CMPSC 301
Data Analytics
Fall 2018

Lab 7: The Statistical Analysis of Economic Data
With an Emphasis on Gender Roles in
International Business Development
16th November 2018

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.

Groupwork

You are to work in a group of not more than four (4) 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.**

GitHub Starter Link: Group work

`https://classroom.github.com/g/6EKjL9cj`

To use this link, please follow the steps below.

- Your group leader will click on the link and accept the assignment and prepare a team name. All other members will later click on the link and select their team's name from the list that will appear.

- Once the importing task has completed, click on the created assignment link which will take you to your newly created GitHub repository for this lab.
- Clone this repository (bearing your name) and work on the lab locally.
- As you are working on your lab, you are to commit and push regularly. You can use the following commands to add a single file, you must be in the directory where the file is located (or add the path to the file in the command):

  ```
  – git commit <nameOfFile> -m ‘‘Your notes about commit here’’
  – git push
  ```

  Alternatively, you can use the following commands to add multiple files from your repository:

  ```
  – git add -A
  – git commit -m ‘‘Your notes about commit here’’
  – git push
  ```

The Article's Themed Questions

It is often likely that the original data from an article is unavailable for public use, or that only a subset of the data has been made available. In this event, we turn to other available data sets in order to study the same types of research questions and themes of the article. Furthermore, using our own data, we are often able to test (and validate) some of the same theories as those of articles which decide not to release their original data sets to the public.

Our data set for this lab originates from the World Bank's Online Data Base [2] and may be conveniently obtained from: `https://datacatalog.worldbank.org/dataset/gender-statistics`. This data 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 (using proxies) to either support or refute the claims made by the four themed questions (listed above in Reading) of the study by Heathcote *et al.* For this work, you are to become acquainted with this article and to understand the authors' position on their conclusions.

For each of the four themes of Heathcote *et al*, shown below, determine the necessary variables from the data set (available from the above link to World Bank's Online Data Base) which could be used to address the claims of the article for women (**of any two countries of your choosing**). Remember, the exact research questions from the article may not be directly answerable by simply running tests over the World Bank data set You will have to develop a strategy in which you will combine several different variables which correlate in some logical (and reasonable) way to be used with linear models (single or multiple). Armed with this analysis from your data, you may proceed to argue your points of view, with regard to those written in blue (below) from the article.

Questions in Blue

Use the below research questions to guide your own analysis for your chosen data of any two countries. Remember that the quality of data may have holes for some countries. Be sure to check before proceeding any further.

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.

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 women 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.
- And other statistics which may be used to argue against a woman's interest in marriage

At Least Two Countries

All of these mentioned variables, when used together in a multi-linear regression model (where you choose relevant dependent variables) could be used to argue for the support or to refute a conclusion made by the article. Remember to chose data for at least two countries to confirm or refute the conclusions since agreement across multiple countries will support your argument.

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 is an interesting result that you are invited to discuss in your report document of your final deliverables, shown below.

Important Details

Note: Please remember to include your name on everything you submit for the class. If there are no included names of the members of the group, then the instructor will be unable to award credit for your work.

Required Deliverables

1. File `src/analysis.r`: The R source code that you used to answer your questions. The instructor should be able to run this file from scratch to completely recreate all of your analysis, graphs, models and p-values. Please be sure that you run this file to make sure that the data is loaded by the script and that there are no bugs in your code. To run your code from the terminal, run the following command.

   ```
   Rscript analysis.r
   ```

2. File `writing/report.md`: The Markdown-formatted report of your conclusions to refute or support each of the above questions in blue from the article. Even though your plots can be recreated from your source file, please attach your plots (with captions) to the report to help describe your points of view. Remember also to address the plots by name in your text.

   In your report, under each of the questions already in the Markdown file, please include the p-values work from t-tests and regression models that you use. Remember, each question will require some kind of model to make your point of view understood. Spend some time to interpret the results by exploring the results, the p-values and then briefly speculate why you think your conclusions are correct.

3. File `writing/report.md`; **Reflection**: 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 the insights from his talk in your reflection document.

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.