5. Panel data methods in marketing research

Natalie Mizik and Eugene Pavlov

The increased availability of longitudinal marketing data collected at the individual level has offered marketing researchers the option to utilize panel data methods to better study marketing phenomena. The term "panel data" refers to data sets that pool time-series data over multiple cross-sections: individuals, households, firms, business units, or brands.

Panel data are more informative than cross-sectional data because they allow for addressing individual heterogeneity, modeling dynamic processes, and assessing effects that are not detectable in pure cross-sections. Panel data are more informative than aggregate time series because they allow tracking of individual histories and eliminate biases resulting from aggregation. Panel data offer more variability and greater efficiency and allow estimation of more complex and insightful models. Importantly, panel data allow researchers to design models that control for omitted and unobservable factors which can often mask causal effects of interest.

Indeed, the role of unobservables (such as individual ability, firm culture, management quality) has long been debated in the marketing and economics literature. A number of strategy perspectives, for example, the Resource-Based and Austrian economics perspectives, highlight the central role of unobservable factors in explaining business performance. Marketing is largely concerned with the development and deployment of intangible assesses. These assets often fall into the category of unobservables. Unobservable factors (which include both true unobservables and factors that are simply difficult to measure) can be posited to be the most influential determinants of business performance (Jacobson 1990).

As we discuss later in this chapter, modeling and controlling for unobservables in panel data often comes at the expense of efficiency. Kirzner (1976), for example, notes that studies placing great emphasis on unobservable factors are often criticized as incapable of saying anything about observed strategic factors. He feels, however, that the truth is the other way around. Only by controlling for unobservables can insights into strategic factors be adequately assessed. According to Kirzner (1976), "The real world includes a whole range of matters beyond the scope of the measuring instruments of the econometrician. Economic science must encompass this realm." As such, empirically assessing marketing impact hinges critically on controlling for the role of unobservable factors and the panel data methods offer tools to achieve this.

In this chapter, we review panel data models popular in marketing applications and highlight some issues, potential solutions, and trade-offs that arise in their estimation. Panel data studies controlling for unobservables often show dramatically different estimates than cross-sectional studies (Mizik and Jacobson 2004). We focus on estimation of models with unobservable individual-specific effects and address some misconceptions appearing in marketing applications. The choice of discussed topics is highly selective and reflects the authors' review of the panel data methods used in the marketing field. We do not cover some important issues (e.g., the weak instruments problem) and recent developments in the causal modeling as these are presented in Chapter 6, "Causal Inference in Marketing Applications." Furthermore, Chapter 17, using pharmaceutical marketing activity and drug prescriptions data, presents an empirical illustration of the models, methods, and issues discussed here.

STATIC PANEL DATA MODELS

Time-invariant Random Effects: The Random-effects Model

Marketing researchers are frequently confronted with the data comprising observations of multiple units (firms, stores, customers) over time. Let y_{it} be the value of the dependent variable for individual or firm *i* at time *t* and let the set of predictor variables be represented by the vector x_{it}.

$$y_{it} = \alpha_0 + \beta x_{it} + u_{it} \tag{5.1}$$

The error term u_{it} in Equation 5.1 reflects the influence of omitted factors affecting y_{it}. Some of these factors reflected in the error term can be posited to be specific to a particular cross-sectional unit *i*. As such, the error term in Equation 5.1 can be expressed as

$$u_{it} = \mu_i + e_{it},$$

where μ_i is an unobservable time-invariant individual-specific factor and e_{it} is a contemporaneous (idiosyncratic) shock. This structure of the error term induces a block diagonal variance-covariance matrix and calls for the use of generalized least squares (GLS). As long as μ_i and e_{it} are uncorrelated with the explanatory factors x_{it} included in the model, OLS and GLS estimation generates consistent coefficient estimates. However, the residuals for a given cross-section *i* are correlated across periods and, as a result, the reported standard errors from OLS estimation will be biased and inconsistent. The GLS model, known as the random-effects model in the panel data literature (e.g., Chamberlain 1984, Hsiao 1986), not only generates consistent standard errors but it is also asymptotically efficient.

For the random-effects model specification to be valid, it should be plausible that all individual effects μ_i are drawn from the same probability distribution. Strong heterogeneity across cross-sections invalidates the random-effects specification. Generally, random-effects models are unattractive for panels with small number of cross-sectional units *N* and for panels with large time dimension *T*.

Time-invariant Fixed Effects: The Fixed-effects Model

The random-effects model assumes zero correlation between the explanatory factors x_{it} and the unobserved individual-specific factor μ_i. Many researchers (e.g., Mundlak 1978) have criticized the random-effects specification because of the restrictiveness of this assumption. Indeed, many theories of firm performance (e.g., the resource-based perspective, Rumelt 1984, Wernerfelt 1984) emphasize the inter-relatedness of invisible assets and strategic choices. The fixed-effects model takes into account the likely correlation of strategic factors with the unobservable factors that persist over time. Allowing for fixed effects of this type requires modeling these effects explicitly:

$$y_{it} = \alpha_i + \beta x_{it} + e_{it} \qquad (5.2)$$

Equation 5.2 differs from equation 5.1 in that it allows for the time-invariant (fixed) unobserved factors that differ across cross-sections *i* to be correlated with the explanatory factors x_{it}. The effect of these fixed factors is reflected in the individual-specific constant α_i. To the extent that fixed effects α_i are correlated with the observed explanatory variables x_{it} included in the model (even if the correlation is with just one of the several explanatory variables included in the set x, see discussion of bias spreading later in the chapter) the OLS or GLS estimation of Equation 5.2 will generate biased and inconsistent coefficient estimates.

Consistent estimation of the static fixed-effects models

For static panel data models, researchers typically choose one of the two common estimation approaches for obtaining consistent estimates of the effects for the observed strategic factors x_{it} in the presence of unobservable fixed effects (α_i). One approach, the *within* (i.e., *mean-difference*) *estimator*, involves analysis of deviations from the individual-specific mean of each variable. That is, the following model is estimated:

$$y_{it} - \overline{y_\iota} = (\alpha_i - \overline{\alpha_\iota}) + \beta(x_{it} - \overline{x_\iota}) + (e_{it} - \overline{e_\iota}) = \beta(x_{it} - \overline{x_\iota}) + (e_{it} - \overline{e_\iota}) \qquad (5.3)$$

Here, $\overline{y_\iota} = \frac{1}{T}\sum_{t=1}^{T} y_{it}$ and the means of other variables are defined similarly. Since $\overline{\alpha_\iota} = \alpha_i$ (α_i is constant over time for a given cross-sectional unit), the within transformation of the data eliminates the individual-specific unobserved effects α_i from the equation. The within estimator for the effects of the time-varying factors $\hat{\beta}$ is numerically identical to the least-squares dummy variable (LSDV) estimator of

$\hat{\beta}$. The advantage of the dummy variable approach is that it does not difference out and provides direct estimates of $\widehat{\alpha_\iota}$. For short panels (small T and large N), however, the estimates of $\widehat{\alpha_\iota}$ are inconsistent (Cameron and Trivedi 2005, 704).

The other common approach to estimating fixed-effects models, the *first-difference estimator*, involves taking first differences of the data. That is, the following model is estimated:

$$y_{it} - y_{it-1} = (\alpha_i - \alpha_i) + \beta(x_{it} - x_{it-1}) + (e_{it} - e_{it-1}) = \beta(x_{it} - x_{it-1}) + (e_{it} - e_{it-1}) \tag{5.4}$$

Taking either the first-differences or the mean-differences removes all time-invariant factors, including fixed effects α_i. Equation 5.3 assesses how the deviations from the mean of the outcome variable y_{it} are affected by the explanatory variables x_{it} deviating from their mean values. Equation 5.4 assesses how the first-difference in the outcome variable y_{it} is affected by the explanatory variables x_{it} deviating from their previous values. If the model is specified correctly (no mis-specification issues are present), these estimators will generate statistically identical estimates. Under certain conditions (discussed below), however, one estimator may be preferred to the other.

