Norges teknisk–naturvitenskapelige universitet
Institutt for matematiske fag

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

EKSAMEN I TMA4315 GENERALISERTE LINEÆRE MODELLER

Tirsdag 6. desember, 2011
Tid: 09:00 – 13:00

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

Sensur: 28. desember, 2011

Oppgave 1 Julegavepreferanser

Nissen vil modellere gavepreferanser; om barn fortrekker myke eller harde gaver. Han har to forklaringsvariable; alder (*age*, gitt i år) og kjønn (*sex*, som er *male* eller $female$). Han har data fra 100 barn, 10 av dem er gitt i Tabell 1. Nissen vurderer tre ulike modeller, alle analysert i `R` (se redigert utskrift under); *modell 1* gir `result1`, *modell 2* gir `result2` og *model 3* gir `result3`.

a) Sett opp den generaliserte lineære modellen (GLMen) brukt for *model 1* matematisk. Spesifiser design matrisa X for de første 6 observasjonene.
Videre presiser hvilken strategi som er brukt for å sikre identifiserbarhet, og diskuter kort alternativ(e) metode(r).

	pref	**sex**	**age**	**agegr**
1	1	female	7.5	3
2	0	male	10.5	3
3	0	female	2.0	1
4	1	female	8.0	3
5	0	male	2.0	1
6	1	female	4.0	2
7	1	female	6.0	2
8	1	male	10.0	3
9	1	female	8.0	3
10	0	female	3.0	1

Tabell 1: Data fra 10 barn i nissens datasett på julegavepreferanser: Om de foretrekker myke gaver ($pref = 0$) eller harde gaver ($pref = 1$). Også tilgjengelig; age (alder i år), sex (kjønn) and aldersgruppe $agrgr$ definert i Oppgave 1d)

b) Forklar matematisk og med ord de tre ulike modellene. Spesielt forklar hvilken modell `R` notasjonen `age*sex` gir.
Videre lag skisser som grafisk illustrerer forskjellene mellom de tre modellene.

```
> summary(result1)

Call:
glm(formula = pref ~ sex + age, family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5356  -1.1459   0.8824   1.0555   1.4217

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.69613    0.18356  -3.792 0.000149 ***
sexmale      0.78254    0.13010   6.015  1.8e-09 ***
age          0.06906    0.02529   2.731 0.006311 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    Null deviance: 1383.4  on 999  degrees of freedom
Residual deviance: 1338.0  on 997  degrees of freedom
AIC: 1344
```

```
> summary(result2)

Call:
glm(formula = pref ~ age, family = binomial(link = "logit"))

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.353  -1.196   1.026   1.129   1.267

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.35167    0.17067  -2.061  0.03935 *
age          0.07195    0.02483   2.898  0.00376 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    Null deviance: 1383.4  on 999  degrees of freedom
Residual deviance: 1374.9  on 998  degrees of freedom
AIC: 1378.9

> summary(result3)

Call:
glm(formula = pref ~ sex * age, family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5287  -1.1495   0.8866   1.0504   1.4280

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.716715   0.237113  -3.023  0.00251 **
sexmale      0.826983   0.348833   2.371  0.01775 *
age          0.072290   0.034541   2.093  0.03636 *
sexmale:age -0.006965   0.050707  -0.137  0.89076
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    Null deviance: 1383.4  on 999  degrees of freedom
Residual deviance: 1338.0  on 996  degrees of freedom
AIC: 1346
```

c) Bruk *modell 1*. Basert på resultatene fra `R`:
Hva er sannsynligheten for at ei 5 år gammel jente fortrekker myke gaver?
Hva er sannsynligheten for at ei 15 år gammel jente fortrekker myke gaver?
Hva er odds ratio mellom en 15 år gammel jente og 15 år gammel gutt for å foretrekke myke gaver?

