UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	ST 213 — Generaliserte lineære modeller.
Eksamensdag:	Fredag 12. desember 1997.
Tid for eksamen:	09.00 – 15.00.
Oppgavesettet er på 5 sider.	
Vedlegg:	Tabeller.
Tillatte hjelpemidler:	Formelsamling for ST 101 og ST 102, kalkulator.

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

Oppgave 1.

Eksponensielle klasser parametriseres ved tettheter (punktsannsynligheter) på formen

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

der $b(\theta)$ og $c(y,\phi)$ er passende funksjoner.

a) Vis at poissonfordelingene utgjør en eksponensiell klasse og identifiser parametrene θ og ϕ samt funksjonene $b(\theta)$ og $c(y,\phi)$.

b) Vis at klassen av normalfordelingen utgjør en eksponensiell klasse og identifiser parametrene θ og ϕ samt funksjonene $b(\theta)$ og $c(y,\phi)$.

c) Finn de momentgenererende og kumulantgenererende funksjoner for en stokastisk variabel Y med tetthet $f(y;\theta,\phi)$.

d) Vis på bakgrunn av c) at $EY = b'(\theta)$ og $\mathrm{Var}(Y) = \phi b''(\theta)$. Benytt dette til å vise følgende identiteter

(1) $E\frac{\partial}{\partial\theta}\ln f(Y;\theta,\phi) = 0$

(2) $E\left[\frac{\partial}{\partial\theta}\ln f(Y;\theta,\phi)\right]^2 = -E\frac{\partial^2}{\partial\theta^2}\ln f(Y;\theta,\phi)$.

(Fortsettes side 2.)

e) Vis at ligningene (1) og (2) holder generelt når Y har en tetthet $f(y;\theta)$ som ikke nødvendigvis tilhører en eksponensiell klasse.

f) La $Y_1, \ldots, Y_n$ være uavhengige med felles tetthet $f(y;\theta)$ og $\widehat{\theta}$ maksimum likelihood estimatoren (MLE) for θ.
Angi den tilnærmede fordeling, når n er stor, for $\widehat{\theta}$ og skisser utledningen.
Angi også, uten bevis, de asymptotiske resultatene for likelihood ratio testen (LRT) for $H_0 : \theta = \theta_0$ mot $H_1 : \theta \neq \theta_0$.

Resultatene i punktene e) og f) krever noen regularitetsbetingelser, men det er ikke meningen at du skal gjengi detaljer vedrørende disse.

Oppgave 2.

Du skal i denne oppgaven se på sammenheng mellom lav fødselsvekt og dødelighet i barnealder. Dataene er hentet fra Medisinsk fødselsregister i perioden 1967–1989 koblet mot Dødsårsaksregisteret i perioden 1968–1991. De omfatter, med noen få unntak, alle barn født i Norge i perioden, ca. 1.250.000. Du skal bare se på dødelighet fra fylte ett år og fram til 11-årsdagen. Det var ialt 4877 som døde blant barna i datamaterialet. Siden det er (mange) overlevende i materialet har vi såkalt sensurerte levetidsdata med sensureringstider

$$c_i = \begin{cases} \text{Alder 31/12–1991 hvis alder} < 11 \text{ år} \\ 11 \text{ hvis alder 31/12–1991} \geq 11 \text{ år} \end{cases}$$

Vi studerer altså levetiden T_i, men registrerer bare

$$X_i = \min(T_i, c_i) - 1 = \text{“observert levetid utover 1 år”}$$

og

$$D_i = \text{indikator for død} = \begin{cases} 1 & \text{dødsfall ved } X_i \\ 0 & \text{levde lenger enn } X_i \end{cases}$$

Formålet er altså å se hvordan dødeligheten påvirkes av kovariater. Vi definerer derfor

$$z_{i1} = \begin{cases} 1 & \text{Fødselsvekt} \leq 2500 \text{ g} \\ 0 & \text{Fødselsvekt} > 2500 \text{ g} \end{cases}$$

$$z_{i2} = \begin{cases} 1 & \text{Jente} \\ 0 & \text{Gutt} \end{cases}$$

$$z_{i3} = \begin{cases} 1 & \text{Født i 1975–1982} \\ 0 & \text{ellers} \end{cases}$$

$$z_{i4} = \begin{cases} 1 & \text{Født i 1983–1989} \\ 0 & \text{ellers} \end{cases}$$

(Fortsettes side 3.)

