ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ

ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS

M.Sc. in Data Science

Course: Probability and Statistics for Data Analysis

Semester: Fall 2024

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

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

Assignment 2

Deadline: 5 January 2025

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. A study of the effect of two drugs on the reduction of cholesterol used 100 volunteers who tested the drugs. Fifty of them were randomly selected to take the first drug (A), while the remaining fifty took the second one (B). We measured the cholesterol levels after receiving the drugs, tested the presence of Myalgia symptoms, and measured the Glucose levels in order to check the side effects of the drugs. The observations are in the file “cholesterol.txt” (available on the e-class assignments site).

(a) Provide a 99% confidence interval for Cholesterol values.

(b) Provide a 95% confidence interval for Cholesterol values after receiving drug A and B, respectively.

(c) Provide a 90% confidence interval for the mean difference in Cholesterol values after receiving drug A and drug B, respectively.

(d) Examine the following hypothesis test:

$$H_0 : \mu_A = \mu_B$$
$$H_1 : \mu_A < \mu_B$$

where μ_A and μ_B are the mean Cholesterol values after receiving drug A and B, respectively. The level of significance is $\alpha = 0.05$.

(e) Provide a hypothesis test ($\alpha = 0.01$) for the equality of variances of Glucose levels after receiving drug A and drug B, respectively.

(f) At a significance level of 5%, test if there is a statistically significant side effect on Glucose levels.

(g) Provide a 95% confidence interval for the proportion of volunteers who had Myalgia symptoms.

(h) Test if the proportion of volunteers who had Myalgia symptoms is statistically greater than 5% at a significance level of 5%.

(i) Test if the drug and the presence of Myalgia symptoms are independent ($\alpha = 0.05$).

(j) Provide a 95% confidence interval for the mean difference $\mu_1 - \mu_2$, where μ_1 and μ_2 are the mean Glucose levels for volunteers with and without Myalgia symptoms, respectively.

2. In the file “data2.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) for 150 cases. Using these data, answer the following questions:

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

- **(i)** Provide a graphical representation of each continuous variable 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 variable).

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

(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 3 levels based on the quantiles of X4. Provide the contingency table of Z 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.