Preparació de les eines: Selecció del llindar

Esther Amores, Anna Costa i Oscar Ortiz

28/5/2021

Abans de començar amb la preparació de les eines, carreguem els paquets necessaris per tal d'executar les instruccions que es troben a continuació, així com també les dades `danish` del paquet `evir()`. Aquestes dades ens serviran com a referència per tal d'estimar els paràmetres d'una distribució power-law, així com també la distància de Kolmogorov-Smirnov.

```r
# Paquets
library(poweRlaw)

# Dades
data("danish", package="evir")
```

Apartat 1

Programeu en `R` la funció que dóna l'estimació màxim versemblant dels paràmetres d'una power-law. Llegiu la secció 3.1 de Clauset. Creem una funció anomenada `ePL` que retorna $\hat{\mu}$, que és el valor mínim de les dades, el paràmetre estimat α i el valor de la log-likelihood.

Donada una mostra $\{x_1, x_2, \dots, x_m\}$ busquem la màxima versemblança d'una distribució power-law.

- La funció de densitat és:
$$f(x; \alpha, \mu) = \frac{\alpha}{\mu}\left(\frac{\mu}{x}\right)^{\alpha+1}$$
- La funció de versemblança és:
$$L(\alpha, \mu) = \alpha^n \mu^{n\alpha} \left(\prod_{i=1}^{n} x_i\right)^{-(\alpha+1)}, \qquad (\mu \leq x_{min})$$
- La funció log-versemblança és:
$$l(\alpha, \mu) = logL(x) = n \cdot log(\alpha) + n \cdot \alpha \cdot log(\mu) - (\alpha+1)\sum_{i=1}^{n} log(x_i)$$

Fixat α el màxim és $\mu = x_{min}$. Així doncs, els paràmetres estimats d'una power-law són:

$$\hat{\mu} = x_{1,n}$$
$$\hat{\alpha} = \left[\frac{1}{n}\sum_{x=1}^{n} log\left(\frac{x_i}{x_{1,n}}\right)\right]^{-1} = \frac{1}{\xi}$$

que anomenarem `xm` i `alpha`, respectivament.

```r
ePL <- function(xdt){
  xm <- min(xdt)
  xi <- mean(log(xdt/xm))
  n <- length(xdt)
  al <- 1/xi
  lpl <- n*log(al)+n*al*log(xm)-(al+1)*sum(log(xdt))
  list(min=xm, alpha=1/xi, lPL=lpl)
}

ePL(danish)
```

```
## $min
## [1] 1
##
## $alpha
## [1] 1.270729
##
## $lPL
## [1] -3353.128
```

Apartat 2

Construïu un generador de nombres aleatoris per a la distribució power-law composant el generador d'uniformes, U(0,1), amb la funció quantil. La funció de distribució d'una power-law és:

$$F(x) = 1 - \left(\frac{x}{\mu}\right)^{-\alpha}$$

La funció quantil d'una distribució power-law es calcula fent $F^{-1}(x) = y$ i aillant la x d'aquesta equació. És a dir:

$$\begin{aligned} F^{-1}(x) = 1 - \left(\frac{x}{\mu}\right)^{-1/\alpha} &= y \\ \Leftrightarrow y - 1 &= -\left(\frac{x}{\mu}\right)^{-1/\alpha} \\ \Leftrightarrow (y-1)^{-1/\alpha} &= -\frac{x}{\mu} \\ \Leftrightarrow x &= \mu(1-y)^{-1/\alpha} \end{aligned}$$

$$Q(y) = F^{-1}(x) = \mu(1-y)^{-1/\alpha}$$

```r
rgpl <- function(mu, alpha, n, seed){
  set.seed(seed)
  y <- runif(n)
  x <- mu*((1-y)^(-1/alpha))
  return(x)
}
```

Apartat 3

Simuleu dades power-law amb la funció de l'apartat 1 i verifiqueu que la funció dóna l'estimació correcta, utilitzant el test de Kolmogorov-Smirnov:

$$D_n = \sup_{x_m < x < \infty} |F_n(x) - F_\alpha(x)|$$

En primer lloc, generem valors d'una distribució power-law amb paràmetres determinats α i μ. Després, calculem l'estadístic de contrast D. Aleshores, si X és un vector aleatori de dades $\{x_1, x_2, \ldots, x_n\}$ i $X \sim PL(\alpha, \mu)$, el test d'hipòtesis que avaluarem és el següent:

$$\begin{cases} H_0 & = X \sim PL(\alpha, \mu) \\ H_1 & = H_0 \text{ falsa} \end{cases}$$

Rebutjarem l'hipòtesi nul · la si $\sqrt{n} D_n > k_\alpha$.

```
mu <- 1
alpha <- c(0.01, 0.05, 0.10)

for(alpha in c(0.01, 0.05, 0.10)){
  # Simulació d'un vector de dades aleatòries
  x <- rgpl(mu=mu, alpha=alpha, n=1000, seed=1)
  # Estadístic de contrast
  X <- sort(x)
  n <- length(X)
  F <- 1-(X/mu)^(-alpha)
  E <- seq(1:n)/n
  D <- max(abs(E-F))
  est.con <- sqrt(n)*D
  print(sprintf("alpha=%.2f, est.con=%.8f", alpha, est.con))
}
```

```
## [1] "alpha=0.01, est.con=0.77051225"
## [1] "alpha=0.05, est.con=0.77051225"
## [1] "alpha=0.10, est.con=0.77051225"
```

α	0.10	0.05	0.01
Punt crític	1.22	1.36	1.63
$\sqrt{n}D_n$	0.7705123	0.7705123	0.7705123

No tenim evidències per rebutjar la hipòtesi nul · la quan $\alpha = 0.01, 0.05$ o 0.10 perquè cap dels estadístics de contrast obtinguts superen el llindar k_α. Per tant, les dades generades segueixen una distribució power-law amb parametres $\mu = 1$, $\alpha = 1$.

Apartat 4

Programeu en `R` la funció que dóna el mètode de selecció del llindar seguint el que explica Clauset a la secció 3.3, basat en la distància de Kolmogorov-Smirnov, estadístic KS.

Apartat 5

Ordeneu la mostra i calculeu per a cada valor considerat com origen la distància KS. L'estimació de x_{min} és el valor que minimitza KS.