UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	ST 202 — Statistiske slutninger for den eksponentielle fordelingsklasse.
Eksamensdag:	Mandag 7. desember 1992.
Tid for eksamen:	09.00 – 15.00.
Oppgavesettet er på 6 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.

En stokastisk variabel Y sies å være negativt binomisk fordelt dersom punktsannsynlighetene til Y er gitt ved

$$P(Y=y) = \binom{y-1}{k-1} p^k (1-p)^{y-k}, \qquad y = k, k+1, \ldots$$

der $p \in [0,1]$ og $k \in \mathbb{N}$. Dette vil betegnes $Y \sim NB(k,p)$.

a) Vis at for gitt k utgjør familien $NB(k,p)$, $p \in [0,1]$ en eksponensiell klasse uten spredningsparameter, dvs. punktsannsynlighetene kan skrives på formen

$$f_Y(y;\theta) = \exp\{y\theta - b(\theta) + c(y)\}, \qquad y = k, k+1, \ldots$$

Gi eksplisitte uttrykk for θ, $b(\theta)$ og $c(y)$. Benytt disse til å vise at

$$\mu = EY = \frac{k}{p}$$

(Fortsettes side 2.)

og

$$\operatorname{Var} Y = k\frac{1-p}{p^2} .$$

Finn variansfunksjonen $V(\mu)$.

Anta at $Y_1, \ldots, Y_n$ er uavhengige med $Y_i \sim NB(k, p_i)$ der k er felles for Y_i-ene og p_i-ene funksjoner av kovariater $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$.

b) Hvilke ytterligere spesifikasjoner kreves for konstruksjon av en generalisert lineær modell? En av modellene spesifisert på denne måten sies å ha kanonisk linkfunksjon. Hva innebære dette? Vis at den kanoniske linkfunksjonen blir

$$g(\mu) = \ln\left(1 - \frac{k}{\mu}\right) .$$

Angi en begrensning på sammenhengen mellom kovariatene og regresjonsparametrene som vil gjøre denne modellen veldefinert.

c) Sett $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$ der $\mu_i = EY_i$ og anta at vi har observert $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ lik $\boldsymbol{y} = (y_1, \ldots, y_n)$. Deviansen er nå definert ved

$$D(\boldsymbol{y}, \boldsymbol{\mu}) = 2\sum_{i=1}^{n}[l_i(y_i; y_i) - l_i(\mu_i; y_i)]$$

der $l_i(\mu_i, y_i)$ er log-likelihood bidraget for observasjon i uttrykt ved μ_i. Hvorfor blir devians og skalert devians i denne modellen like? Finn et eksplisitt uttrykk for deviansen.

d) Anta at vi har tilpasset en generalisert lineær modell som i pkt. b) med regresjonsparametre $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ ved å minimere deviansen $D(\boldsymbol{y},\boldsymbol{\mu})$ mht. $\boldsymbol{\beta}$. La $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)$ være den minimerende verdi av $\boldsymbol{\beta}$. Begrunn at $\hat{\boldsymbol{\beta}}$ er tilnærmet normalfordelt. Angi forventning og kovariansmatrise (ved et generelt uttrykk) i denne normalfordelingen. Hva er de viktigste betingelser for at tilnærmelsen skal holde?

e) Anta at $x_{i1} = 1$, $i = 1, \ldots, n$. Forklar at under tilpasning av kun denne kovariaten blir de estimerte forventningene $\mu_i^* = \mu^*$, dvs. like for alle $i = 1, \ldots, n$. Sett $\boldsymbol{\mu}^* = (\mu^*, \ldots, \mu^*)$ og la $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \ldots, \hat{\mu}_n)$ være de estimerte forventningene etter tilpasning av modellen med alle $p > 1$ kovariater. Angi den tilnærmede fordeling for

$$D(\boldsymbol{Y}, \boldsymbol{\mu}^*) - D(\boldsymbol{Y}, \hat{\boldsymbol{\mu}}) .$$

(Fortsettes side 3.)

når $\beta_2 = \beta_3 = \cdots = \beta_p = 0$. Begrunn svaret, men unngå tekniske utledninger.

Negativt binomiske fordelinger oppstår blant annet på bakgrunn av Bernoulli forsøksrekker på den måte at $Y_i \sim NB(k, p_i)$ dersom vi for uavhengige indikatorvariable I_{ji} med $P(I_{ji} = 1) = p_i$ har $Y_i = y \Leftrightarrow \sum_{j=1}^{y-1} I_{ji} < k = \sum_{j=1}^{y} I_{ji}$. Y_i er altså antall forsøk til nøyaktig k suksesser. (Dette skal ikke vises.)

f) Y_i-ene i modellen i pkt. b) kan altså tenkes å ha oppstått ved observasjonsplanen å observere I_{ji} inntil k suksessen, der p_i er en funksjon av kovariatene $\boldsymbol{x}_i$. I dette punktet skal vi anta en annen observasjonsplan, nemlig at vi observerer $Z_i = \sum_{j=1}^{m_i} I_{ji}$ der m_i er bestemt på forhånd.

