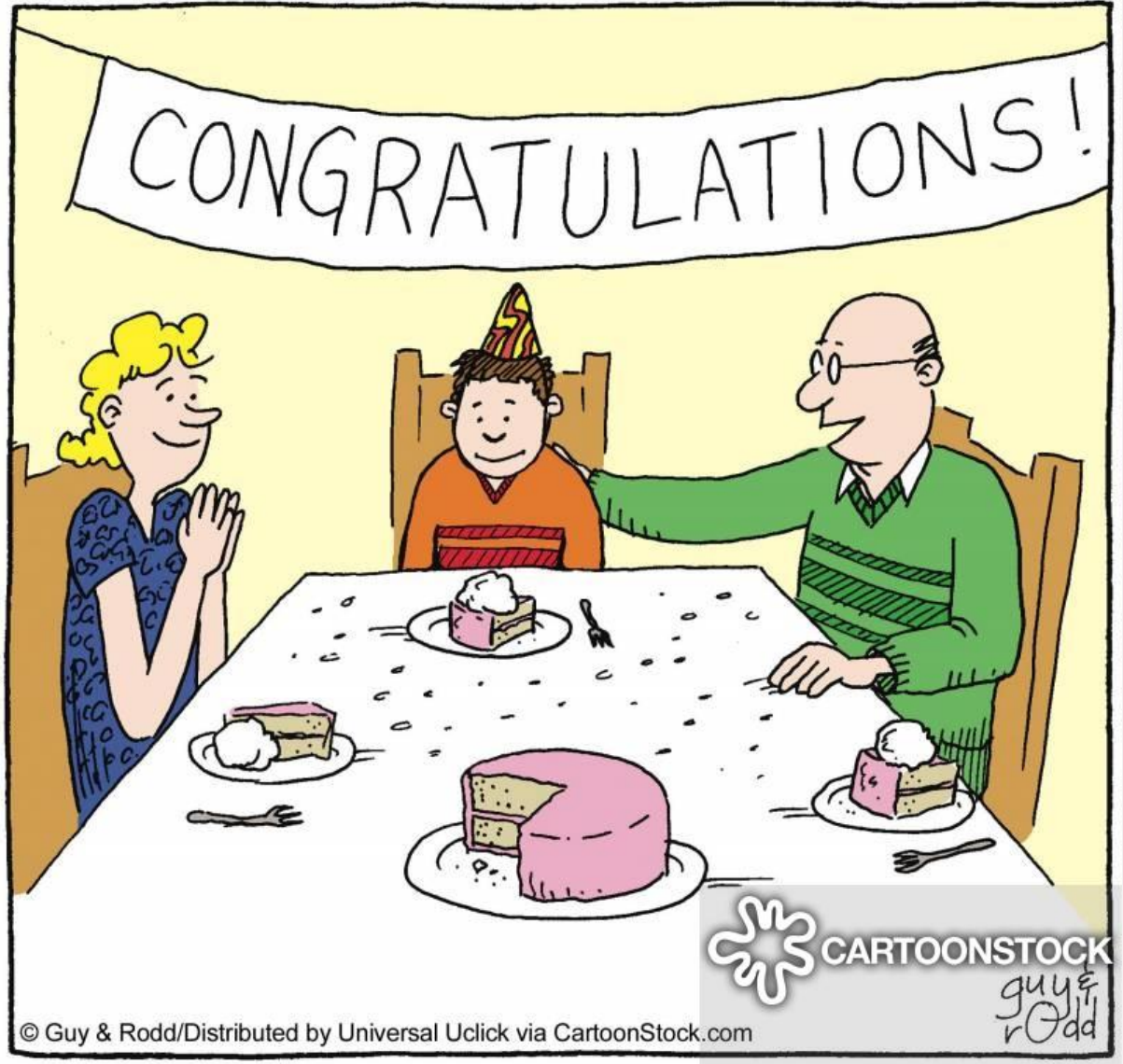
Modelação Ecológica

AULA 21

27th November 2019

No goodies… lame performance!



THE CAKE WAS GREAT AND THE ICE CREAM WAS DELICIOUS, BUT DEEP DOWN INSIDE, HE KNEW THAT SOME DAY HIS PARENTS WOULD DISCOVER THAT "F" WASN'T FOR FANTASTIC, AND THEN NONE OF IT WOULD BE WORTH IT.

