Løsningsforslag til eksamen i

STK2120 13. juni 2016

Oppgave 1

a)

N_i er binomisk fordelt og $E(N_i) = np_i$, der $n = 204$.
Hvis H_0 er sann, er forventningen lik $E_i = n \cdot \frac{1}{6} = 204/6 = 34$ for $i = 1, 2, \ldots, 6$.

Hvis H_0 er sann er $\chi^2 = \sum_{i=1}^{6} \frac{(N_i - E_i)^2}{E_i}$ kji-kvadrat fordelt med $6 - 1 = 5$ frihetsgrader.

For marmorterningen ble $\chi^2 = 25.17$. Fra tabellen over kji-kvadratfordelingen har vi at $\chi^2_{0.005,5} = 16.75$. Siden $\chi^2 > \chi^2_{0.005,5}$ forkaster vi nullhypotesen. Vi kan konkludere at for marmorterningen har ikke de seks mulige utfallene samme sannsynlighet.

b)

For jernterningen ble $\chi^2 = 8.35$. Fra tabellen over kji-kvadratfordelingen har vi at $\chi^2_{0.90,5} = 1.61$ og $\chi^2_{0.10,5} = 9.24$.

Siden $\chi^2_{0.90,5} < \chi^2 < \chi^2_{0.10,5}$ er P-verdien mellom 10% og 90%. Videre er χ^2 ikke så mye mindre enn $\chi^2_{0.10,5}$, så P-verdien er mye nærmere 10% enn 90%.

For jernterningen kan vi ikke forkaste nullhypotesen om at alle de seks utfallene har samme sannsynlighet.

c)

Et alternativ til kji-kvadrat testen, er likelihood ratio testen.

Likelihooden er gitt ved

$$L(p_1, p_2, \ldots, p_6) = \frac{n!}{N_1!\, N_2! \cdots N_6!}\, p_1^{N_1}\, p_2^{N_1}\, \cdots p_6^{N_6},$$

og maksimum likelihood estimatoren for p_i er $\widehat{p}_i = N_i/n$ for $i = 1, 2, \ldots, 6$.

Likelihood ratio observatoren blir dermed

$$\mathrm{LR} = \frac{L(\frac{1}{6}, \frac{1}{6}, \ldots, \frac{1}{6})}{L(\widehat{p}_1, \widehat{p}_2, \ldots, \widehat{p}_6)} = \prod_{i=1}^{6} \left(\frac{\frac{1}{6}}{\widehat{p}_i}\right)^{N_i} = \prod_{i=1}^{6} \left(\frac{E_i}{N_i}\right)^{N_i},$$

og vi har at

$$-2\log(\mathrm{LR}) = 2\sum_{i=1}^{6} N_i \log\left(\frac{N_i}{E_i}\right).$$

Fra generell teori, vet vi at $-2\log(\mathrm{LR})$ er tilnærmet kji-kvadratfordelt med $6 - 1 = 5$ frihetsgrader under H_0

For marmorterningen får vi $-2\log(\mathrm{LR}) = 23.72$, mens vi for jernterningen har at $-2\log(\mathrm{LR}) = 8.99$. (Studentene er ikke bedt om å regne ut disse.)

Oppgave 2

a)

La y_i stå for i-te måling av `logNO2` og la x_{i1}, x_{i2} og x_{i3} være de tilhørende målingene av henholdsvis `logbil`, `temp` og `vind` for $i = 1, 2, \ldots, n$ der $n = 500$. Vi antar at y_i-ene er observerte verdier av stokastiske variabler $Y_1, Y_2, \ldots, Y_n$. Analysen bygger på den lineære regresjonsmodellen

$$Y_i = \beta_0 \ + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i, \tag{L.1}$$

der ϵ_i-ene er uavhengige og $N(0, \sigma^2)$-fordelte. Modellen (L.1) forutsetter:

(i) Lineær effekt av forklaringsvariablene x_{i1}, x_{i2} og x_{i3}.

(ii) Samme varians for alle ϵ_i-ene, og dermed også samme varians for alle Y_i-ene.

(iii) At ϵ_i-ene, og dermed også Y_i-ene, er normalfordelte.

(iv) At ϵ_i-ene, og dermed også Y_i-ene, er uavhengige.

