OVERVIEW OF ESTIMATION FRAMEWORKS AND ESTIMATORS

INSTRUCTOR: KLAUS MOELTNER

Textbook chapters: G14, K1
R material: `mod1_1a`, `mod1_1b`

1. Definitions

Econometric framework
An econometric framework results when a theoretical economic model (e.g. a specific production function) is combined with empirical or simulated data and statistical assumptions related to these data. Some frameworks rest on many such assumptions, others on barely any.

Estimator:
A rule or strategy for using the data to estimate the unknown elements in the econometric framework (usually referred to as parameters). The set of feasible or permissible parameters is called parameter space. (Example: The mean of a normal distribution can be any real number, so its space is $\Re$, but the variance has to be positive, i.e. reside in $\Re_+$). Economic theory can limit the parameter space.

Estimation
Estimation means to (i) specify a *decision criterion* to choose from competing sets of parameters (ex: "minimize the sum of squared residuals", or "maximize the likelihood function"), and (ii) choose a mathematical and / or statistical procedure to implement the criterion (e.g. Ordinary Least Squares, or Maximum Likelihood Estimation using Newton's gradient method, etc.).

Statistic
Any function computed from the data contained in a sample. If the data are interpreted to be manifestations of random variables, the statistic itself will be a random variable. Its distribution is called sampling distribution.

2. Common issues related to all estimators

2.1. Sample vs. population.

- In any econometric research we first need to define the "population of interest"
- Usually, however, we ultimately work with a sample of data < total observations in the population
- In virtually all econometric applications we desire to "infer" information on the population from our sample (in a parametric framework this usually means: we desire to learn more

about population parameters). Examples of inferences are:

- Estimate marginal effects of explanatory variables on the dependent variable
- Test if specific economic theories hold in our population
- Predict future or "out-of-sample" outcomes of the dependent variable

Example:
Population = all employed U.S. residents age 18 - 65 in March 1995; Sample = 1000 randomly drawn individuals

2.2. **Link to economic theory.**

- In "traditional" Econometrics (as opposed to other fields that use statistical analysis) researchers generally attempted to let economic theory and intuition guide them in data collection & analysis.
- In many cases, theory and intuition largely dictated the choice of econometric framework, especially when econometric analysis was supposed to test underlying theory.
- However, a strong link with a specific theoretical model is not always needed for econometric analysis to be useful. For example, it has been found over again that econometric models with weak links to theory are still well suited to accurately predict future economic outcomes.
- In this course we consider any research question that is of economic (and thus policy) relevance to be fair game for econometric analysis. This loosens the link (some might say shackles...) to underlying structural theory, and allows a broader spectrum of analytical tools to enter the scene.

Some recent papers on the position and role of econometrics relative to economic theory and structural modeling: Angrist and Pischke (2010), Keane (2010b), Keane (2010a), Heckman (2010), Nevo and Whinston (2010), and Imbens (2010).

3. Types of Estimation Frameworks

The most straightforward way to distinguish between estimation frameworks is by the number and importance of underlying *assumptions*. Econometricians generally make two types of assumptions:

- Assumptions flowing from underlying economic theory (e.g. "the effect of price on own-demand has to be negative", "all individuals in the sample are assumed to be risk-neutral", etc.)
- Assumptions associated with the statistical properties of the underlying *data generating mechanism*(e.g. "the error term is normally distributed").

> **Data Generating Mechanism (DGM)**
> In most econometric frameworks (for exceptions see below) we view our data as a collection of *random variables*. The DGM is the statistical framework that we believe has "generated" the observed data. The DGM usually takes the shape of one or more specific *density functions* from which we believe the observed data were "drawn".

Making a blatantly wrong assumption is just about the worst thing that can happen in econometric estimation. In most such cases, estimation results will be misleading. Thus the high degree of caution and reluctance with which econometricians make assumption, and the general desire to "get away" with as few assumptions as possible. An estimation framework that builds on few(er) assumptions is called more "robust" than one relying on more assumptions.

However, a minimum of assumptions is often needed to address the research question at hand (especially when the objective of econometric estimation is to test economic theory). Frameworks that rest on few assumptions usually allow only for limited inference (have low "inferential power"). Furthermore, any estimator that relies on correct assumptions will generally have more desirable properties than a more robust counterpart.

The researcher must make a judgment call on the trade-off between assumptions and inferential power. A good basic strategy: Avoid all "risky" assumptions, i.e. assumptions that are not supported by existing research, intuition, and / or a basic statistical inspection of the data.