Dealing with correlation
Random Effects, Mixed Models & Generalized Estimating Equations

Wrapping up Mixed Models

AN EXAMPLE WITH TWO FACTORS

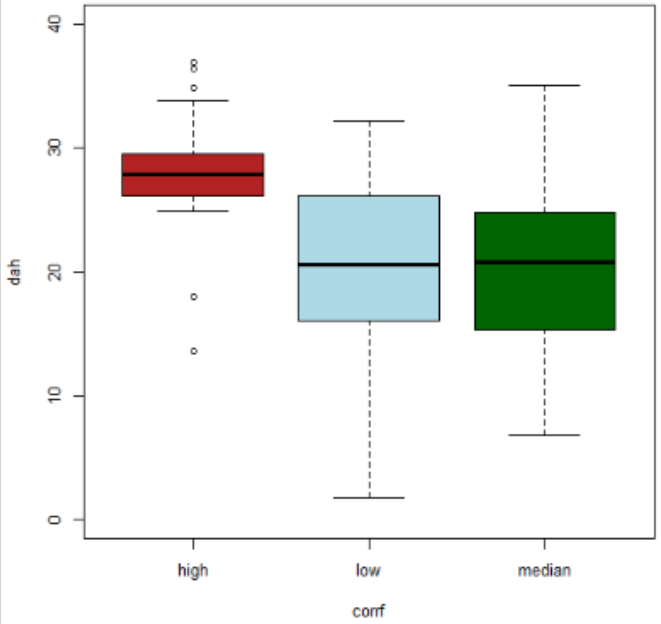
- Density of *Anaecypris hispanica* as a function of current velocity (`corrf`) and river (`riverI`) - dataAHD.txt
- One might be random… you have to explore
- Create a small report describing the data
- Model *A. hispanica* density as a function of the covariates
- Take you conclusions

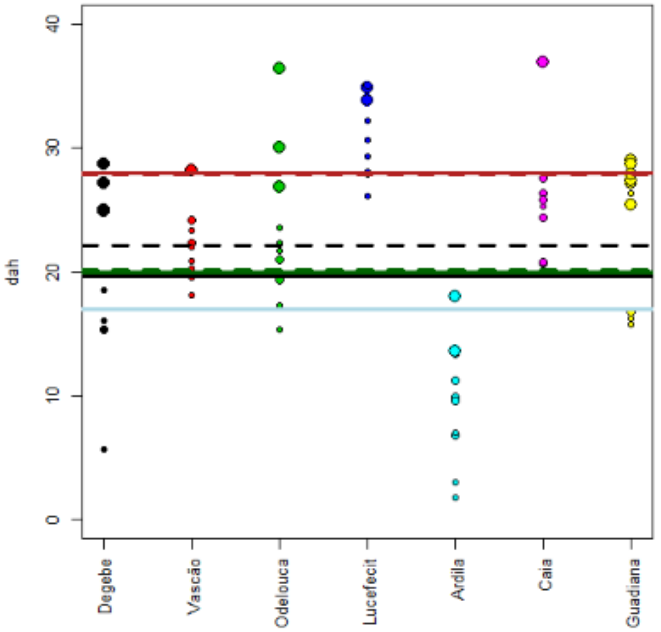
```
> head(dataAHD)
       dah riverI riverI2  corrf     corr
1 28.68104 Degebe       1   high 9.138035
2 24.91680 Degebe       1   high 7.011055
3 15.30073 Degebe       1 median 5.146127
4 16.01281 Degebe       1    low 1.667494
5 27.19677 Degebe       1   high 7.318869
6 25.01242 Degebe       1   high 9.406168
```

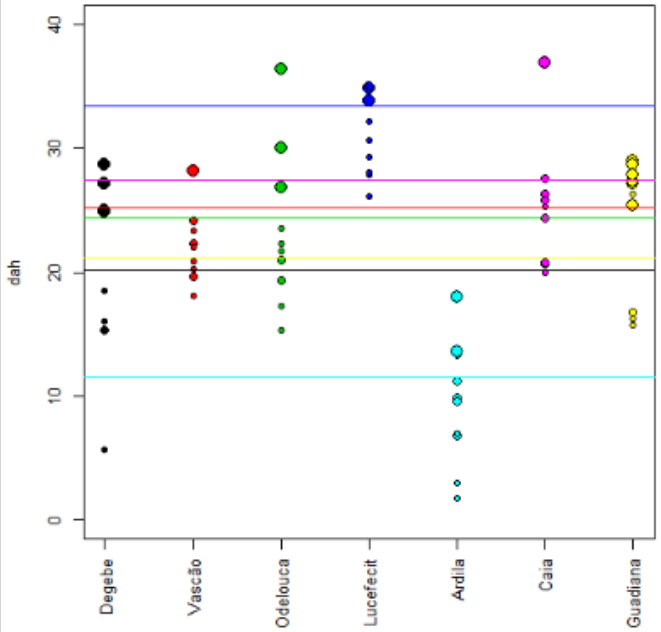
At the end of the class I'll give you (I've given you ;) my code that allows you to see how I generated the data and how the different models retrieve different components of the "truth".

