ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

M.Sc. in Data Science

Course: Probability and Statistics for Data Analysis

Semester: Fall 2023

Instructor: Ioannis Vrontos (`vrontos@aueb.gr`)

Grader: Konstantinos Bourazas (`kbourazas@aueb.gr`)

Assignment 2

Deadline: 3 January 2024

Note: Use R in this assignment and submit your .R code that was used to answer the questions, along with a small report where you will present plots and results for each question of this assignment.

1. In file "data.txt" (available on the e-class assignments site), you will find the recorded variables Y, X1, X2, X3, X4 (continuous), and W (categorical with three levels) on 150 cases. Using these data, answer the following questions:

(a) Run the parametric one-way ANOVA of each of the continuous variables (Y, X1, X2, X3, X4) on the categorical variable (W). Specifically,

(i) provides a graphical representation of each of the continuous versus the categorical variable

(ii) provide the ANOVA output

(iii) check the assumptions.

(b) Provide a scatter-plot matrix of Y, X1, X2, X3, and X4, annotating the different levels of W in each plot using a different color.

(c) Run the regression model of Y on X4

(d) Run the regression model of Y on all the remaining variables (X1, X2, X3, X4, W), including the non-additive terms (i.e., interactions of the continuous predictors with the categorical).

(e) Examine the regression assumptions and provide alternatives if any of them fails.

(f) Use the "stepwise regression" approach to examine whether you can reduce the dimension of the model.

(g) Using the model found in **(f)**, provide a point estimate and a 95% confidence interval for the prediction of Y when: (X1, X2, X3, X4, W) = (120, 30, 10, 90, B)

(h) Using the cut() function, create a categorical variable (named Z) with three levels based on the quantiles of X4. Provide the contingency table of X4 and W.

(i) Run the parametric two-way ANOVA of Y on the categorical variables W and Z (including the interaction term). Provide the fit, examine the assumptions, and comment on the significance of the terms.

2. In the file "weightloss.txt"(available on the e-class assignments site) you will find the recorded variables work (categorical with three levels), diet (categorical with four levels), and loss (continue, in calories). More specifically, the data provide the weight loss per day in a 3×4 factorial experiment. The two factors include 3 types of workout and 4 types of diet. Each combination of the two factors is used to be completely randomized.

(a) Provide boxplots of the weight loss per workout, per diet, and for the combinations of the two categorical factors.

(b) Fit a One-Way ANOVA model with the weight loss as a response and the workout (as a factor). Interpret the model parameters.

(c) In the ANOVA model of **(b)**, is the expected difference between $W2$ and $W3$ significant? [TIP: change the reference level appropriately and refit the ANOVA model of question **(b)**]

(d) Fit a One-Way ANOVA model for the weight loss as response and diet. Interpret the model parameters. Are all treatments significant?

(e) Exclude the non-significant treatments of the model **(d)** and re-estimate the parameters.

(f) Fit a Two-Way ANOVA model of main effects. Provide the interpretation for the parameters.

(g) Exclude the non-significant levels of the factors and refit the model. Provide the interpretation for the parameters of the new, simplified, model.

(h) Fit a Two-Way ANOVA model with interactions. Are all the parameters significant?

(i) Using the stepwise method, choose a model based on the AIC criterion starting from the full model (including the main effects and the interaction

term). Are all coefficients significant?

(j) Provide a graphical representation for the final model.

(k) Compare the constant model against the (full) main effects model and the (full) interaction model. Are the models different?