Renato Sousa Recoder

Sou aluno de doutorado pelo departamento de Zoologia do IB-USP, orientado pelo prof. Miguel Trefaut Rodrigues. Meu projeto tem como foco a variação morfológica e genética em uma linhagem de lagartos microteídeos (tribo Iphisini), visando compreender padrões de diferenciação e especiação.

Currículo Lattes

renatorecoder@gmail.com

Exercícios

exec

Trabalho Final

Proposta 1. Outliers multivariados

Em análises morfométricas, é comum que se realize uma etapa de exploração dos dados antes da realização de análises, com o objetivo de assegurar a ausência de outliers, e conformação com requisitos para análises paramétricas. No entanto é raro que sejam feitas explorações levando em conta a covariância entre variáveis para análises multivariadas, fator que torna-se mais importante quando a espécie de interesse apresenta crescimento alométrico. O objetivo desta proposta é criar uma função capaz de: 1. delimitar outliers multivariados, 2. estimar valores para os outliers com base na curva de crescimento alométrico da espécie (por regressão), e 3. indicar se há conformação com requisitos para análises multivariadas.

Entrada

Um dataframe ou matriz com duas ou mais variáveis quantitativas dependentes e fatores de interesse.

Saída

Tabela com teste de premissas (normalidade e equilíbrio de variâncias multivariados) e teste de covariância de caracteres entre grupos. Tabela com outliers identificados e o valor estimado com base na curva alométrica.

> Sua proposta parece bem útil. Não sei se entendi todos os termos técnicos, então vou pedir pra você dar mais detalhes pra dar pra entender o que exatamente sua função vai fazer.

Em primeiro lugar, como exatamente será essa entrada? Há uma separação entre as variáveis dependentes e os “fatores de interesse”, ou a função tratará todas elas da mesma forma?

Como exatamente será o formato da saída? Uma lista de tabelas? Você pretende testar a normalidade de cada variável? Não entendi o que é uma tabela de teste de covariância de caracteres entre grupos. A idéia é apresentar a matriz de covariâncias, para explicitar quais variáveis têm alta correlação?

Pela sua explicação parece que o ponto central seria identificar os outliers multivariados. Talvez fosse bom você focar nesse ponto. Um problema de fazer uma função que faz muitas análises ao mesmo tempo é que cada pesquisador escolhe que análises fazer com base em uma porção de coisas, então é melhor fazer uma função que faça uma análise bem-feita do que uma que faça diversos testes ao mesmo tempo.

Com isso em mente, quantas variáveis ao mesmo tempo serão usadas pra encontrar esses outliers? Todas? Ou sub-conjuntos delas? Se forem só duas, você poderia mostrar num gráfico esses outliers. Como serão selecionadas as variáveis?

—-*Mali*

Olá Renato, Primeiro, é importante considerar que a função não é um script de análise de dados. Lembre que a tarefa que você quer automatizar depende muito da análise do pesquisador. Tente implementar a função de maneira que i) o usuário tenha consciência disso, o que implica que o help estará muito bem explicado e/ou ii) o usuário possa tomar as decisões, o que implica em deixar alguns argumentos livres para o usuário decidir como prosseguir. Segundo, acho que ainda não está claro como será o passo a passo da implementação da função. Tente pensar neste passo a passo da maneira mais geral possível, dado o contexto da sua função. E mais, tente relevar os pontos que eu e Mali colocamos. A proposta ainda está confusa, mas pode ser implementada de forma interessante.

— *Sara Mortara*

Proposta 2. Erro de medição

Em análises morfométricas é comum a ocorrência de erros de medição, em geral, causados por definições imprecisas de marcos anatômicos ou por variações na acurácia do método de tomada de dados. A identificação de caracteres com variâncias enviesadas por erro de medição é crucial na decisão de incluir/excluir variáveis em análises multivariadas, e assim, evitar resultados espúrios. Esta proposta visa criar uma função que estime o erro de medição associado a variáveis utilizadas,

calculando a proporção da variância explicada pelo erro em comparação com outros fatores de interesse (variância entre grupos).

Entrada

Um ou mais dataframes contendo medições morfométricas com duas ou mais réplicas, e um fator de interesse (sexo, espécie, local...)

Saída

Uma tabela com a variância explicada pelo erro, por um ou mais fatores, e pela interação dos mesmos.

> Esta função está mais clara que a anterior, mas ainda não está claro que tipo de técnica você pretende usar para fazer essa análise. Lambre de explicitar exatamente qual a técnica usada. Por exemplo, para calcular o erro de medição, você usará todos os blocos, gerando uma variância média?
>
> Além disso a função parece simples demais. Tenho algumas sugestões de como incrementá-la: você pode estimar também a precisão da sua estimativa de variância, e pode fazer um gráfico ilustrando seus resultados.
>
> —-*Mali*

> Oi Renato, a proposta é interessante porém simples. De novo, faltou apresentar como será o passo a passo da implementação. Acrescentar uma saída gráfica sempre adiciona um desafio a mais. Você inclusive deixar à escolha do usuário se ele quer ou não plotar o gráfico. Sugiro que você siga em frente com a proposta que te motiva mais. Para qualquer uma das duas, pense melhor na implementação e em como torná-la mais interessante em termos de programação. Boa sorte! — *Sara Mortara*

Proposta de trabalho final: parte dois

Olá! Obrigado Mali e Sara pelos pareceres!

