CHAPTER 13

A Summary View

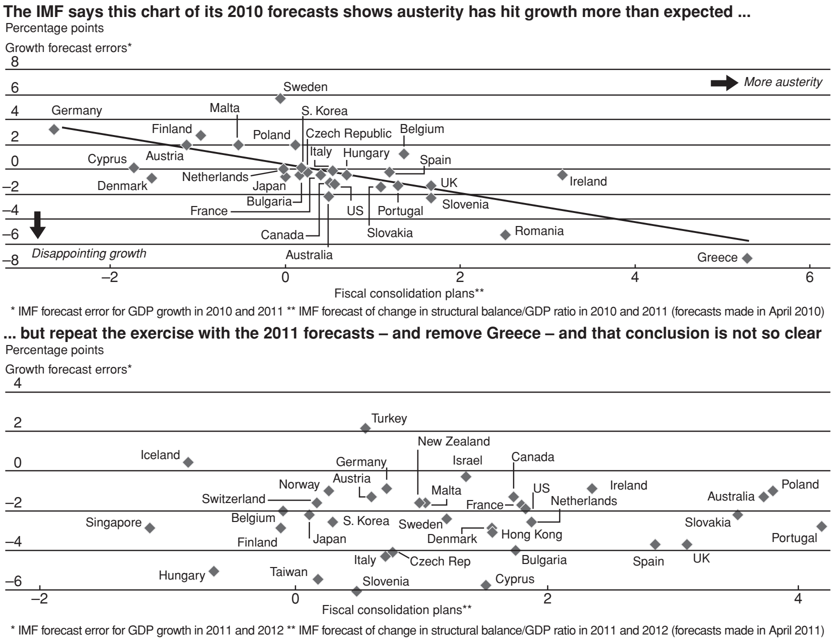
13.1 BENEFITS OF PANEL DATA

As discussed in Chapter 1, panel data provides major benefits for econometric estimation in at least six areas: (1) increasing degrees of freedom and reducing problems of data multicollinearity, (2) constructing more realistic behavioral models and discriminating between competing economic hypotheses, (3) eliminating or reducing estimation bias, (4) obtaining more precise estimates of micro relations and generating more accurate micro predictions, (5) providing information on appropriate level of aggregation, and (6) simplifying cross sections or time series data inferential procedures. In this section we provide a summary view on how different methods discussed in this monograph can be used to achieve these benefits.

13.1.1 Increasing Degrees of Freedom and Lessening the Problem of Multicollinearity

In empirical studies investigators often encounter problems of shortage of degrees of freedom and multicollinearity. That is, the information provided by the sample is not rich enough to meet the requirement of the specified model. To narrow this gap, investigators either often have to impose ad hoc prior restrictions (e.g., Hsiao, Mountain, and Ho-Illman 1995) or to augment sample information. Panel data have many more degrees of freedom than cross-sectional or time series data. Moreover, panel data containing information on both interindividual differences across cross-sectional units and intraindividual dynamics over time can substantially increase the sample information. Pooling procedures to obtain more accurate estimation of common parameters for linear static and dynamic models are discussed in Chapters 3, 4, 9, 11 (Section 11.4), and 12 (Section 12.3); static and dynamic system of equations are in Chapters 5 and 10; nonlinear models are in Chapters 7, 8, and 12 (Sections 12.1 and 12.2). Pooling for heterogeneous individuals is discussed in Chapter 6.

https://doi.org/10.1017/CBO9781139839327.016 Published online by Cambridge University Press



The IMF says this chart of its 2010 forecasts shows austerity has hit growth more than expected ...

* IMF forecast error for GDP growth in 2010 and 2011 ** IMF forecast of change in structural balance/GDP ratio in 2010 and 2011 (forecasts made in April 2010)

... but repeat the exercise with the 2011 forecasts – and remove Greece – and that conclusion is not so clear

* IMF forecast error for GDP growth in 2011 and 2012 ** IMF forecast of change in structural balance/GDP ratio in 2011 and 2012 (forecasts made in April 2011)

Figure 13.1. Robustness of IMF data scrutinized. * *Source:* The top figure is from IMF World Economic Outlook, Oct. 2012. The bottom figure is from Financial Times, Oct 13/14, 2012.

