

混合效应模型-第一部分

理论基础与应用

张 霜

中国科学院生态环境研究中心

2021-05-08



北京凌云翼数据科技有限公司

听完统计基础，是不是感觉各种问题都Under control，
按奈不住，跃跃欲试了？
那么问题来了！

先请大家思考一个问题

问题：想知道一片森林中的常见乔木叶片的虫食程度（%）如何？

方法：对该森林中的12种常见乔木的叶片损伤比率Y（每个物种不止测量一次，但不同物种次数有差异）进行了测量，共95个测量值。

问：该森林中常见乔木叶损伤均值该如何计算？



id	species	y
1	Acacia_obtusata	11.67
2	Acacia_obtusata	14.85
3	Acacia_obtusata	18.13
4	Acanthus_ilicifolius	9
5	Acanthus_ilicifolius	6.8
6	Acanthus_ilicifolius	7.3
7	Acer_platanoides	6.7
8	Acer_platanoides	9.05
9	Acer_platanoides	13.8
10	Acer_platanoides	29.4
11	Acer_platanoides	15.9
12	Acer_platanoides	3.9
13	Acer_platanoides	24.75
14	Acer_platanoides	3.2
15	Acer_platanoides	1.1
16	Acer_platanoides	1.3
17	Acer_platanoides	1.05
18	Acer_platanoides	0.5
19	Acer_platanoides	0.45
20	Acer_platanoides	0.5
21	Acer_platanoides	2.1
22	Acer_platanoides	24.4

文件名：Herbivory sample data

方案1: (Complete-pooling) 认为95个测量值彼此独立

```
fm1<-lm(y~1, data=h1)
summary(fm1)
```

完全忽视物种的影响

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	7.3142	0.7372	9.921	2.7e-16 ***

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 7.185 on 94 degrees of freedom
```

```
> mean(h1$y)
[1] 7.314211
```

缺点，违反了数据独立性原则！！！！

方案2: (No-pooling)

先把属于同一个物种的测量值平均, 然后再求平均值的均值

```
library(plyr)
###首先对同属一个物种的值求平均
h2<-ddply(h1, .(species), summarize, mean=mean(y))
###然后求这些均值的均值
fm2<-lm(mean~1, data=h2)
summary(fm2)

Call:
lm(formula = mean ~ 1, data = h2)

Residuals:
    Min       1Q   Median       3Q      Max
-5.8614 -3.5328 -0.8171  2.2552  9.2211

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    8.541      1.350   6.327 5.62e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.677 on 11 degrees of freedom
```

缺点, 一个物种不管有多少个测量值, 都被归结为一个点 (信息丢失), 对总体均值的贡献相同, 这不公平!

样本量急剧降低, 统计power下降

```
> mean(h2$mean)
[1] 8.541366
```

方案三： 把物种当做一个因素，拿叶损伤比率和物种做回归

```
fm3<-lm(herbivory~species, data=h1)
summary(fm3)
```

```
Call:
lm(formula = herbivory ~ species, data = h1)

Residuals:
    Min       1Q   Median       3Q      Max
-10.351  -4.477  -0.900   1.772  22.895

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    14.883     3.782   3.936 0.000172 ***
speciesAcanthus_ilicifolius -7.183     5.348  -1.343 0.182874
speciesAcer_platanoides    -8.378     3.960  -2.116 0.037377 *
speciesAcer_rubrum        -7.135     4.520  -1.579 0.118248
speciesAcer_saccharum    -11.160     4.102  -2.721 0.007930 **
speciesAegialitis_annulata -9.683     5.979  -1.619 0.109135
speciesAegiceras_corniculatum -4.905     4.632  -1.059 0.292649
speciesAlepis_flavida    -12.203     5.348  -2.282 0.025058 *
speciesAlnus_glutinosa    -6.253     5.348  -1.169 0.245636
speciesAlnus_incana       -1.632     4.520  -0.361 0.718979
speciesAlnus_viridis_fruticosa -10.449     4.367  -2.393 0.018973 *
speciesAngophora_hispida   2.879     5.003   0.576 0.566487
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.55 on 83 degrees of freedom
Multiple R-squared:  0.2663,    Adjusted R-squared:  0.1691
F-statistic: 2.739 on 11 and 83 DF,  p-value: 0.00453
```

```
fm4<-lm(herbivory~species-1, data=h1)
summary(fm4)
```

```
Call:
lm(formula = herbivory ~ species - 1, data = h1)

Residuals:
    Min       1Q   Median       3Q      Max
-10.351  -4.477  -0.900   1.772  22.895

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
speciesAcacia_obtusata    14.883     3.782   3.936 0.000172 ***
speciesAcanthus_ilicifolius  7.700     3.782   2.036 0.044922 *
speciesAcer_platanoides    6.505     1.176   5.529 3.64e-07 ***
speciesAcer_rubrum         7.749     2.476   3.130 0.002414 **
speciesAcer_saccharum      3.723     1.589   2.344 0.021493 *
speciesAegialitis_annulata  5.200     4.632   1.123 0.264783
speciesAegiceras_corniculatum 9.978     2.674   3.732 0.000347 ***
speciesAlepis_flavida      2.680     3.782   0.709 0.480500
speciesAlnus_glutinosa      8.630     3.782   2.282 0.025043 *
speciesAlnus_incana       13.251     2.476   5.353 7.55e-07 ***
speciesAlnus_viridis_fruticosa 4.434     2.183   2.031 0.045449 *
speciesAngophora_hispida   17.762     3.275   5.424 5.64e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.55 on 83 degrees of freedom
Multiple R-squared:  0.6416,    Adjusted R-squared:  0.5898
F-statistic: 12.38 on 12 and 83 DF,  p-value: 5.407e-14
```

- 1) 只能得到各个物种的均值，无法得到总体平均值，答非所问！
- 2) 估计的参数太多，会增大对单个物种均值的参数估计的误差

那么，这个问题该如何解决
混合效应模型—闪亮登场！！！！

问题
完美解决

这就是混合效应模型的魅力！！！！

同时还能给出对各个物种的Y的均值的估计

```
library(lme4)
library(lmerTest)
fm5<-lmer(y~1+(1|species), data=h1)
summary(fm5)
```

```
> summary(fm5)
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: herbivory ~ 1 + (1 | species)
Data: h1
```

REML criterion at convergence: 635.6

Scaled residuals:

Min	1Q	Median	3Q	Max
-1.3273	-0.7279	-0.1871	0.3344	3.4661

Random effects:

Groups	Name	Variance	Std.Dev.
species	(Intercept)	12.21	3.494
Residual		42.94	6.553

Number of obs: 95, groups: species, 12

Fixed effects:

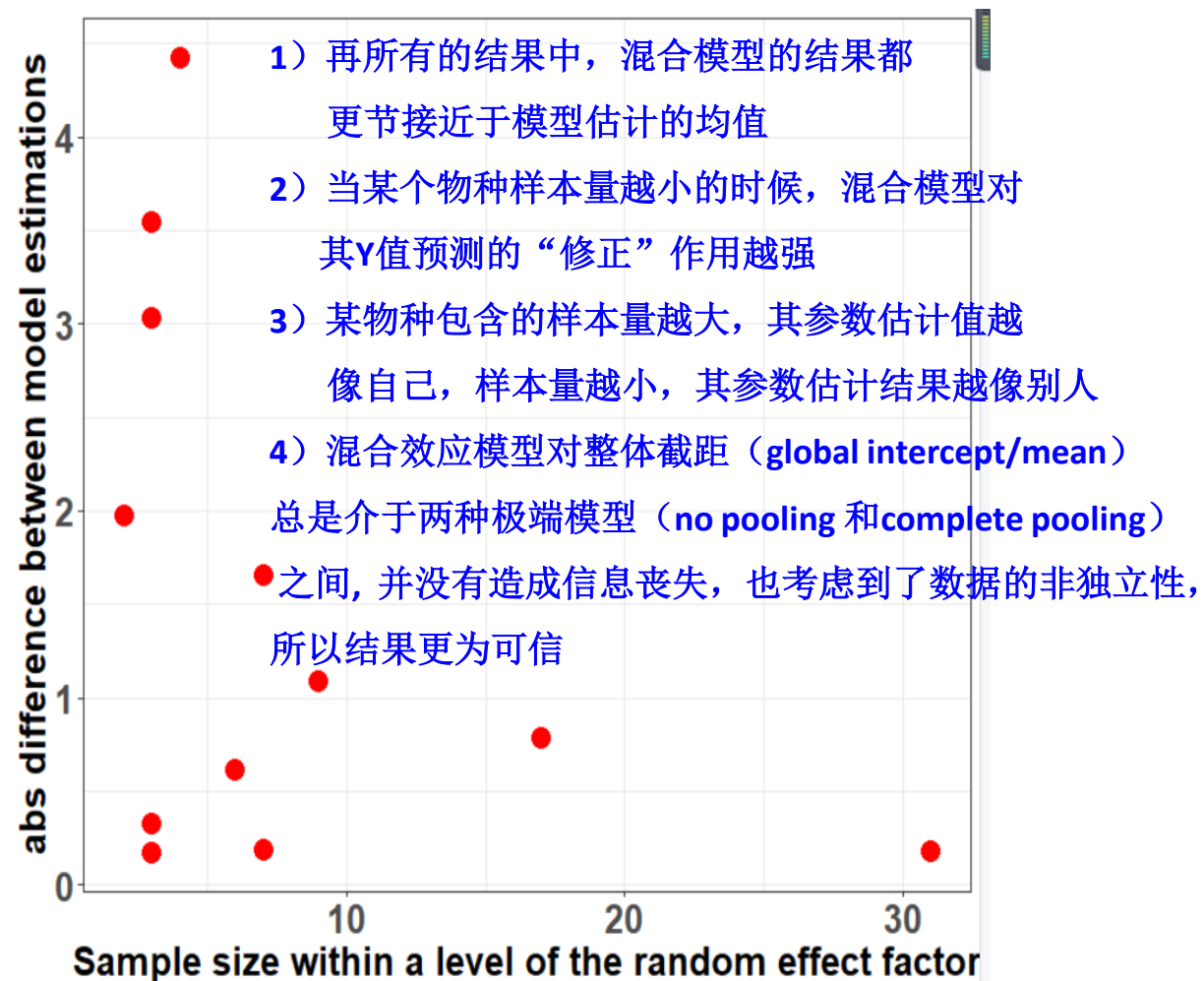
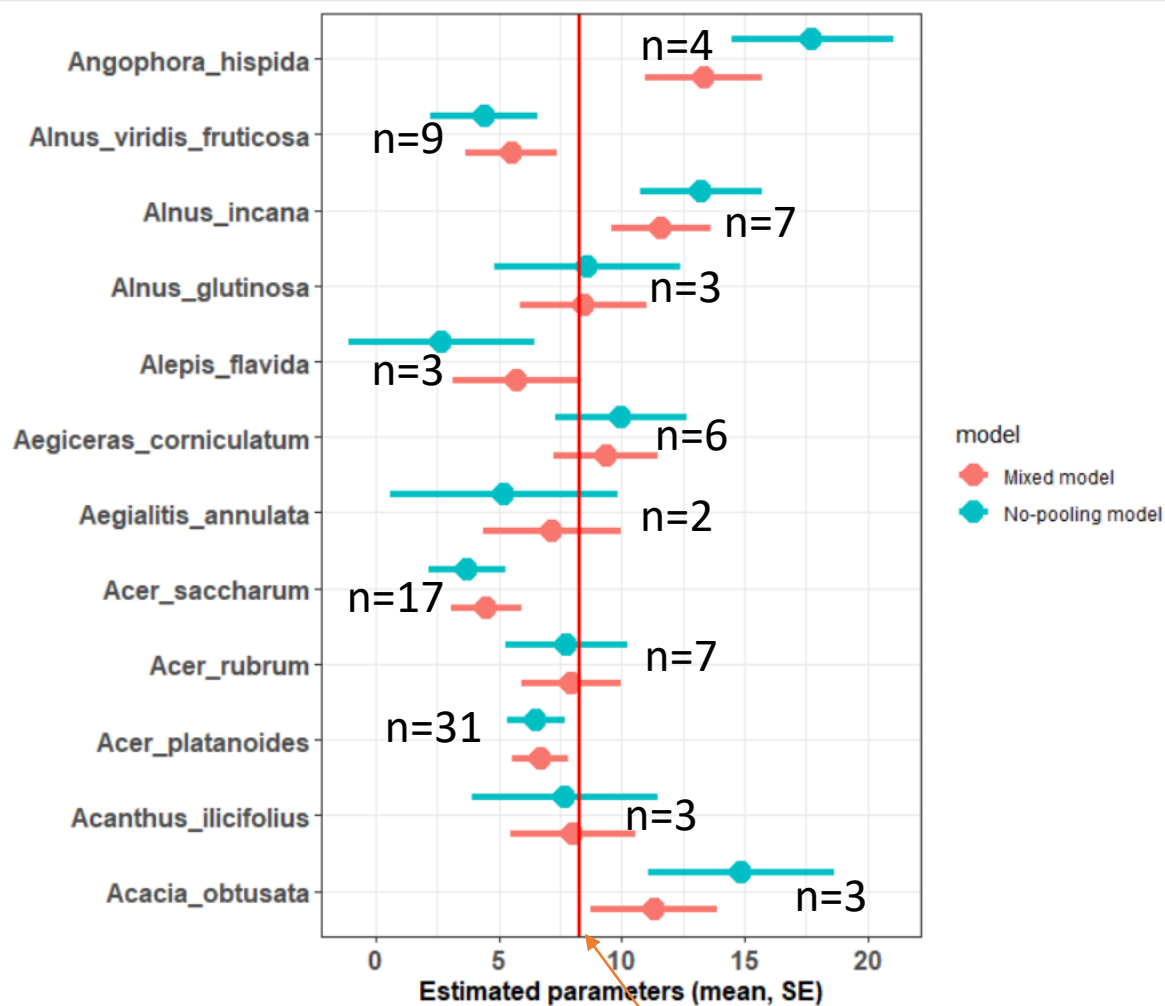
	Estimate	Std. Error	df	t value	Pr(> t)
(Intercept)	8.305	1.307	9.581	6.352	0.000101 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```
> coef(fm5)
$species
(Intercept)
Acacia_obtusata      11.333372
Acanthus_ilicifolius  8.026300
Acer_platanoides      6.688198
Acer_rubrum           7.934519
Acer_saccharum        4.508216
Aegialitis_annulata   7.179054
Aegiceras_corniculatum 9.359912
Alepis_flavida        5.715187
Alnus_glutinosa       8.454455
Alnus_incana          11.597391
Alnus_viridis_fruticosa 5.521750
Angophora_hispida     13.337888
```

给出Y的总体估计值，回答开头提出的问题

混合效应模型—partial pooling造成shrinkage



Mixed 模型均值, 混合模型的结果距离globe 均值的结果总是更近，这就是shrinkage

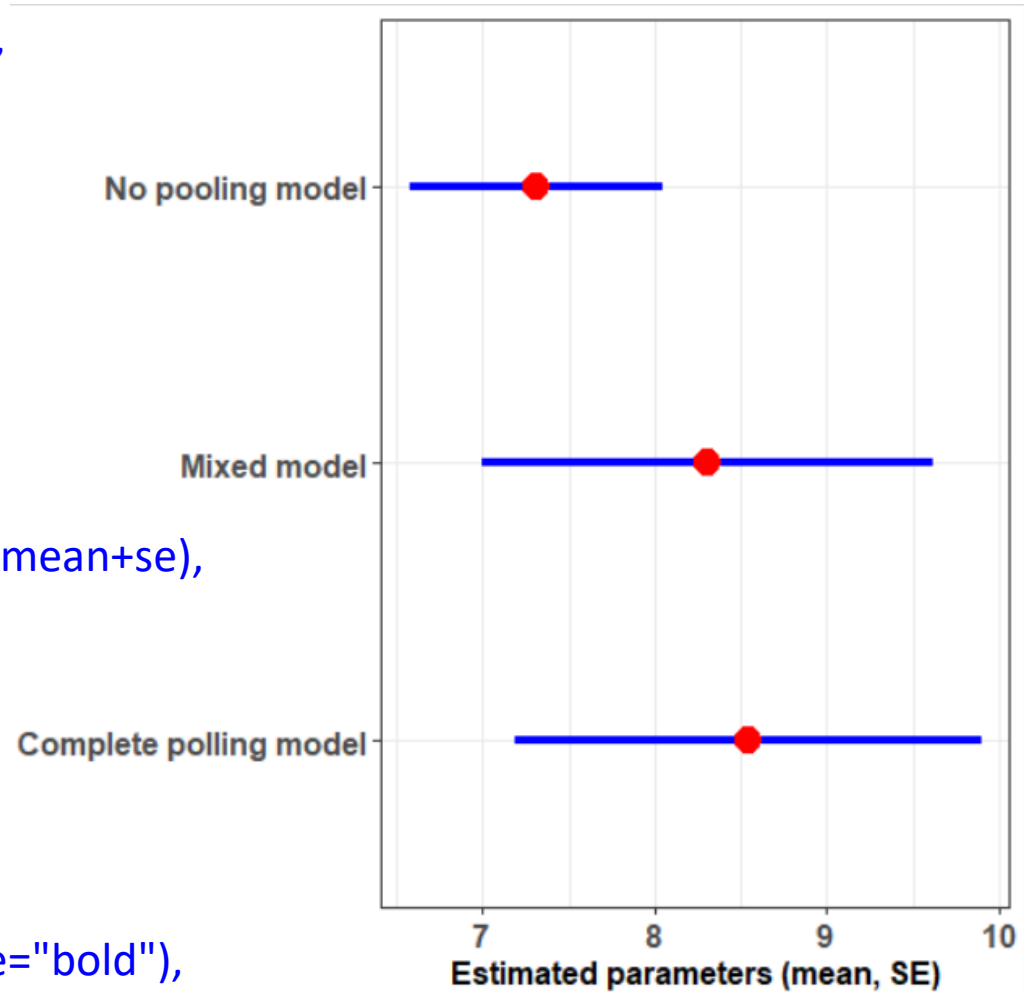
对比几种方案的结果

###根据结果构造一个新数据集

```
rd1 <- data.frame(model =c('No pooling model', 'Complete  
polling model', 'Mixed model'),  
                  mean=c(coef(summary(fm1))[1,1],  
                        coef(summary(fm2))[1,1],  
                        coef(summary(fm5))[1,1]),  
                  se=c(coef(summary(fm1))[1,2],  
                      coef(summary(fm2))[1,2],  
                      coef(summary(fm5))[1,2]))
```

###依据这个新数据集作图

```
library(ggplot2)  
ggplot(rd1, aes(x=model, y=mean)) +  
  geom_errorbar(aes(ymin=mean-se, ymax=mean+se),  
width=0,size=1.5, color="blue") +  
  geom_point(size=4.5,color="red")+  
  coord_flip() +  
  ylab("Estimated parameters (mean, SE)") +  
  xlab("") +  
  theme_bw()+  
  theme(axis.text=element_text(size=12,face="bold"),  
        axis.title=element_text(size=12,face="bold"))
```



