Norges teknisk–naturvitenskapelige universitet
Institutt for matematiske fag

Bokmål

Kontakt under eksamen: Thiago G. Martins 46 93 74 29

EKSAMEN I TMA4315 GENERALISERTE LINEÆRE MODELLER

Torsdag 13. desember, 2012
Tid: 09:00 – 13:00

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

Sensur: 28. desember, 2011

Oppgave 1 Nedbør i Trondheim i morgen?

NTNU matematikkstudenten Konrad Kontrollfrik liker å være forberedt til neste dag, og han vil vite om det blir nedbør eller ikke i morgen tidlig. Han lager derfor statistiske modeller for nedbør forekomst, og vurderer tre forklaringsvariabler:

1. Mengde nedbør (i mm) ifølge værvarselet i morgen ($Fore$).
2. Ein binær variabel $ForeBin$ som indikerer om vervarselet seier nedbør ($ForeBin = 1$) eller ikke nedbør ($ForeBin = 0$).
3. En variabel OF som indikerer hvordan gårsdagens varsel og dagens observasjon stemmer;
   - $OF = 0$: ingen nedbør observert, og ingen nedbør i varsel.

- $OF = 1$:nedbør observert, men ingen nedbør i varsel
- $OF = 2$: ingen nedbør observert, men nedbør i varsel
- $OF = 3$: nedbør observert, og nedbør i varsel

Konrad har fått tak i data for 100 påfølgende dager. Data for dei første ti er gitt i Tabell 1.

Han bruker *R* for å tilpasse disse modellene (se redigert utskrift under); *model 1* gir `result1`, *model 2* gir `result2` og *model 3* gir `result3`.

```
summary(result1)

Call:
glm(formula = Occur ~ -1 + Fore + ForeBin + as.factor(OF),
family = binomial(link = "logit"),data = OccurData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1714  -0.6614  -0.5975   0.7493   2.2624

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
Fore             0.4609     0.2294   2.009 0.044494 *
ForeBin          0.8883     0.6559   1.354 0.175641
as.factor(OF)0  -1.6327     0.4349  -3.755 0.000174 ***
as.factor(OF)1  -1.1941     0.9690  -1.232 0.217818
as.factor(OF)2  -2.5144     0.6764  -3.717 0.000202 ***
as.factor(OF)3  -1.7609     0.5983  -2.943 0.003249 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 138.63  on 100  degrees of freedom
Residual deviance: 101.09  on  94  degrees of freedom
AIC: 113.09

> summary(result2)

Call:
```

```
glm(formula = Occur ~ Fore + ForeBin + OF, family = binomial(link = "logit"),
    data = OccurData)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.6438     0.4280  -3.840 0.000123 ***
Fore          0.5204     0.2224   2.340 0.019286 *
ForeBin       0.7344     0.6396   1.148 0.250832
OF           -0.1355     0.2026  -0.669 0.503589
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 125.37  on 99  degrees of freedom
Residual deviance: 103.19  on 96  degrees of freedom
AIC: 111.19

> summary(result3)

Call:
glm(formula = Occur ~ Fore, family = binomial(link = "logit"),
    data = OccurData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1246  -0.6984  -0.6146   0.7471   1.8759

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.5707     0.3152  -4.983 6.27e-07 ***
Fore          0.6683     0.1856   3.601 0.000317 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 125.37  on 99  degrees of freedom
Residual deviance: 104.81  on 98  degrees of freedom
AIC: 108.81
```

Occure	Fore	ForeBin	OF
1	3.0	1	3
0	0.5	1	3
0	0.0	0	2
0	0.0	0	0
0	0.0	0	0
0	0.0	0	0
0	0.7	1	0
0	0.4	0	2
0	0.0	0	0
1	0.4	0	0

Tabell 1: Data fra ti dager i Konrad sitt datasett om nedbør forekomst, ingen nedbør(Occure= 0) eller nedbør (Occure = 1). Også tilgjengelig; mengde nedbør frå værvarsel i mm, binær varsel og OF variabelen.