Den øverste raden i matriseplottet viser sammenhengen mellom y_i-ene og hver av forklaringsvariablene. Selv om den simultane effekten av forklaringsvariblene kan avvike fra de marginale effektene som er vist i matriseplottet, kan det virke som om effekten av x_{i1} (`logbil`) og x_{i3} (`vind`) ikke er helt lineær.

Utfra matriseplottet er det vanskelig å si noe om de tre siste antagelsene. Men en kan tenke seg at det er avhengighet mellom to målinger av konsentrasjonen av NO_2 som er tatt nær hverandre i tid, slik at det kan være problemer med (iv). Men vi kan ikke sjekke det ut fra de opplysningene som er gitt i oppgaven.

b)

Forklaringen på utskriften er som følger ($j = 0, 1, 2, 3$):

- `Estimate` er minste kvadraters estimater $\widehat{\beta}_j$.
- `Std. Error` er de estimete standardfeilene $s_{\widehat{\beta}_j}$.
- `t value` er t-observatorene $t_j = \widehat{\beta}_j / s_{\widehat{\beta}_j}$ for testing av $H_0 : \beta_j = 0$.
- `Residual standard error` er estimatet for σ gitt ved $s = \sqrt{\text{SSE}/[n - (k+1)]}$. Her er $\text{SSE} = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$ residual kvadratsummen ($\widehat{y}_i = \widehat{\beta}_0 \ + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + \widehat{\beta}_3 x_{i3}$), $n = 500$ er antall observasjoner og $k = 3$ er antall forklaringsvariabler i den lineære regresjonsmodellen.
- `Multiple R-squared` er R^2 og angir hvor stor del av variasjonen i y_i-ene som er "forklart" av regresjonen. Her er $R^2 = 1 - \text{SSE}/\text{SST}$ der $\text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2$ er den totale kvadratsummen.

De tallene som skal stå der det er spørsmålstegn i utskriften er følgende:

- `t value` for `temp` er gitt ved $t_2 = \widehat{\beta}_2 / s_{\widehat{\beta}_2} = -0.038155/0.005673 = -6.7257$.
- `Std. Error` for `vind` er gitt ved $s_{\widehat{\beta}_2} = \widehat{\beta}_2 / t_2 = -0.211491/-10.359 = 0.0204$.
- `Degrees of freedom` for `residual standard error` er gitt ved $\text{df} = n - (k+1) = 500 - (3+1) = 496$.

c)

For den lineære regresjonsmodellen (L.1) vil den forventede responsen øke med β_j når den j-te forklaringsvariabelen øker med én og de to andre forklaringsvariblene holdes konstant. Det betyr at:

- Hvis biltrafikken dobles, slik at `logbil` øker med én, vil vi forvente at `logNO2` øker med 0.409 når temperaturen og vindstyrken er den samme.
- Hvis temperaturen øker én grad, vil vi forvente at `logNO2` blir redusert med 0.038 når biltrafikken og vindstyrken er den samme.
- Hvis vindstyrken øker med 1 m/s, vil vi forvente at `logNO2` blir redusert med 0.211 når biltrafikken og temperaturen er den samme.

Effektene ovenfor gjelder 2-logaritmen til NO_2-konsentrasjonen.

Hvis vi vil se på endringen i NO_2-konsentrasjonen, merker vi oss at en økning på β_j av 2-logaritmen til NO_2-konsentrasjonen, svarer til at selve konsentrasjonen blir multiplisert med en faktor 2^{β_j}. Det svarer til en relativ endring på $100 \cdot (2^{\beta_j} - 1)\%$. Altså har vi at:

- Hvis biltrafikken dobles, slik at `logbil` øker med én, vil vi vente at NO_2-konsentrasjonen øker med $100 \cdot (2^{0.409} - 1) = 32.8\%$ når temperaturen og vindstyrken er den samme.
- Hvis temperaturen øker én grad, vil vi vente at NO_2-konsentrasjonen reduseres med $-100 \cdot (2^{-0.038} - 1) = 2.6\%$ når biltrafikken og vindstyrken er den samme.
- Hvis vindstyrken øker med 1 m/s, vil vi vente at NO_2-konsentrasjonen reduseres med $-100 \cdot (2^{-0.211} - 1) = 13.6\%$ når biltrafikken og temperaturen er den samme.

d)

Vi har tilpasset alle mulige regresjonsmodeller med $m = 1, 2, \ldots, 9$ forklaringsvariabler og for hver m valgt den av dem som har minst residualkvadratsum $\text{SSE}(m)$. På denne måten har vi fått en sekvens av modeller $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_9$ slik det er gitt i oppgaven.

