**Tips til øvingene**

**Kapittel 2**

1. Hvordan en legger noter til en Stata-datafil (datasett), er forklart i kapittel 1, side 20.

2. Det første spørsmålet krever en vurdering av frekvensfordelingen for variabelen som lages med kommandoen tabulate. Det andre spørsmålet kan besvares ved å krysstabulere de to utdanningsvariablene med kommandoen tabulate der begge variablene angis.

3. Omkoding av alder i 5 like store kategorier. Først må en finne de 4 aldrene som gir 5 like store kategorier. Det enkleste er vel å tabulere alder og finne kuttpunktene i den kumulative frekvensfordelingen. Litt mer elegant er det å finne prosentilene. Dette krever imidlertid bruk av kommandoen summarize som først blir gjennomgått i neste kapittel. Deretter lages den nye variabelen ved å sette inn aldersintervallene i parentesene i denne kommandoen:

recode alder ( .. = 1)( .. =2)(.. =3)(.. = 4)( .. =5), gen(alder5)

4. Hvordan dette gjøres, er forklart på side 58.

**Kapittel 3**

1. Filen kan åpnes på flere måter, blant annet fra Fagbokforlagets server:

use "https://nedlasting.fagbokforlaget.no/stata/abu89.dta", clear

Kommandoen summarize gir beskrivende statistikk for fexp, og tabulate kan benyttes for å lage frekvensfordelingen.

De to diagrammene som er aktuelle her, er histogram og boksdiagram. Begge er forklart i kapittel 3 fra side 70.

2. Hvordan stolpediagram og histogram lages, er forklart i kapittel 3, side 67–72. For å besvare spørsmålet er nøkkelen å avgjøre om ed (antall år utdanning) er en kontinuerlig variabel eller ikke.

**Kapittel 4**

1. Krever vel ingen tips, men start med å lage nye dikotome variabler av alder4 og utdniv. Benytt kommandoen på denne formen:

recode gammel (.. =0)(.. =1), gen(ny)

Deretter benyttes tabulate til å lage krysstabellene med aktuelle korrelasjonsmål.

2. Også her kan en benytte samme analyseopplegg som for tabellanalysen i kapittel 4. Først må en klargjøre variablene og avgjøre hvor mange verdier (kategorier) de bør ha. I filen finnes intulik og utdniv, hver med fire kategorier. For å vurdere samspill er det best å benytte korrelasjonsmål.

**Kapittel 5**

a. Skal fexp benyttes som en faktor i en toveis variansanalyse, må den omkodes til en kategorivariabel med et lite antall (f.eks. 3) verdier (kategorier).

b. Gjennomfør en toveis variansanalyse med bare hovedeffekter:

anova promot female fexp3

I tillegg til å se på F-verdiene kan spørsmålet besvares mer presist med postestimeringskommandoen estat esize. Eta-kvadrert vil gi et klart svar på spørsmålet.

c. Legg til samspillet i anova-kommandoen og gjennomfør analysen på nytt. F-testene som følger, viser at svaret er ja. Det er videre fornuftig å vurdere effektstørrelsen både ved økningen i R2 og ved bruk av postestimeringskommandoen ovenfor.

d. Vurder forutsetningene gjengitt på side 105 i boka. Er den avhengige variabelen kontinuerlig? Holder forutsetningen om at variansen i den avhengige variabelen er lik i gruppene etter de uavhengige variablene? Dette gjøres med postestimeringskommandoen estat hettest. I Statistikk-menyen finnes den under Postestimation statistics for anova.

**Kapittel 6**

1. Regresjonsanalyse av bedriftserfaring. Dette er en standard multippel regresjonsanalyse som kan gjennomføres på samme måte som analysen av timelønn. Spørsmålene a, b og d besvares med en regresjon av fexp etter female, age, ed, private. Spørsmål e krever at samspillet mellom kjønn og utdanning legges til, f.eks. på denne måten: female#c.ed

Spørsmål c kan besvares ved bruk av lowess-kommandoen

Påstandene i a og c bør blir bekreftet, mens påstandene i b, d og e ikke får støtte.

f. Benytt margins og marginsplot til grafisk framstilling av samspillet (se eksempel s. 136).

2. Forutsetningene om modellspesifisering er gjengitt på side 126 med påfølgende oppskrifter for dem som kan undersøkes. På side 136 er forutsetningen om residualene gjengitt med påfølgende oppskrifter for forutsetningene 2 og 4. I tillegg kan det vurderes om ekstremverdier har innflytelse på resultatene. Dette er forklart på side 140–146. Endelig er problemet med multikollinearitet gjennomgått på side 146–148.