a) Sett opp matematisk de generaliserte lineære modellene (GLM) som er brukt. Spesifiser og diskuter antagelser. Skriv videre ut designmatrisa X for dei første 6 observasjonene for hver modell.
Når det er relevant, spesifiser hvilken strategi som er brukt for å sikre identifiserbarhet. Beskriv kort forskjellene mellom *model 1* og *model 2*.

b) Basert på resultata fra `R` svar på de følgende spørsmåla:
Ifølge *model 1*: Hva er sannsynligheten for nedbør dersom varselet er på 5mm og $OF = 0$?
Ifølge *model 2*:Hva er sannsynligheten for nedbør dersom varselet er på 5mm og $OF = 3$?
Ifølge *model 3*: Hva er odds ratio mellom en dag med varsel 0mm og ein dag med 5mm?

c) Konrad vil nå sammenligne modeller: Sett opp ei hypotese for å sammenligne modell 1 mot modell 3 ved å bruke likelihood ratio testen (dvs basert på deviance), og utfør testen. Hvilken modell vil du foretrekke, modell 1, modell 2 eller modell 3. Hvorfor?

d) Konrad er også interessert i nedbør forekomst på Trondheim Lufthavn Værnes og på Vassfjellet. Han har observasjoner og varsel også for disse stedene;
Location $\in \{$`Trondheim`, `Vaernes`, `Vassfjellet`$\}$ for de same 100 dagene.

Han tilpasser tre modeller ved å kjøre R kommandoene under:

```
res4 = glm(Occur~Fore, family=binomial(link="logit"),data=OccurDataTVV)
res5 = glm(Occur~Fore + Location, family=binomial(link="logit"),data=OccurDataTVV)
res6 = glm(Occur~Fore*Location, family=binomial(link="logit"),data=OccurDataTVV)
```

Forklar kort de tre modellene, og lag skisser for å illustrerer.

Se på modellen som er brukt for å få `res4`, og diskuter om antagelsene for GLMer er oppfylt nå. Foreslå alternativ modell.

Oppgave 2 Nedbør i Trondheim som snø, sludd eller regn?
Gitt at det blir nedbør i morgen, så er Konrad interessert i hvilken type nedbør det blir. Vi kaller denne størrelsen C, og klassifiserer nedbør i tre klasser; snø ($C = 1$), sludd ($C = 2$) og regn ($C = 3$). Videre vil vi undersøke om temperaturvarselet er en god forklaringsvariabel.

a) La C_i være nedbørsklassen for dag i, og la t_i være temperaturvarselet som er gyldig for dag i.
Foreslå en passende modell for C. Spesifiser modellen matematisk, og forklar den med ord/figur(er). Spesielt forklar hvordan parametrene skal bli tolket.

Oppgave 3 Nedbør i Trondheim, mengde
Vi vil nå modellere mengda av nedbør i løpet av et døgn, gitt at det er nedbør, og kaller denne størrelsen Y. Det er vanlig å modellere Y som gammafordelt $Y \sim Ga(\alpha, \beta)$, med tetthet;

$$f_Y(y) = \frac{1}{\beta^\alpha \Gamma(\alpha)} y^{\alpha-1} \exp(-y/\beta).$$

I denne oppgava skal vi anta det er N observasjoner, hver gammafordelt $Y_i \sim Ga(\alpha_i, \mu_i/\alpha_i)$. Her er $\alpha_i$ene antatt kjent, det vil si at dei er nuisance parametre.

a) Vis at gammafordelinga er medlem i den eksponensielle familien når μ_i er parameteren av interesse.
Bruk dette til å finne uttrykk for forventningverdien og variansen til Y_i, som funksjon av (α_i, μ_i). Fra dette tolk α_i.

b) Forklar hva en mettet (saturated) modell er.
Sett opp log-likelihood funksjonen uttrykt med μ_i, og bruk dette til å finne maximum likelihood estimatorene for μ_i-ene i den mettede modellen.
Finn deviancen (basert på alle N observasjonene).

c) Vi vil nå lage en modell for mengde nedbør (gitt at det er nedbør) med nedbørvarsel som forklaringsvariabel.

La Y_i være mengde nedbør på dag i, og la x_i være nedbørsvarselet som er gyldig for dag i.
Sett opp en GLM for dette, og begrunn valget ditt for link-funksjon og lineærkomponent.