Econ 255

Spring 2014

Prof. Watson

**Group Empirical Project**

The purpose of the group empirical project is to allow you to practice the tools you have learned in class, and to expose you to the challenges and joys of empirical research. In groups of 1-3 people, you will write a 15-page empirical paper on a topic of your choice. You will give a 10-15 minute presentation to the class at the end of the semester explaining your project and highlighting your findings. You will then submit your paper, raw data, and Stata programs, as described by the research protocol below.

**Timeline.**

March 3. Group Selection due.

March 12. Topic Proposal Due.

Week of April 14. Group Meetings.

April 22. Data Appendix due.

May 7. Results Section Due.

Week of May 12. Presentations.

May 20. Final Paper and Protocol Components Due.

**Group selection.**

You will select a group of up to 1-3 people. E-mail me with the names and e-mail addresses of your group by March 3. If you would like to work with someone, but have trouble finding a partner, let me know by March 3 and I will try to match you with others in the same position.

**Topic Proposal.**

The initial project proposal is due March 12. It should be a PDF file of about two pages and include the following:

* Group Member Names and E-mail addresses
* Project Topic, with a specific well-defined question
* Overview of Empirical Strategy – what is your dependent variable, and what other variables will you consider?
* Data Sources, list of data sources that you will use for the project and where you will get them.
* Bibliography, list at least five major papers related to the topic.

**I strongly recommend that you meet with me before submitting your topic proposal.** As you develop the proposal, you are likely to find that data availability limits the scope of questions you can answer. See below for suggestions.

**Data Description.**

We will meet outside of class the week of April 14. By that time you should have collected your data and put them into Stata format. You should clean your data, paying attention to issues like missing values, adjusting for inflation, logs versus levels, etc.

The data appendix is due April 22. It should be a PDF file and include the following:

* Group Member Names and E-mail addresses
* Well-defined question
* A brief description of the empirical strategy, including a regression equation to be estimated
* Create a codebook for each variable you expect to use in your project. For each variable, describe:

1. Stata variable name.

2. Longer description of variable (label).

3. Discussion of where you got the variable or precisely how it was created.

4. Number of observations and discussion of why any observations are missing.

5. Basic summary statistics: Mean, Std. Dev., Min, Max

6. Histogram of variable.

7. Where relevant, graphical presentation of how the mean of the variable changes over time, varies across geographies, or differs across demographic groups of focus.

**Results Section.**

A draft of the results section is due May 7. This should be 3-5 pages of text plus tables. The results section should describe the empirical strategy and results. Include tables of regression output. Before submitting the results section, you should meet with me at least once more to talk about your analysis.

**Presentations.**

You will do a 10-minute group presentation of your research to the class. We will have presentations the last week of classes.

**Final Paper and Protocol Components.**

Each group will submit one final paper. The final paper is due Tuesday, May 20thth. It may not exceed 15 pages of text, plus tables, figures, and references. The final paper should be submitted both in hard copy to my office and on a USB key as described by the research protocol below. The USB key will also include raw data and programs.

**Conducting the Research Project.**

**1. Pick a topic**

First you have to decide what it is you are going to research. What is it that you want to know? Do you want to determine the effect of the minimum wage on poverty rates? Do you want to know whether hot weather causes crime? Obviously there are an infinite number of things you might want to know. But, at this stage in the process you should aim to answer two questions with great clarity.

What question are you asking?

Why is it important?

For example, if you are investigating whether a training program for unskilled workers has “worked” you must be very clear about what it means to have “worked.” Do you mean participants have higher wages? Lower unemployment rates? Higher “job satisfaction”? It is also often useful to describe why you think your question is important. What are the implications of you answer? Would your findings suggest a market failure? Or, would they suggest a particular policy response?

How do you get started? You might identify a topic of interest by reading the newspaper. Often topics that are under-researched or poorly understood are marked by vitriolic disagreement. You might also use the classification system developed by the *Journal of Economic Literature*. This journal provides summaries of research papers for a variety of topics. You might find yourself gravitating toward topics related to fiscal policy, or economic growth, or social welfare. You might be intrigued by an empirical paper you have read and want to learn more. More basically, you might simply be a little introspective. What topics do *you* find interesting? What questions capture *your* imagination?