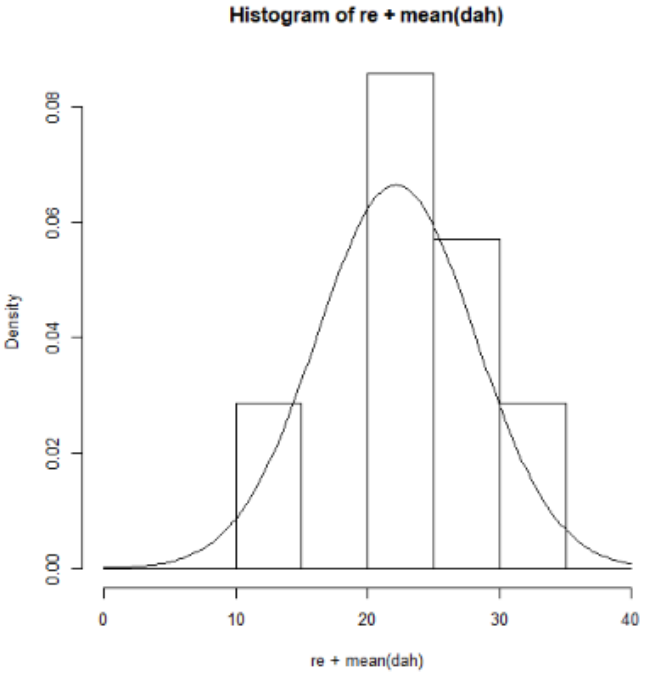
```r
#Data are made up - just for biological context

#density of A hispanica dah in stretches of river
set.seed(444)
set.seed(555)
rivers=c("Degebe","Vascão","Odelouca","Lucefecit","Ardila","Caia","Guadiana")
#number of rivers
nr=length(rivers)
#observations per river
obr=10
riverI=rep(rivers,each=obr)
riverI2=rep(1:nr,each=obr)
#total number of obvservations
n=nr*obr
#standard deviation of the random effect
sdre=6
re=rnorm(nr,mean=0,sd=sdre)
#velocity of current, in m por minuto
corr=runif(n,0,10)
#there's about 1/3 of each type of river
corrf=ifelse(corr<10/3,"low",ifelse(corr>(2*10/3),"high","median"))
#standard deviation of the error = residuals
sderro=3
#density of anaecypris
dah=
14+
3*(corrf=="low")+
6*(corrf=="median")+
14*(corrf=="high")+
(riverI==rivers[1])*re[1]+
(riverI==rivers[2])*re[2]+
(riverI==rivers[3])*re[3]+
(riverI==rivers[4])*re[4]+
(riverI==rivers[5])*re[5]+
(riverI==rivers[6])*re[6]+
(riverI==rivers[7])*re[7]+
+rnorm(n,mean=0,sd=sderro)
```


dah
0
10
20
30
40
high
low
median
corrf


dah
0
10
20
30
40
Degebe
Vascão
Odelouca
Lucefecit
Ardila
Caia
Guadiana


dah
0
10
20
30
40
Degebe
Vascão
Odelouca
Lucefecit
Ardila
Caia
Guadiana


Histogram of re + mean(dah)
Density
0.00
0.02
0.04
0.06
0.08
0
10
20
30
40
re + mean(dah)

Using velocity as a factor variable in the mixed effects model

14+14=28

14+3-(14+14)=3-14=-11

14+6-(14+14)=6-14=-8

```
> summary(lme(dah~corrf,random=~1|riverI,data=dataAHD))
Linear mixed-effects model fit by REML
 Data: dataAHD
       AIC      BIC    logLik
  391.4507 402.4742 -190.7254

Random effects:
 Formula: ~1 | riverI
        (Intercept) Residual
StdDev:    6.811164 3.28974

Fixed effects: dah ~ corrf
                Value Std.Error DF   t-value p-value
(Intercept) 28.588381  2.695333 61 10.606622       0
corrflow    -9.906773  1.019084 61 -9.721255       0
corrfmedian -7.305759  1.159858 61 -6.298837       0
 Correlation:
            (Intr) crrflw
corrflow    -0.235
corrfmedian -0.220  0.563

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-3.06202233 -0.68268447  0.04239849  0.58462158  2.57127416

Number of Observations: 70
Number of Groups: 7
```

