On the Nature of Econometrics and Economic Data

1. Purpose of an Empirical Analysis

A. Description: What's the relationship between Y and X in a given population?

- What's the relationship between grades in Eco2408F and attributes such as grade in intermediate statistics, gender, completion of an advanced UG course in econometrics, obesity, etc.

B. Prediction: Estimate the relationship between Y and X; then use it to predict the value of Y_0 given X_0.

- Use last year's student data to predict your grade in Eco2408 this year.

C. Estimation of a Causal Relationship: Find a relationship between Y and X that is invariant to manipulation of one or more of the components in X.

- By how much would the class average increase if we changed the entrance rules and required all MA students to complete an advanced UG course in econometrics? What would happen to a student's grade if we could change, ceteris paribus, the student's sex?

2. Specification Issues: Economic Components

A. Dependent variable: What's the right choice of Y?

- Suppose you are interested in measuring the teaching output of professors. How should you do it? Teaching evaluations? Or some measure of student performance: Performance in subsequent courses? Successful completion of degree? Starting salary?

B. Independent variables: What's the right choice for X?

- It depends on your purpose. Suppose you decide that you want to look at the relationship between Y=soft-drink consumption and some X variables.
  - If you are only interested in descriptive statistics, then the choice of X doesn't matter. The joint relationship between Y and and just about any X can be analyzed and reported.
  - Previous studies and experience often lead us to identify a set of explanatory variables that have proven useful in the past. It makes sense to include them and other variables you think useful in your study. However, if you want to predict, then it is important to use variables that are available to you at the time of prediction (there is no point next month's weather to predict next month's

soft-drink consumption); you also want to find a relationship that is “stable” so that it will hold out of sample. For example if you have a short sample and horizon, you may decide to include temperature alone as a predictor. With a longer sample and horizon, you may decide to include also relative prices, income, and population. Note, however, that stability considerations tend to favour smaller models.

- To estimate a causal relationship by OLS, you either need to be lucky or very smart. If you are lucky, you have experimental data (real or "natural"). Otherwise, to convince economists that you have uncovered a causal relationship, you need to have a good theory that explains the relationship. So a forecasting model that uses temperature and ignores relative prices and income isn’t going to be convincing (the forecasting model says that price doesn’t matter so you can charge as much as

you want without affecting demand–economists aren't going to believe this). At the very least, you'll probably have to consider as a possibility–but not necessarily your final and preferred choice–a larger model to convince readers that what you've estimated isn't just an artifact of leaving out some other important explanatory model. Part of the larger model could be terms allowing for heterogeneity in coefficients across difference possible subpopulations. Of course, anything you do will be necessary but not sufficient to convince readers of a stable causal relationship. If history is a guide, you'll have to wait until lots of other researchers (both theoretical and empirical) have taken a kick at the can and lots of new data are found that support your results.

3. Economics/Statistical Specification Issues

A. Functional Form

- Economics rarely has much to say about functional form (an exception is the CAPM which predicts that expected returns on risky assets are linear in the asset's "beta"). I include a brief discussion of functional form here because I need to talk about f(X) before I can talk about the substantive (as opposed to purely statistical) issues involved in the joint distribution of X and u. From statistics, there are two obvious candidates for f(X) that may be of interest. In both cases, assume Y and X have finite variances.
  - Suppose we are interested in predicting Y and restrict attention to linear functions of X. For $\beta \in R^K$, find $\beta^* = \arg\min E(Y - X\beta)^2$. If we write $Y = X\beta^* + u^*$,

then by construction we have $cov(u^*, X\lambda) = 0$, for all $\lambda \in R^K$. If our interest is in descriptive statistics, and we've decided to restrict attention to linear relationships between Y and X, then β^* is usually the coefficient vector of interest. It is called "the coefficient of the Best Linear Predictor (BLP) of Y given X", and it is under very general conditions the object that OLS tries to estimate. If we are interested in forecasting, then β^* will still be of interest, as long as the joint distribution of (Y_0, X_0) is the same as the joint distribution of (Y, X).
Note: If we are interested in causal relationships then we may be interested in another coefficient vector. The classic example is demand and supply. Suppose $Q^D = a + bp + u^D$, and $Q^S = c + dp + u^S$. If we regress Q on p, OLS will estimate the coefficient of the BLP, but $\beta \neq b$ and $\beta \neq d$!