Hva slags type generalisert lineære modeller er rimelige for å beskrive Z_i-ene? Hvilken linkfunksjon er kanonisk for denne typen? Denne linkfunksjonen kan uttrykkes gjennom p_i-ene. Derfor vil den også indusere en linkfunksjon i modellen for Y_i-ene. Vis at denne induserte linkfunksjonen blir

$$g_0(\mu) = \ln\left(\frac{k}{\mu - k}\right).$$

og er ulik den kanoniske $g(\mu)$ fra pkt. b). Kommenter på denne bakgrunn begrepet kanonisk link.

g) Anta nå at $Y_i \sim NB(k_i, p_i)$, uavhengige, $i = 1, \ldots, n$, der ikke alle k_i er like, er modellert med samme lineære struktur uttrykt i p_i-ene som i pkt. b). Betrakt variansfunksjonen du fant i pkt. a). Hvilket problem oppstår?

Oppgave 2.

Vi skal i denne oppgaven se på tester i forbindelser med negativt binomiske stokastiske variable som definert i oppgave 1.

a) Anta at $Y_1, \ldots, Y_n \sim NB(k, p)$ uavhengige der k er kjent. Uttrykk den overalt sterkeste α-nivå testen for $H_0 : p \leq p_0$ mot $H_1 : p > p_0$. Begrunn svaret.

(Fortsettes side 4.)

b) Anta at $Y_i \sim NB(k_i, p_i)$ uavhengige, $i = 1, 2$ der k_1 og k_2 er kjent. Utled den overalt sterkeste styrkerette α-nivå testen for $H_0 : p_1 \geq p_2$ mot $H_1 : p_1 < p_2$. Utfør testen for $k_1 = 1$, $k_2 = 2$, $Y_1 = 10$, $Y_2 = 3$ og $\alpha = 0.05$.

Hint: Vis at

$$P(Y_1 = y_1 | Y_1 + Y_2 = z) = \frac{\binom{y_1 - 1}{k_1 - 1}\binom{z - y_1 - 1}{k_2 - 1}\left(\frac{1-p_1}{1-p_2}\right)^{y_1}}{\sum_{y=k_1}^{z-k_2}\binom{y-1}{k_1-1}\binom{z-y-1}{k_2-1}\left(\frac{1-p_1}{1-p_2}\right)^{y}}$$

c) Anta at $Y_i \sim NB(k, p_i)$ uavhengige, $i = 1, \ldots, n$ der k er kjent, kan beskrives med en generalisert lineær modell med kanonisk linkfunksjon og lineær prediktor $\eta_i = \sum_{j=1}^{p} \beta_j x_{ij}$. Forklar hvordan man går fram for å finne den overalt sterkeste styrkerette testen med nivå α for hypoteser av typen $H_0 : \beta_j \leq 0$ mot $H_1 : \beta_j > 0$. Begrunn svaret.

Oppgave 3.

I denne oppgaven skal det ses på situasjonen der hvert individ kan ha begge, det ene eller ingen av to kjennetegn A og B. F.eks. kan kjennetegn A være at individet er HIV-smittet, mens kjennetegn B er at individet har vært hepatitt B-smittet. Vi er spesielt interessert i å se på avhengighet mellom A og B, gitt kovariater $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$, $i = 1, \ldots, n$. La

Y_{1i} = indikatorvariabel for kjennetegn A,
Y_{2i} = indikatorvariabel for kjennetegn B

og

Y_{jki} = indikatorvariabel for $Y_{1i} = j$ og $Y_{2i} = k$, $j, k = 0, 1$

for individ $i = 1, \ldots, n$.

Vi skal i hele oppgaven anta at parrene (Y_{1i}, Y_{2i}) er uavhengige, $i = 1, \ldots, n$.

a) Anta at Y_{si}-ene, $s = 1, 2$, marginalt kan modelleres med logistisk regresjon, dvs.

$$p_{si} = P(Y_{si} = 1 | \boldsymbol{x}_i) = \frac{e^{\boldsymbol{\beta}_s \boldsymbol{x}_i}}{1 + e^{\boldsymbol{\beta}_s x_i}} \tag{1}$$

der $\boldsymbol{\beta}_s = (\beta_{s1}, \ldots, \beta_{sp})$ og $\boldsymbol{\beta}_s \boldsymbol{x}_i = \sum_{j=1}^{p} \beta_{sj} x_{ij}$, $s = 1, 2$. Anta videre at, gitt $\boldsymbol{x}_i$, så er Y_{1i} og Y_{2i} uavhengige.

(Fortsettes side 5.)

