Mestranda em Ecologia, Instituto de Biociências, USP.

Laboratory of Arthropod Behavior and Evolution

Atualmente estudando custos e benefícios do cuidado maternal em uma espécie de opilião (Arachnida, Opiliones) de Mata Atlântica.

Orientador: Glauco Machado

Proposta

Plano A

Meu plano para começar a conquistar o mundo é criar uma função que nos permita explorar um conjunto de dados e acessar rapidamente relações significativas entre as diferentes variáveis que o compõem. Ao trabalhar com dados morfométricos e comportamentais, por exemplo, uma importante análise inicial é a busca por relações significativas entre as diferentes variáveis a serem analisadas. Após transformar o conjunto de dados em uma matriz, a minha super-função fornecerá primeiro uma análise exploratória semelhante à da função summary, além de histogramas de todas as variáveis e boxplots comparando-as. Além disso, minha super-função calculará e fornecerá os resultados numéricos e gráficos dos modelos lineares resultantes das combinações de todas as variáveis entre si ou de um sub-conjunto delas (a ser escolhido e determinado pelo usuário). Além de diminuir o número de passos necessários para explorar um conjunto de dados, minha função destacaria quais dentre os modelos possíveis são significativos. Num mundo ideal minha função também destacaria e excluiria as variáveis entre as quais exista colinearidade, mas eu não sei como fazer isso.[1)]

Comentários

Alexandre

Estou imaginando que a unidade amostral é um indivíduo e os diferentes dados morfometricos relacionados a ele. Tenho uma sugestão para a saída gráfica: 1. Uma única saída gráfica com o número de gráficos nas colunas e linhas igual ao número de variáveis. 2. Na diagonal plotar o histograma da variável, na parte inferior da diagonal plotar o scatterplot das duas variáveis e a linha de tendência com os coeficientes do modelo linear, acima o valor do coeficiente de correlação e o p da correlação entre as duas variáveis. Só isso já ajudaria vc. a tomar a decisão de quais variáveis manter ou descartar... Veja os exemplos no help da função `pairs`, tem algo semelhante lá... Boa sorte!

Paulo

Tem um problema estatístico importante a se levar em conta aí que é o das comparações múltiplas. A significância é calculada para um teste. Fazendo vários com o mesmo conjunto de dados você deve aplicar alguma correção ao valor de *p* obtido. A mais rigorosa e simples é a de Bonferroni: divida o p crítico que vc vai usar pelo número de comparações. Então se vc tem 3 variáveis e vai correlacionar

todas duas a duas então são 3 combinações, e se vc quer manter um alfa de 0,05 para cada combinação vocẽ deve baixar o alfa em cada teste para 0,05/3. Veja a wikipedia para a fórmula do numero de combinações (o R tem uma função para isto, mas vc pode também fazer uma:-)). Tudo isto e muito mais estão super explicados em Manly 2008[2)]. Talvez no seu e-mail haja algo a respeito ...

Enfim, sua função **precisa** levar isto em conta, ou desencanar de significâncias. Mas se vc mantiver vai mesmo dominar o mundo!

Plano B

Caso eu não consiga descrever o mundo e as relações entre as variáveis que o compõem logo de cara, vou me contentar em descrevê-lo, ou seja, vou criar uma função que forneça uma primeira análise exploratória dos dados (semelhante à da função summary), além de histogramas e gráficos qqnorm de todas as variáveis e boxplots comparando-as. Essa função permitira acessar com apenas um comando as principais características de cada variável.

Código da Função

```
 data.explorer <- function(x)
{
x11()
panel.hist <- function(x, ...)
{
   usr <- par("usr"); on.exit(par(usr))
   par(usr = c(usr[1:2], 0, 1.5) )
   h <- hist(x, plot = FALSE)
   breaks <- h$breaks; nB <- length(breaks)
   y <- h$counts; y <- y/max(y)
   rect(breaks[-nB], 0, breaks[-1], y, col="blue3", ...)
 }
panel.cor <- function(x, y, digits=2, prefix="", cex.cor)
 {
    usr <- par("usr"); on.exit(par(usr))
    par(usr = c(0, 1, 0, 1))
    r <- (cor.test(x, y))
    p <- p.adjust(r$p.value, method = "holm", n = length(r$p.value))
    txt <- format(r$estimate, digits=digits)
    txt <- paste(txt, sep="")
    if(p<=0.05)
         {
         if(missing(cex.cor)) cex <- 0.1/strwidth(txt)^2
         text(0.5, 0.7, label=paste("Cor.value = ", txt), cex = cex,
col="green4")
         }
    else
         {
```

```
          if(missing(cex.cor)) cex <- 0.07/strwidth(txt)^2
          text(0.5, 0.7, label=paste("Cor.value = ", txt), cex = cex,
col="red")
          }
     txt1 <- format(p, digits=digits)
     txt1 <- paste(txt1, sep="")
     if(p<=0.05)
          {
          if(missing(cex.cor)) cex <- -2/strwidth(txt1)
          text(0.5, 0.3, label=paste("p Value = ", txt1), cex = cex,
col="green4")
          }
     else
          {
          if(missing(cex.cor)) cex <- -2/strwidth(txt1)
          text(0.5, 0.3, label=paste("p Value = ", txt1), cex = cex,
col="red")
          }
  }
 panel.myfitline <- function(x, y, digits=2, prefix="", cex.cor, ...)
  {
    usr <- par("usr")
    res<-panel.smooth(x,y, col.smooth="blue",...)
    reg <- coef(lm(y ~ x))
    abline(coef=reg,untf=F, col="red")
    const<-format(reg[1], trim = FALSE, digits = NULL, nsmall = 0, justify =
"left", ...)
    const<-paste(prefix, const, sep="")
    slope<-format(reg[2], trim = FALSE, digits = NULL, nsmall = 0, justify =
"left", ...)
    slope<-paste(prefix, slope, sep="")
    if(missing(cex.cor)) cex <- 0.8/strwidth(const)
    text(.8*usr[2], usr[3] +.1*(usr[4]-usr[3]), const, cex=cex)
    if(missing(cex.cor)) cex <- 0.8/strwidth(slope)
    text(usr[1] + .4*(usr[2]-usr[1]), .9*usr[4], slope, cex=cex)
  }
 result <- pairs(x, panel=points, diag.panel=panel.hist,
lower.panel=panel.myfitline, upper.panel=panel.cor)
 sumario <- summary(x)
 resultado <- list("Veja janela gráfica", sumario, "Verde = significativo
para alfa 0,05; Vermelho = não significativo   para alfa 0,05", result)
 return(resultado)
 }
```