混合效应模型均值估计的结果总是位于两种极端模型之间

现实中的数据结构

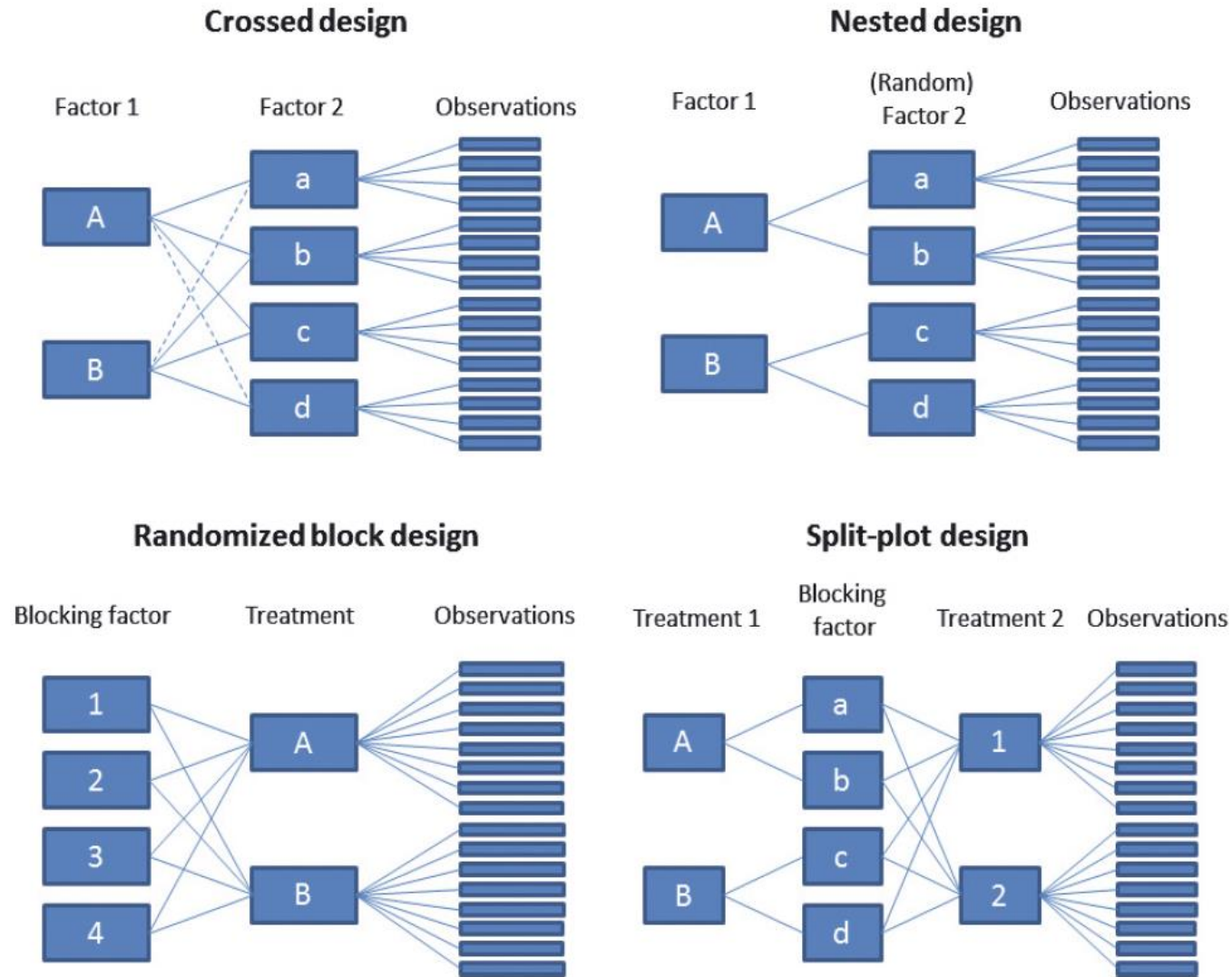
嵌套型（分层、多水平）--违反 y_i 独立性原则

- 来自同一个单元的不同数据点之间彼此更为相似，彼此不独立。
- ✓ 来自同一地点(如样方)的数据
- ✓ 来自同一物种的数据
- ✓ 同一班级，不同同学的数据
- ✓ 同一对象，不同时间测量的数据
- ✓ 来自同一篇文献的数据
- ✓ 同一批次样品的数据
- ✓ 受亲缘关系影响的数据

.....

嵌套结构几乎无处不在，传统回归模型对此无能为力，只能视而不见

常用实验设计中的数据结构分层



Schielzeth, H. & Nakagawa, S. Nested by design: model fitting and interpretation in a mixed model era. *Methods in Ecology and Evolution* **4**, 14-24, doi:<https://doi.org/10.1111/j.2041-210x.2012.00251.x> (2013).

传统回归模型和混合模型之间的关系

嵌套

- 线性模型
(linear regression model)
- 加和模型
(additive model)
- 广义线性模型
(generalized linear model)
- 广义可加模型
(generalized additive model)

- 嵌套数据
- 时间相关
- 空间相关
- 异质性
- 重复测量

当传统模型能处理这些问题时，就变成了混合效应模型

- 线性混合模型
(linear mixed regression model)
- 线性加和混合模型
(additive mixed model)
- 广义线性混合模型
(generalized linear mixed model)
- 广义可加混合模型
(generalized additive mixed model)

混合模型

几乎已经全面接管传统生态学领域的数据分析任务

SHARE REPORT



Electric field control of natural optical activity in a multiferroic helimagnet

Ryoji Masuda¹, Yoshio Kaneko², Yoshinori Tokura^{1,2,3}, Youtarou Takahashi^{1,2,*}

+ See all authors and affiliations

Science 30 Apr 2021:
Vol. 372, Issue 6541, pp. 496-500
DOI: 10.1126/science.aaz4312

2021年4月30号



Science

Vol 372, Issue 6541
30 April 2021

Table of Contents
Print Table of Contents
Advertising (PDF)
Classified (PDF)
Masthead (PDF)

We then used **generalized linear mixed effect models** in the R package ‘glmmTMB’ to analyze how resource types (i.e. equid wells, background waters, and dry controls) and environmental variables influenced vertebrate activity. We chose distributions appropriate for each response variable and **nested date within site as a random effect**.

混合模型

几乎已经全面接管传统生态学领域的数据分析任务



2021年4月16号



We fit linear mixed effects models (LMMs) in R using the package ‘nlme,... We modelled burn depth and % SOL C combusted as a function of successional trajectory and included a random intercept of fire perimeter (3 levels) to account for spatial non-independence of sites within each fire.

- tested the significance of fixed effects and interactions through model reduction using likelihood ratio tests of the full model against the reduced models and confirmed their importance using Akaike information criteria (AIC) ;
- visually inspected residual plots to ensure assumptions of heteroscedasticity and normality were met;
- fit the final model with restricted maximum likelihood;
- For the LMM of aboveground C we also included a variance structure (varExp) to account for increasing variance with age that was apparent in the residuals.
- calculated marginal R² (accounting only for fixed effects) and conditional R² values (accounting for random and fixed effects) using the package ‘MuMIn’ ;
- completed pairwise comparisons of marginal mean intercepts and slopes for each trajectory using a Tukey-Kramer post hoc analyses in the ‘emmeans’ package.

混合模型

几乎已经全面接管传统生态学领域的数据分析任务

nature

[Explore content](#) ▾ [Journal information](#) ▾ [Publish with us](#) ▾

[nature](#) > [articles](#) > [article](#)

2021年3月1号

Article | [Published: 01 March 2021](#)

Warming impairs trophic transfer efficiency in a long-term field experiment

Diego R. Barneche, Chris J. Hulatt, Matteo Dossena, Daniel Padfield, Guy Woodward, Mark Trimmer  & Gabriel Yvon-Durocher 

Nature **592**, 76–79(2021) | [Cite this article](#)

6362 [Accesses](#) | **410** [Altmetric](#) | [Metrics](#)

A **Bayesian hierarchical model**, which accounted for repeated measures throughout the experiment, revealed that...

混合模型

几乎已经全面接管传统生态学领域的数据分析任务



RESEARCH ARTICLE

2021年4月6号



Invasiveness is linked to greater commercial success in the global pet trade

Jérôme M. W. Gippet and Cleo Bertelsmeier

[+ See all authors and affiliations](#)

PNAS April 6, 2021 118 (14) e2016337118; <https://doi.org/10.1073/pnas.2016337118>

Edited by Nils Chr. Stenseth, University of Oslo, Oslo, Norway, and approved February 11, 2021 (received for review August 3, 2020)

We used a negative binomial generalized linear mixed model (glmmTMB function in R package glmmTMB) to test ..., while accounting for species' geographical origin (i.e., in which ecozone species occur) and controlling for phylogenetic effects by using ants' superfamily and genus as nested random effects.

混合模型

几乎已经全面接管传统生态学领域的数据分析任务



RESEARCH ARTICLE

Global tree intrinsic water use efficiency is enhanced by increased atmospheric CO₂ and modulated by climate and plant functional types

 Justin M. Mathias and  Richard B. Thomas

[+ See all authors and affiliations](#)

PNAS February 16, 2021 118 (7) e2014286118; <https://doi.org/10.1073/pnas.2014286118>

Edited by Elizabeth Ainsworth, Agricultural Research Service, Urbana, IL, and approved January 4, 2021 (received for review July 7, 2020)



To address this knowledge gap, we used **linear mixed effects (LME) models** to examine, **with tree species nested within site as random factors**. temporal trends were fit via restricted maximum likelihood and included a **first-order autocorrelation structure** to account for temporal autocorrelation in the chronologies.

2021年2月16号

混合模型

几乎已经全面接管传统生态学领域的数据分析任务

Science Contents ▾ News ▾ Careers ▾ Journals ▾

[Read our COVID-19 research and news.](#)

SHARE **REPORT**

Novel trophic interactions under climate change promote alpine plant coexistence

 Patrice Descombes^{1,3},  Camille Potteluz¹,  Gaëtan Glauser⁴, Emmanuel Defossez⁵, Alan Kergunteuil⁶,  Pierre...

[+ See all authors and affiliations](#)

Science 18 Dec 2020:
Vol. 370, Issue 6523, pp. 1469-1473
DOI: 10.1126/science.abd7015

2020年12月18号

using a linear mixed-effects model fitted by maximum likelihood with the “lme” function implemented in the R package nlme. Treatment and the elevation of the site were set as fixed effects, and site and plot identity as random effects.

Mixed effect model, 当前最为强力的回归模型



Science

Vol 368, Issue 6492

15 May 2020

[Table of Contents](#)

[Print Table of Contents](#)

[Advertising \(PDF\)](#)

[Classified \(PDF\)](#)

[Masthead \(PDF\)](#)

we applied a multilevel-modelling framework with group-level effects for ‘region’ to account for the **spatial non-independence of plots within the same region, using linear mixed-effects models(LMMs)**. Goodness-of-fit was determined by calculating the marginal coefficient of determination (R-squared) as proposed by Nakagawa & Schielzeth (55) using the r.squaredGLMM function in the MuMIn package (56). The **marginal R-squared** describes the variation explained by the fixed factor only, as opposed to the **conditional R-squared**, which describes the variation explained by the fixed and random factors together

2020年5月15号Science封面文章

Mixed effect model, 当前最为强力的回归模型

We used **autoregressive mixed effects models** to assess...

We included a series of random intercepts for:

- dataset
- study area (in cases when sites were clustered in different study areas of the same dataset)
- site (the smallest reported sampling unit)
- within-year time-period (finest resolution: month; when samples were collected repeatedly within year, nested within dataset).

We included random slopes for:

- the effect of year (at the levels of dataset, study area and site)
- temporal autocorrelation by adding an autoregressive term of order 1 (AR1), on which we placed a random effect at the site-level to allow site-level variation in the strength of autocorrelation.



2020年4月24号Science封面文章

Mixed effect model, 当前最为强力的回归模型

SHARE

REPORT



Termites mitigate the effects of drought in tropical rainforest

L. A. Ashton^{1,2,3,*}, H. M. Griffiths^{4,*†}, C. L. Parr^{4,5,6}, T. A. Evans⁷, R. K. Didham^{7,8}, F. Hasan², Y. A. Teh⁹, H. S. Tin¹⁰, C. S. ...



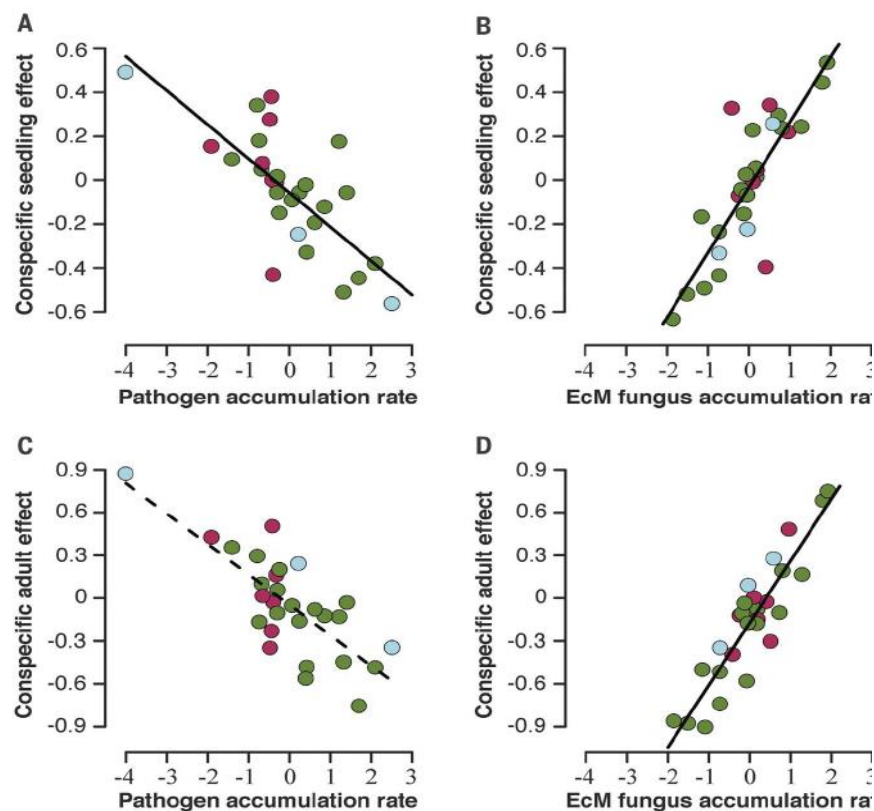
2019年1月11日
Science封面文章

Generalised linear mixed models with a Poisson error structure were used to test for any effect of termite suppression on non-target invertebrate abundance within leaf litter samples. Random slopes models were performed on each major taxonomic group with year | plot included as random factors.

Differential soil fungus accumulation and density dependence of trees in a subtropical forest

Lei Chen¹, Nathan G. Swenson², Niuniu Ji³, Xiangcheng Mi¹, Haibao Ren¹, Liangdong Guo³, Keping Ma^{1*}

We examined neighborhood effects on the...adopting a mixed modelling approach. ... We included log-transformed seedling height and all neighborhood variables in one model after standardizing as fixed effects, To control for temporal, spatial and interspecific variation, we included census year, species identity and 1-m² seedling plot nested in each census station as random factors. We also included seedling individual as a random effect because some individuals had multiple observations as they survived multiple intervals, ...Wald's z test and likelihood ratio tests (LRTs) were applied to determine the significant level of fixed and random effects, as suggested by...



SHARE

RESEARCH ARTICLE



The geography of biodiversity change in marine and terrestrial assemblages



要闻

娱乐

财经

体育

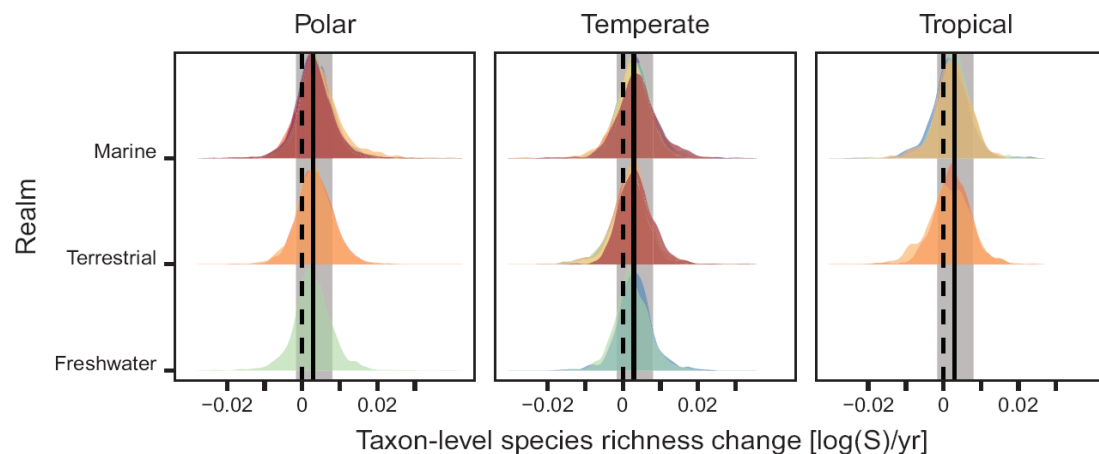
图片

时尚



第六次物种灭绝已经来临？科学家：地球生物多样性正处于危机之中

人类活动导致的气候变化深远地影响了地球上的物种栖息地，并严重破坏了全球生物多样性，即所谓的“第六次物种大灭绝”已经来临，地球物种多样性正在经历快速却又难以被察觉的变化。对此，科学家警告称全球生物正处于危机之中，而首当其冲的便是海洋生物。



We examined geographic patterns of biodiversity change using three complementary *hierarchical linear models*. All models included *site nested in region as random effect* to account for *spatial arrangement and temporal repetitions per site*. *Poisson models* included *an observation-specific random effect* to account for potential *overdispersion*.

SHARE

REPORT



Extinction filters mediate the global effects of habitat fragmentation on animals

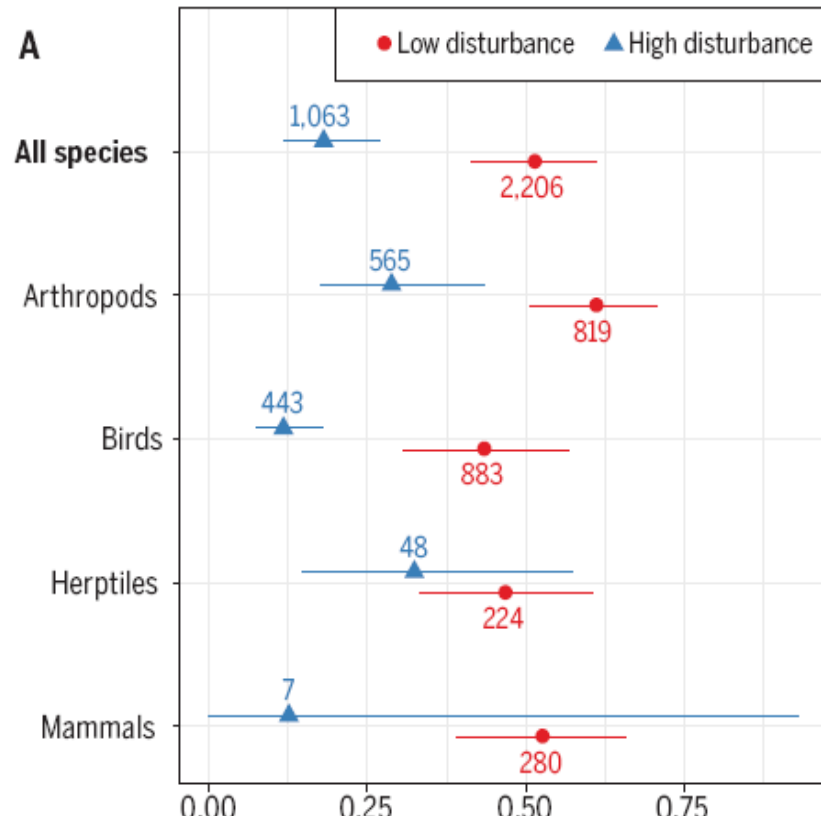


Fig. 2. The proportion of forest species associated with core habitat is mitigated by historical exposure to disturbance. (A) Estimated proportions (with 95% confidence intervals) are based on mixed-effects logistic regression models. The binary disturbance variable (low- versus high-disturbance sites) indicates whether each of the 73 BIOFRAG datasets comes from a location that has had high-severity disturbances of any type (glaciation, tropical storms, crown fires, or >50% historical forest loss). Numbers of species are shown beside point

中国移动 HD 4G+ 信号强度 网速 电量 时间 应用图标

半夜 11:26



“生于忧患，死于安乐”：灭绝过滤假说在保护生态学中的应用 | Science 述评

原创：刘宣 ScienceAAAS 5天前

Estimated proportions
(with 95% confidence intervals) are based
on *mixed-effects logistic regression*
Models..