Eu particularmente prefiro a proposta dois, apesar de ter deixado como plano B caso parecesse simples demais. Eu acho que esta função tem um cálculo/resultado mais direto do que a proposta um e, apesar de parecer tecnicamente simples, envolve calcular as variâncias como um modelo de ANOVA do tipo II e partilhar as proporções da variância total. Basicamente eu pretendo que a função resolva com um comando o cálculo da fórmula de erro de medição de Bailey & Byrnes (1990):

$\%ME = 100\% * s^2_{within}/s^2_{within} + s^2_{among}$

onde a variância entre grupos é uma razão da diferença entre variâncias médias (MS) e o número de repetições nas medições(n):

$s^2_{among} = MS_{among} - MS_{within}/n$

e assim, retorne o valor de erro de medição e outros parâmetros que pudessem ajudar na decisão de manter ou excluir uma variável de análises multivariadas. Eu concordo que dá para incrementar estes cálculos para ficar mais interessante, incluindo gráficos dos desvios (como no exercício da ANOVA) e outras estimativas (como sugerido pela Mali), por exemplo qual o mínimo de repetições devem ser feitas pra se obter uma média confiável apesar do erro, ou quantos indivíduos devem medidos pra obter uma média e variância de um grupo, apesar do erro. Talvez fazer reamostragens pra estimar significância... Mas ainda preciso pensar na parte matemática desta etapa e sugestões são bem-vindas! 😃

Como entrada, pensei que seria mais simples usar um dataframe com uma variável de interesse (e.g. uma medida morfométrica), e como fator os indivíduos com as respectivas repetições (cada exemplar medido com um número n, igual para as k repetições).

HELP DA FUNÇÃO EDM

```
edm { nenhum}                                R Documentation


Função para estimar erro de medição em morfometria


Description:

Calcula o erro de medição/repetibilidade associados à tomada de dados
morfométricos
e simula amostragens de forma a minimizar o erro para análises subsequentes.


Usage:

edm(x, y, data=data, IC=FALSE, estima=FALSE, N=N)


Arguments:

x: um vetor contendo a identificação dos indivíduos e as respectivas
réplicas
(ver detalhes).

y: um vetor contendo uma variável morfométrica representada por valores
numéricos.
```

```
data: um objeto da classe “data frame” contendo os vetores “x” e “y”.

IC: argumento lógico que indica se a função deve (TRUE) ou não (FALSE)
calcular
intervalos de confiança associados à repetibilidade.

estima: argumento lógico que indica se a função deve (TRUE) ou não (FALSE)
retornar
um gráfico do tipo boxplot com N estimativas de repetibilidade para k
réplicas
(ver detalhes).

N: número de estimativas de repetibilidade a ser feitas com base em
reamostragem
de 2:k medições.


Details:

O objeto “x” deve ser representado por um vetor contendo a identificação dos
n indivíduos para as k réplicas, correspondentes aos valores numéricos
(medições
morfométricas) apresentados no objeto “y”.

O vetor “x” deve ser da classe “fator”, e caso não o seja, a função realiza
a
transformação.

A função deve extrair os vetores “x” e “y” de um objeto do tipo “data
frame”.

O intervalo de confiança é estimado usando alfa = 5% de significância.

O argumento "estima" calcula diferentes valores de repetibilidade "r"
através de N
reamostragens sem reposição de k-1 replicas.

Valores de "r" e "EM" variam entre 0 e 1, e são complementares (EM = 1-r).


Value:

edm
Retorna uma lista com o valores calculados para erro de medição de uma
variável
dependente (EM), repetibilidade (r), intervalo de confiança da
repetibilidade (ICsup,
ICinf), tamanho da amostra (n) e número de repetições (k); um boxplot com N
valores
de repetibilidade calculados para diferentes classes de repetições (k-1).
```

```
Author(s):

Renato Sousa Recoder (renatorecoder@gmail.com)


References:

Bailey, R. C., & Byrnes, J. (1990). A new, old method for assessing
measurement
error in both univariate and multivariate morphometric studies. Systematic
Biology, 39(2), 124-130.

Bonett. D.G. (2002). Statistics in Medicine, 21(9), 1331-1335.

Claude, J. (2008). Morphometrics with R. Springer Science & Business Media.

Wolak, M. E., Fairbairn, D. J., & Paulsen, Y. R. (2012). Guidelines for
estimating
repeatability. Methods in Ecology and Evolution, 3(1), 129-137.

Yezerinac, S. M., Lougheed, S. C., & Handford, P. (1992). Measurement error
and
morphometric studies: statistical power and observer experience. Systematic
Biology,
41(4), 471-482.


Examples:

#medidas morfometricas de uma amostra de lagartos Micrablepharus:
#15 individuos, 5 repeticoes.

#implementação mais simples da função edm
micra.morpho <- read.table("micra_morpho.txt", head=TRUE, sep=",", dec=".")
str(micra.morpho)
edm(IND, CRC, data=micra.morpho)

#implementação mais completa da função edm
str(micra.morpho)
edm(IND, CRC, data=micra.morpho, IC=TRUE, estima=TRUE, N=20)
```

