UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	ST 213 — Generaliserte lineære modeller
Eksamensdag:	Fredag 11. desember 1998.
Tid for eksamen:	09.00 – 15.00.
Oppgavesettet er på 6 sider.	
Vedlegg:	Tabeller over standard normal-fordelingen og kji-kvadrat-fordelingen.
Tillatte hjelpemidler:	Formelsamlinger for ST101 og ST102, lommeregner.

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

Oppgave 1.

I en studie av toksisiteten (giftigheten) av karbondisulfid (CS_2) ble en bestemt billeart utsatt for ulike CS_2-konsentrasjoner. Ca. 60 biller ble benyttet ved hver av 8 konsentrasjoner, og antallet som døde ble registrert. Resultatet ble som følger:

Dose ($\log_{10} CS_2 mg\, l^{-1}$)	Antall biller	Antall døde
1,691	59	6
1,724	60	13
1,755	62	18
1,784	56	28
1,811	63	52
1,837	59	53
1,861	62	61
1,884	60	60

(Fortsettes side 2.)

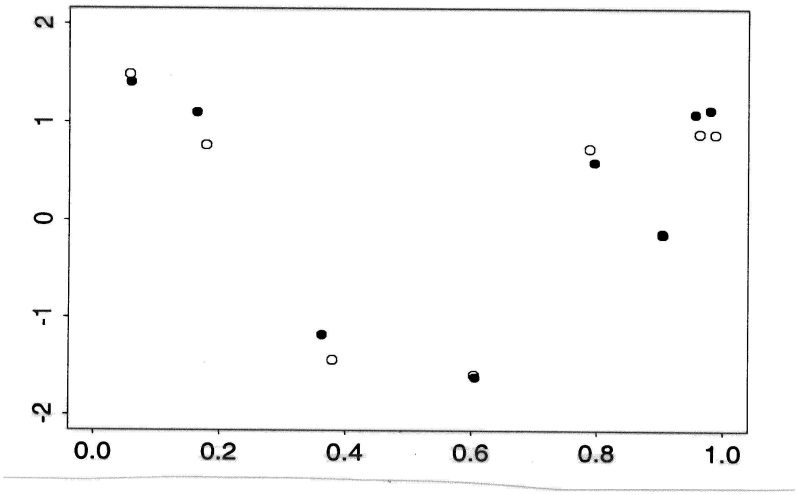
Dataene ble analysert ved en generalisert lineær modell med binomisk fordeling for antall døde og med giftdose som kovariat. Både logistisk-link og probit-link ble benyttet for å beskrive sammenhengen mellom dødssannsynligheten $\pi(x)$ og dosen x. Resultatene av disse analysene ble som følger:

Parameter	Logit-link	Probit-link
Konstant	$-60,72$	$-34,93$
Dose	$34,27$	$19,73$

a) Skriv, både for den logistiske regresjonen og for probitanalysen, opp et uttrykk for den estimerte dødssannsynligheten $\widehat{\pi}(x)$ ved dose x. Bestem de estimerte sannsynlighetene for $x = 1,724$.

b) Definer Pearson residualene, og beregn disse for begge analysene for $x = 1,724$.

c) I figuren nedenfor er Pearson residualene for begge analysene gitt i samme diagram plottet som funksjon av de tilpassede sannsynlighetene $\widehat{\pi}(x)$. Forskjellige symboler er benyttet for de to analysene. Avgjør hvilke av residualene som hører til den logistiske regresjonen og hvilke som hører til probitanalysen. Svaret skal begrunnes.



La Z være dosen som skal til for å drepe en tilfeldig valgt bille (den dødelige dosen), og la μ og σ være forventning og standardavik til Z. Da er $\pi(x) = \mathrm{P}(Z \leq x)$. Hvis Z er normalfordelt has

$$\Phi^{-1}(\pi(x)) = \frac{x - \mu}{\sigma},$$

hvor Φ er den kumulative standard normalfordelingen, mens hvis Z er logistisk fordelt has

$$g(\pi(x)) = \frac{\pi}{\sqrt{3}} \frac{x - \mu}{\sigma},$$

hvor $g(y) = \log\{y/(1-y)\}$. Disse resultatene skal du ikke vise.

(Fortsettes side 3.)

d) Redgjør for hvordan disse resultatene hjelper til å forklare den store forskjellen vi fikk i estimert konstantledd og regresjonsparameter ved bruk av logit-link og probit-link.

Oppgave 2.