Når vi skal velge den beste av modellene $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_9$, kan vi ikke bruke R^2. Grunnen til det er at siden $\text{SSE}(m)$ vil avta med m, så vil $R^2 = 1 - \text{SSE}(m)/\text{SST}$ øke med m.

For å velge den beste modellen, er en mulighet å velge den som har minst verdi av justert R^2. I vår situasjon vil det gi en modell med sju forklaringsvariabler.

En annen (og ofte bedre) mulighet er å velge den verdien av m der Mallows C_p er omtrent lik (eller mindre enn) $m + 1$. For luftforurensningsdataene gir det $m = 6$, som svarer til modellen som har med variablene `logbil`, `temp`, `vind`, `I(logbil^2)`, `I(vind^2)` og `logbil:vind`.

Oppgave 3

a)

Likelihooden blir:

$$\begin{aligned} L(\beta_0,\beta_1) &= \prod_{i=1}^{n}\left(\frac{\lambda(x_i)^{y_i}}{y_i!}e^{-\lambda(x_i)}\right) \\ &= \prod_{i=1}^{n}\left(\frac{(e^{\beta_0+\beta_1 x_i})^{y_i}}{y_i!}\exp\{-e^{\beta_0+\beta_1 x_i}\}\right) \\ &= \left(\prod_{i=1}^{n} y_i!\right)^{-1}\exp\left(\beta_0\sum_{i=1}^{n} y_i+\beta_1\sum_{i=1}^{n} x_i y_i\right)\exp\left\{-\sum_{i=1}^{n} e^{\beta_0+\beta_1 x_i}\right\} \\ &= \left(\prod_{i=1}^{n} y_i!\right)^{-1}\exp\left\{\beta_0\sum_{i=1}^{n} y_i+\beta_1\sum_{i=1}^{n} x_i y_i-\sum_{i=1}^{n} e^{\beta_0+\beta_1 x_i}\right\}. \end{aligned}$$

b)

Log-likelihood funksjonen kan skrives som:

$$l(\beta_0,\beta_1) = -\sum_{i=1}^{n}\log(y_i!)+\beta_0\sum_{i=1}^{n} y_i+\beta_1\sum_{i=1}^{n} x_i y_i-\sum_{i=1}^{n} e^{\beta_0+\beta_1 x_i}$$

Ved å derivere finner vi score-funksjonene:

$$\begin{aligned} s_0(\beta_0,\beta_1) &= \frac{\partial}{\partial\beta_0}l(\beta_0,\beta_1) = \sum_{i=1}^{n} y_i-\sum_{i=1}^{n} e^{\beta_0+\beta_1 x_i} \\ s_1(\beta_0,\beta_1) &= \frac{\partial}{\partial\beta_1}l(\beta_0,\beta_1) = \sum_{i=1}^{n} x_i y_i-\sum_{i=1}^{n} x_i e^{\beta_0+\beta_1 x_i} \end{aligned}$$

c)

Elementene i den observerte informasjonsmatrisen blir:

$$\begin{aligned} J_{00}(\beta_0,\beta_1) &= -\frac{\partial^2}{\partial\beta_0^2}l(\beta_0,\beta_1) = -\frac{\partial}{\partial\beta_0}s_0(\beta_0,\beta_1) = \sum_{i=1}^{n} e^{\beta_0+\beta_1 x_i} \\ J_{01}(\beta_0,\beta_1) &= J_{10}(\beta_0,\beta_1) = -\frac{\partial^2}{\partial\beta_1\partial\beta_0}l(\beta_0,\beta_1) = -\frac{\partial}{\partial\beta_1}s_0(\beta_0,\beta_1) = \sum_{i=1}^{n} x_i e^{\beta_0+\beta_1 x_i} \\ J_{11}(\beta_0,\beta_1) &= -\frac{\partial^2}{\partial\beta_1^2}l(\beta_0,\beta_1) = -\frac{\partial}{\partial\beta_1}s_1(\beta_0,\beta_1) = \sum_{i=1}^{n} x_i^2 e^{\beta_0+\beta_1 x_i} \end{aligned}$$

