Løsningsforslag

Oppgavesettet og løsningsforslaget er utarbeidet av Jarle Møen. Det anbefales at sensorene gir inntil 10 poeng per delspørsmål.

Oppgave 1

a) Både analyse 1 og 2 tester en nullhypotese om lik varians, men i analyse 1 er det tastet inn feil tall (standardavvik istedenfor varians, og antall observasjoner er feil). Fra analyse 2 ser vi at vi klart kan forkaste en hypotese om lik varians For FMNC og DMNC.

b) Både analyse 3 og 4 tester en nullhypotese om like gjennomsnitt – med relevante gjennomsnitt og antall observasjoner inntastet. Analyse 3 forutsetter lik varians i de to utvalgene og analyse 4 forutsetter ulik varians. Dette ser vi enklest av frihetsgradstallet som er 21653=15971+5684-2 i analyse 3 og vesentlig mindre i analyse 4. Siden vi i a) har konkludert med ulik varians bør vi bruke analyse 4. Da kan vi nesten forkaste en nullhypotese om like gjennomsnitt på 5 % signifikansnivå. [Har man bommet i a) bør man velge analyse 3 som med en p-verdi på 0,076 er lengre fra å forkaste nullhypotesen på 5 %-nivå.]

c) Vi kan bruke enveis variansanalyse (enfaktor ANOVA), Kruskal-Wallis-testen eller regresjon med dummier for FMNC og DMNC.

d) Disse formlene kan hentes direkte fra forelesningsnotatene, men med s i stedet for p, T i stedet for I og indeks i i stedet for j.

$ET_i = s \cdot 1 + (1-s) \cdot 0 = \mathbf{s}$

$Var(T_i) = E(T_i^2) - (ET_i)^2 = E(T_i) - (ET_i)^2 = s - s^2 = \mathbf{s(1-s)}$

Variansen til andelen i skatteposisjon blir

$$Var\left(\frac{\sum_{i=1}^{n} T_i}{n}\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(T_i) = \frac{n \cdot s(1-s)}{n^2} = \frac{s(1-s)}{n}$$

e) Vi setter inn den estimerte andelen, 0,721 for s og antall observasjoner, 79 815, for n. i formelen over og får at standardavviket er $\sqrt{0{,}721(1-0{,}721)/79815} = \underline{0{,}00159}$.

f) Kvartilbredden er avstanden mellom 75 og 25-prosent persentilene, altså

16,66-(-0,11)=<u>16,77.</u>

Oppgave 2

a) Det er to grunner til at biler taper seg i verdi over tid. Det ene er bruksslitasje – denne kan vi måle ved antall kjørte kilometer – den andre er hva vi kan kalle ren aldring. Den rene alderseffekten fanger opp at teknologien blir utdatert og at enkelte komponenter på bilene blir dårligere over tid uavhengig av bruk. Antall kjørte kilometer vil åpenbart være sterkt korrelert med bilens alder. Når kilometerstanden er utelatt vil derfor den inkluderte aldersvariabelen fange opp begge effektene. Når vi inkluderer kilometerstanden som en egen variable vil bilens alder bare fange opp ren aldring (effekten av at bilen står i garasjen et år) og dette estimatet blir mindre enn den kombinerte effekten. Oppsummert: I kolonne (1) har vi en utelatt variabel som er positivt korrelert med den inkluderte aldersvariabelen, og vi får et estimat som fanger om ren aldring pluss den gjennomsnittlige bruksslitasjen som følger med alderen.

b) Tolkningen av alderskoeffisienten i kolonne (2) er at prisen faller med 13 508 kroner per år. Tolkningen av alderskoeffisienten i kolonne (3) er at prisen faller med ca. 9,5 % per år.

c) Et prisfall på 13 508 er 9,6 % av prisen på en gjennomsnittsbil, og 9,5 % av gjennomsnittsprisen er 13 306. Vi ser altså at estimert prisfall for en gjennomsnittsbil er svært lite sensitivt for valg av funksjonsform.

d) Fra kolonne (2) har vi at E(Pris)= 208152-(13508*5)-(313*75)=117137.

