Modelos lineares mistos v2018

Esse roteiro foi inicalmente desenvolvido por Melina Leite, Marília Gaiarsa e Lucas Medeiros, e sua versão original está disponível neste site

Para acompanhar o roteiro é importante ter uma compreensão básica de análises estatísticas em R. Isso quer dizer que você já usou e sabe interpretar o output de funções como lm(). Esperamos que ao terminar você seja capaz de:

1. Entender o que são efeitos fixos e aleatórios.

2. Compreender a estrutura básica de um modelo linear misto.

3. Fazer uma análise de modelo misto no **R** usando o pacote `lme4` (Bates et al. 2014)

4. Entender o *output* da função `lmer`.

5. Decidir quais efeitos aleatórios manter no seu modelo final.

6. Tirar conclusões a partir da análise de um modelo misto por meio de teste de hipótese ou seleção de modelos.

7. Fazer uma análise visual de diagnóstico do modelo.

Caso se sinta um pouco perdido com certas terminologias estatísticas ou queira relembrar alguns termos, ao final do roteiro temos um pequeno **glossário** que pode ajudar.

Modelos Mistos

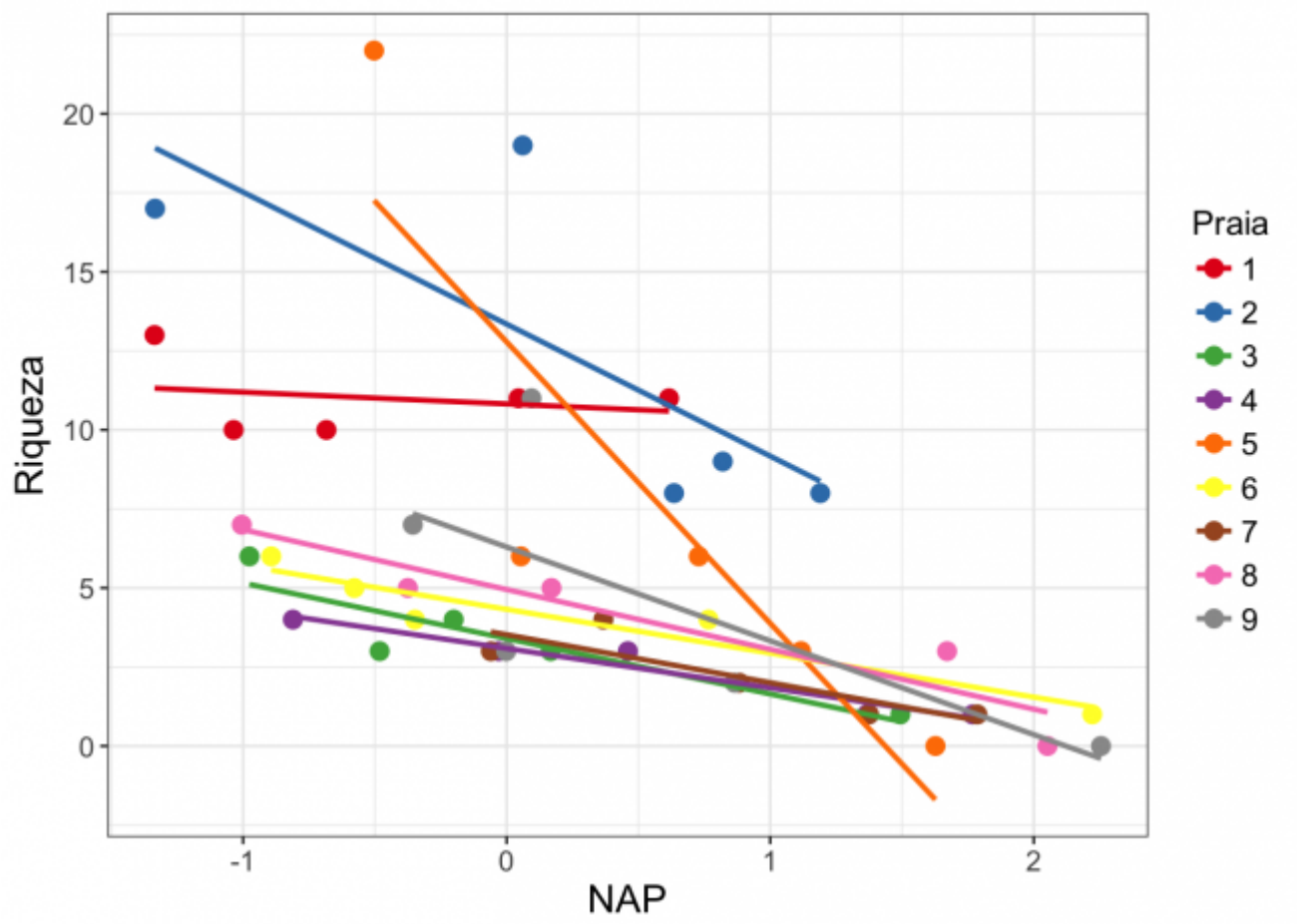
Para explicar o que são modelos mistos e sua importância em ecologia, usaremos como exemplo um conjunto de dados presente no capítulo 5 de Zuur et al. (2009). Esses dados contêm a riqueza de espécies da macro-fauna em 9 praias na costa da Holanda. Em cada uma das praias, os autores coletaram dados em cinco sítios diferentes. Para cada sítio existe informação sobre a altura da estação de amostragem em relação à altura média da maré (`NAP`, variável contínua) e também um índice de exposição da praia (`Exposure`, variável categórica).