We could also compare the estimated value of the random effect associated with each river, and the true random effect value used in the generation of the data

```
> unique(round(fitted(lme1,level = 1)-fitted(lme1,level = 0),2))
[1]  -2.96   1.35   1.18   9.59 -12.30   3.96  -0.82
> round(re,2)
[1]  -1.98   3.02   2.25  11.33 -10.68   5.31  -0.94
```

Understanding this bit of code is actually quite complicated! It requires:

1. looking at "Aula 18" to understand what the different levels of a prediction from a mixed model correspond to
    1. level=0 – population level
    2. level=1 – level of the random effect
2. Realizing that the difference between those gives you the estimated random effect
3. Realizing that the rounding is used just to make the visualization more digestible (and avoiding issues with rounding)

A CONCLUSION ABOUT FIXED vs. RANDOM EFFECTS

- Whether a factor is included as a random effect or not is often a philosophical question. If
  1. One is interested in the specific levels of the factor (e.g. each of the rivers) then it should be a fixed effect
  2. One is interested in the variability across the different levels of the random effect, but not on each river per sem then it should be a random effect

The discussion and the conclusions will necessarily be different!

A CONCLUSION ABOUT RANDOM EFFECTS

- Random effects are often used to "soak up" variation that exists in the data but which we can't describe
- In fact, river is not really a random factor at all (say what?) It is just a proxi for stuff we can't explain!
- What happens is that there are some differences across rivers, e.g.
  - some have dams and some don't,
  - some are wide and some are narrow,
  - some are surrounded by forests some by agricultural fields,
  - Etc.

A CONCLUSION ABOUT RANDOM EFFECTS

- If we had all the (relevant) variables, once these were all included in a model, we would not need river as a random effect
- But because we never do, this is a useful way to remove some of the variability that otherwise unexplained would end up in the error term, but in this way is explained by the random effect
- As a consequence, it makes it more likely that we will find relevant variables amongst the ones we have collected… and that is a good thing ☺

A NOTE ON ML vs. REML

- When fitting mixed models the default is to use REML, not ML
- This is because ML (for technical reasons beyond what I want to torture you about – 4trbwlwttya) produces biased estimates of variances
- REML is shown to have better properties
- However, REML does not allow you to do (in general, it does under certain conditions 4trbwlwttya) likelihood ratio tests, and so model selection might be harder with REML
- This is why e.g. Zuur et al. 2009 recommend the procedure that was referred to in "Aula 18", slide 31, now slightly updated in the next slide

AULA 18, SLIDE 31, UPDATED

Model selection in a mixed model context (a possible top-down approach)

1. Start from a full model with all relevant fixed effects
2. Find best random structure (e.g. via AIC, or because it is the structure that respects your experiment. Comparing two models with nested random structures cannot be done with ML because the estimators for the variance terms are biased)
3. Conditional on that random effect structure, select the relevant (fixed) effects (To compare models with nested fixed effects (but with the same random structure), ML estimation must be used and not REML)
4. Present the final model using REML estimation

e.g. recommended by Diggle et al. (2002) The Analysis of Longitudinal Data. Oxford University Press, see also page 122 of Zuur et al. 2009

Dealing with correlation
Random Effects, Mixed Models & Generalized Estimating Equations

Generalized Estimating Equations

When we consider a LM, GLM or a GAM, we assume the data are independent

This is often not the case, and not accounting for the **correlation structure** will tend to lead to errors

So… why is it relevant to account for the correlation structure?

1. It does often not change much the parameter estimates… but
2. It changes the variance of the parameters, which means that inferences might change!

In particular, we will often find significant predictors than we should!

Why: because with positive correlation (the most common case) we think we have more data than we actually have!

Strong (positive) correlation ⟶ Smaller effective sample size

When we consider a LM, GLM or a GAM, we assume the data are independent

Recall for independent data, the error term: $\mathbf{e} \sim Normal(0, \sigma_e^2)$ is the same as $Normal(0, \sigma_e^2 \mathbf{I})$ where $\mathbf{I}$ is a $N \times N$ diagonal matrix (which means it has a diagonal of 1's and zeros in the off diagonals):

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & ..... & 0 \\ 0 & 1 & 0 & 0 & 0 & ..... & 0 \\ 0 & 0 & 1 & 0 & 0 & ..... & 0 \\ 0 & 0 & 0 & 1 & 0 & ..... & 0 \\ 0 & 0 & 0 & 0 & 1 & ..... & 0 \\ . & . & . & . & . & ..... & . \\ . & . & . & . & . & ..... & . \\ 0 & 0 & 0 & 0 & 0 & ..... & 1 \end{bmatrix}$$

Generalized Estimating Equations (GEE's)

GGEs represent an alternative to mixed models, where you model the relationship between the mean value and the variance (of a response variable), not the actual distribution of the data

These are also called marginal models or population averaged models, because in this case you are not interested in the response at the level of the "random effects" (if so you need a GLMM or a GAMM), you are just interested in modelling the response at the population level but accounting for the **adequate correlation structure** present in the data.