Basert på evalueringen av resultatene for modell 1-3 og nissens erfaring, vil han også prøve en modell der alder blir sett på som en faktor med tre nivåer; `agegr1` (0-3 år), `agegr2` (3.5-7 år), `agegr3` (>7 år). Han tilpasser en modell, *modell 4* som gir `result4` (se under).

```
 > summary(result4)

Call:
glm(formula = pref ~ sex + as.factor(agegr), family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6569  -1.0654   0.7645   1.0962   1.9477

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -1.7341     0.2254  -7.694 1.43e-14 ***
sexmale             0.8865     0.1367   6.486 8.81e-11 ***
as.factor(agegr)2   1.9281     0.2338   8.247  < 2e-16 ***
as.factor(agegr)3   1.3020     0.2346   5.550 2.85e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    Null deviance: 1383.4  on 999  degrees of freedom
Residual deviance: 1261.8  on 996  degrees of freedom
AIC: 1269.8
```

d) Hvordan kan du sammenligne og evaluere modellene basert på resultatene gitt for modell 1-4?
Demonstrer hva du vil gjøre på en modell / et par av modeller.
Beskriv hvordan du kan gå fram for å evaluere modellene grafisk.
Tilslutt foreslå en modell du vil forsøke å tilpasse som neste modell. Begrunn valget ditt.

e) Nissen sin avdeling for kunderelasjoner har registrert at flere og flere barn ønsker seg elektroniske gaver. Hvordan kan din favoritt modell (av *modell 1-4*) bli utvidet for å ta hensyn til dette. Foreslå en modell, og diskuter valgene dine.

Oppgave 2 Antall julegaver

Nissen har en modell han trekker fra for å få antallet gaver for hvert barn. La oss anta at det er en Poisson modell.

a) Vis at Poisson fordelingen er medlem i den eksponensielle familien, og bruk dette til å finne forventningsverdi og varians for poissonfordelte stokastiske variable.

Vi har data på antall gaver (Y_i) og alder (x_i) for $n = 20$ barn. Anta at $Y_1, \ldots, Y_n$ er Poisson med forventning $\mu_i = \exp(\alpha + \beta x_i)$ der α og β er regresjonsparametre.

b) Sett opp log-likelihood funsjonen for dataene beskrevet over.
Vis at score funksjonen kan bli skrevet som

$$U(\alpha, \beta) = \begin{pmatrix} U_1(\alpha, \beta) \\ U_2(\alpha, \beta) \end{pmatrix} = \sum_{i=1}^{n} \begin{pmatrix} 1 \\ x_i \end{pmatrix} (Y_i - \mu_i)$$

Finn også et uttrykk for forventet informasjonsmatrise.

c) Hva er en mettet modell (saturated model)?
Vis at maximum likelihood estimatene for den mettede modellen er $\tilde{\mu}_i = Y_i$ for $i = 1, 2, \ldots, 20$

d) Basert på resulatene over, skriv et uttrykk for deviance for denne modellen.

Under er en editert utskrift fra `R` som analyserer disse dataene. Hva er det manglende tallet (degrees of freedom)?

Videre grei ut om hvordan deviance kan brukes for denne type modeller (generalisert lineær modell med Poisson likelihood)

```
> summary(resut5)

Call:
glm(formula = gifts ~ age, family = poisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8766  -0.9863  -0.4624   0.6776   1.7126

Coefficients:
```

```
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.8952     0.4291   4.417    1e-05 ***
age          -0.3323     0.0935  -3.555    0.000379 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


    Null deviance: 35.941  on 19  degrees of freedom
Residual deviance: 19.805  on ??  degrees of freedom
AIC: 53.599
```

Oppgave 3 Gyldige GLMer

Definer klassen av generaliserte lineære modeller (GLMer), og eksplisitt list opp alle krav for hver del av modellen.

Under er det listet opp tre likelihooder, tre link-funksjoner og tre lineær komponenter. Forklar hvilke kombinasjoner som gir gyldige GLMer, og kommenter spesielle trekk ved disse modellene. (Du trenger ikke matematisk å bevise hvilke modeller som er gyldige.)

Likelihooder:
1. Gaussisk; $Y \sim N(\mu, \sigma^2)$
2. Multinomisk; $Y \sim M([p_1, p_2, p_3, p_4], N)$
3. Poisson; $Y \sim Po(\theta)$

Link-funksjoner:
1. $\eta = \cos(\mu)$
2. $\eta = \mu$
3. $\eta = \log(\mu)$

Lineær komponenter:
1. $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
2. $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2$
3. $\eta = \beta_0 + \beta_1 x_1 + \beta_1^2 x_2$