Vamos supor que estamos interessados em verificar se a variável `NAP` influencia a riqueza de espécies nessas praias, deixando de lado a variável de exposição por hora. Uma primeira ideia que poderíamos ter é construir um modelo linear da seguinte forma:

$$riqueza = \alpha + \beta * NAP + \varepsilon$$

Aqui estamos modelando a riqueza de cada praia em função da variável `NAP`. O coeficiente α é o intercepto do modelo, β é o coeficiente angular (inclinação) de `NAP` e ε é o erro do nosso modelo (resíduos), isto é, a variação da riqueza que não conseguimos explicar com nossa variável preditora `NAP`. No entanto, esse modelo tem um problema. Estamos violando uma premissa fundamental de modelos lineares, a de que os dados são independentes uns dos outros (Winter, 2013). Os dados obtidos em uma mesma praia não são independentes entre si

(dependência espacial). Podemos imaginar diversas características de cada praia que podem influenciar a riqueza de espécies, como o tipo de grão de areia ou a força das ondas. Veja o gráfico abaixo, em que cada cor representa uma praia, para se convencer de que a relação entre riqueza e `NAP` varia entre praias:



Nesta figura, temos as retas do modelo entre `Riqueza` e `NAP` aplicado para cada praia (usando a função `lm()`). Entretanto, nossa pergunta inicial não diz respeito a cada praia separadamente, nós queremos modelar esta relação para todas as praias amostradas, sem pseudoreplicação (não-independência dos dados) e queremos fazer isso de forma a ser possível predizer a riqueza em praias não amostradas.

Portanto, precisamos de um modelo linear que incorpore o fato de que nossos dados estão agrupados em praias. Chamamos de um **efeito aleatório** uma variável que agrupa nossos dados e que geralmente seu efeito sobre a variável resposta não nos interessa diretamente (nesse exemplo não interessa, mas dependendo da análise, pode interessar - veja discussão em McGill 2015). Nesse caso, os autores amostraram nestas 9 praias, mas poderiam ter feito o mesmo estudo escolhendo outras praias. O efeito aleatório `praia` "organiza" a parte da variação nos nossos dados que não conseguimos explicar, presente no erro ε do modelo (Winter, 2013). Por outro lado, as variáveis preditoras que estamos acostumados a encontrar em modelos lineares são chamadas de **efeitos fixos**. No exemplo das praias, a variável `NAP` é um efeito fixo e estamos diretamente interessados em seu efeito sobre a variável resposta. O nome "misto" advém do fato de que existe ao menos um efeito fixo e um efeito aleatório no modelo.

Em ecologia, é comum encontrarmos delineamentos amostrais ou experimentais que geram dados com algum tipo de agrupamento (delineamento hierárquico ou aninhado). Por exemplo, quando uma amostragem é feita por parcelas ou quando um experimento é separado em diferentes blocos. Nesses

casos, não podemos tratar nossos dados como independentes e a abordagem estatística mais adequada é a de **modelos mistos**.

