LØSNINGSFORSLAG MET4 H23

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

Oppgave 1

(a) Vi skal teste følgende nullhypotese om lik andel:

$$H_0 : p_m = p_k \quad \text{mot} \quad H_a : p_m \neq p_k$$

der p_m og og p_k er andelen med lang utdanning blant hhv. menn og kvinner. Vi ser at 82 av totalt $n_m = 132 + 82 = 214$ menn har lang utdanning, mens 64 av $n_k = 222 + 64 = 286$ kvinner har lang utdanning. De estimerte andelene er derfor hhv. $\hat{p}_m = 82/214 = 0.383$ og $\hat{p}_k = 64/286 = 0.224$. Den samlede andelen med lang utdanning er $p = (82 + 64)/500 = 0.292$. Testobservatoren er da gitt ved:

$$Z = \frac{\hat{p}_m - \hat{p}_m}{\sqrt{(1/n_m + 1/n_k)p(1-p)}} = \frac{0.383 - 0.224}{\sqrt{(1/214 + 1/286)0.292(1-0.292)}} = 3.87$$

Kritisk verdi er gitt ved 0.975 percentilen i en standardnormalfordeling-fordeling. Her kan man bruke hukommelsen, tabell eller R-utregningen i vedlegg 1 for komme frem til 1.96.

Siden $Z = 3.87 > 1.96$, kan vi forkaste nullhypotesen om like andeler.

(b) Her er det 4 ulike grupper og vi nummerer dem med tall 1-4:

- Gruppe 1: Menn med kort utdanning
- Gruppe 2: Menn med lang utdanning
- Gruppe 3: Kvinner med kort utdanning
- Gruppe 4: Kvinner med lang utdanning

Indeksene under referer til gruppetilhørigheten over.

Vi skal utføre en *goodness-of-fit* test om observasjonene fra 2019 kan komme fra fordelingen som ble observert i 2009:

$$H_0 : p_1 = 0.315, \quad p_2 = 0.150 \quad p_3 = 0.439 \quad p_4 = 0.096$$

$$H_a : \text{minst to sannsynligheter er forskjellige fra dette}$$

Dersom fordelingen fra 2009 stemmer forventer vi følgende frekvenser blant de 500 individene: $e_1 = 500 \cdot 0.315 = 157.4$, $e_2 = 500 \cdot 0.15 = 75$, $e_3 = 500 \cdot 0.439 = 219.5$ og $e_4 = 500 \cdot 0.096 = 48$.

Testobservatoren er gitt ved

$$\chi^2 = \sum_{i=1}^{4} (f_i - e_i)^2/e_i = (132-157.4)^2/154 + (82-75)^2/75 + (222-219.5)^2/219.5 + (64-48)^2/48 \approx 10.1$$

Kritisk verdi ved 5% signifikansnivå er gitt ved 0.95 percentilen i en χ^2-fordeling med $4-1$ frihetsgrader. Her kan vi bruke en tabell eller følgende R-utregningen fra vedlegg 1: `qchisq(.95, df = 3) = 7.814`.

Siden $\chi^2 = 10.1 > 7.814$ kan vi forkaste nullhypotesen om at fordelingen fra 2009 også stemmer for 2019. Det har vært en endring i andelene siden da.

(c) Antall personer som svarer ja kan da skrives som $Y = X_1 + X_2 + \dots X_n = \sum_{i=1}^{n} X_i$. Dermed kan vi formulere den estimerte andelen som

$$\hat{p} = Y/n = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$$

og vi ser at $\hat{p}$ er gjennomsnittet av dummyvariablene og sentralgrenseteoremet forteller oss at gjennomsnitt er tilnærmet normalfordelt for store utvalg.

INFO: Dersom denne sanne andelen er p må fordelingen til dummyvariablene være gitt ved $P(X_i = 1) = p, P(X_i = 0) = 1 - p$, som vi kunne ha brukt til å vise at $\hat{p}$ er tilmærmet normalfordelt med forventning p og varians $p(1-p)/n$ som er grunnlaget for hypotesetestingen vi gjør for andeler.

Oppgave 2

(a) Vi skriver for enkelhets skyld

$$p = P(\text{liteUtbygging} \mid \text{fritidsbeboer} = x_1, \text{lokalbeboer} = x_2, \text{alder} = x_3, \text{utdanning} = x_4, \text{kvinne} = x_5)$$

.

Den estimerte modellen er da:

$$p = \frac{\exp(z)}{1 + \exp(z)}$$

hvor

$$z = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \widehat{\beta}_3 x_3 + \widehat{\beta}_4 x_4 + \widehat{\beta}_5 x_5 = -0.583 + 0.078 x_1 - 0.346 x_2 + 0.017 x_3 + 0.089 x_4 + 0.518 x_5.$$

(b) Husk: Generelt for logistisk regresjon kan vi si at en enhets økning i x er assosiert med at oddsen for utfallet vi ser på endres med en multiplikativ faktor e^{β} som svarer til en prosentendring i oddsen på $(e^{\beta} - 1)100\%$.

Konstantleddet i denne modellen har ingen reell tolkning da vi må anta at de andre forklaringsvariablene, inkludert alder, er 0.

Vi ser at fritidsbeboere i NR-område og personer fra Oslo og Viken (referanse gruppen) ikke har signifikant forskjellig holdning til lite utbygging. Oddsen for at lokalbeboere foretrekker lite utbygging er 29% lavere sammenlignet med disse to gruppene ($(e^{-0.346} - 1)100\% \approx -29\%$). Det kan virke som at de som bor i området er positive til mer utbyggingen.