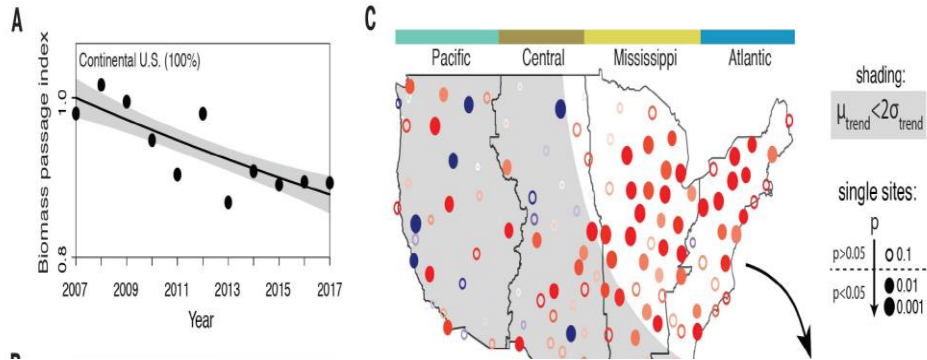
Betts *et al.*, *Science* **366**, 1236–1239 (2019)

6 December 2019

BIODIVERSITY LOSS

Decline of the North American avifauna

Kenneth V. Rosenberg^{1,2*}, Adriaan M. Dokter¹, Peter J. Blancher³, John R. Sauer⁴, Adam C. Smith⁵, Paul A. Smith³, Jessica C. Stanton⁶, Arvind Panjabi⁷, Laura Helft¹, Michael Parr², Peter P. Marra^{8†}



We fitted **generalized linear mixed models** using R package **lme4**, including **radar station as a random offset**, and **region and the interaction year:region as fixed effects...**

Model*	Formula	AIC	df	
5	index ~ region + year:flyway + (1 radar) + dualpol [†]	338	11	a
6	index ~ region + year:flyway + (1 radar) + mode [‡]	340	12	a
7	Index ~ region + year:flyway + (1 radar) + superres	343	11	b
8	Index ~ region + year:flyway + (1 radar)	361	10	c

*Family=Gamma(link=log)

[‡]mode is a factor variable with levels “legacy”, “superres” and “dualpol”, distinguishing the three time periods in which the radar acquired legacy, super-resolution and dual-polarization data. Note that the dual-polarization upgrade

Models significantly different according to a Chi-squared likelihood ratio test are labelled by different letters

《科学》公布 2019 年度 “十大突破” 与 “四个退步”

有色技术平台 2019-12-20 16:44:13

“四大退步” 之一

但是，越强力的模型，越需要正确使用

Modern mixed effect models offer an unprecedented opportunity to explore complex biological problems by explicitly modelling non-Normal data structures and/or nonindependence among observational units. **However, the LMM and GLMM toolset should be used with caution.**

Harrison, X. A., L. Donaldson, M. E. Correa-Cano, J. Evans, D. N. Fisher, C. E. D. Goodwin, B. S. Robinson, D. J. Hodgson, and R. Inger. 2018. A brief introduction to mixed effects modelling and multi-model inference in ecology. PeerJ 6:e4794.

混合效应模型—使用有风险

1. Arnqvist, G. 2020. Mixed models offer no freedom from degrees of freedom. *Trends in Ecology & Evolution* **35**:329-335.
2. Silk, M. J., X. A. Harrison, and D. J. Hodgson. 2020. Perils and pitfalls of mixed-effects regression models in biology. *PeerJ* **8**:e9522.
3. Harrison, X. A., L. Donaldson, M. E. Correa-Cano, J. Evans, D. N. Fisher, C. E. D. Goodwin, B. S. Robinson, D. J. Hodgson, and R. Inger. 2018. A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ* **6**:e4794.
4. Zuur, A. F. and E. N. Ieno. 2016. A protocol for conducting and presenting results of regression-type analyses. *Methods in Ecology and Evolution* **7**:636-645.
5. Bolker, B. M., M. E. Brooks, C. J. Clark, S. W. Geange, J. R. Poulsen, M. H. H. Stevens, and J.-S. S. White. 2009. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution* **24**:127-135.
6. Zuur, A. F., E. N. Ieno, and C. S. Elphick. 2010. A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution* **1**:3-14.

Generalized linear mixed models: a practical guide for ecology and evolution

[BM Bolker](#), [ME Brooks](#), [CJ Clark](#), [SW Geange](#)... - *Trends in ecology & ...*, 2009 - Elsevier

How should ecologists and evolutionary biologists analyze nonnormal data that involve random effects? Nonnormal data such as counts or proportions often defy classical statistical procedures. Generalized linear mixed models (GLMMs) provide a more flexible approach for analyzing nonnormal data when random effects are present. The explosion of research on GLMMs in the last decade has generated considerable uncertainty for practitioners in ecology and evolution. Despite the availability of accurate techniques for estimating GLMM ...

☆ 被引用次数: 7022 相关文章 所有 41 个版本

[\[PDF\] u-psud.fr](#)

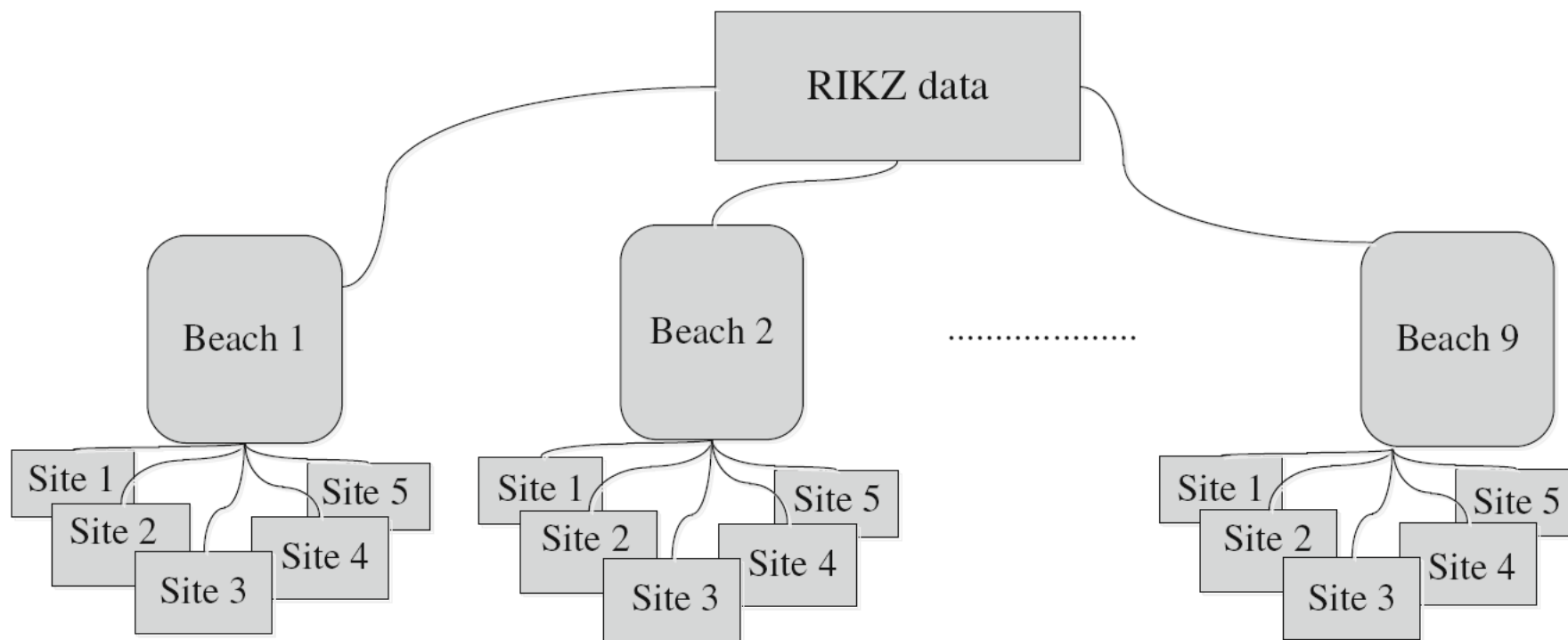
混合模型的优势

- 1) 解决数据的不独立性问题，partial pooling 结果更可靠
- 2) 提高对group水平参数估计的准确性
- 3) 可以估计不同组分的方差
- 4) 对随机效应的未测量水平的值进行估计
- 5) 可以考虑X的层次性对Y的影响的交互作用

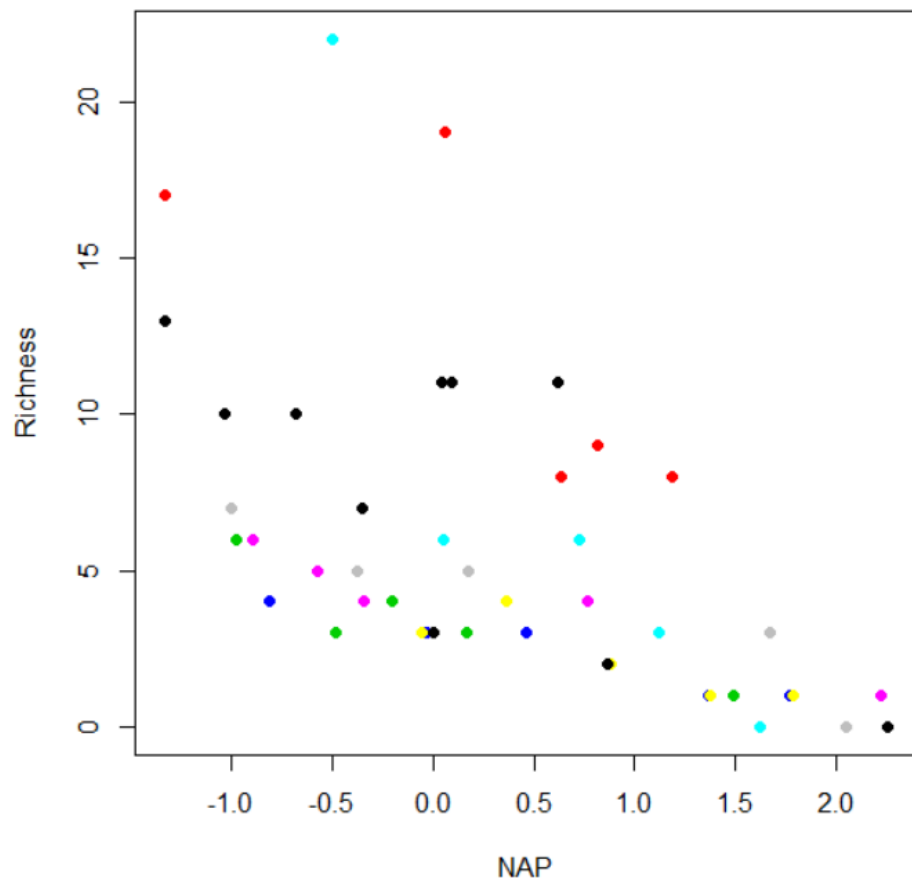
.....

还有很多

现实数据的常见结构—分层（嵌套）



生态数据的常见结构—分层（嵌套）

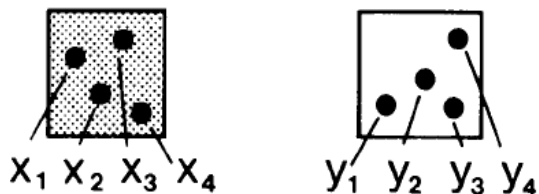


- ✓ 来自同一地点(如样方)的数据
- ✓ 来自同一物种的数据
- ✓ 同一班级，不同同学的数据
- ✓ 同一对象，不同时间测量的数据
- ✓ 来自同一篇文献的数据
- ✓ 同一批次样品的数据
- ✓ 受亲缘关系影响的数据
- ✓

数据单元不止一个水平
嵌套结构几乎无处不在

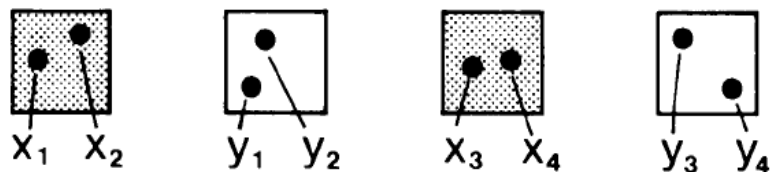
分层（或嵌套）的本质是伪重复问题

A. SIMPLE PSEUDOREPLICATION

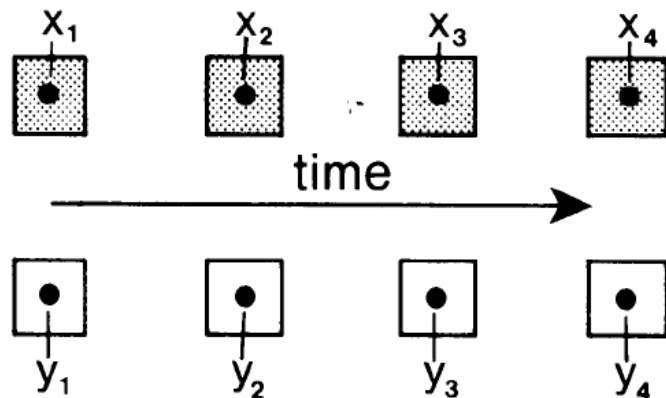


同一实验单元内的不同数据点彼此不独立

B. SACRIFICIAL PSEUDOREPLICATION



C. TEMPORAL PSEUDOREPLICATION



Hurlbert, S. H. 1984. Pseudoreplication and the design of ecological field experiments. *Ecological Monographs* **54**:187-211.

截至2021年5月7号，9203次引用

线性混合效应模型 VS 线性回归模型

- Mixed effect model/ Multilevel model/Hierarchical model
- 线性模型(截距, 斜率为常数, 固定值)

$$y_i = \alpha + \beta x_i + \varepsilon$$

- 混合效应模型 (随机截距模型)

$$y_i = \alpha_{j[i]} + \beta x_i + \epsilon_i$$

- 混合效应模型 (随机斜率, 随机截距斜率模型)

$$y_i = \alpha_{j[i]} + \beta_{j[i]} x_i + \epsilon_i$$

混合模型—为什么叫混合？

以随机截距混合模型为例

$$R_{ji} = \mu + \alpha_j + \beta \times NAP_{ji} + \varepsilon_{ji}$$

常数（固定效应）

随随机因素的水平而变化
（随机效应）

常数（固定效应）

- 依赖于因素（如Beach）而变化，其值 $\sim N(0, sd^2)$
- 即 α_j 服从均值为零，方差为 sd^2 的正态分布

混合效应模型(固定+随机=混合)

$$y_{ij} = \mu + \alpha_j + \beta x_{ji} + \varepsilon_{ji}$$

$$\text{其中 } \alpha_j \sim N(0, sd^2); \varepsilon_{ji} \sim N(0, \sigma^2)$$

公式也可以写成

$$\text{第一层 } y_{ji} = \textit{int}_j + \beta * x_{ji} + \varepsilon_{ji} \quad \varepsilon \sim N(0, \sigma^2)$$

$$\text{第二层 } \textit{int}_j = \mu + \alpha_j, \alpha_j \sim N(0, sd^2)$$

**Mixed effect model = multilevel model = Hierarchical model
=Nested model**

混合模型为何更可靠: shrinkage

SHRINKAGE

- Random effects assume each observation is drawn from a grand mean
- This we can **draw strength** from samples in other groups to estimate a group mean
- It **reduces variance** between groups by biasing results towards grand mean