Kovariatene sammenfattes ved $z_i = (z_{i1}, z_{i2}, z_{i3}, z_{i4})$. Definer også for de ulike mulige verdiene z av z_i, akkumulert levetid

$$X_{\bullet z} = \sum_{i:\, z_i = z} X_i$$

og akkumulert antall døde

$$D_{\bullet z} = \sum_{i:\, z_i = z} D_i \, .$$

a) Vi skal anta at T_i er eksponensialfordelt med tetthet $\lambda_i e^{-\lambda_i t}$ der intensiteten eller raten λ_i er gitt ved

$$\lambda_i = \exp(\alpha + \beta_1 z_{i1} + \beta_2 z_{i2} + \beta_3 z_{i3} + \beta_4 z_{i4})$$

Forklar hva de ulike β_j og e^{β_j} representerer. Hvordan fortolker du e^{α}?

Du kan regne som kjent at (D_i, X_i) har fordelingsfunksjon (pkt.sanns./tetthet) $f_{(D_i,X_i)}(d, x) = \lambda_i^d e^{-\lambda_i x}$.

b) Forklar hvorfor modellen kan analyseres som om D_i er poissonfordelt med forventning $X_i \lambda_i$. Hvordan kan man da benytte programvare for generaliserte lineære modeller til å tilpasse modellen?

c) Argumenter for at man kan benytte tilsvarende modell på akkumulert antall døde $D_{\bullet z}$. Hvorfor blir forventingen i poissonfordelingen en benytter for $D_{\bullet z}$ lik $X_{\bullet z} \exp(\alpha + \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3 + \beta_4 z_4)$ der $z = (z_1, z_2, z_3, z_4)$?

d) Under er angitt regresjonstabellen over estimatene for α og β_j-ene i modellen i punkt a). Forklar hva de estimerte sammenhengene består i og angi signifikanser.

e) Finn et 95% konfidensintervall for e^{β_1} og sammenhold med signifikansvurdering i forrige punkt.

	Value	Std. Error	t value
(Intercept)	-7.585	0.024	-314.795
z1	0.813	0.063	12.970
z2	-0.412	0.032	-12.850
z3	-0.333	0.034	-9.649
z4	-0.079	0.056	-1.427

f) Vurder på grunnlag av ANOVA tabellen under, der Lvekt svarer til z_{1i}, Kj til z_{2i} og Faar til (z_{3i}, z_{4i}), om noen av kovariatene kan fjernes

(Fortsettes side 4.)

fra modellen og/eller om det bør inkluderes interaksjonsledd i modellen. Diskuter spesielt om interaksjon mellom kjønn og lav fødselsvekt bør inkluderes. Her bør du også benytte estimatene fra modellen med denne interaksjonen inkludert angitt under.

	Df	Deviance	Resid. Df	Resid. Dev
NULL			11	415.264
factor(Lvekt)	1	131.257	10	284.006
factor(Faar)	1	168.838	9	115.168
factor(Kj)	2	96.739	7	18.429
factor(Lvekt):factor(Kj)	1	4.674	6	13.755
factor(Lvekt):factor(Faar)	2	5.586	4	8.170
factor(Kj):factor(Faar)	2	3.130	2	5.040

	Value	Std. Error	t value
(Intercept)	-7.578	0.024	-312.358
z1	0.692	0.086	8.005
z2	-0.431	0.033	-12.944
z3	-0.332	0.034	-9.644
z4	-0.079	0.056	-1.426
z1*z2	0.272	0.126	2.168

g) Av to forskjellige grunner er det unødvendig å ta med interaksjonen mellom alle tre kovariater i ANOVA tabellen. Hvilke to grunner er dette?