The choice of the estimator for a fixed-effects model: first-difference vs. mean-difference

If the panel consists of two periods only, the within and the first-difference estimators (equations 5.3 and 5.4, respectively) are algebraically identical. For T > 2, mean-differencing (the within estimator) is more efficient under the assumption of homoscedastic and serially uncorrelated disturbances. The within estimator also has an advantage in that it does not eliminate a portion of the data as a result of differencing. First-order differencing eliminates N out of N*T observations, second-order differencing (i.e., $y_{it} - y_{it-2}$) eliminates 2N, and so on. For these reasons, the within estimator (mean-difference) is a more popular method of removing α_i in static panel data models and is the default method of fixed-effects panel data regressions in many software packages. It is implemented in Stata with the *xtreg, fe* command.

The relative efficiency of the within versus the first-difference estimator depends on the statistical properties of the idiosyncratic error term e_{it}. The within estimator is more efficient when the idiosyncratic errors e_{it} are serially uncorrelated. If $e_{it} \sim$ iid $[0, \sigma_e^2]$, then taking first difference generates the error term Δe_{it} which follows an MA(1) process and has a first-order autocorrelation coefficient of –0.5. As such, the first-difference estimator, while still unbiased, is less efficient. However, if e_{it} follows a random walk (exhibits high levels of autocorrelation), the first differenced error term Δe_{it} is serially uncorrelated, and the first-difference estimation is more efficient. The first-difference estimation can be implemented in Stata with the *regress d1.Y d1.X* command, where operator d1 denotes first-differencing.

In situations where the error term e_{it} is somewhere between a random walk and the iid process, it is more difficult to decide between the first-difference versus the within estimators. Wooldridge (2006, 487) suggests examining the autocorrelation patterns of the differenced errors Δe_{it} in order to decide between the first-difference versus the mean-difference estimators. He also suggests performing estimation using both methods to compare the results and then to try to identify the sources of any differences in the estimates. If the first-difference and the mean-difference estimates differ significantly (i.e., the difference cannot be attributed to a sampling error), the strict exogeneity assumption ($E(e_{it}|x_{is}, \alpha_i) = 0, s = 1, \dots, T$) might be violated.

Any of the standard endogeneity problems (measurement error, omitted variables, simultaneity) can induce contemporaneous correlation between the error term e_{it} and the explanatory variables x_{it}. A contemporaneous correlation causes the first-difference and the within estimators to be inconsistent and to have different probability limits. In some applications, it is also possible for the errors e_{it} to be correlated with the past or future values of x_{it}. Correlation between e_{it} and x_{is}, for s ≠ t also causes both estimators to be inconsistent. If s < t (error is correlated with the past values of explanatory variables), including lags of x_{it} and interpreting the equation as a distributed lag model solves the problem. The correlation of e_{it} with the future values of explanatory variables x_{is} (i.e., s > t) is more problematic as it rarely makes economic sense to include future explanatory variables into the estimation model.

Another consideration when choosing the estimator for a fixed-effects model is the potential presence of a measurement error. When measurement error is present in the explanatory variables, the severity of attenuation bias differs for the first-difference and the within estimators (Griliches and Hausman 1986). We address this issue in more detail later in the chapter. In summary, the higher the autocorrelation in the mis-measured explanatory variable, the greater the attenuation bias under the first-difference estimator, compared to the bias under the within estimator. However, if the time dimension T is sufficiently high, taking higher-order differences can potentially remedy the problem.

Our discussion of the first-difference versus the within estimator so far pertained to static panels only. Once dynamics are introduced into the model and a lagged dependent variable is added to the right-hand-side of a model, the time-difference-based estimator becomes the estimator of choice. In dynamic panels, the within estimator is always biased (Nickell 1981). Time-differencing is the core of instrumental variable-based estimation in dynamic panels (e.g., Anderson and Hsiao 1981, Arellano and Bond 1991).

Choosing Between Random-effects and Fixed-effects Specification in Static Panel Data Models

An important issue in static panel data models is whether a random-effects or a fixed-effects model is appropriate. The most important advantage of the fixed-effects model is that it allows for a non-zero correlation between unobserved individual effects α_i and explanatory variables x_{it}, hence delivering consistent estimates regardless of whether assumption $cov(\alpha_i, x_{it}) = 0$ truly holds. The random-effects model, on the other hand, relies on the zero correlation assumption and delivers inconsistent estimates if the assumption is violated. Only if the zero correlation assumption ($cov(\alpha_i, x_{it}) = 0$) holds, is the random-effects specification more desirable than the fixed-effects specification because it generates more efficient parameter estimates.

Some researchers prefer random-effects models because they allow identifying parameters on time-invariant regressors (e.g., gender). Indeed, in the fixed-effects model, where all time-invariant effects are differenced out, it is impossible to distinguish between the effects of time-invariant observables (individual-specific characteristics) and the unobservable fixed effects. This motivation alone, however, is never a legitimate reason for selecting random-effects over fixed-effects specification.

The choice between random-effects and fixed-effects model specification should be driven by the validity of the assumption of no correlation between the unobservable factors α_i and the explanatory factors x_{it} (i.e., $cov(\alpha_i, x_{it}) = 0$). Other considerations should not drive the choice between random-effects versus fixed-effects model specification (Wooldridge 2006, 493). Specification tests for choosing fixed-effects versus random-effects exist and the Hausman (1978) test is the most popular among them. It is focused on assessing the validity of the $cov(\alpha_i, x_{it}) = 0$ assumption. We describe the test, its interpretation, and limitations later in the chapter.

DYNAMIC PANEL DATA MODELS

In dynamic panel data models, a lag of the dependent variable enters the right-hand-side of the estimating equation as another explanatory variable. Researchers are often compelled to include a lagged dependent variable as a predictor when estimating regression models for longitudinal panel data. The reason is that in most situations, the best predictor of what happens at time *t* is what happened at time *t – 1*. Many marketing processes and data series marketing researchers work with (sales, earnings, etc.) have fixed effects and also exhibit high levels of persistence (autocorrelation) and, as such, warrant the inclusion of lagged dependent variables into the model:

$$y_{it} = \alpha_i + \phi y_{it-1} + \beta x_{it} + e_{it} \tag{5.5}$$

Models with lagged dependent variables are known as *dynamic panel data* models and econometricians have long emphasized that lagged dependent variables can cause major estimation problems and lead to severe biases, particularly when individual-specific effects are present. OLS, random-effects, and within estimators generate biased estimates in dynamic panel data models and instrumental variable-based estimators (Anderson and Hsiao 1981, Arellano and Bond 1991) are preferred for dynamic panel data models with individual-specific effects. Unfortunately, some of the estimation issues in the dynamic panel data models are not widely known or appreciated in marketing applications.

Problems with OLS, Within, and Random-effects Estimators in Dynamic Panel Data Models

When a lagged dependent variable enters the model with unobserved individual effects, standard OLS, within, and random-effects estimators are not appropriate, as we describe below.

OLS

The OLS estimator generates biased and inconsistent estimates of model 5.5. The intuition is straightforward. Consider the OLS estimation of the model 5.5:

$$y_{it} = \alpha_0 + \phi y_{it-1} + \beta x_{it} + \alpha_i + e_{it} \qquad (5.6)$$

Both y_t and y_{t-1} depend on α_i. This means that the lagged dependent variable y_{t-1} and α_i, which is a part of the composite OLS error ($\alpha_i + e_{it}$), are correlated. As such, the exogeneity assumption is violated and the estimate of ϕ, as well as the estimates for the other explanatory variables correlated with regressor y_{t-1}, are biased. Hsiao (2014, 86) formally derives the bias for the OLS estimator of ϕ in a simple autoregressive model with fixed effects and reports that OLS tends to overestimate the magnitude of the autoregressive coefficient. Higher variance of individual-specific effects σ_α^2 increases the magnitude of the bias.

Trognon (1978) provides OLS bias formulas for a dynamic panel data model with exogenous regressors and for an autoregressive process of order p. Adding exogenous explanatory variables does somewhat reduce the magnitude, but does not alter the direction or the bias in ϕ: in the first-order autoregressive model with exogenous regressors, the OLS estimate of ϕ remains biased upward and the effects of the exogenous factors are underestimated (their estimates are biased toward zero). The direction of the asymptotic bias for higher-order autoregressive models is difficult to postulate a priori.