Some examples of previous projects are:

The Effect of Alcohol Advertising on Alcohol Consumption

Wage Discrimination in Major League Baseball

The Effect of Weather on Crime

Predicting Oscar Winners

The Determinants of Wine Prices

The Effect of Enterprise Zones on Employment

Religion and Teen Pregnancy

Does Defense Win Games?

Are science majors more likely to be employed after graduation?

Note that your empirical approach should use the micro-econometric tools we will have studied in Econ 255. **This rules out time series analysis and some macro topics.** You may choose to usepanel data techniques we learn in class, but these are best suited for situations where the number of observations greatly exceeds the number of time periods.

**2. Learn what others have learned about this topic.**

Once you have identified a topic that you believe is interesting you should spend some time learning what others have done on that topic. There are a variety of places to turn and we will only mention a few that are particularly helpful. First, you might turn to the online index *Econlit* which provides a searchable database for research done in economics. This is available through the library website in the Articles and Databases section. Another useful source is the *Social Science Citation Index*. This index will let you learn which authors have cited a particular study in their own work. This helps you to track which papers seem to be “seminal” to the literature in which you are interested. Which papers seem to be cited time and time again? You might also check the *National Bureau of Economic Research* ([www.nber.org](http://www.nber.org)) which publishes current working papers on a variety of topics. Again, the *Journal of Economic Literature* provides helpful summaries of a wide range of economic research. After spending a few hours with these resources you will begin to get a feel for the intellectual “lay of the land” for the topic you have chosen.

**3. Search for Data**

You should be sure that your proposal includes data that are attainable. For instance, tax records are never made publicly available and do not expect the College to hand over students’ grades. Likewise, an original survey is a major undertaking which requires IRB approval. Your best bet is to look for existing, publicly available data. You may find that data availability requires you to restructure your plan – it is fine to refine your topic once you have seen the data possibilities.

Places to look for data:

ICPSR ([www.umich.icpsr.edu](http://www.umich.icpsr.edu)) A data clearing house with hundreds of data sets.

IPUMS (<http://www.ipums.umn.edu/usa>). Individual level Census/ACS data. The website is user friendly and allows you to make extracts. For confidentiality reasons the data set does not identify neighborhoods, but has a lot of information about demographics, labor market outcomes, housing, etc. IPUMS also houses user-friendly versions of the Integrated Health Interview Survey and the Current Population Survey.

The Census website ([www.census.gov](http://www.census.gov) ) has Census data aggregated to various levels of geography. The Census CD published by Geolytics and available in Sawyer includes similar information.

The NBER data page ([www.nber.org/data](http://www.nber.org/data)) has numerous data sets.

If you have a specific data need, you are likely to have some luck searching the web or asking me.

**4. Have an empirical strategy.**

At this juncture, you are going to have to think more carefully about how you will do your econometric analyses. Are you quantifying some relationship, testing a theory, or generating a forecast? Often the empirical exercise can be framed simply as: X affects Y. You have to decide what Y is, what X is, and the nature of their relationship. But, as a first step you will need data that include Y and X. Hopefully your investigation of the existing literature will have been helpful on this score. At this point you will likely apply the methods you have learned in this class. You will gather data and use the appropriate econometric techniques. You may compare means, estimate regression models, and so on. It might be useful to imagine yourself as Sherlock Holmes. There is a question you are trying to answer and you carefully and creatively bring a variety of evidence to bear in doing so.

### 5. Interpret your results.

Once you have implemented your empirical strategy on actual data you will need to interpret your results. What do they tell you? Are the results clear or ambiguous? Do they shed light on what is or is not appropriate policy? Can the results be generalized beyond the context implicit in your study? Are the results robust - do small changes in your empirical approach or the way you classify your data have a noticeable effect on your results? At this point you should be candid. You are trying to answer a question. You are *not* trying to prove *your* prior point of view on the matter. Be scientific. Report what you find with nothing “up the sleeves.”

