ECON3150/4150 INTRODUCTORY ECONOMETRICS

Lecture note no. 1

Erik Biørn,
Department of Economics

Version of January 14, 2013

ON MODELS AND DATA TYPES IN ECONOMETRICS

REMARK: *In the following we use some concepts and approaches that may – initially – be unknown, or strange, to you. They will be explained later in the course. A good advice therefore is to try to grasp the main message now and postpone reading of the details until later.*

A statistical model to be used in analyzing an economic relationship, should be *in agreement with the data situation at hand.* This may sound a rather trivial statement, but is related to a very fundamental matter when constructing an econometric model. We should know what kind of data we have access to before we could formulate the model. In this note – intended as a part of the general introduction to the course – we shall take a closer look at the relation between model type and data type.

Two basic types of Economic Data

The two most important types of data an econometrician, or an economist performing correlation studies, is occupied with are:

Cross-Section Data
Time Series Data

Cross-Section Data (Tverrsnittsdata): These are data from units observed at the same time or in the same time period. The data may be single observations from a sample survey or from all units in a population. Examples of Norwegian cross-section data are the Household Budget Survey for the year 1999, The Manufacturing Statistics for the year 2000, the Population Census for the year 2001.

Time-Series Data (Tidsseriedata): These are data from a unit (or a group of units) observed in several successive periods. Examples of Norwegian time-series data are National Accounts data (production, private and public consumption, investment, export, import etc.), the Index of Manufacturing Production, the Consumer Price Index and Financial statistics (money stock, exchange rates, interest rates, bank deposits, etc.)

Most often *cross-section data* are data for *micro units* – individuals, households, firms, companies, etc. But macro-like cross-section data may well occur; examples are cross-section data for municipalities, other local units, counties or or even countries. In cross-section data, *all data variation goes across units*; we have variation across space (spatial variation).

Most often *time-series data* are *macro data or macro-type data*, for example time-series for macro-economic variables from the National Accounts. But micro-data may also occur as time-series, for example time-series for a particular household or time-series for a particular firm. In time-series data *the data variation goes over time periods*; we have variation over time (time serial variation).

Cross-Section data show spatial variation:
Variation across units (individuals, households, firms, ...)

Time-Series data show temporal variation:
Variation over periods (years, months, weeks, seconds, ...)

Example: Demand Function for a consumption commodity

Assume that our theory postulate that the expenditure on a consumption commodity (y) depends on the consumer's income (x), the price of the commodity (p), the number households member (z) and a disturbance (u), which captures unspecified explanatory variables etc. in the following way:

$$y = a + bx + cp + dz + u. \tag{1}$$

This is our general theory. *It is not accommodated to a specific data situation.* We want to estimate the intercept (constant term) a and the coefficients b, c and d.

Let now i indicate the number of the household and t the number of the observation. *If we had had observations for all households in a successive number of years*, we would have designed the model description as follows:

$$y_{it} = a + bx_{it} + cp_t + dz_i + u_{it}, \tag{2}$$

specified the range of the subscripts i and t and a suitable set of assumptions for the disturbances. In formulating (2) we have assumed that all households in each period have been confronted with the same commodity price, and that the number of persons in each specific household has not changed from year to year.

We now consider specializations of (2) to three data types.

Specialization I: Cross-Section Data: Assume that we have cross-section data for one single year, year $t = 1$, from a sample of M households drawn randomly from a population, say a population containing all Norwegian households. Equation (2) translated to this data situation then becomes

$$y_{i1} = (a + cp_1) + bx_{i1} + dz_i + u_{i1}, \qquad i = 1, \ldots, M. \tag{3}$$

Since the price does not vary across the data set, we can combine the price term cp_1 with the genuine intercept a and interpret $a + cp_1$ as a *cross-section intercept for year 1.* The variables in the cross-section data set therefore become y, x, z, with observation set $\{y_{i1}, x_{i1}, z_i\}_{i=1}^{i=N}$. The disturbance u_{i1} varies across households.