Generalized Estimating Equations (GEE's)

GEEs can be used to analyze repeated measurements (either or not over time, in the later case often called longitudinal repeated measurements) data.

Much material in these slides was blatantly stolen from material kindly shared by my good friend Monique MacKenzie – so many thanks Mon ☺

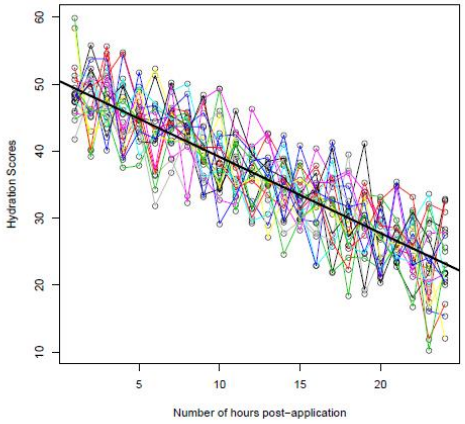


Figure 63: Scatter plots of the number of hours post-application vs hydration scores for independent data and with correlated errors for 20 subjects.

https://moniquemackenzie.wixsite.com/drmoniquemackenzie

These can be used to model continuous, binary, proportional, or count data (so, essentially the same type of data we have already dealt with in a GLM or GAM framework).

A good way to model response variable accounting for correlation structures in the data when we are not really interested about the random effects is using GEEs

	package	function
GEEs can be implemented in R via:	`geepack`	`geeglm`

R syntax

```
geeglm(formula, id, data, corstr, family)
```

- defines groups
- defines the correlation structure
- defines mean-variance relation

GEEs can include linear terms but also non-linear smooth terms… but that is not trivial to do.

So the key thing is to decide what is the expected correlation structure inside each "unit"

For illustration, imagine we have just 4 observations for an individual ($n_i = 4$), and an AR(1) model is a sensible representation of the correlated pattern within individuals. This would give a $n_i \times n_i$ AR(1) block ($n_i = 4$):

$$\begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$

So, if our data set contained just 2 individuals and 4 observations for each, the 8×8 block-diagonal correlation matrix would look like:

$$\begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & 0 & 0 & 0 & 0 \\ \rho & 1 & \rho & \rho^2 & 0 & 0 & 0 & 0 \\ \rho^2 & \rho & 1 & \rho & 0 & 0 & 0 & 0 \\ \rho^3 & \rho^2 & \rho & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & \rho & \rho^2 & \rho^3 \\ 0 & 0 & 0 & 0 & \rho & 1 & \rho & \rho^2 \\ 0 & 0 & 0 & 0 & \rho^2 & \rho & 1 & \rho \\ 0 & 0 & 0 & 0 & \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$

Each one of these is a block of the block correlation matrix

e.g. For two medical centres, each with 4 individuals measured just once each, an exchangeable/compound symmetry structure looks like:

$$\begin{bmatrix} 1 & \rho & \rho & \rho & 0 & 0 & 0 & 0 \\ \rho & 1 & \rho & \rho & 0 & 0 & 0 & 0 \\ \rho & \rho & 1 & \rho & 0 & 0 & 0 & 0 \\ \rho & \rho & \rho & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & \rho & \rho & \rho \\ 0 & 0 & 0 & 0 & \rho & 1 & \rho & \rho \\ 0 & 0 & 0 & 0 & \rho & \rho & 1 & \rho \\ 0 & 0 & 0 & 0 & \rho & \rho & \rho & 1 \end{bmatrix}$$

Journal of Statistical Software
January 2006, Volume 15, Issue 2. http://www.jstatsoft.org/

The R Package geepack for Generalized Estimating Equations

Ulrich Halekoh
Danish Institute of Agricultural Sciences

Søren Højsgaard
Danish Institute of Agricultural Sciences

Jun Yan
University of Iowa

Defining the relation between mean value and variance is via argument `family`

4.1. Variance and link functions (`family`)

The variance function is specified by the **`family`** argument in **`geeglm`** and is identified by the name of the corresponding distribution in a generalized linear model. The available variance functions are given in Table 4. The available link functions for the mean are the same as those in **`glm`** with the exception of the **`cauchit`** link for the **`binomial`** family.

name	function $v(\mu)$
`gaussian`	identity
`binomial`	$\mu(1-\mu)$, $\mu \in (0,1)$
`poisson`	μ, $\mu > 0$
`gamma`	μ^2, $\mu > 0$

Table 4: Variance functions in **`geeglm`**.

