SML 2023, big programming assignment 1

Dorota Celińska-Kopczyńska

The goal of this assignment is to provide statistical analysis of the data from file `earnings.csv` available here.

Dataset: You will work with a sample of the data provided by Central Statistical Office of Poland. The sample is obtained from the Structure of Wages and Salaries by Occupations (SWS) database from October 2010 (a part of Z12 programe, for more information see [1], [2], [3]). SWS database covers non-financial entities of the national economy that employ more than nine employees. Data on earnings from SWS are highly reliable because they are reported by corporate accounting departments. The original database contains data on wages and their components as well as selected characteristics of the companies and employees. The available variables for the needs of our project are:

- *id* – observation id;
- *base* – total of base salaries;
- *bonus* – statutory bonuses, awards and discretionary bonuses;
- *overtime_pay* – overtime pay;
- *other* – remuneration in the form of employee remuneration, additional annual remuneration for employees of public sector entities, payments for participation in profit or in the balance sheet surplus in cooperatives;
- *sector* – economic sector (1 – public, 2 – private);
- *section_07* – NACE section (1 – Public Administration and Defence; Compulsory Social Security, 2 – Education, 3 – Human Health and Social Work Activities);
- *sex* – the sex of the employee (1 – man, 2 – woman);
- *education* – highest educational level obtained by the employee (1 – doctorate, 2 – higher, 3 – post-secondary, 4 – secondary, 5 – basic vocational, 6 – middle school and below);
- *contract* – type of employment contract (1 – for an indefinite period, 2 – for a definite period);
- *age* – age of the employee as in 2010;
- *duration_total* – total duration of employment;
- *duration_entity* – duration of employment in the reporting entity;
- *duration_nominal* – the time actually worked in nominal hours;
- *duration_overtime* – time actually worked overtime.

Desired output: You are supposed to submit 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. They will be graded by lab assistants of respective groups.

Deadline: 3.01.2024, 11:59PM CET

Total points to obtain: 25

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 any missing data? (0.5point)
   - Provide and describe appropriate frequency tables or descriptive statistics for the variables (take into account the type of the variables!) (0.5point).
   - Present and discuss (where appropriate) variables' distributions, especially compare them with the normal distribution (e.g. with histograms, density functions, qqplots...). (2points)
2. Analyze if there are associations among the variables: visualize and compute proper correlation coefficients (justify your choices); find out whether the possible dependencies are significant. Discuss the results. (3points)
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 the variables related to the salary structure against the duration of employment in the reporting entity
   - A boxplot for a quantitative variable of choice in division by the type of employment contract of the respondents
   - A heatmap of the (appropriate!) correlation coefficients among the quantitative variables in the dataset

   However, we encourage you to provide additional types of the plots! (3 points in total for the minimum set, 1point for each plot: 0.25point for the graph and 0.75 for the commentary; possible 1 extra point in the discretion of the lab assistant for the outstanding additional visualizations.)
4. Compute the confidence intervals at the confidence level $1 - \alpha = 0.99$ for the *age* of the employees for the following parameters:
   - mean and the variance
   - median

   Write down the assumptions you made for the needs of your analysis and discuss if they are justifiable. (1point in total: 0.25point for mean, 0.25point for variance, 0.5 point for median)

   Hint: For CI for variance, check if the variance estimator you use is unbiased (division by $n-1$) or biased (division by n). In the lecture, we provided the formula for the biased estimator, if you use the unbiased one, instead of multiplying by n you need to multiply by $n-1$ (quantile of the χ^2 distribution stays the same). Using bootstrap for computing CI for median is not mandatory (we have introduced a different tool that you may creatively use here!).
5. Choose and conduct the appropriate statistical tests to verify the following hypotheses:

   (a) There are significant differences between the base salary of the employees of the public and private sector

   (b) Among employees that are younger than 30 years old, the mean total duration of the employment is equal to the mean total duration of the employment in the reporting entity.

   (c) A shorter total duration of the employment is correlated with the longer time actually worked overtime.

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. (3 points in total, 1point for each hypothesis).

6. Verify additional 3 hypotheses for the provided problem descriptions. Pick some variables that would fit problem description and state the H0 and H1 yourself.

   - Comparison whether two qualitative variables of your choice are independent in division by different types of employment sectors. One of those variables has to be coded by yourself using a quantitative variable of your choice (justify your choices and describe the process).
   - Comparison of the distributions of the components of the salaries from administration section against all other sections.
   - Hypothesis on the goodness-of-fit with a given parametric distribution (e.g., "variable A comes from the exponential distribution with the parameter $\lambda = 10$").

   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. (3 points in total, 1point for each hypothesis).

7. In the dataset, we included employees from three NACE sections dominated by the public sector: administration, education, and healthcare. Education and Healthcare are known for their employees' strikes related to working conditions and wage demands. Using linear regression model, analyze the factors that may influence the total salary of the employee.

   - Compute the total salary of the employee (by summing the components of the salary) (0.5point)
   - Estimate a preliminary model including all the variables from the original dataset (apart from id) where the total salary is the dependent variable. Discuss the R^2, individual and joint significance of the independent variables. Check if the assumptions of the linear regression model are satisfied for the preliminary model. (2.5points)
   - Improve the model so that it satisfies as many assumptions of the linear regression model as possible. Do not forget to check for the outliers and transform data if needed! Describe the steps you took to improve the model and present your "best" model (3points)
   - Provide the insights based on the "best" model – provide interpretation of the individually significant parameters (2points)
   - What are the descriptive characteristics of the employees whose salaries belong to 10% bottom predictions of the total salary in your "best" model? Inspect and discuss. (1point)

References:

[1]: `https://stat.gov.pl/download/cps/rde/xbcr/gus/pw_struktura_wynagr_wg_zawodow_10_2010.pdf` (only in Polish)

[2]: `https://stat.gov.pl/cps/rde/xbcr/gus/pw_strukt_wynagrodzen_wg_zawodow_X_2010.pdf` (only in Polish)

[3]: `https://stat.gov.pl/en/topics/labour-market/working-employed-wages-and-salaries-cost-structure-of-wages-and-salaries-by-occupations-in-october-2010,4,2.html`