- If we are interested in predicting Y and consider any (measurable and square-integrable) function $f(X)$, then it is natural to pay special attention to the one that solves $f^{\circ}(X) = \arg\min E(Y - f(X))^2$. The solution is $f^{\circ}(X) = E(Y|X)$. If we write $Y = f^{\circ}(X) + u^{\circ}$, then by construction $cov(u^{\circ}, f(X)) = 0$, for all (measurable, square-integrable) $f(X)$. The conditional expectation function (and analogues such as the conditional median function) is of natural interest if our purpose is descriptive statistics. With the usual caveat about stability, it is also a sensible objective to uncover if our interest is in forecasting. Again, it may or may not be of interest if we are after causal relationships.
- The distinction between linear and nonlinear models is not as sharp as may first appear. The model $f(X) = X\beta$ is *linear in parameters*. By suitably defining the

components of X to include functions (polynomials, logarithms, etc.) of some smaller set of underlying explanatory variables, we can do a pretty good job of approximating the conditional expectation function. In most applications, we try to exploit nonlinearities in this fashion. In the last two decades, there has been an explosion of interest in estimating nonlinear objects directly, rather than working with a finite-dimensional approximation. Unfortunately, we will have little to say about such nonparametric estimation methods in this course.

- Unless it involves an affine transformation (such as a change in units), transformations of the dependent variable matter because they determine the object of interest. Choosing to use wages or its logarithm as the dependent variable has the same character as choosing

wages or hours worked as the object to investigate! Are we interested in estimating derivatives or elasticities? They're not the same object! Because economics rarely provides any guidance on whether or not it is more interesting to explain, say, the level or the logarithm of the dependent variable, this choice is often made on statistical grounds. For some transformation of the dependent variable, such as the log or square root of wages, a linear relationship with X may be a better approximation to the conditional expectation, or the errors may have "better" statistical properties. It such a case, it makes sense to conduct the empirical analysis using such a transformation, but always translate results into objects (such as derivatives or elasticities at a point) that have a familiar economic interpretation.

B. First-order Statistical Properties

- First-order restrictions on the joint distribution of u and X identify the object of statistical inference. It's important to make sure that this is also the object of *economic* interest! In what follows, I assume that we are restricting attention to a linear relationship $Y = X\beta + u$. Think of identifying assumptions as "assertions"–they can't be tested. Overidentifying assumptions are testable.
  - If we assume $cov(X, u) = 0$, then we identify β as the coefficient of the BLP of Y given X to be the object of interest.

- If we assume $E(u|X) = 0$, then we implicitly identify the conditional expection of Y given X as the object of interest, AND we claim that it is linear (which is testable). A stronger, but easier to interpret assumption is that (u, X) are independent. Dependence between (u, X) could come from not measuring all the variables that affect Y, or because variations in u may affect X directly.
- If we assume the existence of some instrument Z such that $cov(Z, u) = 0$ and $Cov(Z, X)$ has full column rank, then we identify as the coefficient vector β of interest to be the one that generates residuals uncorrelated to Z (and we make the testable assumption that Z is systematically related to the regressors X). Such assumptions are commonly made by economists in attempting to uncover causal relationships from observational data.

C. Second-order statistical properties

- In most economic applications, second-order statistical properties are "nuisance" assumptions. They have no economic importance. They matter for efficiency and correct inference, but NOT for the interpretation of which β is the object of interest.
  - If we write $E(uu') = \sigma^2 I$, then we are making assumptions that will matter for our statistical analysis, i.e. how to estimate β efficiently and how to make correct inferences about β.
- In finance, the second-order moments are sometimes the focus of attention (eg. ARCH models). That means we should re-interpret the dependent variable Y as the volatility (variance) in sections A-C above.