Så langt er det antatt at levetidene T_i er eksponensialfordelt med konstant intensitet λ_i. Dette er urealistisk. Vi skal utvide modellen til at intensiteten/raten er "stykkevis konstant" dvs.

$$\lambda_i(t) = \exp(\alpha_j + \beta_1 z_{i1} + \cdots + \beta_p z_{ip})$$

når $t \in \langle t_{j-1}, t_j]$, $t_0 < t_1 < \cdots < t_q$. Her skal det benyttes intensiteter som er konstante på ett års intervaller slik at $t_0 = 1$, $t_1 = 2, \ldots, t_{10} = 11$, $q = 10$. Definer nå nye akkumulerte størrelser

$$X_{\bullet zj} = \text{total levetid med kovariat } z \text{ på alderstrinn } j.$$

og

$$D_{\bullet zj} = \text{totalt antall døde med kovariat } z \text{ på alderstrinn } j.$$

Modellen kan tilpasses som om alle $D_{\bullet zj}$ er uavhengige og poissonfordelte med forventning

$$X_{\bullet zj} = \exp(\alpha_j + \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3 + \beta_4 z_4)$$

Du skal ikke vise dette.

(Fortsettes side 5.)

h) Under er angitt regresjonstabellen fra modellen med stykkevis konstante intensiteter, med raten ved alder $\langle 1, 2]$ år som referanse. Beskriv sammenhengen mellom alder og dødelighet.

Beskriv også sammenhengen mellom dødelighet og de andre kovariatene og sammenlign med resultatene i punkt d). Gi en forklaring på at endringene blir helt marginale når det gjelder kjønn og lav fødselsvekt, men at det er en markant effekt på tidsintervall for fødsel.

	Value	Std. Error	t value
(Intercept)	-6.771	0.037	-182.054
factor(Ald)2	-0.345	0.049	-7.017
factor(Ald)3	-0.505	0.052	-9.670
factor(Ald)4	-0.799	0.058	-13.691
factor(Ald)5	-0.921	0.062	-14.893
factor(Ald)6	-1.042	0.066	-15.859
factor(Ald)7	-1.274	0.073	-17.436
factor(Ald)8	-1.318	0.076	-17.433
factor(Ald)9	-1.642	0.088	-18.622
factor(Ald)10	-1.750	0.094	-18.579
z1	0.811	0.063	12.969
z2	-0.412	0.032	-12.847
z3	-0.382	0.034	-11.080
z4	-0.538	0.057	-9.470

Oppgave 3.

a) Anta at Y er en lognormal stokastisk variabel, dvs. $V = \ln Y$ er normalfordelt med $EV = \gamma$ og $\operatorname{Var} V = \sigma^2$.
Vis at med $\mu = EY$ så kan man skrive $\operatorname{Var}(Y) = \mu^2\phi$. Hvordan avhenger ϕ av σ^2?
Hint: Du kan gjøre bruk av momentgenererende funksjon for normalfordelingen.

b) Anta at $Y_1, \ldots, Y_n$ er lognormale med samme spredningsledd ϕ, men $EY_i = \mu_i = e^{\alpha+\beta x_i}$. Hvordan vil du finne MLE for α, β og ϕ?

c) Anta at $Y_1, \ldots, Y_n$ er som i punkt b), bortsett fra at $\mu_i = \alpha + \beta x_i$.
Hvordan kan du benytte teorien for generaliserte lineære modeller til å estimere α, β og ϕ? Kan metoden modifiseres til å tilpasse modellen i punkt b)?

d) Begrunn at estimatorene du får med metoden i punkt c) er asymptotisk normalfordelte. Er estimatorene asymptotisk effisiente?

SLUTT