UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i	STK3100/4100 — Innføring i generaliserte lineære modeller
Eksamensdag:	Torsdag 6. desember 2012.
Tid for eksamen:	14.30 – 18.30.
Oppgavesettet er på 4 sider.	
Vedlegg:	Tabell over 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

En stokastisk variabel Y sies å ha fordeling i den eksponensielle fordelingsklasse dersom tettheten (eller punkt sannsynligheten) til Y kan skrives på formen

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

For videre utregninger får du oppgitt at hvis

$$M_Y(t) = \mathrm{E}[\exp(Yt)] = \int \exp(yt)f(y)dy$$

eksisterer for alle t i et omegn om 0, så er

$$\mathrm{E}[Y^r] = M_Y^{(r)}(0)$$

der $M_Y^{(r)}(\cdot)$ er den r-te deriverte av $M_Y(t)$ mhp t.

(a) Regn ut forventning og varians i den eksponensielle fordelingsklasse.

Vi vil i resten av denne oppgaven se på den inverse Gaussiske fordeling, gitt ved

$$f(y) = \frac{1}{\sqrt{2\pi y^3\sigma}}\exp\left\{-\frac{1}{2y}\left(\frac{y-\mu}{\mu\sigma}\right)^2\right\}, \quad y > 0$$

(Fortsettes på side 2.)

(b) Vis at denne fordelingen tilhører den eksponensielle familie og vis at $\theta = -1/(2\mu^2)$ og $a(\theta) = -\sqrt{-2\theta}$. Identifiser også ϕ og $c(y;\phi)$.

(c) Finn forventning og varians i den inverse Gaussiske fordeling. Bruk dette til å diskutere i hvilke situasjoner en slik fordeling kan være nyttig å bruke.

Hva slags begrensninger ligger det på parametrene som er involvert?

(d) Anta nå $Y_1, ..., Y_n$ er uavhengige variable fra en generalisert lineær modell (GLM) med invers Gaussisk fordeling som respons fordeling. Forklar hva dette betyr.

Forklar generelt hva devians betyr og diskuter hva devians kan brukes til i GLM-sammenheng.

(e) Forklar hva vi mener med kanonisk link og hvilke fordeler det har å bruke denne.

Hva blir den kanoniske link for den inverse Gaussiske fordeling?

Oppgave 2

Vi skal i denne oppgaven se på modeller med følgende struktur:

$$\begin{aligned} Y_{ij} =& \beta_0 + b_{0,i} + (\beta_1 + b_{1,i})x_{ij} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2) \\ \boldsymbol{b}_i =& (b_{0,i}, b_{1,i})^T \sim N(\mathbf{0}, \boldsymbol{D}) \end{aligned} \qquad (*)$$

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 at alle b- og ε-variable er uavhengige av hverandre.

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

Diskutér 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?

(c) Forklar hovedprinsippene ved REML estimering.

Diskutér fordeler og ulemper med maksimum likelihood (ML) estimering sammenliknet med REML estimering. Spesifiser spesielt i hvilke tilfeller en vil bruke de ulike metodene.

Davidian og Giltinan (1995) beskriver et datasett for å sammenlikne vekstmønstre for to typer av soyabønner. Datasettet består av 412 observasjoner fordelt på 48 jordstykker med 8-10 observasjoner innen hvert jordstykke. I tillegg til vekt (`weight`) og type soyabønne (`Variety`, to typer)

(Fortsettes på side 3.)

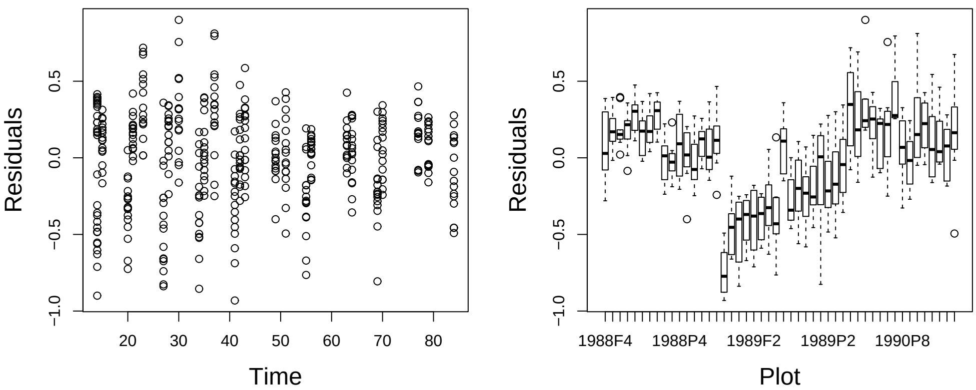
er også tidspunkt for innsamling av observasjon (dager etter planting, `Time`) angitt. Variabelen `Plot` angir jordstykke. I tillegg innfører vi variabelen `Time2` som er `Time` kvadrert.

(d) Vi vil først se på en enkel modell der vekt på log-skala er brukt som responsvariabel mens `Variety`, `Time` og `Time2` er inkludert som forklaringsvariable i tillegg til interaksjon mellom `Variety` og `Time`. Vi skriver denne modellen generelt som

$$\log(\text{Weight}_{ij}) = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \varepsilon_{ij} \tag{M0}$$

der i angir jordstykke mens j angir replikasjon innen jordstykke.

Figuren nedenfor viser boksplot av residualer gruppert etter jordstykker og plot av residualer mot `Time`. Kommentér plottene og argumenter hvorfor en modell tilsvarende $(*)$ kan være nyttig i dette tilfellet.



(e) Modellen i foregående deloppgave er så utvidet til to alternative modeller

$$\log(\text{Weight}_{ij}) = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + b_i + \varepsilon_{ij} \tag{M1}$$

$$\log(\text{Weight}_{ij}) = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + b_{0,i} + b_{1,i}\text{Time}_{ij} + \varepsilon_{ij} \tag{M2}$$

der $b_i \sim N(0, d^2)$ i modell M1 og $\boldsymbol{b}_i = (b_{0,i}, b_{1,i})^T \sim N(\boldsymbol{0}, \boldsymbol{D})$ i modell M2. Log-likelihood verdiene (innsatt REML estimater) for de tre modellene er gitt nedenfor:

Modell	M0	M1	M2
Loglik	-130.92	-13.42	6.16

Basert på dette, begrunn hvorfor modell M2 er å foretrekke.

(f) Nedenfor er resultatet av en tilpasning av modell M2 gjort med ML estimering. Diskutér om det er behov for å forenkle modellen.

```
Linear mixed-effects model fit by maximum likelihood

Random effects:
```

(Fortsettes på side 4.)

```
 Formula: ~1 + Time | Plot
            StdDev      Corr
(Intercept) 0.373119650 (Intr)
Time        0.002970683 -0.999
Residual    0.190066092

Fixed effects: log(weight) ~ Variety + Time + Time2 + Variety:Time
                  Value  Std.Error  DF   t-value p-value
(Intercept)   -5.202444 0.09214869 361 -56.45707  0.0000
VarietyP       0.478989 0.11677518  46   4.10181  0.0002
Time           0.204265 0.00236086 361  86.52125  0.0000
Time2         -0.001317 0.00002330 361 -56.52636  0.0000
VarietyP:Time -0.003079 0.00124599 361  -2.47102  0.0139
```

SLUTT