Vis at dette er ekvivalent med et spesialtilfelle av den multinære logistiske regresjonsmodell for Y_{jki} , $j, k = 0, 1$, dvs.

$$\pi_{jki} = P(Y_{jki} = 1|\boldsymbol{x}_i) = \frac{\exp\{\boldsymbol{\gamma}_{jk}\boldsymbol{x}_i\}}{\sum_{j=0}^{1}\sum_{k=0}^{1}\exp\{\boldsymbol{\gamma}_{jk}\boldsymbol{x}_i\}} \tag{2}$$

der $\boldsymbol{\gamma}_{jk} = (\gamma_{jk1}, \ldots, \gamma_{jkp})$ og $\gamma_{00l} = 0$, $l = 1, \ldots, p$. Identifiser sammenhengen mellom $\boldsymbol{\gamma}_{jk}$-ene og $\boldsymbol{\beta}_1$ og $\boldsymbol{\beta}_2$ i dette spesialtilfellet.

b) Sett $\boldsymbol{\pi}_i = (\pi_{00i}, \pi_{01i}, \pi_{10i}, \pi_{11i})$ og $\boldsymbol{Y}_i = (Y_{00i}, Y_{01i}, Y_{10i}, Y_{11i})$ og la $\boldsymbol{\pi} = (\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_n)$ og $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$. Hvordan defineres deviansen $D_M(\boldsymbol{Y}, \boldsymbol{\pi})$ i den multinære logistiske regresjonsmodellen (2) uttrykt ved $\boldsymbol{\pi}_i, \boldsymbol{Y}_i$ og log-likelihood bidragene $l_i = l(\boldsymbol{\pi}_i, \boldsymbol{Y}_i)$. Vis at denne eksplisitt kan skrives som

$$D_M(\boldsymbol{Y}, \boldsymbol{\pi}) = 2\sum_{i=1}^{n}\sum_{j=0}^{1}\sum_{k=0}^{1} Y_{jki} \ln\left[\frac{Y_{jki}}{\pi_{jki}}\right]$$

der $0 \cdot \ln(0) = 0$ per definisjon.

Betrakt $\boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\gamma})$ som en funksjon av $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_{01}, \boldsymbol{\gamma}_{10}, \boldsymbol{\gamma}_{11})$ og sett $\hat{\boldsymbol{\pi}} = \boldsymbol{\pi}(\hat{\boldsymbol{\gamma}})$ og $\boldsymbol{\pi}^* = \boldsymbol{\pi}(\gamma^*)$ der $\hat{\boldsymbol{\gamma}}$ og $\boldsymbol{\gamma}^*$ er sannsynlighetsmaksimeringsestimatorene for $\boldsymbol{\gamma}$ i den multinære logistiske regresjonsmodell (2) henholdsvis generelt og under uavhengighetsmodellen (1).

Hvorfor vil $D_M(\boldsymbol{Y},\boldsymbol{\pi}^*) - D_M(\boldsymbol{Y}, \hat{\boldsymbol{\pi}})$ være tilnærmet χ^2-fordelt under uavhengighetsmodellen (1)? Hva blir antall frihetsgrader i denne χ^2-fordelingen?

c) La $D_s(\boldsymbol{Y}_s, \boldsymbol{p}_s)$ være deviansen for $\boldsymbol{Y}_s = (Y_{s1}, \ldots, Y_{sn})$ uttrykt ved $\boldsymbol{p}_s = (p_{s1}, \ldots, p_{sn})$, $s = 1, 2$ i uavhengighetsmodellen (1). Betrakt $\boldsymbol{p}_s = \boldsymbol{p}_s(\boldsymbol{\beta}_s)$ som en funksjon av $\boldsymbol{\beta}_s$ og la $\boldsymbol{\beta}_s^*$ være sannsynlighetsmaksimeringsestimatoren for $\boldsymbol{\beta}_s$ under (1). Hvilken sammenheng tilfredstiller $\boldsymbol{\gamma}^*, \boldsymbol{\beta}_1^*$ og $\boldsymbol{\beta}_2^*$? Vis at man kan sette, med $\boldsymbol{p}_s^* = \boldsymbol{p}_s(\boldsymbol{\beta}_s^*)$,

$$D_M(\boldsymbol{Y}, \boldsymbol{\pi}^*) = D_1(\boldsymbol{Y}_1, \boldsymbol{p}_1^*) + D_2(\boldsymbol{Y}_2, \boldsymbol{p}_2^*)\,.$$

Forklar hvordan man på denne bakgrunn kan gå fram for å teste at Y_{1i}-ene og Y_{2i}-ene, gitt x_i-ene, er uavhengige.

d) Vis at under den generelle multinære logistiske regresjonsmodellen (2) så vil modellene for Y_{jki} gitt at $Y_{00i} + Y_{jki} = 1$, $i = 1, \ldots, n$, $jk =$

(Fortsettes side 6.)

$01, 10, 11$, være gitt som vanlige logistiske regresjonsmodeller. Forklar hvordan dette kan benyttes til å estimere $\boldsymbol{\gamma}_{jk}$-ene selv om man ikke har programvare for multinær logistisk regresjon tilgjengelig.

e) Vis at generelt, dvs. under (2), så vil modellen for Y_{1i} gitt $Y_{2i} = k$, $k = 0, 1$, kunne beskrives ved vanlig logistisk regresjon. Angi kovariatene i denne modellen samt sammenhengen mellom regresjonsparametrene og $\boldsymbol{\gamma}_{jk}$-ene.

SLUTT