UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	STK3100/4100 — Innføring i generaliserte lineære modeller.
Eksamensdag:	Tirsdag 18. desember 2007.
Tid for eksamen:	14.30 – 17.30.
Oppgavesettet er på 3 sider.	
Vedlegg:	Ingen
Tillatte hjelpemidler:	Godkjent kalkulator og formelsamling for STK1100/STK1110 og for STK1120.

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

Oppgave 1.

En eksponensiell klasse kan parametriseres ved tetthet / punktsannsynligheter på formen

$$f(y;\theta) = \exp(\theta y - c(\theta) + d(y))$$

(a) Vis at Poissonfordelingene kan bringes på denne formen.

(b) Vis at eksponensialfordelingene også kan skrives på denne formen.

(c) Finn forventning og varians i eksponensialfordelingen ved hjelp av karakteriseringen i punkt (b). Beskriv spesielt hvordan variansen avhenger av forventningen.

(d) En Paretofordelt Y variabel har kumulativ fordelingsfunksjon $F(y;\theta) = 1 - (1/y)^\lambda$ når $y > 1$. Vis at $V = \log(Y)$ er eksponensialfordelt med forventning $\mu = 1/\lambda$.

(e) Anta at $Y_i, i = 1, \ldots, n$ er uavhengige og Paretofordelte med parameter $\lambda_i = \exp(\alpha + \beta x_i)$ for kjente kovariater x_i. Begrunn at man kan bruke et statistikkprogram med en vanlig implementasjon av generaliserte lineære modeller (GLM) som f.eks. `R` til å estimere parametrene α og β.

(Fortsettes side 2.)

(f) Man kan noe mer generelt definere eksponensielle klasser ved at tettheten kan skrives på formen

$$g(y;\theta) = \exp(\theta a(y) - c(\theta) + d(y))$$

Vis at med denne definisjonen tilhører Paretofordelingene den eksponensielle fordelingsklasse.

(g) Anta at Y har tetthet på formen $g(y;\theta) = \exp(\theta a(y) - c(\theta) + d(y))$. Utled et uttrykk for tettheten til $V = a(Y)$. Angi spesielt sammenhengen med parametriseringen $f(v;\theta)$ fra innledningen til denne oppgaven.

Oppgave 2.

I denne oppgaven skal vi se på risikoen for dødelighet i den såkalte "postneonatale perioden", fra 28. levedag til ett-årsdag. Vi skal bare se på dødelighet av SIDS (Sudden infant death syndrome, plutselig spebarnsdød) og vi modellerer sannsynligheten for SIDS-død i perioden, gitt at barna ikke dør av en annen årsak, ved logistisk regresjon.
Vi skal spesielt se på hvordan SIDS-dødeligheten avhenger av fødselsår, kjønn og fødselsvekt. Fødselsår (`kohort`) er angitt som en kategorisk variabel med 5 nivåer der nivå 1 angir 1967-1974, nivå 2 1975-1979, nivå 3 1980-1984, nivå 4 1985-1989 og nivå 5 1990-1995. Kjønn (`kjonn`) er kodet som 1 for gutter og 2 for jenter. Fødselsvekt (`vekt`) benyttes som kontinuerlig variabel angitt i kilo.

(a) Under er det gjengitt en R-utskrift av en devianstabell for dataene hvor endel størrelser er byttet ut med "?". Fyll ut verdiene som er byttet ut og fortolk resultatene. Når det gjelder p-verdier er det nok å angi om vi har statistisk signifikante effekter eller ikke.

Gi en begrunnelse for at signifikanstesting kan gjennomføres på bakgrunn av deviansanalysetabellen.

```
Analysis of Deviance Table
Model: binomial, link: logit

                         Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                                       570    1101.92
vekt                      ?        ?       569     842.33         ?
factor(kohort)            ?        ?         ?     527.74         ?
kjonn                     ?        ?         ?     434.93         ?
vekt:factor(kohort)       ?        ?         ?     428.56         ?
vekt:kjonn                ?        ?         ?     428.37         ?
factor(kohort):kjonn      ?        ?         ?     413.05    0.0041
vekt:factor(kohort):kjonn ?        ?         ?     407.80         ?
```

(Fortsettes side 3.)

(b) Under er det gjengitt en `R`-utskrift hvor det bare er tatt hensyn til hovedeffektene av kovariatene `vekt, kjonn` og `kohort`. For `kohort` er det benyttet en hjørnepunktparametrisering med nivå 1967-1974 som referanse. Fortolk resultatene ved hjelp av odds-ratio-estimater avledet fra tabellen.

Beregn også et 95% konfidensintervall for odds-ratioen svarende til fødselsvekt.

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -4.37607    0.15833 -27.639  < 2e-16 ***
vekt              -0.67110    0.03758 -17.859  < 2e-16 ***
kjonn             -0.47371    0.04981  -9.511  < 2e-16 ***
factor(kohort)2    0.56224    0.08629   6.515 7.25e-11 ***
factor(kohort)3    0.90941    0.08105  11.220  < 2e-16 ***
factor(kohort)4    1.07958    0.07743  13.943  < 2e-16 ***
factor(kohort)5    0.11049    0.08958   1.233    0.217
---
```

(c) Vi har så langt ignorert dødelighet av andre årsaker enn SIDS. Anta nå at det totalt er J ulike dødsårsaker og at vi har en multinomisk regresjonsmodell med $J+1$ utfall (inkludert de overlevende) med π_{ij} lik sannsynligheten for at individ i dør av årsak $j = 1, \ldots, J$ og π_{i0} lik sannsynlighet for at individ i overlever. Med kovariatvektorer x_i angis modellen ved

$$\pi_{ij} = \frac{\exp(\beta_j' x_i)}{1+\sum_{k=1}^{J} \exp(\beta_k' x_i)} \text{ for } j = 1, \ldots, J$$

$$\pi_{i0} = \frac{1}{1+\sum_{k=1}^{J} \exp(\beta_k' x_i)}$$

der β_j er vektorer av regresjonsparametre. Vi angir SIDS som årsak $j = 1$.

Vis at vi da har en vanlig logistisk regresjonsmodell for å dø av SIDS gitt at vi (som i punkt (a) og (b)) bare ser på de som dør av SIDS og de som overlever. Angi spesielt parametrene i denne logistiske regresjonsmodellen.

Diskuter ulemper og fordeler med å analysere dataene med logistisk regresjon framfor den fulle multinomiske regresjonsmodellen.

SLUTT