SML 2024, big programming assignment 1

Dorota Celińska-Kopczyńska

The goal of this assignment is to provide statistical analysis of the data from file `project1_data.csv` available in the moodle.

Dataset: you will work with simulated data. The sample comes from a survey on the consumer habits of residents of the fictional Byteland. The survey was conducted on a representative sample[1] The data may contain random errors. The variables available for the needs of our project are:

- *id* – id of an observation, it does not contain any additional information
- *weight* – weight of the respondent (in kg)
- *height* – height of the respondent (in cm)
- *sex* – sex as stated in the respondent's national id (1 – "female", 2 – "male")
- *cats* – number of cats in the household (in entities)
- *age* – age of the respondent (in years)
- *income* – declared income of the respondent in the month of the survey (in bythalers)
- *savings* – declared savings of the respondent in the month of the survey (in bythalers, negative values indicate that the expenses were greater than the income)
- *single* – household status (1 – "single person household", 0 – "multi-person household")
- *place* – size of the town where the respondent lives (1 – "up to 10,000 inhabitants", 2 – "from 10,000 inhabitants to 100,000 inhabitants", 3 – "over 100,000 inhabitants")
- *expenses* – declared expenses on food by the respondent in the month of the survey (in bythalers)

Desired output: You are supposed to submit a jupyter notebook with the solutions, commentary, and results by Moodle. Please make sure your notebook opens and works in Google Colab it will not be graded otherwise. Solutions will be graded by lab assistants of respective groups. Exercise 6 will be graded by the lecturer.

The report and comments must be sufficient to understand and reproduce the steps you have taken without having to read your codes. Each action taken that significantly modifies the database (e.g., deleting records, modifying and introducing new variables) must be justified and described. In each task, you can use ready-made implementations.

Deadline: 22.12.2024, 11:59 PM CET

Total points to obtain: 20

[1] A representative sample is a sample whose structure, in terms of the studied features (variables), is similar to the structure of the statistical population from which it comes.

1. Download and load the data, describe and summarize it in a few sentences. Leading questions:

   - how many observations are there in the sample? Discuss the structure of the dataset: how many quantitative and how many qualitative variables do we have? Are there missing data? (0.5point)
   - Provide and describe appropriate frequency tables or descriptive statistics for the variables (take into account the type of the variables!) (1 point).
   - Present and discuss (where appropriate) variables' distributions, especially compare them with the normal distribution (e.g. with histograms, density functions, qqplots, etc.). (1point)

2. Analyze if there are associations among the variables: visualize (with heatmaps) and compute **proper** correlation coefficients (justify your choices); find out whether the possible dependencies are significant. Discuss the results. (1.5point)

3. Summarize the data with at least three different types of plots (do not forget to provide a commentary!). The minimum set of plot types includes:

   - Scatter plots for all quantitative variables against the expenses on food
   - A boxplot for a quantitative variable of choice in division by the size of the town the respodent lives in
   - A stacked bar chart for sex against whether the respondent is from single-person household

   However, we encourage you to provide additional types of the plots! (1.5 points in total for the minimum set, 0.5point for each plot: 0.125point for the graph and 0.375 for the commentary; possible 1 extra point in the discretion of the lab assistant for the outstanding additional visualizations.)

4. Byteland's sociologists divide Byteland's society into four wealth classes:

   - lower class (income below the 25th percentile of the income distribution)
   - middle class (income equal to or higher than the 25th percentile of the income distribution and lower than the 75th percentile of the income distribution)
   - upper middle class (income equal to or higher than the 75th percentile of the income distribution and lower than the 90th percentile of the income distribution)
   - upper class (income equal to or higher than the 90th percentile of the income distribution)

   Discuss and compare the differences in expenses on food in the above-mentioned wealth classes (1.5 points: 0.25 point for performing the division, 0.75 point for calculating the correct measure of variability, 0.5 point for commenting and discussing the results).

5. Choose and conduct the most appropriate statistical tests to answer the following research questions:

   (a) Do women declare higher savings than men?

   (b) Does lower proportion of food expenses to income correlate with higher savings?

   (c) Is the mean weight of women greater than 56 kg?

Additionally:

(d) verify an additional (sensible) hypothesis on the goodness-of-fit with a given (sensible) parametric distribution for the selected variable (e.g., "variable A has a Poisson distribution with parameter 1").

Assume significance level of $\alpha = 0.01$. For each statistical test: provide the assumptions (and justify them), state null and alternative hypotheses, justify the choice of the statistical test, present the results and decide if you reject the null hypothesis. It is fine to use built-in statistical tests instead of coding them yourself. (4 points in total, 1 point for each hypothesis).

6. Conduct a study on food expenses using variables from the database. Assume a significance level of $\alpha = 0.01$. To do this:

   - Estimate a preliminary model containing all variables from the original database (except id) and a constant, where the variable *expenses* is the dependent variable. Remember to decode qualitative variables. Discuss the R^2, individual and joint significance of the independent variables in the preliminary model. (1 point)
   - Check whether the preliminary model meets the assumptions of the Linear Regression Model. Pay special attention to the issues of linearity of the functional form, homoscedasticity, and lack of autocorrelation of the random disturbance, and the distribution of the random disturbance. (2 points)
   - Improve the model so that it satisfies as many assumptions of the linear regression model as possible. Describe the steps you took to improve the model and present your "best" model (4points)

     Hint: We work with simulated data, so an extremely high R^2 is possible.
   - Provide a **proper** quantitative interpretation of the selected two individually significant coefficients in the "best" model. Remember that the constant is not interpretable. (1 point)
   - What are the descriptive characteristics of people whose food expenses belong to 10% top predictions of food expenses in your "best" model? Inspect and discuss (1 point).