UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i: STK3100 — Innføring i genera..., løsningsforslag.

Eksamensdag: Mandag 6. desember 2010

Tid for eksamen: 14.30 – 18.30

Oppgavesettet er på 4 sider.

Vedlegg: Tabell over normalfordeling og χ^2-fordeling

Tillatte hjelpemidler: Godkjent kalkulator og formelsamling for STK1100/STK1110 og STK1120/STK2120

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

Løsningsforslag

Oppgave 1

a) Betegn respons med y og kovariatene med $x_1, \ldots, x_p$
Generell form for GLM:
Tetthet/punktsannsynlighet for respons: $f(y;\theta,\phi) = c(y,\phi)\exp(\frac{y\theta - a(\theta)}{\phi})$
Lineær prediktor: $\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$
Link: $\eta = g(\mu); \mu = E(y)$

I dette tilfellet:

Tetthet for respons: $\pi^y(1-\pi)^{1-y} = \exp(y\log(\frac{\pi}{1-\pi}) + \log(1-\pi))$ dvs. $\theta = \log(\frac{\pi}{1-\pi}), \pi = \frac{\exp(\theta)}{1+exp(\theta)}, a(\theta) = -\log(1-\pi) = \log(1+exp(\theta)), c(y,\phi) = \phi = 1.$
Lineær prediktor: $\eta = \beta_0 + \beta_1 x$
Logit link: $\eta = g(\mu) = \log(\frac{\pi}{1-\pi}); \mu = E(y) = a'(\theta) = \frac{\exp(\theta)}{1+exp(\theta)} = \pi.$

b) $\pi(30) = \frac{\exp(\beta_0+\beta_1 30)}{1+\exp(\beta_0+\beta_1 30)}$.
Herav log odds $\log(\frac{\pi(30)}{1-\pi(30)}) = \beta_0 + \beta_1 30$, og log oddsforhold $\log(\frac{\pi(30)(1-\pi(40))}{(1-\pi(30))\pi(40)}) = \beta_1(30-40) = -10\beta_1$,
som innsatt gir oddsforholdet $OR_1 = \exp(-10 \times 0.07038) = 0.4946881$.

Et 95% konfidensintervall for log oddsforholdet er gitt ved $-10 \times \hat{\beta}_1 \pm 1.96 \times 10 \times s_{\hat{\beta}_1}$ der $s_{\hat{\beta}_1}$ er den estimerte standardfeilen til $\hat{\beta}_1$. Fra utskriften $-10 \times 0.07038 \pm 1.96 \times 10 \times 0.02667$, som gir intervallgrensene -1.226519 og -0.1811365. For oddsforholdet er konfidensintervallet derfor
(exp(-1.226519),exp(-0.1811365))=(0.293311841,0.8343214).

(Fortsettes på side 2.)

c) Den predikerte sannsynligheten er $\hat{\pi}(40) = \frac{\exp(\hat{\beta}_0+\hat{\beta}_1 40)}{1+\exp(\hat{\beta}_0+\hat{\beta}_1 40)}$ som innsatt gir $\frac{\exp(-2.21358+40\times 0.07038)}{1+\exp(-2.21358+40\times 0.07038)} = 0.6460518$. Et 95% konfidensinterval for $\beta_0 + \beta_1 40$ er gitt ved $\hat{\beta}_0 + \hat{\beta}_1 40 \pm 1.96\sqrt{varest}$ der er $varest$ er estimatet for $Var(\hat{\beta}_0 + 40\hat{\beta}) = Var(\hat{\beta}_0) + 40^2 Var(\hat{\beta}) + 2 \times 40 \times Cov(\hat{\beta}_0, \hat{\beta}_1)$, dvs $0.99874^2 + 1600 \times 0.02667^2 + 2 \times 40 \times (-0.906) \times 0.99874 \times 0.02667 = 0.2039967$. Konfidensintervallet for $\beta_0 + \beta_1 40$ er derfor $-2.21358 + 40 \times 0.07038 \pm 1.96 \times \sqrt{0.2039967}$ eller (-0.2835082, 1.486966). Siden $\exp(x)/(1 + \exp(x)$ er en voksende funksjon blir konfidensintervallet for den predikerte sannsynligheten $\exp(-0.2835082)/(1 + \exp(-0.2835082), \exp(1.486966)/(1 + \exp(1.486966) = (0.4295939, 0.8156225)$.

d)
```
Analysis of Deviance Table

Model 1: sore ~ 1
Model 2: sore ~ I(duration)
Model 3: sore ~ I(duration) + factor(type)
Model 4: sore ~ I(duration) + I(duration^2) + factor(type)
  Resid. Df Resid. Dev Df Deviance
1        34     46.180
2        33     33.651  1   12.528
3        32     30.138  1    3.513
4        31     30.133  1    0.005
```