I like this cite from Kennedy (2003), p. 2 (based on Malinvaud, 1966):

"The art of the econometrician consists in finding the set of assumptions which are both sufficiently specific and sufficiently realistic to allow him to take the best possible advantage of the available data."

4. `R` practice: Illustration of assumptions and generality

R scripts `mod1_1a` and `mod1_1b` illustrate the consequences of making assumptions of different levels of generality in model specification.

Usually, a *more general* (or, alternatively put, *less restrictive*) set of assumptions calls for the estimation of a larger set of parameters. A more restricitve model usually forces some of these parameters to share the same value, or otherwise follow some pre-defined mathematical relationship.

In script `mod1_1a` we generate wage data with different means and standard deviations for female and male workers. Thus, a more general regression model would allow for different sets of parameters for male and female workers. Forcing the two groups to share one or more common parameters is called "pooling", or a "parameter equality restriction". In this case, pooling violates the underlying DGP, and leads to misleading results.

`mod1_1a` shows that in some cases, the more general model can be the correct (and thus a very efficient) model, even though it must estimate more parameters than the naive model, and is based on an (inherently inefficient) 2-step method.

For `mod1_1b` we assume that the moments of the wage distributions are identical for both genders. As a result, the basic pooling model (the most restrictive model in `mod1_1a`) now becomes the correct, and also the most efficient model. The other models, while more general, have larger standard errors for estimated parameters.

Note that a big advantage from estimating a more general model is that we can usually *rigorously test* for the legitimacy of parameter restrictions. By running a restricted model from the onset, we deprive ourselves of this additional opportunity to learn from the data.

5. Broad categories of estimation frameworks

(Fully) Parametric:
From the very onset a parametric model makes assumptions on the density and parameters of the data generating process (DGP). Examples: Maximum Likelihood Estimation, Least Squares (with density assumption on the error term)

Semi-parametric: Work with parameters, but avoid assumption on some or all the densities specified in the fully parametric model. Examples: Least Squares without density assumptions (only conditional mean of y is specified) Method of Moments (MoM), Generalized Method of Moments (GMM))

Non-parametric: Discards all population assumptions on functional form and distribution. Works entirely without parameters. Estimates densities solely from information contained in sample. Naturally, this guards against (almost) any wrong assumption on model structure and densities. Drawback: Less efficient than a correctly specified parametric model. Nonparametric regression suffers from the "curse of dimensionality" - it becomes intractable with more than a handful of (continuous) regressors. Thus it has somewhat limited inferential ability.

6. Criteria for Choosing Amongst Frameworks and Estimators

In most cases the underlying economic research question will require a semi- or fully parametric framework. If the same question can be answered semi-parametrically, this will usually be preferred to a parametric approach based on robustness considerations.

However, in most situations a parametric framework will be required to fully address common research questions. Within this broad framework, the winner from a competing set of estimators can be chosen based on one or more of the following criteria:

- Rigorous statistical hypothesis tests (we'll learn a bunch of those)
- Computational feasibility
- Data availability
- Consistency (A key criterion that relates to the performance of an estimator under (hypothetically) increasing sample size).
- Efficiency (A more efficient estimator utilizes the data better than a less efficient one - this usually manifests itself in a smaller variance for the sampling distribution of estimated parameters, i.e. smaller standard errors and larger t-values). In general, the "efficiency" criterion only applies to a set of competing estimators that have already been found to be "consistent". Efficiency is a moot concept in the absence of consistency.

References

Angrist, J. D. and Pischke, J. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics, *Journal of Economic Perspectives* **24**: 3–30.

Heckman, J. (2010). Building bridges between structural and program evaluation approaches to evaluating policy, *Journal of Economic Literature* **48**: 356–398.

Imbens, G. (2010). Better late than nothing: Some comments on deaton (2009) and heckman and urzua (2009), *Journal of Economic Literature* **48**: 399–423.

Keane, M. (2010a). A structural perspective on the experimentalist school, *Journal of Economic Perspectives* **24**: 4758.

Keane, M. (2010b). Structural vs. atheoretic approaches to econometrics, *Journal of Econometrics* **156**: 3–20.

Kennedy, P. (2003). *A Guide to Econometrics*, 5th edition edn, MIT Press.

Nevo, A. and Whinston, M. (2010). Taking the dogma out of econometrics: structural modeling and credible inference, *Journal of Economic Perspectives* **24**: 69–82.