Equation (3) shows the following: *It is impossible to estimate the price coefficient c from the cross-section data. This is because the price only varies 'along the time dimension', not 'along the cross-sectional dimension'.* What we are able to estimate – for example by applying the Ordinary Least Squares (OLS) method on (3), provided that the u_{i1}'s $(i = 1, \ldots, N)$ satisfy classical assumptions – are b, d and the composite intercept $(a + cp_1)$. Even if we know p_1, this is not sufficient to derive an estimate for c. We say that we are unable to *identify* c in equation (1) from our data. This also has a positive side: *In cross-section data we do not need to be concerned with, or bothered with, correlation between the income x and the price p.*

Specialization II: Micro Time Series Data: Assume that we have time-series data for one single household, household $i = 1$, for T successive years. Equation (2) translated to this data situation then becomes

$$y_{1t} = (a + dz_1) + bx_{1t} + cp_t + +u_{1t}, \qquad t = 1, \ldots, T, \tag{4}$$

Since the number of household members does not show any variation across the data set, we can combine the household size term dz_1 with the genuine intercept a and interpret $a + dz_1$ as a *time-series intercept for household 1.* The variables in the time-series data set therefore become y, x, p, with observation set $\{y_{1t}, x_{1t}, p_t\}_{t=1}^{t=T}$. The disturbance u_{1t} varies over years.

Equation (4) shows the following: *It is impossible to estimate the household size coefficient d from the time-series data. This is because the number of households members only varies 'along the cross-sectional dimension', not 'along the time dimension'.* What we are able to estimate – for example by applying the OLS method on (4), provided that the u_{1t}'s $(t = 1, \ldots, T)$ satisfy classical assumptions – are b, c and the composite intercept $(a + dz_1)$. Even if we know z_1, this is not sufficient to derive an estimate for d. We say that we are unable to *identify* d in equation (1) from our data. This also has a positive side: *In time-series data we do not need to be concerned with, or bothered with, correlation between the income x and the household size z.*

Altogether Equations (3) and (4) show:

1) Estimation of the price coefficient c from cross-section data is impossible, because the price varies only over time.

2) Estimation of the household size coefficient d from time-series data is impossible, because the household size varies only across households.

Specialization III: Aggregate Time-Series Data: Next, assume that we have time-series data for T years, aggregated across all households in the population, say all Norwegian households. Let the number of households in year t be N_t. How should we translate or accommodate (2) to this data situation? With aggregation we here understand simple summation across households. We sum on both sides of the equality sign across i and get

$$\sum_{i=1}^{N_t} y_{it} = aN_t + b\sum_{i=1}^{N_t} x_{it} + cN_t p_t + d\sum_{i=1}^{N_t} z_i + \sum_{i=1}^{N_t} u_{it},$$

or

$$Y_t = aN_t + bX_t + cN_t p_t + dZ_t + U_t, \tag{5}$$

using the following symbols for the aggregate (y, x, z, u)-variables:

$$Y_t = \sum_{i=1}^{N_t} y_{it}, \quad X_t = \sum_{i=1}^{N_t} x_{it}, \quad Z_t = \sum_{i=1}^{N_t} z_i, \quad U_t = \sum_{i=1}^{N_t} u_{it}. \tag{6}$$

In this situation it will, in principle, be variables attached to all coefficients in (5). The variable attached to the original intercept a is the number of households N_t, the variable attached to the price coefficient is the product of the number of households and the commodity price, while the variable attached to the household size coefficient, $\sum_{i=1}^{N_t} z_i$, is the number of individuals in the population. The disturbance U_t has t-dependent variance if N_t varies over t and the micro disturbance u_{it} has constant variance: If u_{it} has constant variance σ^2 and is uncorrelated across individuals and over time, then $\text{var}(U_t) = N_t\sigma^2$. Multicollinearity problems may, however, easily arise by using (5) since the number of individuals and the number of households often vary closely.

Let us therefore, for simplicity, consider the case where *the number of households is constant* over the T years (or that this holds as a good approximation), i.e. $N_t = N$ for all t. It then follows from (5) and (6) in particular that

$$Y_t = A + bX_t + Cp_t + U_t, \tag{7}$$

where

$$Y_t = \sum_{i=1}^{N} y_{it}, \quad X_t = \sum_{i=1}^{N} x_{it}, \quad U_t = \sum_{i=1}^{N} u_{it}$$

are variables and

$$A = aN + dZ, \qquad C = cN, \qquad Z = \sum_{i=1}^{N} z_i$$

are coefficients. From (7) we can estimate the income coefficient b, the macro price coefficient C and the composite macro intercept A.

