Economics 304 Fall 2014
Country-Analysis Project Part 3: Data Collection

Introduction

In this part of the project, you will collect basic macroeconomic data for your country. In some ways this is the most important part of the project because without a good database you will be unable to obtain satisfactory results on later parts. It is also the most potentially frustrating because data collection is subject to the infamous 90/10 rule: 10 percent of the task requires 90 percent of the work. If you find the data collection for your country to be rough going, you should seek the early assistance of the instructor or of Reed Library data specialist Ryan Clement.

There is a set of variables for which it is essential that you have a usable sample and others that are desirable but not necessary. Every dataset will be incomplete or unsatisfactory in one way or another. You cannot be a perfectionist about the data themselves, but you *must be* a perfectionist about the careful documentation of precisely what data series you are using and its source. You should get your basic dataset assembled at this point of the semester, but it is likely that you will need to modify or augment your data as the semester goes on and you encounter specific issues in later assignments. Be flexible.

You will have to make many small decisions about variables based on very imperfect information. There will be many variants of some variables from multiple sources. In many cases it may not make much difference, but learning to make such choices is part of what you are learning in this project. There are a few general principles that you should follow in making these decisions:

- First, to repeat what I said above, be scrupulously careful in documenting your choices; based on your documentation another student should be able to follow your footsteps and obtain the same data (or an updated version). Document both the proximate source from which you obtained the data and the original source that collected the series.
- Second, be internally consistent whenever possible, gathering related series from the same source and data family. When using secondary data sources such as ProQuest, pay careful attention to the original source of the series because that is the consistency that matters. For some countries, many series will be available from the OECD, so you might get all possible series there, then fill in missing series from the IMF, World Bank, and other sources.
- Third, pay attention to annotations in the dataset indicating breaks in the series. These may be minor changes in how the data are collected that have been "patched" by the data collector to make them (arguably) usable as a continuous series. But they

may also be radical changes in the series that make the numbers before and after the break incomparable, at least without adjustment.

- Fourth, pay careful attention (and document) the units of measurement. Is the series nominal or real? Is it aggregate or per capita? Is it an index number (and what is the base period)?

Data frequency and sample period

Annual data are typically used for looking at long-run phenomena such as growth and inflation. Quarterly data are often preferred for examination of business cycles. You will do both, so ideally you would have two, separate datasets: one with annual data and one with quarterly data. Some variables (often monetary aggregates, interest rates, unemployment, and some price indexes) are published at monthly frequency. I will not expect any analysis of monthly data (though you are welcome to do this if you want), so you should collect annual and quarterly versions of variables when possible. There may be situations where variables are published in monthly but not quarterly form. For these variables, download the monthly versions and use Stata to aggregate them into quarterly form. (I'll provide directions for doing this.)

Sub-annual data are always subject to potential seasonal fluctuations. We would like to ignore these seasonal movements, so you should collect "seasonally adjusted" data when they are available. If there is no seasonally adjusted series, it might mean that the variables exhibit no seasonal movement (interest rates, for example) or it might mean that the publisher of the data has not gone to the trouble to adjust the series. There are easy ways that we can test for the presence of seasonality and slightly less easy ways to correct for it if it is present.

At a bare minimum you will need 20–25 years of data in your sample. This means that most series must be available on a consistent and comparable basis back before 1990. Ideally, your data would go back to 1950, but 1970 or 1980 is more realistic for some countries. If you can get most of the essential series back to about 1980, you'll probably have enough raw material to complete the project.

We will generally want variables that are measured in currency units to be in "real" terms, adjusted for price-level changes. You can use the GDP price index to convert any nominal variables into real ones. Be careful to use the same price index for all such conversions. Some series may be available in international currencies (usually dollars) converted at "purchasing-power parity." This means that the variable has been converted to dollars using an exchange rate based on the real purchasing power of the two currencies rather than the market exchange rate. This is generally preferable, but unless you can get all of your data in PPP form, it may be safer to use the domestic currency versions.

Some series may be available in per-capita as well as aggregate form. The per-capita variant will be useful at times, but it's probably better to collect the basic series in general form and calculate the per-capita version using a population variable.

Essential variables

The bare minimum set of variables would include the following (with suggested variable names in brackets):

- Real ("constant prices") GDP and a price index for it (or real and nominal GDP, from which you can calculate the price index) [*gdp, gdpn, pgdp*]
- Component breakdown of GDP into real expenditure components: consumption, investment, government spending, exports, imports [*cons, inv, gov, ex, im*]
- Employment, labor force, and unemployment: total numbers, from which you can calculate rates, or an unemployment rate and one of these numbers, which would allow you to infer the levels [*emp, lf, unemp, urate*]
- Average hourly or monthly wage rate [*wage*]
- Interest rates: ideally a one-month or three-month government borrowing rate, a long-term government bond rate, whatever short-term rate is used by the central bank as a monetary-policy target, and perhaps a long-term corporate rate [*rg1m, rg3m, rg10y, rg30y, rcorp*]
- Money supply variables (assuming your country has its own currency): narrow and broad money supply, monetary base [*m1, m2, mbase*]
- Government finance variables: Government spending, taxes, transfers (or net taxes = taxes minus transfers), government debt and deficits [*gov, tax, trans, taxn, gdebt, gdef*]
- Foreign variables: "effective" nominal and real exchange rates (in units of home currency per unit of foreign currency, assuming your country has its own currency), imports, exports, international capital flows [*xrn, xrr, im, ex, icf*]
- Population: so that you can put your series into per-capita form when desired [*pop*]

Additional desirable variables

These variables would be a useful to supplement some parts of the project, but you can work around their absence if necessary:

- Capital stock [*cap*]
- Consumer price index or equivalent [*cpi*]
- Indexes of stock-market prices [*stkp*]
- Breakdown of GDP by income categories (labor earnings, capital income, etc.)
- Breakdown of GDP by industry (value added)

Data sources

A vast array of data series are available electronically through the Reed Library. Most can be downloaded as an Excel spreadsheet, which can then be copied to a Stata data file if

necessary. You will need to be on the Reed campus network to access many of these datasets. In all cases, you should not have to log in; the site should recognize that you are a Reed user and use the library's subscription to grant you access to the data.

