UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i	STK3100 — innføring i generaliserte lineære modeller.
Eksamensdag:	Tirsdag 15. desember 2009.
Tid for eksamen:	09.00 – 12.00.
Oppgavesettet er på 2 sider.	
Vedlegg:	Tabell over normalfordelingen.
Tillatte hjelpemidler:	Godkjent kalkulator og formelsamling for STK1100/STK1110 og for STK1120.

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

NB. Oppgavesettet består av to oppgaver.

Oppgave 1

(a) Vis at binomialfordelingen (n, π) tilhører den eksponensielle fordelingsklassen, der tettheten / punktsannsynligheten kan skrives på formen

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

Finn $a(\theta)$ og ϕ, og bruk dette til å vise de kjente formlene for forventning og varians i den binomiske fordelingen.

(b) Definer begrepet generalisert lineær modell (GLM). Hva menes med kanonisk link? Finn kanonisk link i en binær regresjonsmodell, dvs. en modell der responsen har en binær fordeling / Bernoullifordeling. Hvilke andre linkfunksjoner er vanlige for slike fordelinger?

(c) Definer begrepene mettet modell og devians. Se på en logistisk regresjonsmodell med regresjonsdel $\eta_i = \alpha + \beta x_i$, med binær responsvariabel y_i og med forklaringsvariabel x_i, $i = 1, 2, ..., n$. Hvordan vil du i denne modellen teste hypotesen $\mathrm{H}_0 : \beta = 0$ ved å gjøre bruk av devians-begrepet?

(d) Sett opp loglikelihoodfunksjonen for modellen i c). Vis at scorefunksjonen med hensyn til parametrene α og β har komponenter

$$\sum_{i=1}^{n}(y_i - \frac{e^{\eta_i}}{1+e^{\eta_i}}) \quad \text{og} \quad \sum_{i=1}^{n}(x_i y_i - \frac{x_i e^{\eta_i}}{1+e^{\eta_i}}).$$

Finn Fisherinformasjonsmatrisen. Fortell hvordan du ut fra disse størrelsene kan sette opp en iterasjonsprosedyre for maximum likelihood estimatoren av vektoren $(\alpha\ , \beta)^T$.

(Fortsettes på side 2.)

Oppgave 2

I en tysk spørreundersøkelse har man innhentet data om antall legebesøk siste 3 måneder (`numvisit`) og om ulike faktorer som kan ha betydning for legebesøk blant 1100 kvinner. Blant de innsamlede kovariatene skal vi bare se på en indikator for selvrapportert dårlig helse (`badh`) og alder (`age`) i år.

(a) Under er det gjengitt resultater fra en R-kjøring av en Poissonregresjon med log-link for antall legebesøk mot kovariatene `badh` og `age`. Spesifiser modellen matematisk og fortolk parameterne.

```
> M0<-glm(numvisit~age+badh,family=poisson)
> summary(M0)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.9452  -2.0348  -0.8169   0.5191  12.5571

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.588731   0.064318   9.153  < 2e-16 ***
age         0.005556   0.001676   3.316 0.000914 ***
badh        1.140908   0.039858  28.625  < 2e-16 ***
---
Signif. codes:  0 ë***í 0.001 ë**í 0.01 ë*í 0.05 ë.í 0.1 ë í 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 4779.4  on 1099  degrees of freedom
Residual deviance: 3975.3  on 1097  degrees of freedom
```

(b) Finn, basert på `R`-utskriften, estimater for rate-ratioer samt 95% konfidensintervall for legebesøk mellom

(i) kvinner med selvrapportert dårlig helse og selvrapportert god helse.

(ii) kvinner på 50 år og 40 år

Estimer også raten for legebesøk for en 40 årig kvinne med god helse. Hvilken ytterligere informasjon trenger du for å finne et konfidensintervall for (den teoretiske) raten?

(c) En mer generell modell tillater overspredning i forhold til Poissonmodellen i punkt a) via en spesifikasjon $\mathrm{Var}(Y) = \phi\mu$ der Y er responsen antall legebesøk og μ er forventningen til Y. En metode for å ta hensyn til overspredning består i å anta at $Y|Z$ er Poissonfordelt med forventning $Z\mu$ der Z er en latent variabel. Hvilken fordeling vil Y ha marginalt (ubetinget) hvis Z er gammafordelt (med forventning lik 1). Hvordan vil du begrunne svaret?

SLUTT