ST523/ST813 - Statistical Modelling

E2024

Home Assignment 2

This assignment should be **submitted electronically** as a single **pdf file** via **itslearning.** See itslearning for the submission deadline.

It is expected that you **work independently** on the assignment. You are not permitted to discuss the solutions with anyone (this includes your fellow students as well as posting questions about the exercises on the internet - OBS: the latter does not refer to the use of the exam-discussion board on the course's itslearning page) or to use generative models such as ChatGPT. Any suspicious similarities of the solutions or any other concerns will be investigated and interactions will be considered as exam fraud.

Unless stated differently, **R** can be used to solve the exercises. Then, the output should be explained in detail, and all replies to a question should be stated separately in the text. Output alone without further explanation will not be accepted as valid answer.

Your submission should not exceed 8 pages.

Exercise 1

Biomass is an essential climate and biodiversity variable which is referred to in six UN Sustainable Development Goals. Investigations of biomass are therefore of high interest in environmental science. Forests are an important part of terrestrial vegetation and field measurements are the most accurate ways to learn about forest biomass. However, field measurements are very labour- and cost-intensive. See `https://www.nature.com/articles/sdata201770` for more details.

This exercise focusses on statistical modelling of foliage biomass which describes biomass contained in leaves and needles. The main interest lies in the development of a statistical model that can be used to relate foliage biomass to more easily measurable quantities.

The data used for this exercise was created on the basis of the following public database `https://doi.pangaea.de/10.1594/PANGAEA.871491`.
In this exercise we work with a smaller subset of the published data that has been slightly modified. Especially, our data represents measurements from individual trees in Eurasia of the species Pinus sylvestris.

To access the data for this exercise, save the file `Data_Assignment2_Ex1_E2024.rdata` in your **R** working directory and type the following command in **R** :
`load("Data_Assignment2_Ex1_E2024.rdata")`

The dataset includes the following variables:

`biomfoliage`	live biomass from leaves/needles (in kg)
`age`	age (in years)
`DBH`	diameter at breast height (in cm)
`height`	tree height (in m)
`origin`	categorical (natural vs. planted)
`nrtrees`	number of trees per hectar at the location
`latitude`	geographical latitude (8N to 72N)
`longitude`	geographical longitude (8W to 160E)
`ecoregion`	categorical indicator for ecoregions, see `https://en.wikipedia.org/wiki/Ecoregion`
`country`	country code (KAZ=Kazakhstan, RUS=Russia, UKR=Ukraine)

1. (Data Exploration) Explore and describe the dataset.

   a) Create a table containing relevant summary measures for each of the variables individually.

   b) Make suitable plots that show the relation between foliage biomass and the other variables.

2. (Model Selection) Select a suitable generalized regression model for `biomfoliage` that can be used to relate foliage biomass to the other available variables.

   Thereby, give explicit details about following aspects:

   a) State and explain which main classes of GLM are applicable in this situation,

   b) as well as the link functions you considered together with these.

   c) In case there should be several possible alternatives for a) and b), use appropriate methods from the course to make an informed choice between the different alternatives. Explain your reasoning and provide details.

   d) Apply appropriate methods from the course to investigate the functional form in which explanatory variables are included in the linear predictor and whether selected interactions should be taken into account.

   e) Select a parsimonious model, i.e. reduce model complexity and number of included explanatory variables as much as possible. Explain which principles you use for model selection.

3. (Final model)

   a) Represent your final model, which you would use as a substitute for direct measures of foliage biomass. Represent this model in a table specifying estimates, standard errors and p-values (care about good readability and use roundings).

   b) Write down the resulting formula for the calculation of fitted values from your final model.

   Which fitted value do you obtain for a 20-year old tree with DBH=$10cm$, height=$10m$, from a natural Russian forest at a location with 1000 trees per hectar, at longitude 85, latitude 56 and ecoregion 80817.

   c) Make an exemplary plot that can illustrate your model predictions: e.g. show the relation between `DBH` and `BiomFoliage` while keeping other variables fixed.

   d) Which values do you obtain for the total deviance and AIC?

e) What is the estimated dispersion parameter? Does this value indicate over- or underdispersion?

Exercise 2

Consider the family of transformations given by

$$g_\alpha(Y) = \begin{cases} \frac{e^{\alpha Y}-1}{\alpha} & \text{for } \alpha \neq 0 \\ Y & \text{for } \alpha = 0 \end{cases}$$

for $Y \in \mathbb{R}$. Analogous to the estimation of the Box-Cox parameter λ, the parameter $\alpha \in \mathbb{R}$ can be estimated using a profile likelihood approach. Let $Y_1, \ldots, Y_n \in \mathbb{R}$ be independent responses together with corresponding predictors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^p$.

1. Assume for α there exist $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\sigma^2 > 0$ such that $g_\alpha(Y_i) \sim N(\boldsymbol{x}_i^T\boldsymbol{\beta}, \sigma^2)$ for $i = 1, \ldots, n$. Derive the density f_{Y_i} of the untransformed observations Y_i.

2. Write down the log-likelihood function $\ell(\alpha, \boldsymbol{\beta}, \sigma^2; y_1, \ldots, y_n)$, i.e. $\sum_{i=1}^n \log f_{Y_i}(y_i)$.

3. What are the ML-estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ for fixed α? Provide your answers as formulas.

4. Substitute $\boldsymbol{\beta}$ and σ^2 with these ML-estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ to obtain the profile likelihood function
$$l(\alpha; y_1, \ldots, y_n) = \ell(\alpha, \hat{\boldsymbol{\beta}}, \hat{\sigma}^2; y_1, \ldots, y_n)$$
and simplify this function.

5. Write an **R** function that calculates the profile log-likelihood for a given set of observations.

6. Load the dataset `Data_Assignment2_Ex2_E2024.rdata`. The data contains two variables: a response Y and a single predictor x. Consider additionally an intercept term.

   Calculate and plot the profile log-likelihood for different values of α using your function.

7. Determine the ML-estimate of α visually from this plot.

8. Finally, which transformation would you consider for the analysis of the data?