Bioberegninger - notat 3: Anvendelser av Newton's metode

20. februar 2004

1 Euler-Lotka ligningen

La oss tenke oss en populasjon bestående av individer av ulik alder. La n være maksimal alder. La m_i være antall avkom en hunn produserer på alderstrinn i (fekunditeten) og la l_i være overlevelse opp til alder i. Anta at disse parameterne ikke endrer seg over tid. Dette vil kunne være tilfelle så lenge den totale populasjonsstørrelsen er liten slik at det ikke er konkurranse om plass og andre begrensede ressurser. Selv om en slik antakelse kanskje ikke er realistisk vil det kunne være interessant å undersøke implikasjonen av en disse antakelsene. Et spørsmål er hvor raskt en slik populasjon vil vokse.

La $N_{t,0}$ betegne antall nyfødte i år t. Disse nyfødte er avkom av hunner født i tidligere år — av antall hunner født i år tilbake, $N_{t-i,0}$, vil en andel l_i være i live ved tidspunkt t. Hver av disse vil bidra med m_i avkom slik at det totale bidraget fra årsklassen født i år tidligere blir $N_{t-i,0}l_im_i$ avkom. Totalt antall avkom født i år t kan dermed skrives som

$$N_{t,0} = \sum_{i=1}^{n} N_{t-i,0}l_im_i. \tag{1}$$

Hvordan vil $N_{t,0}$ endre seg over tid? Det virker rimelig å anta at $N_{t,0}$ etter hvert vil vokse eller avta eksponentielt med t, altså at en mulig løsning vil kunne være på formen

$$N_{t,0} = Ke^{rt}. \tag{2}$$

Setter vi (2) inn i (1) får vi at

$$Ke^{rt} = \sum_{i=1}^{n} Ke^{r(t-i)}l_im_i. \tag{3}$$

Deler vi begge sider med Ke^{rt} får vi

$$1 = \sum_{i=1}^{n} e^{-ri} l_i m_i. \tag{4}$$

som er den såkalte Euler-Lotka ligningen. Hvis det eksisterer en r som er løsning av (4) betyr det at en løsning på formen (2) vil passe i (1), altså at populasjonen vil kunne vokse eksponentielt med vekstrate r.

Bare når $n \leq 2$ kan (4) løses ved vanlige metoder — for $n > 2$ ser vi at (4) er en n'te grads ligning i e^r. Denne må generelt løses numerisk, f.eks. ved bruk av Newton's metode. Vi må da skrive om (4) slik at den er på formen

$$f(r) = 0, \tag{5}$$

finne $f'(r)$, og så bruke iterasjonsligningen

$$r' = r - \frac{f(r)}{f'(r)}, \tag{6}$$

for å regne oss fram til løsningen.

Euler-Lotka ligningen og alderstrukturerte modeller generelt er nyttige i mange sammenhenger, f.eks. kan slike modeller brukes i befolkningsframskrivninger for å predikere fremtidige befolkningsstørrelser under ulike scenarier. Innen evolusjonsbiologi er slike modeller viktige for å forstå evolusjon av ulike livshistoriestrategier — slike modeller gir oss da innsikt i hva som er årsaken til fenomen som aldring, hva som påvirker alder for kjønnsmodning o.l.

Alderstrukturerte modeller kan håndteres mer generelt ved bruk av metoder fra lineær algebra. Dette vil vi se nærmere på senere i kurset.

2 Sannsynlighetsmaksimering og Newton's metode

La oss tenke oss en levetidsmodell med dødsrate (intensitet)

$$\lambda(t) = \frac{1}{t + a}, \tag{7}$$

hvor parameter a er en positiv konstant. Dødsraten, altså sannsynligheten for at et individ dør i et lite tidsintervall $t, t + \Delta t$, gitt at det er i live ved tidspunkt t avtar altså med tiden t og går mot null når t går mot uendelig.