Journal of Statistical Software
January 2006, Volume 15, Issue 2. http://www.jstatsoft.org/

The R Package geepack for Generalized Estimating Equations

Ulrich Halekoh
Danish Institute of Agricultural Sciences

Søren Højsgaard
Danish Institute of Agricultural Sciences

Jun Yan
University of Iowa

Defining the relation between mean value and variance is via argument `family`

4.1. Variance and link functions (`family`)

The variance function is specified by the **`family`** argument in **`geeglm`** and is identified by the name of the corresponding distribution in a generalized linear model. The available variance functions are given in Table 4. The available link functions for the mean are the same as those in **`glm`** with the exception of the **`cauchit`** link for the **`binomial`** family.

Defining the correlation structure is via argument `corstr`

name	$R(\alpha)$
`independence`	$\mathrm{COR}(Y_{it}, Y_{it'}) = 0, \quad t \neq t'$
`exchangeable`	$\mathrm{COR}(Y_{it}, Y_{it'}) = \alpha, \quad t \neq t'$
`ar1`	$\mathrm{COR}(Y_{it}, Y_{it'}) = \alpha^{\lvert t-t' \rvert}$
`unstructured`	$\mathrm{COR}(Y_{it}, Y_{it'}) = \alpha_{tt'}, t \neq t'$

Table 5: Working correlations in **`geeglm`**.

Cada elemento dos blocos tem uma correlação (potencialmente) diferente

Unstructured correlation matrix…

$$\begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1K} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{K1} & \cdots & \sigma_{K,K-1} & \sigma_K^2 \end{bmatrix}$$

- For a p-dimensional covariance matrix, $p(p+1)/2$ parameters are required, becoming large very rapidly as p increases

This is the hardest to fit and not recommended if you don't know exactly what you are doing. The large number of parameters means that the model might become unstable.

