README

Neli Noykova

June 29, 2018

Data Analysis Project, Summer 2018, University of Helsinki

Brief description

Here I have described the results on my work on Data Analysis Project in June 2018 at Helsinki University.

The data used in this work are taken from [**Statistics Finland**] (https://tilastokeskus.fi/tup/mikroaineistot/aineistot_en.html). These are individual level combined employer-employee data, or so-called FLEED (Finnish Longitudinal Employer-Employee Data). The original data contain information on population of working age, which can be combined with enterprise and establishment level data.

The goal is to find out how the employee's family status and the size of the employing company (measured by its turnover) are related to the size of the earned income, or in other words - how one numerical and two categorical variables are related.
The total earned income is assumed as dependent variable, while the two categorical variables - family status and size of the company - are assumed as dependent variables. For this work only a small data subset is used for year=2 (the data are observed for 15 years).

During the work some basic statistical techniques in univariate, bivariate and multivariate analyses are demonstrated and applied. Big attention is paid to data visualization, especially during the initial step in exploratory analysis when some initial inferences and dependences could be revealed. The one way ANOVA is applied in bivariate analysis, while two ways ANOVA is used in multivariate analyses.

In this work first the dependences between the original variables are investigated. The second part include similar investigations, but with grouped categories of both independent variables. For this particular task it appeared that grouping categories helped a lot to reveal the some tendencies in the investigated relationships. All provided analyses showed that both assumptions about the homogeneity of the variance and normal distribution are probably violated.
For deeper understanding it will be necessary to analyze other subsets for different years, and compare the results. Also some other inferences and other possibilities for future work are discussed.

About this document

This is R Markdown document, created in R Studio (package Rmarkdown should be installed).

Link to my report

Here is the link to my course report:

https://noykova.github.io/Data-Analysis-Project/