Norges teknisk–
naturvitenskapelige universitet
Institutt for matematiske fag

Faglig kontakt under eksamen:
Bo Lindqvist, tlf. 975 89 418

EKSAMEN I FAG TMA4315 GENERALISERTE LINEÆRE MODELLER

Torsdag 14. desember 2006

Tid: 09:00–13:00

Tillatte hjelpemidler:
Alle trykte og håndskrevne hjelpemidler. Godkjent enkel kalkulator.

Sensur: 11. januar 2007

Oppgave 1

Tabell 1 viser resultatene av en studie for sammenligning av to injeksjonsplaner for det nevroleptiske preparatet perphenazine decanoate (fra P. Knudsen, L. B. Hansen, K. Højholdt, N. E. Larsen, Acta Psychiatrica Scandinavica, 1985).

En gruppe på 19 psykotiske pasienter ble gitt injeksjoner hver annen uke, mens en annen gruppe på 19 pasienter ble gitt injeksjoner hver tredje uke. Pasientene ble fulgt i seks måneder, og effekten av behandlingen ble da evaluert. Kliniske evalueringer ble gjort med en seks-punkts skala kalt CGI (Clinical Global Impression), der høyere score betyr dårligere tilstand for pasienten.

De 12 radene i Tabell 1 svarer til 12 ulike kombinasjoner av de tre forklaringsvariablene

$$
\begin{aligned}
x_1 &= \begin{cases} 0 \text{ hvis injeksjoner gis hver annen uke} \\ 1 \text{ hvis injeksjoner gis hver tredje uke} \end{cases} \\
x_2 &= \begin{cases} 0 \text{ hvis kvinne} \\ 1 \text{ hvis mann} \end{cases} \\
x_3 &= \text{CGI ved start av behandling (initiell CGI).}
\end{aligned}
$$

Behandlings-intervall (uker)	Kjønn	Initiell CGI	Endelig CGI 0	1	2
2	Kvinne	2	1	0	0
2	Kvinne	3	3	1	0
2	Kvinne	4	0	1	0
2	Mann	3	4	4	1
2	Mann	4	0	2	1
2	Mann	5	0	0	1
3	Kvinne	2	1	0	0
3	Kvinne	3	2	1	0
3	Kvinne	4	1	2	0
3	Mann	2	3	1	0
3	Mann	3	0	5	0
3	Mann	4	0	3	0
"x_1"	"x_2"	x_3	y_0	y_1	y_2

Tabell 1: Data for nevroleptisk behandling.

De tilhørende responser er opptellinger for hver kombinasjon av forklaringsvariabler:

$$\begin{aligned} y_0 &= \text{antall som etter behandling har CGI} = 0 \\ y_1 &= \text{antall som etter behandling har CGI} = 1 \\ y_2 &= \text{antall som etter behandling har CGI} = 2 \end{aligned}$$

Merk at ingen pasienter hadde endelig CGI (i det følgende betegnet kun "CGI") over 2 etter behandlingen. Anta at CGI for en pasient med forklaringsvektor $\boldsymbol{x} = (x_1, x_2, x_3)$ har verdier $0, 1, 2$ med sannsynligheter

$$\pi_j = P(CGI = j|\boldsymbol{x}) \text{ for } j = 0, 1, 2.$$

Antallene $\boldsymbol{y} = (y_0, y_1, y_2)$ for en rad i tabellen antas å være resultatet av en multinomisk fordelt vektor $\boldsymbol{Y} = (Y_0, Y_1, Y_2)$ med sannsynlighetsvektor $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3)$ som altså avhenger av $\boldsymbol{x}$. Merk at x_3 i hele oppgaven skal betraktes som en numerisk kovariat, dvs. ikke som en faktor.

a) Sett opp og begrunn kort en proportional odds modell for de gitte dataene. Anta at det ikke er interaksjoner mellom de tre forklaringsvariablene x_1, x_2, x_3.

Uttrykk π_j for $j = 0, 1, 2$ som funksjoner av β-parametrene i modellen og forklaringsvektoren $\boldsymbol{x}$.

Modell	Devians
I	17.68
B+I	17.66
K+I	10.64
B*I	14.43
K*I	10.63
B+K+I	10.56
B*K+I	10.33
B*I+K	8.52
B+K*I	10.51
B*K+B*I	8.33
B*K+I*K	10.31
B*I+K*I	8.47
B*K+B*I+I*K	8.28
B*K*I	8.28

Tabell 2: Devianser for alle modeller som inneholder initiell CGI.

b) Kvotientene $P(CGI \leq j \mid \boldsymbol{x})/P(CGI > j \mid \boldsymbol{x})$ for $j = 0, 1$ kalles de kumulative odds ratioer for en pasient med forklaringsvektor $\boldsymbol{x}$.

Vis at dersom initiell CGI øker med 1 i modellen i (a), vil de kumulative odds ratioene multipliseres med e^{β_3}, der β_3 er koeffisienten til x_3 i de lineære prediktorene i modellen. Gi en praktisk fortolkning av verdien på e^{β_3}.

Gi tilsvarende fortolkninger av verdiene på e^{β_1} og e^{β_2}.

c) Beskriv den saturerte modell for dataene i denne oppgaven. Hvor mange frie parametre har den?

