Regression Analysis

Exam Project

KU Leuven - academic year 2019-2020

Practical information

The project consists of the analysis of one dataset by using the tools you have learned throughout the course. The dataset is available on Toledo. Descriptions of the dataset and tasks are provided below. **The report and an appendix should be uploaded as 1 pdf document on Toledo on 5 January 2020 at the latest**. It is not necessary to send your report by email.

Writing guidelines

The report should be written in an **answer-to-question format**. You should **not** provide an introductory section, goals of the study, and general conclusions. Each answer to a question should contain an overview of the relevant results and a discussion of those specific results. A rationale behind question-specific analysis choices should be provided when appropriate. Do not copy-paste R output into the report (except figures), but similarly as in journal articles, e.g., present the important results in a table, which you add and refer to in the text. Do not recite theory from the course. The report should not exceed 5 pages (font size 12 pt), including the relevant figures and tables. Attach your R code as an appendix. Do not report R code within the actual text. Moreover, make sure the code is structured and readable.

Questions

Questions can be posted on the discussion forum on Toledo, or in case of questions that are specific to your unique dataset (more info about this below), you can e-mail me (thomas.neyens@kuleuven.be). Note that I will check Toledo/my e-mails only sporadically from Dec 25 until Jan 5.

Study background

Consider the dataset *invertebrate.txt*. It holds 400 records of flying nocturnal invertebrate (e.g., insects such as moths, beetles, mosquitos, etc.) biodiversity, based on 400 sampling events in 2017 in the Belgian province of Limburg. Each sampling event was carried out on a patch, being a small nature area that has been subject to nature management for at least 3 months. Each patch was visited only once (so the 400 data points reflect 400 different locations). We assume spatially uncorrelated data, as is indicated by an exploratory test for spatial autocorrelation (not shown). Sampling was carried out by catching the invertebrates in a net in the evening. Sampling events were carried out from the beginning of Spring until the middle of Summer, when sampling stopped because of extreme heat. Note that multiple sampling events could be carried out on the same day by different persons. Also note that the dataset is ordered chronologically. However, information about the exact dates is not provided.

The outcome of interest is the Shannon-Wiener index (SWI), which is a measure of biodiversity. Within the context of this project, it suffices to know that SWI is a non-negative metric that is usually in practice smaller than 4.5. Low values denote low diversity, while higher values denote higher diversity. In addition to SWI, a number of explanatory variables are collected as well. An overview of all variables in the data is given in Table 1.

Table 1: Biodiversity data: variable overview.

variable name	explanation
SWI	the Shannon-Wiener index for (flying nocturnal) invertebrate diversity on the patch (non-negative, larger values denote higher diversity).
SWF	an (adjusted) Shannon-Wiener index for floristic diversity on the patch. The interpretation of this metric is the same as for SWI (non-negative, larger values denote higher diversity).
temperature	temperature at the sampling event (in degrees Celsius)
size	the size of the sampling patch (in m^2)
management	the number of years that the patch has been subject to nature management
duration	the duration of a sampling event (in minutes)

Tasks

The goal of this analysis is to build a model of SWI as a function of SWF, temperature, size, and management. You will build the model via a training dataset and validate it via a validation dataset, containing all variables (including duration). Do this as follows:

```
data.full = read.table("invertebrate.txt",header=T)
set.seed(0012345)
d.test <- sample(1:dim(data.full)[1], 200 )
data.test <- data.full[d.test, ]
data.training <- data.full[-d.test, ]
```

In this code, **0012345 should be replaced by your student number**. Next, answer to the following questions:

Note that the variable *duration* should not be considered until you reach question 6!

1. **Training dataset**: before fitting any model, conduct an exploratory data analysis and report peculiarities that might be important to account for in the analysis.
2. **Training dataset**: fit a linear first-order regression model with SWI as outcome and SWF, temperature, size, and management (not duration!) as predictors.
   (a) Check whether a first-order model adequately captures the variability in the outcome.
   (b) Check the Gauss-Markov conditions (except independence, which we assume to be met).
   (c) Check whether there is (severe) multicollinearity.
   (d) Check whether there are influential outliers. Use both traditional methods and a robust diagnostic plot.
3. **Training dataset**: based on the results in the previous question, build a good linear regression model. This model may contain higher-order terms, interactions, transformed variables and/or other methods to improve the model assumptions. A few remarks:
   - Focus primarily on remedial methods for assumptions in (a) and (b) from the previous question.

- In case of severe multicollinearity (c), delete one of the highly correlated variables from the model. Note that this choice is made within the context of this exercise and is not necessarily the best option in other data analyses.
- Since you have no information about the sampling process, you should not delete influential observations from the data. Due to incompatibility of the ltsReg function for robust regression with specific methods that remedy departures from Gauss-Markov assumptions, you should here only consider traditional methods to pinpoint influential observations. Reflect whether possible problems with influential outliers have changed by remedying other departures from model assumptions.
- If you do not succeed in meeting all the conditions, which is possible, explain the shortcomings of your model.
- Note that in most cases there is not 1 correct answer, but some choices are obviously better than others.

Variable selection can, when appropriate, be used to pinpoint the best model. In addition, pick a model of your second choice.

4. Fit both models to the **validation data**. Investigate and compare their performance in the way that you consider to be appropriate for the models at hand.

5. Based on the validation process, fit your ultimate model of preference to the **full dataset**. Which insights do you gain? Which variables have influence on SWI and how?

6. Consider again the **training dataset**: We are interested in investigating possible association between duration (outcome) and temperature (predictor).

   (a) Fit non-parametric models with k=1 and k=2, for spans 0.25, 0.5, and 0.75 and choose the best-fitting model.

   (b) Fit a quadratic linear model.

   (c) Provide a plot of the data, the best nonparametric fit, and the linear fit. Which model fits the data best, according to your visual interpretation?

   (d) Test whether the non-parametric model of your choice fits the data better than the quadratic model.

   Only do this for the training dataset.