Na seção a seguir veremos como explorar modelos mistos no **R**. Veremos que a forma mais simples de incorporar o efeito aleatório em nosso exemplo é definir que cada praia apresenta um intercepto diferente no modelo. Ou seja, queremos ajustar uma reta para nossa relação entre `riqueza` e `NAP`, mas permitir que cada praia tenha sua própria "retinha", com intercepto variando em relação ao intercepto da reta "principal" (efeito fixo). Nosso modelo será:

$$riqueza_i = \alpha + b_i + \beta * NAP + \varepsilon_{i}$$

A riqueza da praia `i` é explicada pelo efeito fixo $\beta*NAP$ e pelo efeito aleatório b_i, que se soma ao valor do intercepto fixo do modelo, α, para formar o intercepto da praia `i`. Veja que o índice `i` está atrelado a b_i e, portanto, o modelo permite que cada uma das praias apresente uma relação diferente entre riqueza e NAP. Já β faz parte do efeito fixo, e não possuí nenhum índice `i` atrelado a ele. Finalmente, ε representa o erro associado a cada uma das amostras na mesma praia.

Mãos à massa! Modelos mistos no R

Existem vários pacotes disponíveis no R para realizar análises de modelos mistos. Neste roteiro usaremos o `lme4` (Bates et al. 2014), que possui funções para analisar modelos lineares mistos, modelos lineares mistos generalizados e modelos mistos não lineares. A seguir, vamos colocar a mão na massa e analisar os dados provenientes do capítulo 5 de Zuur et al. (2009) que exploramos na seção anterior.

Antes de tudo, precisamos baixar e instalar o pacote `lme4`:

```
install.packages("lme4")
library(lme4)
```

Os dados estão disponíveis no site do livro do Zuur et al. (2009) (baixe aqui o zip com os dados - "data files").

```
dados <- read.table("RIKZ.txt", header = TRUE, row.names = 1, as.is = TRUE)
head(dados) #observando as primeiras linhas de dados
```

OBS: Antes de prosseguir, temos que modificar a variável `Exposure` nos dados. Originalmente esta variável tem 3 níveis, mas porque o valor mais baixo foi observado apenas em uma praia, nós iremos reclassificar esta praia para o segundo valor mais baixo (Zuur et al. 2009). Depois vamos transformar essas variáveis em fatores, dando o nome de "low" e "high" para o índice de exposição abaixa e alta, respectivamente.

```
#descobrindo qual é o valor de `Exposure` (colunas) em cada praia (linhas)
table(dados$Beach, dados$Exposure)

#renomeando a praia 2 de Exposure = 8 para Exposure = 10
dados$Exposure[dados$Exposure == 8] <- 10
```

```
# criando uma nova coluna com a variável exposure fator
dados$fExposure[dados$Exposure == 10] <- "low"
dados$fExposure[dados$Exposure == 11] <- "high"
dados$fExposure <- as.factor(dados$fExposure)
```

Agora, vamos construir nosso modelo. Relembrando, estamos interessados em ver se existe um efeito de `NAP` na riqueza de espécies. Nossos dados contam com nove diferentes praias e cinco diferentes amostras para cada uma delas. Dessa forma, temos como efeito fixo a variável `NAP`, e como efeito aleatório a variável praia (`Beach`), que significa que cada praia terá um intercepto diferente. Assim, nosso modelo é:

```
modelo.riqueza <- lmer(Richness ~ NAP +  (1 | Beach), data = dados)
```

Em seguida, vamos dar uma olhada no output do modelo:

```
summary(modelo.riqueza)
```

Vamos dar uma olhada na parte dos efeitos aleatórios (**Random effects**). O desvio padrão é uma medida do quanto a variablidade da nossa variável dependente - riqueza - é devida aos dois efeitos aleatórios que estamos analisando (os interceptos das praias e os resíduos). Podemos ver o desvio padrão associado às diferenças de intercepto entre praias (o b_i do modelo). A última linha nos dá o resíduo, que indica o quanto da variabilidade não é prevista pela praia nem pelo `NAP`, que nada mais é do que o ε acima explicado.

