Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

ST2304 Statistisk modellering for biologer og bioteknologer

Løsningsforslag eksamen 24. mai, 2013

Oppgave 1

a) `qf(c(.025,.975),df1=9,df2=19)`

b) `2*pf(13.5/5.2,df1=9,df2=19,lower.tail=FALSE)`

c) Siden F er kontinuerlig fordelt er $P(F = 1) = 0$. Sannsynlighetstettheten i $F = 1$ kunne vært beregnet med uttrykket

```
df(1,df1=9,df2=19)
```

men dette er det ikke spurt om.

d) `hist(rf(1000,df1=9,df2=19))`

Oppgave 2

a) Modellen vi har tilpasset antar at sammenhengen mellom responsvariabelen y og x er gitt ved

$$y = b_0 + b_1 x + b_2 x^2 + e \tag{1}$$

hvor e er normalfordelt med forventning 0 og varians σ^2. Videre antas det at de ulike observerte y'ene er uavhengige stokastiske variable. Siden p-verdiene for både x og x^2 (lik henholdsvis $8.95 \cdot 10^{-7}$ og $9.41 \cdot 10^{-6}$) er mindre enn 0.05 har begge variablene en statistisk signifikant effekt på y.

b) For negative b_2 beskriver ligningen over en konkav kurve som har sitt maksimum der

$$\frac{dy}{dx} = b_1 + 2b_2 x = 0, \tag{2}$$

med andre ord i $x_0 = -\frac{b_1}{2b_2}$.

Sannsynlighetsmaksimeringsestimatet av x_0 blir $\hat{x}_0 = -\hat{b}_1/(2\hat{b}_2) = 0.8160/(2 \cdot 0.03135) = 12.94$ som virker rimelig ut i fra vedlagt figur av de observerte dataene.

c) De to partiellderiverte blir

$$\frac{\partial f}{\partial b_1} = -\frac{1}{2b_2} \tag{3}$$

og

$$\frac{\partial f}{\partial b_1} = \frac{b_1}{2b_2^2}. \tag{4}$$

I $b_1 = \hat{b}_1$ og $b_2 = \hat{b}_2$ har disse verdi 15.94 og 415.12. I følge deltametoden er så variansen til $\hat{x}_0$ tilnærmet

$$\begin{aligned} \operatorname{Var} \hat{x}_0 &\approx \left(\frac{\partial f}{\partial b_1}\right)^2 \operatorname{Var} \hat{b}_1 + \left(\frac{\partial f}{\partial b_1}\right)^2 \operatorname{Var} \hat{b}_2 + 2\left(\frac{\partial f}{\partial b_1}\right)\left(\frac{\partial f}{\partial b_1}\right) \operatorname{Cov}(\hat{b}_1, \hat{b}_2) \\ &= 15.94^2 0.01189 + 415.12^2 0.000025 + 2 \cdot 15.94 \cdot 415.12 \cdot (-0.000534) \\ &= 0.2619, \end{aligned} \tag{5}$$

og standardfeilen $\mathrm{SE}(\hat{x}_0) = 0.51$. Dette virker igjen rimelig ut i fra de observerte dataene.

Oppgave 3

a) Dersom ulike personer passerer observasjonspunktet med en viss rate λ og antall personer som passerer i ulike disjunkte delintervall er uavhengige variable, og ikke mer enn én person kan passere på et gitt tidspunkt, vil passeringene være beskrevet ved en Poissonprosess og antall passeringer i løpet av en dag vil være Poissonfordelt. Hvis raten λ varierer med tid på dagen, har vi en inhomogen Poissonprosess hvilket også vil føre til at det totale antallet som passerer blir Poissonfordelt.

På en helgedag øker forventet antall passerende med en faktor lik $e^{0.69} = 1.95$ eller med andre ord, med 95%, i forhold til på en hverdag.

Forventet antall passerende på en helgedag med 5mm nedbør blir $\exp(2.12 + 0.69 - 0.08 \cdot 5) = 10.69$.

b) Om vi setter parameterene som representer effekten av henholdsvis hver hverdag og hver helgedag like hverandre ser vi at modell A er et spesialtilfelle av og dermed nøstet i modell B. Antall estimerte parametere for modell A og B er henholdsvis $p_0 = 3$ og $p_1 = 8$. For nøstede generaliserte lineære modeller gjelder det at endring i devians under H_0 (modell A) er kji-kvadratfordelt med $p_1 - p_0$ (her 5) frihetsgrader. Observert verdi av denne testobservatoren blir $108.11 - 106.33 = 1.78$ som er mindre enn kritisk verdi lik 11.07 (øvre 0.05-kvantil i en kji-kvadratfordeling med 5 frihetsgrader). Vi velger derfor H_0 (modell A) som beste modell.

c) Den observert deviansen under modell A lik 108.11 er større enn kritisk verdi lik 76.77 og vi kan derfor konkludere med at det er overdispersjon i dataene.

Mekansimer som kan generere overdispersjon kan i dette tilfelle være at viktige forklaringsvariable som hvorvidt det er fellesferie eller ikke (inntreffer i juli), temperatur, skydekke o.s.v. Videre kan ulike person være i følge (en form for klumping) slik at antakelsene for Poissonprosessen ikke er oppfyllt og variansen til responsvariabelen blir større en Poissonantakelsen tilsier.