### 6. Write a report.

To write up your report you can simply describe the steps you have taken. State the question you are addressing and why it is important. This can be done in a section titled “Introduction” Summarize the appropriate existing literature in a section titled “Review of the Literature” or “Background.” Describe how you’ve tackled your problem empirically and outline your econometric strategy in a section titled “Empirical Strategy”. Summarize your data, including the variables used and any restrictions you imposed in selecting your sample. Call this section “Data.” Interpret your results with a focus on the question you initially posed. Call this section “Empirical Results.” Identify any weaknesses in your analyses and provide ideas for future research. Call this section “Final Comments.” End with a summary of your main finding in a section called “Conclusions.” The format may be altered if modifications are required by your project. Tables and figures and a list of references should be at the end of the document.

There is a strict page limit of 15 pages of text (12-point font, double spaced). This page limit does not include any tables and figures or the reference page. In many cases you will be able to describe your work in as little as 10-12 pages – the emphasis should be on quality rather than quantity.

Follow the standard practices of good writing. Put thought into the organization of your paper. You should make clear early on in the paper what the question is that you are addressing, and why we should care about that question. Organize your paper into paragraphs, each of which addresses a particular coherent point. Make use of topic sentences to start you paragraphs and concluding sentences to end them. It is insulting to the reader to have a sloppy paper that has not been proofread. I take your writing seriously and expect you to do the same. You are strongly encouraged to make use of the writing tutors.

Tables and figures should be well-labeled and clear. Do not use Stata variable names; instead provide short variable explanations in your table. A good explanation of key points should be described in the text. What should the reader infer from looking at the tables? The reader should be able to get a good sense about your project just by looking at the tables, or just by reading the text.

**7. Present your project.**

Develop a 10-minute power point presentation of your project. You will not be able to present every detail of your project. Focus on the highlights.

**Sources.**

You will cite the relevant literature in your paper. The sources should be listed on the reference page at the end of the paper. The reference list should include the title, author, journal or book title, date published, and page numbers if appropriate. For internet references, include title, author, hosting organization, date published (if possible), and complete http:// address. For information in class that you have not found elsewhere, cite as in the example below. See sample reference citations below.

Sample citations for reference page:

Havemann, Judith, "Mass. City Gets a New Lease on Life; Cambridge Booms as Rent Control Ends, but the Poorer Pay a Price," *Washington Post*, September 19, 1998.

Troesken, Werner, “Race, Disease, and the Provision of Water in American Cities, 1889-1921,” *The Journal of Economic History,* LXI (2001), pp. 750-756.

Chin, James, *Control of Communicable Diseases Manual,* Washington D.C.: American Public Health Association, 2000.

Cutler, David M., and Ellen Meara, “The Medical Costs of the Young and Old: A Forty Year Perspective” in David Wise, ed., *Frontiers in the Economics of Aging*, Chicago:The University of Chicago Press, 1998, pp. 215-246.

Evans, William N, and Julie H. Topoleski, “The Social and Economic Impact of Native American Casinos,” (http://www.bsos.umd.edu/econ/evans/), 2002.

Watson, Williams College, Economics 255 Lecture, February 27, 2011.

In addition to the page of references at the end of the paper, you should cite in the text to give credit for any idea, piece of evidence, or quotation that is not your own. Citations in the text should put the author and the year in parentheses, e.g. (Ellwood, 2000). Every citation in the text should have a corresponding entry in the list of references. Any phrases taken from another author (including the textbook) must be enclosed in quotes and cited properly. If your ideas emerged from discussions with a fellow student outside your group, include a footnote stating “This point emerged through helpful discussions with David Ellwood.”, or something similar.

Here are some examples of acceptable citations:

* Gary Burtless (1991) argues that although international trade has indeed grown as a share of the U.S. economy in recent decades, the increase is too small to explain much of the rise in wage inequality over this period.
* Although international trade has indeed grown as a share of the U.S. economy in recent decades, the increase is too small to explain much of the rise in wage inequality over this period (Burtless 1991).