13.1.2 Identification and Discrimination between Competing Hypotheses

Aggregate time series data are not particularly useful for discriminating between hypotheses that depend on individual attributes. A single individual time series data set is not possible to provide information on the effects of different socio-demographic factors. Cross-sectional data, though containing information on microeconomic and demographic variables, cannot take account the (unobserved) heterogeneity across individuals. A fundamental assumption inherent in studies using cross-sectional data is that $E(y_i \mid \mathbf{x} = \mathbf{x}^*) = E(y_j \mid \mathbf{x} = \mathbf{x}^*)$. This homogeneity (conditional on $\mathbf{x}$) assumption can lead to grossly misleading or sensitive inference on the impact of $\mathbf{x}$ on y. For instance, with many economics in fiscal consolidation model since financial crises broke out in 2008, a debate has been raging about the size of fiscal multipliers. The smaller the multipliers, the less costly the fiscal consolidation. Under rational expectations and if the correct forecast model has been used, there should be no relation between the forecast error for real GDP growth and planned fiscal consolidation (measured as changes in the structural fiscal balance (due to tax rises and spending cuts) as a percentage of potential GDP). The top figure of Figure 13.1 shows the International Monetary Fund (IMF) estimate of the effect of austerity plans

https://doi.org/10.1017/CBO9781139839327.016 Published online by Cambridge University Press

on the 2010–11 forecast error of the real GDP growth rate (World Economic Outlook IMF, October 9, 2012). The figures shows a large, negative relation. The baseline estimate suggests that a planned consolidation of 1 percent of GDP is associated with a growth forecast error of about 1 percentage point. Based on this analysis, IMF concludes that the assumed multipliers of about 0.5 underlying the forecast models have been too low by about 1. The short-term fiscal multipliers should be in the range of 0.9 to 1.7. The bottom figure shows the Financial Times (October 13/14, 2012) estimates after removing Greece and Germany. The relationship between deficit reduction efforts and forecast error of the growth rate is simply not there. One reason that these results are so sensitive to the includion and exclusion of certain countries is because the cross-sectional analysis cannot take account the effects of (unobserved) country-specific factors.

In economics, as in other branches of the social and behavioral sciences, often there are competing theories. Examples of these include the effect of collective bargaining on wages, the appropriate short-term policy to alleviate unemployment (Chapters 1 and 7), the effects of schooling on earnings (Chapter 5), and the question of causal ordering such as "Does delinquency lead to low self-esteem or does low self-esteem lead to delinquency?" (e.g., Jang and Thornberry 1998). Economists on opposite sides of these issues generally have very different views on the operation of the economy and the influence of institutions on economic performance. Some economists believe unions indeed raise wages or that advertising truly generates greater sales. Adherents of the opposite view tend to regard the effects more as epiphenomena than as substantive forces and believe that observed differences are due mainly to sorting of workers or firms by characteristics (e.g., Allison 2000).

Proper recognition of the sources of variation can provide very useful information for discriminating individual behavior from average behavior or for identifying an otherwise unidentified model. For instance, in the foregoing collective bargaining example, even if information on worker quality is not available, if a worker's ability stays constant or changes only slowly, the within correlation between the union-status dummy and the worker-quality variable is likely to be negligible. Thus, the impact of worker quality can be controlled through the use of within estimates (Chapter 3). The resulting coefficient for the union-status dummy then will provide a measure of the effect of unionism. In the income schooling model, the availability of family groupings can provide an additional set of cross-sibling covariances via a set of common omitted variables. These additional restrictions can be combined with the conventional slope restrictions to identify what would otherwise be unidentified structure parameters (Chapter 5, Section 5.4).

Panel data providing sequential observations for a number of individuals allow an investigator to distinguish interindividual differences from intraindividual differences and construct a proper causal structure (Chapters 5, 9 [Sections 9.3, 9.4], and 10). Furthermore, addition of the cross-sectional dimension to the time series dimension provides a distinct possibility to identify the pattern of serial correlations in the residuals or to identify the lag adjustment

https://doi.org/10.1017/CBO9781139839327.016 Published online by Cambridge University Press

patterns when the conditioning variables are changed without having to resort to imposing prior parametric restrictions (Chapters 3 [Section 3.8] and 11 [Section 11.4]) or to identify a model subject to measurement errors (Chapter 12, Section 12.6).

