IMM, 2004 -02-27
Poul Thyregod

1 Opgave om devianser

Betragt observationerne hhv $y_1 = 4$ og $y_2 = 14$

For hver af disse to observationer skal du tegne deviansen

$$\lambda d(y, \mu)$$

hvor $d(y, \mu)$ er enhedsdeviansen og λ er præcisionsparameteren (evt. $\lambda = 1/\sigma^2$).

Deviansen tegnes som funktion af μ for hver af nedenstående fordelingsantagelser

a) $\mathrm{N}(\mu, \sigma^2)$, $\sigma^2 = 4$

b) Gamma-fordeling, $\mathrm{G}(\alpha, \mu/\alpha)$ for $\alpha = 1$

c) Den inverse Gauss-fordeling, $\mathrm{IG}(\mu, \lambda)$ for $\lambda = 1$

d) Poissonfordelingen, $\mathrm{P}(\mu)$

Angiv endvidere variansen i fordelingen af Y svarende til $\mu = 4$ og $\mu = 14$ for de fire fordelinger.

2 Opgave Kundetilfredshed

2.1 Data

Datasaettet `bustilfr` indeholder et sammendrag af svarene fra 12 233 kvindelige buspassagerer, der blev spurgt om deres tilfredshed med bussens overholdelse af køreplanen. De mulige svarkategorier var hhv `‘‘Meget utilfreds’’`, `‘‘utilfreds’’`, `‘‘Både og’’`, `‘‘Tilfreds’’ og ‘‘Meget tilfreds’’`. Samtidigt med indsamlingen af svarene registreredes bussens eventuelle forsinkelse,

hhv 0, 2, 5 og 7 minutter.

Datasættet indeholder for hver af disse fire forsinkelsestider `fors`, den variable `Utilfr`, der angiver antallet af svar i kategorien "Meget utilfreds" eller "Utilfreds", endvidere indeholder den variable `Ialt` det totale antal respondenter. Desuden er der beregnet de variable `andel` og `logit`, der angiver hhv. andelen af utilfredse ("Utilfreds" eller "Meget Utilfreds") og den empiriske logit svarende til denne andel, og endelig indeholder den variable `vegt` vægten $ialt \times andel \times (1 - andel)$ til brug for spørgsmål 2.

2.2 Opgave

- Udfør en logistisk regression med henblik på at beskrive en eventuel sammenhæng mellem utilfredshed og forsinkelse.
- Prøv også at udføre en almindelig regressionsanalyse (normalfordeling) af de empiriske logit'er og sammenlign estimaterne under denne model med estimaterne fra den logistiske regression
  (Prøv eventuelt en vægtet regression, jvf nedenstående vejledning)
- Bestem den lineære prædiktor svarende til en forsinkelse på 4 minutter, og bestem usikkerheden (variansen) på denne lineære prædiktor,
- Bestem estimatet for andelen af utilfredse svarende til en forsinkelse på 4 minutter
  og bestem eventuelt variansen på denne (brug fejlophobningsloven)

2.3 Vejledning til udførelse af vægtet regression

Baggrund:
Hvis den underliggende responssandsynlighed er p, er variansen på den observerede andel, Y,

$$\mathsf{V}[Y] = p(1-p)/n$$

Men nu ser vi jo på logit'en,

$$g(Y) = \ln(Y/(1-Y))$$

og så bruger vi fejlophobningsloven :

$$\mathsf{V}[g(Y)] \approx (g'(p))^2 \mathsf{V}[Y]$$

idet $\mathsf{E}[Y] = p$.
Men $g(p) = \ln(p/(1-p))$, hvorfor

$$g'(p) = \frac{1}{p} - \frac{1}{1-p} = \frac{1}{p(1-p)}$$

og

$$\mathsf{V}[g(Y)] \approx \left(\frac{1}{p(1-p)}\right)^2 \times \frac{p(1-p)}{n} = \frac{1}{np(1-p)}$$

så hvis vi vil lave en vægtet (mindste kvadraters) regressionsanalyse, hvor vi vægter de observerede logit'er med estimeret præcision, kan vi altså bruge vægten

$$n \times andel \times (1 - andel)$$

, og så er der næsten ikke forskel på parameterestimaterne, men det er nu alligevel mere tilfredsstillende at bruge den generaliserede lineaere model med alle dens krummelurer.

3 Opgave Fødsler

3.1 Data

Datasættet `fodsl` indeholder for hvert af årene 1963-1976 to observationer, med de variable `aar, STA fodsl dod, STATUS, andel, myaar`, der angiver hhv årstallet, moderens ægteskabelige status (gift ugift), antallet af fødsler for denne gruppe det pågædende år, antallet (blandt disse), der resulterede i et dødt barn, moderens status (numerisk kodet som nul eller een), andelen af dødfødte, år efter 1963.

3.2 Opgave

Foretag en analyse med henblik på at beskrive udviklingen i andelen af dødfødte børn i de to statusgrupper den betragtede periode og sammenlign udviklingen i de to grupper.

Dvs kør en generaliseret lineær model (binomialfordeling, logit) af andel dødfødte mod år (myaar) i lighed med bilpris-undersøgelsen.

En eventuel fortolkning af analysen skal ses i lyset af periodens samfundsnormer - og absolut ikke i dagens normer. Hvis man ser på udviklingen i antal fødsler i de to grupper ser man jo tydeligt ændringen i "samfundets" holdning til ugifte mødre, så en rigtig relevant model ville jo lade de to kurver nærme sig hinanden - men her bruger vi bare disse data som en øvelse i brug af SAS.