Any words that were written by anyone except for you must be enclosed in quotes, and in addition must include a proper citation to the source in the text. If you explain another author’s idea completely in your own words, you just need to provide a citation in the text of your paper. But if you use any sentences, fragments of sentences, phrases, or anything else that were written by someone else, you absolutely must put them in quotes. When something is not in quotes, you are saying that it is your own completely original writing. Even if you provide a citation, anything that is not in quotes is being presented as if you are explaining the author’s idea completely in your own original words. If it is not in fact your own original writing, you will be brought before the honor committee. The standard punishment handed down by the honor committee for presenting someone else’s words as if they were your own is failure in the course.

**Honor code.**

It is essential that you give proper credit in your paper for any idea or piece of information that is not completely your own original contribution. Failure to do so is a serious violation of the honor code and will be reported to the honor committee. A typical punishment handed down by the honor committee for such an infraction is failure in the course. Details about citations, references, and quotations are described above. If you are unclear about whether you need to cite a particular idea or piece of information, err on the side of caution and/or ask.

You may not hand in a memo that is substantively similar to work turned in for another course. If you are unclear about whether this applies to you, ask.

You are encouraged to make use of the writing tutors. If you do so, please include a footnote stating that you worked with them.

Each member of the group is responsible for the group following the Honor Code and may be held responsible for any violation on the group project. If any problem arises, please come talk to the professor as soon as possible.

**Getting help.**

You are encouraged to set up appointments with the professor to talk about the project as often as needed. Do not wait until the last minute; it may be too late at that point. If all of the group members cannot make it to a meeting, it is usually fine to have a subset there.

**Grading.**

The project is worth 20% of your final grade in the course. Of this, approximately 50% will be determined by your presentation and 50% will be determined by the paper itself. Please take both seriously. If a member of the group is not pulling his or her weight, please feel free to talk to me about it confidentially.

**Research Protocol**

For this project, you will use the Project TIER Research Protocol. This protocol, recently developed at Haverford College, improves transparency, replicability, and self-discipline in the research process. The final paper will be submitted with the raw data files (in the original file format and converted to Stata) and all do files necessary to implement the analysis in the project. After your project is submitted, I will run the do files you provide to confirm that the results you present in the paper can be replicated.

**I. Overview**

This protocol provides instructions for assembling a set of electronic files that document all the steps of data management and analysis you conduct for your empirical research project. This documentation includes four kinds of files: data, metadata, computer command files, and a readme file.

To begin, create a new folder and give it the name “Documentation.” Save this folder on your personal computer or some other site that you can easily access. This is the folder in which you will be storing documentation files as you create and modify them throughout the course of your research, and it is the folder in which the final versions of those files will be preserved.

**II. Data**

First create a new folder called “Data,” and place it in the top level of your “Documentation” folder. Then create two more new folders with the names “Original” and “Importable,” and place them in the top level of your “Data” folder.

***Original data files.***

In the “Original” data folder, you should save a copy of every file from which you obtained statistical data that you used for your project. These original data files should be preserved in exactly the format they were in when you first obtained them.

If all the data you used were contained in a single file when you first obtained them, then that file will be your only original data file; if your data were originally contained in two or more files, then each individual original data file should be included in the documentation. (If you use data from several different worksheets contained in a single Excel workbook, you should create an individual original data file for each worksheet.)

Your original data files should be given names of the form ***original\_'name'.'ext'***, where ***'name'*** is an informal name you choose to refer to the source and/or content of the data, and ***'ext'*** is the extension determined by the format of the file. For example, an original data file with data on US presidential elections that was in Stata’s **.dta** format when you first obtained it could be named ***original\_elections.dta***; a tab-delimited text file obtained from the Penn World Tables could be named ***original\_pwt.txt***; and an Excel-formatted file with data on development aid from the UN Common Database could be named ***original\_UNaid.xls***.