Os efeitos fixos indicam os coeficientes estimado pra cada um dos fatores que estamos considerando como fixos. No caso, temos o intercepto que é a riqueza quando o `NAP` é zero (6.5819), e o coeficiente angular de NAP (-2.5684).

```
#criando objeto com os coeficientes do modelo (efeitos fixos)
cof <- fixef(modelo.riqueza)
```

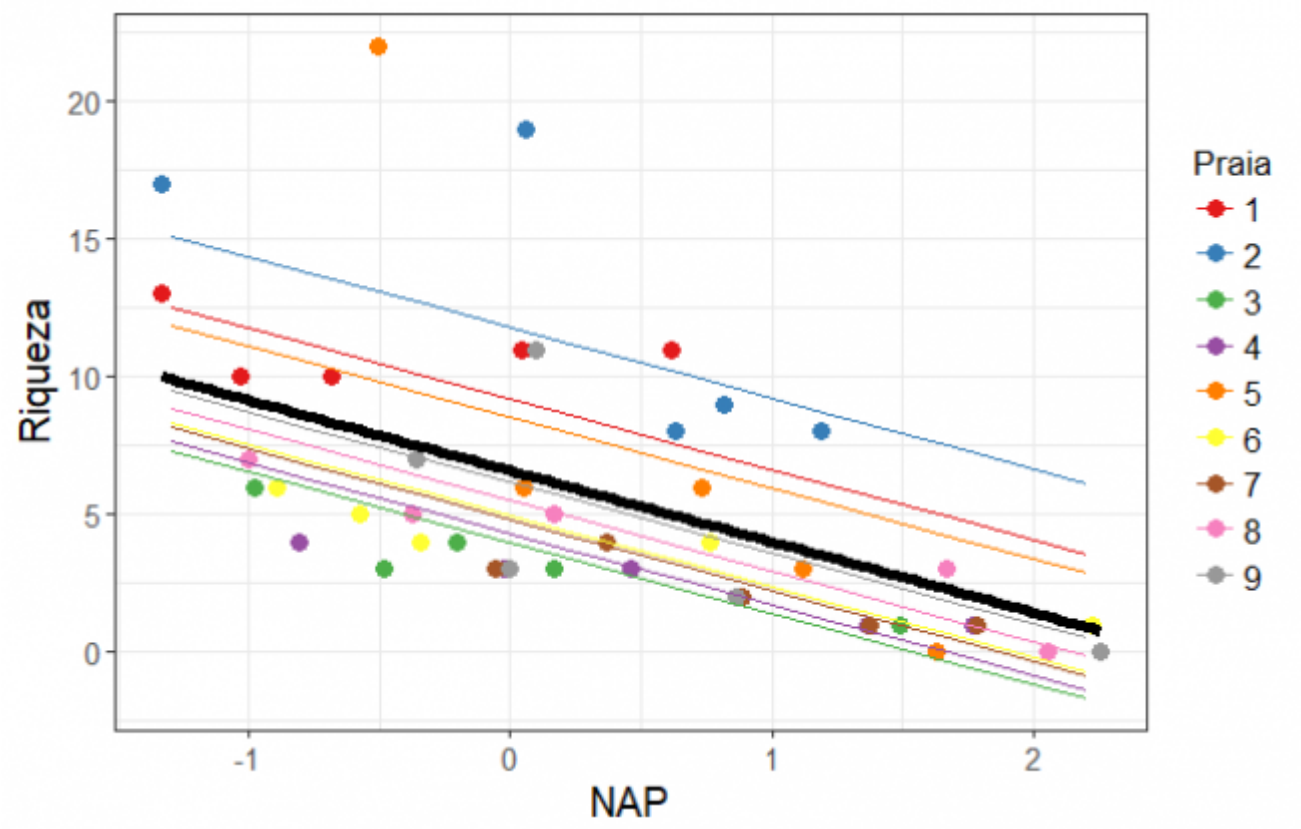
Para fazer o gráfico de nosso modelo ajustado, com a reta (predição) dos efeitos fixos e as “retinhas” preditas para cada praia, vamos primeiro calcular os valores preditos para cada praia [1)].

```
# Primeiro criamos um novo conjunto de dados com as características no
antigo
novo <- expand.grid(Beach = unique(dados$Beach),
                    NAP = seq(-1.3,2.2,0.5))

#agora calculamos os valores preditos para esse novo conjunto de dados
preditos <- predict(modelo.riqueza, newdata = novo)

#guardando tudo em um novo data.frame
dados.preditos <- data.frame(pred = preditos, novo)
```

