LØSNINGSFORSLAG MET4 V23

Til sensor: Hver deloppgave teller likt og gir maksimalt 10 poeng.

Oppgave 1

(a) For det første ser vi at dummyvariablene til de fleste bilmodellene, med få unntak, ikke har estimerte regresjonskoeffisienter som er statistisk signifikant forskjellig fra null. Koeffisienten til `log(price)` er negativ, og med logaritme på begge sider av likhetstegnet estimerer vi at en økning i prisen på 1% henger sammen med at det selges 1.1% færre biler av denne modellen. Men denne koeffisienten er ikke statistisk signifikant forskjellig fra null på 5% signifikansnivå. Det er heller ikke koeffisientene for fysisk størrelse og motoreffekt, drivstoffkostnad eller CO_2-utslipp. Det er naturlig her å tenke at disse variablene kan være sterkt korrelerte (dyre biler er store og har stor motor – og vice versa), mer om det i oppgave **(c)**. Vi ser at elektriske biler selger signifikant bedre enn fossilbiler, alt annet likt. Modellen har en forholdsvis høy forklaringsgrad på over 50%.

(b) Residualene ser ut til å være heteroskedastiske. Variansen øker med predikert verdi. Vi har ikke en tidsdimensjon, og dermed ikke et autokorrelasjonsplott. Det hadde kanskje ikke vært så dumt å se på likevel, dersom modeller fra samme bilprodusent ligger etter hverandre i datasettet. Det kan være avhengighet mellom observasjonene av den grunn. Det er noen få uteliggere til venstre i fordelingen. Ellers ser residualene ut til å være tilnærmet normalfordelt.

(c) Nei.

- For det første er det estimert en negativ sammenheng mellom pris og antall solgte. Det at den har p-verdi rett over 5% er ikke substansielt forskjellig fra om den hadde vært rett under 5%. Det er ingenting magisk med det terskelnivået.
- For det andre har vi allerede kommentert muligheten for multikolinearitet. Pris/størrelse/motoreffekt er gjerne positivt korrelert med hverandre som gjør at en lineær regresjonsmodell får problemer med å skille effektene fra hverandre. Da kan estimatene bli ustabile (få høy varians), som igjen fører til høyere p-verdier.
- For det tredje er det ikke sikkert at log-log sammenhengen mellom pris og antall solgte biler er en god beskrivelse av virkeligheten. Det *kan* være en sterk sammenheng mellom variablene som modellen vår ikke klarer å fange opp.
- For det fjerde kan vi uansett ikke konkludere med kausale sammenhenger ut fra statistisk signifikans i en multippel regresjonsmodell (eller i dette tilfellet: fravær av kausal sammenheng som følge av fravær av statistisk signifikans).

Til sensor: Her holder det med et godt argument som konkluderer riktig. Man trenger ikke liste opp alle grunnene som står over.

(d) Moms: $0.25 \cdot (1\,008\,990 - 500\,000) = 127\,247.50$. Vektavgift: $12.50 \cdot (2\,533 - 500) = 25\,412.50$. Total prisøkning: kr 152 660. Det er bare prisen som forandrer seg etter innføring av nye avgifter, så høyresiden på regresjonsligningen forandrer seg med $-1.11(\log(1\,161\,650) - \log(1\,008\,990)) \approx -0.157$. Det svarer til at vi forventer at logaritmen til antall solgte biler etter avgiftsøkningen ($\log(Y_2) - \log(Y_1)$) øker med -0.157. Fra regnereglene til logaritmer kan vi se at $\log(Y_2) - \log(Y_1) = \log(Y_2/Y_2) = -0.157$. Opphøyer vi begge sider i e får vi $Y_2/Y_1 = e^{-0.157} \approx 0.85$, slik at vi forventer at antallet solgte biler av denne modellen reduseres med om lag 15% ut fra denne modellen.

Argumentet over er "eksakt" i den forstand at vi ikke brukte "prosenttolkningen" til log-log-sammenhengen (som er en tilnærming). Vi kan gjøre det i stedet. Avgiftsøkningen fører til en prisøkning på $152\,660/1\,008\,990 = 15.1\%$, som i følge prosenttolkningen skal henge sammen med en endring i salgstall på $-1.11 \cdot 15.1\% = -16.8\%$. Vi får omtrent det samme svaret.

Til sensor: Begge disse utregningene gir like god uttelling.

