Andre obligatoriske oppgave i STK3100 Høst 2010

Utlevering: Tirsdag 19. oktober

Innleveringsfrist: Torsdag 4.november, kl. 14:30

Besvarelsen leveres i kassa for obligatoriske oppgaver i gangen i 7. etasje, Niels Henrik Abels hus.

Dette er det andre settet med obligatoriske innlevering i STK31000 høsten 2010. Oppgavesettet består av en oppgave. Det er valgfritt om du vil skrive besvarelsen for hånd eller om du vil bruke et tekstbehandlingsprogram. Der du bruker R (eller et annet program), må utskrifter legges ved eller limes inn. Hvis flere studenter samarbeider om å løse oppgavene, må likevel hver student levere sin selvstendige besvarelse. Det må gå fram av besvarelsen hvem du har samarbeidet med. Se ellers "Regelverk for obligatoriske oppgaver" som er gitt på kursets hjemmeside.

Oppgave 1

I denne oppgaven skal vi analysere noen data fra Baxter et al. (1980) *Transactions 21 Congress of Actuaries* **2-3**, 11-29 om skadetilfeller i en portefølje av forsikrede privatbiler i et middels stort engelsk forsikringsselskap tredje kvartal 1973. Det er registrert antall skadetilfeller oppdelt etter tre tarifferingsfaktorer, hver med fire nivåer. Tarifferingsfaktorene er kodet som følger:

- Forsikringstagers alder:
  1 = under 25 år.
  2 = 25-29 år.
  3 = 30-35 år.
  4 = over 35 år.
- Bilens motorvolum:
  1 = under 1 liter.
  2 = 1-1,5 liter.
  3 = 1,5-2 liter.
  4 = over 2 liter.
- Distrikt:
  4 = London og andre store byer.
  1–3 = andre distrikt.

Dataene ligger på filen `claims` på hjemmesiden til kurset. Første søyle i filen angir alder, andre søyle bilens motorvolum, tredje søyle distrikt, fjerde søyle antall forsikrede i gruppen og femte søyle antall skader.

a) Anta at for hver forsikringstaker inntrer skader med en rate (som kan avhenge av forsikringstakers alder og bosted/distrikt og bilens motorvolum). Hvorfor er det rimelig å tenke seg at antall skader angitt i

datafilen er Poissonfordelt med forventning proporsjonalt med antall forsikringstakere.

b) Vis at kanonisk link for Poissondata er $g(\mu) = \log(\mu)$ (der μ er forventningen i en Poissonfordeling). Vi skal videre i oppgaven, bortsett fra i punktene k) og l), bare bruke denne linkfunksjonen.

c) For de foreliggende dataene er altså forventet antall skader proporsjonalt med antall forsikringstakere. Forklar hvorfor dette antallet inngår i lineær prediktor som en 'offset'.

d) Diskuter om de tre kovariatene bør modelleres som kategoriske forklaringsvariable (faktorer) eller numeriske forklaringsvariable. Diskuter fordeler og ulemper ved begge tilnærminger.

e) Sett opp en deviansanalysetabell for dataene der alder, motorvolum og distrikt er modelleret som faktorer. Forklar hvorfor modellen med alle hovedeffekter, men ingen interaksjoner, er adekvat for dataene.

f) Utfør en analyse av dataene som klargjør betydningen av alder, motorvolum og distrikt modelleret som faktorer.

g) Forklar hvorfor parametrene i modellen har fortolkning som logaritmen til rateratioer. Estimer rateratioene og beregn tilhørende konfidensintervall.

h) Undersøk om det er en lineær trend i alder og motorvolum ved å tilpasse modeller der disse tarifferingsfaktorene tas med som kvantitative kovariater ('variates' i GLM terminologi). Er det mulig å foreta en forenkling i modelleringen av effekten av distrikt? (Dvs. kan denne modelleres med færre parametere?)

i) Foreta en analyse av residualene i den 'endelige modellen' du har kommet fram til over. Er det noe ved residualene som tyder på at denne ikke er tilfredsstillende?

j) Estimer raten for skadetilfeller i 3. kvartal for en forsikret i alder 25-29 år med bil med motorvolum 1,5-2 liter bosatt i London. Finn også 95% konfidensintervall for denne raten.

k) Undersøk om det er overspredning (evt. underspredning) i forhold til Poissonfordeling i dette datasettet.

l) Vi har sålangt benyttet log-link der logaritmen til antall forsikringtakere inngår som en offset. For andre linkfunksjoner må en da gå litt annerledes til verks. Først vis at når $Y \sim \text{Po}(n\lambda)$ blir

$$\text{E}[Y/n] = \lambda \quad \text{og} \quad \text{Var}[Y/n] = \frac{\lambda}{n}.$$

Lag nå en avledet responsvariabel lik antall skader delt på antall forsikringstakere. Tilpass så en Poissonregresjonsmodell med denne responsvariablen hvor du også legger inn antall forsikringstakere som vekter: `weights=antforsikrede`. Sammenlign estimater og devians med analysen der dette antallet ble lagt inn ved 'offset'. Sammenlign også med det du får ved å bruke opsjonen `family=quasi(link="log",variance="mu")` i stedet for `family=poisson(link="log")`

m) Forklar hvorfor loglinken kan betraktes som et grensetilfelle av 'powerlinkene' $g_\rho(\mu) = \mu^\rho$ når $\rho \to 0$. Tilpass modeller for et par ulike ρ (bruk `family=quasi(link="power(rho)",variance="mu")` for passende valg av `rho`). Sammenlign de estimerte koeffesientene.