Agora podemos plotar os dados com o ajuste do nosso modelo [2)].




```
#não esqueça de instalar o pacote
#install.packages("ggplot2")

library(ggplot2)

funi <- function(x){cof[1] + cof[2]*x} # para plotar a predição dos efeitos
fixos
dados$Praia = as.factor(dados$Beach)# transformando praia em fator

ggplot(data = dados, aes(x = NAP, y = Richness, color = Praia)) + # dados e
eixos
  geom_point(size = 3, shape = 19) +  # plotando os pontos das praias
  geom_line(data = dados.preditos, aes(y = pred, x = NAP,
                              col = as.factor(Beach))) + # retas de cada praia
  stat_function(fun = funi, col = "black", size = 2) +  # reta do modelo
fixo
  scale_color_brewer(palette = "Set1") + # a partir daqui estética do
gráfico
  theme_bw() +
  theme(axis.text = element_text(size = 13),
        axis.title = element_text(size = 15),
        axis.text.x = element_text(size = 11),
        axis.text.y = element_text(size = 11),
        legend.key.size = unit(0.6, "cm"),
        legend.text = element_text(size = 12),
        legend.title = element_text(size = 13)) +
  xlab("NAP") +
  ylab("Riqueza")
```

Podemos ver nessa figura a predição do nosso modelo em relação aos parâmetros fixos (reta em preto), e as predições para cada praia separadamente. Como o efeito aleatório do nosso modelo estava apenas variando os valores de intercepto das praias, as retas para cada praia são paralelas.

Existem, entretanto, diversas maneiras de criar seu modelo, escolher os parâmetros importantes, e decidir quais são fixos e quais são aleatórios. Nesse exemplo, poderíamos colocar a variável `Exposure` também como um efeito fixo. Outra complicação que poderíamos inserir no nosso modelo é inserir mais um efeito aleatório de interação entre a `praia` (efeito aleatório) e `NAP` (efeito fixo), fazendo com que cada praia possa também ter sua relação com a NAP (inclinações de reta diferentes).

Por isso, é importante testar diferentes modelos antes de decidir qual é o melhor. Na próxima parte abordaremos como selecionar o melhor modelo dadas todas as possibilidades.

Escolha dos efeitos aleatórios

Existem modelos e, portanto, perguntas e delineamentos amostrais, que requerem apenas um efeito aleatório para indicar o agrupamento dos dados. Entretanto, como colocado no final da seção anterior, há também modelos que podem incluir mais de um efeito aleatório. Esse é o caso da interação entre `praia` e `NAP` mencionada acima. Se olharmos para o primeiro gráfico, veremos que cada praia parece ter seu próprio intercepto e inclinação, o que torna plausível pensarmos que o modelo com a interação se ajuste melhor aos nosso dados.

Para modelos que podem ter mais de um efeito aleatório, Zuur et al. (2009), sugere um protocolo para a escolha da melhor estrutura de efeitos aleatórios. Vamos aos passos:

1. A primeira coisa a ser feita é ajustar um modelo com todos os efeitos fixos que estão sendo testados (modelo “completo”). No nosso exemplo tínhamos apenas `NAP`, mas vamos incluir `Exposure` (como fator) para exemplificar melhor, e vamos incluir a interação entre `NAP` e `Exposure`. Nosso modelo “completo” é:

```
Richness ~ fExposure * NAP + "efeito(s) aleatório(s)"
```

2. Depois que escolhemos o modelo “completo”, adicionamos com os efeitos aleatórios plausíveis. No nosso exemplo, escolhemos duas possibilidades de efeito aleatório, variações no intercepto entre praias (`(1|Beach)`) e a interação entre praia e NAP, resultando também na variação de inclinação entre praias (`(NAP|Beach)`). Podemos também ajustar um modelo sem efeito aleatório (usando a função `lm`) e ver se a inclusão do efeito aleatório resulta num melhor ajuste do modelo:

```
# os modelos possíveis do nosso exemplo
m0 <- lm(Richness ~ fExposure * NAP, data = dados) #sem efeito aleat.
m1 <- lmer(Richness ~ fExposure * NAP + (1|Beach), data = dados)
m2 <- lmer(Richness ~ fExposure * NAP + (NAP|Beach), data = dados)
```

Um ponto importante é que estes modelos serão ajustados utilizando uma forma diferente de estimação, ao invés da máxima verossimilhança (ML), vamos usar a **máxima verossimilhança restrita** (REML). Isso acontece porque a ML é enviesado para as estimativas de variância do modelo e a REML corrige este enviesamento. Na prática, nós não precisamos fazer nada, pois a função `lmer` já usa, por padrão, a REML (argumento `REML = T`).

