Norges teknisk–naturvitenskapelige universitet
Institutt for matematiske fag

Kontakt under eksamen:
Ingelin Steinsland (92 66 30 96)

EKSAMEN I TMA4315 GENERALISERTE LINEÆRE MODELLER

Fredag 10. desember, 2010
Tid: 09:00 – 13:00

Tillatte hjelpemiddler:
Tabeller og formler i statistikk, Tapir Forlag
K. Rottmann: Matematisk formelsamling
Calculator HP30S / CITIZEN SR-270X
Gult, stempla A4-ark med egne handskrevne notat.

Sensur: 28. desember, 2010

Oppgave 1 Antall busspassasjerer

En busssjåfør vil modellere hvor mange passasjerer han får fra busstoppet rett ved studentbyen. Han kan tenke seg tre forklaringsvariable; hvilken rute (route) det er (8 am eller 9 am), om det er i løpet av semesteret eller ikke, og temperaturen. Han har data for 20 dagar, gitt i tabellen under. Han vurderer tre ulike modeller, alle analysert i `R` (se editert utskrift under): *modell 1* gir `result1`, *modell 2* gir `result2` og *modell 3* gir `result3`.

a) Sett opp den generaliserte lineære modellen (GLM) *modell 1* bruker matematisk, spesifiser antagelsene og design matrisa X for de første 6 observasjonene. Spesifiser også hvilken strategi som er brukt for å sikre identifiserbarhet, og diskuter kort alternativ(ene).
Forklar, matematisk og med ord, hvilken modell `R`-notasjonen `temp*semester` gir (brukt i *modell 2*).

	Passengers	route	semester	temp
1	3	8am	semester	8.8
2	1	9am	nonSemester	11.5
3	1	8am	nonSemester	12.0
4	3	8am	semester	14.8
5	0	8am	nonSemester	-1.2
6	0	8am	nonSemester	7.8
7	0	8am	nonSemester	6.9
8	1	9am	nonSemester	7.5
9	6	8am	semester	7.7
10	2	8am	semester	5.5
11	1	8am	nonSemester	13.7
12	1	8am	nonSemester	13.1
13	0	9am	nonSemester	14.2
14	2	9am	nonSemester	0.2
15	4	8am	nonSemester	-4.7
16	0	9am	nonSemester	26.3
17	3	9am	semester	3.1
18	2	8am	semester	-4.0
19	1	9am	nonSemester	18.4
20	2	8am	nonSemester	-5.0

b) Bruk nå *modell 1*. Basert på resultatene fra `R`:
Hva er forventet antall passasjerer på 9 am ruta, i semesteret når det er $5.4^{O}C$.
Hva er forventet antall passasjerer på 8 am ruta, utenfor semesteret når det er $-15.2^{O}C$.

c) Vi vil nå sammenlikne modeller: Sett opp en hypotese for å teste *modell 2* mot *modell 1* ved å bruke likelihood ratio testen (dvs basert på deviance), og utfør testen.
Hvilke av modellene *model 1*, *model 2* eller *model 3*, vil du foretrekke. Hvorfor?

d) La $Y_1, \ldots Y_N$ være uavhengige responser med $Y_i \sim Po(\lambda_i)$. For modellen av interesse, med $p < N$ parameter, la $\hat{y}_i$ være de tilpassede verdiene når man bruker maximum likelihood estimatene. Finn et uttrykk, basert på y_i og $\hat{y}_i$, for deviancen i dette tilfellet.

```
> result1 = glm(Passengers~temp+semester, family=poisson(link="log"))
> summary(result1)
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       0.25406    0.30667   0.828  0.40741
temp             -0.03451    0.02462  -1.401  0.16107
semestersemester  1.08499    0.35365   3.068  0.00216 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    Null deviance: 30.406  on 19  degrees of freedom
Residual deviance: 17.677  on 17  degrees of freedom
AIC: 62.03

> result2 = glm(Passengers~temp*semester, family=poisson(link="log"))
> summary(result2)
Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           0.44315    0.29124   1.522   0.1281
temp                 -0.07445    0.03384  -2.200   0.0278 *
semestersemester      0.54611    0.46383   1.177   0.2390
temp:semestersemester 0.10002    0.05316   1.881   0.0599 .

    Null deviance: 30.406  on 19  degrees of freedom
Residual deviance: 13.981  on 16  degrees of freedom
AIC: 60.334

> result3 = glm(Passengers~temp+semester+route, family=poisson(link="log"))
> summary(result3)
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       0.28227    0.32780   0.861  0.38918
temp             -0.03345    0.02501  -1.338  0.18095
semestersemester  1.06849    0.36035   2.965  0.00303 **
route9am         -0.09713    0.42224  -0.230  0.81806

   Null deviance: 30.406  on 19  degrees of freedom
Residual deviance: 17.623  on 16  degrees of freedom
AIC: 63.976
```

Oppgave 2 Negativ binomial fordeling

Sannsynlighetstettheten for en negativ binomisk fordelt stokastisk variabel er;

$$f_y(y;\theta,r) = \frac{\Gamma(y+r)}{y!\Gamma(r)}(1-\theta)^r\theta^y$$

for $y = 0, 1, 2, \ldots$, $r > 0$ og $\theta \in (0, 1)$. $\Gamma()$ står for gamma funksjonen. (Det er også andre parametriseringer for negativ binomial fordeling, men bruk denne nå.)

a) Vis at negativ binomial fordelinga er en medlem i den eksponensielle familien.
Du kan i dette punktet se på r som en kjent konstant.

b) Bruk generelle formler for eksponensiell familie for å vise at $E(Y) = \mu = r\frac{\theta}{1-\theta}$ og $Var(Y) = \mu\frac{1}{1-\theta}$

c) Sett opp en GLM for datasettet i oppgave 1 med negativ binomial responsfunksjon og samme lineærkomponent som i *modell 1.*
Begrunn ditt valg av link-funksjon.
Hvilken rolle har r?
I hvilken situasjon vil det kunne vere fordelaktig å bruke negativ binomial som responsfunksjon i stedet for Poisson-responsfunksjon?