13.1.3 Reducing Estimation Bias

A fundamental statistical problem facing every econometrician is the *specification problem*. By that we mean the selection of variables to be included in a behavioral relationship as well as the manner in which these variables are related to the variables that affect the outcome but appear in the equation only through the error term. Empirical findings are often criticized on the grounds that the researcher has not explicitly recognized the effects of omitted variables that are correlated with the included explanatory variables (in the union example, the omitted variable, worker quality, can be correlated with the included variable, union status). If the effects of the omitted variables are correlated with the included explanatory variables, and if these correlations are not explicitly allowed for, the resulting regression estimates could be seriously biased (Chapter 3, Sections 3.4 and 3.5). To minimize the bias, it is helpful to distinguish four types of correlations between the included variables and the error term. The first type is due to the correlation between the included exogenous variables and those variables that should be included in the equation but are not, either because of a specification error or because of unavailability of data (Chapters 3, 7, and 8). The second type is due to the dynamic structure of the model and the persistence of the shocks that give rise to the correlation between lagged dependent variables and the error term (Chapters 4 and 10). The third type is due to the simultaneity of the model, which gives rise to the correlation between the jointly dependent variables and the error terms (Chapters 5 and 10 [Section 10.4]). The fourth type is due to measurement errors in the explanatory variables (Chapter 12, Section 12.6). Knowing the different sources of correlations provides important information for devising consistent estimators. It also helps one avoid the possibility of eliminating one source of bias while aggravating another (e.g., Chapter 5, Section 5.1).

Panel data can help identify these four sources of correlations. For instance, if the effects of these omitted variables stay constant for a given individual through time or are the same for all individuals in a given time period and the model is linear (e.g., Chapters 3 and 4), the omitted-variable bias can be eliminated by one of the following three methods when panel data are available: (1) differencing the sample observations to eliminate the individual-specific and/or time-specific effects, (2) using dummy variables to capture the effects of individual invariant and/or time-invariant variables; and (3) postulating a conditional distribution of unobserved effects given observed exogenous variables, then integrating out the unobserved effects to make inferences based on the marginal distribution of observables. The first two approaches are commonly referred as the *fixed-effects inference* and the third approach is referred as the *random-effects inference*.

https://doi.org/10.1017/CBO9781139839327.016 Published online by Cambridge University Press

Panel data can also help to identify the correlations between the regressors and errors that are due to simultaneity or to the correlations between the unobserved individual- or time-specific effects and the regressors. The standard approach to eliminate simultaneity bias is to use instrumental variables to purge the correlations between the joint dependent variables and the error of the equation. However, if there exist correlations between the regressors and the unobserved individual- or time-specific effects, what are generally considered as valid instruments may not be valid any more (Chapter 5, Sections 5.3 and 5.4.)

Measurement errors in the explanatory variables create correlations between the regressors and the errors of the equation. If variables are subject to measurement errors, the common practice of differencing out individual effects eliminates one source of bias but may aggravate the bias due to measurement errors. However, different transformation of the data can induce different and deducible changes in the estimated regression parameters, which can be used to determine the importance of measurement errors and obtain consistent estimators of parameters of interest (Chapter 12, Section 12.6).

13.1.4 Generating More Accurate Predictions for Individual Outcomes

If individual behaviors are similar conditional on certain variables, panel data provide the possibility of learning an individual's behavior by observing the behavior of others. Thus, it is possible to obtain a more accurate description of an individual's behavior by supplementing observations of the individual in question with data on other individuals (e.g., Chapter 6).

13.1.5 Providing Information on Appropriate Level of Aggregation

A model is a simplification of reality, not a slavish reproduction of all real-world data. The real-world detail is reduced through aggregation of "homogeneous" units or through the "representative agent" assumption. However, if micro units are heterogeneous, not only can the time series properties of aggregate data be very different from those of disaggregate data (e.g., Granger 1980; Lewbel 1992, 1994; Stoker 1993), policy evaluation based on aggregate data can be grossly misleading. Furthermore, the prediction of aggregate outcomes using aggregate data can be less accurate than the prediction based on microequations (e.g., Chapter 6, Section 6.8.2 or Hsiao, Shen, and Fujiki 2005). Panel data containing time series observations for a number of individuals is ideal for investigating the "homogeneity" versus "heterogeneity" issue. Moreover, when "homogeneity" in panel is rejected, the variable coefficient models discussed in Chapter 6 provides a feasible alternative to make inferences about the population while taking account of the heterogeneity among micro units.