***Importable data files.***

For every file in your “Original” data folder, you should create a corresponding importable version, and store it in your “Importable” data folder. The importable version should be as similar to the original data file as possible. Changes should be made to the original data file only if the original data file is neither in Stata’s **.dta** format nor in a format that can be imported into Stata. When such changes are necessary, the file should be modified only in the minimal ways required to make it possible for Stata to open or import the data.

The modifications (if any) that you make to an original data file when you create the corresponding importable version will depend on the format of the of the original data file. Four cases are commonly encountered: (i) original data files that are in Stata’s **.dta** format, (ii) original data files that are in a format (such as tab-delimited or comma-delimited text) that can be imported to Stata with a command like insheet or infix, (iii) original data files that are in the format of one of the major statistical packages other than Stata (such as SPSS, SAS or R), (iv) original data files that are in Excel’s **.xls** or **.xlsx** format.

(i) If an original data file is in **.dta** format, the corresponding importable data file you create should simply be an exact copy of the original data file.

(ii) When an original data file is stored in tab- or comma-delimited text or some other format that can be imported to Stata with insheet, infix or a similar command, creating the corresponding importable file usually requires little or no modification of the original.

If the original file contains nothing but rectangular data organized into (delimited) columns representing variables and rows representing cases (with or without variable names in the first row), the importable version should simply be an exact copy of the original data file.

If the original data file contains additional information, like variable definitions, citations of the sources of the data, the URL of the website from which the file was downloaded, or any other explanatory notes or comments, then everything other than the rows and columns of data (perhaps with variable names in the first row of each column) should be

deleted from the importable version. The importable data file should then be saved in the same format (tab- or text-delimited) as the original version.

(iii) For original data files that are stored in the format of one of the major statistical packages other than Stata, we recommend using a program called Stat/Transfer to create the importable version of the file. With Stat/Transfer, you can easily convert a data file stored in the format of any of the major packages (including SPSS, SAS and R) into Stata’s **.dta** format. (See www.stattransfer.com for information about Stat/Transfer. It is available on certain Jesup computers.)

(iv) There are several possible ways of handling original data files that are in Excel’s **.xls** or **.xlsx** format.

One approach is simply to create a delimited text version of the file using Excel’s “save as” function, specifying the storage format as either tab-delimited text or comma-separated values (**.csv**). The only other modification that would then be necessary would be to remove any extraneous notes from the file, as described for (ii) above.

A second approach would be to use Stat/Transfer, which converts Excel files to Stata format exactly as it does for data files in SPSS, SAS or R format.

Finally, if you are using Stata 13, that version of the program is able to import data from Excel-formatted files; provided there are no extraneous notes or comments in the original Excel file, the importable version can simply be an exact copy of the original.

If an importable data file is an exact copy of an original data file, it should have the same name as the original. (For the sake of keeping your work organized, however, you should keep two copies of the file—one in your “Original” data folder and one in your “Importable” data folder.) If you modify an original data file in any way in the process of creating the importable version, the importable file should be given a new name. The new name should be similar to the name of the original data file but the ***original*** prefix in the name of the original file should be changed to ***importable***, and the extension should be changed as appropriate. For example, ***original \_pwt.txt*** would be renamed ***importable\_pwt.txt***, and ***original\_UNaid.xls*** would be renamed ***importable\_UNaid.csv***.

**II. Metadata**

Begin by creating a new folder called “Metadata,” and save it in the top level of your “Documentation” folder. Next create another new folder called “Supplementary Metadata,” and save it in the top level of your “Metadata” folder. Here the term metadata refers to information about or documentation of your original data files. What kinds of metadata are appropriate and necessary will vary a great deal, depending on the nature of your original data file. In some cases, documentation that accompanied the original data file, such as a codebook and/or a users’ guide, will contain all or some of the necessary metadata. We will refer to documentation that the producers or distributors made available with a data file as “native” metadata. When no native metadata are available, or when the available metadata need to be supplemented with additional information or explanatory comments, you will need to write some or all of the metadata yourself.