Página de Ajuda

```
 data.explorer                                          package: none
R.Documentation
```

Análise da relação entre as diferentes variáveis numéricas de uma matriz de dados.

Description:

Produz uma saída gráfica contendo os histogramas das diversas variáveis numéricas contidas em uma matriz, suas correlações duas a duas (com índices de significância corrigidos pelo método de Holm (1979)) e gráficos de dispersão entre todos os pares de variáveis, com suas retas e coeficientes de regressão linear.

Usage:

data.explorer(x)

Arguments:

x: Matriz de dados com ao menos 3 colunas. As variáveis em estudos devem estar representadas pelas colunas e as unidades amostrais pelas linhas.

Details:
O resultado da função retorna uma janela gráfica, o sumário da matriz de dados original e a legenda das cores utilizadas na saída gráfica. A função retorna também retorna um fator “NULL”, que não tem relação alguma com o resultado obtido.

A função retorna , na diagonal da janela gráfica, os histogramas de todas as variáveis da matriz original. Na porção superior à diagonal a função retorna os valores das correlações entre as variáveis, em cores diferentes de acordo com o seu resultado. Correlações com índice de significância maior que 0,05 são apresentadas em vermelho, ao passo que correlações com índice de significância menor que 0,05 são apresentadas em verde. O tamanho da fonte em que os resultados das correlações são apresentados também varia em função de sua significância. Na porção inferior à diagonal a função retorna gráficos de dispersão combinando as variáveis duas a duas, juntamente com suas linhas de tendência (em azul) e suas retas de regressão linear (em vermelho). A função retorna também, na própria janela gráfica, o resultado da função “coef”, ou seja, o valor de intercepto (porção inferior da janela gráfica) e o coeficiente de regressão (porção superior da janela gráfica) do modelo linear. Por enquanto, este modelo segue o padrão “modelo <- lm(coluna 2 ~coluna 1)”. Considerando que esta é uma função em fase de desenvolvimento, espera-se poder, em breve, obter também na mesma janela gráfica o resultado do modelo “modelo <- lm(coluna 1~coluna 2).

See also:

```
A função pairs retorna uma matriz de scatterplots.
O fórum de discussão do R retorna muitas coisas interessantes.

 Example:
 ### Definicao da matriz de exemplo:
 largura=c(1.5, 2.7, 1.4, 1.7, 1.9, 2.2, 2.5, 2.3, 1.6, 1.7)
 comprimento=c(2.3, 3.2, 2.3, 2.9, 3.0, 3.1, 3.1, 3.2, 2.7, 2.5)
 massa=c(0.3, 0.6, 0.4, 0.5, 0.5, 0.7, 0.9, 1.0, 0.7, 0.5)
 comp.perna=c(4, 3.5, 3.9, 4.2, 3.7, 3.8, 4,1, 3,6, 4, 3,5)
 nomes=c("largura","comprimento","massa", "comp.perna")
 individuos=c(1,2,3,4,5,6,7,8,9,10)
 morfo=c(largura, comprimento, massa, comp.perna)
 morfo
 exemplo=matrix(morfo, nrow=10, ncol=4, byrow=FALSE)
 rownames(exemplo)=individuos
 colnames(exemplo)=nomes
 exemplo
 ### teste da função:
 data.explorer(exemplo)


 Author(s):

 Marie-Claire Monier Chelini
 mcchelini@gmail.com
```

[1)]
veja comentário no formato de edição da página

[2)]
Manly, B. 2008. Statistics for Environmental Sciences and Management. 2nd Ed. Chapman & Hall.