Introduction to Data Science

Assignment 2: Extensions of the Linear Model

Due: 23 December 2023

This assignment consists of two analyses. Each of them are to be written like two small research reports, each of approximately 3 pages. Any output from R/Python needs to be accompanied by an explanation about its relevance.

1. **Mixed effect ANOVA.** You are provided with a dataset that examines the performance scores of students from multiple schools under different teaching methods. The dataset is available on iCorsi (`test.csv`) and organized as follows:

   - Y_{ijk}: the performance score (`test`) of the i-th student taught using the j-th teaching method (`ttt`) in the k-th `school`.

   The data were obtained via a randomized study whereby 21 schools were randomly assigned to three teaching methods.

   Perform an analysis of variance (ANOVA) with random effects to investigate whether there are significant differences in student performance due to teaching methods and schools. Assume a two-factor model with one factor as a fixed effect (teaching method) and the other as a random effect (school).

   (a) $\boxed{5}$ **Research problem.** Formulate a research question (HINT: involve causality).

   (b) $\boxed{10}$ **Exploratory analysis.** Perform an informal analysis of the data (i.e. tables and/or plots of the data) relevant for answering the research question. Accompany every figure/table with an explanation of what can be seen and why it is relevant.

   (c) $\boxed{20}$ **Formal analysis.** Perform a formal analysis of the data.

      i. Formulate the appropriate statistical model for this two-factor ANOVA with random effects. Clearly define the fixed and random effects, and express the model equation.

      ii. State the null and alternative hypotheses for testing the significance of the fixed effect (teaching method) and the random effect.

      iii. Provide a brief summary of the dataset, including descriptive statistics and any relevant visualizations that can help in understanding the data.

      iv. Perform the ANOVA analysis using the dataset. Report the relevant F-statistics, p-values, and any other necessary output.

   (d) $\boxed{5}$ **Conclusions.** Answer the initial research problem.

   (e) $\boxed{5}$ **Discussion.** Provide a brief discussion about some critical aspects in the analysis.

2. **Logistic additive regression.** A study on 768 adult female Pima Indians living near Phoenix was undertaken to investigate the causes of diabetes and the potential for its diagnosis via easily available variables. Data was obtained via an observational study and is available on iCorsi (`diabetes.csv`).

   The data include various variables, which may be related to a subject's risk of diabetes, such as (i) the number of times the women were `pregnant`, (ii) the plasma `glucose` concentration at 2 hours in an oral glucose tolerance test, (iii) the `diastolic` blood pressue (mm Hg), (iv) the `triceps` skin fold thickness (mm), (v) the 2-hour serum `insulin` level (mu U/ml), (vi) the body mass index (`bmi`), (vii) `age` in years, and (viii) the `diabetes` Pedigree Function (DPF) is a genetic score developed by Smith et al. (1988) to provide a synthesis of the diabetes mellitus history in relatives and the genetic relationship of those relatives to the subject. It provides a measure of the expected genetic influence of affected and unaffected relatives on the subject's eventual diabetes risk.

   The variable `test` measure whether the patient shows signs of diabetes (1) or not (0).

   (a) $\boxed{5}$ **Research problem.** Formulate a research question (HINT: it is necessary to focus on two aspects).

   (b) $\boxed{10}$ **Exploratory analysis.** Perform an informal analysis of the data (i.e. tables and/or plots of the data) relevant for answering the research question. Accompany every figure/table with an explanation of what can be seen and why it is relevant.

   (c) $\boxed{20}$ **Formal analysis.** Perform a formal analysis of the data based upon several predictors.

      i. Consider using linear and non-linear functions of the covariates for modelling the response. In case of the latter, use an additive model. (HINT: use the `gam` function from the `mgcv` package with the `s(covariate)` option and the appropriate **family** description).

      ii. For the linear predictors interpret the regression coefficients. For the non-linear predictors interpret the plot of the predictor function.

      iii. Use model selection to find a parsimonious model.

      iv. Do the model assumptions hold?

   (d) $\boxed{5}$ **Conclusions.** Answer the initial research problems.

   (e) $\boxed{5}$ **Discussion.** Provide a brief discussion about some critical aspects in the analysis.