Oppgave 2

(a) Dette histogrammet har særlig to iøynefallene karakteristikker; opphopningen ved null og seks Mwh. Opphopningen ved null MWh kan både skyldes timer hvor vindhastigheten er for lav til at vindturbinen kan produsere og timer hvor vindhastigheten er så høy at vindturbinen må stoppes. Opphopningen ved seks MWh skyldes timer hvor vindhastigheten er mindre enn terskelverdien for stans, men høy nok til å produsere ved makskapasitet.

Dette er sentralgrenseteoremet i praksis. Kraftproduksjonen i løpet av en måned er summen av kraftproduksjonen for alle timene i løpet av en måned. Både en sum og et gjennomsnitt vil være tilnærmet normalfordelt, bare på forskjellig skala. Derfor skal det mye mer til at vi får slike opphopninger vi så i histogrammet per time. En opphopning på 0 MWh ville f.eks bety at vindhastigheten var veldig lav eller over terskelverdien for produksjon alle timene i løpet av måneden. Tilsvarende er det lite sannsynlig vindmturbinen ville produsert på makskapasitet hver time hele måneden.

(b) Vi skal teste følgende nullhypotese om lik varians:

$$\sigma^2_{UN} = \sigma^2_{SN2} \quad \text{mot} \quad \sigma^2_{UN} \neq \sigma^2_{SN2}$$

der σ^2_{UN} og σ^2_{SN2} er populasjonsvariansen i kraftproduksjon for henholdsvis Utsira Nord og Sørlige Nordsjø 2. Her er det lurte å sette den største estimerte variansen (SN2) i telleren av testobservatoren, slik at en en kun trenger å sammenligne testobservatoren mot kritisk verdi i høyre hale av F-fordeling (0.975 percentilen).

Testobservatoren blir da

$$F = \frac{S^2_{SN2}}{S^2_{UN}} = 0.60^2/0.58^2 = 1.07$$

Kritisk verdi er gitt ved 0.975 percentilen i en F-fordeling med $288 - 1$ og $288 - 1$ frihetsgrader. Her kan vi bruke en tabell eller R:

```
qf(1 - 0.05/2, df1 = 288 - 1, df2 = 288 - 1)
```

```
[1] 1.260926
```

Siden $F = 1.07 < k^{287,287}_{0.975} = 1.26$, kan vi ikke forkaste nullhypotesen om lik varians.

(c) Det fremgår ikke av oppgaven at vi ønsker å teste i en spesifikk retning, så det er ikke noen spesiell grunn til å gjøre en ensidig test. For begge spørsmålene utfører vi derfor følgende tosidige hypotesetest:

$$\mu_{SN2} = \mu_{UN} \quad \text{mot} \quad \mu_{SN2} \neq \mu_{UN}$$

der μ_{UN} og μ_{SN2} er forventet kraftproduksjon per måned for Utsira Nord og Sørlige Nordsjø 2. Vi antar lik varians når vi skal utføre en to-utvalgs t-test, ut fra resultatet fra forrige oppgave.

Vi får følgende testobservator:

$$T = \frac{\overline{X}_{SN2} - \overline{X}_{UN}}{\sqrt{S^2_P\left(\frac{1}{n_{SN2}} + \frac{1}{n_{UN}}\right)}} = \frac{2.20 - 2.03}{\sqrt{0.3482\left(\frac{1}{288} + \frac{1}{288}\right)}} = 3.46,$$

der vi har regnet ut en felles varians ved hjelp av følgende formel:

$$S^2_P = \frac{(n_{SN2} - 1)S^2_{SN2} + (n_{UN} - 1)S^2_{UN}}{n_{SN2} + n_{UN} - 2} = \frac{(288 - 1)0.60^2 + (288 - 1)0.58^2}{288 + 288 - 2} = 0.3482$$

Under nullhypotesen er testobservatoren tilnærmet normalfordelt for så mange observasjoner så kritisk verdi for en tosidig test er 1.96. Testobservatoren er lik 3.45, så **vi forkaster nullhypotesen om lik forventet kraftproduksjon ved de to lokasjonene**.

(d) Antall vindturbiner som blir plassert på UN er da $5000 \cdot w$. Dersom vi antar at den ene vindturbinen er representativ for alle disse vil derfor den samlede produksjonen fra UN være $5000 \cdot w \cdot U$. Tilsvarende vil produksjonen for SN2 være $5000 \cdot (1-w) \cdot V$ og den totale produksjonen fra disse to lokasjonene blir derfor