3. Regresjonsanalyse av sosial (generell) tillit

Først må filen åpnes (merk at det er en trykkfeil i filnavnet i oppgaven der tallet «7» er falt ut).

use "https://nedlasting.fagbokforlaget.no/stata/ESS7NO\_Pol2.dta", clear

a. Bruk tab1 ppltrst pplfair, missing til å bekrefte at de to tillitsvariablene har det samme verdisettet. Dette viser at begge variablene varierer fra 0 til 10 der 10 betyr høy tillit til folk, og at manglende informasjon (missing data) ikke er noe problem. Er variablene korrelerte og trekker i samme retning? Kommandoen corr ppltrst pplfair vil vise at (Pearsons-) korrelasjonen er 0,56. Dette er rimelig tilfredsstillende. Cronbachs  kan også sjekkes med kommandoen:

alpha ppltrst pplfair

Dette gir også tilfredsstillende resultat, og skalaen kan nå konstrueres som et gjennomsnitt av de to variablene:

egen tillit = rowmean (ppltrst pplfair)

b. En regresjonsanalyse av sosial tillit

Bakgrunnsvariabler: kjønn, alder, utdanning

Blant andre mulige uavhengige variabler: toleranse målt med freehms.

c. Bruk margins og marginsplot til å framstille effekten av en av de uavhengige variablene grafisk (se eksempel på s. 129–130).

d. Vurder om regresjonsmodellens forutsetninger er tilfredsstilt. Gjør de mest vanlige sjekkene av: linearitet og additivitet, homoskedastisitet og fordelingen av residualene.

e. Rapportere resultatene i en APA-tabell. Eksempler finnes på nett ved å søke etter «apa 6th table format». Denne nettadressen har eksempler: <https://graduateschool.camden.rutgers.edu/writing/apa-styling-general-overview/tables/>

**Kapittel 7**

1. Én eller flere skalaer (indekser) for politisk tillit?

Aller først bør en se på variablenes fordelinger: summ trstprl-trstprt

a. Faktoranalysen (eksplorerende) er det viktigste redskapet her. Husk at du bør velge denne kommandoen samt nødvendige postestimeringskommandoer.

factor trstprl-trstprt, pcf

Dette er egentlig nok til å svare på spørsmålet om én eller to dimensjoner. Se på egenverdiene, og lag gjerne et skredplott. En endimensjonal løsning krever at bare den første faktoren har en egenverdi over 1,0, og at den forklarer mye, helst mer enn 50 % av variansen i variablene som inngår i analysen. I tillegg må vi vurdere om det er teoretisk mest rimelig med én eller to dimensjoner.

Cronbachs  kan i tillegg vise om skalaen har tilstrekkelig reliabilitet (> 0,7). Denne kommandoen estimerer  i en skala der alle fem spørsmålene er med:

alpha trstprl-trstprt

b. Start med en CFA-modell med én faktor. Følg oppskriften i kapittelet. Bruk gjerne SEM Builder. Kriterierelatert validitet kan vurderes ved å utvide CFA-modellen med en strukturmodell slik som i modellen på side 184. Alternativt lages én eller to tillitsskalaer ved hjelp av kommandoen egen, og disse korreleres med variabler det er teoretisk rimelig skal ha en sammenheng med politisk tillit.

2. Utradisjonell politisk aktivitet. Beskrivende statistikk vil vise at problemet med manglende informasjon er ubetydelig. Alle variablene har verdiene 1 og 2. Lag gjerne nye variabler med verdiene 0 og 1 før du går videre.

a. Som i øving 1 er det naturlig å starte med en eksplorerende faktoranalyse. Hvis analysen gir mer enn én faktor, må en rotere løsningen med en postestimeringskommando og gjerne undertrykke svake faktorladninger:

rotate, blanks(.3)

Cronbachs kan estimeres med kommandoen:

alpha *variabelliste*, item

b. Lag CFA-modeller med én og to faktor(er) og vurder deres modelltilpasning.

c. De 7 variablene (contplt – bctprd) har bare to verdier hver. Tilfredsstiller dette kravet til målenivå? Se gjerne framstillingen av den lineære sannsynlighetsmodellen i neste kapittel. Alternative løsninger peker egentlig utover temaet for denne boka, men sjekk help gsem.

d. Dette besvares enklest ved å lage én eller to skala(er) ved hjelp av kommandoen på denne formen:

egen skala1 = rowmean *variabelliste*

Deretter kan fordelingen(e) beskrives ved statistiske mål og framstilles grafisk i histogram og boksdiagram.

**Kapittel 8**

1. Forutsetninger og innflytelsesrike observasjoner i eksempel 1.

Forutsetningene som skal undersøkes, er gjengitt på side 212. De er stort sett samsvarende med forutsetningen for lineær regresjonsanalyse. Framgangsmåten er vist for eksempel 2 i boka. I tillegg bør modellen vise en god tilpasning. Hosmer-Lemeshows test og ROC-kurven kan benyttes her.

Framgangsmåten for å undersøke forutsetningene som går på modellens spesifisering, spesielt om sammenhengene er lineære og additive, er den samme som for lineær regresjon.

Blant andre forutsetninger er kravet om at det ikke må være perfekt diskriminering spesielt for logistisk regresjon. Les nøye forklaringen i boka. Symptomene er unormalt store oddsratioer som ikke er statistisk signifikante.

Analyse av residualer og innflytelse i logistisk regresjon er inspirert av tilsvarende analyser i vanlig lineær regresjon, men de statistiske størrelsene og diagrammene har gjerne andre navn. Når det gjelder innflytelse, knyttes dette til kovariatmønstre og ikke til individuelle observasjoner som ved vanlig regresjon. Ellers er det bare å følge framgangsmåten som er benyttet for eksempel 2 i kapitlet.

