# UNIVERSITETET I OSLO

## Det matematisk-naturvitenskapelige fakultet

| | |
|---|---|
| Eksamen i: | ST 202 — Statistiske slutninger for den eksponentielle fordelingsklasse. |
| Eksamensdag: | Fredag 17. desember 1993. |
| Tid for eksamen: | 09.00 – 15.00. |
| Oppgavesettet er på 4 sider. | |
| Vedlegg: | Ingen. |
| Tillatte hjelpemidler: | Formelsamlinger for ST 101, ST 102, ST 103, lommeregner. |

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

## Oppgave 1.

Vi har responsen $Y_1, \ldots, Y_n$ fra en generalisert lineær modell (GLM), dvs. $Y_1, \ldots, Y_n$ er uavhengige og $Y_i$ har tetthet $f_{Y_i}(y; \theta_i, \phi) = \exp\left(\frac{\theta_i y - b(\theta_i)}{a(\phi)} + c(y, \phi)\right)$ og lineær prediktor $\eta_i = \boldsymbol{\beta}' \boldsymbol{x}_i = \sum_{j=1}^{p} \beta_j x_{ji} = g(\mu_i)$ for linkfunksjon $g(\cdot)$ og $\mu_i = EY_i$. La dessuten $\boldsymbol{Y}' = (Y_1, \ldots, Y_n)$ og $\boldsymbol{\mu}' = (\mu_1, \ldots, \mu_n)$.

a) Definer devians $D(\boldsymbol{Y}, \boldsymbol{\mu})$ og skalert devians $D^*(\boldsymbol{Y}, \boldsymbol{\mu})$. Angi hvordan og begrunn hvorfor $D^*(\boldsymbol{Y}, \boldsymbol{\mu})$ kan benyttes til å teste hypotesen (der $1 \leq q < p$)

$H_0 : \beta_{q+1} = \cdots = \beta_p = 0$ mot $H_1$ : Alle $\beta_j$ vilkårlige.

b) Finn eksplisitte uttrykk for $D(\boldsymbol{Y}, \boldsymbol{\mu})$ og $D^*(\boldsymbol{Y}, \boldsymbol{\mu})$ når

(i) $f_{Y_i}(y) = \frac{\mu_i^y}{y!} e^{-\mu_i}$, dvs. $Y_i$ poissonfordelt

(ii) $f_{Y_i}(y) = \left(\frac{\nu}{\mu_i}\right)^{\nu} \frac{y^{\nu-1}}{\Gamma(\nu)} e^{-\frac{\nu}{\mu_i} y}$, dvs. $Y_i$ gammafordelt.

*(Fortsettes side 2.)*

c) Vis at estimeringsligningene for sannsynlighetsmaksimeringsestimatoren $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ kan skrives

$$\sum_{i=1}^{n} \frac{Y_i - \hat{\mu}_i}{V(\hat{\mu}_i)} \, \frac{x_{ji}}{g'(\hat{\mu}_i)} = 0 \qquad j = 1, \dots, p$$

der $V(\mu_i) = \operatorname{Var} Y_i / a(\phi)$ og $\hat{\mu}_i = g^{-1}(\hat{\boldsymbol{\beta}}' x_i)$.

d) Vis at hvis

A. $g(\mu) = \ln(\mu)$ samt at $\eta_i = \beta_1 + \sum_{j=2}^{p} \beta_j x_{ji}$, dvs. $\eta_i$ inneholder et konstantledd $\beta_1$, så holder

$$\sum_{i=1}^{n} \frac{(Y_i - \hat{\mu}_i)\hat{\mu}_i}{V(\hat{\mu}_i)} = 0 \tag{1}$$

Vis at formelen (1) også holder hvis

B. $g(\mu) = \mu^{\alpha}$ for gitt $\alpha \neq 0$ også når det ikke er konstantledd i $\eta_i$.

(Hint: Benytt at $\hat{\eta}_i = \sum_{j=1}^{p} \hat{\beta}_j x_{ji} = g(\hat{\mu}_i)$.)

e) Anta at formelen (1) holder. La $\hat{\boldsymbol{\mu}}' = (\hat{\mu}_1, \dots, \hat{\mu}_n)$. Finn forenklede uttrykk for $D^*(\boldsymbol{Y}, \hat{\boldsymbol{\mu}})$ og $D(\boldsymbol{Y}, \hat{\boldsymbol{\mu}})$ i forhold til dem du fant i punkt b) for

(i) $Y_i$ poissonfordelt

(ii) $Y_i$ gammafordelt.

# Oppgave 2.

Anta $Y_1$ og $Y_2$ er uavhengige og binomisk fordelte med $EY_i = n_i p_i$ og $\operatorname{Var} Y_i = n_i p_i (1 - p_i)$. La $\theta_i = \ln\left(\frac{p_i}{1-p_i}\right)$ og $\psi = \exp(\theta_1 - \theta_2)$.

a) Gi en fortolkning av $\psi$. Vis at sannsynlighetsmaksimeringsestimatoren for $\psi$ er

$$\psi^* = \frac{Y_1(n_2 - Y_2)}{(n_1 - Y_1)Y_2} \, .$$

Angi en estimator for variansen til $\ln(\psi^*)$.

