Introduction to Statistical Modeling
Project 2: Non-Linear Modeling

Semester 2, Academic year 2023-24

Contents

1 Data and models

The decomposition of crop residue in soil unfolds in two distinct phases, beginning with a rapid breakdown of easily degradable compounds and microbial cells, followed by a slower phase involving the decay of more resistant crop residue components and stabilized microbial products (see Figure 1).

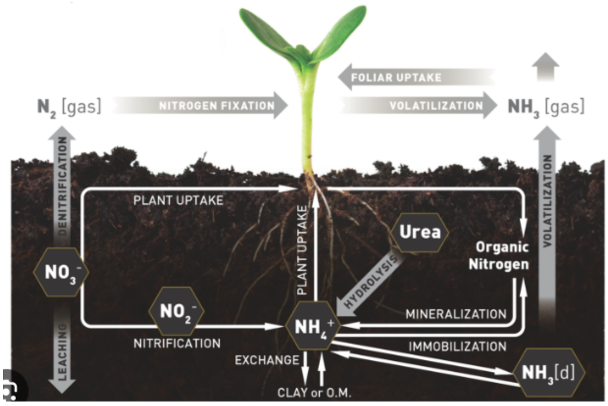

Figure 1: Nitrogen circle.

Research has predominantly focused on the immediate effects of soil nitrogen dynamics post-residue incorporation, yet long-term studies, particularly those extending beyond a decade, remain sparse. In a study by Laberge et al. [2006], the authors assess the decline of residual organic nitrogen and its availability to plants over a period of up to 16.5 years after the residue is added to the soil.

The data from this study is available to you in the file `exp1.dat`. This dataset describes 8 observations over a period of 16.5 years, and for each observation the following variables are recorded:

- `time` – the time (in years) at which the observation was taken.
- `Nremaining` – the remaining residual organic nitrogen.
- `stdev` – a measure of uncertainty for the observation.
- `norep` – this variable will not be used.

For this assignment, you will consider the three nonlinear models listed below. Your task is to ascertain which model provides the best fit for the given data.

1. The **biexponential** model, given by

$$F(t) = A_1 e^{(-e^{k_1 t})} + A_2 e^{(-e^{k_2 t})},$$

   with parameters:

   - A_1 and A_2: initial amplitudes of the decay components.
   - k_1 and k_2: decay constants for the first and second components, respectively.

2. The **Gompertz** model, given by

$$G(t) = Ae^{-ce^{rt}}$$

   with parameters:

   - A: the horizontal asymptote as $t \to +\infty$.
   - c: a constant related to displacement.
   - r: rate of decay.

3. The **exponential** model, given by

$$H(t) = Ae^{-kt} + B.$$

   For this model, you should figure out the interpretation of the parameters A, B, and k by yourself.

2 Data analysis and model fitting

Work out the following questions and compile the code and answers in your RMarkDown script. You are allowed to use functions from the PC practicals, but make sure that the implementation of these functions is included in your RMarkDown script.

1. Read in the dataset and make a plot of the remaining amount of residual organic nitrogen as a function of time.

   Use the column `stdev` to plot error bars for the individual observations. The error bars should extend from the value of the observation minus one standard deviation to the value of the observation plus one standard deviation. To draw the error bars, use the approach described here: `https://stackoverflow.com/a/22037078/394770`.

2. Fit the biexponential and Gompertz models to the data and report the fitted parameters for each model.

   For this question, there are two important things to take into account:

   (a) You must use the `nls` function to do the fitting. Do not implement the objective function and the optimization by yourself.

   (b) You must use a self-starting model for both the biexponential and Gompertz fit. These self-starting models are provided with R; look for functions that start with `SS`.

   One of the PC practicals used a self-starting model with the `nls` command; please review that practical if this question is not clear.

3. Fit the exponential model to the data and report the fitted parameters.

   There is no self-starting exponential model. Instead, implement the model by yourself, select some good starting values for the parameters, and fit the model. You can do the fitting procedure however you want, by using the `nls` function or "manually", by implementing the objective function and finding its minimum, just like you did in the PC practical.

4. Re-create your plot from question 1 and now also plot the three fitted models. Make the plot look nice: use color to distinguish the three fits and provide an appropriate legend.

5. Compare the model fits using the Bayes Information Criterion (BIC). Which model provides a better fit? Discuss your findings.

6. For the best fitting model, investigate the quality of the fit. Make a plot of the residuals versus time, as well as a QQ-plot of the residuals. Is the model well fit?

7. Given the amount of uncertainty in the data, do you think it makes sense to say that one of the fits is "better" than the others? Why (not)? Provide some brief discussion based on the plots that you made before. For this question, you do not need to write any code.

3 Guidelines

Write the report as an **R Markdown** (`.rmd`) file. Structure your report well, with headers of different levels for the sections and research questions. There is a specific syntax in Markdown to write headers, bullet points, numbered items and tables. Please use this. You can find a complete guide on R Markdown here: `https://rmarkdown.rstudio.com/lesson-1.html`. A cheatsheet summarizes the most important syntax: `https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf`. Use LaTeXsyntax inside R Markdown for mathematical notations. During your work, regularly convert your .rmd file to HTML to avoid unexpected errors in the end.

Feel free to use R functions which haven't been used in the practicals, for example to make nicer graphs. Supplement graphs with informative main titles and axis titles with variable units.

Note on the use of Generative AI

You must make sure that **all project members understand and can explain everything that is written in the report**. You may be quizzed on the content of your report, and your final grade for the project may be affected (negatively) if it is clear that you did not write the project yourself or if you have no clear understanding of its contents.

4 Practical information

- **Deadline:** The project can be handed in until **27 May 2024 (6pm)**. The project is an obligatory part of the exam and will count for 15% of the final score. If no (or no decent) project is handed in, the maximal obtainable grade for the entire course is 7/20.

- **Groups:** Work is done in groups of **3 or 4 persons**. The project is scored as a group effort, everyone in a group receives the same mark.

- **Submission:** Each group should submit an RMarkdown file, and a corresponding HTML file. Both files have to be zipped and the zip-file submitted through Ufora. The names of the uploaded files must have the following structure: `Group_X_project2`, where `X` corresponds to your group number.

- Do not forget to mention the names of the group members in your project.

Good luck!

References

G. Laberge, P. Ambus, H. Hauggaard-Nielsen, and E. S. Jensen. Stabilization and plant uptake of N from 15N-labelled pea residue 16.5 years after incorporation in soil. *Soil Biology and Biochemistry*, 38(7):1998–2000, 2006.