Her ser vi at en test for model 3 mot model 4, dvs H_0 at leddet x^2 kan sløyfes ikke er signifikant, p-verdien er 0.95 i en χ^2 fordeling med 1 frihetsgrad. Heller ikke neste forenkling, dvs at hypotesen at koeffesienten for faktoren `type` er lik null, er spesielt signifikant, p-verdien er 0.06=P(X>3.513) i en χ^2 fordeling med 1 frihetsgrad. Ytterligere forenklinger gir derimot sterk signifikans, p-verdien er 0.0004=P(X>12.528) i en χ^2 fordeling med 1 frihetsgrad.

e) Ved biomisk respons er deviansen

$$2\sum_{i=1}^{n}[y_i \log(\frac{y_i}{\hat{\mu}_i}) + (n_i - y_i)\log(\frac{n_i - y_i}{n_i - \hat{\mu}_i})]$$

der y_i er observerte og $\hat{\mu}_i$ er tilpassede verdier.

Ved binær respons er $n_i = 1, i = 1, \cdots, n = 35$ og $\hat{\mu}_i = \hat{\pi}_i$ slik at deviansen blir

$$2\sum_{i=1}^{35}[y_i \log(\frac{y_i}{\hat{\pi}_i}) + (1 - y_i)\log(\frac{1 - y_i}{1 - \hat{\pi}_i})].$$

Siden y_i har verdiene 0 eller 1, blir uttrykk av formen $y_i \log(y_i)$ og $(1 - y_i)\log(1 - y_i)$ lik 0. Deviansen reduseres derfor til

$$-2\sum_{i=1}^{35}[y_i \log(\frac{\hat{\pi}_i}{1 - \hat{\pi}_i}) + \log(1 - \hat{\pi}_i)].$$

(Fortsettes på side 3.)

Ligningene for bestemmelse av SME er $X'D(y - \hat{\mu}) = 0$, jfr ligning (5.4) på side 68 i læreboka, der X er designmatrisen, y er vektoren av observasjoner og $\hat{\mu} = \hat{\pi}$ er vektoren av tilpassede verdier. Matrisen D er diagonalmatrisen med diagonalelementer av formen $1/V(\hat{\mu}_i)g'(\hat{\mu}_i), i = 1, \cdots, n$, der V er variansfunksjonen og g linkfunksjonen. I dette tilfellet reduseres D derfor til identitetsmatrisen slik at ligningene blir $X'y = X'\hat{\pi}$.

Da er den første summen i deviansen $-2y'\hat{\eta}$ der $\eta = X\hat{\beta}$ er vektoren av lineære prediktorer. Derfor er $-2y'\hat{\eta} = -2y'X\hat{\beta} = -2(X'y)'\hat{\beta} = -2(X'\hat{\pi})'\hat{\beta} = -2\hat{\pi}'X\hat{\beta} = -2\hat{\pi}'\hat{\eta}$, som gir resultatet.

Vi ser at deviansen bare avhenger av observasjonene gjennom de tilpassede verdiene. Det går derfor ikke an å sammenligne de observerte vardiene og de tilpassede verdiene dvs. vurdere føyningen ved å se på deviansen.

Oppgave 2

a) Antallet dødsfall av denne typen kan også ses som antallet suksesser"(i dette tilfellet dødsfall), i et stort antall forsøk. Hvert personår er et forsøk. Det betyr at antallet suksesser er binomisk fordelt. Her er suksess-sannsynheten, p, liten og antallet forsøk, m, stort, og da er fordelingen til antallet tilnærmet Poissonfordelt med forventing λ når mp er nær λ.

Forventet antall kan uttrykkes som en rate som er proposjonal med størrelsen på området, intervallet eller populasjonen antallet angis for, dvs $\lambda = N\mu$ der N er størrelsen og μ raten. I dette tilfellet angis størrelsen med antall personår.

Modellen blir derfor at responsene $y_1, \cdots, y_n$ er uavhengige Poisson-fordelte variable med forventning $N_i\mu_i$ der N_i er antallet personår i gruppene og μ_i er ratene.