Within estimator

The within estimator is not appropriate for the dynamic panel data models with individual-specific effects either. The within transformation of the data in the dynamic panel data models leads to biased estimates. If we apply the within estimator to model (5.5), we would regress $(y_{it} - \overline{y_\iota})$ on $(y_{it-1} - \overline{y_\iota})$ and $(x_{it} - \overline{x_\iota})$:

$$y_{it} - \overline{y_\iota} = (\alpha_i - \overline{\alpha_\iota}) + \phi(y_{it-1} - \overline{y_\iota}) + \beta(x_{it} - \overline{x_\iota}) + (e_{it} - \overline{e_\iota}) =$$

$$\phi(y_{it-1} - \overline{y_\iota}) + \beta(x_{it} - \overline{x_\iota}) + (e_{it} - \overline{e_\iota}) \qquad (5.7)$$

This regression has an error term equal to $(e_{it} - \overline{e_\iota})$. By construction, y_{it} is a function of e_{it} and y_{it-1} is a function of e_{it-1}. But e_{it-1} enters the calculation of the mean of errors $(\overline{e_\iota})$ and, as such the lagged mean-differenced dependent variable regressor $\left(y_{it-1} - \overline{y_{\iota,-1}}\right)$ is correlated with the mean-differenced error term $(e_{it} - \overline{e_\iota})$. Specifically, y_{it-1} and $\overline{e_\iota}$ are correlated because they share a common component (e_{it-1}). This correlation of the lagged mean-differenced dependent variable with the mean-differenced error term gives rise to the dynamic panel bias (Nickell 1981).

Nickell (1981, 1422) derives the general expression for the within estimator bias in dynamic panels. For the arbitrary T and ϕ the bias is equal to

$$\underset{N \to \infty}{plim}\left(\hat{\phi} - \phi\right) = \frac{-(1+\phi)}{T-1}\left\{1 - \frac{1}{T}\frac{(1-\phi^T)}{1-\phi}\right\} \cdot \left\{1 - \frac{2\phi}{(1-\phi)(T-1)}\left(1 - \frac{1}{T}\frac{(1-\phi^T)}{1-\phi}\right)\right\}^{-1}$$

The magnitude of the bias can be significant. For example, when the true value of $\phi = 0.5$ and T=10, the bias is equal to –0.167. This implies a 33.4 percent deviation from the true value (i.e., –0.167/0.5). As long as ϕ is positive, the sign of the bias is always negative and the within estimator underestimates the magnitude of ϕ.

The severity of the bias for the within estimator is greater for shorter panels. The bias diminishes for longer time series because as T→∞, the contribution of e_{it-1} to $\overline{e_\iota}$ decreases and $\left(y_{it-1} - \overline{y_{\iota,-1}}\right)$ becomes asymptotically uncorrelated with $(e_{it} - \overline{e_\iota})$, reducing the dynamic panel bias of the mean-difference (within) estimator. For large T, the asymptotic bias is approximated by:

$$\underset{N \to \infty}{plim}\left(\hat{\phi} - \phi\right) \cong \frac{-(1+\phi)}{T-1}$$

Random effects

A random-effects specification is generally not appropriate in dynamic panel data models because the assumption of no correlation between the unobservable factors μ_i and the explanatory factors is violated. The logic is straightforward. If we add a lagged dependent variable to the set of explanatory variables in a random-effects model (5.1), we obtain the following model:

$$y_{it} = \alpha_0 + \phi y_{it-1} + \beta x_{it} + \mu_i + e_{it} \tag{5.8}$$

In the random-effects models the random intercept (μ_i) is assumed to be independent of all other variables on the right-hand side. μ_i represents the combined effect on y_{it} of all unobserved variables that are constant over time. Because the model applies at all time points, μ_i also has a direct effect on y_{it-1}:

$$y_{it-1} = \alpha_0 + \phi y_{it-2} + \beta x_{it-1} + \mu_i + e_{it-1} \tag{5.9}$$

That is, y_{it-1} is not statistically independent of μ_i, which is a component of the composite error in the equation (5.8) above. This violation of the zero correlation assumption in the random-effects model biases both the coefficient for the lagged dependent variable y_{it-1} and the coefficients of all other explanatory variables x_{it} correlated with y_{it-1}.

For a summary discussion of the required assumptions about the initial conditions, and the resulting consistency/inconsistency of the maximum likelihood (MLE), generalized least-squares (GLS), instrumental variables (IV), and generalized method of moments (GMM) estimators in models with individual effects see Hsiao (2014). Different assumptions about initial conditions (Hsiao 2014, 87, outlines four different cases and six subcases) imply different likelihood functions and generate different results. It is often not possible to make an informed choice regarding the initial conditions, and an incorrect choice results in inconsistent estimates. Anderson and Hsiao (1981) proposed a simple consistent estimator that is independent of initial conditions, and it became the foundation for the development of a set of consistent estimators preferred in empirical applications with dynamic panel data models.

Consistent Instrumental Variable-based Estimation of Dynamic Panel Data Models with Individual-Specific Effects

The first-difference instrumental variable-based estimator developed by Anderson and Hsiao (1981) and its extensions (e.g., Arellano and Bond 1991) became dominant for estimating dynamic panel data models with individual effects.

Consider the first-difference transformation of equation 5.5:

$$y_{it} - y_{it-1} = \phi(y_{it-1} - y_{it-2}) + \beta(x_{it} - x_{it-1}) + (e_{it} - e_{it-1}) \qquad (5.10)$$

By construction, y_{it-1} is correlated with e_{it-1} and ϕ is biased. As such, an instrument Z is required for the regressor $(y_{it-1} - y_{it-2})$. An instrumental variable candidate should exhibit the properties of relevance (i.e., $cov(Z, y_{it-1} - y_{it-2}) \neq 0$) and validity (i.e., $cov(Z, e_{it} - e_{it-1}) = 0$). Anderson and Hsiao (1981) pointed out that y_{it-2} is a valid instrument for $(y_{it-1} - y_{it-2})$ because it is not correlated with e_{it-1}. The estimation can be carried out in a two-stage least squares (2SLS) procedure:

Step 1: Regress $(y_{it-1} - y_{it-2})$ on y_{it-2} and obtain predicted values $\Delta\widehat{y_{it-1}}$. Since y_{it-2} is a valid instrument, $\Delta\widehat{y_{it-1}}$ is a portion of $(y_{it-1} - y_{it-2})$ uncorrelated with e_{it-1}.

Step 2: Regress $(y_{it} - y_{it-1})$ on $\Delta\widehat{y_{it-1}}$ and $(x_{it} - x_{it-1})$. The resulting estimates $\hat{\phi}$ and $\hat{\beta}$ are consistent.

Other valid instruments also exist. For example, $(y_{it-2} - y_{it-3})$ is also a valid instrument for $(y_{it-1} - y_{it-2})$. Using $(y_{it-2} - y_{it-3})$ rather than y_{it-2}, however, requires an additional time period of data and leaves the researcher with N fewer observations in the final estimation step. The strength of a particular instrumental variable is an empirical question, and can be examined in the first stage of 2SLS estimation. The Anderson-Hsiao estimator is implemented in Stata with *xtivreg*, *fd* command.

Extending this logic of Anderson and Hsiao (1981) further, any level or difference of y_{it}, appropriately lagged, is a valid instrumental variable for $(y_{it-1} - y_{it-2})$. The pool of such potential instrumental variables grows with increasing T. Certain optimal combinations of instrumental variables might deliver more efficient estimates. Identification of this optimal combination is at the core of Arellano and Bond (1991) estimator.

The Arellano-Bond GMM estimator specifies a system of equations (one equation per time period) and allows the instruments to differ for each equation (e.g., additional lags are available as instruments in later periods). As we have many instruments and only one variable that requires instrumentation $(y_{it-1} - y_{it-2})$, the system will be overidentified, calling for the use of Generalized Method of Moments (GMM).

