Вероятности и статистика с R

Асен Чорбаджиев

April 26, 2017

1 Ковариация и корелация

Когато трябва да се опишат стистическите свойства на случайна функция често се налага тя да бъде измерена с експеримент от n различни количества, $n \geq 2$. Примери за двумерен случай с количествени вектори X,Y, които може да описват например възраст и височина на случайно избрани индивиди от една популация или измерен един и същ признак в различен период друг признак. Това се описва с n-мерна функция на разределение. В двумерен случай това се описва с двумерна функция на разпределение:

$$F_{X,Y}(x, y) = P(X \leq x, Y \leq y) \tag{1}$$

или със съответните функции на вероятност или плътност.
Такива две случайни величини X, Y са независими за всяко x,y, ако:

$$F_{X,Y}(x, y) = F_X(x)F_Y(y) \tag{2}$$

Важен концепт е ковариацията между двете случайни променливи X и Y, която е равна на:

$$C[X, Y] = E[(X - EX)(Y - EY)] = E(XY) - EXEY \tag{3}$$

Нормализираната стойност се нарича корелация и е равна на:

$$\rho[X, Y] = \frac{C[X, Y]}{\sqrt{V(X)V(Y)}} \tag{4}$$

За корелацията между две случайни величини е вярно:

- X,Y са корелирани $\rho \neq 0$
- X,Y са некорелирани $\rho = 0$
- Ако X,Y са независими, то $\rho = 0$. Обратното твърдение не е вярно.

Когато се пресмята корелацията на емпирични данни се използва коефициент на Пирсън:

$$r = \frac{\sum_i (x_i - \overline{X})(y_i - \overline{Y})}{\sqrt{\sum_i (x_i - \overline{X})^2 (y_i - \overline{Y})^2}} = \frac{1}{n-1}(\sum_i (\frac{x_i - \overline{X}}{s_x})(\frac{y_i - \overline{Y}}{s_y})) \tag{5}$$

функциите в R за ковариация и корелация са cov() и cor(). За графики се използват функции в R за scatterplot или функцията pairs() за многомерен случай. Основните графики за корелационни зависимости изглеждат така:

- Графика за $\rho = 1$
  x=c(1:10)
  y=c(1:10)
  cor(x,y)
  plot(x,y)

- Графика за $\rho = -1$
  x=c(1:10)
  y=c(10:1)
  cor(x,y)
  plot(x,y)
  plot(y,x)

- Графика за близко до нула ρ
  x=rnorm(100)
  y=rexp(100)
  cor(x,y)
  plot(x,y)

2 Многомерно Нормално разпределение

Многомерната (n-мерна) случайна векторна величина $X = (X_1, ...X_n)$ има нормално разпределение N(m,K) с плътност:

$$f_X(x) = \frac{1}{\sqrt{(2\pi)^n detK}} exp\{-\frac{1}{2}(x-m)^T K^{-1}(x-m)\} \tag{6}$$

$K = E((X-m)(X-m)^T)$ - ковариационна матрица Пример: Използване на пакета mvtnorm за чертаене на двумерно нормално разпределение.
library(mvtnorm)
x=y=seq(-5,5, length=50)

f=function(x,y)dmvnorm(cbind(x,y))
z=outer(x,y,f)
persp(x,y,z, theta=10, phi=20,expand=0.5, col="light blue")
persp(x,y,z, theta=10, phi=20,expand=0.5, col="light blue", shade=0.75) Параметрите theta и phi са полярни координати.

3 Упражнения:

- **Зад. 1.**: Да се пресметне корелацията между данните на таблицата iris. Да се начертае съвместна графика с pairs().

- **Зад. 2.**: Да се пресметне корелационната матрица на ковариацията K_Z на векторните случайни величини $X = (X_1, X_2, X_2)$:

$$K_Z = \begin{bmatrix} 16 & -14 & 12 \\ -14 & 49 & -21 \\ 12 & -21 & 36 \end{bmatrix} \tag{7}$$