UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i	STK3100 — Innføring i generaliserte lineære mo
Eksamensdag:	Mandag 5. desember 2011.
Tid for eksamen:	14.30 – 18.30.
Oppgavesettet er på 3 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) Modellen kalles *random intercept model*

Slike modeller er nyttige for å bygge inn korrelasjoner mellom variable som kommer fra samme individ/gruppe.

(b) Vi har at

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + b_i\mathbf{1} + \boldsymbol{\varepsilon}_i, \quad \boldsymbol{\varepsilon}_i \sim N(0, \sigma^2\boldsymbol{I})$$

som gir at

$$\mathbf{Y}_i \sim N(\mathbf{X}_i\boldsymbol{\beta}, \sigma_b^2\mathbf{1}\mathbf{1}^T + \sigma^2\boldsymbol{I})$$

Ved at vi har et eksplisitt uttrykk for den marginale fordeling for $\mathbf{Y}_i$, har vi også et eksplisitt uttrykk for likelihooden

$$L(\boldsymbol{\beta}, \sigma_b^2, \sigma^2) = \prod_{i=1}^{N} f(y_i; \boldsymbol{\beta}, \sigma_b^2, \sigma^2)$$

og denne kan så puttes inn i en numerisk optimerer for å finne ML-estimatene (evt REML estimater).

(c) Vi har

Parameter	β_1	β_1	β_2	σ_b	σ
Estimat	-4.028899	2.873710	-0.002898	0.040365	0.135060

Vi har at korrelasjonen mellom to variable fra samme fangst er

$$\frac{\sigma_b^2}{\sigma_b^2 + \sigma^2} \overset{estimert}{\approx} \frac{0.0016294}{0.0016294 + 0.0182412} = 0.082$$

(Fortsettes på side 2.)

(d) Vi har at $\hat{\boldsymbol{\beta}}$ er tilnærmet normalfordelt og dermed blir

$$\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 \log(66) + \hat{\beta}_2 \log(0.46)$$

også tilnærmet normalfordelt med varians

$$\begin{aligned}\text{Var}[\hat{\mu}] =& \text{Var}[\hat{\beta}_0] + \log(66)^2\text{Var}[\hat{\beta}_1] + \log(0.46)^2\text{Var}[\hat{\beta}_2] + \\ & 2\log(66)\text{Cov}[\hat{\beta}_0, \hat{\beta}_1] + 2\log(0.46)\text{Cov}[\hat{\beta}_0, \hat{\beta}_2] + \\ & 2\log(66)\log(0.46)\text{Cov}[\hat{\beta}_1, \hat{\beta}_2] \\ =& 0.2009^2 = 0.04037467\end{aligned}$$

Dermed er et 95% konfidensintervall for μ lik

$$\hat{\mu} \pm 1.96\text{SE}(\hat{\mu}) = [7.619373, 8.407036]$$

Siden $L < \mu < U \Leftrightarrow \exp(L) < \exp(\mu) < \exp(U)$ blir dermed et 95% konfidensintervall for θ lik $[2037.283, 4478.465]$

(e) Strategi

(a) Start med modell med alle forklaringsvariable og så mange interaksjoner som mulig

(b) Finn optimal struktur på tilfeldige effekter.
Her bør REML brukes!

(c) Finn optimal struktur for faste effekter.
Her bør ML brukes!

(d) Presenter endelig modell med REML estimering.

Oppgave 2

(a) Vi har at den eksponensielle klasse er gitt ved

$$f(y; \theta, \phi) = c(y; \phi) \exp\left(\frac{y\theta - a(\theta)}{\phi}\right)$$

For den binomiske fordeling har vi

$$\begin{aligned}f(y; \pi) =& \binom{n}{y} \pi^y (1-\pi)^{n-y} \\ =& \binom{n}{y} \exp(y\log(\pi) + (n-y)\log(1-\pi)) \\ =& \binom{n}{y} \exp(y\log(\pi/(1-\pi)) + n\log(1-\pi)) \\ =& c(y; \phi)\exp(y\theta - a(\theta))\end{aligned}$$

(Fortsettes på side 3.)

med $\theta = \log(\pi/(1-\pi), \phi = 1$ og $a(\theta) = -n\log(1-\pi) = -n\log(1+\exp(\theta)), c(y;\pi) = \binom{n}{y}$.

Kanonisk link: $g(\pi) = \log(\pi/(1-\pi))$

Vi modellerer π gjennom $g(\pi) = \eta = \mathbf{x}^T\boldsymbol{\beta}$. Med kanonisk link mener vi at $\theta = \eta$, som forenkler matematikken involvert (score-likningene og informasjonsmatrisene blir enklere)

(b) AIC $= -2\log L(\hat{\theta}) + 2q$ der L er likelihood verdi innsatt estimat på de ukjente parametre θ og q er antall parametre i modellen. Dvs vi har et straffeledd for kompleksitet av modellen.

En velger den modell som har minst AIC verdi. Her blir det modellen med probit link-funksjon.

Her vil vi sammenlikne modeller som ikke er nøstet i hverandre. Dermed vil LR og Wald-type tester ikke være egnede. AIC kan imidlertid brukes mer generelt.

Her vil vi i praksis velge den modell som gir høyest likelihoodverdi da kompleksiteten (antall parametre) er de samme for alle valg av link-funksjoner.

(c) Likelihoodfunksjonen er i dette tilfellet vanskelig å beregne pga den latente tilfeldige effekten, som medfører at likelihooden er gitt som et integral. Ved å skrive integranden som $e^{g_i(\mathbf{b}_i)}$ og deretter gjøre en (2. ordens) Taylor tilnærming av g_i, får en Laplace approksimasjonen. Denne vil være betraktelig enklere å beregne og kan da optimeres for å finne estimater.

(d) Vi har at $-2LR = -2(-240.8 + 240.6) = 0.4$. Vi har her én parameter mer i modellen med tilfeldige effekter. Vi må imidlertid ta hensyn til at vi her tester en H_0 som ligger på randen av parameter-rommet. Dermed må vi bruke en mikstur av χ^2_0 og χ^2_1 for beregning av P-verdi. Dette svarer til å beregne P-verdien under χ^2_1 fordelingen og så dele på 2. Nå er $P(\chi^2_1 > 0.4) = 0.527$ som dermed gir en P-verdi på 0.2635. Dette tilsier at vi ikke har grunnlag for å forkaste H_0 i dette tilfellet.

(e) Da `log(haulsize)` har en liten z verdi (ikke signifikant) både for modellen(e) i oppgave 1 og i oppgave 2 så antyder dette at denne variabelen ikke er så viktig. Merk dog at vi kun bruker et delsett av det totale datasettet, og det kan være vi finner denne variabelen til å være signifikant hvis vi brukte det fulle datasett.

SLUTT