Writing an Empirical Project

This handout offers general guidelines on writing an empirical paper or project. This discussion is followed by two project topics (including data). You should choose one of these two project topics.

Note that, in your EC 306 project, you do not have to include every section I talk about below (I make explicit what you do not have to do below). However, I include a complete discussion of what an empirical paper involves since you may have to do one as a student (e.g. for your dissertation) or in your future job. Hence, I think it useful that you see the complete picture, rather than just the precise steps necessary for the EC 306 project.

Description of a Typical Empirical Project

Economists are engaged in research in a wide variety of areas today. Undergraduate and graduate students, academic economists, policymakers working in the civil service and central banks, professional economists working in private sector banks or industry -- may all need to write reports that involve analyzing economic data. Depending on the topic and intended audience, the form of these reports can vary widely, so that there is no one correct format for an empirical paper. With this in mind, we provide common elements of economic reports below as a guideline for future empirical work. Note, however, that, in the context of your own undergraduate projects or careers, it may not be necessary for you to include all of these elements in your report(s).

1. **Introduction**. Most reports begin with an introduction that briefly motivates and describes the issue being studied and summarizes the main empirical findings. The introduction should be written in simple non-technical language, with statistical and economic jargon kept to a minimum. A reader who is not an expert in the field should be able to read and understand the general issues and findings of the report or paper. [For EC 306 this can be very brief, 1 page at most]

2. **Literature Review.** This should summarize related work that others have done. It should list and very briefly describe other papers and findings that relate to yours. [For EC 306 do not include this section]

3. **Economic Theory.** If the report is academic in nature and involves a formal theoretical model, then it is often described in this section. For policy reports you may not need to include a formal mathematical model, but this section allows you to describe the economic or institutional issues of your work in more detail. This section can be more technical than the preceding ones and will typically include some mathematics and economic jargon. In short, you can address this section solely to an audience of experts in your field. [For EC 306 do not include this section]

4. **Data**. In this section you should describe your data, including a detailed discussion of its sources. [For EC 306, since I am giving you the data, you do not have to include this section]

5. **The Model to be Estimated.** In this section you should discuss how you use the data to investigate your particular economic questions of interest. The exact form of this section might vary considerably, depending on the topic and on the intended audience. For instance, you may want to argue that a particular regression is of interest for the study, that a certain variable will be the dependent variable and that other variables will be the explanatory variables. Similarly, in a macroeconomic time series exercise, you may wish to argue that your economic theory implies that your variables should be cointegrated and that, for this reason, a test of cointegration will be carried out. In short, it is in this section that you should justify the techniques used in the next section. [This is an important section for EC 306. Perhaps 2 pages long.]

6. **Empirical Results.** This section is typically the heart of any report. At this stage you should describe your empirical findings and discuss how they relate to the economic issue(s) under investigation. It should contain both statistical and economic information. By "economic" information we refer, for example, to coefficient estimates or to a finding of cointegration between two variables, and what these findings may imply for economic theory. In contrast, "statistical" information may include: results from hypothesis tests that show how coefficient estimates are significant; a justification for choice of lag length; an explanation for deleting insignificant explanatory variables; a discussion of model fit (e.g. the R^2 or outliers); etc. Much of this information can be presented in charts or graphs. It is not uncommon for papers to begin with some simple graphs (e.g. a time series plot of the data) and then follow with a table of descriptive statistics (e.g. the mean, standard deviation, and minimum/maximum of each variable, and a correlation matrix). Another table might include results from a more formal statistical analysis, such as OLS coefficient estimates, together with t-statistics (or P-values), R^2s and F-statistics for testing the significance of the regression as a whole. [This is the most important part of the EC 306 project, perhaps 3-4 pages long]

7. **Conclusion.** This should briefly summarize the issues addressed in the paper, specifically, its most important empirical findings. [For EC 306, this can be brief, probably less than 1 page]

General Considerations

The following contains a discussion of a few of the issues that you should keep foremost in your mind while carrying out an empirical project. In particular, it discusses what constitutes good empirical science and how you should present your results.

The first thing worth stressing is that there are no right or wrong empirical results. *Empirical results are what they are and you should not be disappointed if they do not show what you had hoped they would.* In an ideal world, a researcher comes up with a new theory then carries out empirical work that supports this new theory in a statistically significant way. *The real world very rarely approaches this ideal.*

In the real world, explanatory variables that you expect to be statistically significant often aren't significant. Variables you expect to be cointegrated often aren't cointegrated. Coefficients you expect to be positive often turn out to be negative. These results are obtained all the time -- even in the most sophisticated of studies. They should not discourage you! Instead, you should always keep an open mind. *A finding that a theory does not seem to work is just as scientifically valid as a finding that a theory does work.*

Furthermore, empirical results are often unclear or confusing. For instance, one statistical test might indicate one thing while another the opposite. Likewise, an explanatory variable that is significant in one regression might be insignificant in another regression. There is nothing you can do about this, except to report your results honestly and try (if possible) to understand why such conflicts or confusions are occurring.

