DAT320: Sequential and time series data analysis — Autumn 2023

Compulsory assignment 0 — **Submission deadline: 26.09.2023**

This compulsory assignment covers some of the prerequisites of the DAT320 course, as well as some basic R programming skills. The goal is to revisit statistical fundamentals required throughout the course.

Notation: vectors are written as non-capital letters in bold print, e.g. $\boldsymbol{y}$, while matrices are written as capital letters in bold print, e.g. $\boldsymbol{X}$. $\boldsymbol{X}^T$ indicates the transposed matrix of $\boldsymbol{X}$.

Exercise 1 (R syntax & data structures)

The goal of this exercise is to become familiar with the syntax and data structures in R. In particular, the R packages `dplyr` and `ggplot2` can be useful when working with R `data.frames`. The dataset used in this exercise contains a time series of weight gain in 50 chicks on different diets. The goal is to determine empirically, whether the diets have an influence on the weight of the chicks.

Material provided:

- `ChickWeight` dataset from package `datasets`

Tasks:

(a) *Load the dataset and convert the group variable* `Diet` *to type* `factor`*. The factor levels should be renamed to "A", "B", "C", and "D". Print a summary of the variables in the data frame.*

(b) *Give a summary of the different diets as follows: For each diet, compute the number of chicks the diet was applied to, as well as their minimum, mean, and maximum observation times.*

(c) *Plot the time series of each single chick (*`Time` *versus* `weight`*), colored by Diet, as a line plot. Further, plot the averaged weight across all chicks per time point (*`Time` *versus avg.* `weight`*) as a line plot for each Diet. Add error bars or ribbons indicating ± one standard deviation.*

(d) *Finally, interpret the results: Based on your analyses, what can we say about the influence of the different diets on the weight gain of chicks?*

Exercise 2 (Elementary data analysis)

In this exercise, we will perform an analysis of a dataset on fish toxicities. The dataset contains 908 samples and 7 variables, where one target variable (LC50, representing the toxicity level) should be modeled based on 6 molecular descriptors containing information about chemicals (samples). The goal is to characterize chemicals, which are associated with high toxicity levels.

Material provided:

- `fish.csv`: dataset

Tasks:

(a) *Compute and plot the empirical (univariate) distributions of the variables in the dataset. Provide the mean values, variances, and skewness, as well as minimum and maximum values of each variable. Plot a relative histogram along with a kernel density estimate. Which family of probability distributions could be used to model each of the given variables?*

(b) *Print the correlation matrix of the variables and plot it as a heatmap. Explain what you see (you may disregard absolute correlations below the level of 0.4)! Which consequences between input variables may correlations have in a predictive model?*

(c) *Perform an 75% - 25% train-test split and train a linear regression model with intercept to predict the variable LC50 (target variable) from all other variables (input variables). Which parameters are relevant for the model (i.e., have a parameter estimate, which is significantly unequal to 0)? Check the model assumptions and plot the residuals! Compute the following evaluation metrics on train and test set, respectively:*

- *RMSE (root mean squared error)*
- *MAE (mean absolute error)*
- *coefficient of determination (R^2 score)*

(d) *Train a new regression model with the same train-test split as in (c), but remove the variable `NdssC`. Compare the model summaries, evaluation metrics on the train and test set, as well as the log-likelihood, AIC and BIC values. Which model should be used?*

Exercise 3 (Covariance and correlation)

In this exercise, we investigate properties of correlations. Let X, Y and Z be continuous random variables. Let $\rho_{..} = \text{cor}(.,.)$ denote the Pearson correlation between each pair of random variables.

Tasks:

(a) *Let two transformed random variables W_1 and W_2 be defined as $W_1 = \alpha_1 \cdot X$ and $W_2 = \alpha_2 \cdot Y$ with real-valued scalars $\alpha_1, \alpha_2 \neq 0$. How does the correlation of the transformed variables $\rho_{W_1 W_2}$ relate to the original random variables ρ_{XY}? Explain your answer!*

(b) *The so-called "partial correlation" between random variables X and Y given Z is given as*

$$\rho_{XY|Z} = \frac{\rho_{XY} - \rho_{XZ}\rho_{ZY}}{\sqrt{1-\rho_{XZ}^2}\sqrt{1-\rho_{ZY}^2}}.$$

Do research on the concept and explain in which scenarios it is necessary to use it instead of the default correlation coefficient ρ_{XY}. Give an example.

(c) *Does the same property as in task (a) hold for the partial correlation concept, as well, i.e. for the partial correlation of $\rho_{W_1 W_2|Z}$? Justify!*

(d) *A sample from the random variables X,Y and Z delivers the following values:*
X: `1,1,1,1,0,0,0,0,1,1,1,1`
Y: `3,1,3,1,1,-1,1,-1,3,1,3,1`
Z: `5,5,9,1,1,1,5,-3,5,5,9,1`
Compute the pairwise correlations $\rho_{XY}, \rho_{XZ}, \rho_{YZ}$, as well the partial correlation $\rho_{X,Y|Z}$.

Exercise 4 (Likelihood of linear regression models)

The goal of this exercise is to become familiar with the concept of likelihoods and information criteria in linear regression models. A multiple linear regression model (without intercept) is given by the following formula:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \sum_{j=1}^{m} \beta_j \boldsymbol{x}^{(j)} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{y} \in \mathbb{R}^n$ is a vector of target values, $\boldsymbol{\beta} \in \mathbb{R}^m$ is the parameter vector, $\boldsymbol{X} \in \mathbb{R}^{n \times m}$ is the (input) data matrix with column vectors $\boldsymbol{x}^{(j)}$, $j = 1, \ldots, m$, and $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ is the model error. By default, we assume that $\boldsymbol{\varepsilon}$ is an independent and identically distributed vector, and ε_i follows a Gaussian distribution $\varepsilon_i \sim N(0, \sigma^2)$ with a constant standard deviation σ for all $i \in \{1, \ldots, n\}$.

Tasks:

(a) *Derive the log-likelihood of the model, denoted as $\ell(\boldsymbol{\beta}; \boldsymbol{x}) = \log L(\boldsymbol{\beta}; \boldsymbol{x})$. Give a step-by-step explanation (it should be possible to understand each step with ordinary knowledge in calculus and linear algebra). The result should be:*

$$\log L(\boldsymbol{\beta}; \boldsymbol{x}) = -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2\sigma^2}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right)^T\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right)$$

(b) *Explain, which of the steps / mathematical operations in (a) requires the following model assumptions:*

- *independence of errors ε_i,*
- *mean 0 of all ε_i,*
- *equal variances σ^2 of all ε_i (independent of sample index i).*

(c) *A common way to compute the parameters of the linear regression model is maximum-likelihood estimation. Derive the maximum-likelihood estimator $\hat{\boldsymbol{\beta}}$ by maximizing $\log L$. Compare with the Ordinary Least Squares (OLS) estimate. Hint: compute the first derivative of $\ell(.;.)$ with respect to $\boldsymbol{\beta}$ and set it to 0.*

(d) *The Bayesian Information Criterion (BIC), which is a common tool for model selection. Derive the expression for the BIC in the case of the given linear regression model. Hint: compute the derivative of $\ell(.;.)$ w.r.t. σ and set it to 0. Thereby, derive* *an expression for σ, and insert it in the formula for the log-likelihood derived in (a). For simplicity, you may substitute $(\boldsymbol{y} - \boldsymbol{X\beta})^T(\boldsymbol{y} - \boldsymbol{X\beta})$ by a variable "RSS", since this term is not dependent on σ.*