$$5000 \cdot w \cdot U + 5000 \cdot (1-w) \cdot V = 5000[wU + (1-w)V] = Y$$

Variansen til denne variabelen er gitt ved

$$\begin{aligned}
\text{Var}(Y) =& \text{Var}(5000[wU + (1-w)V]) = 5000^2 \text{Var}(wU + (1-w)V) \\
&= 5000^2 \left(w^2 \text{Var}(U) + (1-w)^2 \text{Var}(V) + 2w(1-w)\text{cov}(U,V) \right) \\
&= 5000^2 \left(w^2\sigma_u^2 + (1-w)^2\sigma_v^2 + 2w(1-w)\sigma_{uv} \right) \\
&= 5000^2 \left(w^2\sigma_u^2 + \sigma_v^2 - 2w\sigma_v^2 + w^2\sigma_v^2 + 2w\sigma_{uv} - 2w^2\sigma_{uv} \right) \\
&= 5000^2 \left[(\sigma_u^2 + \sigma_v^2 - 2\sigma_{uv})w^2 + 2(\sigma_{uv} - \sigma_v^2)w + \sigma_v^2 \right]
\end{aligned} \tag{1}$$

(e) Vi finner vekten som minimerer varians ved å sette

$$\begin{aligned}
&\frac{\partial}{\partial w}\text{Var}(Y) = 0 \\
&\implies \frac{\partial}{\partial w} 5000^2 \left[(\sigma_u^2 + \sigma_v^2 - 2\sigma_{uv})w^2 + 2(\sigma_{uv} - \sigma_v^2)w + \sigma_v^2 \right] = 0 \\
&\implies 5000^2 \left[2(\sigma_u^2 + \sigma_v^2 - 2\sigma_{uv})w + 2(\sigma_{uv} - \sigma_v^2) \right] = 0 \\
&\implies 2(\sigma_u^2 + \sigma_v^2 - 2\sigma_{uv})w + 2(\sigma_{uv} - \sigma_v^2) = 0 \\
&\implies w = \frac{-2(\sigma_{uv} - \sigma_v^2)}{2(\sigma_u^2 + \sigma_v^2 - 2\sigma_{uv})} = \frac{\sigma_v^2 - \sigma_{uv}}{\sigma_u^2 + \sigma_v^2 - 2\sigma_{uv}}
\end{aligned} \tag{2}$$

Den empiriske variansen til Y får vi ved å erstatte populasjonsverdiene over med de empiriske motpartene som vi finner i tabellen fra oppgave a) (i tillegg til empirisk kovarians som blir oppgitt til å være 0.29). Vekten som minimerer denne er derfor gitt ved:

$$w = \frac{\hat{\sigma}_v^2 - \hat{\sigma}_{uv}}{\hat{\sigma}_u^2 + \hat{\sigma}_v^2 - 2\hat{\sigma}_{uv}} = \frac{0.60^2 - 0.29}{0.58^2 + 0.60^2 - 2 \cdot 0.29} = 0.60 \tag{3}$$

Det betyr at 60 % av vindturbinene da må plasseres på UN.

Til sensor: Rent formelt må man vise at dette faktisk er et bunnpunkt for variansen til Y, men vi legger ikke vekt på dette i sensuren.

Oppgave 2

(a) Fra autokorrelasjonsplottet ser vi at det er en tydelig sesongkomponent i månedlig kraftproduksjon. Den positive autokorrelasjon med det som skjedde for 12 måneder siden reflekterer at en sesong gjentar seg med 12 måneders mellomrom (dah!). Den negative autokorrelasjon med det som skjedde for 6 måneder siden reflekterer at en i løpet av et år svinger mellom en høy og en lav kraftproduksjon.

(b) Dette er en ARMA(4,0)-modell. Hvis vi lar R_t være notasjonen for residualtidsrekken ved tid t kan denne skrives som

$$R_t = 0.0110R_{t-1} - 0.1227R_{t-2} - 0.2633R_{t-3} - 0.1452R_{t-4} + U_t$$

Hvor U_t er hvit støy med varians 0.128.

(c) Vi predikerer først residualtidsrekken

$$\begin{aligned}\hat{R}_{t+1} &= 0.0110R_t - 0.1227R_{t-1} - 0.2633R_{t-2} - 0.1452R_{t-3} \\ &= 0.0110 \cdot (-1.05) - 0.1227 \cdot (-0.45) - 0.2633 \cdot (0.34) - 0.1452 \cdot (0.46) = -0.1126\end{aligned}$$

Dersom vi lar trend og sesong ha henholdsvis notasjon T_t og S_t kan modellen for tidsrekken skrives

$$Y_t = T_t + S_t + R_t$$

Prediksjonen for neste måneds kraftproduksjon blir derfor

$$\hat{Y}_{t+1} = \hat{T}_{t+1} + \hat{S}_{t+1} + \hat{R}_{t+1} = 1.79 + 0.42 - 0.1126 = 2.0974$$