Generelt (se notat om levetidsfordelinger fra brukerkurset i sannsynlighetsregning) er kumulativ fordeling til T uttrykt ved dødsraten gitt ved

$$F_T(t) = P(T \leq t) = 1 - e^{-\int_0^t \lambda(u)du}. \tag{8}$$

For modell (7) får vi etter litt regning at dette blir

$$F_T(t) = 1 - \frac{a}{t+a}. \tag{9}$$

Deriverer vi får vi at tettheten til T blir

$$f_T(t) = \frac{a}{(t+a)^2}. \tag{10}$$

La oss tenke oss at $t_1, t_2, \ldots, t_n$ er observerte levetider fra modellen over og at vi ønsker å estimere parameteren a ved bruk av sannsynlighetsmaksimering. Vi trenger da likelihoodfunksjonen, altså sannsynligheten for dataene gitt a, som blir

$$L(a) = \prod_{i=1}^{n} \frac{a}{(t_i+a)^2}, \tag{11}$$

og log-likelihoodfunksjonen

$$\ln L(a) = n \ln a - 2 \sum_{i=1}^{n} \ln(t_i + a)^2. \tag{12}$$

I likelihoodfunksjonens maksimum er

$$\frac{\partial}{\partial a} \ln L(a) = 0, \tag{13}$$

slik at

$$\frac{n}{a} - 2 \sum_{i=1}^{n} \frac{1}{t_i + a} = 0. \tag{14}$$

Løsningen av denne ligningen er sannsynlighetsmaksimeringsestimatet av a. Vi ser at (14) er en $(n+1)$'te gradsligning i a som derfor ikke kan løses ved vanlige metoder med mindre $n = 1$. Derfor må denne estimeringsligningen løses numerisk. Vi ser at vi har en ligning på formen $f(a) = 0$ slik at vi kan finne løsningen ved å bruke Newton's metode. Vi trenger da $f'(a)$ som blir

$$-\frac{n}{a^2} + 2 \sum_{i=1}^{2} \frac{1}{(t_i + a)^2}. \tag{15}$$

En funksjon som tar en vektor av observasjoner som innargument og beregner SME av a, $\hat{a}$, ved hjelp av Newton's metode blir dermed:

```
ahat <- function(t,a=1,tol=1e-6) {
   n <- length(t)
   forrigea <- a-1
   while (abs(a-forrigea)>tol) {
      forrigea <- a
      a <- a - (n/a-2*sum(1/(t+a)))/(-n/a^2+2*sum(1/(t+a)^2))
      print(a)
   }
   return(a)
}
```

3 Inversjonsmetoden

La oss fortsette eksempelet i del 2. For å undersøke om vår estimeringsmetode fungerer trenger vi å beregne et estimat $\hat{a}$ på grunnlag av simulerte data. Da bør estimatet $\hat{a}$ ligge nærme den sanne verdien av a dersom utvalgsstørrelsen n er stor. Vi har tidligere laget oss data fra ulike fordelinger ved å bruke innebygde funksjoner i R for å simulere fra kjente fordelinger. Fordelingen gitt ved (10) svarer imidlertid ikke til noen fordeling som R kjenner. Vi trenger derfor en metode for å simulere fra (10).

Mer generelt skal vi se på en metode for å simulere fra kontinuerlige fordelinger ved bruk av den såkalte inversjonsmetoden. Metoden bygger på at vi er i stand til å simulere $U \sim \text{Unif}(0,1)$ (dette kan vi gjøre med funksjonen `runif` i R). Hvis vi kan finne en passende transformasjon $Y = g(U)$ slik at Y får fordelingen vi søker har vi løst problemet. Vi kan da simulere fra fordelingen til Y ved å simulere fra uniform fordeling og så beregne $Y = g(U)$. La $F(y)$ være den kumulative tettheten til fordelingen vi ønsker å simulere fra.

