom expnstu og mealpct. Lagre modellene som objekter.
8) Vis hvordan du kan lage et histogram med residualene fra modellene. Lag deretter qq-plot basert på modellene (du trenger ikke å bruke en pakke til dette, det finnes en enkel funksjon i base-R som gjør jobben). Er residualene noenlunde normalfordelt? Forklar.
9) Kjør en flernivåanalyse med variabelen testscr som avhengig variabel. Den hierarkiske strukturen: de 420 skolene er lokalisert i 45 counties (variabelen county). Kjør først en modell med bare random intercept, og skriv ned variansen på nivå 1 og 2 i en kommentar. Sett deretter opp en modell med random intercept, samme gruppering og avh.var, og med random slope for elpcts, og ellers de samme kontrollvariablene som i den første lineære regresjonsmodellen i oppgave 7. Du trenger en pakke for å løse denne oppgaven.
10) Sammenlign de to modellene fra forrige oppgave med utgangspunkt i log likelihood, hvilken modell passer best til data?