Students of your generation are tempted to attack such problems by simply searching in Google, typing in, for example, "Ecuador GDP data" and following whatever links come up. This is ***not*** a good strategy and should be used only as a last resort. While there are many Web sites that may reproduce data, few will have historical data and those that do may not update them when the collectors of the data issue revisions (which happens frequently). You are responsible for the reliability of your data; the locations cited below are definitive macroeconomic data sources and should be used unless you find that something you need is unavailable there. If you use alternative data sources, document them closely and provide an assessment of their likely reliability.

- **OECD**

http://www.oecd-ilibrary.org/content/data/data-00285-en
The Organization for Economic Cooperation and Development includes most of the richest countries of the world. Some were original members and have data going back into the 1950s; others joined later and will have shorter samples. The countries currently available on the OECD site are Australia, Austria, Belgium, Canada, Chile, Czech Republic, Denmark, Estonia, Finland, France, Germany, Hungary, Iceland, Ireland, Israel, Italy, Japan, Korea (South), Luxembourg, Mexico, Netherlands, New Zealand, Norway, Poland, Portugal, Slovak Republic, Slovenia, Spain, Sweden, Switzerland, Turkey, United Kingdom, United States, plus China and Colombia.

GDP and other aggregates are under "National Accounts." There are many other categories of data available from the OECD, and the quality of data is generally very high. If you are working on an OECD country, this should be your first stop.

- **ProQuest**

http://datasets.proquest.com/
ProQuest is a commercial data aggregator to which our library subscribes. It provides access to many national databases of individual countries as well as to some of the major international databases (e.g., Eurostat, International Labour Organization). But it does not seem to include the OECD, IMF, and World Bank macroeconomic databases on a comprehensive basis. It has a search facility that makes it fairly easy to find a particular series and also a reasonable "browse" feature in which you can limit your search to individual countries, subjects, and/or data providers. ProQuest is useful, but not a panacea for your data needs.

- **IMF**

http://www.imf.org/external/data.htm

The International Monetary Fund collects some basic macroeconomic data along with detailed international-transaction data for a huge number of countries. The database of greatest use to macroeconomists is International Financial Statistics (IFS). This is probably your best source for data on monetary aggregates, interest rates, and exchange rates.

The IMF data interface can be a little confusing because (like most data sources) it is designed primarily for people interested in current/recent data, not for those of us who want a long time period. The tool I found most useful on this site is the "Query across Datasets" tool. This will allow you to specify a country, a set of series, and a time period. It will display the data in a spreadsheet, which you can then download and manipulate in Excel.

- **Eurostat**

http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home
Eurostat is the official statistical bureau of the European Union. It compiles lots of data on member countries, though much of it does not go back before individual countries joined the EU. If you click on the Statistics Database button, you'll get a "Data Navigation Tree" through which you can look for data series. Eurostat data are supposed to be included in ProQuest, so it may be unnecessary to go directly to the Eurostat site. Additional monetary data for European countries may be found at the European Central Bank site, although these data series are often extremely limited in historical length because the ECB has only existed for a little more than a decade.

- **Other sources**

Other data-publishing organizations include the International Labour Organization (which has definitive "harmonized" unemployment rates that are more comparable across countries than other series), the United Nations and its various sub-organizations, and the World Bank (which focuses primarily on lower-income countries). You may wish to explore the Web sites of these organizations to fill in difficult series for your country.

Notes on variables

It can be awkward to work with variables whose values are very large numbers or very small numbers. I recommend rescaling variables into thousands, millions, or billions in order to keep the actual number in a reasonable range (say, between 1 and 100,000). It will be easiest to interpret if you put rates into percentage terms by multiplying by 100 when necessary—downloaded unemployment, interest, and inflation rates will probably already be in percentages but if you calculate them yourself (say, but dividing the number unemployed by the labor force) you'll need to do that manually.

To the extent that you end up using Stata for your analysis, variable-naming issues are quite important. Variable names in Stata are case sensitive and most people find it easiest to use lower-case letters only. It would be convenient (for the instructor) if all members of the class used the same basic set of variable names, based on the ones listed in the variable lists above.

Stata allows underscores in variable names, so they are convenient dividers when creating variations on a variable. For example, make the general unemployment rate *urate* and if you want to look at male unemployment, make it *urate_male* or something similar. When converting a variable that is initially in nominal terms (such as the money supply), you might use *m1_r* for the real version. When calculating the growth rate of GDP, you could use *gdp_g*.

What your end product should look like

The outcome of this part of the project is:

- Two datasets, one annual and one quarterly (or three, if you retain monthly data). Each dataset should be in an Excel spreadsheet. The year or year and quarter should be in the first columns, and the variables in the subsequent columns. (I recommend having the year in the first column and the quarter number in the second for the quarterly dataset. This is easy to move into Stata.) The variable names should be in the first row with data below. Missing observations should be empty cells.
- A document describing each variable: what it is, its source (including all details required to find the variable on the Web site you used), its units of measure, and any details or anomalies about the variable (including any discontinuities).