Vi ser at disse uttrykkene kan skrives som:

$$J_{jk}(\beta_0,\beta_1) = \sum_{i=1}^{n} x_i^{j+k}\, e^{\beta_0+\beta_1 x_i} \quad \text{for} \quad j,k\in\{0,1\}.$$

d)

For å bestemme maksimum likelihood estimatene ved Newton-Raphsons metode, velger vi startverider $\beta_0^{(0)}$ og $\beta_1^{(0)}$. For eksempel kan vi velge $\beta_0^{(0)} = \beta_1^{(0)} = 0$. Deretter beregner vi nye verdier $\beta_0^{(k+1)}$ og $\beta_1^{(k+1)}$ ved algoritmen

$$\begin{bmatrix} \beta_0^{(k+1)} \\ \beta_1^{(k+1)} \end{bmatrix} = \begin{bmatrix} \beta_0^{(k)} \\ \beta_1^{(k)} \end{bmatrix} + \mathbf{J}\left(\beta_0^{(k)}, \beta_1^{(k)}\right)^{-1} \begin{bmatrix} s_0(\beta_0^{(k)}, \beta_1^{(k)}) \\ s_1(\beta_0^{(k)}, \beta_1^{(k)}) \end{bmatrix}$$

for $k = 0, 1, 2, \ldots$. Vi stopper iterasjonen når både $|\,\beta_0^{(k+1)} - \beta_0^{(k)}\,| < \delta$ og $|\,\beta_1^{(k+1)} - \beta_1^{(k)}\,| < \delta$ for en gitt nøyaktighet δ (for eksempel $\delta = 10^{-6}$). Maksimum likelihood estimatene $\widehat{\beta}_0$ og $\widehat{\beta}_1$ er de verdiene vi har for $\beta_0^{(k+1)}$ og $\beta_1^{(k+1)}$ når vi stopper iterasjonen.

e)

Av generell teori har vi at maximum likelihood estimatorene $\widehat{\beta}_j$ er tilnærmet $N(\beta_j, \sigma^2_{\widehat{\beta}_j})$-fordelte. Vi har også at vi kan estimere $\sigma^2_{\widehat{\beta}_j}$ med $s^2_{\widehat{\beta}_j} = J^{jj}(\widehat{\beta}_0, \widehat{\beta}_1)$, der $J^{jj}(\widehat{\beta}_0, \widehat{\beta}_1)$ er element (j, j) i den inverse informasjonsmatrisen $\mathbf{J}(\widehat{\beta}_0, \widehat{\beta}_1)^{-1}$.

De estimerte standardfeilene til maksumum likelihood estimatene er derfor:

$$s_{\widehat{\beta}_0} = \sqrt{0.015883} = 0.1260$$

$$s_{\widehat{\beta}_1} = \sqrt{0.000500} = 0.0224$$

Av resultatet om fordelingen til maksimum likelihood estimatorene, følger det at

$$\frac{\widehat{\beta}_1 - \beta_1}{s_{\widehat{\beta}_1}}$$

er tilnærmet standardnormalfordelt. Fra dette følger det på vanlig måte at et tilnærmet 95% konfidensintervall for β_1 er gitt som

$$\widehat{\beta}_1 \pm 1.96 \cdot s_{\widehat{\beta}_1}$$

Innsatt tall får vi $0.0899 \pm 1.96 \cdot 0.0224$, dvs. fra 0.0460 til 0.1338.

f)

Av (3) har vi at

$$\lambda(x+1) = e^{\beta_0 + \beta_1(x+1)} = e^{\beta_1}\, e^{\beta_0 + \beta_1 x} = e^{\beta_1}\, \lambda(x)$$

Den relative økningen i forventet antall barn når alderen øker med ett år er derfor:

$$\frac{\lambda(x+1) - \lambda(x)}{\lambda(x)} = \frac{e^{\beta_1}\, \lambda(x) - \lambda(x)}{\lambda(x)} = e^{\beta_1} - 1$$

Fra punkt e) har vi at $[0.0460\,,\, 0.1338]$ er et tilnærmet 95% konfidensintervall for β_1. Siden funksjonen $f(x) = e^x - 1$ er strengt voksende, følger det at

$$[e^{0.0460} - 1\,,\, e^{0.1338} - 1] = [0.0471\,,\, 0.1432]$$

er et tilnærmet 95% konfidensintervall for $\theta = e^{\beta_1} - 1$.