Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

Statistisk modellering for biologer og bioteknologer, ST2304

12. august, 2013

Løsningsforslag

Oppgave 1

a) Antall frihetsgrader i en to-utvalgs t-test er $n_1 + n_2 - 2$, i dette tilfelle $10 + 13 - 2 = 21$. Det må strengt tatt forutsette at testobservatoren defineres slik at vi forkaster for store verdier av T. Testens kritiske verdi kan beregnes ved følgende uttrykk i R.

```
qt(0.05,df=21,lower.tail=FALSE)
```

b) Gitt at testobservatoren tar verdien 2.31 kan sannsynligheten for at testobservatoren tar samme eller en mer ekstrem verdi (testens p-verdi) når testen er ensidig beregnes ved uttrykket

```
pt(2.31,df=21,lower.tail=FALSE)
```

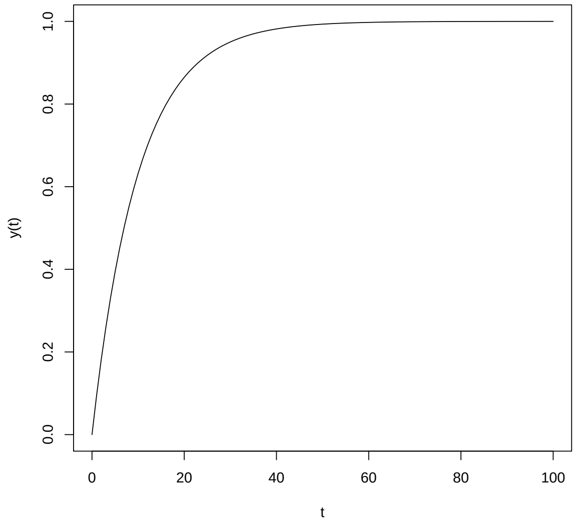
c) Testobservatoren er kontinuerlig fordelt. Sannsynlighetstettheten i $t = 0$ kan beregnes ved uttrykket

```
dt(0,df=21)
```

Oppgave 2

a) Tilnærmede standardfeil til parameterestimatene finner vi ved å ta kvadratroten av elementene langs diagonalen til den inverse Hessiske matrisa. SMEer av ukjente parametere i modellen og tilhørende standardfeil blir $\hat{k}_1 = 0.1026 \pm 0.0028$, $\hat{k}_2 = 0.000954 \pm 0.000092$ og $\sigma = 0.044 \pm 0.0030$.

b) Hvis $k_2 = 0$ blir $y(t) = x_0(1 - e^{-k_1 t})$ som går mot x_0 når t går mot uendelig.



Siden denne enklere modellen fremkommer ved å sette parameteren $k_2 = 0$ i modellen beskrevet ved ligning (1) er modellen nøstet i denne mer kompliserte modellen.

c)
```
lnL0 <- function(p,y,t) {
  k1 <- p[1]
  sigma <- p[2]
  -sum(dnorm(y,mean=1-exp(-k2*t),sd=sigma,log=T))
}
fit0 <- optim(c(0.1,.1),lnL0,t=t,y=y,lower=c(-Inf,0),hessian=TRUE)
```

d) La H_0 betegne den forenklede modellen ($k_2 = 0$) og H_1 modellen beskrevet ved ligning (1) ($k_2 \neq 0$). Under H_0 er da $2(\ln L_1 - \ln L_0)$ kji-kvadrat med $p_1 - p_0$, her $3 - 2 = 1$ frihetsgrader. Vi forkaster hvis denne testobservatoren er større enn kritisk verdi lik øvre 0.05-kvantil i kji-kvadratfordelingen lik 3.84. De maksimale (negative) log-likelihoodene finner vi fra `value`-komponenten i utskriften. Vi får dermed

$$2(\ln L_1 - \ln L_0) = 2(168.91 - 135.11) = 67.6 \tag{1}$$

og vi kan dermed klart forkaste H_0 til fordel for H_1. Dette virker rimelig ut i fra de observerte dataene hvor vi ser en viss tendens i nedgang i konsentrasjonen av B konsistent med H_1. Hadde H_0 vært riktig skulle B gått mot $x_0 = 1$ (en kjent parameter) noe som langt fra ser ut til å være tilfelle. Beste modell er dermed H_1 og estimatet av k_1 under denne modellen er å foretrekke.

Oppgave 3

a) Det er fornuftig å tenke seg at individene vi finner av den aktuelle arten innenfor hver rute er forekomster i en romlig Poissonprosess med en viss rate λ bestemt av de ulike forklaringsvariablene. Antall forekomster innenfor hver rute (responsvariabalen i modellen) blir da Poissonfordelt med parameter λA.

For å sikre at λ kun tar meningsfulle positive verdier bruker vi log som link-funksjon.

Numeriske forklaringsvariable er breddegrad (bredde), høyde over havet (høyde). Terrengorientering (terreng, nord eller syd) vil være en kategorisk variabel. Dersom vi forventer en lineær trend over ulike år vil det kunne være fornuftig å inkludere år også som numerisk variabel og eventuellt forkaste en slik modell mot en utvidet modell hvor år er inkludert som kategorisk variabel (altså at tettheten av individ varierer vilkårlig mellom ulike år.

Fordi vi a priori vet at det er direkte proposjonalitet mellom areal og forventet antall individ i hver rute bør log(areal) inkluderes som offset-variabel i modellen.

b) I matematisk notasjon er modellen som følger. Antall individ y i hver rute $i = 1, 2, \ldots, n$ er Poissonfordelt med forventning μ hvor

$$\log(\mu) = \beta_0 + \beta_B \text{bredde} + \beta_H \text{høyde} + \beta_{\text{terreng}} + \beta_{\text{år}} + \log(\text{areal}). \qquad (2)$$

c) Klumpvis fordeling av ulike individ vil kunne føre til overdispersjon. På den annen side vil konkurranse mellom ulike individ kunne føre til underdispersjon.