2. Gir den lineære sannsynlighetsmodellen samme resultat som logistisk regresjon i eksempel 2 der dep, med verdiene 0 og 1, er den avhengige variabelen? Løsningen på dette er å estimere modellen i boka med både logistisk regresjon og med vanlig regresjon og benytte margins til å beregne predikerte sannsynligheter for grupper etter forklaringsvariablene. Dermed gjenstår det å sammenligne resultatene og se om de samsvarer eller ikke.

Spørsmålet om å teste likheter i regresjonskoeffisientene for kvinner og menn kan besvares ved å gjøre separate regresjonsanalyser for kvinner og menn og benytte kommandoen suest, med påfølgende tester av likheter mellom parameterestimatene. I boka (s. 218–220) er denne framgangsmåten forklart. Predikerte sannsynligheter for grupper av de andre uavhengige variablene for menn og kvinner finnes ved hjelp av kommandoen margins.

3. Sammenlign vanlige regresjonsanalyser (med de samme uavhengige variablene) av dep (dikotom) og cesd8 (kontinuerlig). Svaret på spørsmålet bør baseres på sammenligning av regresjonskoeffisientenes fortegn og statistisk signifikans. De standardiserte regresjonskoeffisientene kan også sammenlignes.

Det andre spørsmålet her er identisk med spørsmål 2.

**Kapittel 9**

1. Bruk datafilen wage89 til en flernivåanalyse av timelønn. I oppgaven i boka mangler «89» i filnavnet.

a. Kjønnsforskjellen i timelønn er regresjonskoeffisienten til female. Spørsmålet er om denne regresjonskoeffisienten varierer statistisk signifikant mellom virksomhetene. Denne kommandoen kan benyttes:

mixed wage edyears c.age##c.age female || firmno: female, cov(un)

Estimatene bør lagres med: estimates store rand

Deretter estimeres modellen der *ikke* regresjonskoeffisienten til female varierer mellom virksomhetene, og estimatene lagres med: estimates store uten

Deretter utføres en LR-test slik som forklart på side 237: lrtest uten rand

Du bør finne at testen gir et statistisk signifikant resultat.

b. Gir en enklere variabel for virksomhetsstørrelse som bare skiller mellom små og store virksomheter, en annen konklusjon enn bruk av lnsize? Virksomhetsstørrelse (size) varierer fra 1 til 5414. Hvis vi for eksempel velger 200 som skille, blir resultatet at rundt 70 % blir definert som små, mens 30 % blir definert som store virksomheter. Det kan være lurt å prøve 2–3 ulike kriterier. Denne kommandoen benytter 200 som skille:

recode size (1/199=0)(200/5414=1),gen(stor)

c. Forutsetningen om linearitet for lnsize. Jeg ville estimert lowess-linjen og eventuelt testet om andregradsleddet til lnsize var statistisk signifikant. Dette bør gi et klart svar på spørsmålet.

d. Har lnsize store verdier som er innflytelsesrike?

Dette spørsmålet kan besvares på enten en enkel eller en komplisert måte. Bare den enkle måten kan benyttes i mixed. Denne er også den beste måten siden spørsmålet er knyttet til en spesifisert variabel. Det første trinnet er å ta utgangspunkt i fordelingen til lnsize og lage en ny variabel som har verdien 1 for spesielt store virksomheter, og som har verdien null for alle andre. En mulighet er å velge de største bedriftene som til sammen har rundt 5 % av utvalget. Dette gir et skille på lnsize = 7,5. Benytt recode-kommandoen slik som i oppgave b til å lage den nye variabelen som jeg kaller stor2. Deretter utføres mixed-kommandoen på nytt uten de store virksomhetene med dette tillegget: if stor2==0. Dette gir en flernivåanalyse med 3557 nivå-1-enheter mot 3759 for modellen i boka. Prøv å sette skillet noe lavere og se om det gjør noen forskjell.

Den kompliserte måten å svare på er å benytte teknikkene fra kapittel 6, men da må også vanlig regresjon benyttes fordi de fleste statistiske størrelsene og diagnosediagrammene ikke kan lages fra mixed.

e. Dette er spørsmål om det er samspill mellom variablene private og edyears. Dette besvares ved å legge samspillsleddet inn i en flernivåmodell, for eksempel den på side 240.

f. Hvordan samspillseffekter framstilles grafisk, er forklart i kapittel 6, side 136.

g. Er de to stokastiske effektene normalfordelte? Modellen dette gjelder, er gjengitt på side 241 i boka. Estimer denne, og benytt deretter tipset i oppgaveteksten for hvordan de stokastiske effektene kan lagres i filen. Disse kan benyttes til å gjenskape regresjonskoeffisientene og regresjonskonstantene i virksomhetene. Lag til slutt to histogrammer med innlagt normalfordelingskurve. Du bør finne at fordelingen av de stokastiske effektene for regresjonskoeffisienten til utdanning kommer nærmest normalfordelingen.