The first step in assembling your metadata is to decide what information or documentation you should provide for each of your original data files. The general principle guiding these decisions is that the metadata for a data file should contain all the information a user would need to understand the contents of the file, such as variable definitions, units of measurement, coding schemes, and sampling methods. In practice, making these decisions requires judgment.

Once you have decided what metadata you will provide for each of your original data files, you should create a document titled ***metadata.pdf***, which should be stored in the top level of your “Metadata” folder. This document should consist of one section, or “entry,” for each of your original data files. Each entry should begin with a brief bibliographic citation of the original data file, in a format that would be appropriate for the reference list of a research paper.

The entry for the data file should provide information about any native metadata for the file that is included in your documentation. Give a bibliographic citation for each item of native metadata, and describe briefly the relevant information it contains. If the item is available from a stable and publicly accessible source, explain how a user can obtain it. If the item is not easily available from a public source, save a copy of it in the “Supplementary Metadata” folder you created, and make a note in the ***metadata.pdf*** file indicating that it can be found there.

If you have written metadata for an original data file yourself, that information can usually be included under the entry for the data file that appears in the ***metadata.pdf*** file. If presenting the metadata you have composed for a data file in a separate document would be more effective (which might be the case, for example, if it is very long or consists of tables or figures), then create a separate document containing this metadata, and save it in the “Supplementary Metadata” folder. In the ***metadata.pdf*** file, under the entry for the relevant data file, give the name of the metadata file you created, describe briefly the information contained in that file, and make a note indicating that the metadata file can be found in the “Supplementary Metadata” folder.

**III. Do files**

The number of do-files you include in your documentation and how they are organized may vary for a variety of reasons, including the number of data files you have and how they are organized. The following instructions are written for relatively simple situations, in which the number of data files is not too large and no unusual complications arise. To begin, create a new folder with the name “Do-files,” and save it in the top level of your “Documentation” folder.

In the simplest case, you will create three do-files to include with your documentation: one (titled ***import.do***) that imports the data from the importable files you created and then saves them in Stata’s **.dta** format; one (titled ***cleaning.do***) that combines and processes the data as necessary to create the final data set used in your analysis; and one (titled ***results.do***) containing the commands needed to generate all the results you report in your thesis. All three of these files should be saved in your “Do-files” folder.

**It is very important to include thorough comments throughout all of the do-files you create.** These comments should be detailed and clear enough to make it possible for someone not familiar with your project to understand every step of data management and analysis executed by the commands in the do-file.

**1. import.do.** The purpose of the ***import.do*** file is to import the data from each of your importable data files that is not in **.dta** format, and then to save the imported data in a new file that is in **.dta** format. After creating and running ***import.do***, you will therefore have a **.dta**-formatted version of each of your raw data files.

If all of your importable data files are in **.dta** format, you do not need to create an ***import.do*** file.

For your importable data files that are not in **.dta** format, ***import.do*** should contain commands that import the data from the importable file into Stata (usually with the insheet or infix command) and then save it in **.dta** format. Each new **.dta**-formatted file created by ***import.do*** should be given a name that corresponds to the name of the importable version, but with the prefix “***import*\_**” dropped from the name and with the extension changed to **.dta**. For example, if the name of the importable file is ***import\_UNaid.txt***, the file created by ***import.do*** should be saved with the name ***UNaid.dta***. two or more of your importable data files were not in **.dta** format, your ***import.do*** file should contain one block of commands for each of your importable data files. Each block of commands should instruct Stata, as described above, to import the data from one of the importable files, and save the file in **.dta**.

Note that any **.dta**-formatted files that were created by commands in the ***import.do*** file should not be included in the electronic documentation that you turn in with your thesis; it is not necessary to include these with your documentation because anyone interested in replicating your analysis can create them simply by running ***import.do***.

**2. cleaning.do.** The purpose of the ***cleaning.do*** file is to process your data in whatever ways are necessary to create the final data set or sets that you will use for the analysis you present in your thesis.

