Norges teknisk–naturvitenskapelige universitet
Institutt for matematiske fag

Bokmål

Faglig kontakt under eksamen:
Håkon Tjelmeland 73 59 35 38

EKSAMEN I EMNE TMA4315 GENERALISERTE LINEÆRE MODELLER

Torsdag 29. november 2007
Tid: 09:00–13:00

Hjelpemidler: *Tabeller og formler i statistikk*, Tapir forlag,
K. Rottmann, Matematisk formelsamling,
Ett gult A4-ark med IMF-stempel med egne håndskrevne formler og notater,
Godkjent enkel kalkulator.

Sensur er senest ferdig: Torsdag 20. desember 2007.

Oppgave 1

Tabell 1 viser resultatene av en undersøkelse om lungekrefttilfeller i en populasjon i et bestemt år. Første kolonne i tabellen angir aldersintervall for personene som var med i undersøkelsen. I andre kolonne angis røykestatus som en av fire muligheter, ikke-røyker, røyker sigar og/eller pipe, røyker både sigaretter og sigar/pipe, og røyker kun sigaretter. Den tredje kolonnen angir antall personer (i 100 000) som var med i undersøkelsen for de ulike kombinasjoner av aldersgruppe og røykestatus, og den siste kolonnen angir tilhørende observert antall lungekrefttilfeller.

Vi skal benytte poisson-regresjon til å modellere observerte antall lungekrefttilfeller. En (noe modifisert) utskrift fra R for en tilpasset modell der aldersintervall og røykestatus er benyttet som forklaringsvariabler er gitt under.

age	smoke	pop	dead
40-44	no	656	18
45-49	no	359	22
50-54	no	249	19
55-59	no	632	55
60-64	no	1067	117
65-69	no	897	170
70-74	no	668	179
75-79	no	361	120
80+	no	274	120
40-44	cigarPipeOnly	145	2
45-49	cigarPipeOnly	104	4
50-54	cigarPipeOnly	98	3
55-59	cigarPipeOnly	372	38
60-64	cigarPipeOnly	846	113
65-69	cigarPipeOnly	949	173
70-74	cigarPipeOnly	824	212
75-79	cigarPipeOnly	667	243
80+	cigarPipeOnly	537	253
40-44	cigarrettePlus	4531	149
45-49	cigarrettePlus	3030	169
50-54	cigarrettePlus	2267	193
55-59	cigarrettePlus	4682	576
60-64	cigarrettePlus	6052	1001
65-69	cigarrettePlus	3880	901
70-74	cigarrettePlus	2033	613
75-79	cigarrettePlus	871	337
80+	cigarrettePlus	345	189
40-44	cigarretteOnly	3410	124
45-49	cigarretteOnly	2239	140
50-54	cigarretteOnly	1851	187
55-59	cigarretteOnly	3270	514
60-64	cigarretteOnly	3791	778
65-69	cigarretteOnly	2421	689
70-74	cigarretteOnly	1195	432
75-79	cigarretteOnly	436	214
80+	cigarretteOnly	113	63

Tabell 1: Lungekreftdata.

```
glm(formula = dead ~ age + smoke, family = poisson("log"), data = data,
    offset = log(pop))

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         -3.63222    0.06783 -53.552  < 2e-16
age45-49             0.55388    0.07999   6.924 4.38e-12
age50-54             0.98039    0.07682  12.762  < 2e-16
age55-59             1.37946    0.06526  21.138  < 2e-16
age60-64             1.65423    0.06257  26.439  < 2e-16
age65-69             1.99817    0.06279  31.824  < 2e-16
age70-74             2.27141    0.06435  35.296  < 2e-16
age75-79             2.55858    0.06778  37.746  < 2e-16
age80+               2.84692    0.07242  39.310  < 2e-16
smokecigarretteOnly  0.36915    0.03791   9.737  < 2e-16
smokecigarrettePlus  0.17015    0.03643   4.671 3.00e-06
smokeno             -0.04781    0.04699  -1.017    0.309

Residual deviance:   21.487  on ??  degrees of freedom
```

a) For personer som er 53 år og ikke-røykere, hva blir det estimerte antall lungekrefttilfeller per 100 000 personer?