The method of moments estimator uses moment conditions of the type: $E[Z'_{1it}(\Delta y_{it} - \phi\Delta y_{it-1} - \beta\Delta x_{it})] = 0$, which reflects the validity of a particular instrument: $Z_{1it} \perp (e_{it} - e_{it-1})$. $\frac{1}{N}\sum_{i=1}^{N} Z'_{1it}\Delta e_{it} = 0$ are sample analogues of these moment conditions. The goal of the method of moments estimator is to find values β, ϕ such that sample moment conditions are satisfied. If the system is overidentified (i.e., there are more instruments than variables that require instrumentation), it is often impossible to find values β, ϕ that strictly satisfy all orthogonality conditions. Instead, the idea underlying the GMM approach is to find β, ϕ that minimize a certain (loss) function of all sample moment conditions. Such objective function often takes the form:

$$J(\beta, \phi) = \ \bar{g}(\beta, \phi)' \ W \ \bar{g}(\beta, \phi)$$

Here, $\bar{g}(\beta, \phi)$ is $l \times 1$ vector of l stacked sample moment conditions, l is the number of instruments, and W is a $l \times l$ weighting matrix. As long as W is positive-definite, GMM estimates of β, ϕ will be consistent (Wooldridge 2002, 422). However, certain choices of Wcan also deliver efficiency of β, ϕ. The optimal weight corresponding to a specific moment condition is typically inversely proportional to the variance of this moment condition.

The Arellano-Bond (1991) estimator is defined as:

$$\widehat{\beta_{AB}} = \left[\left(\sum_{i=1}^{N} \breve{X}_{\iota}{}'Z_i\right) W_N \left(\sum_{i=1}^{N} Z_i{}'\breve{X}_{\iota}\right)\right]^{-1} \left(\sum_{i=1}^{N} \breve{X}_{\iota}{}'Z_i\right) W_N \left(\sum_{i=1}^{N} Z_i'\breve{y}_{\iota}\right)$$

$\breve{X}_{\iota}$ is the matrix of regressors where row t is $[\Delta y_{it-1}, \Delta x_{it}{}']$ (t=3,…, T), $\breve{y}_{\iota}$ is a vector of the dependent variable with Δy_{it} in row t, and Z_i is a matrix of instruments:

$$Z_i = \begin{bmatrix} z_{i3}{}' & 0 & \dots & 0 \\ 0 & z_{i4}{}' & \vdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & z_{iT}{}' \end{bmatrix}$$

The z_{it} element of Z_i is $[y_{it-2}, y_{it-3}, \dots, y_{i1}, \Delta x_{it}']$, and the number of rows of Z_i equals to $T - 2$. For example, if $T = 5$,

$$Z_i = \begin{bmatrix} y_{i1} & \Delta x_{i3} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & y_{i2} & y_{i1} & \Delta x_{i4} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & y_{i3} & y_{i2} & y_{i1} & \Delta x_{i5} \end{bmatrix}$$

The intuition underlying this structure is as follows. Suppose we observe five years of panel data, 2011 to 2015. For 2011 and 2012 we do not have valid instruments (e.g., we do not observe years 2009 and 2010). Thus, only years 2013–2015 will enter the Arellano-Bond estimation procedure. We have only one valid instrument for 2013 coming from 2011. For 2014 we have two valid instruments from 2011 and 2012. For 2015 we have three valid instruments. Arellano-Bond GMM-based estimator utilizes information more efficiently (compared to Anderson and Hsiao 1981), especially for longer panels as the pool of available instruments grows in T. When T is large, a researcher might wish to limit the maximum number of lags of an instrument. The Arellano-Bond (1991) estimator is implemented in Stata with the *xtabond* routine.

One weakness of the Arellano-Bond ("Difference GMM") estimator is that lagged levels sometimes can be rather weak instruments for the first-differenced variables. The problem is particularly pronounced when the variables exhibit high autocorrelation (e.g., random walk). Arellano and Bover (1995) and Blundell and Bond (1998) developed the so-called System GMM estimator, which incorporates lagged differences, along with lagged levels of y_{it}, into the matrix of instruments Z_i. Incorporating additional information contained in lagged Δy_{it} allows to further increase efficiency of the estimator. The Blundell and Bond (1998) estimator is implemented in Stata with *xtdpd*.

Lags of independent variables as instruments are consistent under the assumption that idiosyncratic errors e_{it} are not serially correlated. This assumption is testable through the Arellano-Bond (1991) test for serial correlation in errors. If e_{it} are iid, then Δe_{it} exhibit negative first-order serial correlation and zero serial correlation at higher orders. That is, when the null hypothesis of no serial correlation is rejected at order 1, but is not rejected at higher orders, the validity of Arellano-Bond instruments is supported. The test is implemented in Stata with *estat abond* command which should be run after *xtabond* (or *xtdpd* in case of system GMM estimation). The Sargan/Hansen test of overidentifying restrictions (Sargan 1958, Hansen 1982) assesses the joint validity of instruments in a given model. *xtabond2* command reports Sargan and Hansen statistics separately after model estimation. Roodman (2009) offers discussion of the tests and their interpretation.

The two-step Arellano-Bond estimation has been shown to generate downward biased standard errors (the one-step implementation does not have this issue). Arellano and Bond found that "the estimator of the asymptotic standard errors of GMM2 shows a downward bias of around 20 percent relative to the finite-sample standard deviations" (1991, 285). The Windmeijer (2005) finite sample correction resolves the issue. It is available in Stata with the *xtabond, twostep vce(robust)* command syntax.

SPECIFICATION TESTING

How can a researcher choose an appropriate model specification and estimator for the data at hand? Hausman (1978) suggested a specification test designed to assist researchers in choosing between potential alternative estimators. The test relies on the observation that two consistent estimates will not differ systematically. The Hausman specification test can be used to determine the possible presence of the different types of unobservable factors and their correlation with the explanatory factors. The hypothesis of no time-invariant effects, for example, can be assessed by comparing the estimates of the fixed-effects model with the random-effects model. Similarly, the fixed-effects estimator can be compared to the fixed-effects/ instrumental variable estimator to test for the presence of contemporaneous shocks correlated with the error term and the fixed-effects/instrumental variable model can be compared to the fixed-effects/instrumental variable/ serial correlation model to test for the presence of an autocorrelated error term. In the discussion below we use the test for random versus fixed-effects specification as an illustration.

The following logic underlies the Hausman specification test. Fixed-effects estimates are *assumed* to be consistent whether the assumption of $cov(\alpha_i, x_{it}) = 0$ holds or not, because they directly account for time-invariant individual-specific unobserved heterogeneity. The random-effects model estimates are consistent and efficient (i.e., minimum variance) under the null hypothesis that the fixed effects and the contemporaneous shocks are uncorrelated with the explanatory factors. However, under the alternative hypothesis of omitted fixed effects correlated with the explanatory factors included in the model, the random-effects estimates will be biased and inconsistent (see Table 5.1).

Under the null hypothesis of the time-invariant individual-specific effects α_i being uncorrelated with the explanatory factors x_{it} (i.e., $cov(\alpha_i, x_{it}) = 0$), the estimates from a random-effects model should not differ significantly from the estimates obtained from a fixed-effects model. If a statistically significant discrepancy between random-effects and fixed-effects model estimates is not detected, the finding is interpreted as evidence in favor of the assumption that individual effects are (approximately) uncorrelated with the regressors. In such a case, random-effects estimates are consistent and the random-effects model is preferred to fixed-effects models because the random-effects estimates are efficient and the coefficients on time-invariant regressors can be identified. However, if a significant discrepancy between random-effects and fixed-effects model estimates is found, random-effects estimates are deemed inconsistent and the fixed-effects model is preferred.

Table 5.1. Hausman Test for Fixed-Effects vs. Random-Effects Specification

	FE estimator:	RE estimator:	Implication
H_0: $(cov(a_i, x_{it}) = 0$	Consistent	Consistent and efficient	RE model preferred
H_1:$(cov(a_i, x_{it}) \neq 0)$	Consistent	Inconsistent	FE model preferred

The Hausman test statistic can be computed as:

$$H = \frac{\left(\hat{\beta}_{FE} - \hat{\beta}_{RE}\right)^2}{Var(\hat{\beta}_{FE}) - Var(\hat{\beta}_{RE})}$$

Under the null hypothesis H follows χ_M^2 distribution, where M is the dimensionality of the coefficient vector. The test can be performed for the whole set of coefficients on time-varying regressors (time-invariant regressors are not identified in the fixed-effects model) or for a subset of the coefficients of interest. In Stata, this test is implemented with the *hausman* command.

Before interpreting the Hausman test and using it to choose between estimators, however, it is important to understand the underlying assumptions and limitations of this test.

