Applied Statistics

Instructions:

- The assignment must be carried out in a group consisting of two students or, exceptionally, individually if pairing is not possible.
- The resolution of the assignment consists of answering the questions in the statement and an R file with the obtained results.
- The choice of a group and the dataset must be made by opening a topic in the Moodle discipline forum with the title "Group ...", where information can be exchanged, and doubts clarified regarding the assignment. The group selection can be indicated even without choosing the dataset. A dataset will only be officially assigned after confirmation by me that it is admissible. As this choice is made, a list of datasets already used will be updated three times a week on Moodle.
- The written assignment must be a maximum of 10 pages and delivered in PDF format at the time of the presentation, as well as an R file, both submitted on Moodle. Presentation dates will be scheduled through a Doodle made available for that purpose. Each presentation must not exceed 30 minutes.

1. Find a dataset that fits into a linear regression context and satisfies the following conditions:

   - Inludes at least 3 explanatory variables;

   - Contains at least one continuous explanatory variable;

   - Contains at least one categorical explanatory variable with more than 2 categories. If no categorical variables are present, categorize one of the continuous variables.

   You can use your own dataset or one already available in R. You should use a dataset contained in one of the R packages made available in 2024:

   - `http://cran.r-project.org/web/packages/available_packages_by_date.html`

   When using an R dataset, indicate the name and the respective package.

   Clearly and precisely formulate the problem corresponding to the chosen data and fit a linear regression model that seems most appropriate. Build a final report that includes:

   (a) a statistical, numerical, and graphical description of the data;

   (b) a well-founded discussion on the final model selection;

(c) for a continuous variable X_1 and a categorical variable with more than two categories X_2 included in the final model (or the initial model in case the final model does not have this type of variable):

  i. interpret the corresponding raw and adjusted effects;

  ii. interpret the effect caused on the response by a change from the third category of X_2 to the second, as well as 95% and 90% confidence intervals for this effect;

  iii. investigate the existence of a significant interaction between X_1 and X_2.

2. Consider a transformation of the previously used response variable, separating values greater than or equal to its median from the lower values. Repeat the previous exercise assuming a logistic regression model with the transformed response variable.