

UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i STK2120 — Statistiske metoder og dataanalyse 2

Eksamensdag: Mandag 6. juni 2011.

Tid for eksamen: 09.00 – 13.00.

Oppgavesettet er på 5 sider.

Vedlegg: Tabeller for normal- og t-fordeling.

Tillatte hjelpemidler: Lærebok: Moore & McCabe "Introduction to the practice of

Kontroller at oppgavesettet er komplett før du begynner å besvare spørsmålene.

Oppgave 1

(a) Modell:

$$X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

der ε_{ijk} er uavhengige, normalfordelte med forventning 0 og varians σ^2 .
For å unngå overparametrisering, har vi begrensningene

$$\sum_j \alpha_i = \sum_j \beta_j = \sum_i \gamma_{ij} = \sum_j \gamma_{ij} = 0$$

der de to siste gjelder for alle j og alle i , henholdsvis.

Estimat på variansen er 5.99 i dette tilfellet, (MS for residuals).

(b) Tredje linje svarer til $H_0 : \gamma_{ij} = 0$, for alle i, j . Naturlig å først teste denne hypotesen. Her vil vi ikke forkaste hypotesen (med P-verdi lik 0.251).

Gitt at vi ikke fikk signifikante interaksjonsledd, er det fornuftig å gå videre og teste hovedeffekter.

Første linje svarer til $H_0 : \alpha_i = 0$ for alle i . Forkastes med P-verdi lik 0.0018

Andre linje svarer til $H_0 : \beta_j = 0$, for alle j . Forkastes med P-verdi tilnærmet lik 0.

En kan da forenkle modellen til

$$X_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$$

dvs ingen interaksjonsledd.

(Fortsettes på side 2.)

Oppgave 2

- (a) For å bruke χ^2 testen bør en ha en forventet verdi på minst 5 for hver kategori. De to siste kategoriene har mye mindre forventninger, men hvis vi slår de sammen med den 3. kategorien, vil alle kategorier tilfredsstillte dette.

Med 3 kategorier og en fri parameter har vi $3-1-1=1$ frihetsgrad. Dette betyr at vi vil forkaste H_0 : Data følger Poisson fordelingen hvis $\chi^2 > \chi_{\alpha,1}^2 = 3.841$ hvis $\alpha = 0.05$. I dette tilfellet er testobservatoren langt større og vi får dermed forkastning. Siden også $\chi^2 > \chi_{0.001,1}^2 = 10.82$, så er P-verdien mindre enn 0.001.

Vi kan dermed konkludere med at vi kan forkaste hypotesen om at dataene følger en Poisson fordeling på signifikansnivå 0.05 (P-verdi < 0.001).

- (b) Da $\theta = \Pr(Y = 0t)$, må vi ha $0 \leq \theta \leq 1$.

For $\theta = e^{-\lambda}$ blir

$$\Pr(Y = y) = \frac{\lambda^y e^{-\lambda}}{y!}$$

for alle y , som svarer til Poisson fordelingen.

- (c) Vi har at

$$\begin{aligned} L(\boldsymbol{\theta}) &= \prod_{i=1}^n p(y_i) = \prod_{i=1}^n \prod_{y=0}^{\infty} p(y)^{I(y_i=y)} \\ &= \prod_{y=0}^{\infty} p(y)^{\sum_{i=1}^n I(y_i=y)} = \prod_{y=0}^{\infty} p(y)^{N_y}. \end{aligned}$$

Dermed blir

$$\begin{aligned} l(\boldsymbol{\theta}) &= \sum_{y=0}^{\infty} N_y \log(p(y)) \\ &= N_0 \log(p(0)) + \sum_{y=1}^{\infty} N_y \log(p(y)) \\ &= N_0 \log(\theta) + \\ &\quad \sum_{y=1}^{\infty} N_y [\log(1 - \theta) + \lambda - \log(e^\lambda - 1) + y \log(\lambda) - \lambda - \log(y!)] \\ &= \text{Konst} + N_0 \log(\theta) + (n - N_0) \log(1 - \theta) \\ &\quad - (n - N_0) \log(e^\lambda - 1) + \log(\lambda) \sum_{y=1}^{\infty} y N_y. \end{aligned}$$

(Fortsettes på side 3.)

- (d) Score-funksjonen er de deriverte av log-likelihood funksjonen mhp parametrene. Vi har at

$$\begin{aligned} s_1(\boldsymbol{\theta}) &= \frac{\partial}{\partial \theta} l(\boldsymbol{\theta}) \\ &= \frac{N_0}{\theta} - \frac{n - N_0}{1 - \theta} \end{aligned}$$

og

$$\begin{aligned} s_2(\boldsymbol{\theta}) &= \frac{\partial}{\partial \lambda} l(\boldsymbol{\theta}) \\ &= -(n - N_0) \frac{e^\lambda}{e^\lambda - 1} + \frac{1}{\lambda} \sum_{y=1}^{\infty} y N_y \end{aligned}$$

Videre er

$$E[s_1(\boldsymbol{\theta})] = \frac{n\theta}{\theta} - \frac{n - n\theta}{1 - \theta} = 0.$$

Tilsvarende er

$$\begin{aligned} E[s_2(\boldsymbol{\theta})] &= -(n - n\theta) \frac{e^\lambda}{e^\lambda - 1} + n(1 - \theta) \frac{1}{\lambda} \frac{e^\lambda}{e^\lambda - 1} \lambda \\ &= -n(1 - \theta) \left[\frac{e^\lambda}{e^\lambda - 1} - \frac{e^\lambda}{e^\lambda - 1} \right] = 0 \end{aligned}$$

Den generelle teorien sier også at forventningen skal være null.