Assumption of Consistency of $\hat{\boldsymbol{\beta}}_{FE}$ under Both the Alternative and the Null Hypothesis

The Hausman test relies on the assumption that the fixed-effects estimator $\hat{\beta}_{FE}$ is consistent. That is, it assumes that there is no correlation between x_{it} and e_{it} in any time period, once fixed effects are accounted for. This assumption can be violated. For example, it is violated if relevant variables are omitted or the unobserved heterogeneity in the model is time-variant and the unobserved effect varies over time (α_{it}). In this case, a fixed-effects estimator is not consistent, and cannot serve as an appropriate benchmark in the Hausman test. Under time-varying unobserved heterogeneity, neither fixed-effects nor random-effects estimators are appropriate and the Hausman test would not indicate that.

In the classic interpretation of the Hausman test, the difference between the random-effects and fixed-effects model estimates is attributed to a single issue, namely, the correlation between the unobserved fixed effects and the explanatory factors. Often, in empirical applications the discrepancy between the fixed-effects and random-effects estimators can be driven by other factors.

For example, when the right-hand-side variables are subject to measurement error, a fixed-effects estimator can be subject to a greater attenuation bias compared to a corresponding cross-section estimate. The fixed-effects estimator removes all cross-sectional variation in the data, which is good because it removes the biases due to unobserved individual heterogeneity. However, is also removes useful information about the variables of interest. Depending on the characteristics of particular data, the change in signal-to-noise ratio as a result of applying a fixed-effects estimator is ambiguous, and in many cases is disadvantageous. When measurement error is present, a researcher undertaking a Hausman test might find that fixed-effects estimates are lower in absolute magnitude compared to the alternative random-effects or OLS estimates. The difference might be due to the unobserved heterogeneity biases in random-effects and OLS, or, it can be due to the attenuation bias exacerbated by the differencing of the data in the fixed-effects estimation. In such case, rather than relying on the Hausman test to choose between fixed-effects and random-effects estimators, a researcher should undertake steps to investigate and tackle the potential measurement error problem (e.g., through IV methods).

Assumption of Efficiency for the Random-effects Estimator

A fundamental assumption of the Hausman test for the random-effects estimator is that individual effects are distributed independently of the idiosyncratic error and regressors. The assumption of efficiency is violated when the data are clustered. In empirical applications where cluster-robust standard errors are preferred over classical errors, a robust Hausman test procedure might be required (Cameron and Trivedi 2009, 261). Such a situation might occur, for example, if there are no distinct individual fixed effects, but rather the errors u_{it} for a given panelist *i* exhibit significant autocorrelation.

Cameron and Trivedi (2009) suggest the following procedure for a robust Hausman test. Test H$_0$: $\gamma = 0$ in the following regression:

$$\left(y_{it} - \hat{\theta}\overline{y_{\imath}}\right) = \left(1 - \hat{\theta}\right)\alpha + \left(x_{it} - \hat{\theta}\overline{x_{\imath}}\right)\beta + (x_{it} - \overline{x_{\imath}})\gamma + v_{it},$$

where x_{it} refers to time-varying regressors and $\hat{\theta}$ is an estimate of $\theta = 1 - \sqrt{\sigma_e^2/(T\sigma_\alpha^2 + \sigma_e^2)}$, the relative proportion of how much between versus within variation is used by the random-effects estimator ($\theta = 0$ corresponds to a pooled OLS estimate, $\theta = 1$ corresponds to a fixed-effects estimate–i.e., within variation only). $\hat{\theta}$ could be estimated beforehand using random-effects estimation (e.g., it is a part of a standard output in *xtreg, re* command in Stata). The interpretation of rejecting H$_0$: $\gamma = 0$ is similar to that in the classic Hausman test.

"All or Nothing" Assumption Regarding Exogeneity in the Model

The null and the alternative hypotheses in the Hausman test refer to extreme cases where either all covariates are exogenous (i.e., the random-effects estimator is appropriate), or none of the regressors are exogenous (a fixed-effects model is required). Baltagi (2005, 19) notes that one should probably not immediately proceed with fixed-effects estimation if the classic Hausman test rejects H$_0$. Instead, he advises researchers to explore models that allow for only *some* regressors to be correlated with the fixed effects α_i, while still maintaining the assumption (that all regressors are uncorrelated with idiosyncratic shocks e_{it}).

Hausman and Taylor (1981) developed an estimator which allows some of the regressors in the set x_{it} to be correlated with α_i. The Hausman and Taylor (HT) estimator is an instrumental variable-based estimator (implemented in Stata with command *xthtaylor*). It combines the elements of both fixed-effects and random-effects estimators and offers a range of benefits. The HT procedure gives researchers additional flexibility: when it is appropriate, it delivers consistent estimates that are more efficient than fixed-effects and it allows for identification of time-invariant regressors. As such, it generates better estimates than either the random-effects or the fixed-effects estimators.

Baltagi (2005, 132) suggests the following sequence of steps in applying the HT pre-test estimator:

Step 1: If H_0 of the standard Hausman test (fixed-effects vs. random-effects) is not rejected, a random-effects model should be chosen.

Step 2: If H_0 of the standard Hausman test is rejected, HT estimation is implemented, and another Hausman test (fixed-effects vs HT) is run.

1) If H_0 of the second Hausman test is not rejected (no systematic difference between fixed-effects and HT estimates), HT model should be used;
2) If H_0 of the second Hausman test is rejected, a fixed-effects model should be used.

Power Issues

The Hausman test is a statistical test derived under large sample properties. The denominator of the Hausman statistic relies on the asymptotic variances of coefficient estimates. The betas are assumed to be normally distributed with means $\hat{\beta}_{FE}$ and $\hat{\beta}_{RE}$ and the asymptotic variances $Var(\hat{\beta}_{FE})$ and $Var(\hat{\beta}_{RE})$. The Hausman test computed for small samples should be viewed with additional caution because the variances $Var(\hat{\beta}_{FE})$ and $Var(\hat{\beta}_{RE})$ calculated based on small samples can be far from their asymptotic counterparts.

MEASUREMENT ERROR IN PANEL DATA MODELS

Measurement error is a well-known problem in the empirical literature. Its consequences can be more severe in panel data setting.

The error-in-variables problem typically refers to measurement error in the independent variables. An immediate consequence of the error-in-variables problem is the so-called attenuation bias in the estimated coefficient of interest. That is, a bias toward zero. Measurement error in the dependent variable has less severe consequences. It causes loss of efficiency, but it does not cause bias in the estimates. In the discussion that follows we focus on the measurement error in the independent variables and potential solutions for obtaining consistent estimates.

Errors in Variables in Cross-sectional Settings

To introduce the problem, let us begin with a simple cross-sectional illustration of measurement error in the independent variable. Consider the following model:

$$y_i = \alpha_0 + \beta x_i + e_i \qquad (5.11)$$

We are interested in estimating β, which measures the relationship between x_i and y_i. However, we can only observe x_i^*, which is our measure of x_i combined with a classical measurement error v_i ($x_i^* = x_i + v_i$). That is, v_i is iid noise with a mean of zero and variance σ_v^2 and is uncorrelated with x_i and e_i. Because $cov(x_i, v_i) = 0$ and because $x_i^* = x_i + v_i$, it follows that our observed measure x_i^* is correlated with v_i. The magnitude of their covariance is equal to the variance of the measurement error v_i: $cov(x_i^*, v_i) = E(x_i^*, v_i) = E(x_i, v_i) + E(v_i^2) = \sigma_v^2$.

The covariance between our observed measure x_i^* and measurement error v_i causes a non-zero correlation between the regressor and the composite error in the model:

$$y_i = a + \beta(x_i^* - v_i) + e_i = a + \beta x_i^* + (e_i - \beta v_i)$$

Because $cov(x_i, v_i) = 0$, $var(x_i^*) = var(x_i) + var(v_i) = \sigma_x^2 + \sigma_v^2$, and $cov(x_i^*, e_i - \beta v_i) = -\beta cov(x_i^*, v_i) = -\beta\sigma_v^2$, we can derive the OLS estimator as:

$$plim\left(\hat{\beta}\right) = \beta + \frac{cov(x_i^*, e_i - \beta v_i)}{var(x_i^*)} = \beta - \frac{\beta\sigma_v^2}{\sigma_x^2 + \sigma_v^2} = \beta\left(\frac{\sigma_x^2}{\sigma_x^2 + \sigma_v^2}\right) \qquad (5.12)$$

Unless $\sigma_v^2 = 0$, the multiplier term $\left(\frac{\sigma_x^2}{\sigma_x^2 + \sigma_v^2}\right)$ is always less than 1 and $\hat{\beta}$ is biased (toward zero) and inconsistent. This result is known as attenuation bias. The magnitude of the bias depends on the *signal-to-noise ratio* $\left(\frac{\sigma_x^2}{\sigma_x^2 + \sigma_v^2}\right)$: The greater the variance of the measurement error (noise) relative to the variance of the true regressor x_i (signal), the greater the magnitude of the bias.