Fra kolonne (3) har vi at E(lnPris)= 12,283-(0,0954*5)-(0,00189*75)= 11,664.

e) Vi trenger feilleddsvariansen, og finner den ved å kvadrere standardavviket som er oppgitt i nederste rad i kolonne (3). Da får vi

E(Pris)= $e^{11,664+0,5*0,135*0,135}$= $e^{11,673}$=117 360. Vi noterer oss at denne er svært lik forventet pris med lineær spesifikasjon i kolonne (2).

f) De fire øverste figurene sjekker normalitet i feilleddet. Vi ser at begge feilleddsfordelingene er klokkeformet, men viser klare avvik fra normalitet. Fra de to øverste figurene ser vi at feilleddet i regresjon (2) har størst testobservator – og dermed sterkest avvik fra normalitet, men samtidig ser vi at denne fordelingen er mer symmetrisk enn den andre. De to nederste figurene hjelper oss å ta stilling til om feilleddene er homoskedastiske. Det er ikke noe åpenbart heteroskedastisitetsmønster verken i regresjon (2) eller (3), men man kan kanskje ane at det er litt større feilleddsvarians for biler med lav forventet pris enn for biler med høy forventet pris i regresjon (3). Siden det også er enklere å tolke prediksjonene når vi får de ut i kroner direkte, taler det for å satse videre på spesifikasjon (2). Om man på den annen side gjerne vil ha koeffisienter med prosent-fortolkning, er det så lite som skiller de to spesifikasjonene at man like gjerne kan jobbe videre med de log-transformerte prisene.

g) Alder og kilometer er allerede grundig behandlet, men vi ser at begge estimatene blir litt lavere når vi inkludere flere kontrollvariabler. Subaru Impreza skiller seg ikke signifikant fra Suzuki Baleno i pris, men Citroen Berlingo er i gjennomsnitt 17 000 billigere, og Renault Scenic er i gjennomsnitt 40 000 dyrere. Biler som annonseres som billige har en forventet rabatt på 5000, men effekten er ikke signifikant. Bruktbiler som selges gjennom forhandler er 19 500 kroner dyrere enn de som selges privat, biler med større motor enn standard har et gjennomsnittlig pristillegg på 9000 kroner og firehjulsdrift har et gjennomsnittlig pristillegg på 16000 kroner. Alle disse effektene er signifikante og så langt er alle fortegn som forventet. Den store overraskelsen er at biler som averteres med ekstrautstyr er 10 500 kroner **billigere** enn andre biler, alt annet likt. Det er vanskelig å finne en overbevisende forklaring på dette, og det kan være en anomali. (Konstantleddet er bokstavelig fortolket estimert pris på en helt ny Suzuki Baleno 2WD med standard motor som selges brukt, privat, uten ekstrautstyr, men det er liten grunn til å vie det oppmerksomhet.) Forklaringsgraden er 87 %, og den forventede prediksjonsfeilen er 14 000. Det taler for at vi har en god modell. (Det kan likevel være naturlig å kaste ut variablene «billig» og «D_ekstra» før modellen tas i bruk.)

h) Vi kan teste om firehjulstrekk påvirker verditapet ved å inkludere en interaksjon mellom variabelen alder og dummyen for firehjulstrekk – altså en ny variabel (Alder*Fourwd)

i) Antall ulykker påvirkes ikke bare av bilens sikkerhet og veigrep, men også av bilens bruk og sjåførens atferd. De to siste faktorene vil normalt være uobserverbare, men det er godt mulig at sjåfører som velger å ha bil med firehjulstrekk både bruker bilen til mer krevende kjøring (det er derfor de trenger 4WD) og er mer risikosøkende (eller tar mer risiko enn ellers fordi de stoler på bilens gode kjøreegenskaper). I en regresjon der man skal forklare ulykker med firehjulstrekk kan det altså være en uobserverbar komponent i feilleddet som er positivt korrelert med dummyen for 4WD, og det er mulig at denne uobserverbare, korrelerte effekten dominerer den isolert sett ulykkesreduserende effekten av 4WD. For forsikringsselskapene er det bare den samlede effekten som teller. De vil prise inn all risiko som er korrelert med bilens observerbare egenskaper. (Om modeller med 4WD gjennomgående har høyere forsikringspremie vites ikke, men all informasjon i oppgaven er reell.)