

LINEAR MODELS

Exam Project

KU Leuven - academic year 2021-2022

Practical information



The project consists of the analysis of one dataset using the tools you have learned **throughout the course**. The dataset is available on Toledo. In case of practical questions, please send an email to thomas.neyens@kuleuven.be, thomas.neyens@uhasselt.be, **and** alejandro.rozoposada@uhasselt.be. We will not give feedback on the methodology to be used to conduct these analyses.

Writing guidelines


The report should contain no more than 8 sheets in total, including the title page - a sheet has two pages, i.e., if you print your report it should not exceed 8 sheets (16 pages). Please use an A4 page format, a Calibri 12 font and a 1.2 spacing between lines. Each group sends an electronic copy of the report to both professors (thomas.neyens@kuleuven.be, thomas.neyens@uhasselt.be, **and** ariel.alonsoabad@kuleuven.be), as well as the R code used in the analysis at latest on **Dec 28 (before 23:00)**. The title page of the report should contain the number of the group and a list with the names and student numbers of all the members of the group.


The report should contain a description of the analysis of the case study, a discussion of the results, model checking, etc. Think about it as a report that you will give to a client. As detailed below, the study at hand is positioned within an ecological setting. You do not have to write a detailed overview of ecological concepts or a literature study of the species under investigation, but an introductory paragraph can be good to set the scene. You can look for alternative analyses and techniques in the literature, if the models studied in the course are not appropriate to answer the scientific questions of this study. The idea is to mimic the real work of a statistical consultant in academia or the industry.

Study background

The dwarf squeaker (*Arthroleptis xenodactyloides*) is a frog that lives in cloud forests of the Eastern Arc Mountains in Africa. Its ecology is not well understood, but researchers believe that is a generalist, i.e., a species that can thrive in many habitats. They believe that the species benefits from holes in the cloud forest's canopy or shrub layers. These holes allow sunlight to reach and warm up the ground floor of the forests, which this particular species, in contrast to others, is expected to benefit from. These holes, especially those in the canopy, can exist naturally but are appearing in increasing numbers, due to recent illegal logging of single trees.

A biologist studies this species in Kenya. She has investigated a large number of sites in three forests in which the species is prevalent in large numbers (Ngangao South, Ngangao North, and Chawia; all forests lie far enough from each other to assume spatial independence). These sites are typically named *patches*, the longitude-latitude location of which was randomly chosen. She visited these coordinates and defined the patch as the area surrounding this randomly sampled location in which she would be able to search for the species (e.g., rock or dense tree formations would make this difficult). The size of each patch was then measured. At each patch, she searched until she found one individual, which she examined in her field lab (and re-released afterwards). For each individual, the following data are available for you to analyse. Note that due to their large abundance, she always found an individual in each patch.

Variable	Description
Length	Body length (cm)
Sex	<i>male</i> or <i>female</i>
Size	Size of patch (m ²)
Canopy	Proportion of patch covered by canopy 
Shrub	Proportion of patch covered by shrubs
Effort	The time it took to find the individual (minutes)
Natural	Indicator whether the site was naturally intact, i.e., unspoilt by any human activity (<i>yes</i> or <i>no</i>)
Forest	<i>Ngangao N</i> , <i>Ngangao S</i> , or <i>Chawia</i>

Canopy and shrub covers are based on image analysis of a picture of the respective canopy and shrub layers at each patch. Effort is not an ecological variable, but it is likely associated with the prevalence of the species, possibly that of large individuals since they are assumed to be found more easily. 

Tasks

The goal of this analysis is to build a model of *Length* as a function of the other variables that are available in your dataset. As a group, you should build the best model to analyse the data, and based on that, discuss the results that you have obtained (the model, the effects of the predictors, etc.). You should build the model via a training dataset and validate it via a validation dataset. Do this as follows:

```
data.full = read.table("dataset.txt",header=T)
set.seed(01995xx)
d.test <- sample(1:dim(data.full)[1], round(dim(data.full)[1]/2) )
data.test <- data.full[d.test, ]
data.training <- data.full[-d.test, ]
```

In this code, ***xx*** should be replaced by your group number. Pick a selection of three best training models to test at the validation stage. When building the model, do not forget to carefully investigate model assumptions and other problems that may violate your results.