Hvor mange frihetsgrader har deviansen til den tilpassede modellen? Vil du si at modellen gir en god tilpasning til dataene? (Begrunn svarene)

To alternative modeller er også tilpasset det samme datasettet. I resten av oppgaveteksten skal vi kalle modellen diskutert over for modell 1, og de øvrige to for henholdsvis modell 2 og modell 3. I modell 2 benytter vi kun røykestatus som forklaringsvariabel og en (forkortet) versjon av R-utskrift for denne modellen er:

```
glm(formula = dead ~ smoke, family = poisson("log"), data = data,
    offset = log(pop))

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         -1.47319    0.03099 -47.532  < 2e-16
smokecigarretteOnly -0.31219    0.03576  -8.729  < 2e-16
smokecigarrettePlus -0.43013    0.03468 -12.402  < 2e-16
smokeno             -0.36678    0.04669  -7.855 3.98e-15

Residual deviance: 3910.7  on 32  degrees of freedom
```

I modell 3 benytter man røykestatus og aldersintervallnummer som forklaringsvariabler, der aldersintervallnummer er definert som 1 for aldersgruppen 40 til 44, som 2 for aldersgruppen 45 til 49 og så videre opp til 9 for aldersgruppen 80+. En forkortet versjon av R-utskrift for denne modellen er gitt under.

```
Call:
glm(formula = dead ~ ageLevel + smoke, family = poisson("log"),
    data = data, offset = log(pop))

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         -3.705950   0.050717 -73.071  < 2e-16
ageLevel             0.333006   0.005591  59.559  < 2e-16
smokecigarretteOnly  0.405019   0.037463  10.811  < 2e-16
smokecigarrettePlus  0.203426   0.035996   5.651 1.59e-08
smokeno             -0.032927   0.046894  -0.702    0.483

Residual deviance:   75.734  on 31  degrees of freedom
```

b) Hvilken av de tre modellene vil du velge som "den beste"? Begrunn svaret ved å formulere relevante hypoteser og utføre de tilhørende hypotesetestene.

Uansett hva din konklusjon ble i punktet over skal vi i resten av denne oppgaven igjen betrakte modell 1. La $\mu(a, s)$ betegne forventet antall lungekrefttilfeller per 100 000 personer i aldergruppe a og med røykestatus s, og for to ulike røykestatuser s_1 og s_2 definer videre

$$r(a, s_1, s_2) = \frac{\mu(a, s_1)}{\mu(a, s_2)}.$$

c) Forklar hvorfor $r(a, s_1, s_2)$ **ikke** varierer som funksjon av a i modell 1.

For s_1 lik "cigarPipeOnly" og s_2 lik "cigaretteOnly", finn estimert verdi for $r(a, s_1, s_2)$. Finn også et (tilnærmet) 90%-konfidensintervall for $r(a, s_1, s_2)$ i dette tilfellet. Vil du si at det er en signifikant forskjell i sannsynligheten for å få lungekreft avhengig av om en person røyker sigaretter eller sigar/pipe?

Oppgave 2

Anta at $Y_1, \ldots, Y_N$ er uavhengige kontinuerlig fordelte stokastiske variabler og at sannsynlighetstettheten til Y_i er gitt ved

$$f(y_i; \theta_i) = \begin{cases} \frac{\theta_i^2}{2} y_i e^{-\theta_i y_i} & \text{for } y_i \geq 0, \\ 0 & \text{ellers,} \end{cases}$$

der θ_i er en skalar parameter.

a) Vis at Y_i-ene har fordelinger som kommer fra samme eksponensielle familie.

Benytt generelle formler for en eksponensiell familie til å vise at

$$\mathrm{E}[Y_i] = \frac{2}{\theta_i} \quad \text{og} \quad \mathrm{Var}[Y_i] = \frac{2}{\theta_i^2}.$$

Anta en generalisert lineær modell for $Y_1, \ldots, Y_N$ der fordelingen for Y_i-ene er som spesifisert over og linkfunksjonen er gitt ved

$$\eta = g(\mu) = \ln(\mu) = x^T \beta,$$

der $x, \beta \in \mathbb{R}^p$.

b) Benytt generelle formler fra pensum til å bestemme skårvektoren $U(\beta) = [U_1(\beta), \ldots, U_p(\beta)]^T$ og informasjonsmatrisa $\mathcal{J}(\beta) = [\mathcal{J}_{ij}(\beta)]_{i,j=1}^p$ uttrykt ved $y_1, \ldots, y_N$, β, N, samt forklaringsvariablene.

Skriv opp rekursjonsligningen som kan benyttes til å finne sannsynlighetsmaksimeringsestimatoren for β.

c) Skriv opp log-likelihoodsfunksjonen for den spesifiserte modellen. Ta utgangspunkt i denne for å finne deviansen, D, for modellen uttrykt ved $y_1, \ldots, y_N$ og $\widehat{y}_1, \ldots, \widehat{y}_N$, hvor $\widehat{y}_i$ betegner estimert forventningsverdi for y_i.

Skriv også opp uttrykk for deviansresidualene, d_i, uttrykt ved y_i og $\widehat{y}_i$.

d) Finn pearson-residualene, r_i, uttrykt ved y_i og $\widehat{y}_i$.

La $X^2 = \sum_{i=1}^N r_i^2$ betegne Pearsons kjikvadratobservator. Vis at $D \approx X^2$.