NTNU
Kunnskap for en bedre verden

ST2304 Statistisk modellering for biologer og bioteknologer

Løsningsforslag – Eksamen 21. mai 2014

NB: For noen oppgaver kan det være flere måter å løse oppgaven på.

Oppgave 1:

a) `qchisq(p=c(0.025,0.975),df=8)`

Median: `qchisq(p=0.5,df=8)`

b) Kji-kvadrattestens signifikanssannsynlighet:

```
pchisq(q=14.3,df=8,lower.tail=F)
```

Testobservatoren er kontinuerlig fordelt, slik at sannsynligheten for å få verdi eksakt lik 8 er null.

c) Plott:

```
curve(dchisq(x,df=8),0,20)
```

Oppgave 2:

a) Antallet individer som har hatt hjerteattakk er binomisk fordelt med sannsynlighet p. Den matematiske notasjonen for modell `mod.1` blir:

$$\log it(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 CK$$

Vi forutsetter at vi har en binomisk fordelt respons med uavhengighet mellom pasientene og en konstant sannsynlighet p for gitt `CK` nivå. Videre forutsetter vi at sammenhengen mellom p og `CK` kan beskrives ved en enkel lineær modell, og at vi ikke har noe over- eller underdispersjon.

Andel pasienter med CK=100 som forventes å ha hatt et hjerteattakk:

$$\eta = \log it(p) = -2.758358 + 0.031244 \times 100 = 0.366042$$

$$p = \frac{1}{1+e^{-\eta}} = \frac{1}{1+e^{-0.366042}} = 0.5905$$

b) Under nullhypotesen om ingen overdispersjon er residual deviance tilnærmet kji-kvadratfordelt med 10 frihetsgrader. Beregnet residual deviance er 36.929 noe som er mye større enn forventningsverdien på 10, så det er absolutt tegn til overdispersjon i dataene. Signifikanssannsynlighet (p-verdi) for testen er

```
pchisq(36.929,df=10,lower=F)= 5.82e-05
```

Mulige mekanismer for overdispersjon kan være at andre viktige faktorer som påvirker sannsynligheten er utelatt fra modellen, eller at den funksjonelle sammenhengen mellom logit(p) og CK har en annen form enn den enkle lineære. Det er nok lite sannsynlig at det er avhengigheter mellom pasienter mht. sannsynligheten for å ha hatt hjerteattakk.

c) Foretrekker mod.2 siden vi nå har mye lavere AIC, og den lavere residual deviance tyder på at vi ikke lenger har overdispersjon. Vi kan også se fra figurene at mod.2 passer bedre til observasjonene, spesielt for første og tredje punkt (fra venstre). De to modellene kan også testes mot hverandre ved anova(mod.1,mod.2,test='Chisq').

Residualanalyse: En måte å vurdere residualene til modellen på er å plotte residualene mot tilpassede (fitted / predicted) verdier for å se om det er noen trender eller mønster. Hvis dere skriver plot(mod.2) får dere opp flere residualplott for en modell, men dere har ikke lært å tolke alle disse.

Oppgave 3

a) Ja, både supp og dose har signifikant effekt på tannlengde ($\Pr(>F) < 0.05$).

Nei, det er ikke mulig å lese ut hvor stor effekt en faktor har på responsen fra en ANOVA tabell. Variansanalysen tester kun om det er forskjell i gjennomsnitt mellom grupper (kombinasjoner av faktornivå) og ikke hvor stor denne forskjellen kan være.

Antall observasjoner i datasettet kan leses ut fra antall frihetsgrader i modellen. For residualene er antall frihetsgrader (56) lik totalt antall frihetsgrader ($n-1$) minus antall frihetsgrader brukt for å estimere forskjeller mellom supp verdier (1) og dose nivåer (2): $n-1-1-2=56 \rightarrow n=60$

b) Intercept i modellen gir forventet respons for dose=low og supp=OJ, slik at forventet tannlengde for lav dose, gitt som appelsinjuice, er 12,455.

suppVC gir forventet endring fra nivået beskrevet i intercept (suppOJ – vitamin C gitt som appelsinjuice) til vitamin C gitt som askorbinsyre. Så et marsvin som får appelsinjuice forventes å ha 3,7 lengre tenner enn et som får askorbinsyre.

c) Se s. 1 i Handout 1 (lineære modeller) for mer detaljer. Nullhypotesen er at vi har en modell uten forklaringsvariable, dvs. responsen beskrives kun av gjennomsnittet, mens alternativ hypotese er at responsen beskrives bedre av den tilpassede modellen. Testobservatoren (se likning 6 i H1) er da F-fordelt med (p-1) og (n-p) frihetsgrader, hvor p=4 er antall parametere i modellen og n=60 er antall observasjoner. Konklusjonen for testen blir at nullhypotesen forkastes (p<2.2e-16), dvs. en stor del av den totale variasjonen i tannlengde er forklart av modellen.