3. Colocamos estes modelos ajustados para “concorrer” usando o Critério de Informação de Akaike (AIC), como um critério para a seleção de modelos (menor AIC, melhor modelo) (ver Burnham & Anderson 2002). Como temos poucos dados vamos usar o AICc - que é uma correção do AIC para pequeno tamanho amostral.

```
# usamos a função AICctab do pacote bbmle
library(bbmle)

AICctab(m0, m1,m2, base = T, weights = T)
```

Bom, agora sabemos que a melhor estrutura de efeito aleatório é apenas a variação no intercepto entre as praias. Assim, podemos prosseguir com a verificação dos efeitos fixos através de teste de hipóteses e/ou seleção de modelos.

Inferência e diagnóstico do modelo

Nessa parte, depois de já escolhermos a estrutura aleatória do nosso modelo, podemos averiguar qual a real influência dos efeitos fixos na riqueza de espécies. Vou apresentar duas abordagens de inferência, o Teste de Hipóteses através da comparação de modelos pela tabela de ANOVA. E, depois de selecionado o modelo que melhor se ajusta aos dados, vamos fazer o diagnóstico dos resíduos deste modelo para ver se ele atende às premissas de um modelo linear misto.

Teste de hipótese

Você pode perceber que o output da função `lmer` não dá as estatísticas t e o valor de P, dos parâmetros fixos do modelo como faz um `lm` (veja o summary do nosso primeiro modelo e compare com o `m0`). Isso porque a chamada “estatística Wald” não é recomendada para modelos mistos (sobre isso melhor olhar nas referências recomendadas). O que se faz para saber se uma variável é significativa ou não é construir modelos aninhados (ou seja, retirando um parâmetro do modelo com mais parâmetros) e comparando por uma tabela de Análise de Variância.

Nesse caso, precisamos ajustar nosso modelo por máxima verossimilhança (ML), pois é o indicado para compararmos modelos com diferentes efeitos fixos mas com mesmo efeito aleatório. Então colocamos o argumento `REML = F` no nosso modelo:

```
# modelo com interação entre Exposure e NAP
m1 <- lmer(Richness ~ fExposure * NAP + (1|Beach), data = dados, REML = F)

# modelo sem interação entre exposure e NAP
m3 <- lmer(Richness ~ fExposure + NAP + (1|Beach), data = dados, REML = F)
```

E aplicamos a função `anova` nos nossos modelos aninhados:

```
anova(m1,m3)
```

Como o resultado da comparação entre os modelos foi significativo, nós paramos por aqui, ou seja, o modelo com interação é significativamente diferente do modelo sem interação. O que podemos fazer

é assegurar que o modelo com interação é também diferente do modelo nulo (sem efeitos fixos):

```
# modelo nulo
m6 <- lmer(Richness ~ 1 + (1|Beach), data = dados, REML = F)

anova(m1,m6)
```

Sim! Então podemos concluir que tanto `Exposure` quanto `NAP` influenciam na riqueza de espécies. Caso não tivesse havido diferença entre o modelo com interação e sem, nós deveríamos prosseguir com a seleção de modelos fazendo a comparação entre o modelo com as duas variáveis e modelos com cada variável separadamente.

Vamos observar o resumo do nosso modelo selecionado:

