
Course: Basics of R programming language for statistical analysis

Instructor: Marina FERENT [marinaferent@gmail.com]

Meeting 7

You are working for Multicultural Business Institute. They provide coaching for people in search for a job. One of their promoted theories is that the higher your beginning salary in a company, the higher your future salary will be in that company. As such, they encourage people to stress for higher beginning salaries when in a first interview.

Your manager gathered data on the beginning and current salaries of the previous participants in their coaching programs – see **Wages.csv**[i]. She asked you to perform **a cross-tab analysis on the salary and salbegin** variables so to have a first image on their theory (your manager doesn't yet know that correlation is not causality 😊).

You have 45 minutes to provide your insights.

Data set: Wages.csv

Variables:

- salary (quantitative continuous variable)
- salbegin (quantitative continuous variable)

Cross-tab analysis (statistics cheat sheet):

1. Numerical representation: Cross-tab (bi-dimensional distribution) in absolute or relative frequencies -> omit this time
2. *Graphical representation: scatter plot* (<=the most suitable graph for 2 quantitative continuous variables)
3. *Correlation: Pearson's correlation coefficient* (<=the most suitable for 2 quantitative continuous variables)

Additionally, you know your manager is not keen in statistics. As such, many times she underestimates the time needed to provide a statistical analysis. She is full of theories and ideas, though. Most probably she will ask you again to provide a cross-tab analysis for 2 other quantitative continuous variables. In the time you have at your disposal try to **"automate" one interpretation/reporting aspect of your cross-tab analysis** so to have it at hand for future use. In doing so, use any R functions, for loops, if statements, write your own function etc.

Some ideas on what to "automate" – provide one of:

- Add the Pearson's correlation coefficient value on your scatter plot.
- Export your graph into a .pdf, .png etc. file.

- Based on a conditional statement of your choice print:
  - “Positive correlation” if Pearson’s correlation coefficient > 0
  - “Negative correlation” if Pearson’s correlation coefficient < 0
  - “No correlation” if Pearson’s correlation coefficient = 0
- Based on a conditional statement of your choice print:
  - “High correlation” if abs(Pearson’s correlation coefficient) >=0.7
  - “Medium correlation” if abs(Pearson’s correlation coefficient) <0.7 and >=0.3
  - “Low correlation” if abs(Pearson’s correlation coefficient) <0.3
- Based on a conditional statement of your choice print: “low positive correlation”, “low negative correlation”, “medium positive correlation” etc.
- Based on a conditional statement of your choice and paste0 function print: - e.g. “The correlation is 0.24 => low positive correlation between salary and salbegin.”, “The correlation is -0.24 => low negative correlation between salary and salbegin”, “The correlation is 0.34 => medium positive correlation between salary and salbegin” etc.
- Save into a matrix and export into a .csv file the value of Pearson’s correlation coeffient and the interpretation – e.g.:

Correlation	Interpretation
0.78	There is high positive correlation between salary and salbegin.

- Add the Pearson’s correlation coefficient value and the interpretation on your scatter plot.
- Create an interpretation function.
- Compute the correlation matrix for all the variables in your data set.
- Compute the correlation matrix for all the quantitative continuous variables in your data set.
- Using a for loop, plot the scatter plots of salary against all the other quantitative variables in your data set.

[i] The data set is a slightly altered version of *engin* data from *Wooldridge, Jeffrey M. (2013). Introductory econometrics: a modern approach. Mason, Ohio: South-Western Cengage Learning*. *Wooldridge Source: Thada Chaisawangwong, a former graduate student at MSU, obtained these data for a term project in applied econometrics. They come from the Material Requirement Planning Survey carried out in Thailand during 1998*.
The original data set is available for download at:
(1)https://www.cengage.com/cgi-wadsworth/course_products_wp.pl?fid=M20b&product_isbn_issn=9781111531041 Or
(2) https://cran.r-project.org/web/packages/wooldridge/wooldridge.pdf
Current data set changed the definition of the gender variable and created the variable job category for instructional purposes (Manager=those employees that have higher than Q3 total experience and salary; Admin = those employees that have higher than Q2 salaries or total experience; Worker=the rest). I also dropped some of the variables of the original data set.