We then are in a similar situation as when using the micro time-series relation (4). It is impossible to estimate the household size coefficient d from the time-series data. This is because the number of household members only varies along the cross-sectional dimension, not along the time dimension, also in the aggregate data set. Even if we know Z, this is not sufficient for deriving an estimate of d. On the other hand, we do not need to be concerned with, or bothered with, correlation between the income X and the population size $\sum z_i$ in the macro data set.

Panel Data: A third important data type

We may also imagine that we have a data set consisting of time-series data for several observation units, for example consumption data from $M\,(M \geq 2)$ households observed over $T\,(T \geq 2)$ years. With this specialization the model takes the form

$$y_{it} = a + bx_{it} + cp_t + dz_i + u_{it}, \qquad i = 1, \ldots, M;\ t = 1, \ldots, T. \tag{8}$$

Such a data set, with MT observations, is called a panel data set, because we observe a 'panel' of M households over T years. Alternative terms are combined time-series/cross-section data or longitudinal data. The variables in a panel data set can vary both across the spatial dimension and over and time dimension. But some of them may vary along one dimension only, as z and p in our basic example.

Panel data show both spatial and temporal variation.

This sets us in *a position to estimate a, b, c and d simultaneously.* This is an important difference from pure cross-section data and pure time-series data, from which we are unable to estimate all coefficients simultaneously.

Panel data have, over the years, become a gradually more important and more frequently used data type for analyzing economic relationships. This has several explanations: (i) Panel data is a 'richer' data type than (pure) cross-section data and (pure) time-series data. (ii) The development of the data collection and data processing methods. (iii) The development in computer technology.

Using panel data, which exhibit both spatial and temporal variation, we are able to estimate a, b, c and d jointly.

Panel data set may well be large. For example, $M = 5000$ households observed over $T = 20$ years give a data set with $MT = 100\,000$ observations. Handling so large bodies of data poses strong requirements on computer technology and computer software, but is well within the reach for being handled by modern computers, even lap-tops.

Final remark

Attempts to estimate the same economic coefficient (i) from cross-section data, e.g., the income coefficient b in (3), and (ii) from time-series data, e.g., the income coefficient b in (4), often give 'systematically different' results. Possible explanations of this have been much discussed. Panel data may set us in a position to study such differences mode closely. Biases in the estimation of b and d from cross-section data may reflect omitted (and often unobservable) consumption motivating variables that are correlated with x_{i1} and z_i *across the cross-section*, say tastes and preferences. Biases in the estimation of b and c from time-series data may reflect omitted (and often unobservable) consumption motivating variables that are correlated with x_{1t} and p_t *over time*, say the consumers' moods and expectations about the future business-cycle conditions. *The Gross/Net Coefficient-problem (a concept to be discussed later on) may therefore enter the scene differently and have different consequences in the two data types.* Panel data 'let loose' the variation in the x's, the z's and the p's at the same time. But panel data also set the researcher in a position to examine both (i) correlation over i in each period, for $t = 1, \ldots, T$ separately and (ii) correlation over t for each individual, for $i = 1, \ldots, N$ separately. This may help him or her approach explanations of discrepancies between cross-sectional based and time-serial based estimates of presumptively the same parameter. Panel data may also help to form 'purer' estimators than those obtainable from the two simple data types. This often requires the use of specific estimation methods, a topic studied in more advanced econometrics.

What has been said above underpins, inter alia, the following conclusion: When discussing correlation between economic variables in relation to an econometric investigation, it is *important to be precise about what the correlation goes across*. This enters as an important characteristic of the data type used in the investigation. Correlation between income and wealth across a cross-section has a different meaning than correlation between income and wealth over time, and such correlation coefficients often turn out to have markedly different size. The nature of multicollinearity problems (also a concept to be discussed later on) when using cross-section data and when using time-series data may therefore become widely different.

Supplementary readings:

Erik Biørn: *Økonometriske emner. En videreføring.* Oslo: Unipub 2008. Kapittel 1: Datatyper and modeltyper.

Zvi Griliches: *Handbook of Econometrics, Vol. III.* Amsterdam: North-Holland, 1986. Chapter 25: Economic Data Issues, sections 1, 2, 3.