I denne oppgaven skal vi studere hvordan dødeligheten blant danske diabetikere er i forhold til dødeligheten i den danske befolkningen forøvrig. Dataene vi skal benytte omfatter alle de 1499 insulinavhengige diabetikerne som bodde på Fyn 1. juli 1973. Disse ble fulgt opp via det sentrale personregisteret, og alle dødsfall før 1. januar 1982 ble registrert[1].

La $h_i(t)$ være dødsintensiteten (hasardraten) for i-te diabetiker, hvor t er alder i år. Vi er interessert i å studere hvordan denne er i forhold til befolkningsdødeligheten (slik den gis i offentlige statistiske publikasjoner) og hvordan dette forholdet avhenger av kovariatene:

$$x_{i1} = \begin{cases} 0 & \text{hvis } i \text{ er en kvinne} \\ 1 & \text{hvis } i \text{ er en mann} \end{cases}$$

$x_{i2} = a_i - 50$, hvor a_i er alderen til i pr. 1/7-73 (i år)

$$x_{i3} = \begin{cases} 0 & \text{hvis } i \text{ har hatt diabetes i 15 år eller mer pr. 1/7-73} \\ 1 & \text{hvis } i \text{ har hatt diabetes i mindre enn 15 år pr. 1/7-73} \end{cases}$$

Dette vil vi gjøre ved å tilpasse proporsjonale hasardmodeller av formen

$$h_i(t) = e^{\eta_i} \lambda(t; x_{i1}). \tag{1}$$

Her er $\lambda(t;0)$ og $\lambda(t;1)$ de kjente dødsintensitetene for kvinner og menn i den allmenne befolkningen, mens η_i er en lineær prediktor som angir hvordan kovariatene påvirker dødeligheten. For en modell med bare hovedeffekter har vi for eksempel $\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$.

Vi observerer i-te pasient fra alder a_i til død i alder T_i, hvis dødsfallet skjer før 1/1-82, eller til 1/1-82. I det siste tilfellet observerer vi bare den sensurerte levealderen $a_i + 8,5$ år. Våre data for den i-te pasienten er derfor

$$Y_i = \min\{T_i, a_i + 8,5\}$$

[1] Disse dataene ble også benyttet til den obligatoriske oppgaven høsten 1998. I den obligatoriske oppgaven studerte vi imidlertid den *absolutte* dødeligheten til diabetikerne, mens vi her studerer den *relative* dødeligheten, dvs. deres dødelighet i forhold til dødeligheten i den allmenne danske befolkning. En annen forskjell fra den obligatoriske oppgaven er at vi her benytter informasjon om alder ved eventuelt dødsfall, og ikke bare om en person døde eller ikke mellom 1/7-73 og 1/1-82.

(Fortsettes side 4.)

$$D_i = \begin{cases} 0 & \text{hvis } i \text{ lever lenger enn } Y_i \\ 1 & \text{hvis } i \text{ dør ved alder } Y_i \end{cases}$$

En kan vise at (Y_i, D_i) har fordelingsfunksjon

$$f_{(Y_i,D_i)}(y_i, d_i) = h_i(y_i)^{d_i} \exp\left\{-\int_{a_i}^{y_i} h_i(u)du\right\},$$

for $y_i > a_i$ og $d_i \in \{0, 1\}$. (Denne er en blanding av en tetthet og en punktsannsynlighet.) Du skal ikke vise dette.

a) Vis at likelihooden er proporsjonal med

$$\prod_{i=1}^{n} \frac{\mu_i^{d_i}}{d_i!} e^{-\mu_i},$$

hvor

$$\mu_i = e^{\eta_i} \int_{a_i}^{y_i} \lambda(u; x_{i1})du,$$

og forklar hvorfor modellen (1) kan analyseres som om D_i er Poisson-fordelt med forventning μ_i. Forklar også hvordan en kan benytte programvare for generaliserte lineære modeller for å tilpasse modellen.

Tabellen nedenfor gir devianser og frihetsgrader for ulike tilpassede modeller. Under "Modell" er den lineære prediktoren η_i i (1) angitt med vanlig notasjon. Her står K for kjønn (x_{i1}), A for alder (x_{i2}) og V for varighet (x_{i3}).