Hvordan ville du gå fram for å beregne deviansen for modellen i punkt (a)? Du skal her kun skissere en fremgangsmåte uten å gjøre alle beregninger.

Hvor mange frihetsgrader har deviansen i dette tilfellet? Begrunn svaret.

Tabell 2 viser utregnede devianser for alle proportional odds modeller som inneholder forklaringsvariabelen initiell CGI (dvs. x_3). Her er B=Behandlingsintervall, K=Kjønn, I=Initiell CGI. Når bokstaver knyttes med produkttegn *, betyr det at de tilsvarende forklaringsvariablene er med i modellen, sammen med alle interaksjoner (produkter) mellom to (eller tre) av dem. For eksempel betyr B+K*I modellen der x_1, x_2, x_3, x_2x_3 er med som forklaringsvariabler, mens B+K+I er modellen der x_1, x_2, x_3 er med i modellen uten interaksjoner (dvs. modellen i punkt (a)) og B*K*I er modellen der alle de tre variabler, alle produkter av to av dem, og produktet av alle tre, er med i modellen.

d) Beskriv modellen som svarer til B*K+B*I. Hvor mange parametre har den? Hva blir antall frihetsgrader for deviansen for denne modellen?

En statistiker har plukket ut modellene K+I, B+K+I, B*I+K og B*K+B*I som kandidater til "beste modell". Hvilken av disse modellene ville du velge som "den beste" basert på de gitte devianser? Begrunn ved å formulere relevante hypoteser og utføre de tilsvarende hypotesetester.

En (noe redigert) R-utskrift fra modellen B+K+I er gitt nedenfor. Du skal bruke den i besvarelsen av punkt (e) nedenfor.

```
Call:
vglm(formula = cbind(y0, y1, y2) ~ behand + kjonn +
    initCGI, family = cumulative(parallel = TRUE), data = neuro)

Coefficients:
                 Value Std. Error  t value
(Intercept):1  8.47538    3.34065  2.53705
(Intercept):2 12.87221    3.90906  3.29292
behand        -0.21992    0.75607 -0.29088
kjonn         -2.15758    0.88752 -2.43103
initCGI       -2.27249    0.69850 -3.25339

Number of linear predictors:  2

Dispersion Parameter for cumulative family:   1

Residual Deviance: 10.55521 on ?? degrees of freedom

Log-likelihood: -24.18958 on ?? degrees of freedom
```

Estimert kovariansmatrise for estimatorene, basert på utskrifter fra R:

```
11.1599   12.7043   -1.7892   -1.3418   -1.8574
12.7043   15.2808   -1.8714   -1.6839   -2.2749
-1.7892   -1.8714    0.5716    0.0921    0.0985
-1.3418   -1.6839    0.0921    0.7877    0.1982
-1.8574   -2.2749    0.0985    0.1982    0.4879
```

e) Anta at modellen fra punkt (a) gjelder.

Bruk R-utskriften til å estimere e^{β_k} for $k = 1, 2, 3$.

Finn et tilnærmet konfidensintervall for e^{β_1}. Gi en kommentar i lys av ditt valg av modell i punkt (d).

Estimer sannsynligheten for å få endelig CGI lik 0 for en kvinne som får injeksjoner hver annen uke og har initiell CGI lik 5.

Forklar hvordan du kan finne et estimert standardavvik for dette estimatet. Du trenger her ikke å gjøre alle beregningene.

Oppgave 2

La $Y_1, Y_2, \ldots, Y_N$ være uavhengige og eksponensialfordelte stokastiske variabler, der Y_i har tetthet

$$f(y_i; \theta_i) = \theta_i e^{-\theta_i y_i} \text{ for } y_i > 0,\ \theta_i > 0,\ i = 1, 2, \ldots, N.$$

a) Vis at $Y_i$ene har fordelinger som kommer fra den samme eksponensielle famile.

Bruk generelle formler fra pensum til å finne $\mu_i = \mathrm{E}(Y_i)$ og $\sigma_i^2 = \mathrm{var}(Y_i)$ uttrykt ved θ_i.

b) Vis at log-likelihood funksjonen for dataene $y_1, \ldots, y_n$ kan skrives

$$l = \sum_{i=1}^{N} \{-\theta_i y_i + \ln \theta_i\}$$

Bruk dette til å vise at deviansen for en generalisert lineær modell med estimerte forventninger $\hat{\mu}_i = \hat{y}_i$, er

$$D = 2 \sum_{i=1}^{N} \left\{ \frac{y_i - \hat{y}_i}{\hat{y}_i} - \ln\left(\frac{y_i}{\hat{y}_i}\right) \right\}$$

c) Med notasjon fra forrige punkt, sett opp uttrykk for Pearson-residualene, r_i, i den gitte situasjonen.

La $X^2 = \sum_{i=1}^{N} r_i^2$ være Pearsons kjikvadratobservator. Begrunn at $D \approx X^2$.

d) Anta at man vil teste nullhypotesen

$$H_0 :\ \theta_1 = \theta_2 = \cdots = \theta_N$$

mot den alternative hypotesen at alle θ_i er fritt varierende parametre, som ikke alle er like.

Bruk punktene (b) og (c) til å finne to testobservatorer for dette problemet.

Hvilke tilnærmede fordelinger har de under H_0?