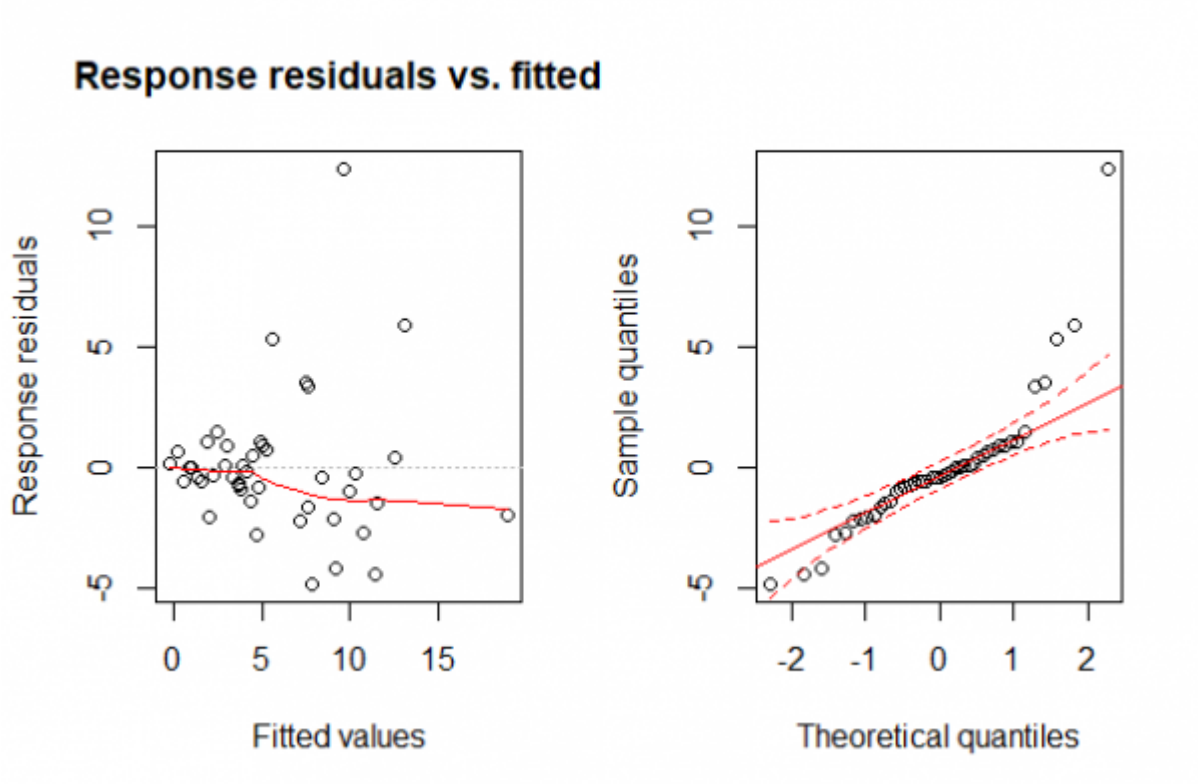
```
summary(m1)
```

Depois de selecionado o modelo que melhor se ajusta aos nossos dados, vamos avaliar o ajuste deste modelo e suas premissas.

Diagnósticos dos modelos

O primeiro diagnóstico do modelo é verificar se os resíduos são normalmente distribuídos, para isso geralmente usamos um qqplot (gráfico de quantil-quantil da distribuição normal) e possivelmente um teste de normalidade (por exemplo o Shapiro-Wilks). Além disso podemos também checar visualmente a homogeneidade de variância dos resíduos (homocedasticidade), ou seja, se a variablidade dos resíduos se mantém constante em relação aos valores ajustados.

Para isso usamos a função `plotresid` do pacote `RVAidememoire`:



```
library(RVAideMemoire)

plotresid(m1, shapiro = T)
```

> **OPS!** Olhando o gráfico da esquerda, vemos que os dados não são tão homocedásticos como gostaríamos, vemos algo parecido com um funil se abrindo da esquerda para a direita, o que indica que estamos violando esta premissa. Olhando o gráfico da direita (o qqplot), temos que os valores extremos dos dados não se comportam muito bem como uma distribuição normal (as linhas em vermelho no gráfico indicam a área em que os pontos deveriam estar para que pudéssemos considerar os resíduos como normalmente distribuídos). Isso fica evidente quando fazemos o teste de Shapiro e encontramos que os dados são significativamente diferentes de uma distribuição normal (nesse teste se dá significativo é que não é normal).
>
> **E agora??!** Bem, não estamos totalmente perdidos, existe um caminho! O problema foi que nós assumimos que a riqueza de espécies poderiam ser modelados como pertencentes a uma distribuição normal. Entretanto, dados de contagem (nesse caso, número de espécies) são geralmente modelados usando a distribuição de Poisson, que também leva em consideração que a média é igual à variância.
>
> Então, para fazermos a modelagem correta dos nossos dados teremos que usar um modelo com distribuição de Poisson, que está implementado no pacote `lme4` com a função `glmer`.
>
> Mas isso é tema para outro roteiro... (de GLMM!)