Inclusion of additional regressors into model 5.11 increases the magnitude of the attenuation bias and the bias spreads to additional regressors. Please see next section for a discussion of measurement error bias in multivariate setting and bias spreading.

Errors in Variables in Static Panel Data Models

Measurement error can be significant in the cross-sectional setting, but in the panel data setting, the attenuation bias due to measurement error can become even more severe, particularly when the researcher utilizes the mean-difference or the first-difference panel data estimators to control for time-invariant individual-specific fixed effects α_i. Under strict exogeneity in the classical errors-in-variables model, differencing removes the omitted variable (fixed effects) bias but exacerbates measurement error bias. The intuition behind this phenomenon is straightforward: while eliminating the effect of α_i, the within and the first-differencing estimators remove a large portion of variation in the data, both the noise and the signal. For a wide variety of data generating processes underlying the x_{it} and v_{it} series, the signal-to-noise ratio decreases when the within or the first-difference estimators are applied, making the attenuation bias in the estimates more pronounced. The measurement error and the resulting attenuation bias may be responsible for the within and the first-difference estimators generating small and insignificant estimates in many empirical settings (Angrist and Pischke 2008).

Measurement error bias in OLS and first difference-estimators in static panels

Let us consider the following static panel data model with measurement error in the independent variable:

$$y_{it} = \alpha_i + \beta x_{it} + e_{it}, \qquad (5.13)$$

where $x_{it}^* = x_{it} + v_{it}$.

Here, x_{it} is the true regressor of interest, and x_{it}^* is its observed value which is measured with measurement error v_{it}. For generality, let us allow x_{it} series to be autocorrelated with the autocorrelation parameter γ_x ($\gamma_x < 1$) and the measurement error v_{it} series to be autocorrelated with the autocorrelation parameter γ_v ($\gamma_v < 1$), such that $cov(v_{it}, v_{it-1}) = \gamma_v \sigma_v^2$, where $Var(v_{it}) = \sigma_v^2$. Further, let's assume that the measurement error v_{it} is not correlated with the true regressor x_{it}, the unobserved individual effect α_i, and the idiosyncratic error e_{it}. Estimating model 5.13 by OLS yields the following probability limit for the estimate $\widehat{\beta_{OLS}}$:

$$plim_{\,N\to\infty}\, \widehat{\beta_{OLS}} = \beta \frac{\sigma_x^2}{\sigma_v^2+\sigma_x^2} + \frac{Cov\,(x_{it},\alpha_i)}{\sigma_v^2+\sigma_x^2} \qquad (5.14)$$

The total bias of $\widehat{\beta_{OLS}}$ consists of two components. The first term, multiplier $\left(\frac{\sigma_x^2}{\sigma_v^2+\sigma_x^2}\right)$, is the familiar attenuation bias caused by the presence of the measurement error. The second term ($Cov\,(x_{it}, \alpha_i)/[\sigma_v^2 + \sigma_x^2]$) is the omitted variable bias caused by the failure to account for the individual heterogeneity.

Individual-specific heterogeneity effects α_i can be eliminated from model 5.13 through first-differencing and estimating the model: $\Delta y_{it} = \beta \Delta x_{it} + \Delta e_{it}$. In this formulation, the expected value of $\hat{\beta}$ can be derived similarly to that in equation 5.2 as:

$$plim\,(\hat{\beta}) = \beta \left(\frac{\sigma_{\Delta x}^2}{\sigma_{\Delta x}^2+\sigma_{\Delta v}^2}\right), \text{ where}$$

$$\sigma_{\Delta x}^2 = Var(x_{it} - x_{it-1}) = Var(x_{it}) - 2cov(x_{it}, x_{it-1}) + Var(x_{it-1})$$

Assuming that x_{it} is stationary means that moments of x_{it} distribution are the same for any t. In particular, $Var(x_{it}) = Var(x_{it-1})$. Then, $\sigma_{\Delta x}^2 = 2\sigma_x^2 - 2cov(x_{it}, x_{it-1}) = 2\sigma_x^2(1 - \gamma_x)$. If v_{it} is stationary as well, then $\sigma_{\Delta v}^2 = 2\sigma_v^2(1 - \gamma_v)$. Hence, the probability limit of the first-difference estimate under measurement error (Pischke 2007) is

$$plim_{\ N\to\infty}\ \widehat{\beta_{FD}} = \beta \frac{\sigma_x^2(1-\gamma_x)}{\sigma_x^2(1-\gamma_x)+\sigma_v^2(1-\gamma_v)} \qquad (5.15)$$

We can compare the magnitude of the bias in the OLS (equation 5.14) and first-difference (equation 5.15) estimates. If there is no measurement error ($\sigma_v^2 = 0$), the first-difference estimate is unbiased while OLS is biased because it fails to account for individual heterogeneity. If $\sigma_v^2 > 0,$ both estimators are subject to attenuation bias, and the relative size of the biases depends on γ_v and γ_x, the degree of autocorrelation in the measurement error and explanatory variable, respectively. If $\gamma_x > \gamma_v$, attenuation bias is greater for $\widehat{\beta_{FD}}$ than for $\widehat{\beta_{OLS}}$ (Hsiao 2014, 56). If x_{it} is autocorrelated stronger than the measurement error v_{it} (i.e., $\gamma_x > \gamma_v$), first-differencing x_{it} results in a reduction in the signal-to-noise ratio making the attenuation bias more severe compared to the attenuation bias component in the OLS estimate. When v_{it} resembles white noise (no persistence), the attenuation bias of the first-difference estimator is large, especially for higher γ_x. On the other hand, as the persistence in the measurement error increases (γ_v goes to 1), the attractiveness of the first-difference estimator increases.

Measurement error bias in mean-difference and first-difference estimators in static panels

Griliches and Hausman (1986) compared attenuation biases of the mean-difference (the within) and first-difference estimators. Both estimators address the individual heterogeneity issue by differencing out α_i, but they have different implications for the magnitude of the measurement error bias. Griliches and Hausman (1986) point out that, while the attenuation bias in the first-difference estimator does not depend on the lengths of the time series dimension T (if $N \to \infty$), it does so for the within estimator because the mean-differencing transformation for the within estimation is calculated taking into account all periods. As such, the relative advantage of a particular estimator depends on T, ρ_j (the j-th order autocorrelation coefficient of the true regressor), and r_j (the j-th order of the autocorrelation coefficient in the measurement error).

Under $r_j = 0$ (for all j), higher ρ_j results in larger attenuation bias for the first-difference estimator since first-differencing removes “more of a signal” in the variable with higher autocorrelation (Griliches and Hausman 1986, 98). The relationship between the biases under the within and the first-difference estimators is summarized in Table 5.2 below.

Table 5.2. When is the attenuation bias smaller for the within estimator versus the first-difference estimator (adapted from Griliches and Hausman 1986, p. 99):

T	Conditions for $\lvert b_w \rvert > \lvert b_{fd} \rvert$
2	Biases are identical
3	$\rho_2 < \rho_1$
4	$\frac{2}{3}\rho_2 + \frac{1}{3}\rho_3 < \rho_1$
…	
T→∞	$\frac{2}{T}(\rho_1 + \rho_2 + ..) < \rho_1$

The condition for the within estimator to be less biased than the first-difference estimator depends on the decay pattern in the x_{it} correlogram: the steeper the decline in the autocorrelation function of x_{it}, the greater the attenuation bias under first-differencing, compared to the bias under the within estimator.