Transformasjonen vi søker, g, skal ha intervallet fra 0 til 1 som definisjonsområdet og samme verdiområde som variabelen Y. Dette vil være tilfelle for den inverse av den kumulative fordelingsfunksjonen til fordelingen vi søker, F^{-1}, denne transformerer tall på intervallet fra 0 til 1 til tall som ligger i verdiområdet til Y. Får vi rett fordeling om vi lar $Y = F^{-1}(U)$? Kumulativ tetthet til transformasjon blir

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P(F^{-1}(U) \leq y) \\ &= P(U \leq F(y)) \\ &= F_U(F(y)). \end{aligned} \tag{16}$$

Fordi U har kumulativ tetthet $F_U(u) = u$ for $0 \leq u \leq 1$ blir kumulativ tetthet til Y

$$F_Y(y) = F(y), \tag{17}$$

altså har Y fordelingen vi søker.

3.1 Generell algoritme

Hvis vi kjenner kumulativ fordeling $F(y)$ til en variabel Y og den inverse av F kan vi altså simulere en realisasjon av Y på følgende måte:

1. Simuler $U \sim \text{Unif}(0, 1)$.
2. Beregn $Y = F^{-1}(U)$.

3.2 Eksempel

I levetidsmodelleksempelet i del 2 hadde vi at

$$F_T(t) = 1 - \frac{a}{t+a}. \tag{18}$$

I følge inversjonsmetoden skal

$$T = F_T^{-1}(U) \tag{19}$$

har rett fordeling dersom $U \sim \text{Unif}(0, 1)$. Ligning (19) er ekvivalent med at

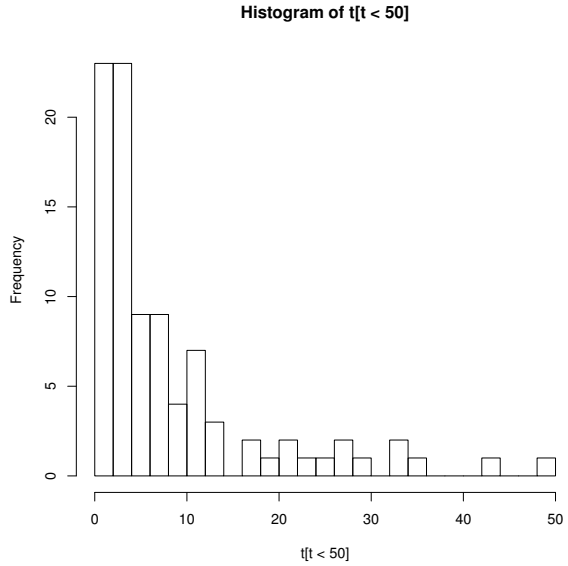
$$\begin{aligned} U &= F_T(T), \\ U &= 1 - \frac{a}{T+a}, \end{aligned} \tag{20}$$

altså transformasjonen

$$T = \frac{aU}{1-U}. \tag{21}$$

Bruker vi denne metoden kan vi simulere 100 observasjoner fra (10) på følgende måte. La oss anta at $a = 5$.

```
> a<-5
> u <- runif(100)
> t <- a*u/(1-u)
> range(t)
[1] 6.380138e-02 1.294394e+03
> hist(t[t<50],breaks=20)
```



Fordi dødsraten $\lambda(t)$ (7) går mot null når t blir stor får fordelingen en lang tynn øvre hale. I eksempelet over blir den lengste simulerte levetiden hele 1294 — langt større enn der alle fleste observasjonene.[1] Det er hensiktsmessig å ta bort disse observasjonene fra histogrammet.

Vi kan nå undersøke om estimatoren vår fra del 2 fungerer

```
> ahat(t)
[1] 1.659555
[1] 2.562604
[1] 3.596082
[1] 4.461676
[1] 4.863254
[1] 4.923986
[1] 4.925144
[1] 4.925144
[1] 4.925144
```

Vi ser at algoritmen konvergerer rimelig raskt og at estimatet ser ut til å bli liggende i nærheten $a = 5$ slik det bør når n er stor.

[1] Det kan og nevnes at forventningen ikke eksisterer. Forsøker vi å beregne forventningen $E(T) = \int_0^\infty t f_T(t) dt$ finner vi at dette integralet ikke konvergerer.