I modellen som er tilpasset angis alder med et andregradspolynom, og røyking er en faktor med to nivåer. I tillegg er det et samspillsledd for koeffesienten foran 1'te gradsleddet og faktoren røyking. Det betyr at raten har formen

$$\mu = \begin{cases} \exp(\beta_0 + \beta_1 age + \beta_2 age^2) & \text{for ikke} - \text{roeykere} \\ \exp(\beta_0 + \beta_3 + (\beta_1 + \beta_4)age + \beta_2 age^2) & \text{for roeykere} \end{cases}$$

Vi ser fra modeltilpasningen at både alder og røyking ser ut til å ha betydning, men samspillsleddet gjør sammenhengen mer komplisert.

b) Siden fornetningen utrykkes som $N_i\mu_i = \exp(\log(N_i))\mu_i$ vil den lineære prediktoren i tillegg til den lineære kombinasjonen beskrevet ovenfor, inneholde et ledd av typen $\log(N_i)$, med andre ord et ledd hvor koeffesienten er kjent og lik 1. Ledd av denne typen, der koeffesientene ikke skal estimeres, betegnes som "offset".

c) Fra uttykket i punkt a) ser vi at forholdet mellom ratene for røykere og ikke-røykere blir

$$\frac{\exp(\beta_0 + \beta_3 + (\beta_1 + \beta_4)age + \beta_2 age^2)}{\exp(\beta_0 + \beta_1 age + \beta_2 age^2)} = \exp(\beta_3 + \beta_4 age)$$

(Fortsettes på side 4.)

som altså varierer med alderen. For leger på 40 år er det estimerte forholdet $\exp(\hat\beta_3 + \hat\beta_4 \times 40) = \exp(2.370 - 0.03084 \times 40) = 3.114493$ og leger på 70 $\exp(\hat\beta_3 + \hat\beta_4 \times 70) = \exp(2.370 - 0.03084 \times 70) = 1.234766$. Dødeligheten på grunn av hjertesykdommer som kan tilskrives røyking, er altså nesten en tredjedel for 70 åringer i forhold til 40 åringer. En forklaring kan rett og slett være at de som er mest utsatt på grunn av røking har en overdødelighet i yngre alder.

d) Betrakt nullhypotesen $H_0 : C\beta = r$ der C er en $q \times 6$ matrise, r er en $q \times 1$ vektor og $\beta' = (\beta_0, \beta_1, \cdots, \beta_5)$ er vektoren av ukjente parametre. Wald resten gir forkastning for store verdier av $(C\hat\beta - r)'(C\Sigma_{\hat\beta}C')^{-1}(C\hat\beta - r)$ der $\hat\beta$ er sannsynlighetsmaksimeringsestimatoren og $\Sigma_{\hat\beta}$ er den estimarte kovariansmatrisen til $\hat\beta$. Under H_0 er testobservatoren tilnærmet χ^2_q-fordelt. I dette tilfellet er $H_0 : \beta_4 = \beta_5 = 0$, som svarar til $q = 2$, $r = (0, 0)'$. C matrisen har rader $(0, 0, 0, 0, 1, 0)$ og $(0, 0, 0, 0, 0, 1)$ slik at $C\Sigma_{\hat\beta}C'$ er den estimerte kovariansmatrisen til $(\hat\beta_4, \hat\beta_5)$, altså den som er oppgitt i oppgaveteksten. Da er Wald-observatoren

$$(-0.0975518230, 0.0005195636) \begin{pmatrix} 1.143363e-02 & -8.807653e-05 \\ -8.807653e-05 & 6.844424e-07 \end{pmatrix}^{-1} \begin{pmatrix} -0.0975518230 \\ 0.0005195636 \end{pmatrix}$$

som er lik 9.85051 og gir klar forkastning ved sammenligning med en χ^2-fordeling med 2 frihetsgrader.

e) I modellene ovenfor brukes kanonisk link, dvs $\eta = \log(\mu) = \theta = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$ slik at likelihooden er

$$\prod_{i=1}^{n} \exp(y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) - \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}))$$

og log-likelihood

$$l = \sum_{i=1}^{n} [y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) - \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})].$$

Herav

$$\frac{\delta l}{\delta \beta_j} = \sum_{i=1}^{n} [y_i x_{ij} - x_{ij} \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})]$$

og

$$\frac{\delta^2 l}{\delta \beta_j \delta \beta_k} = -\sum_{i=1}^{n} x_{ij} x_{ik} \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})$$

for alle $j, k = 0, \ldots, p$, der $x_{i0} = 1, i = 1, \ldots, n$. Den observerte informasjonsmatrisen har elementer $-\frac{\delta^2 l}{\delta \beta_j \delta \beta_k}$, som alle er ikke-tilfeldige. De forventede verdiene blir derfor de samme. Siden den forventede informasjoensmatrisen er forventningen av den observerte, må de i dette tilfellet bli like.

SLUTT