Globe mean

Group mean

让group mean 的估计值向globe Mean 靠拢

https://biol609.github.io/lectures/07_random_effects.html#/shrinkage

混合模型：shrinkage 系数

混合模型对随机效应
不同水平下y的均值的估计

$$\hat{Y}_j = \rho_j \bar{y}_j + (1 - \rho_j) \bar{y}$$

where ρ_j is the shrinkage coefficient

shrinkage 系数

$$\rho_j = \frac{\tau^2}{\tau^2 + \frac{\sigma^2}{n_j}}$$

随机效应造成的方差

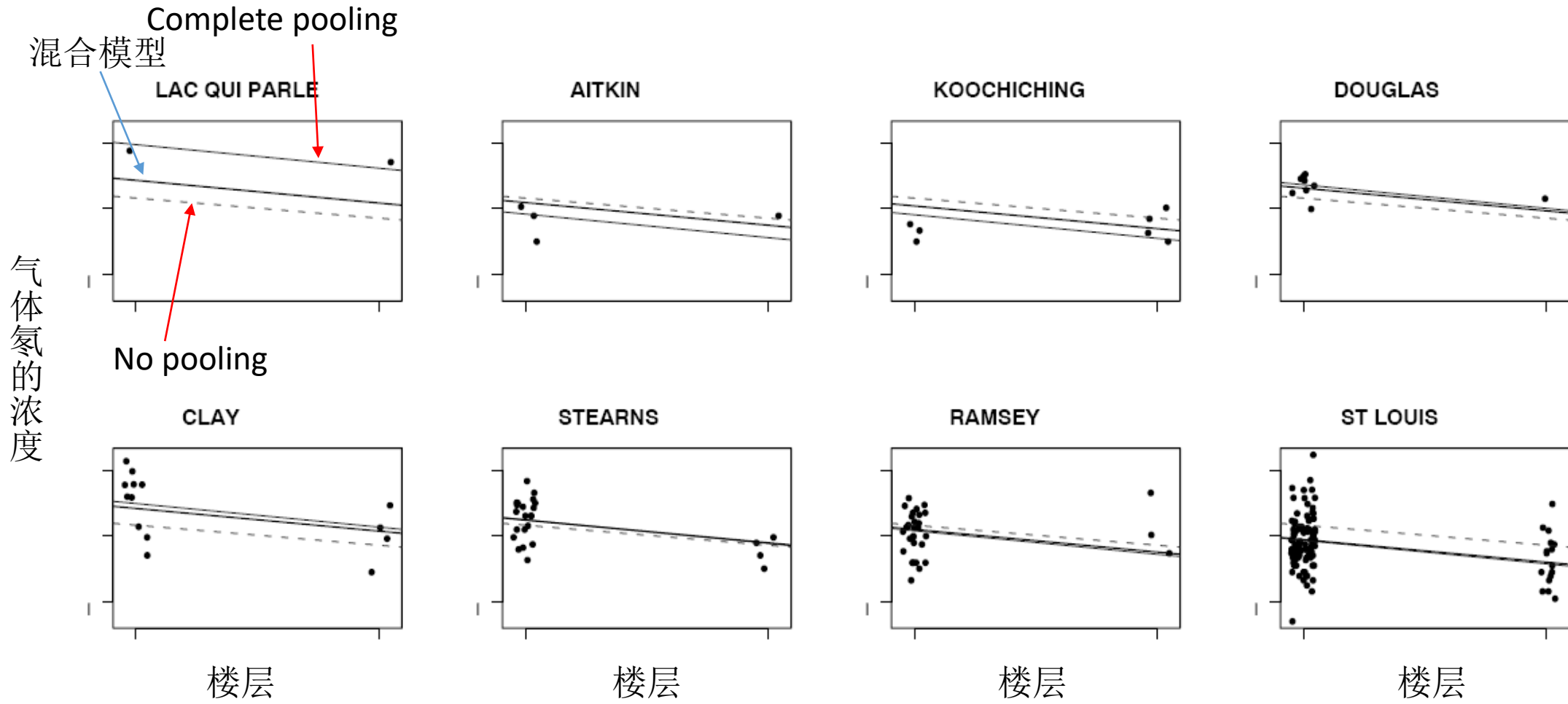
残差的方差

随机效应处于水平j时的样本量

Where τ^2 is the variance of the random effect, and n_j is the group sample size

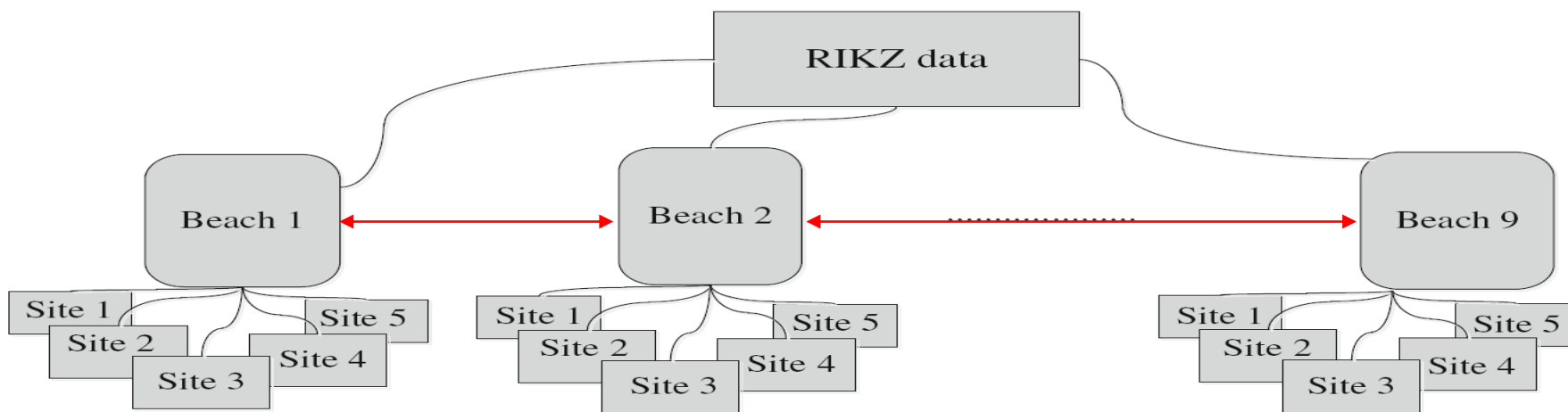
n_j 越大，shrinkage 系数越小，对 $group_j$ 的均值的估计就越接近其自身均值，
 n_j 越小，shrinkage 系数越大，对 $group_j$ 的均值的估计就越接近总体均值（globe mean）

混合效应模型参数估计的结果总是位于complete-pooling和no-pooling 之间，所以更为可信



哪些因素适合作为随机效应？

- 1) 水平多，甚至无限，实验用到的水平只不过是从无限多个水平中的抽样，如样方，样地，物种，文献，病人、学生、班级、学校、城市等等
- 2) 彼此可换性，不同水平意义对等，其数值都是从一个符合一定分布的总体中的随机抽样。
- 3) 不同水平间差异的来源和意义不明确



哪些因素不宜作为随机因素

➤具有明确意义的变量绝对不能作为随机因素

如温度，降雨，生物量，施肥量，距离，物种多样性，年龄

➤意义明确，因素水平数有限的分类变量

性别，是否是成年人，是否吃药，吃药方式，年龄阶段
是否增温，增温幅度，

➤水平数不超过五个的因素

比如，在一个研究中，只有两个物种、只有两个样地、只有三个样方

混合效应模型中，随机效应的设置至关重要，绝不可随意设置

哪些因素不宜作为随机因素

水平数不超过五个的因素

- Silk, M. J., Harrison, X. A. & Hodgson, D. J. Perils and pitfalls of mixed-effects regression models in biology. *PeerJ* **8**, e9522, doi:10.7717/peerj.9522 (2020).
- Arnqvist, G. Mixed models offer no freedom from degrees of freedom. *Trends Ecol. Evol.* **35**, 329-335 (2020).
- Harrison, X. A. *et al.* A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ* **6**, e4794, doi:10.7717/peerj.4794 (2018).
- Gomes, D. G. E. Including random effects in statistical models in ecology: fewer than five levels? *bioRxiv*, 2021.2004.2011.439357, doi:10.1101/2021.04.11.439357 (2021).

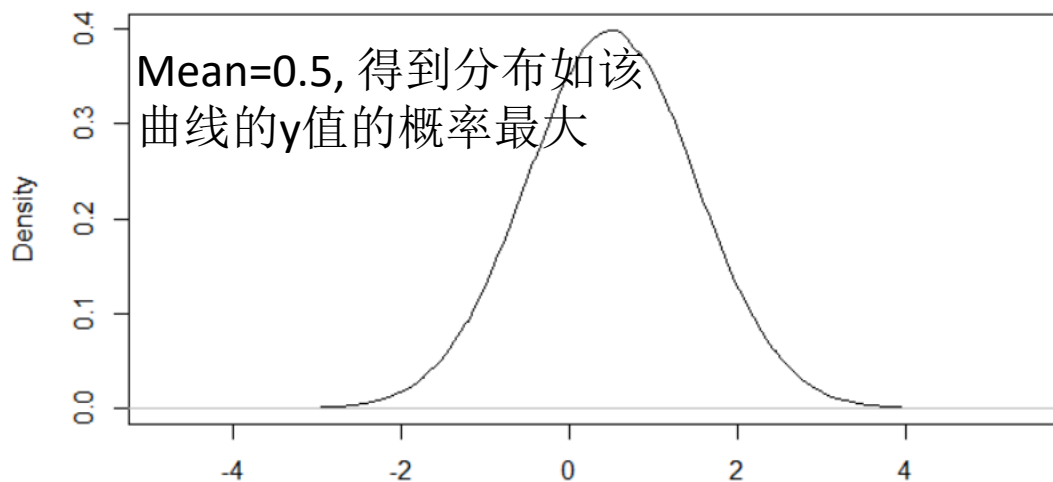
Mixed model 与Im model 的关系

- 随机效应的每个水平只含有一个数据时，混合效应模型的结果与线性回归完全相同，因为这时候不存在来自于随机效应的同一个水平的不同点之间的不独立问题。
- 如测量身高对血压的影响，随机选取了40个成年人，但对每个人只进行了一次身高和血压的测量，这时便没有必要设置人为随机效应。
- 所以一般线性回归只是混合效应模型的一个特例（即随机因素的每个水平只进行了一次测量）。

混合效应模型的参数估计方法

回归模型	参数估计方法
一般线性回归	最小二乘法
线性混合模型	最大似然法 (maximum likelihood, ML) , 限制性最大似然法 (restricted maximum likelihood, REML)
广义线性混合模型	Laplace approximation, Penalized quasi-likelihood等
任意回归模型	贝叶斯法 (Markov chain Monte Carlo)

Linear mixed effect model参数估计: 最大似然法(ML)与限制性最大似然(REML)



$$f_i(Y_i, X_i, \alpha, \beta, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(Y_i - \alpha - \beta \times X_i)^2}{2\sigma^2}}$$

ml:让取得所有观测值的可行性最大化时,模型的参数的值

似然值 (likelihood value) :
给定参数和 x , 取得当前观测到的 y 值的概率

最大似然法(ML): 参数为何值时, 给定 x , 得到当前所观测到取得当前观测到的 y 值的概率最大。

限制性最大似然法(REML), 参数估计过程中对自由度进行了校正。

Linear mixed effect model参数估计方法： 最大似然法(ML) 与限制性最大似然法(REML)

- 如果模型的误差相互独立，均质（随拟合值的变化没有明显格局），且符合正态分布，那么，最小二乘法与最大似然法的参数估计相同
- 混合模型中：如果随机效应嵌套，或只有一个随机效应，且数据结构平衡（随机效应每个水平包含的样本量相等），那么可用ML，否则都建议REML.

R中最常用的混合模型包

Should I use `aov()`, `nlme`, or `lme4`, or some other package?

- `aov()` (in the `stats` package in base R: balanced, orthogonal designs only (R analogue of SAS PROC GLM))
- `nlme` (analogue of SAS PROC MIXED / NLMIXED)
 - allows more complex designs than `aov` (unbalanced, heteroscedasticity and/or correlation among residual errors)
 - more mature than `lme4`
 - well-documented (Pinheiro and Bates 2000)
 - implements R-side effects (heteroscedasticity and correlation)
 - estimates “denominator degrees of freedom” for F statistics, and hence p values, for LMMs (but see above)
- `lme4` (also SAS PROC MIXED / NLMIXED):
 - fastest
 - best for crossed designs (although they are possible in `nlme`)
 - GLMMs
 - cutting-edge (for better or worse!)
 - likelihood profiles
 - use `lme4` for GLMMs, or if you have big data (thousands to tens of thousands of records)

混合效应模型的写法 (lme4包)

##线性模型

```
m0<-lm(Richness~NAP,data=RIKZ)
```

##混合模型

##lme4包--随机截距模型，

```
m1<-lmer(Richness~NAP+(1|fBeach),data=RIKZ, REML=T )
```

```
m1_1<-lmer(Richness~NAP+(1|fBeach),data=RIKZ)
```

```
m1_2<-lmer(Richness~NAP+(1|fBeach),data=RIKZ, REML=F)
```

lme4包--随机截距， 随机斜率模型

```
m1<-lmer(Richness~NAP+(NAP|fBeach),data=RIKZ)
```

Bates, D., M. Mächler, B. Bolker, and S. Walker. 2015. Fitting Linear Mixed- Effects Models using lme4. Journal of Statistical Software **67**:1-48. 截至2021年5月7日已被引用35971次（谷歌学术）

混合效应模型的写法（nlme包）

####nlme包中(随机截距模型)

```
m00<-lme(Richness~NAP, random=~1|Sample,data=RIKZ, method="REML")
```

####随机截距，随机斜率模型

```
m00<-lme(Richness~NAP, random=~NAP|Sample, data=RIKZ)
```

####问题，设置交叉型随机效应有困难

the R Development Core Team (2013) nlme: linear and nonlinear mixed effects models

随机效应的写法(lme4)

Formula	Alternative	Meaning	
$(1 g)$	$1+(1 g)$	Random intercept with fixed mean	随机斜率模型
$0+\text{offset}(o)+(1 g)$	$-1+\text{offset}(o)+(1 g)$	Random intercept with <i>a priori</i> means	
$(1 g1/g2)$	$(1 g1)+(1 g1:g2)$	Intercept varying among $g1$ and $g2$ within $g1$	截距随着 $g1$, 以及嵌套在 $g1$ 下的 $g2$ 的不同而变化
$(1 g1)+(1 g2)$	$1+(1 g1)+(1 g2)$	Intercept varying among $g1$ and $g2$	截距随着 $g1$ 与 $g2$ 的不同而变化
$x+(x g)$	$1+x+(1+x g)$	Correlated random intercept and slope	随机截距且随机斜率模型
$x+(x g)$	$1+x+(1 g)+(0+x g)$	Uncorrelated random intercept and slope	随机截距且随机斜率模型，但二者不相关

$g1/g2$: 两个随机因素， $g1$ 为大单位， $g2$ 为小单位， $g2$ 嵌套于 $g1$ 中

Mixed effect model—嵌套型随机效应 (nested random effect structure)

注意如果随机效应命名情况为：那么写法

school	class
1	1
1	1
1	1
1	2
1	2
1	2
1	2
1	2
2	3
2	3
2	3
2	4
2	4
2	4
2	4
2	5
2	5
3	6
3	6
3	6
3	6
3	7
3	7
3	7

$(1 | \text{school/class})$ 等同于 $(1 | \text{school}) + (1 | \text{class})$

因为class的名字在不同的school中没有重复，这种情况下，来自一个class的数据自动被认嵌套在同一school之下。

在不同的school下，class的命名没有重复

多重嵌套 $(1 | \text{family/genus/species})$

每个随机效应的每个水平都进行唯一化命名

Mixed effect model—交叉型随机效应结构

- 随机效应之间无彼此嵌套关系
- 如物种和地点

$(1|\text{site}) + (1|\text{species})$

Linear mixed effect model— 随机截距模型 (Random intercept model)

```
m1<-lmer(Richness~NAP+(1|fBeach),data=RIKZ)
```

允许不同的Beach内，Richness 与NAP的回归关系的
截距不同，但斜率必须相同

Linear mixed effect model—随机截距模型结果解读

参数估计方法REML

```
> ##1)随机截距模型
> m1<-lmer(Richness~NAP+(1|fBeach),data=RIKZ)
> summary(m1)
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: Richness ~ NAP + (1 | fBeach)
Data: RIKZ
```

REML criterion at convergence: 239.5

Scaled residuals: 残差的分布特征

Min	1Q	Median	3Q	Max
-1.4228	-0.4848	-0.1576	0.2519	3.9794

Random effects:

Groups	Name	Variance	Std.Dev.
fBeach	(Intercept)	8.668	2.944
Residual		9.362	3.060

Number of obs: 45, groups: fBeach, 9

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(> t)
(Intercept)	6.5819	1.0958	8.3487	6.007	0.000271 ***
NAP	-2.5684	0.4947	37.0388	-5.192	7.77e-06 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
(Intr)
NAP -0.157

总体上Richness 和NAP的关系
 $\text{Richness} = 6.5819 - 2.5684 * \text{NAP}$

随机效应部分结果

随机效应(海滩)造成的方差及SD
残差的方差及SD
随机效应的水平数

总体样本量

固定效应参数
(截距和斜率)与
分别与0的差异是否
显著

也可以用boot法重新得到参数估计的置信区间

固定效应的
斜率和截距间的相关性

固定效应部分结果

Linear mixed effect model—随机截距模型

--结果与公式的对应

```
> fixef(m1)
(Intercept)
6.581893
```

NAP
-2.568400

$$\text{Richness}_{ij} = \mu + \alpha_j + \beta * \text{NAP}_{ji} + \varepsilon_{ji}$$

其中 $\alpha_j \sim N(0, sd^2)$; $\varepsilon_{ji} \sim N(0, \sigma^2)$

```
> ranef(m1)
$fBeach
(Intercept)
1      2.621519
2      5.199608
3     -2.615780
4     -2.275618
5      1.950179
6     -1.629402
7     -1.765477
8     -1.061665
9     -0.423364
```

Random effects:

Groups	Name	Variance	Std.Dev.
fBeach	(Intercept)	8.668	2.944
Residual		9.362	3.060

在不同的海滩（beach）上，截距的变异符合均值为0，方差为 σ^2 的正态分布，本案例中， σ^2 为8.668

也可以 boot法重新确定参数的置信区间

```
c1 <- confint(m1,method="boot", nsim = 1000)
```

```
c1
```

```
> c1
```

	2.5 %	97.5 %
.sig01	1.092816	4.670959
.sigma	2.301077	3.757961
(Intercept)	4.560886	8.753053
NAP	-3.477531	-1.565256

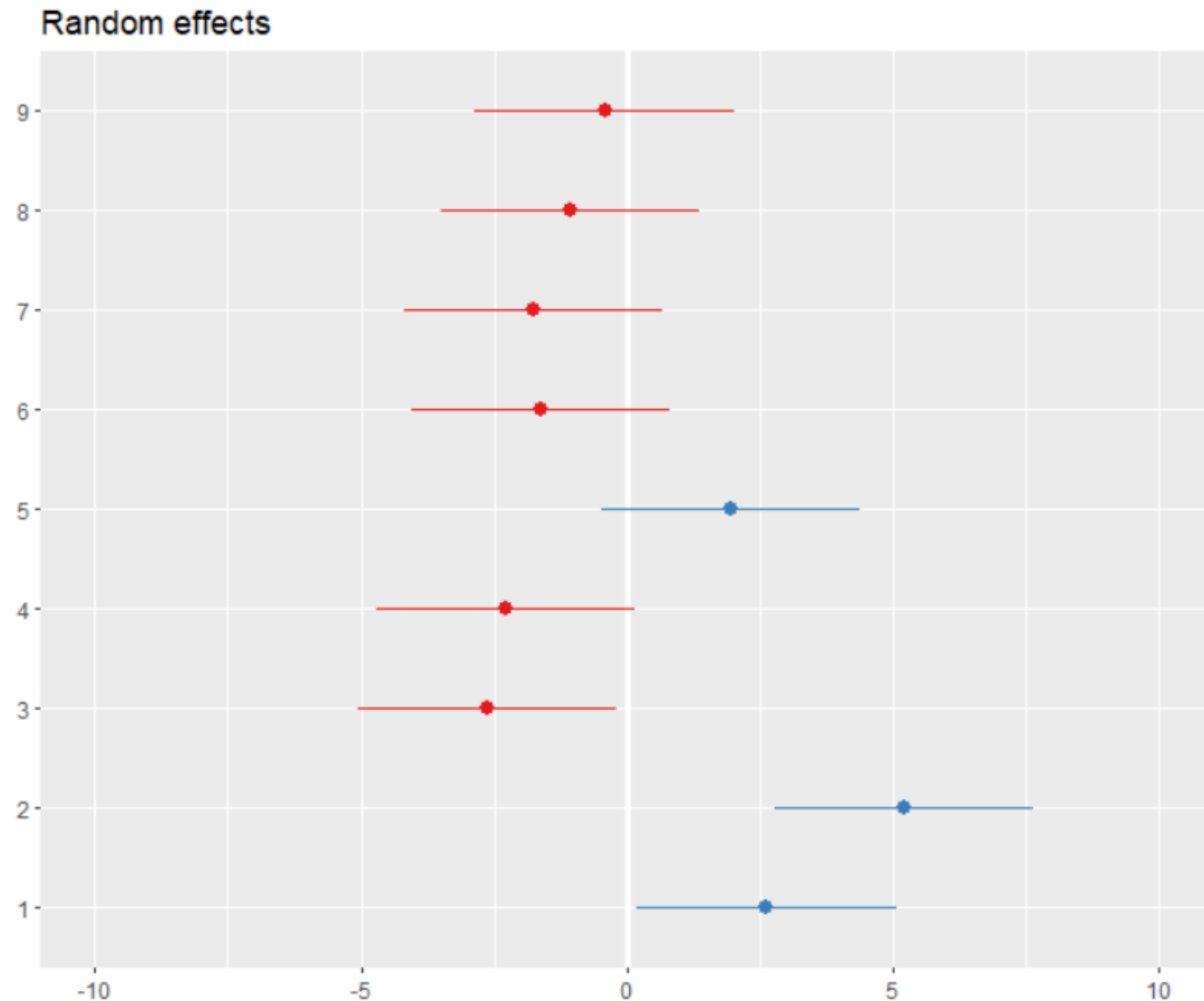
boot: 装大师的必备技能

Linear mixed effect model—随机截距模型

--随机效应部分结果作图

```
library(sjPlot)
##对随机效应作图
plot_model(m1, type='re')
##对固定效应作图
plot_model(m1, type = "eff",
terms = "NAP")
```

```
> ranef(m1)
$fBeach
(Intercept)
1      2.621519
2      5.199608
3     -2.615780
4     -2.275618
5      1.950179
6     -1.629402
7     -1.765477
8     -1.061665
9     -0.423364
```