Modell	Devians	Frihetsgrader
1	1623,2	1498
K	1622,9	1497
A	1567,2	1497
V	1607,3	1497
K + A	1565,6	1496
K + V	1607,1	1496
A + V	1552,5	1496
K + A + V	1550,9	1495
K + A + V + K.A	1549,4	1494
K + A + V + K.V	1550,8	1494
K + A + V + A.V	1548,6	1494
K + A + V + K.A + K.V	1549,4	1493
K + A + V + K.A + A.V	1547,3	1493
K + A + V + K.V + A.V	1548,5	1493
K + A + V + K.A + K.V + A.V	1547,3	1492
K + A + V + K.A + K.V + A.V + K.A.V	1544,6	1491

(Fortsettes side 5.)

b) Forklar hvordan en deviansanalyse leder fram til modellen $\eta_i = \beta_0 + \beta_2 x_{i2} + \beta_3 x_{i3}$, dvs. modellen A + V med alder og varighet som hovedeffekter og ingen interaksjoner eller effekt av kjønn. Er deviansanalysen tilstrekkelig til å vurdere om modellen er tilfredsstillende? Diskuter.

Estimatene for modellen i punkt b er

Parameter	Estimat	Standardfeil
Konstant (β_0)	$1,742$	$0,073$
Alder (β_2)	$-0,0233$	$0,0030$
Varighet (β_3)	$-0,347$	$0,090$

med estimert korrelasjonsmatrise (nedre halvdel)

$$\begin{array}{lll} 1{,}000 & & \\ -0,536 & 1{,}000 & \\ -0,591 & 0{,}0253 & 1{,}000 \end{array}$$

(Parametrene er gitt i samme rekkefølge som i tabellen over.)

c) Gi en fortolkning av estimatene i lys av modellformuleringen (1).

Det kan vises at sannsynligheten for at en 40 år gammel person skal bli minst 60 år er $\exp\left\{-\int_{40}^{60} h(u)du\right\}$, hvor $h(t)$ er personens dødsintensitet. Du skal ikke vise dette.

d) For den allmenne danske befolkning er sannsynligheten 0,916 for at en 40 år gammel kvinne skal bli minst 60 år, mens den tilsvarende sannsynligheten for en mann er 0,870. Bruk dette og den tilpassede modellen til å estimere sannsynligheten for at en 40 år gammel diabetiker skal bli minst 60 år under følgende forutsetninger:

- Diabetikeren er en kvinne som fikk sykdommen da hun var 18 år.
- Diabetikeren er en mann som fikk sykdommen da han var 33 år.

e) Bestem et 95% konfidensintervall for den første av sannsynlighetene i punkt d.

Oppgave 3.

Vi betrakter en stokastisk variabel Y med tetthet/punktsannsynlighet $f_Y(y;\theta)$, hvor θ er en skalar parameter. Vi innfører log-likelihood funksjonen

$$l = l(\theta; Y) = \log\left\{f_Y(Y;\theta)\right\}.$$

(Fortsettes side 6.)

a) Vis at

$$E_\theta\left(\frac{\partial l}{\partial \theta}\right) = 0$$

$$E_\theta\left(\frac{\partial^2 l}{\partial \theta^2}\right) + E_\theta\left\{\left(\frac{\partial l}{\partial \theta}\right)^2\right\} = 0.$$

(under passende regularitetsbetingelser).

På lignende måte som i punkt a, kan det også vises at

$$E_\theta\left(\frac{\partial^3 l}{\partial \theta^3}\right) + 3E_\theta\left\{\frac{\partial^2 l}{\partial \theta^2}\cdot\frac{\partial l}{\partial \theta}\right\} + E_\theta\left\{\left(\frac{\partial l}{\partial \theta}\right)^3\right\} = 0 \qquad (2)$$

Du skal ikke vise dette.

I resten av oppgaven skal vi betrakte den eksponensielle fordelingsklassen hvor Y har tetthet/punktsannsynlighet på formen

$$f_Y(y;\theta) = \exp\{y\theta - b(\theta) + c(y)\} \qquad (3)$$

for en skalar parameter θ og funksjoner $b(\theta)$ og $c(y)$.

b) Vis at den binomiske fordeling med parametre (m, π) kan skrives på formen (3) for passende valg av θ, $b(\theta)$ og $c(y)$.

c) Vis at for fordelinger av formen (3) har vi

$$\mu = \mathrm{E}(Y) = b'(\theta) \quad \text{og} \quad \mathrm{Var}(Y) = b''(\theta).$$

d) Variansfunksjonen $V(\mu)$ for fordelinger av formen (3) er $\mathrm{Var}(Y)$ uttrykt som funksjon av forventningsverdien μ (siden vi her ikke har noen spredningsparameter). Vis at $\mathrm{E}\{(Y-\mu)^3\} = V(\mu)V'(\mu)$.

[*Vink:* Bruk først (2) til å vise at $\mathrm{E}\{(Y - b'(\theta))^3\} = b'''(\theta)$.]

e) Bestem tredje sentralmoment $\mathrm{E}\{(Y-\mu)^3\}$ for den binomiske fordelingen.

SLUTT