UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i	STK3100 — Innføring i generaliserte lineære mo
Eksamensdag:	Torsdag 6. desember 2011.
Tid for eksamen:	14.30 – 18.30.
Oppgavesettet er på 0 sider.	
Vedlegg:	Tabell over normal, χ^2 og t fordeling
Tillatte hjelpemidler: STK1100/STK1110 og STK2120	Godkjent kalkulator og formelsamling for

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

Oppgave 1

(a) Vi har at

$$\begin{aligned}
M_Y(t) &= \int \exp(yt)c(y,\phi)\exp(\frac{y\theta - a(\theta)}{\phi})dy \\
&= \int c(y,\phi)\exp(\frac{y(\theta+t\phi) - a(\theta)}{\phi})dy \\
&= \int c(y,\phi)\exp(\frac{y(\theta+t\phi) - a(\theta+t\phi) + a(\theta+t\phi) - a(\theta)}{\phi})dy \\
&= \exp(\frac{a(\theta+t\phi)-a(\theta)}{\phi})(y,\phi)\exp(\frac{y(\theta+t\phi)-a(\theta+t\phi)+}{\phi})dy \\
&= \exp(\frac{a(\theta+t\phi)-a(\theta)}{\phi})
\end{aligned}$$

der vi har brukt at $f(y;\theta,\phi)$ integrerer seg til 1 for alle verdier av θ. Dermed blir

$$\begin{aligned}
M'_Y(t) &= \exp(\frac{a(\theta+t\phi)-a(\theta)}{\phi})a'(\theta+t\phi) \\
&= M_Y(t)a'(\theta+t\phi) \\
E[Y] &= M'_Y(0) = a'(\theta) \\
M''_Y(t) &= M_Y(t)a'(\theta+t\phi)^2 + M_Y(t)a''(\theta+t\phi)\phi \\
E[Y^2] &= M''_Y(0) = a''(\theta)^2 + \phi a''(\theta) \\
\mathrm{Var}[Y] &= \phi a''(\theta)
\end{aligned}$$

(Fortsettes på side 2.)

Alternativt kan en bruke at

$$\frac{\partial}{\partial\theta}\int_y f(y;\theta,\phi)dy = 0 \int_y \frac{1}{\phi}(y-a'(\theta))f(y;\theta,\phi)dy = 0$$

som gir $E[Y] = a'(\theta)$ og tilsvarende for varians. Dette krever at vi kan bytte om integrasjon og derivasjon, men vi har ikke diskutert de formelle kriterier for når dette er gyldig i kurset.

(b) Vi har at

$$\begin{aligned} f(y) =& \frac{1}{\sqrt{2\mu y^3\sigma}} \exp\left\{-\frac{y^2-2\mu y+\mu^2}{2y\mu^2\sigma^2}\right\} \\ =& \frac{1}{\sqrt{2\mu y^3\sigma}} \exp\left\{-\frac{y\frac{1}{2\mu^2}-\frac{1}{\mu}+\frac{1}{2y}}{\sigma^2}\right\} \\ =& \frac{1}{\sqrt{2\mu y^3\sigma}} \exp\{-\frac{1}{2y\sigma^2}\} \exp\left\{\frac{-y\frac{1}{2\mu^2}+\frac{1}{\mu}}{\sigma^2}\right\} \end{aligned}$$

som viser at

$$\begin{aligned} \theta =& -\frac{1}{2\mu^2} \\ a(\theta) =& \frac{1}{\mu} = -\sqrt{-2\theta} \\ \phi =& \sigma^2 \\ c(y;\phi) =& \frac{1}{\sqrt{2\mu y^3\phi}} \exp\{-\frac{1}{2y\phi}\} \end{aligned}$$

(c) Vi har at

$$\begin{aligned} E[Y] =& a'(\theta) = \frac{1}{\sqrt{-2\theta}} = \mu \\ \mathrm{Var}[Y] =& \phi a'(\theta) = \phi(-2\theta)^{-3/2} = \phi\mu^3 \end{aligned}$$

I tilfeller hvor variansstrukturen er tilnærmet kubisk som funksjon av forventning vil dette være nyttig. I tillegg er det en nyttig fordeling for responser som er positive.

Vi må ha at $\theta \leq 0$ som svarer til at $\mu \geq 0$. I tillegg må selvfølgelig $\phi \geq 0$.

(d) I dette tilfellet betyr det at hver observasjon har en forventning avhengig av forklaringsvariable gjennom

$$g(\mu_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}$$

der g er en såkalt *link*-funksjon.

(Fortsettes på side 3.)

Devians er formelt definert som $2*(\tilde{l}-l)$ der l er log-likelihood innsatt ML estimater og $\tilde{l}$ er log-likelihood for den *mettede* modell.