Videre er ett års økning i alder er assosiert med at oddsen for å foretrekke lite utbygging øker med $(e^{0.017} - 1)100\% = 1.7\%$. Det å gå fra et utdanningsnivå til det neste er assosiert med at den samme oddsen økes med $(e^{0.089} - 1)100\% = 9.3\%$, men denne effekten er bare signfikant ved et 10% signfikansnivå. Merk at dette er en ordinal forklaringsvariabel som vi strengt tatt burde behandlet som en kategorisk variabel i regresjonsmodellen. Det virker som at eldre og personer med høy utdanning foretrekker lite utbygging. Til slutt ser vi at oddsen for at kvinner foretrekker lite utbygging er $(e^{0.518} - 1)100\% = 67.9\%$ større enn for menn. For disse tolkningene antar vi at de andre forklaringsvariablene er uendret.

(c) Med samme notasjon som i oppgave a) blir det lineære leddet:

$$z = -0.583 + 0.078 \cdot 1 - 0.346 \cdot 0 + 0.017 \cdot 37 + 0.089 \cdot 6 + 0.518 \cdot 1 = 1.176$$

Sannsynligheten for at denne respondenten foretrekker lite utbygging er derfor

$$p = \frac{\exp(1.176)}{1 + \exp(1.176)} \approx 0.76$$

(d) Med samme notasjon som i oppgave a) er nå x_3 ukjent. Det lineære leddet kan da skrives:

$$z = -0.583 + 0.078 \cdot 1 - 0.346 \cdot 0 + 0.017 \cdot x_3 + 0.089 \cdot 6 + 0.518 \cdot 1 = 0.547 + 0.017 \cdot x_3$$

Vi skal altså finne den verdien av x_3 som er slik at at

$$p = \frac{\exp(0.547 + 0.017 \cdot x_3)}{1 + \exp(0.547 + 0.017 \cdot x_3)} = 0.80$$

Vi kan enten løse denne ligningen med rå makt eller utnytte at inversen til den logistiske funksjon[1] (den såkalte logit funksjonen) er gitt ved $\text{logit}(p) = \log(p/1-p)$ som da betyr at

$$0.547 + 0.017 \cdot x_3 = \log \frac{0.80}{1 - 0.80}$$

$$x_3 = (\log \frac{0.80}{1 - 0.80} - 0.547)/0.017 = 49.37$$

Kvinnen må derfor være i overkant av 49 år for at modellen skal tildele henne en sannsynlighet på 0.8 for å foretrekke lite utbygging.

Oppgave 3

(a) Sesongkomponenten viser tydelig den årlige svingningen mellom kaldt om vinteren og varmt om sommeren. Ser vi på enheten på y-aksen til trendfiguren ser vi at er det omtrent ikke er noe trend.

(b) Her har det blitt tilpasset en AR(1)-modell:

$$R_t = 0.8147 R_{t-1} + u_t$$

hvor u_t er hvit støy med forventning 0 og varians 2.357. Koeffisienten i denne AR(1)-prosessen er $\phi = 0.8147$. En AR(1)-prossess er stasjonær dersom $|\phi| < 1$, altså er dette en stasjonær tidsrekke.

(c) Prediksjonen av R_{t+1} er i følge modellen:

$$\hat{R}_{t+1} = 0.8147 R_t + \hat{u}_{t+1} = 0.8147 R_t = 0.8147 \cdot 0.25 \approx 0.20$$

siden vår beste prediksjon av den hvite støyen er 0. Prediksjonen av temperaturen blir derfor:

$$\hat{\text{temperatur}}_{t+1} = \hat{T}_{t+1} + \hat{S}_{t+1} + \hat{R}_{t+1} = 8.40 - 4.80 + 0.20 = 3.8$$

(d) Fra forrige oppgave har vi at $\hat{R}_{t+1} = 0.8147 R_t$. Skulle vi predikert to dager frem kan vi tenke likt:

$$\hat{R}_{t+2} = 0.8147 R_{t+1}$$

Men siden vi ikke ennå har informasjon om R_{t+1} ved tid t er det beste vi kan gjøre å plugge inn prediksjonen $\hat{R}_{t+1} = 0.8147 R_t$ slik at

$$\hat{R}_{t+2} = 0.8147 \hat{R}_{t+1} = 0.8147^2 R_t \approx 0.17$$

Og slik kan vi fortsette:

$$\hat{R}_{t+3} = 0.8147 \hat{R}_{t+2} = 0.8147^3 R_t \approx 0.14$$

[1] dersom $p = e^x/(1+e^x)$ så er $x = \log(p/(1-p))$.

$$\hat{R}_{t+4} = 0.8147\hat{R}_{t+3} = 0.8147^4 R_t \approx 0.11$$

$$\vdots$$

$$\hat{R}_{t+365} = 0.8147\hat{R}_{t+364} = 0.8147^{365} R_t = 0$$

Prediksjonen av temperaturen blir derfor

$$\hat{\text{temperatur}}_{t+1} = \hat{T}_{t+365} + \hat{S}_{t+365} + \hat{R}_{t+365} = 8.40 - 4.80 + 0 = 3.6$$

Vi ser at prediksjonen av R_t raskt nærmer seg det stasjonære gjennomsnittet som vi ut fra R-utskriften kan se er 0. Dette reflekterer at når vi skal predikere virkelig langt frem i tid så er det sesong- og trend-komponentene som dominerer, mens eventuelle ARIMA komponenter ikke har noen prediktiv styrke.