It would be rare for an economist to completely falsify his/her results. Often, however, s/he may be tempted to do slightly dishonest things in order to show that results are indeed as economic reasoning anticipated. For instance, it is common for a researcher to run a large number of regressions with many different explanatory variables. On the whole, this is a very wise thing; a sign that the researcher is exploring the data in detail and from a number of angles. However, if the researcher presents only the regression that supports a particular theory and not the other regressions that discredit it, s/he is intentionally misleading the reader. Always avoid this temptation to misrepresent your results!

On the issue of how results should be presented, we cannot stress enough the importance of clarity and brevity. Whether it is a good thing or a bad thing, it is undoubtedly the case that university lecturers, civil servants, policymakers and employers are busy people who do not want to spend a lot of time reading long, poorly organized and verbose reports.

One key skill that writers of good reports show is selectivity. For example, you may have many different coefficient results and tests statistics from your various regression runs. An important part of any report is to decide what information is important and what is unimportant to your readership. Select only the most important information for inclusion in your report and -- as always -- report honestly and openly the results that you obtain.

Project Topics

The following are two project topics that you may wish to undertake. You may obtain the data sets through http://eu.wiley.com/legacy/wileychi/koopdata2ed/.

Project 1: The Equity Underpricing Puzzle

Background

Investors and financial economists are interested in understanding how the stock market values a firm's equity (i.e. shares). In a fundamental sense, the value of a firm's shares should reflect investors' expectations of the firm's future profitability. However, data on expected future profitability is non-existent. Instead, empirical financial studies must use measures such as current income, sales, assets and debt of the firm as explanatory variables.

In addition to the general question of how stock markets value firms, a second question is also receiving considerable attention by financial economists in recent years. By way of motivating this problem, note that most of the shares traded on the stock market are old shares in existing firms. However, many old firms will issue some new shares in addition to those already trading -- what are referred to as "seasoned equity offerings" or SEOs. Furthermore, some firms that have not traded shares on the stock market in the past may decide to now issue such shares (e.g. a computer software firm owned by one individual may decide to "go public" and sell shares in order to raise money for future investment or expansion). Such shares are called "initial public offerings" or IPOs. Some researchers have argued on the basis of empirical evidence that IPOs are undervalued relative to SEOs (although very recent work has suggested the opposite).

In this project, you are asked to empirically investigate these questions using the following data set. With this project topic, it is important to discuss the issue of heteroskedasticity.

Data

Excel file EQUITY.XLS contains data on N=309 firms who sold new shares in the year 1996 in the US. Some of these are SEOs and some are IPOs. Data on the following variables is provided. All variables except SEO are measured in millions of US dollars.

- VALUE = the total value of all shares (new and old) outstanding just after the firm issued the new shares. This is calculated as the price per share times the number of shares outstanding.
- DEBT = the amount of long-term debt held by the firm.
- SALES = total sales of the firm.
- INCOME = net income of the firm.
- ASSETS = book value of the assets of the firm (i.e. what an accountant would judge the firm's assets to be worth).
- SEO = a dummy variable that equals 1 if the new share issue is an SEO and equals 0 if it is an IPO.

Project 2: Wage-Setting Behaviour

Background

This project allows you to investigate wage-setting behavior using time series data. The general issue of interest in such analyses is how wages depend on macroeconomic factors such as the price level, GDP and variables reflecting employment and the labor force. An empirical analysis of such data must involve a discussion of issues such as unit roots and cointegration.

Data

Excel file WAGE.XLS contains annual UK data from 1855 through 1987. The natural logarithm of all variables has been taken. Data on the following variables is provided.

- W = the log of nominal wages.
- P = the log of consumer price index.
- GDP = the log of real GDP.
- E = the log of total employment.
- L = the log of total potential labor force.

Further Background

In addition to the general issue of wage-setting behavior, economic interest often focuses on some functions of the variables provided here. If you remember the properties of the logarithm operator, such as $\ln\left(A/B\right) = \ln(A) - \ln(B)$ and $\ln(1+A) \approx A$, you can derive the following relationships:

- the log of real wages = W-P.
- the log of productivity per worker = GDP-E.
- the log of the unemployment rate ≈ L-E.
- log of the share of wages in GDP = W-P-GDP+E.

One issue you may be interested in investigating is whether the relationships above are cointegrating relationships. In the lectures, we considered estimating the cointegrating regression using OLS techniques – something you may want to explore in your project. You may also wish to use the relationships above to tell you what the coefficients in the cointegrating regression might be. For instance, if the log of real wages equation above is a cointegrating relationship, then the regression of W on P should be:

$$W_t = P_t + e_t.$$

In other words, $\alpha=0$ and $\beta=1$. You can either estimate the regression of W on P, or impose $\alpha=0$ and $\beta=1$ and see whether these values imply cointegration. In this project, I suggest that you consider using both strategies. That is, you can either estimate a regression using OLS and then test the residuals for a unit root or you can impose a possible cointegrating relationship and then test the residuals for a unit root.

The previous material does not focus directly on the issue of wage-setting behaviour. You may want to do other tests or estimate other regressions in addition to (or instead of) the tests suggested above.