The intuition of this result generalizes to the case when measurement error is autocorrelated with coefficient r_j. Generally, if $\rho_j > r_j > 0$ for all j (i.e., the serial correlation is greater in the explanatory variable than in the measurement error) and the decline in the autocorrelation function of x_{it} is steeper than that in the autocorrelation function of v_{it}, the within estimator is less biased than first-difference estimator. For exact conditions under which the within is less biased than the first-difference estimator under correlated errors, see Griliches and Hausman (1986, 101).

Errors in Variables in Dynamic Panel Data Models

In many empirical settings with measurement error problem, the within estimator may be more consistent compared to the first-difference estimator. However, the within estimator is not appropriate in dynamic panel data models. In dynamic models where measurement error is suspected, the researcher can consider long-difference estimators to assess the problem and reduce the measurement error bias.

If measurement error is not autocorrelated ($r_j = 0$, for all j), then a long-difference estimator with order $j = T - 1$ (i.e., $x_{it} - x_{iT-1}$) is optimal (it is also less inconsistent than the within estimator in static models). For differences of orders longer than 1 and shorter than $T - 1$, the situation is more ambiguous, and the outcome depends on T and the speed of autocorrelation decay of x_{it}. If the measurement error is autocorrelated, then the optimal order of the difference estimator (i.e., the differencing of order j which minimizes attenuation bias in the long-difference estimator) is one that maximizes the expression $\left(1 - \rho_j\right)/\ (1 - r_j)$ (Griliches and Hausman 1986, 101). Depending on the data-generating processes underlying x_{it} and v_{it}, optimal j might be 1, T – 1, or something in-between.

Assessing and Managing Measurement Error Problem in Panel Data Models

To assess the potential presence of measurement error, the researcher can compare results from the within, the first-difference, and the long-difference estimators. Under no measurement error in static fixed-effects models, the estimates should be roughly the same since all three estimators are consistent as they eliminate the unobserved individual effect α_i If first-difference estimates are lower in magnitude compared to within estimates, and the discrepancy in magnitude dissipates/reverses when longer differences are used, this pattern might indicate the presence of a measurement error. Similarly, in dynamic models, an increase in the magnitude of the estimates between the first-difference and long-difference estimators may indicate the presence of measurement error.

Dealing with the measurement error problem in panel data models typically requires finding instruments. First, one can look for external instruments that are correlated with the true underlying variable x_{it}, but uncorrelated with the measurement error v_{it}. Such instruments are often difficult to find. Second, depending on the statistical properties of x_{it} and v_{it}, one might be able to use certain lags/leads of the observed variable x_{it}^* as an instrument. In particular, if v_{it} is iid and if x_{it} is serially correlated, one could potentially use x_{it-2}^* and/or Δx_{it-2}^* to instrument for Δx_{it} in first-difference estimation (Hsiao 2014, 456). In general, if v_{it} is known/assumed to exhibit a certain structure, a consistent IV-based estimation should be available provided that the panel at hand is long enough. For a further reading and applications of IV-based measurement error treatments in panels, we refer the reader to Hsiao (2014), Griliches and Hausman (1986), and Biørn (2000).

BIAS SPREADING IN MULTIVARIATE MODELS

One common misconception about the violation of the exogeneity (zero correlation) assumption is that if only one of the variables in the set of explanatory factors is correlated with the error term, then the other coefficients will still be consistently estimated. This is incorrect. The estimates for all explanatory variables included in the model will be biased, unless they are perfectly orthogonal. The bias effectively spreads from the endogenous regressor to the other estimates.

To provide a quick intuition for bias spreading, consider the fixed-effects model (5.2) and assume that only one, the first, variable in the set x_{it} (x_{1it}), is correlated with the individual-specific effect α_i. That is, $cov(\alpha_i, x_{1it}) \neq 0$. If the researcher chooses to estimate model $y_{it} = \alpha_0 + \beta x_{it} + u_{it}$ without explicitly addressing the fixed effects α_i, we have the situation where $u_{it} = \alpha_i + e_{it}$ and $E(u_{it}|x_{1it}) \neq 0$, with $\sigma_{x_1,u} = \sigma_{x_1,\alpha_i} \neq 0$.

Because $\hat{\beta} = \beta + (X'X/N)^{-1}(X'U/N)$, $plim(\hat{\beta}) = \beta$ requires $plim(X'U/N) = 0$. If this does not hold, the estimator is inconsistent.

In our case,

$$plim\frac{1}{N}X'U = \begin{bmatrix} \sigma_{x_1,\alpha_i} \\ 0 \\ .. \\ 0 \end{bmatrix}$$

and

$$plim(\hat{\beta} - \beta) = plim(X'X/N)^{-1}(X'U/N) =$$

$$plim\frac{1}{N}(X'X)^{-1}\begin{bmatrix} \sigma_{x_1,\alpha_i} \\ 0 \\ .. \\ 0 \end{bmatrix} = \sigma_{x_1,\alpha_i}\begin{bmatrix} q^{11} \\ q^{21} \\ .. \\ q^{K1} \end{bmatrix} =$$

$$\sigma_{x_1,\alpha_i} \times [1st\ column\ of\ Q^{-1}],$$

where $Q = plim\frac{1}{N}X'X$. Effectively, the bias is “smeared” over to all other estimates. It affects not only the estimate for x_1, but to the extent x_1 is correlated with the other explanatory variables, the estimates for the other explanatory variables are affected as well, even though they are uncorrelated with the unobserved time-invariant factor α_i.

Endogeneity Bias Spreading in Multivariate Setting

The following illustrates bias spreading from endogenous to exogenous variables in a two-variable model. Consider the following true model:

$$y = x_1\beta_1 + x_2\beta_2 + q\gamma + e \tag{5.16}$$

Assume that the regressors x_1, x_2, and q are uncorrelated with the error term e, i.e., $plim\frac{1}{N}\ q'e = 0$ and $plim\frac{1}{N}\ x_j'e = 0$ for j=1, 2. Also assume that x_1 is uncorrelated with q, while x_2 is correlated with q. That is, $plim\frac{1}{N}\ x_1'q = 0, plim\frac{1}{N}\ x_2'q \neq 0$. Further, assume that q is unobserved and is omitted in the estimation. The estimating equation becomes

$$y = x_1\beta_1 + x_2\beta_2 + \eta, \tag{5.17}$$

where $\eta = q\gamma + e$

As such, x_1 is an exogenous regressor, while x_2 is endogenous.

The Frisch–Waugh–Lovell theorem states that coefficients from a multiple regression can be reconstructed from a series of bivariate regressions. Specifically, β_1 in the equation (5.17) above can be obtained by first regressing y on x_2 (step 1), then regressing x_1 on x_2 (step 2), and finally regressing the residuals from step one on residuals from step two (step 3).

Let us define the projection matrix P_2 and the residual-making matrix M_2 (aka annihilator matrix) as follows:

$$P_2 = x_2(x_2'x_2)^{-1}x_2{}'$$

$$M_2 = I - P_2, \text{ where}$$

I is an identity matrix. P_2 and M_2 are symmetric ($M_2 = M_2', P_2 = P_2{}'$) and idempotent ($P_2 = P_2P_2$, $M_2 = M_2M_2$), and $P_2x_2 = x_2$, $M_2x_2 = 0$ by construction (Hayashi 2000, 9).

