A PRIMER ON CONDUCTING EMPIRICAL RESEARCH

For those that have never done it, empirical work seems pretty straightforward. You get the data, compute the statistics of interest, make a couple of tables and figures, write up what you did, and send the finished product off to a journal. In practice, things are a lot more complicated than this.

Empirical work is messy. It is not a tidy deductive process like proving a theorem with a clear beginning and end. The data always have problems. There are always areas where discretion comes into play. And there is no way to know for sure whether you are right about something -- empirical papers never end with a Q.E.D. symbol. Some people, by disposition, fundamentally can't handle this. If you are such a person, it is probably best to become a theorist or seek work elsewhere. The point of this note is to provide those with the necessary constitution some tips for keeping the mess manageable.

Empirical work consists of three stages: 1) project formulation, 2) obtaining/building the data, 3) data analysis. Each stage requires a different set of skills. Formulating an effective project requires creativity, some familiarity with the existing literature, and knowledge of available datasets and econometric techniques. Building the data requires sharp attention to detail, patience, and excellent computer skills. Finally, the best data analysts have a deep intuitive understanding of econometrics, a natural ability to detect and effectively summarize empirical regularities, and a commitment to standards of scientific objectivity. Because the skills required by these three stages are so different, it is increasingly common for economists to work in teams and exploit the gains to specialization. For example, one often sees an "idea guy" team up with a less mercurial "implementation guy" who makes sure things get done correctly. Part of your graduate education is trying out all three of these roles and learning your comparative advantage.

Project formulation

Perhaps the defining characteristic of an empirical project is that it has the potential to change the audience's mind about some feature of the world. Good empirical work provides facts which convincingly rule out certain patterns that one might have thought would have been present in the data. Ideally, those conjectures are thought to be consequential by a wide array of economists.

Tips for a good project:

Consider the Bottom Line For a project to be worth pursuing, someone, somewhere (maybe not in economics) has to believe that your results are important. Think hard about the set of conclusions you might reach and their broader implications for the literature. You are going to spend two years on this paper. Would anyone care about your results?

One sided tests are risky. Some projects will only be deemed interesting if they yield a certain result. Such projects are usually a bad idea, especially for graduate students as they create huge incentives for the researcher to find a particular result. A good empirical project is publishable even if the researcher's priors turn out to be wrong.

Measure the variable of interest The quality of an empirical paper is limited by the quality of the data. If you want to study corruption, it is best to have data on corruption. If you want to study unemployment, it is best to have data on unemployment. Try to conceptualize what the ideal data would like and then work hard to figure out what is available in the real world. This is more than a day long exercise. Read the literature, learn what datasets others have used, and scour the globe for the best alternatives.

Power over bias It is common for graduate students to propose projects with fancy research designs such as the regression discontinuity design that are nonparametrically identified but woefully underpowered. No one will be interested in your nonparametrically identified local average treatment effect if the standard errors are gigantic. Empirical work is about ruling things out. Nothing can be ruled out unless the sample size is large.

The build process

The first step in working with data is to construct a *build script*. This is a program that takes the raw data, reads it into your statistical package of choice, cleans it, and delivers a *workfile* used to conduct the main analysis. Getting the build process right is not glamorous but your credibility as an empiricist depends upon it. In many cases the build process takes longer, and involves more important decisions, than the corresponding analysis process. Take it seriously.

The build script may actually involve many programs and it is often a good idea to have a master build script which automates the process of calling all of the relevant subscripts in order and delivering the final workfile. The key is for there to be a clearly documented series of scripts which a third party could use to replicate your analysis sample exactly, for example, by typing 'stata-se –b do masterbuild.do'.

Tips for a good build:

Look at the raw data. List some records in your data and verify they make sense. Compute descriptive statistics and compare them to those in other published sources. Think carefully about the units your variables are in. Are the magnitudes you are finding plausible?

Never try to build the workfile interactively. Doing so dramatically increases the chances of making a mistake and makes it virtually impossible for your work to be replicated. Submit jobs in batch mode

Verify assertions about the built data. Have your script check that the data obey the properties you intend them to. For example, if you have two race groups (white and black) and have computed racial shares by neighborhood you can check that those shares sum to 1 in Stata with the command 'assert blackshr+whiteshr==1'. If this condition is violated, your do-file will fail.

Keep logs of the build process and scrutinize them. The larger your script becomes the more tempted you may be to turn various features of the output off. This is a bad idea. You want to be able to see what is going on. Sometimes you don't know what you are looking for and the log file is a good way to serendipitously discover a mistake.

Never store your raw data or the build scripts on your laptop or any other device with a short expected lifespan. The Berkeley economics department provides you with access to its Unix server which is backed up monthly. Learn to use it for storage.

Calm down. It is natural to be in a hurry to get to the main analysis and see the results. You must resist this temptation. Never hurry through the build file. If the build is wrong, it is pointless to do any analysis. Try to view the build process as itself an interesting analysis, which it typically is.

The analysis process

Once the workfile has been built, it is time to get to your analysis script. The structure of this script is going to depend heavily on the nature of what exactly you are trying to do. Again, you will eventually want a master analysis script which, when run, takes the workfile and delivers the estimates used in your paper.

Analysis Tips:

Start simple. Even if the contribution of your paper is a new technique, it is critical to start with the older simpler technique so that readers have something to benchmark your new method against. Simple cross-tabs and regressions can be informative in a bewildering array of environments of interest and are easy to compare across studies.

Change one thing at a time. You will want to engage in a variety of robustness checks. Do this in a controlled manner. Try to enumerate the changes under consideration and implement them one at a time in order to determine their relative contribution.

Automate tables. Learn to automate table creation. This helps to ensure you didn't make any mistakes copying and pasting. As a bonus you can reduce the risk of repetitive stress injuries to your wrists. A good choice in Stata is the program estout which can export to Excel, Latex, and many other formats.

Always report the number of observations. Any serious paper reports summary statistics in the first table and the number of observations used in each specification of each subsequent table.

Tell the truth. Economists go into projects expecting to find certain results. Do not become wedded to your expectations. The data often disappoint us. In such cases it is our responsibility to share that disappointment with the world accurately. Some of the profession's greatest empirical advances have come from the careful documentation of disappointing results.

The Labor Economist's Creed

These are my data. There are many like it, but these data are mine.

My data are my best friend. They are my life. I must master them as I must master my life.

My data, without me, are useless. Without my data, I am useless. I must build my workfile correctly. I must not make mistakes. I must replicate myself before I am replicated. I will...

My data and myself know that what counts is not the sophistication of our estimator or the elegance of our theoretical model. We know that it is the facts that count. We will report them faithfully...

My data are human, even as I, because they are my life. Thus, I will learn them as a brother. I will learn their weaknesses, their strengths, their parts, their accessories, their codebook and their origin. I will ever guard them against the ravages of weather and damage as I will ever guard my legs, my arms, my eyes and my heart against damage. I will keep my data clean and ready. We will become part of each other. We will...

Before God, I swear this creed. My data and myself are the defenders of my profession. We are the masters of our enemy. We are the saviors of my life.

So be it, until economics is a science…