Since there is a great deal of variation in the number and structure of the data files that students use and in the how the data need to be organized in preparation for analysis, it is impossible to give a comprehensive description of the specific commands that should be contained in ***cleaning.do***. In general terms, however, ***cleaning.do*** should take the data from the **.dta**-formatted versions of your data files, and then clean, merge, manage and organize them as required to create and save the final data set or sets that will be used for your analysis.

If your analysis will be conducted using just one final data file, the command in ***cleaning.do*** that saves the final data file should give it the name ***final.dta***. If you need to use data from more than one processed file to generate your results (e.g., some of your results might be generated using a file with data on individuals from different countries, while other results are generated from a file in which the data has been aggregated to the country level), the commands in your ***cleaning.do*** file that save the processed data files should give them names that include the prefix “***final\_***” (e.g., ***clean\_individual.dta*** and ***clean\_aggregated.dta***). Your ***cleaning.do*** file should contain just the minimal set of commands necessary to create your final data set or sets. You will almost certainly spend a good deal of time exploring and experimenting with your data before deciding exactly what analyses you want to present in your paper and how your final data need to be organized so that you can conduct the analyses you choose. In the course of that exploration and experimentation, it is likely that many commands will accumulate in your ***cleaning.do*** file, including many that turn out to be dead-ends or unnecessary for what you ultimately decide to do with your data. All such extraneous commands should be deleted, so that the ***cleaning.do*** file you turn in with only commands that do something necessary to prepare your data for analysis.

Note that the clean data file or files created when you run ***cleaning.do*** should not be included in the electronic documentation you turn in with your thesis. An independent researcher could create them by running ***cleaning.do*** (after running ***import.do***).

3. results.do. Your ***results.do*** file should contain commands that generate the results you report in your paper, using data from your final data file or files.

For every numerical result, figure or table presented in your paper, ***results.do*** should contain a command that opens the appropriate clean data file, followed by a command that generates the output showing the result. Each command that generates the output for a result presented in the thesis should be preceded by a comment that indicates where the result appears in the paper.

**IV. A readme file.** The documentation for your project should include a **.pdf** document, titled ***readme.pdf***, that gives an overview of the various files that you have assembled to document your project. In particular, the readme file should:

(i) list all the files included in the documentation, describe the content and format of each file, and outline the organization of the files into folders and subfolders; and

(ii) explain how the files included in the documentation can be used to replicate the results reported in the paper.

The instructions for replicating the results of the project [described in item (ii) above] should be precise and detailed. The objective is to ensure the transparency of all the steps required to access, process and analyze your data, so that it would be a straightforward task for an independent researcher to replicate all those steps

In the end, you will submit a USB drive with the following folder structure and contents:

**DOCUMENTATION\_lastname**

*readme.pdf (description of what is contained in documentation folder and sub-folders)*

**/DATA**

* . (Here replicability is used to mean using the same data to generate the same results.)
* . (Here replicability is used to mean using the same data to generate the same results.)

/**Original**

*This folder contains all raw data exactly as first obtained. (e.g. births.raw)*

* . (Here replicability is used to mean using the same data to generate the same results.)

**/Importable**

*This folder contains data files from the Original folder that have been converted to machine-readable form. There should be a 1:1 correspondence of files in the original folder and this one. Modifications should be only those required to convert the file to machine-readable format. (e.g. births.csv)*

/**METADATA**

*metadata.pdf (description of each original data file with links to source, bibliographic citation, and description of relevant supplementary files in supplementary metadata file)*

**/Supplementary Metadata**

*This folder contains codebooks and other materials referenced in metadata.pdf.*

*It may not be necessary to include documents with permanent web links.*

**/DO\_FILES**

*import.do (imports importable file into Stata format e.g. births.dta)*

*cleaning.do (works with Stata dataset to create variables and samples used in final analysis)*

*results.do (creates summary tables, runs regressions)*

**FINAL\_lastname**

*finalpaper\_lastname.pdf (final complete paper)*

*dataappendix\_lastname.pdf (data appendix described below)*

**REPLICATE**

*This file includes copies of .do files and importable data files*