Devians kan brukes for sammenlikning av modeller, der forskjell i devians svarer til likelihood ratio (på log skala).

For kjent spredningsparameter kan devians brukes til en "goodness of fit" test for å se om en gitt modell er god nok (det siste bør brukes med varsomhet).

(e) Definer $\eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$. Vi har at $\beta \rightarrow \eta_i \leftrightarrow \mu_i \leftrightarrow \theta_i$. Sammenhengen mellom μ_i og θ_i er definert gjennom fordeling. Sammenheng mellom η_i og μ_i er definert gjennom link-funksjon. Hvis link-funksjonen velges slik at $\eta_i = \theta_i$, så forenkles mye av matematikken, og log-likelihood funksjonen blir penere (konkav). Det medfører også at observert informasjon blir lik forventet informasjon. I dette tilfellet svarer det til at

$$\begin{aligned} g^{-1}(\eta) &= \mu \\ \mu &= a'(\theta) \end{aligned}$$

som gir at vi må ha

$$g^{-1}(\theta) = a'(\theta) = \frac{1}{\sqrt{-2\theta}}$$

eller $g(\mu) = -1/(2\mu^2)$

Oppgave 2

(a) Modellen kalles *random intercept and slope model*

Slike modeller er nyttige for å bygge inn korrelasjoner mellom variable som kommer fra samme individ/gruppe og der korrelasjoner avhenger av noen kovariater. De er også nyttige når vi ønsker å gjøre prediksjon for grupper der vi ikke har observasjoner.

(b) Vi har at $\boldsymbol{Y}_i$ er (multivariat) normal fordelt. Videre er

$$\begin{aligned} E[\boldsymbol{Y}_i] &= \beta_0\mathbf{1} + \beta_1\boldsymbol{x}_i \\ \text{Var}[\boldsymbol{Y}_i] &= \boldsymbol{X}_i\boldsymbol{D}\boldsymbol{X}_i^T + \sigma^2\boldsymbol{I} \end{aligned}$$

der $\boldsymbol{X}_i = (\mathbf{1}, \boldsymbol{x}_i)$.

Dette gir oss eksplisitte uttrykk for likelihooden noe som gjør det rimelig enkelt å bruke en numerisk optimerer for å finne ML estimater.

(c) Et problem med ML estimering er at de gir forventningsskjeve estimater for varianser. Denne skjevhetet kommer av at variansestimater

(Fortsettes på side 4.)

benytter seg av ulike residualer som må beregnes basert på estimater av regresjonskoeffisienter. REML ideen er å (lineær) transformere data slik at de transformerte data har en fordeling som ikke avhenger av regresjonskoeffisientene. Så brukes ML estimering på de transformerte data. Det er uendelig mange slike transformasjoner. Vi ønsker imidlertid å utnytte data så mye som mulig. Derfor transformeres de kun ned til dimensjon $n-p$ hvis det er p β'er. Sålenge transformasjonen har rang $n-p$, er resultatet invariant mhp hvilken transformasjon vi velger.

Fordelen med REML er altså forventningskorrigering av variansestimater. REML brukes derfor når en vil sammenlikne modeller med ulike variansstrukturer/tilfeldige effekter. REML har imidlertid lavere effisiens i forhold til ML, og ML brukes heller når en konsentrerer seg om faste effekter.

(d) Boksplottet viser klart at det er store variasjoner fra jordstykke til jordstykke som indikerer at innkludering av $b_{0,i}$ er fornuftig. Det andre plottet gir en viss indikasjon på at variasjonen endrer seg med tid, noe som kan fanges opp med $b_{1,i}$ leddet.

(e) Et mulig kriterie for modell-valg er AIC. Her vil det være henholdsvis 6, 7 og 9 parametre i modellene, som gir AIC verdier

Modell	M0	M1	M2
Loglik	273.84	40.85	5.68

som viser at modell M2 gir den klart laveste AIC verdi og dermed er å foretrekke.

En kunne alternativt brukt LR test, men da må en passe på at en tester på parameterverdier som ligger på randen av parameterrrommet. En bør da bruke en blanding av χ^2 fordelinger for å beregne P-verdier.

(f) Da alle faste effekter er såpass signifikante, er det ingen grunn til å fjerne noen av disse. Merk imidlertid at det er en svært høy korrelasjon mellom $b_{0,i}$ og $b_{1,i}$. Nå viste vi i forrige deloppgave at det var bedre å ha med begge enn bare $b_{0,i}$. En kunne imidlertid undersøke om det er hensiktsmessig å bare ha med $b_{1,i}$.

(En tilpasning med bare $b_{1,i}$ ga dog mye dårligere AIC verdi)

SLUTT

(Fortsettes på side 5.)