4. Types of Data

A. Important for first-order statistical properties

- Experimental (controlled, randomized treatments, “natural”) vs. nonexperimental or observational data.
  - This dichotomy is very important for estimating causal effects. Experimental data allow for empirical discovery of “laws”. The experiment can be constructed to guarantee that our statistical assertions (u is independent of X) are true. Some argue that it is not possible to identify causal effects from observational data. All agree that it’s hard to do! But experimental data are not a panacea. Although more common than in the past, democratic societies place strong limits on the ability of researchers to “role the dice” in a way that affects

people's lives, so lots of questions cannot be investigated with experiments. Moreover, experiment findings have no basis for extrapolation to other situations. Theories provide a way to extrapolate findings and to use observational data to identify structural relations (those that are invariant to interventions in the variables X). Economists have developed a rich set of tools, based on instrumental variables to identify causal relationships using theory and observational data. But economic theories are controversial and often not well-enough developed to convince skeptics. Many econometric studies can make no claim beyond descriptive statistics.

B. Important for second-order statistical properties (joint distribution of $Y_t = X_t\beta + u_t, \quad t = 1..T$)

- Cross-sectional data: *t* indexes households, firms, individuals, etc.
  - With *random sampling*, then u_s is independent of u_t for $s \neq t$. The order of the observations doesn't matter (we could shuffle the data and not lose any information). This leads to the most straightforward set of econometric issues and is the focus of most of the Wooldridge book. Often *T* is very large, so standard large-sample theory provides a useful framework for approximate inference. This is the sort of data often assumed to be encountered in applied micro (labour, public, health, etc.).

- Time-series data: *t* indexes time.
  - In time-series applications, the data may be measured at annual, quarterly, monthly, weekly, or daily frequency. In financial applications, the data may even be measured "tick-by-tick". Special econometric problems arise in time-series applications. Dynamics may enter the regression–*X* may contain lags of the dependent variable as well as lags of explanatory variables. Alternatively, dynamics may be introduced through a rich correlation structure for the errors. Note that including lags of *Y* in *X* means that we have to be careful about what sort of conditioning we use in defining *u*.

C. Important for both first and second-order statistical properties

- Pooled cross-sections
  - A collection of cross-sections for different years, if each year is collected from a random sample, allows one to investigate some dynamics and to consider first-order models that cannot be identified from a single cross-section. For example, are the coefficents stable over time or do they display a noticeable trend?
- Panel data
  - A collection of time-series for each micro observation will have dependencies that must be addressed for proper inference. More important, it allows us to use the same agent in a different time period as a "control" in calculating the effect of a change in X.

D. Special features of the support of the data

- As a first approximation, special features of the explanatory data X introduce no special statistical issues, as long as we believe that the parameter β is truly a constant for all observations. In practice, however, we notice that estimates of β vary from sample to sample. In trying to reconcile this variation (or in thinking about whether to use your estimated β for an out-of-sample forecast or policy prediction), it is useful to think of each individual observation as having it's own coefficient β_t. Then OLS will estimate some sort of 'average' of the β_t in your sample. If your data only contain observations with small values for X, then this average may be quite different than the β_t relevant if X is large (for example, the effect of increasing the legal minimum age to leave school may be very different for weak students than for strong students). Think about the source of variation in your

sample X. Is it across people? Is it across time? Is it for the same person across time? Each of these sources of variation in X may pick up a different weighted average of β_t as the object of interest. With one sample, there is not much you can do about this, but you should be aware of it, especially in interpreting variations across samples and in thinking about using your results out of sample.

- Features of the support of Y matter a good deal in building an appropriate statistical model, and point to different econometric strategies.
  - Multivariate (vs univariate) data refers to the situation where the basic observation of interest is a vector rather than a scalar (eg. expenditure shares on food, clothing, rent, transportation, etc.)
  - Indicator or dummy variables take on only the values 0 or 1. They are used to indicate the absence or presence

of an attribute (eg. employed, own a car, university degree, etc.)

- Count data refers to the situation where the variable can take on only non-negative integer values (eg., number of assets owned, or number of children).
- Truncated data refers to the case where the support of the data is a strict subinterval of the real line (eg. wages must be strictly positive). Usually, truncated also is used to mean that some data has been suppressed (eg. we don't observe wages of non-workers).
- Censored data refers to the case where the support of the data is a strict subinterval of the real line, but the density has a "spike" at one or more endpoints. For example, we measure $Y = Y^*$ if $Y^* > 0$, and $Y = 0$ if $Y^* \leq 0$. An application might be expenditure on cigarettes (a large fraction of the sample will report zero expenditures).

- Limited dependent variable refers to any situation where the support of Y is not the real line (including the four examples above). Unfortunately, we will not be covering limited dependent variables in this course.