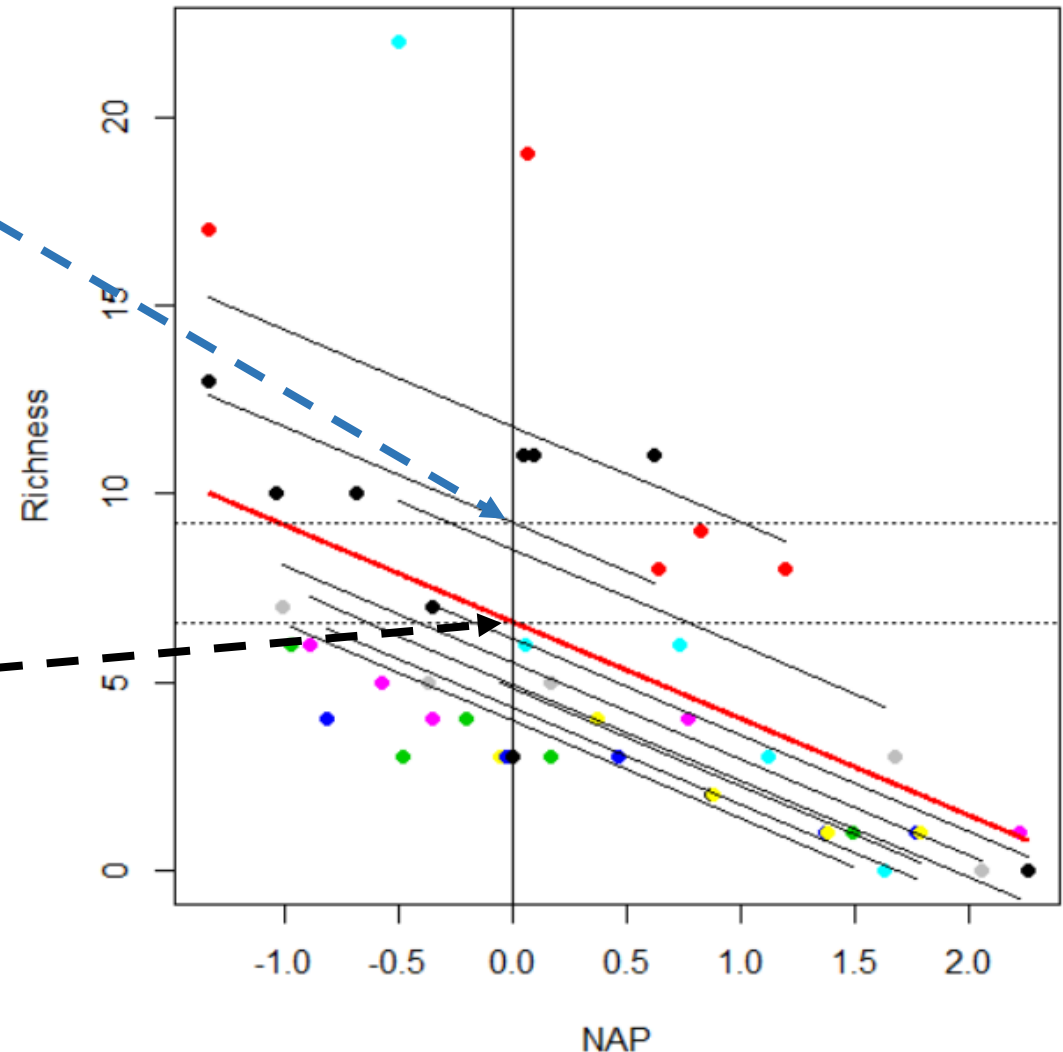


Linear mixed effect model—随机截距模型

--随机截距的意义

```
> ranef(m1)
$fBeach
(Intercept)
1      2.621519+6.581893
2      5.199608+6.581893
3     -2.615780+6.581893
4     -2.275618+6.581893
5      1.950179+6.581893
6     -1.629402+6.581893
7     -1.765477+6.581893
8     -1.061665+6.581893
9     -0.423364+6.581893
```

```
> fixef(m1)
(Intercept)
6.581893
NAP
-2.568400
```

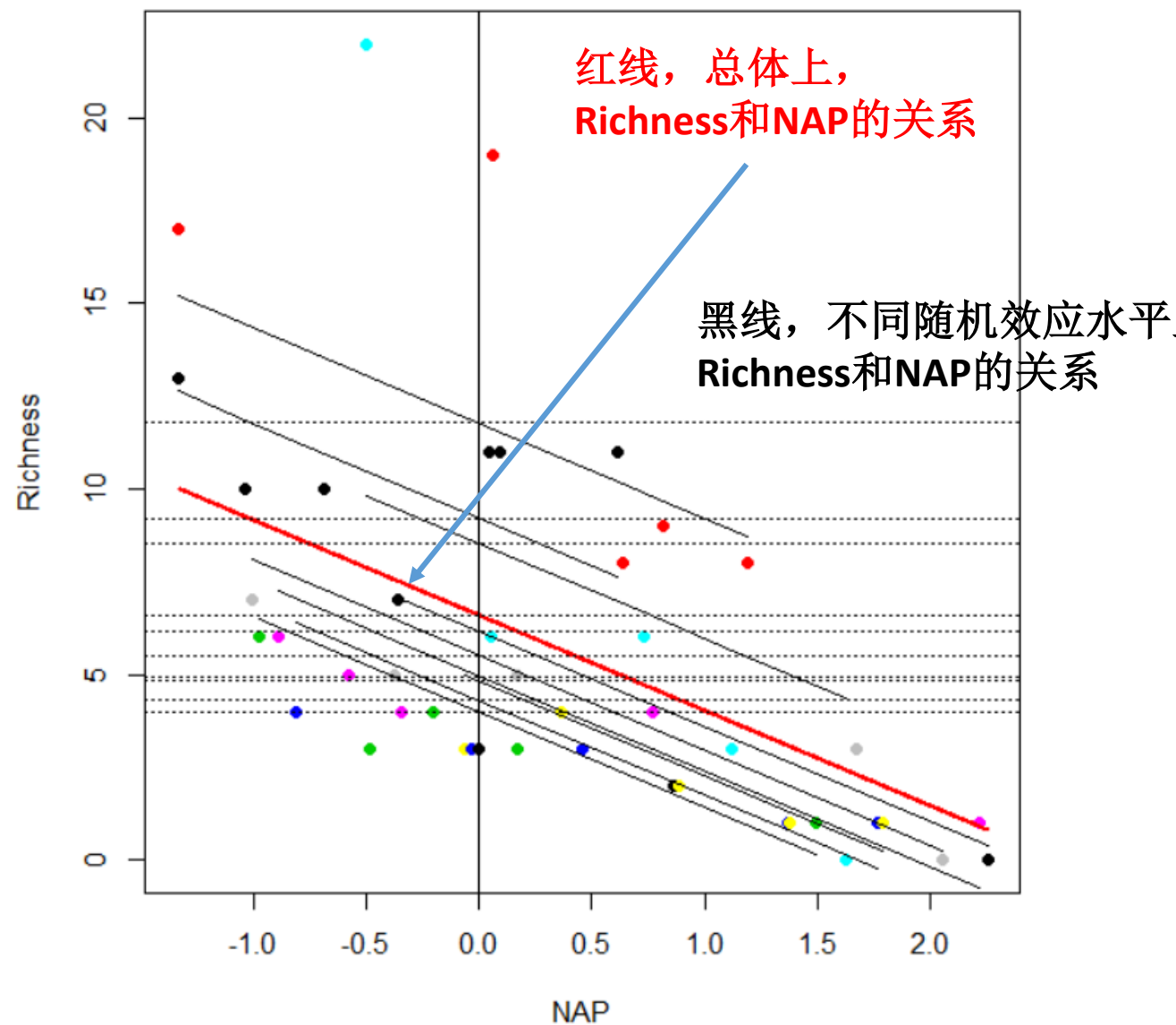


Linear mixed effect model—随机截距模型

--随机截距的意义

用coef命令直接查看在不同随机效应（海滩）水平下，Richness和NAP的关系

```
> coef(m1)
$fBeach
  (Intercept)      NAP
1    9.203415 -2.568399
2   11.781507 -2.568399
3    3.966110 -2.568399
4    4.306272 -2.568399
5    8.532074 -2.568399
6    4.952488 -2.568399
7    4.816414 -2.568399
8    5.520227 -2.568399
9    6.158528 -2.568399
```



Linear mixed effect model—随机截距模型

结果作图（固定效应部分）

###定义一个新的数据集，里面含变量NAP

```
new_NAP<-data.frame(NAP=seq(-1.336,2.2550,length=1000))
```

###用predict命令得到拟合线, 即与NAP对应的Richness值

```
pred_R<-predict(m1,newdata=new_NAP, re.form=~0)
```

###用bootMer得到拟合线的95%置信区间

```
ci_line<-bootMer(m1,FUN=function(.)  
  predict(.,newdata=new_NAP,  
          re.form=~0),nsim=1000)
```

```
ci_R<-apply(ci_line$t,2,function(x) x[order(x)][c(25,975)])
```

```
lb_R<-ci_R[1,]
```

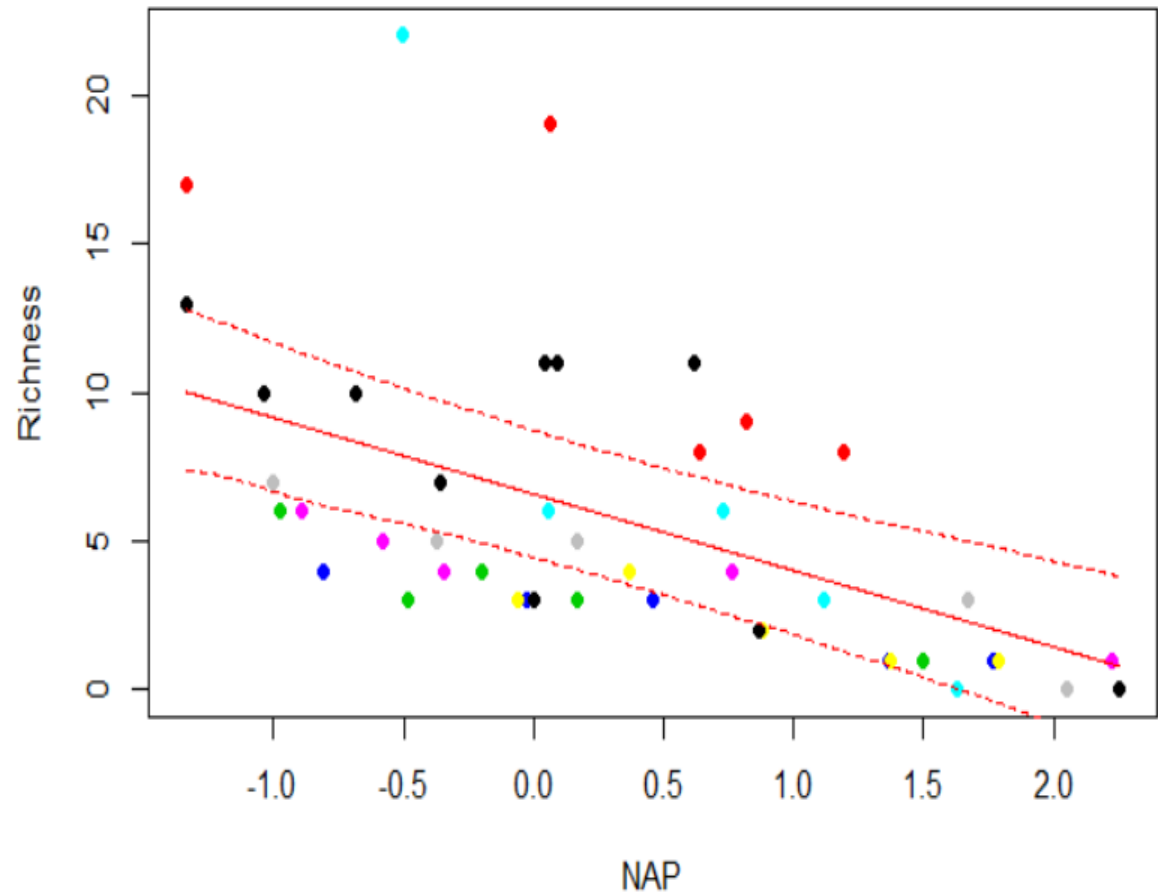
```
ub_R<-ci_R[2,]
```

```
plot(Richness~NAP,data=RIKZ,col=fBeach, pch=19)
```

```
lines(new_NAP$NAP,pred_R,lwd=1,col="red")
```

```
lines(new_NAP$NAP,lb_R,lty=2,col="red")
```

```
lines(new_NAP$NAP,ub_R,lty=2,col="red")
```



Linear mixed effect model结果作图—方法2

--add_ci命令

```
library(ciTools)
```

```
RIKZ<-RIKZ[order(RIKZ$NAP),]
```

```
plot(Richness~NAP,data=RIKZ,col=fBeach, pch=19)
```

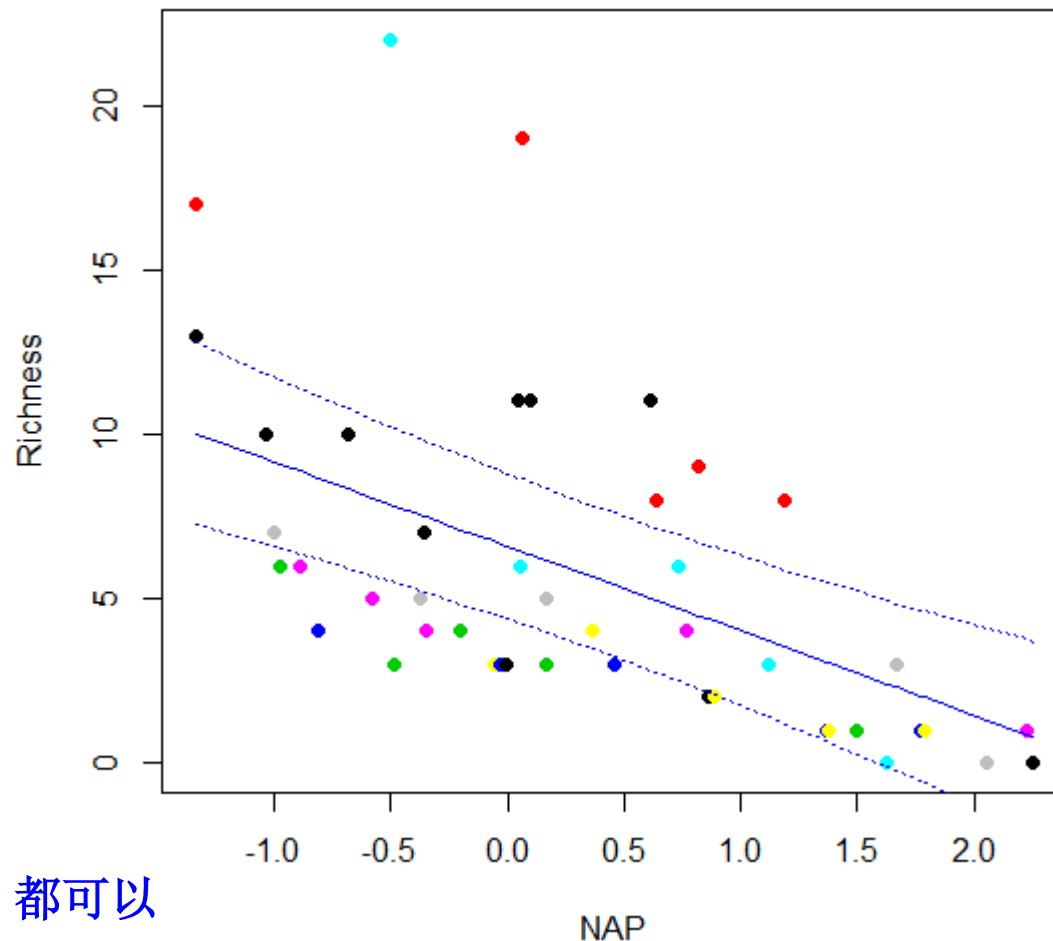
```
mm1_ci<-add_ci(RIKZ, m1, alpha = 0.05, type =  
"parametric",names = c("lwr", "upr"), includeRanef  
=F, nSims = 1000)
```

```
lines(mm1_ci$NAP, mm1_ci$pred, col = "blue", lwd  
= 1,lty="solid")
```