https://doi.org/10.1017/CBO9781139839327.016 Published online by Cambridge University Press

13.1.6 Simplifying Computation and Statistical Inference

Panel data involve at least two dimensions, a cross-sectional dimension and a time series dimension. Under normal circumstances the computation of panel data estimator or inference would be more complicated than cross-sectional or time series data. However, on many occasions, the availability of panel data actually simplifies computation and inference. For instance, in the analysis of time series properties of a variable, first one will need large number of time series observations to properly distinguish stationary time series from nonstationary time series. Second, when time series data are not stationary, the large sample approximation of the distributions of the least-squares or maximum likelihood estimators are no longer normally distributed (e.g., Anderson 1959; Dickey and Fuller 1979, 1981; Phillips and Durlauf 1986). But if panel data are available, one can invoke the central limit theorem across cross-sectional units to show that the limiting distributions of many estimators remain asymptotically normally distributed. Moreover, even only a small number of time series observations are available, an investigator making use of information on cross-sectional dimension may be able to distinguish unit roots or cointegration processes from stationary process (Chapter 10).

Another example is in the evaluation of the impact of social program. When only cross-sectional data are available, the control of the impact of *selection on observables or unobservables* could be complicated (Chapter 9, Section 9.6.2). However, if panel data are available and if individual units are cross-sectionally dependent, then one can use cross-sectional units information to construct the counterfactuals for the evaluation of the impact of social program without the need to worry about the issues of selection on observables or unobservables which may considerably simplify the analysis (Chapter 9, Sections 9.6.3 and 9.6.4).

13.2 CHALLENGES FOR PANEL DATA ANALYSIS

Although panel data offer many advantages over a cross-sectional or time series data set, there are many interesting and unresolved issues remain such as (1) how best to model unobserved heterogeneity across individuals and/or over time; (2) controlling the impact of unobserved heterogeneity to obtain valid inference for nonlinear models; (3) modeling cross-sectional dependence; (4) multidimensional asymptotics; and (5) sample attrition, etc.

13.2.1 Modeling Unobserved Heterogeneity

As discussed in the introduction (Chapter 1, Section 1.3), panel data focus on individual outcomes over time. Factors affecting individual outcomes could be numerous. One of the most challenging issues in panel data modeling is how to model the unobserved heterogeneity across individuals and over time that are not captured by the conditional variables $\mathbf{x}$. This monograph essentially follows

https://doi.org/10.1017/CBO9781139839327.016 Published online by Cambridge University Press

the approach of letting part of the parameters characterizing the conditional distribution of y_{it} given $\mathbf{x}_{it}$ to vary across i and over t, $f(y_{it} \mid \mathbf{x}_{it}, \boldsymbol{\beta}, \boldsymbol{\gamma}_{it})$, where $\boldsymbol{\beta}$ is assumed identical over i and t and $\boldsymbol{\gamma}_{it}$ vary across i and over t. To control the impact of $\boldsymbol{\gamma}_{it}$ on the inference of $\boldsymbol{\beta}$, $\boldsymbol{\gamma}_{it}$ is further decomposed into components that are individual-specific, $\boldsymbol{\alpha}_i$, and component that are time-specific, $\boldsymbol{\lambda}_t$. Is this the best way to model unobserved heterogeneity? When time series dimension or cross-sectional dimension becomes large, is the assumption of time-invariance in $\boldsymbol{\alpha}_i$ or individual-invariance of $\boldsymbol{\lambda}_t$ still reasonable? Further, the function of a variable could have very different meanings at different time. For instance, Friedman (1969) found that there was a stable relation between the $M2$ and nominal GDP in the 1960s. However, with technological development there are many financial instruments today that can also perform the function of currency and demand deposits. Do today's M2 still have the same economic implication as M2 in the 1960s or should these close substitutes be also included in the analysis of money?

