CMPSC 390
Data Analytics
Fall 2017

Lab 8: The Statistical Analysis of Economic Data with an Emphasis on Gender Roles in International Business Development
Save this lab assignment to: `labs/lab8`

Objectives

To explore statistical tools which are relevant for the evaluation of economic data. At this point in the class, there has already been much exposure to using many different types of statistical tools to handle different formats of data. It is therefore expected that the student will able to research code development and will be able to resolve data formatting issues during while working in the analysis of data. In particular, these skills comprise the ability to research R-statistics software packages for the application to the particular contexts for which they were designed and to extract knowledge from the produced visualizations and extracted interpretation of results. Furthermore, the student will be able to explain credible results and conclusions which are to be entirely supported by the methods and data.

Reading Assignment

In the article by Heathcote *et al.* [1], female labor supply is discussed for its impact on the United States economy between the years of 1967 to 2002. The authors produce a model which was designed to help study and measure the below factors which are associated to women in the workforce.

The Article's Themed Questions

1. The decline in marriage rates,
2. The narrowing gender wage gap,
3. The preference (or cultural) shift towards market work, and
4. The change in womens bargaining power within the household.

Our data originates from [2] and concerns the lifestyles, living conditions and economic contributions made by women on the world's stage between the years of 1960 to 2017. In this lab, you are to use this data to gather direct or indirect proof to either support or refute the claims made by the four themed questions (listed above) of the study by Heathcote *et al.*

For each of its four themes, determine the necessary variables from the data set in the spreadsheet which could be used to support or refute the claims of the article. Remember, the exact questions from the article may not be directly answerable by simply running tests over the included data, and so you will have to develop a strategy in which you will use the correlations of several variables, or multi-linear models to argue your point. For this step, your selected variables

will amass proof that could be used to describe scenarios or conditions which will support or refute the conclusions of the authors.

Create a Strategy and an Argument

You will have to be creative as you approach each theme using variables from your data set. For instance, the first theme of the article's study is to investigate whether there is a *decline of marriage rates* in populations of women. This implies that fewer women are either married or are getting married for some unknown reason (an unknown mechanism). In your analysis, you will have to choose variables which could indicate a mechanism leading to a possible decline in marriage. Finding such evidence in your data will enable you to agree with the conclusions of Heathcote *et al.*, even though your data set was completely different. If you are able to refute the article's claims, then this is also an excellent result.

Perhaps a good place to start this analysis in the study of *marriage decline* would be to choose all necessary variables that may contribute to the argument of a woman's general loss of interest in marriage. In this case, evidence could include:

- The declining numbers of married women between 1960's to present,
- The rising numbers of financially autonomous women in society, the populations of women who own their own houses and lands,
- The number of woman who have invested in their own businesses.
- The number of women who work professionally
- The counts of women who deny their husbands the authority to beat them.

At Least Two Countries

All of these mentioned variables, when used together in a multi-linear regression model (where you choose a relevant dependent variable) could be used to argue for the support or to refute a conclusion made by the article. Remember to chose run the data for at least two countries to confirm or refute the conclusions. It is very likely that the conclusions of the article are not generally true all over the world which implies that there is a country for which the conclusions is completely incorrect. Finding a country where a conclusion is correct and, another country where the conclusion is incorrect would be an ideal result for your work.

Groupwork

You are to work in a group of not more than three people for this lab. Be sure to discuss each of the tasks and proceed after the group has come to a complete agreement. **Each person is to turn in his or her own report and code, however all lab partners should be listed in the submission.**

Important Details

Lab directory structure: Make sure you have placed your submission materials for this lab into `labs/lab7` directory in your Bitbucket repository (`cs390f2017-billb`).

Note: Please remember to include your name on everything you submit for the class. Although the instructor collects your work from Bitbucket, each work must be graded outside of the Bitbucket directory and so without adding your name, the instructor will be unable to award you credit for your work.

Required Deliverables

By the 1st December at lab time (2:30pm), submit electronic versions of the following deliverables through your Bitbucket repository (`cs390f2017-billb`) by correctly using using appropriate Git commands, such as `git add -A`, `git commit -m ''your message''` and `git push`. When you have finished, please ensure that the Bitbucket Web site has your pushed work. Please contact the instructor if you have any questions about assignment submission.

1. The R source code that you used to answer your questions.

2. Report that includes the questions, the data descriptions, exploratory questions, correlation analysis, at least two t.tests and at least two linear regression tests to corroborate or debunk the results gained in Heathcote *et al.* Write your report so that it is clear which parts of your code and discussion addresses what particular question. In addition, this report is to contain the visualizations and your brief written evaluation after each to describe the knowledge that you have gained from each visualizations and /or test, in light of the p-values.

   Your report document should be in an open office format and your visualizations should be included directly into your report document.

3. You are also to include a reflection portion to your report document where you describe how data analysis research is different (likely different) between the disciplines of psychology and economics. For instance, you can describe how the software tools and packages may differ between both disciplines. Follow your notes from the talk given by Dr. Steven Onyeiwu and include insights from his talk in your reflection document.

4. **Please do not assume that the instructor will know which files are relevant to your submission. Please place all plots and visualizations which are necessary for this work, into a clearly (and obviously) labeled Libre Office document file. No MS Word files, please.**

References

[1] Jonathan Heathcote, Kjetil Storesletten, and Giovanni L Violante. The macroeconomics of the quiet revolution: Understanding the implications of the rise in womens participation for economic growth and inequality. *Research in Economics*, 2017. `https://www.semanticscholar.org/paper/`

The-Macroeconomics-of-the-Quiet-Revolution-Underst-Heathcote-Storesletten/96752a31855fa0e93b66959a56e9f765f2ff5425.

[2] The World Bank. Gender statistics. https://data-worldbank-org.ezproxy1.allegheny.edu/data-catalog/gender-statistics, October 2017. National and Regional Data, data@worldbank.org.