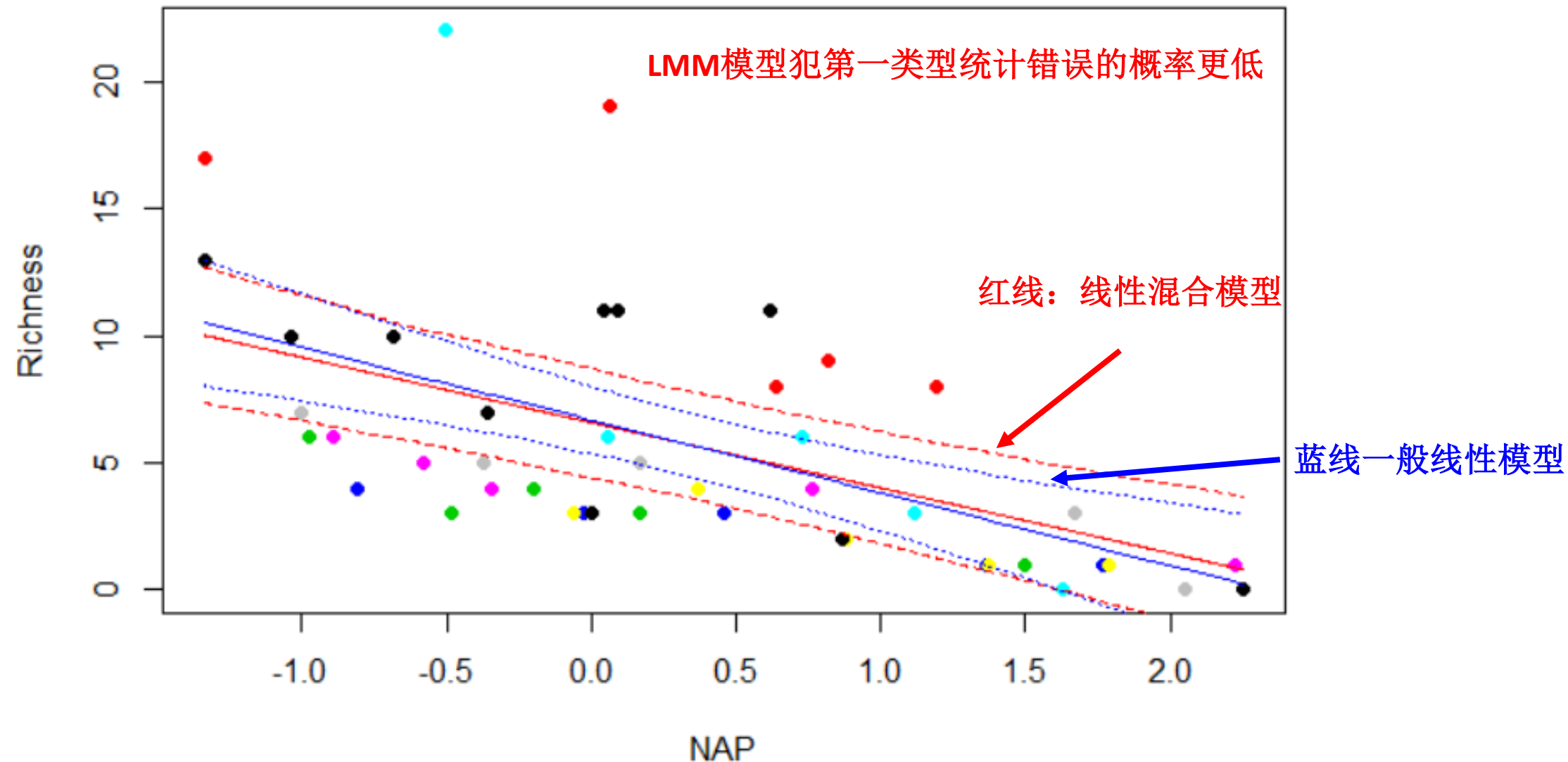
```
lines(mm1_ci$NAP, mm1_ci$upr, col = "blue", lwd =  
1,lty="dotted")
```

```
lines(mm1_ci$NAP, mm1_ci$lwr, col = "blue", lwd =  
1,lty="dotted")
```

type = “**parametric**“只能用于随机截距模型， type=“boot” 都可以



线性混合模型(lmm)和一般线性回归(lm)的结果对比



Linear mixed effect model—随机截距模型

--模型的 R^2

```
> library(MuMIn)
> r.squaredGLMM(m1)
```

	R2m	R2c
	0.2656394	0.6186737

R2m (marginal R^2):

由于固定效应造成的 R^2

R2c (conditional R^2):

固定效应和随机效应共同造成的 R^2

<i>Predictors</i>	Richness		
	<i>Estimates</i>	<i>CI</i>	<i>p</i>
(Intercept)	6.58	4.43 – 8.73	<0.001
NAP	-2.57	-3.54 – -1.60	<0.001
Random Effects			
σ^2	9.36		
$\tau_{00} \text{ fBeach}$	8.67		
ICC	0.48		
$N \text{ fBeach}$	9		
Observations	45		
Marginal R2 / Conditional R2	0.266 / 0.619		

ICC: Intraclass-Correlation Coefficient, 也称为variance partition coefficient (vpc) 指模型的随机变异部分中, 随机效应的方差所占的比重:
 $ICC = 8.67 / (8.67 + 9.36) = 0.48$

Nakagawa, S. and H. Schielzeth. 2013. A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution* 4:133-142.

截至2020年11月20日已被引用5782次（谷歌学术）

Linear mixed effect model—随机截距模型

结果制表

tab_model(m1)

Richness			
Predictors	Estimates	CI	p
(Intercept)	截距 → 6.58	4.43 – 8.73	<0.001
NAP	斜率 → -2.57	-3.54 – -1.60	<0.001
Richness=6.58-2.57*NAP			
Random Effects			
σ^2	9.36	← 残差的方差	
$\tau^2_{00} \text{ fBeach}$	8.67	← 随机效应(beach)造成的方差	
ICC fBeach	0.48	← 随机效应造成的方差在总随机方差中所占比率=8.67/(8.67+9.36)	
Observations	45	← 样本量	
Marginal R ² / Conditional R ²	0.266 / 0.619	(固定效应的R ² /(固定效应+随机效应的R ²))	

一般线性回归只是混合模型的特列

- 当随机效应的每个水平只有一个数据的时候，混合模型与线性回归结果一致

```
> m0<-lm(Richness~NAP,data=RIKZ)
> summary(m0)
```

Call:

```
lm(formula = Richness ~ NAP, data = RIKZ)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-5.0675	-2.7607	-0.8029	1.3534	13.8723

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	6.6857	0.6578	10.164	5.25e-13	***
NAP	-2.8669	0.6307	-4.545	4.42e-05	***

```
> m00<-lme(Richness~NAP,random=~1|Sample,data=RIKZ)
```

```
> summary(m00)
```

Linear mixed-effects model fit by REML

Data: RIKZ

	AIC	BIC	logLik
	260.201	267.2458	-126.1005

Random effects:

Formula: ~1 | Sample

(Intercept) Residual

StdDev: 3.895063 1.460648

Fixed effects: Richness ~ NAP

	Value	Std.Error	DF	t-value	p-value
(Intercept)	6.685662	0.6577579	43	10.164320	0
NAP	-2.866853	0.6307186	43	-4.545376	0

Correlation:

(Intr)

NAP -0.333

注意: lme4中，无法进行随机效应的每个水平只有一个值的运算

Linear mixed effect model— 随机效应的显著性检验

```
> m0<-lm(Richness~NAP, data = RIKZ)
> m100<-lme(Richness~NAP, random =~1|fBeach,data =RIKZ,method="ML")
> exactLRT(m=m100, m0=m000)
No restrictions on fixed effects. REML-based inference preferable.
```

simulated finite sample distribution of LRT. (p-value based on 10000 simulated values)

```
data:
LRT = 12.124, p-value = 2e-04
```

```
> AIC(m0,m100)
      df      AIC
m0      3 259.9535
m100    4 249.8291
```

加入随机效应后模型差异显著

we added random effects for region on ... To test **whether region had a significant effect on** the slope of ..., we used the function **exactLRT** from package **RLRsim** (Sirami et al. 2019, PNAS).

Linear mixed effect model—随机截距模型

--模型诊断

多种模型诊断方法

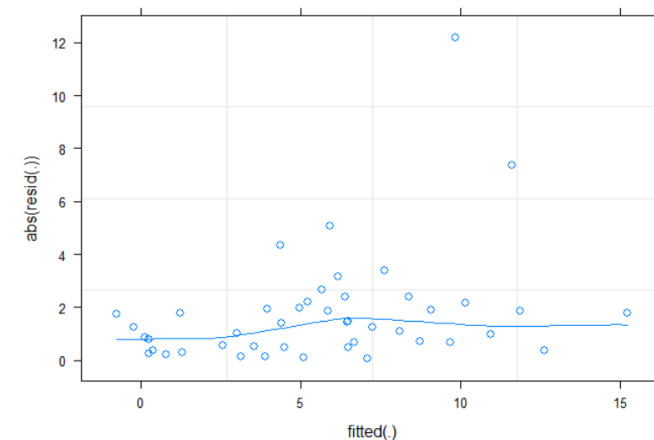
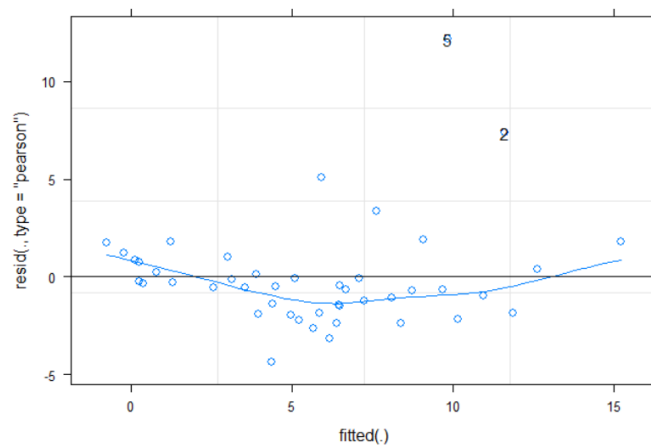
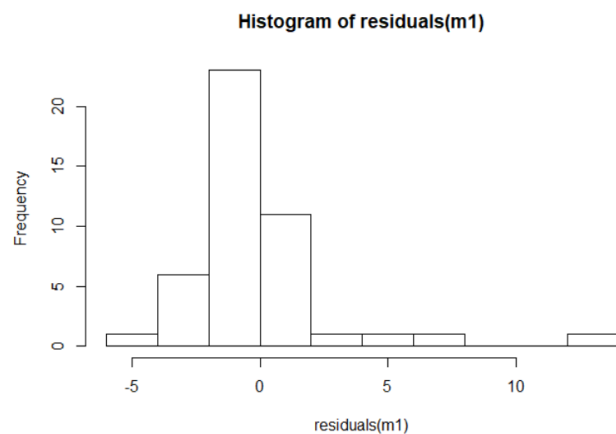
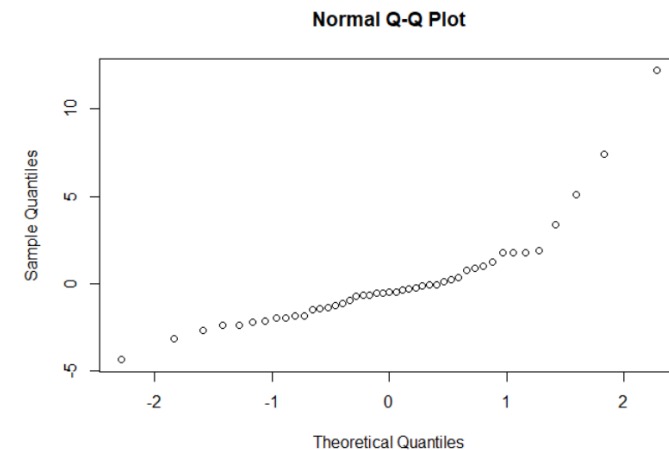
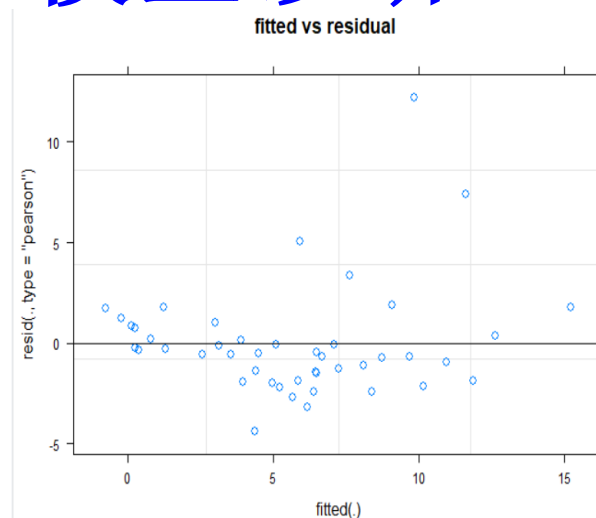
```
plot(m1)
```

```
qqnorm(residuals(m1))
```

```
hist(residuals(m1))
```

```
plot(m1, type = c("p", "smooth"), id = 0.05)
```

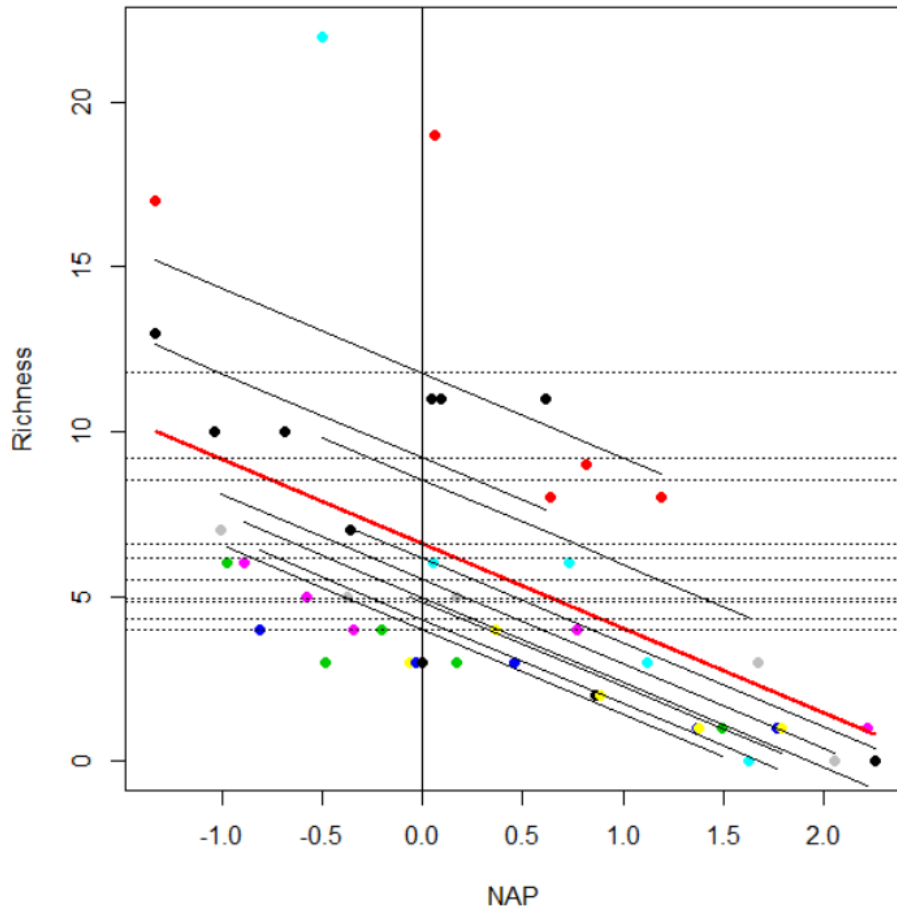
```
plot(m1, abs(resid(.)) ~ fitted(.), type=c('p', 'smooth'))
```



该模型残差明显不符合正态分布

Linear mixed effect model— 随机截距且随机斜率模型

Random intercept model



那么是否也可以允许斜率也随机波动呢？

Random intercept and slope model

```
m2<-lmer(Richness~NAP+(NAP|fBeach),data=RIKZ)
```

允许不同的Beach内，Richness 与NAP的回归关系的
截距和斜率均不同

Linear mixed effect model—随机截距、随机斜率模型

不同的fBeach内， NAP对Richness影响的截距和斜率都有变化

`m2<-lmer(Richness~NAP+(NAP|fBeach),data=RIKZ)`

`summary(m2)`

`anova(m2)`

```
> m2<-lmer(Richness~NAP+(NAP|fBeach),data=RIKZ)
> summary(m2)
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: Richness ~ NAP + (NAP | fBeach)
Data: RIKZ

REML criterion at convergence: 232.4

Scaled residuals:
    Min       1Q   Median       3Q      Max
-1.8213 -0.3411 -0.1675  0.1921  3.0397

Random effects:
 Groups   Name                Variance Std.Dev. Corr
fBeach    (Intercept)    12.596     3.549
          NAP             2.941     1.715    -0.99
Residual              7.305     2.703
Number of obs: 45, groups: fBeach, 9

Fixed effects:
              Estimate Std. Error    df t value Pr(>|t|)
(Intercept)    6.5887     1.2648  7.5715   5.209 0.000968 ***
NAP            -2.8300     0.7229  5.6942  -3.915 0.008705 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
      (Intr)
NAP  -0.819
> anova(m2)
Type III Analysis of Variance Table with Satterthwaite's method
      Sum Sq Mean Sq NumDF  DenDF  F value    Pr(>F)
NAP  111.95   111.95     1  5.6942   15.324 0.008705 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

随机截距
随机斜率

随机斜率和随机截距之间的相关性

采用`lmerTest::anova`评判x对y影响显著性：
自由度计算采用Satterthwaite approximation 法估算自由度

```
> anova(m2, type="2")
Type II Analysis of Variance Table with Satterthwaite's method
      Sum Sq Mean Sq NumDF  DenDF  F value    Pr(>F)
NAP  111.95   111.95     1  5.6942   15.324 0.008705 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

F检验有明确的分母自由度，可知随机效应结构设置是否合理
数据不平衡或者广义混合模型，`anova`不适用，可以选用log likelihood ratio test（LRT）
来检验某因素的显著性，但LRT的问题在于模型的自由度未知，随机效应设置是否合适容易被忽略。

Arnqvist, G. 2020. Mixed models offer no freedom from degrees of freedom.
Trends in Ecology & Evolution 35:329-335.

采用`car::Anova`评判x对y影响显著性：
自由度计算采用`Kenward-Roger approximation`法估算自由度

```
> library(car)
> Anova(m2, test.statistic = "F")
Analysis of Deviance Table (Type II Wald F tests with Kenward-Roger df)

Response: Richness
          F Df Df.res    Pr(>F)
NAP 14.804  1 7.7509 0.005197 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

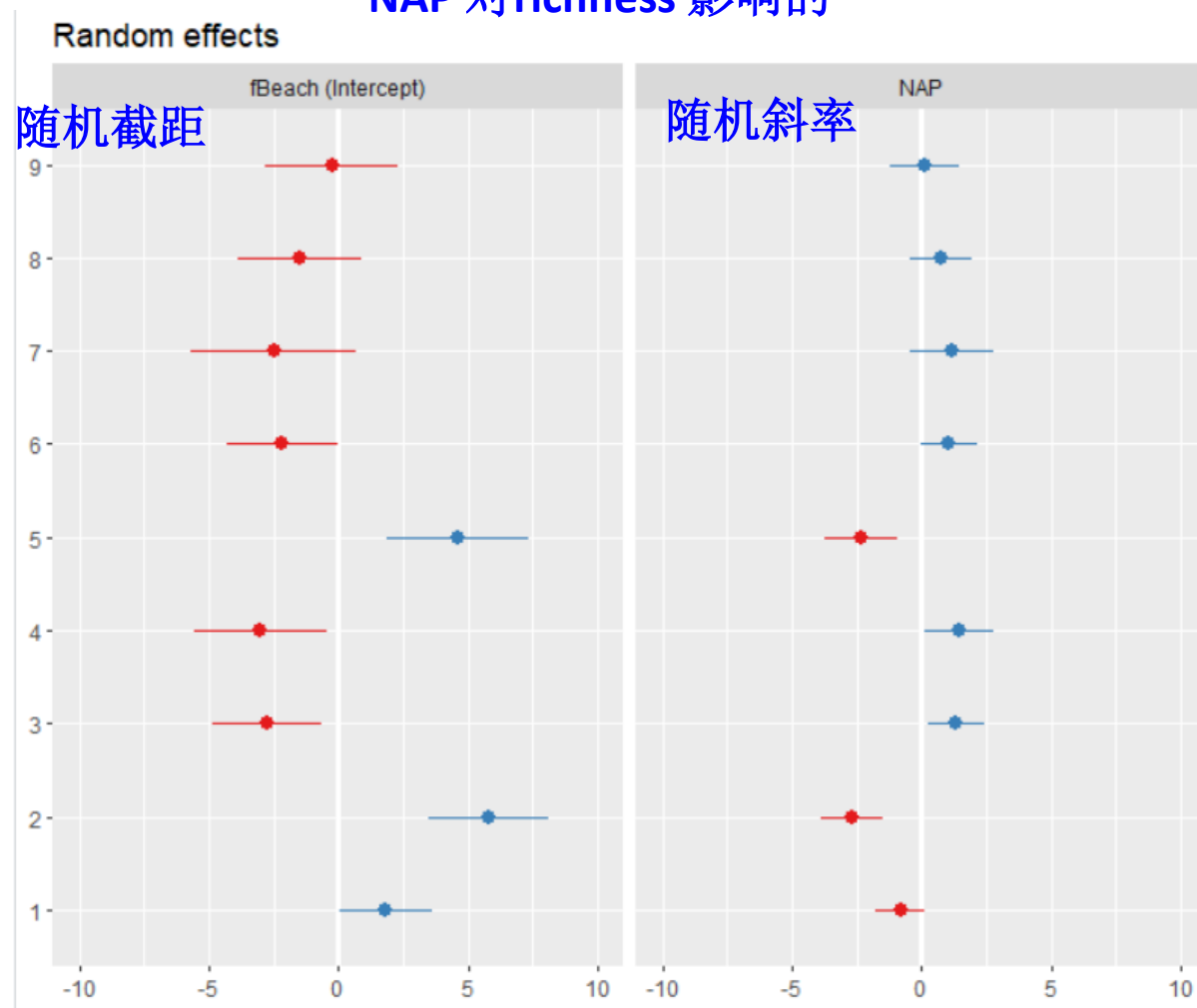
Linear mixed effect model— 随机截距、随机斜率模型的意义

plot_model(m2, type = "re")

```
> ##查看固定效应的估计值  
> fixef(m2)  
(Intercept)      NAP  
  6.588703    -2.830026
```

```
> ranef(m2)  
$fBeach  
  (Intercept)      NAP  
1  1.8323466 -0.8262550  
2  5.7747956 -2.7068019  
3 -2.7820616  1.3243100  
4 -3.0262829  1.4440640  
5  4.6114353 -2.3072473  
6 -2.1624277  1.0543432  
7 -2.5057621  1.1857008  
8 -1.4888102  0.7231694  
9 -0.2532329  0.1087167
```

