UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	STK2000 — Sentrale statistiske metoder og modeller.
Eksamensdag:	Fredag 10. desember 2004.
Tid for eksamen:	09.00 – 12.00.
Oppgavesettet er på 3 sider.	
Vedlegg:	Tabell over normalfordeling, tabell over χ^2-fordeling.
Tillatte hjelpemidler:	Formelsamling for ST100 og ST110, godkjent kalkulator.

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

Oppgave 1.

Betrakt en eksponensiell fordelingsklasse der tettheten har formen

$$f(y;\theta) = \exp[yb(\theta) + c(\theta) + d(y)] \tag{1}$$

der θ er en reell parameter, og b, c og d er kjente funksjoner.

a) Uttrykk $E(Y)$ ved $c(\theta)$ og $b(\theta)$. Her er Y en tilfeldig variabel med tetthet av formen (1). Forklar hvorfor $c(\theta) = -\log\{\int \exp[yb(\theta)+d(y)]dy\}$.

b) Forklar hvordan tettheten til en tilfeldig variabel som er normalfordelt med forventning μ og varians 1, kan skrives på formen (1). Hvordan ser funksjonene b, c og d ut i dette tilfellet?

Hva er de tilsvarende uttrykkene for punktsannsynligheten til en tilfeldig variabel som er Poisson-fordelt med forventning λ?

(Fortsettes side 2.)

c) Anta at $Y_1, \ldots, Y_N$ er uavhengige og identisk fordelte observasjoner med tetthet av formen (1). Vis at sannsynlighetsmaksimeringsestimatoren (SME) for θ vil tilfredsstille

$$\bar{Y} = -\frac{c'(\hat{\theta})}{b'(\hat{\theta})} \text{ der } \bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i.$$

Hva blir SME i situasjonene der variablene er normal- og Poisson-fordelte som i punkt b)?

d) Hva er den tilnærmede fordelingen til SME $\hat{\theta}$ når N er stor? Forklar i hovedtrekk hvordan denne tilnærmede fordelingen utledes når tettheten til observasjonene har formen (1).

Oppgave 2.

Nedenfor er det gjengitt en $2 \times 2 \times 2$ tabell basert på 123 pasienter med diabetes. De er klassifisert etter hvorvidt de var avhengige av injeksjon av insulin (1 = avhengig, 2 = ikke avhengig), etter om noen i familien også hadde diabetes (1= slektninger med diabetes, 2 = ingen slektninger med diabetes) og etter alder da de ble diagnostisert (1 = under 45 år, 2 = 45 år eller mer).

Vi betegner faktorene med "insulin", "diafam" og "alder".

		Diabetes i familien			
		Ja		Nei	
		Alder ved diagnose		Alder ved diagnose	
		< 45	≥ 45	< 45	≥ 45
Avhengig av injeksjon av insulin	Ja	6	6	16	8
	Nei	1	36	2	48

La observasjonene være y_{jkl} der j, k, l betegner henholdsvis nivå for insulin, nivå for diafam og nivå for alder.

a) Formuler en logistisk regresjonsmodell med

- respons: Avhengighet av injeksjon av insulin
- kovariater: Alder ved diagnose (alder) og om slektninger har diabetes (diafam)

Som linkfunksjon bruker du en logit link. Forklar hva det er. Angi også et uttrykk for den lineære prediktoren når du bruker en såkalt "cornerpoint" parametrisering med $k = l = 1$ som referansekategori.

b) Nedenfor er det angitt del av en deviansanalysetabell. Kolonnen med frihetsgrader for deviansen er fjernet. Fyll ut det som mangler. Bruk tabellen til å begrunne at en modell med "konstantledd + alder" er tilfredsstillende.

(Fortsettes side 3.)

```
           Terms Resid. Df Resid. Dev        Test  Df   Deviance
               1         3   50.03359
          alder          2    0.04667                ?   49.98693
 alder + diafam          1    0.03929       +diafam  ?    0.00738
 alder * diafam          0    0.00000 +alder:diafam  ?    0.03929
```

c) Estimatene for parametrene er

```
Coefficients:
               Value Std. Error   t value
(Intercept)  1.99243  0.6153449  3.237908
      alder -3.78419  0.6796930 -5.567498
```

Uttrykk den estimerte modellen ved passende oddsforhold, og forklar hvordan resultatet skal tolkes. Her er det brukt en såkalt "cornerpoint" parametrisering med referansekategori 1, slik at den angitte koeffisienten er estimatet for parameteren for nivået "$\geq$ 45 år".

d) Beregn også 95% konfidensintervall for de aktuelle oddsforholdene.

e) Forklar hva som menes med deviansresidualer, og hva de brukes til. Beregn deviansresidualet for modellen "konstantledd + alder" svarende til kovariatkombinasjonen "alder mindre enn 45 år" og "ingen slektninger med diabetes".

I resten av oppgaven vil de tre faktorene bli behandlet symmetrisk, og den vil dreie seg om log-lineære modeller.

f) Forklar hvordan man formulerer en log-lineær modell for å analysere $2 \times 2 \times 2$ tabellen, og angi sammenhengen mellom parametrene i den logistiske modellen "konstantledd+alder+diafam" og den tilsvarende log-lineære modellen. Forklar også hvordan likelihooden for den logistiske modellen "konstantledd + alder + diafam" kan maksimeres ved å maksimere en passende Poisson likelihood.

SLUTT