A real life example

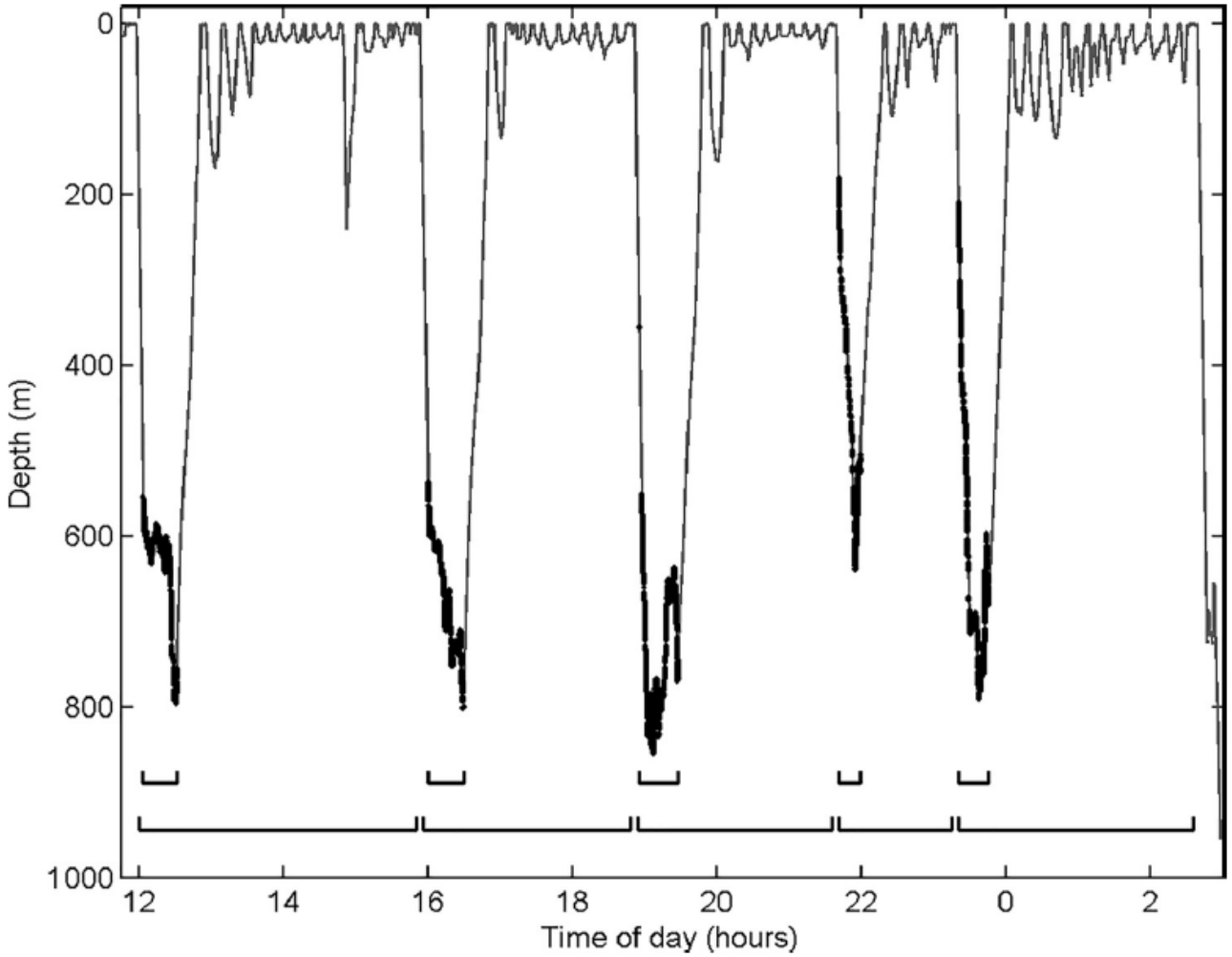


FIG. 1. Example dive profile of a Blainville's beaked whale tagged in the waters adjacent to El Hierro, Canary Islands. Bold sections indicate the presence of foraging clicks. Shorter, upper markers delineate vocal periods, while lower, longer markers indicate the lengths of individual dive cycles. The final dive featured tag detachment and was not analyzed.

Warren, V. E.; Marques, T. A.; Harris, D.; Tyack, P. L.; Thomas, L.; de Soto, N. A.; Hickmott, L. & Johnson, M. P. 2017 Spatio-temporal variation in click production rates of beaked whales: implications for passive acoustic density estimation The Journal of the Acoustical Society of America 141: 1962-1974