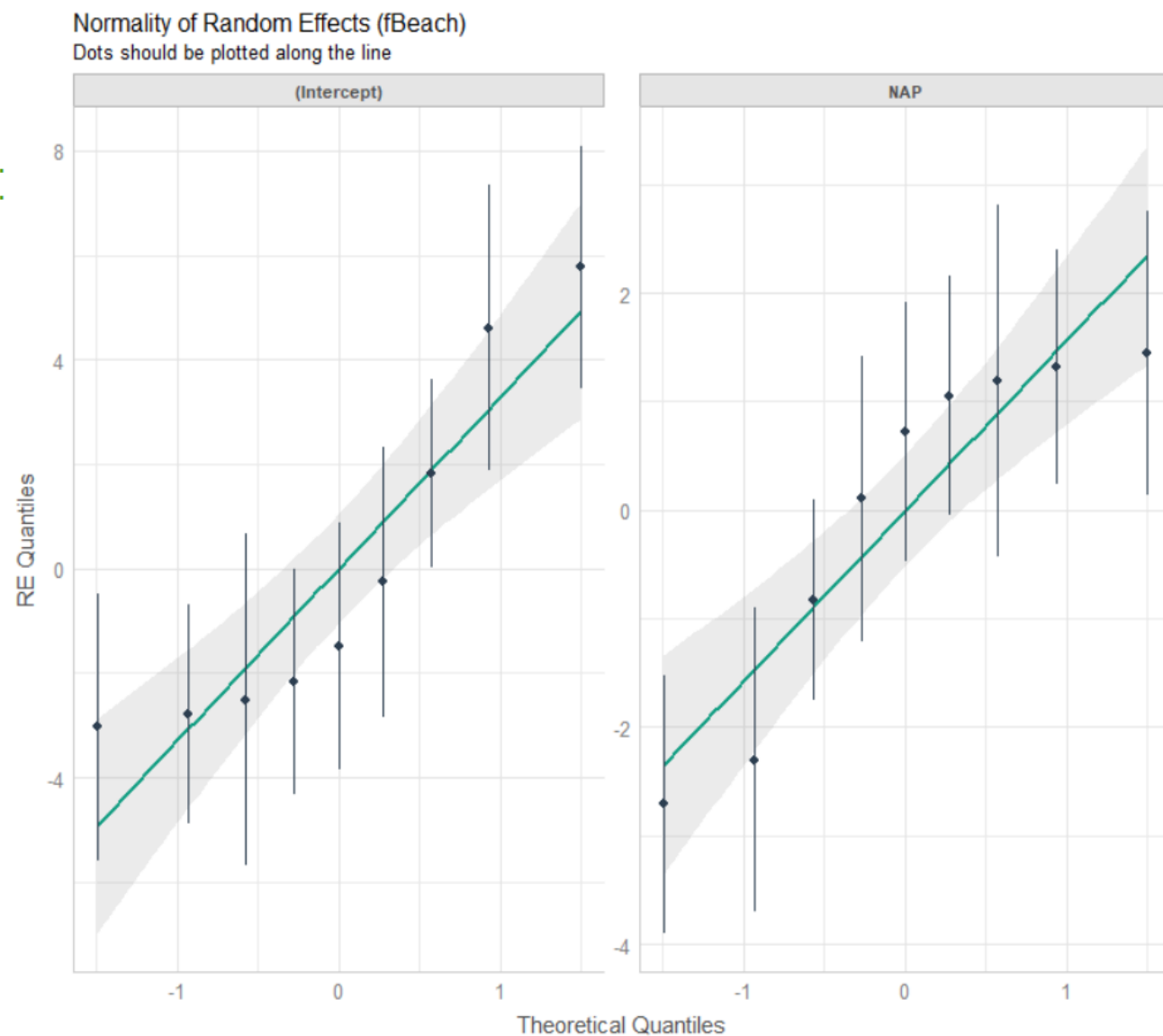
NAP 对 richness 影响的



随机效应是否真的服从正态分布？

```
> library(performance)
> result <- check_normality(m2, effects = "random")
Group: fBeach
(Intercept) OK: random effects appear as normally distributed (p = 0.063).
NAP OK: random effects appear as normally distributed (p = 0.051).
> plot(result)
```

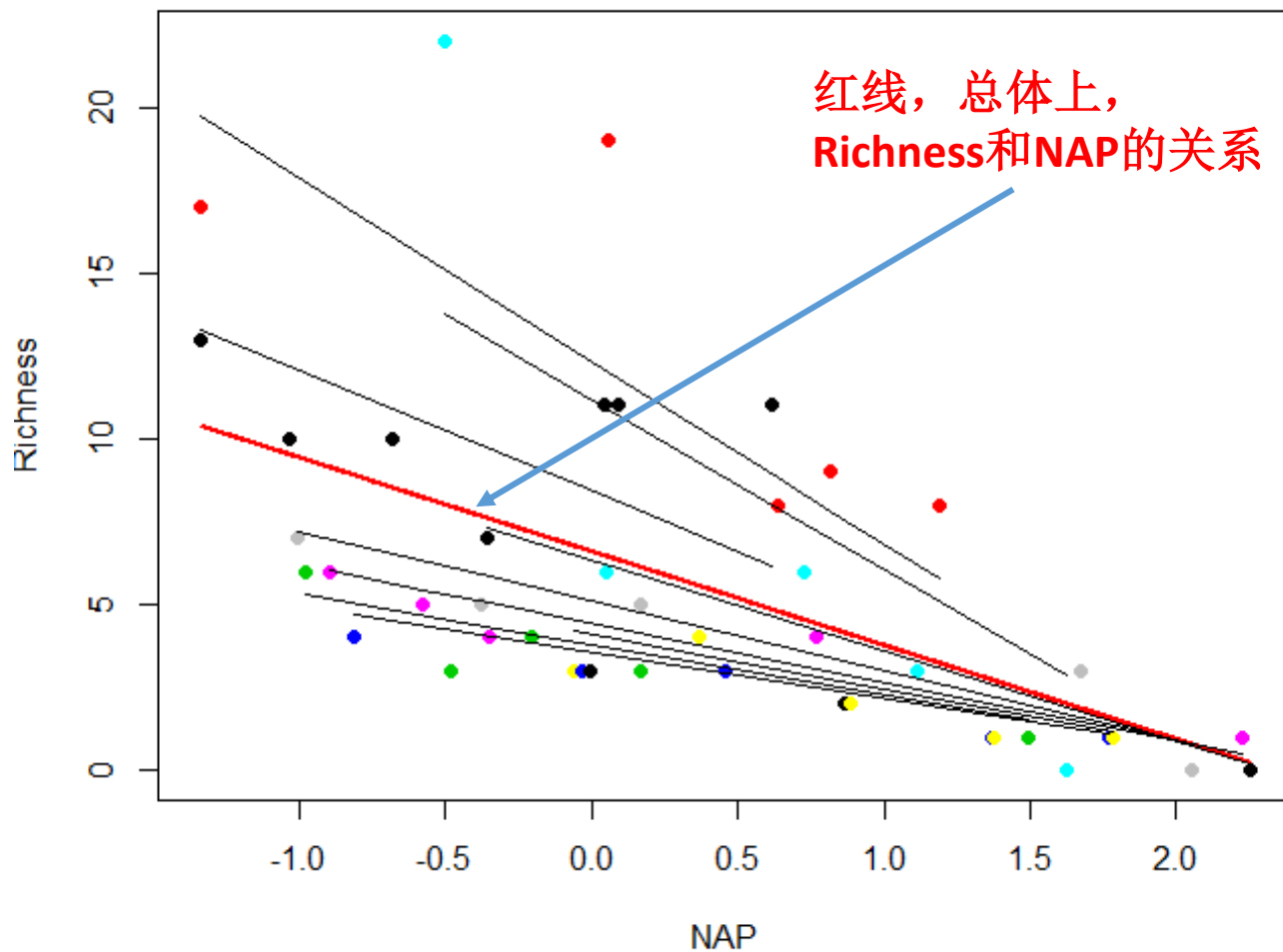
该检验在实际应用中较少见



Linear mixed effect model— 随机截距、随机斜率模型的意义

用coef命令直接查看随机效应（海滩）不同水平下，
Richness 和NAP的回归关系

```
> coef(m2)
$fBeach
  (Intercept)      NAP
1      8.421    -3.656
2     12.363    -5.537
3      3.807    -1.506
4      3.562    -1.386
5     11.200    -5.137
6      4.426    -1.776
7      4.083    -1.644
8      5.100    -2.107
9      6.335    -2.721
```



Linear mixed effect model— 随机截距、随机斜率模型的结果制表

tab_model(m2)

marginal R^2 :

由于固定效应造成的 R^2

conditional R^2 :

固定效应和随机效应共同造成的 R^2

Richness			
<i>Predictors</i>	<i>Estimates</i>	<i>CI</i>	<i>p</i>
(Intercept)	6.59	4.11 – 9.07	<0.001
NAP	-2.83	-4.25 – -1.41	<0.001
Random Effects			
σ^2	7.30	残差的方差	
τ_{00} fBeach	12.63	随机截距造成的方差	
τ_{11} fBeach.NAP	2.94	随机斜率造成的方差	
ρ_{01} fBeach	-0.99	随机截距和随机斜率之间的相关性	
ICC	0.61	ICC= (12.63+2.94/(12.63+2.94+7.30)=0.61	
N fBeach	9	随机效应的水平数	
Observations	45		
Marginal R2 / Conditional R2	0.295 / 0.728		

Linear mixed effect model— 随机截距、随机斜率模型总体效应结果作图

```
new_NAP<-data.frame(NAP=seq(-1.336,2.2550,length=1000))
```

###predict命令与bootMer命令相结合

```
pred_R2<-predict(m2, newdata=new_NAP, re.form=~0)
```

```
ci_line2<-bootMer(m2,FUN=function(.)  
predict(.,newdata=new_NAP,  
re.form=~0),nsim=1000)
```

```
ci_R2<-apply(ci_line2$t,2,function(x) x[order(x)][c(25,975)])
```

```
lb_R2<-ci_R2[1,]
```

```
ub_R2<-ci_R2[2,]
```

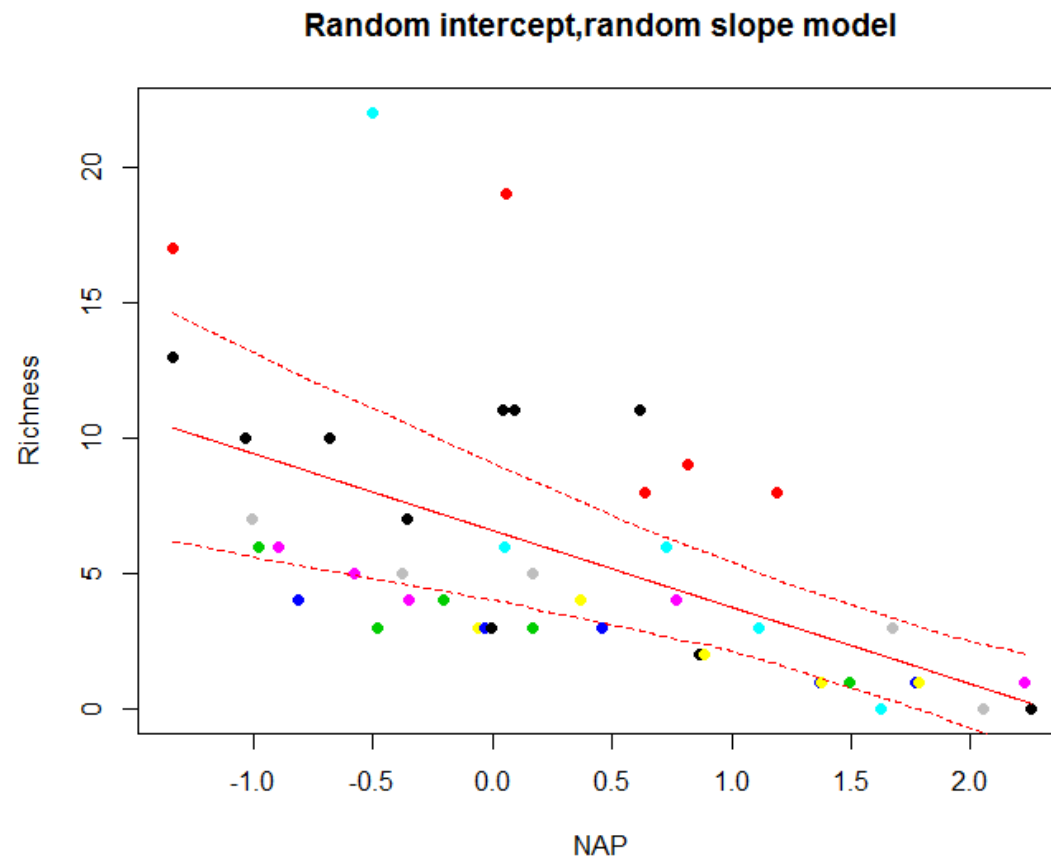
```
plot(Richness~NAP,data=RIKZ,col=fBeach, pch=19)
```

```
lines(new_NAP$NAP,pred_R2,lwd=1,col="red")
```

```
lines(new_NAP$NAP,lb_R2,lty=2,col="red")
```

```
lines(new_NAP$NAP,ub_R2,lty=2,col="red")
```

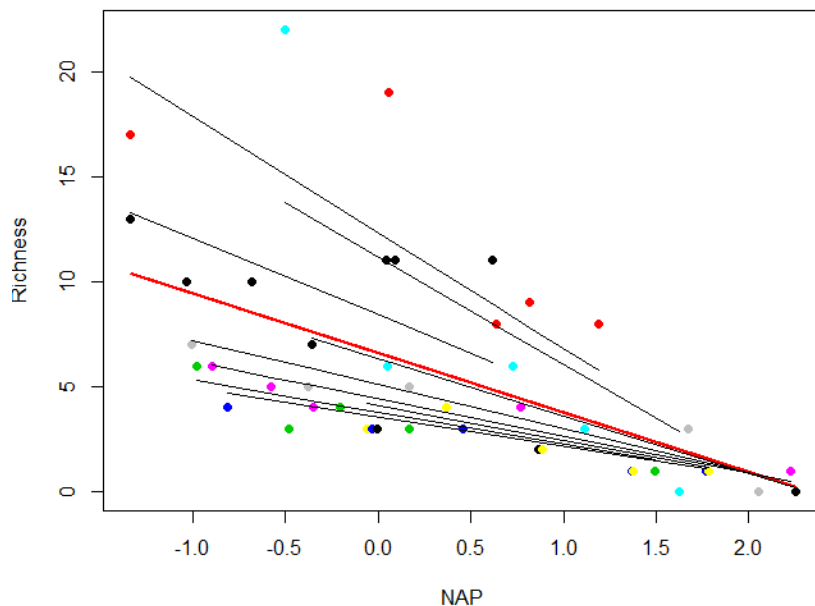
```
title("Random intercept,random slope model")
```



对固定效应部分作图

Linear mixed effect model— 随机截距、随机斜率模型—随机部分结果作图

随机截距和随机斜率之间的相关性



```
Random effects:
Groups   Name              Variance Std.Dev. Corr
fBeach   (Intercept)  12.596   3.549
NAP      NAP          2.941   1.715   -0.99
Residual                    7.305   2.703
Number of obs: 45, groups: fBeach, 9
```

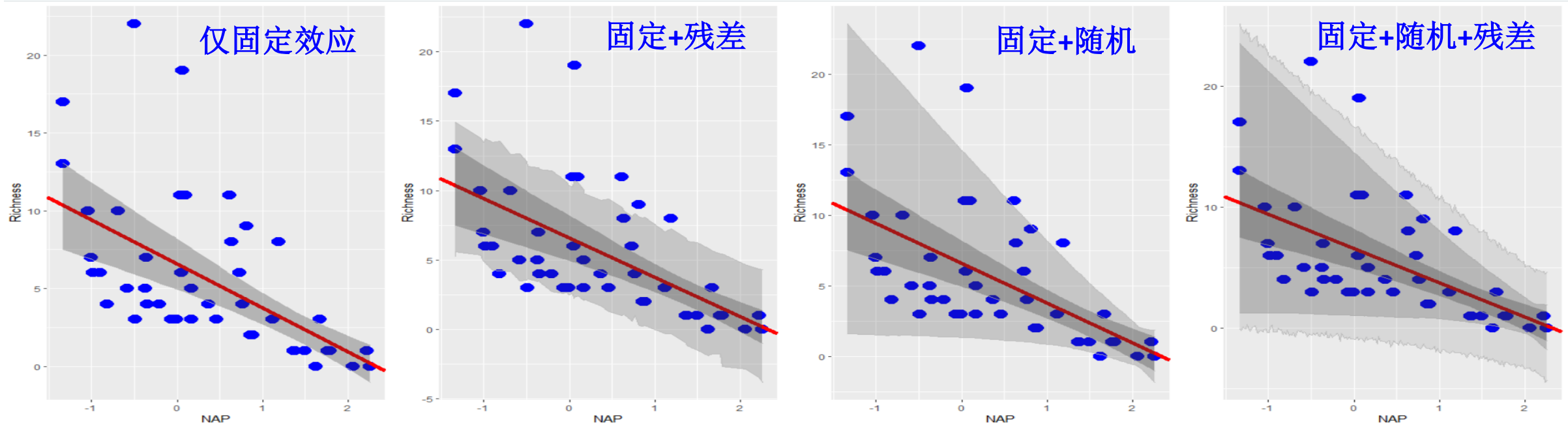
也可以指定随机截距和斜率之间没有相关性

Then, we added random effects for region on.... **We assumed that the effects of region on the intercept and slopes were uncorrelated** (Sirami et al. 2019, PNAS).