CÓDIGO DA FUNÇÃO EDM

```

#Codigo da funcao "edm" com intervalos de confianca e graficos

edm <- function (x, y, data=data, IC=FALSE, estima=FALSE, N=10)
{
  #chama os vetores contendo as variaveis de interesse "x" e "y" e produz um
"data
```

```
  #frame".
  edm.call <- Call <- match.call()
  x <- as.character(edm.call[[2L]])
  y <- as.character(edm.call[[3L]])
  dados <- data.frame(data[x], data[y])
  #numero de reamostragens em caso de multiplas estimativas de erros.
  N <- N
  #testa se a variavel independente pertence a classe fator, e faz a
transformacao
  #em caso negativo
  if (!is.factor(dados[, 1]))
  {
    dados[, 1] <- as.factor(dados[, 1])
  }
  #estima o numero de amostras e de replicas.
  n <- length(unique(dados[, 1]))
  k <- length(dados[, 1])/n
  #realiza uma anova com as variaveis de interesse, das quais sao retiradas
os
  #valores de desvios quadrados medios, utilizados no calculo de "EM" e "r".
  funcao <- formula(dados[, 2] ~ dados[, 1])
  varis <- anova(aov(funcao, data = dados))
  num.df <- n-1
  denom.df <- n*(k-1)
  MSa <- varis [1, 3]
  Sw <- MSw <- varis [2, 3]
  Sa <- (MSa - MSw)/k
  #calcula o valor de repetibilidade.
  r <- Sa/(Sw + Sa)
  #calcula o valor de EM.
  EM <- Sw/(Sw + Sa)
  #argumento que chama a estimativa de multiplas repetibilidades.
  if (estima == TRUE)
  {
    result3 <- data.frame(rep(2:k, each=N), rep(NA, ((k-1)*N)), rep(NA,
((k-1)*N)))
    #produz N estimativas de repetibilidade para subamostras das 2:k
replicas das
    #medicoes.
    for(l in 2:k){
      result2 <- data.frame(rep(NA, N), rep(NA, N))
      #refaz a estimativa de repetibilidade N vezes.
      for(j in 1:N){
        result <- data.frame(rep(1:n, each=l) , rep(NA, n*l))
        #loop que calcula a repetibilidade
        for(i in 1:n){
          result[, 2][which(result[, 1]==i)] = sample((dados[which(dados[,
1]==i), ]
          [, 2]), size=l)
          result[, 1] <- as.factor(result[, 1])
        }
```

```
        funcao <- anova(aov(result[, 2] ~ result[, 1], data=dados))
        n <- length(unique(result[, 1]))
        rep <- length(result[, 1])/n
        MSa <- funcao [1, 3]
        Sw <- MSw <- funcao [2, 3]
        Sa <- (MSa - MSw)/k
        r <- Sa/(Sw + Sa)
        result2[, 1][j] = r
        result2[, 2][j] = l
      }
      result3[, 2][which(result3[, 1]==l)] = result2[, 1][1:N]
      result3[, 3][which(result3[, 1]==l)] = result2[, 2][1:N]
    }
    #abre uma janela e plota grafico boxplot com as estimativas de
repetibilidade.
    x11()
    boxplot(result3[, 2] ~ result3[, 1])
    mtext("Repetibilidade", side=2, cex=1.5, line=2.5, las=0)
    mtext("Número de réplicas", side=1, cex=1.5, line=2.5)
  }
  #argumento que controla o calculo de intervalos de confianca
  if (IC == TRUE)
  {
    #calculo dos intervalos de confianca segundo Bonett(2002).
    F.stat <- varis [1, 4]
    low.F <- qf(0.05/2, num.df, denom.df, lower.tail = FALSE)
    up.F <- qf(0.05/2, denom.df, num.df, lower.tail = FALSE)
    FL <- F.stat/low.F
    FU <- F.stat * up.F
    ICLow <- (FL - 1)/(FL + k - 1)
    ICUp <- (FU - 1)/(FU + k - 1)
    ICdif <- ICUp-ICLow
    #retorna a lista com valores calculados pela funcao
    return(list(edm = EM, repetibilidade = r, ICsup = ICUp, ICinf = ICLow,
ICdif =
    ICdif, individuos = n, replicas = k))
  }
  #retorna resultados simplificados, sem calculo de ICs.
  else
    return(list(edm = EM, repetibilidade = r, individuos = n, replicas = k))
}
```

ARQUIVOS DA FUNÇÃO EDM

help.edm.txt

codigo_edm.txt

micra_morpho.txt