UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i	STK3100/4100 — Innføring i generaliserte lineære modeller.
Eksamensdag:	Mandag 5. desember 2011.
Tid for eksamen:	14.30 – 18.30.
Oppgavesettet er på 4 sider.	
Vedlegg:	Tabell over normal, χ^2 og t fordeling
Tillatte hjelpemidler:	Godkjent kalkulator og formelsamling for STK1100/STK1110 og STK2120.

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

De ulike delpunktene kan stort sett løses uavhengige av hverandre. Hvis du står fast på et punkt, gå derfor heller videre til neste punkt.

Oppgave 1

Vi skal i denne oppgaven se på følgende modell:

$$\begin{aligned} Y_{ij} &= \mathbf{x}_{ij}^T\boldsymbol{\beta} + b_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2) \\ b_i &\sim N(0, \sigma_b^2) \end{aligned}$$

der $i \in \{1, ..., N\}$ er en gruppe-indeks mens $j \in \{1, ..., n_i\}$ er en indeks for repeterte målinger innen gruppe. Vi antar her alle tilfeldige variable er uavhengige av hverandre.

(a) Hva kalles denne modellen? (Bruk gjerne det engelske navnet)

Diskuter nytten av slike modeller.

(b) Hva blir den *marginale* modellen for $\mathbf{Y}_i = (Y_{i1}, ..., Y_{in_i})$?

Hvilke fordeler har det at vi har et eksplisitt uttrykk for den marginale fordelingen til $\mathbf{Y}_i$ når det gjelder estimering?

Vi skal i den resterende delen av oppgaven se på et konkret datasett fra Havforskningsinstituttet i Bergen. Dette datasettet er en liten del av et større datasett som benyttes for å kartlegge bestander av fisk i Barentshavet. Vårt datasett vil begrense seg til observasjoner på torsk fra år 2000 innenfor et spesifikt område.
Følgende variable er tilgjengelige på et tilfeldig utvalg innenfor hver fangst (et utkast av tråd)

(Fortsettes på side 2.)

- `length` Lengde på fisk (i cm)
- `weight` Vekt på fisk (i gram)
- `age` Alder til fisk (i år)
- `haulsize` Størrelse på total fangst (i tonn)

Vi vil i det etterfølgende la i være en indeks for fangst (haul) mens j er indeks for en individuell fisk innen en fangst. Aldersvariabelen vil først bli brukt i neste oppgave.
Vi vil starte med å se på en modell

$$\log(\text{weight}_{ij}) = \beta_0 + \beta_1 \log(\text{length}_{ij}) + \beta_2 \log(\text{haulsize}_i) + b_i + \varepsilon_{ij}$$

der $b_i \sim N(0, \sigma_b^2), \varepsilon_{ij} \sim N(0, \sigma^2)$ og alle tilfeldige effekter er uavhengige av hverandre. Nedenfor er en utskrift fra en tilpasning av denne modellen:

```
Linear mixed model fit by REML
Formula: log(weight) ~ log(length) + log(haulsize) + (1 | haul)
   Data: d.4
    AIC    BIC logLik deviance REMLdev
 -985.3 -961.4  497.6    -1012  -995.3
Random effects:
 Groups   Name        Variance  Std.Dev.
 haul     (Intercept) 0.0016294 0.040365
 Residual             0.0182412 0.135060
Number of obs: 885, groups: haul, 11

Fixed effects:
               Estimate Std. Error t value
(Intercept)   -4.028899   0.236925  -17.00
log(length)    2.873710   0.036685   78.33
log(haulsize) -0.002898   0.028023   -0.10

Correlation of Fixed Effects:
             (Intercept) log(length)
log(length)     -0.654
log(haulsize)   -0.732     -0.034
```

(c) Skriv opp estimatene til alle parametrene som inngår i modellen.

Hva blir korrelasjonen mellom to vekt-variable fra samme fangst (haul)?

(d) Vi vil i denne deloppgaven være interessert i

$$\theta = \exp\{\beta_0 + \beta_1 \log(66) + \beta_2 \log(0.46)\}$$

(Fortsettes på side 3.)

Du får her oppgitt at

$$\text{SE}\left(\hat{\beta}_0 + \hat{\beta}_1 \log(66) + \hat{\beta}_2 \log(0.46)\right) = 0.2009$$

der SE her står for standardfeil.