随机效应的写法由(x|Region) 变为(1|Region)+(0+x|Region)即可

Sirami, C. *et al.* Increasing crop heterogeneity enhances multitrophic diversity across agricultural regions. *Proc. Natl. Acad. Sci. USA*, 201906419, doi:10.1073/pnas.1906419116 (2019).

混合效应模型，纳入不同来源的不确定性



`library(merTools)`

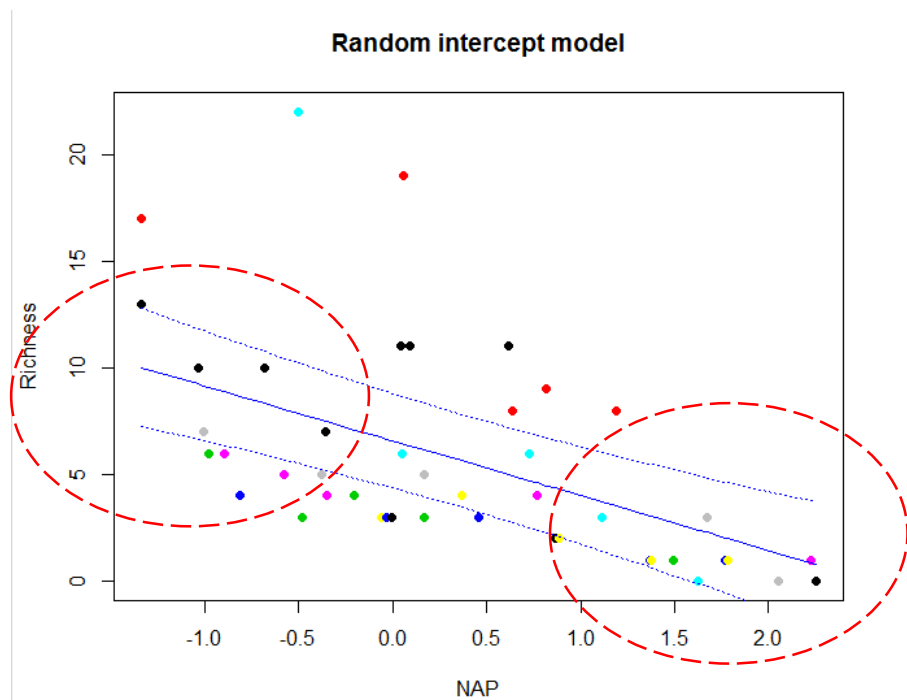
```
prd1<- predictInterval(m2, newdata=RIKZ, which="fixed", include.resid.var = F)
```

```
prd2<- predictInterval(m2, newdata=RIKZ, which="fixed", include.resid.var = T)
```

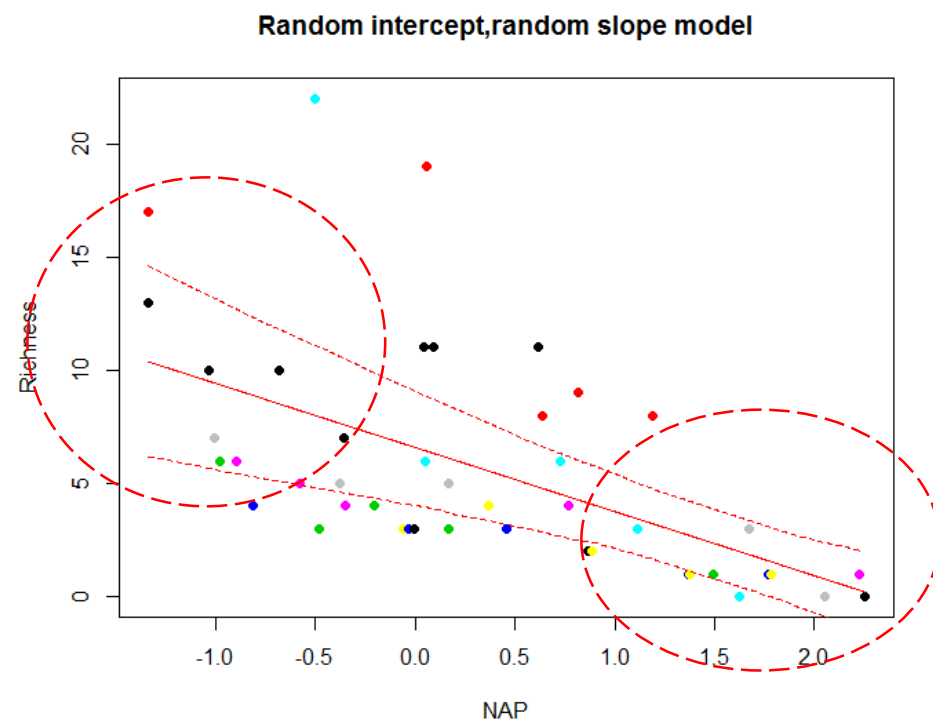
```
prd3<- predictInterval(m2, newdata=newd1, which="full", include.resid.var = F)
```

```
prd4<- predictInterval(m2, newdata=newd1, which="full", include.resid.var = T)
```

Linear mixed effect model— 随机截距模型与随机截距、随机斜率模型的优劣对比



`m1<-lmer(Richness~NAP+(1|fBeach), data=RIKZ)`



`m2<-lmer(Richness~NAP+(NAP|fBeach),data=RIKZ)`

模型对比：在数据一定的前提下，哪一种结构的模型更好

似然值、对数似然比检验与AIC值

似然值： $L = p(y|\beta, \sigma, X) = \prod_{i=1}^n N(y_i|X_i\beta, \sigma^2)$ 给定 x , β 得到当前所观测到的 y 值的概率

最大似然估计： β 取何值时， L 最大

对数似然： $\log L$

对数似然比检验(log-likelihood ratio test):

1) 对于同一组数据，假如有两个模型(如包含的固定效应不同)

2) 其似然值分别为 $L1$, $L2$ ，那个这两个模型的对数似然值比

即: $\text{likelihood ratio} = \ln(L1) - \ln(L2) = \ln(\frac{L1}{L2})$,

应服从自由度为 $df = |df1 - df2|$ 的卡方分布

3) 其中 $df1$, 和 $df2$ 分别为这两个模型的参数个数

所以可以依据log-likelihood ratio test的结果作为
判定某一因素显著性的指标

Linear mixed effect model

参数估计方法与模型选择

随机效应不变，对比固定效应：必须ML法

固定效应不变，对比随机效应：优先REML

```
m1<-lmer(Richness~NAP+(1|fBeach),data=RIKZ) ##默认为REML法
```

```
m2<-lmer(Richness~NAP+(1|fBeach),REML=F, data=RIKZ) ##ML法
```

AIC (*Akaike information criterion*)值的意义

$$AIC=2K-2\ln(L)$$

- **K**: 模型参数的个数，代表了模型的复杂性
- **L**: 模型的似然值，代表了模型拟合的结果与实际数据的匹配程度，越大，则说明模型拟合越准确。
- 所以AIC值是模型简洁性和准确性的综合
- 好模型的特征：简洁（参数少，K小）准确（L值大）
- 所以AIC值越小说明模型越好，一般以2为阈值

模型拟合优度对比

一般线性模型: `m0<-lm(Richness~NAP,data=RIKZ)`

混合: 随机截距

`m1<-lmer(Richness~NAP+(1|fBeach),data=RIKZ)`

混合: 随机截距, 随机斜率

`m2<-lmer(Richness~NAP+(NAP|fBeach),data=RIKZ)`

```
> AIC(m0,m1,m2)
```

	df	AIC
--	----	-----

m0	3	259.9535
----	---	----------

m1	4	247.4802
----	---	----------

m2	6	244.3839
----	---	----------

随机效应模型的结构确定

GENERAL PROTOCOL FOR MODEL FITTING

1. Start with model with all fixed and random effects that may be important. Evaluate with diagnostics.
2. Evaluate random effects with full model of all fixed effects using REML(χ^2 , RLRT, cAIC, etc.)
3. Evaluate fixed effects with reduced random effects (F Tests using ML fit)
4. Model diagnostics again...
5. Draw inference from model using Wald tests, lsmeans, visualization, etc.

一定要先确定随机效应的结构，
然后再确定固定效应的结构

用对数似然比检验判定某因素的显著性

log-likelihood ratio test

```
m1<-lmer(Richness~NAP+(1|fBeach),data=RIKZ)
```

```
m2<-lmer(Richness~NAP+(NAP|fBeach),data=RIKZ)
```

```
> anova(m1,m2)
refitting model(s) with ML (instead of REML)
Data: RIKZ
Models:
m1: Richness ~ NAP + (1 | fBeach)
m2: Richness ~ NAP + (NAP | fBeach)
   Df    AIC    BIC logLik deviance Chisq  Chi Df Pr(>Chisq)
m1   4 249.83 257.06 -120.92   241.83      7.173    2 0.02769 *
m2   6 246.66 257.50 -117.33   234.66      7.173    2 0.02769 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

注意LRT 中自由度是两模型参数数量的差异，而并非模型本身的自由度

```
> library(lmtest)
> lrtest(m1, m2)
Likelihood ratio test

Model 1: Richness ~ NAP + (1 | fBeach)
Model 2: Richness ~ NAP + (NAP | fBeach)
  #Df LogLik Df  Chisq Pr(>Chisq)
1    4 -119.74
2    6 -116.19  2  7.0964   0.02878 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

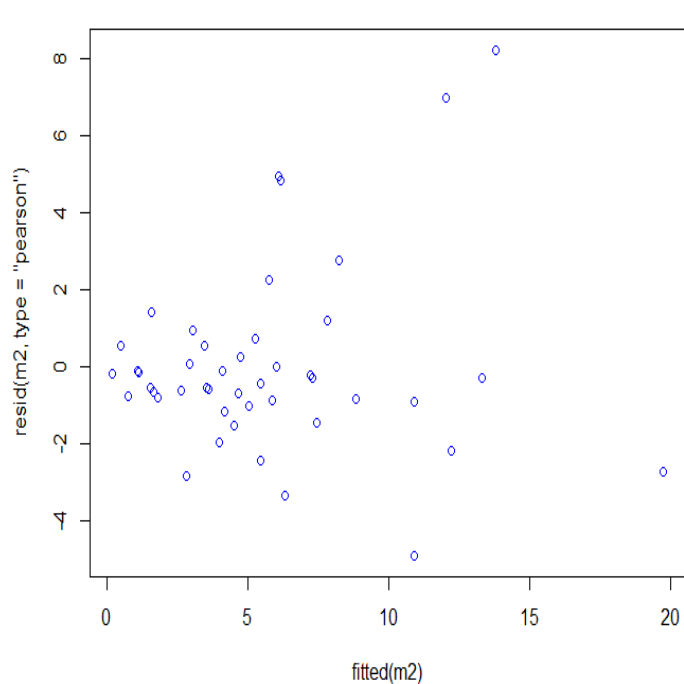
```
m1_ml<-lmer(Richness~NAP+(1|fBeach),data=RIKZ, REML=F)
m2_ml<-lmer(Richness~NAP+(NAP|fBeach),data=RIKZ, REML=F)
```

```
> library(lmtest)
> lrtest(m1_ml, m2_ml)
Likelihood ratio test

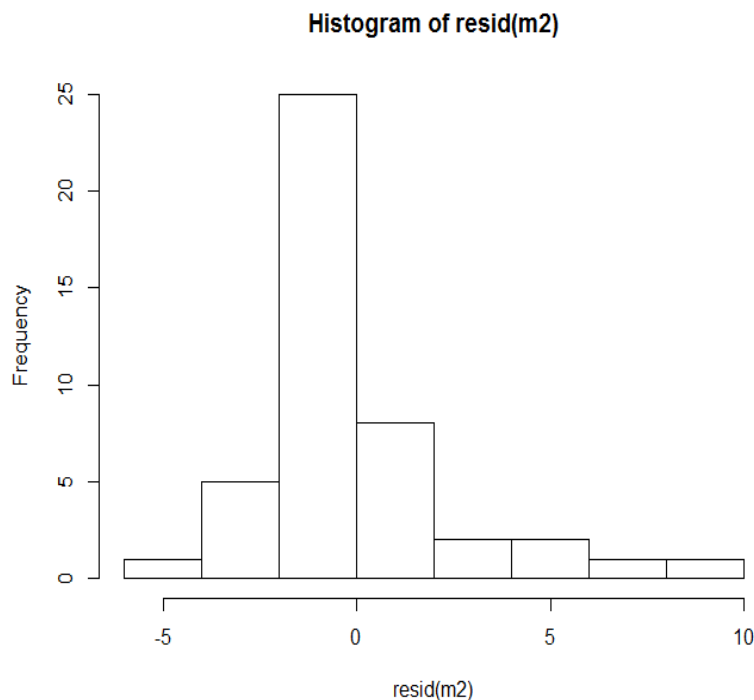
Model 1: Richness ~ NAP + (1 | fBeach)
Model 2: Richness ~ NAP + (NAP | fBeach)
  #Df LogLik Df  Chisq Pr(>Chisq)
1    4 -120.92
2    6 -117.33  2  7.173   0.02769 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

随机截距、随机斜率模型（m2）—模型诊断

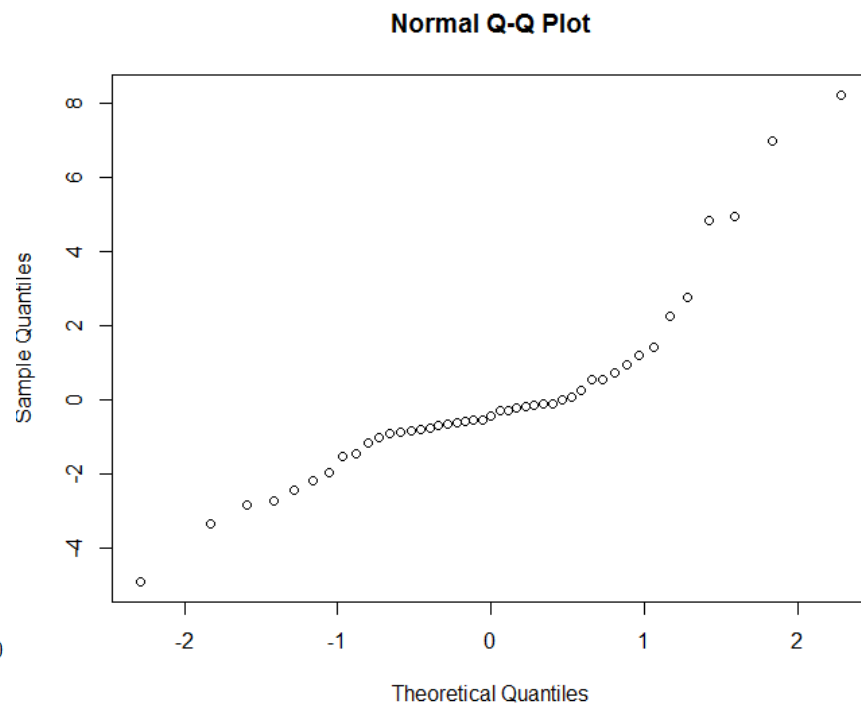
```
m2<-lmer(Richness~NAP+(NAP|fBeach),data=RIKZ)
```



```
plot(fitted(m2),resid(m2,  
type="pearson"),col="blue")
```



```
hist(resid(m2))
```



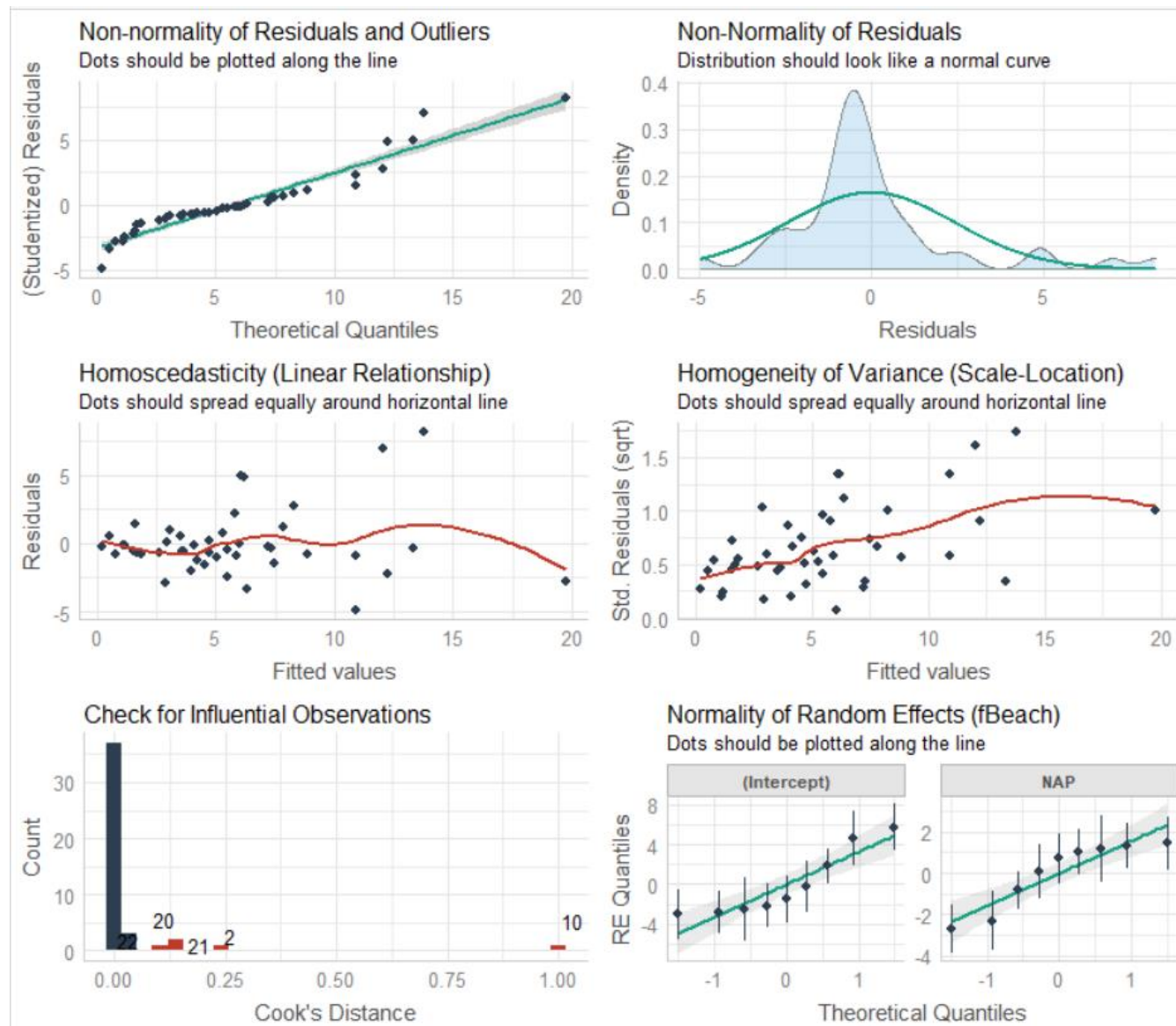
```
qqnorm(resid(m2))
```

该模型残差明显不符合正态分布

模型诊断的另一方法: performance包

```
m2<-lmer(Richness~NAP+(NAP|fBeach),data=RIKZ)
performance::check_model(m2)
```

总体上模型m2并不理想
需要改进



休息一下，稍后进入广义线性混合模型