# ST523/ST813 - Statistical Modelling

E2025

Home Assignment 2

This assignment should be **submitted electronically** as a single **pdf file** via **itslearning.**

**R** (or another statistics software) can be used to solve the exercises. The **output from the software should be explained in detail**, and all **replies to a question should be stated separately in the text**. Copied output alone without further explanation will not be accepted as valid answer.

Your final submission **must not exceed 13 pages** (additional pages will be disregarded). The page count is number of pdf pages (A4) using normal font size 11 or 12.
You can choose to include pieces of code, but it is recommended to NOT include unimportant code (e.g. for generating figures, data handling or formatting). Avoid long text passages, formulate shortly and precisely.

It is expected that you **work independently** on the assignment. Interactions will be considered as exam fraud.

## Generative AI:

- It is **NOT ALLOWED** to use generative AI **for carrying out any part of the statistical analysis** of the datasets that are studied during this assignment.

**Exercise 1**

Continuous glucose monitoring (CGM) sensors are wearable medical devices used by diabetic patients to track blood glucose levels. Each sensor has an individual lifetime, ending when adhesion degrades, electronics fails, or calibration drift becomes too large.

The aim of the exercise is to analyse how sensor lifetime is related to environmental, physiological, and technical factors, and to identify a parsimonious statistical model describing these relations. Additionally, it is of special interest to understand whether and how user experience influences lifetime.

Consider data from a single CGM sensor model collected from different users contained in the file `Data_Assignment2_Ex1_E2025.rdata`.

The dataset includes the following variables:

| | |
|---|---|
| `lifetime` | lifetime of the CGM sensor (days) |
| `skin_temp` | average local skin temperature (°C) |
| `humidity` | average ambient humidity (%) |
| `activity_level` | average daily activity intensity (0(low)–10(high)) |
| `adhesive_type` | adhesive class (A, B, C) |
| `sweat_rate` | average sweat rate (ml/hr) |
| `calibration_error` | electronics drift index |
| `patient_bmi` | body mass index |
| `experience` | years of CGM use |

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 lifetime and the other variables.

2. (Model Selection) Select a suitable generalized regression model that can be used to `lifetime` to the other available variables.

   Thereby, give explicit details about following aspects:

   a) State and explain which main classes of GLM are most appropriate for the given data,

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

   c) Consider several possible alternatives for a) and b), explain how you made an informed choice between the different alternatives.

   d) Apply suitable 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) Explain which principles you used in order to select a parsimonious model and indicate the sequence of actual steps taken by your approach including intermediate models and the rationale for transitions between them.

3. (Final model)

   a) Present your final model for CGM lifetime by 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 (on the scale of the response) from your final model.

      What is the expected lifetime of the sensor if the user has a BMI of 30, 1 year prior experience and an activity level of 4 (use the sample means for the other covariates)?

      And by how much does expected lifetime change if prior experience increases by 1 year while the other covariates remain unchanged?

   c) Make exemplary plots that can illustrate your model predictions: e.g. show the relation between `experience` and `lifetime` while keeping other variables fixed at their sample means.

   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?

   f) Which final conclusions do you draw from your statistical analysis about how sensor lifetime is related to environmental, physiological, and technical factors, and how important user experience is for sensor lifetime? Summarize your findings in one or two sentences.

**Exercise 2**

Consider the family of transformations given by

$$g_\alpha(Y) = \begin{cases} \frac{(Y+1)^\alpha - 1}{\alpha} & \text{for } y \geq 0, \alpha \neq 0 \\ \log(Y+1) & \text{for } y \geq 0, \alpha = 0 \\ -\frac{(-Y+1)^{2-\alpha}-1}{2-\alpha} & \text{for } y < 0, \alpha \neq 2 \\ -\log(-Y+1) & \text{for } y < 0, \alpha = 2 \end{cases}$$

for $Y \in \mathbb{R}$. Analogous to the estimation of the Box-Cox parameter $\lambda$, 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 $\alpha$ 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 $\alpha$? Provide your answers as formulas.

4. Substitute $\boldsymbol{\beta}$ and $\sigma^2$ 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_E2025.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 $\alpha$ using your function.

7. Determine the ML-estimate of $\alpha$ visually from this plot.

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