Se você se interessou pelos modelos mistos e acha que eles se encaixam no seu problema, não deixe de conferir as referências que a gente listou abaixo para se aprofundar nesse universo!!

Glossário

Efeitos fixos e aleatórios

Existe **muita** discussão na literatura sobre como se diferencia efeitos fixos de aleatórios (veja sugestões de leitura no final do roteiro - principalmente McGill 2015). De maneira geral, efeitos fixos são constantes em toda sua amostra, não variam de amostra pra amostra. Além disso, os diferentes níveis existentes no efeitos fixos não variam se você incluir mais amostras (Bates et al 2014). Um exemplo claro é sexo: normalmente sua amostra conterá machos e fêmeas, e aumentar o seu tamanho amostral a quantidade de níveis em seus fatores permanecerá constante. Já os efeitos aleatórios variam de amostra para amostra, por exemplo as praias que foram amostrados no dados que analisamos. Os pesquisadores poderiam ter ido a outras diferentes praias para coletar dados.

Máxima verossimilhança

Máxima verossimilhança (ML) é uma abordagem estatística que estima os parâmetros de um modelo a partir de um dado conjunto de dados.

Nos modelos mistos normais (que assume distribuição normal dos resíduos), utiliza-se a máxima verossimilhança restrita (REML) para estimar os parâmetros, pois a ML enviesa as estimativas de variância do modelo.

Modelo linear

Modelos lineares descrevem a resposta de uma variável dependente, a que você está interessado em explorar, em função de uma variável preditora, explanatória ou independente. $$dependente \sim preditora $$

Em ecologia, um dos usos mais comuns de modelos lineares é o modelo de regressão, por exemplo. Para fazer um modelo linear no R normalmente usamos a função `lm`:

```
lm(dependente ~ preditora, data = seus dados)
```

Modelo aninhado

É o tipo de modelo que utilizamos modelos mistos, no qual um modelo mais geral está aninhado dentro de outros modelos de modo que as variáveis independentes do modelo mais específico formam um subconjunto das variáveis do modelo mais geral. No exemplo das praias isso é facilmente visualizado dado que o conjunto de dados é composto de nove praias, e para cada uma das nove praias existem cinco amostras. O modelo linear simples não considera esse aninhamento dos dados e o fato de que, muito provavelmente, a riqueza em cada uma das cinco amostras está muito mais relacionada entre si do que entre praias.

Referências e recomendações

Bates, et al. 2014.**Fitting linear mixed-effects models using lme4**. arXiv preprint arXiv:1406.5823. (publicação do pacote `lme4`)

Burnham, K. & Anderson, D. 2002.](http://gen.lib.rus.ec/book/index.php?md5=0572C2F65088CFA05EC3757297DBC173) Model selection and multimodel inference: a practical information-theoretic approach. 2nd edn. New York: Springer-Verlag. (Livro sobre a abordagem de seleção de modelos baseada em Teoria da Informação)

McGill, B. 2015. **Is it a fixed or random effect?** Blog Dynamic Ecology. (Uma boa discussão sobre o que são efeitos fixos e aleatórios)

Winter, B. 2013.[http://arxiv.org/pdf/1308.5499.pdf| **Linear models and linear mixed effects models in R with linguistic applications**. arXiv:1308.5499.]] (Nesse excelente roteiro, o autor explica modelos lineares e depois apresenta modelos mistos de uma forma bem didática)

Zuur, A., Ieno, E., Walker, N., Saveliev, A. & Smith, G. 2009. Mixed effects models and extensions in ecology with R. (Livro muito bom e completo sobre modelos mistos e aditivos)

1)
esse valor predito é o que se considera ingênuo (**naive**), quando não estamos incorporando a variabilidade dos efeitos aleatórios. Como introdução é válido calcular os valores preditos desta maneira, mas para se aprofundar no tema sugerimos ler Bates et al. (2014) para formas mais apropriadas de predição.

[2)]

estamos usando o pacote `ggplot2` para fazer o gráfico, que tem uma sintaxe um pouco diferente de um gráfico do pacote base