Forklar hvordan denne er blitt beregnet utifra opplysninger fra utskriften over (du behøver ikke å gjøre de faktiske utregninger).

Bruk dette til å lage et 95% konfidensintervall for θ.

(e) Anta nå vi ønsker å sammenlikne ulike modeller både med hensyn på hvilke tilfeldige effekter som bør være med *og* hvilke faste effekter som bør inkluderes. Skriv opp en generell strategi for å utføre modell-valg for slike modeller.

Oppgave 2

(a) Vis at den binomiske fordelingen er innenfor den eksponensielle klasse.

Hva er kanonisk link for den binomiske fordelingen innenfor GLM modellene?

Hva menes med kanonisk link, og hvilke fordeler har det å bruke kanonisk link?

Vi vil igjen se på data om fisk, men nå være interessert i alder. Som tidligere vil vi la indeks i stå for fangst (haul) mens j er indeks for individuell fisk innen fangst.
I utgangspunktet er alder en kategorisk variabel som varierer fra 3 til 13 år i dette datasettet. For å få det til å passe i vårt rammeverk, vil vi forenkle problemstillingen noe ved at vi vil definere

$$A_{ij} = \begin{cases} 1 & \text{hvis age}_{ij} > 9 \\ 0 & \text{ellers} \end{cases}$$

og bruke denne som responsvariabel.
Vi vil så se på følgende modell:

$$\begin{aligned} A_{ij} \sim & \text{Binom}(1, \pi_{ij}) \\ g(\pi_{ij}) = & \beta_0 + \beta_1 \log(\text{length}_{ij}) + \log(\text{haulsize}_i) \end{aligned}$$

der $g(\cdot)$ er en passende linkfunksjon og der alle A_{ij} er uavhengige.

(b) Tabellen nedenfor viser AIC verdier for 3 ulike link-funksjoner.

Link-funksjon	AIC
log	488.6953
probit	487.7484
cloglog	489.6841

(Fortsettes på side 4.)

Forklar hva AIC er og argumenter for hvorfor det er fornuftig å bruke et slikt kriterium (kontra andre typer tester vi har diskutert i kurset) i akkurat denne situasjonen.

Basert på disse verdiene, hvilken link-funksjon vil du foretrekke?

En utvidelse av modellen ovenfor er

$$\begin{aligned} A_{ij}|c_i &\sim \text{Binom}(1, \pi_{ij}) \\ g(\pi_{ij}) &= \beta_0 + \beta_1 \log(\text{length}_{ij}) + \beta_2 \log(\text{haulsize}_i) + c_i \\ c_i &\sim N(0, \sigma_c^2) \end{aligned}$$

Utskriften nedenfor svarer til denne modellen (link-funksjonen som er brukt her er ikke spesifisert og er ikke nødvendig å vite for å løse de følgende oppgaver, men er den optimale i forhold til AIC tabellen ovenfor).

```
Generalized linear mixed model fit by the Laplace approximation
Formula: A ~ log(length) + log(haulsize) + (1 | haul)
   Data: d.4
   AIC   BIC logLik deviance
 489.3 508.5 -240.7    481.3
Random effects:
 Groups Name        Variance Std.Dev.
 haul   (Intercept) 0.013427 0.11587
Number of obs: 885, groups: haul, 11

Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -45.7253     3.7311 -12.255   <2e-16 ***
log(length)     9.6381     0.7910  12.184   <2e-16 ***
log(haulsize)   0.2781     0.1513   1.838   0.0661 .

Correlation of Fixed Effects:
            (Intercept) log(length)
log(length)    -0.965
log(haulsize)  -0.319      0.062
```

(c) Forklar hva som menes med at modellen er tilpasset med Laplace approksimasjon.

(d) Log-likelihood verdien for modellen uten tilfeldig effekt er -240.8. Bruk dette til å utføre en likelihood-ratio test på $H_0 : \sigma_c^2 = 0$. Beregn tilhørende P-verdi og konkluder.

(e) Havforskningsinstituttet mener at størrelsen på fangst er viktig for å modellere alder og lengde av fisk. Basert på utskriftene i både denne og foregående oppgave, hva er din mening om dette?

SLUTT