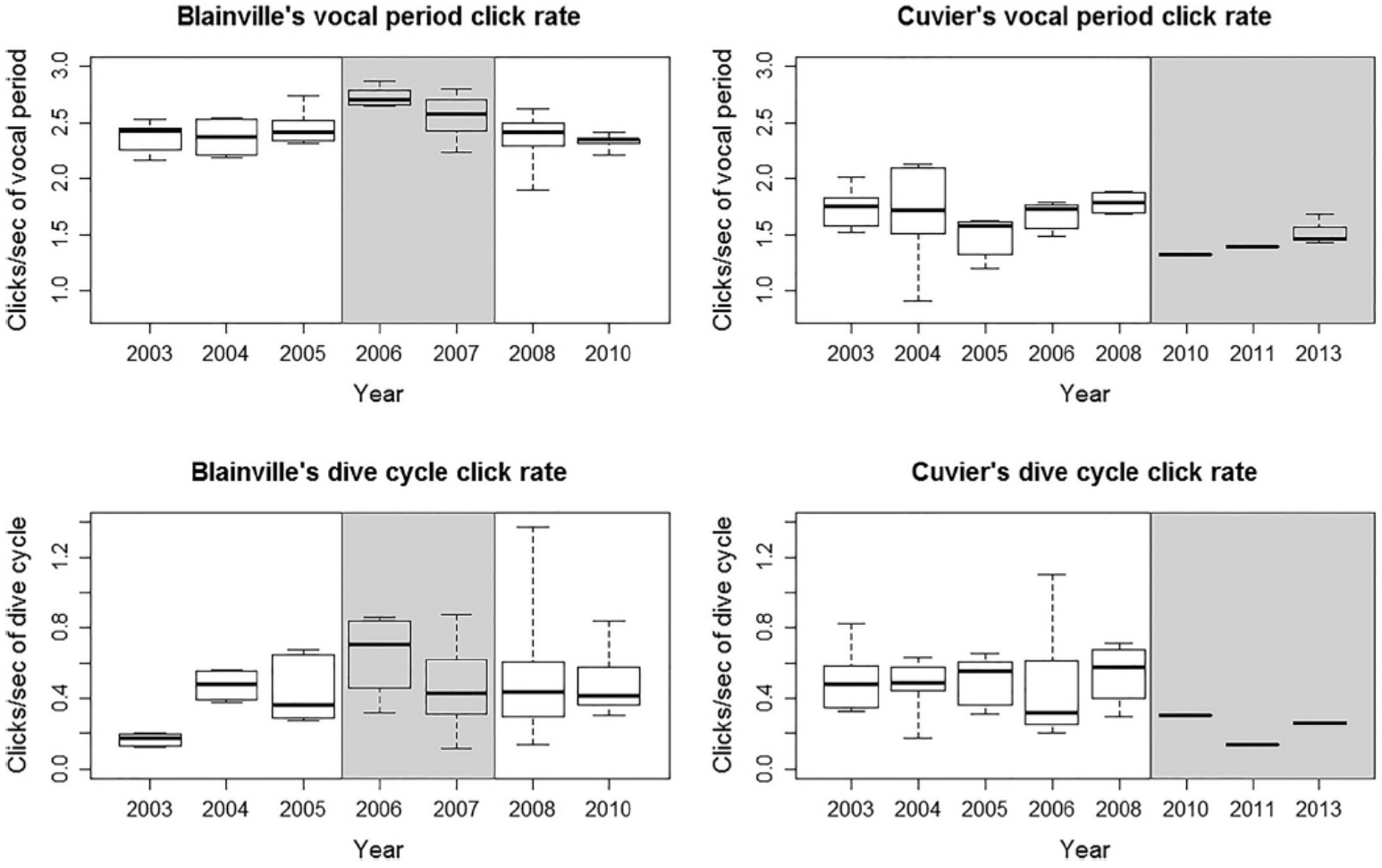


FIG. 4. Inter-annual variation in vocal period and dive cycle click production rates for Blainville's (left) and Cuvier's (right) beaked whales. Box plots consist of median, interquartile range and maximum/minimum extremes. In the Blainville's data, boxes in white areas represent animals tagged in El Hierro and boxes in grey areas (2006 and 2007) indicate tags deployed in the Bahamas. In the Cuvier's plots, boxes in the white area represent Liguria, and boxes in the grey area (2010, 2011, and 2013) are southern California deployments. See Table I for respective sample sizes. Y axes scales differ between vocal period plots (upper) and dive cycle plots (lower).

Warren, V. E.; Marques, T. A.; Harris, D.; Tyack, P. L.; Thomas, L.; de Soto, N. A.; Hickmott, L. & Johnson, M. P. 2017 Spatio-temporal variation in click production rates of beaked whales: implications for passive acoustic density estimation The Journal of the Acoustical Society of America 141: 1962-1974

Runs tests revealed the presence of weak autocorrelation within model residuals due to longitudinal sampling, i.e., multiple observations of the same animal over time. Generalized Estimating Equations (GEEs) were therefore used in R (version 3.3.1; package "geepack," version 1.2–0; R core Team, 2015; Højsgaard *et al.*, 2006), with "Tag ID"

• • •

2016, for a similar approach). GEEs are appropriate for data containing a large number of clusters (tag deployments) with relatively few observations (dives or dive cycles) per cluster (Bailey *et al.*, 2013).

Warren, V. E.; Marques, T. A.; Harris, D.; Tyack, P. L.; Thomas, L.; de Soto, N. A.; Hickmott, L. & Johnson, M. P. 2017 Spatio-temporal variation in click production rates of beaked whales: implications for passive acoustic density estimation The Journal of the Acoustical Society of America 141: 1962-1974

TODAY's TASK

Revisit two datasets from FT7b4ME 20 11 2019.pdf in "Aula 19"

7. Find a GLM that best fits the data "Owls.txt", where you are trying to explain the begging behavior of owls offspring when the parents are absent from the nest. The variable "SiblingNegotiation" represents the number of calls produced by the chicks in the nest during a 30 second period, while "BroodSize" represents the size of the brood. More details about this data can be found in Zuur et al. 2009.

Account for variation over time in the same nest

8. The data "DeerEcervi.txt" contains the incidence of *E. cervi* parasites in deer pellets, and we have also the corresponding sex, length and farm the deer were on. How many farms were available? Ignore them for now, and model the presence/absence of parasites in pellets as a function of deer characteristics. This is a dataset also used by Zuur et al. 2009.

Account for variation across farms

If you are looking at this slide,
I have pressed click one time too many

!