There is also an issue of whether to treat the unobserved heterogeneity as fixed and different (fixed effects) or as random draws from a common (conditional) distribution (random effects). In general, if the unobserved heterogeneity can be viewed as random draws from a common population, then it is more appropriate to postulate a random-effects model. If the unobserved heterogeneity is correlated with explanatory variables or comes from heterogeneous population, then it is more appropriate to postulate a fixed-effects model unless the interaction between observables and unobservables are known to investigators. The fixed-effects formulation makes inference conditional on the specific effects; hence it has the advantage of not requiring one to postulate the distribution of the effects. However, there is also a loss of efficiency in conditional inference because of the loss of degrees of freedom in estimating the specific effects. It may even introduce incidental parameters problem if the dimension of the effects increase at the same rate as sample size (Chapters 4, 7, and 8). The advantages of a random effects specification is that the probability function of the effects in general depends only on a finite number of parameters and there is no incidental-parameter problem, and efficient inference is possible. The disadvantage is that it requires explicit knowledge about the way in which observables and unobservables interact. In general, the advantages of fixed effects formulation are the disadvantages of random effects formulation and the disadvantages of the fixed effects formulation are the advantages of the random effects formulations (e.g., see the discussions in Chapters 3 and 4). Unfortunately, without explicit knowledge about the way in which observables and unobservables interact it is hard to decide which approach to adopt.

13.2.2 Controlling the Impact of Unobserved Heterogeneity in Nonlinear Models

There is a very fundamental difference between the linear and nonlinear models. If a model is linear, one can condition the effects on the observables and apply a minimum distance type estimator. If the model is nonlinear, the assumptions

https://doi.org/10.1017/CBO9781139839327.016 Published online by Cambridge University Press

for the conditional distribution of the effects need to be very specific. However, the effects are unobservable. It is hard to specify the conditional distribution of the effects without explicit assumptions about how the observables and unobservables interact. Moreover, the derivation of random-effect estimator often would involve multidimensional integration which can be very complicated even with today's computing capacity.

If the effects are treated as fixed, and if the number of unknown specific effects increases at the same rate as the sample size, attempts to estimate the specific effects creates the incidental-parameter problem. For general nonlinear models, there does not exist a generally applicable framework to implement the Neyman–Scott (1948) principle of separating the estimation of the common coefficients from the estimation of the specific effects. To devise consistent estimators of the structural parameters, one has to exploit the specific structure of a nonlinear model. The three most commonly used approaches are: (1) the conditional approach that conditions on the minimum sufficient statistics of the effects, (2) the semiparametric approach that exploits the latent linear structure of a model (Chapters 7 and 8), and (3) reparameterization of the model so that the information matrix of the reparameterized individual effects are uncorrelated with the reparameterized structural parameters (Lancaster 2001). The first two approaches apply classical sampling inference to a model that no longer involve incidental parameters. The transformation of a model containing incidental parameters to a model without incidental parameters is obtained through exploiting the specific structure of the original model. The third approach is from a Bayesian perspective. It can be shown that when the information matrix of the structural (or common) parameters are orthogonal to the incidental (or individual-specific) parameters, taking a uniform prior for the incidental parameter reduces the bias (Arellano and Bonhomme 2009). However, for most nonlinear models there does not appear that simple transformations to achieve information orthogonality exist. Whether any of these approaches will yield consistent estimators has to be considered case by case. Moreover, even in the case that consistent estimators exist, the conditions imposed on the data are so restrictive that hardly any data set can meet them (e.g., Chapter 7, Section 7.5).[1]

13.2.3 Modeling Cross-Sectional Dependence

If panel data are not conditional independent across cross-sectional units, ignoring cross-sectional dependence can lead to misleading inference. Contrary to time series observations there is no natural ordering of cross-sectional units. Chapter 9 surveyed some of the popular approaches that have been tried econometrically. Each approach has its merits and also limitations. In particular, in

[1] For instance, in the dynamic logit model considered in Chapter 7, Section 7.5, the conditions for the existence of consistent estimator requires at least (1) four times series observations for each individual; (2) indivdiuals switch position during the two intermediate periods; and (3) the value of the exogenous variable has to be equal in period 3 and period 4.

https://doi.org/10.1017/CBO9781139839327.016 Published online by Cambridge University Press

the case when N is large and T is small or the model is nonlinear, methods to take account cross-sectional dependence remain to be developed.