*(Fortsettes side 3.)*

b) Vis at $Y_1$ betinget mhp. $Z = Y_1 + Y_2 = z$ er ikke-sentralt hypergeometrisk fordelt, dvs. har tetthet

$$f_{Y_1|Z}(y) = \frac{\binom{n_1}{y}\binom{n_2}{z-y}\psi^y}{\sum\binom{n_1}{\nu}\binom{n_2}{z-\nu}\psi^\nu}$$

der summen tas over $\nu \in \{\max(0, z - n_2), \min(n_1, z)\}$.

c) Angi den sterkeste styrkerette $\alpha$-nivå testen for

$$H_0 : \psi \geq \psi_0 \qquad \text{mot} \qquad H_1 : \psi < \psi_0$$

der $\psi_0$ er en gitt størrelse.

d) Utfør testen for $\psi_0 = 2$, $n_1 = n_2 = 10$, $Y_1 = 2$, $Z = 8$ og $\alpha = 0.05$. Du kan benytte at $\sum_{\nu=0}^{8}\binom{10}{\nu}\binom{10}{8-\nu}2^\nu \approx 2.73 \cdot 10^6$.

# Oppgave 3.

Anta at $Y$ er binær med $\pi_j = P(Y = j)$, $\pi_1 + \pi_0 = 1$, $0 < \pi_j < 1$. Anta videre at for gitt $Y = j$ vil $X$ være normalfordelt med $E[X|Y = j] = \mu_j$ og $\text{Var}[X|Y = j] = \sigma^2$.

a) Vis at gitt $X = x$ er

$$p(x) = P(Y = 1 \mid X = x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}$$

der

$$\alpha = \ln\frac{\pi_1}{\pi_0} + \frac{\mu_0^2 - \mu_1^2}{2\sigma^2}$$

og

$$\beta = \frac{\mu_1 - \mu_0}{\sigma^2}\ .$$

b) Anta at $(X_i, Y_i)$ $i = 1, \ldots, n$ er uavhengige par fordelt som $(X, Y)$. La $n_j = \#\{Y_i = j\}$, $\overline{X}_j = \frac{1}{n_j}\Sigma X_i$ og $S_j^2 = \Sigma(X_i - \overline{X}_j)^2$ der summene tas over de $i$ med $Y_i = j$.

Hvorfor er

$$\hat{\beta} = \frac{\overline{X}_1 - \overline{X}_0}{S_1^2 + S_0^2}(n - 2)$$

en rimelig estimator for $\beta$? Angi en tilsvarende estimator $\hat{\alpha}$ for $\alpha$. Vis at $\hat{\alpha}$ og $\hat{\beta}$ er konsistente estimatorer. Utled et tilnærmet uttrykk for $\text{Var}(\hat{\beta})$.

*(Fortsettes side 4.)*

c) Den vanlige antagelsen i logistisk regresjon vil være å ignorere fordelingen til kovariatene $X_i = x_i$ og maksimere (den betingede) likelihooden

$$L = \prod_{i=1}^{n} p(x_i)^{Y_i}(1 - p(x_i))^{1-Y_i}$$

der $p(x)$ er gitt i punkt a).

Sett opp estimeringsligningene basert på $L$ og utled et uttrykk for informasjonsmatrisen. Hvordan vil du gå fram for å finne et tilnærmet uttrykk for $\mathrm{Var}(\beta^*)$ der $\beta^*$ er estimatoren for $\beta$ basert på $L$?

Du kan i dette punktet gjøre bruk av resultatet i oppgave 1c).

d) Sett nå at antagelsen om at $X_i$-ene gitt $Y_i = j$ er normalfordelt er feilaktig. Vi skal se på effekten av fortsatt å benytte estimatoren $\hat{\beta}$ fra punkt b). Anta, som et ekstremt tilfelle, at $X_i$-ene faktisk er binære. La antallene der $Y_i = j$ og $X_i = k$ være gitt i henhold til følgende tabell

| | | $X_i$ | |
|---|---|---|---|
| | | 1 | 0 |
| $Y_i$ | 1 | $A$ | $B$ |
| | 0 | $C$ | $D$ |

I henhold til oppgave 2 følger det at sannsynlighetsmaksimeringsestimatoren for $\beta$ basert på $L$ er gitt ved $\beta^* = \ln(A \cdot D/B \cdot C)$. Uttrykk $\hat{\beta}$ ved $A, B, C$ og $D$.

Vis at $\hat{\beta} = 0$ hvis og bare hvis $\beta^* = 0$ samt at $\hat{\beta}$ og $\beta^*$ har samme fortegn. Påvis, ved eksempel, at $\hat{\beta}$ og $\beta^*$ generelt er ulike.

Hva konvergerer $\beta^*$ mot når $n = A + B + C + D$ vokser?

Hva er den tilsvarende grensen for $\hat{\beta}$?

SLUTT