Mellomregning, ikke krevd i oppgaven

Merk at siden s_1 ikke avhenger av λ , så er kryss-leddene i informasjonsmatrisene lik 0.

$$-\frac{\partial^2}{\partial \theta^2} l(\boldsymbol{\theta}) = \frac{N_0}{\theta^2} + \frac{n - N_0}{(1 - \theta)^2}$$

med forventning

$$E\left[-\frac{\partial^2}{\partial \theta^2} l(\boldsymbol{\theta})\right] = \frac{n\theta}{\theta^2} + \frac{n - n\theta}{(1 - \theta)^2} = \frac{n}{\theta} + \frac{n}{1 - \theta}$$

mens

$$-\frac{\partial^2}{\partial \lambda^2} l(\boldsymbol{\theta}) = -(n - N_0) \frac{e^\lambda}{(e^\lambda - 1)^2} + \frac{1}{\lambda^2} \sum_{y=1}^{\infty} y N_y$$

med forventning

$$\begin{aligned} E\left[-\frac{\partial^2}{\partial \lambda^2} l(\boldsymbol{\theta})\right] &= -(n - n\theta) \frac{e^\lambda}{(e^\lambda - 1)^2} + \frac{1}{\lambda^2} n(1 - \theta) \frac{e^\lambda}{e^\lambda - 1} \lambda \\ &= n(1 - \theta) \frac{e^\lambda}{e^\lambda - 1} \left[\frac{1}{\lambda} - \frac{1}{e^\lambda - 1} \right] \end{aligned}$$

(Fortsettes på side 4.)

- (e) Et problem med Newton-Raphson algoritmen er at \mathbf{J} ikke behøver å være positiv definit. Hvis den ikke er det, vil den kvadratiske tilnærmingen ikke ha et maksimumspunkt, og oppdateringen kan gi en dårligere gjett på maksimumet enn forrige verdi.

Scoring-algoritmen er en modifikasjon av N-R, ved at \mathbf{J} erstattes med dens forventning, \mathbf{I} . Siden \mathbf{I} er kovariansmatrisen til scoring funksjonen, vil den alltid være positivt definit, og dermed løse denne svakheten ved N-R.

At informasjonsmatrisene er diagonale, betyr at de to parametrene oppdateres uavhengige av hverandre noe som både forenkler og kan gi raskere konvergens (dette har vi ikke diskutert i kurset).

- (f) Vi har at

$$\begin{aligned}\theta^{(s+1)} &= \theta^{(s)} + \frac{\frac{N_0}{\theta^{(s)}} - \frac{n-N_0}{1-\theta^{(s)}}}{\frac{n}{\theta^{(s)}} - \frac{n}{1-\theta^{(s)}}} \\ &= \theta^{(s)} + \frac{N_0(1-\theta^{(s)}) - (n-N_0)\theta^{(s)}}{n(1-\theta^{(s)}) - n\theta^{(s)}} \\ &= \theta^{(s)} + \frac{N_0}{n} - \theta^{(s)} \\ &= \frac{N_0}{n}\end{aligned}$$

Denne verdien avhenger ikke av $\theta^{(s)}$ slik at vi vil få den samme verdien på hver eneste iterasjon.

- (g) Kan løse $s_1(\theta) = 0$ direkte. Gir

$$\begin{aligned}\frac{N_0}{\theta} - \frac{n-N_0}{1-\theta} &= 0 \\ \Downarrow \\ N_0(1-\theta) - (n-N_0)\theta &= 0 \\ \Downarrow \\ \theta &= \frac{N_0}{n}\end{aligned}$$

Kan dermed sette denne løsningen inn direkte og behøver kun å optimere likelihood funksjone mhp én parameter. Dette gir som regel mye raskere konvergens til max-punkt.

Dette er et rimelig estimat siden θ er sannsynligheten for $y = 0$ og N_0/n er andelen med $y = 0$.

Ved å sette inn dette estimatet direkte kunne en bruke dimensjonsreduksjon i optimering som da ville gi en enklere algoritme.

(Fortsettes på side 5.)

- (h) Siden informasjonsmatrisene er diagonale, er også de inverse diagonale. Dermed vil de asymptotiske kovariansmatrisene være diagonale og estimatene $\hat{\theta}$ og $\hat{\lambda}$ er tilnærmet uavhengige.

Et 95% konfidensintervall for θ er $[\hat{\theta} \pm z_{0.025}s_{\hat{\theta}}]$. Med $z_{0.025} = 1.960$ og $s_{\hat{\theta}} = \sqrt{1/I_{11}} = 0.0289$, får vi et intervall $[0.696, 0.810]$.

Et 95% konfidensintervall for λ er $[\hat{\lambda} \pm z_{0.025}s_{\hat{\lambda}}]$. Med $z_{0.025} = 1.960$ og $s_{\hat{\lambda}} = \sqrt{1/I_{22}} = 0.164$, får vi et intervall $[0.651, 1.294]$.

- (i) Parametrisk bootstrapping.

Bootstrap intervall for θ : $\delta_L = 0.695 - 0.753 = -0.058$, $\delta_U = 0.812 - 0.753 = 0.059$ som gir intervallet $[0.695, 0.812]$.

Bootstrap intervall for λ : $\delta_L = 0.656 - 0.972 = -0.316$, $\delta_U = 1.298 - 0.972 = 0.326$ som gir intervallet $[0.646, 1.288]$.

Svært like intervaller. Indikerer at normaltilnærming er rimelig.

- (j) Ønsker å teste $H_0 : \theta = \exp(-\lambda)$.

Kan lage et bootstrap konfidensintervall for $\theta - \exp(-\lambda)$. Hvis dette ikke dekker 0, betyr det at θ er signifikant forskjellig fra $\exp(-\lambda)$ og vi kan forkaste H_0 .