13.2.4 Multidimensional Asymptotics

This monograph focuses on panels that contain a cross-sectional dimension (N) and a time-series dimension (T). The majority of the discussions are on the case that there are a few observations in one dimension (usually the time dimension) and a great many observations in another dimension (usually the cross-sectional dimension), but there are panels where N and T are of similar magnitude. It is important to understand the properties of inferential procedures when a panel with only one dimension observations that are large or a panel that both or multidimensional observations are large, say both N and $T \to \infty$, and the relative speed of their increase. On the basis of this information, one can then determine which parameters can, and which parameters cannot, be consistently estimated from a given panel or where the asymptotic bias comes from. For instance, in a linear dynamic model with the error composed of the sum of two components, one being individually time-invariant and the other being independently distributed, then the individual time-invariant effects can be eliminated by differencing successive observations of an individual. We can then use lagged dependent variable (of sufficiently high order) as instruments for the transformed model to circumvent the issues of the serial dependence of the residual (Chapter 4, Sections 4.3 and 4.5). When T is fixed and N is large, the resulting estimator is consistent and asymptotically normally distributed. However, when T increases with N and $\frac{T}{N} \to c \neq 0$ as $N \to \infty$, although the resulting estimator is consistent, there is an asymptotic bias term when the estimator is multiplied by the scale factor, $\sqrt{NT}$, that needs to be corrected to obtain asymptotic valid inference (e.g., Chapters 4 and 10, Appendix 4B or Alvarez and Arellano 2003; Phillips and Moon 1999).

Computing speed and storage capability have enabled researchers to collect, store and analyze data sets of very high dimensions. Multidimensional panel will become more available. Classical asymptotic theorems under the assumption that the dimension of data is fixed (e.g., Anderson (1985)) appear to be inadequate to analyze issues arising from finite sample of very high dimensional data (e.g., Bai and Silverstein 2004). For example, Bai and Saranadasa (1996) proved that when testing the difference of means of two high-dimensional populations, Dempsters (1959) nonexact test is more powerful than Hotellings (1931) T^2-test even though the latter is well defined. Many interesting and important issues remain to be worked out. Statistic theorems providing insight to finite sample issues for high dimensional data analysis can be very useful to economists and/or social scientists (e.g., Bai and Silverstein 2006).

13.2.5 Sample Attrition

Panel data follows a number of individuals over time. As Table 1.1 shows, as time goes on, a number of individuals drop out. If sample attrition is random,

https://doi.org/10.1017/CBO9781139839327.016 Published online by Cambridge University Press

it does not pose serious issues on panel data model, as one can simply focus on the remaining samples that have complete history. If a test (Chapter 11, Section 11.1) indicates that sample attrition is behaviorably related, ignoring the attrition issues could result in misleading inference. Baltagi and Song (2006), and Hirano et al. (2001) show the potential of using refreshment samples to distinguish between various forms of attrition. However, to properly take account of sample attrition, one will have to have explicit knowledge of why individuals drop out. Moreover, as Hausman and Wise (1977) (see Chapter 8, Section 8.2) or Ridder (1990) illustrates, computationally it could be a formidable task to take into account the sample attrition issue.

13.3 A CONCLUDING REMARK

This monograph hopes to provide an overview of the many statistical tools developed to analyze panel data and demonstrate the many advantages panel data may possess. In choosing the proper method to exploit the richness and unique property of panel data, it is helpful to keep several factors in mind. First, what advantages do panel data offer us in adapting economic theory for empirical investigation over data sets consisting of a single cross section or time series? Second, what are the limitations of panel data and the econometric methods that have been proposed for analyzing such data? Third, the usefulness of panel data in providing particular answers to certain issues depends critically on the compatibility between the assumptions underlying the statistical inference procedures and the data-generating process. Fourth, when using panel data, how can we increase the efficiency of parameter estimates? “Analyzing economic data requires skills of synthesis, interpretation and empirical imagination. Command of statistical methods is only a part, and sometimes a very small part, of what is required to do a first-class empirical research” (Heckman 2001). Panel data are no panacea. Nevertheless, if “panel data are only a little window that opens upon a great world, they are nevertheless the best window in econometrics” (Mairesse 2007).

https://doi.org/10.1017/CBO9781139839327.016 Published online by Cambridge University Press