Applying projection and annihilator matrices to estimating equation (5.17) yields representation of β_1 as a function of residuals from two bivariate regressions. To see this, multiply both sides of (5.17) by M_2:

$$M_2y = M_2x_1\beta_1 + M_2x_2\beta_2 + M_2\eta \tag{5.18}$$

Because $M_2x_2 = 0$ ($M_2x_2 =(I - x_2(x_2'x_2)^{-1}x_2{}'x_2 = x_2 - x_2(x_2'x_2)^{-1}x_2{}'x_2 = 0$), equation (5.18) becomes

$$M_2y = M_2x_1\beta_1 + M_2\eta \tag{5.19}$$

Redefining $M_2y = \tilde{y}$, $M_2x_1 = \tilde{x}$ and $M_2v = \tilde{\eta}$, equation (5.19) can be written as

$$\tilde{y} = \tilde{x}\beta_1 + \tilde{\eta} \tag{5.20}$$

Then

$$\widehat{\beta_1} = (\tilde{x}'\tilde{x})^{-1}(\tilde{x}'\tilde{y}) = (x_1'M_2{}'M_2x_1)^{-1}(x_1'M_2{}'M_2y) \tag{5.21}$$

Because M_2 is symmetric ($M_2 = M_2{}'$) and idempotent ($M_2 = M_2M_2$), (5.21) can be rewritten as:

$$\widehat{\beta_1} = (x_1'M_2x_1)^{-1}(x_1'M_2y) =$$

$$(x_1'M_2x_1)^{-1}(x_1'M_2(x_1\beta_1 + x_2\beta_2 + q\gamma + e\ ))=$$

$$(x_1'M_2x_1)^{-1}x_1'M_2x_1\beta_1 + (x_1'M_2x_1)^{-1}x_1'M_2x_2\beta_2 + (x_1'M_2x_1)^{-1}x_1'M_2(q\gamma + e\ ) \tag{5.22}$$

Since $M_2x_2 = 0$, second term becomes zero, and (5.22) is simplified to

$$\widehat{\beta_1} = \beta_1 + (x_1'M_2x_1)^{-1}x_1'M_2(q\gamma + e\ ) \tag{5.23}$$

$(x_1'M_2x_1)^{-1}x_1'M_2(q\gamma + e\ )$ is the "smeared bias" term. To derive the probability limit of this bias let us first simplify the $x_1'M_2(q\gamma + e\ )$ component:

$$x_1'M_2(q\gamma + e\ ) = x_1'(I - x_2(x_2'x_2)^{-1}x_2{}')(q\gamma + e\ ) =$$

$$x_1'(q\gamma + e\ ) - x_1{}'x_2\ (x_2'x_2)^{-1}x_2{}'\ q\gamma - x_1{}'x_2\ (x_2'x_2)^{-1}x_2{}'\ e =$$

$$\left(-\gamma\frac{cov_{x_1,x_2}cov_{x_2,q}}{V_{x_2}}\right) \tag{5.24}$$

Employing the exogeneity assumption on x_1 (i.e., $plim\frac{1}{\mathrm{N}}x_1'e = 0$ and $plim\frac{1}{N}\ x_1'q = 0$) and assumption $plim\frac{1}{N}\ x_2'e = 0$, the terms $x_1'(q\gamma + e\ )$ and $x_1{}'x_2\ (x_2'x_2)^{-1}x_2{}'e$ cancel out. $(x_2'x_2)^{-1}x_2'x_1$ is an OLS estimate from bivariate regression of x_1 on x_2, which equals $cov_{x_1,x_2}/V_{x_2}$.

Now, let us rewrite $x_1'M_2x_1$ as follows:

$$x_1'M_2x_1 = x_1'(I - x_2(x_2'x_2)^{-1}x_2')x_1 = x_1'x_1 - x_1'x_2(x_2'x_2)^{-1}x_2'x_1 =$$

$$V_{x_1} - \frac{cov_{x_1,x_2}^2}{V_{x_2}} = \frac{V_{x_1}V_{x_2} - cov_{x_1,x_2}^2}{V_{x_2}} \tag{5.25}$$

Combining (5.24) and (5.25), the asymptotic bias is equal to

$$plim(\widehat{\beta_1} - \beta_1) = \left(\frac{V_{x_1}V_{x_2} - cov_{x_1,x_2}^2}{V_{x_2}}\right)^{-1}\left(-\gamma\frac{cov_{x_1,x_2}cov_{x_2,q}}{V_{x_2}}\right) = -\gamma\frac{cov_{x_1,x_2}cov_{x_2,q}}{V_{x_1}V_{x_2} - cov_{x_1,x_2}^2}$$

This expression could be further simplified to aid interpretation. Since $cov_{x_1,x_2} = \rho_{x_1,x_2}\sigma_{x_1}\sigma_{x_2}$ and $cov_{x_2,q} = \rho_{q,x_2}\sigma_q\sigma_{x_2}$ (where ρ is correlation coefficient):

$$plim\left(\widehat{\beta_1} - \beta_1\right) = -\gamma\frac{\rho_{x_1,x_2}\sigma_{x_1}\sigma_{x_2}\,\rho_{q,x_2}\sigma_q\sigma_{x_2}}{V_{x_1}V_{x_2}\left(1-\rho^2_{x_1,x_2}\right)} = -\gamma\frac{\rho_{x_1,x_2}\rho_{q,x_2}\sigma_q}{\sigma_{x_1}(1-\rho^2_{x_1,x_2})} \tag{5.26}$$

This expression is generally non-zero. Hence, even though x_1 is exogenous, coefficient β_1 would still be biased as long as x_1 is correlated with the endogenous regressor x_2. The sign of the bias is determined by the signs of cov_{x_1,x_2}, $cov_{x_2,q}$ and γ. The magnitude of the bias is amplified when x_1 and x_2 are highly correlated. Only in the special case when x_1 and x_2 are orthogonal ($\rho_{x_1,x_2} = 0$), the bias equals to zero. Strict orthogonality of x_1 and x_2, however, almost never holds in economic settings.

Measurement Error Bias Spreading in Multivariate Setting

Smearing of the bias also occurs in multivariate regression in the case of measurement error. Consider a two-variable model where one of the regressors is mismeasured:

$$y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2 + e \tag{5.27}$$

where x_1 is measured with error and x_2 is measured without error. That is, we observe $x_1^* = x_1 + v$. If equation (5.27) is estimated by OLS then both estimates, $\widehat{\beta_1}$ and $\widehat{\beta_2}$, are biased and inconsistent (Greene 2017):

$$plim\,\widehat{\beta_1} \; = \beta_1\left(\frac{1}{1+\sigma_v^2\sigma^{11}}\right) \tag{5.28}$$

$$plim\,\widehat{\beta_2} \; = \beta_2 - \beta_1\left(\frac{\sigma_v^2\sigma^{12}}{1+\sigma_v^2\sigma^{11}}\right) \tag{5.29}$$

where σ^{ij} is the ij-th element of the inverse of the covariance matrix and σ_v^2 is the variance of the measurement error v.

$\widehat{\beta_1}$ is still subject to attenuation bias as in the bivariate case: the magnitude of the estimate is smaller than the true β_1. As long as x_1 and x_2 are correlated, the magnitude of the attenuation bias is greater in the multivariate setting than in the bivariate setting. The intuition for this result is that the additional variable x_2 in the regression will serve as a proxy for a part of the signal in the mismeasured regressor x_1. As such, the partial correlation between y and x_1 will be attenuated even more.

$\widehat{\beta_2}$ is biased and the direction of the bias can be either upward or downward, depending on the sign of β_1 and covariance between the two regressors.

CONCLUSION

Panel data allow researchers to design insightful models and control for the effects of unobservable factors. We advise caution and careful testing of alternative specifications before selecting models and estimators and suggest steps to avoid common errors in panel data modeling. Misspecification can lead to significant biases and erroneous conclusions about the economic effects of marketing or public policy activities.

REFERENCES

Anderson, Theodore Wilbur and Cheng Hsiao (1981), “Estimation of Dynamic Models with Error Components,” *Journal of the American Statistical Association,* 76 (January), 598–606.

Angrist, Joshua D. and Jörn-Steffen Pischke (2008), *Mostly Harmless Econometrics: An Empiricist’s Companion*. Princeton, NJ: Princeton University Press.

Arellano, Manuel and Stephen Bond (1991), “Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations,” *Review of Economic Studies*, 58(2), 277–297.

Arellano, Manuel and Olympia Bover (1995), “Another look at the instrumental variable estimation of error-components models,” *Journal of Econometrics*, 68(1), 29–51.

Baltagi, Badi (2005), *Econometric Analysis of Panel Data.* New York: John Wiley & Sons.

Biørn, Erik (2000), “Panel Data with Measurement Errors: Instrumental Variables and GMM Procedures Combining Levels and Differences,” *Econometric Reviews*, 19(4), 391–424.

Blundell, Richard and Stephen Bond (1998), “Initial Conditions and Moment Restrictions in Dynamic Panel Data Models,” *Journal of Econometrics*, 87(1), 115–143.

Cameron, A. Colin and Pravin K. Trivedi (2005), *Microeconometrics: Methods and Applications*. New York: Cambridge University Press.

Cameron, A. Colin and Pravin K. Trivedi (2009), *Microeconometrics Using Stata* (Vol. 5). College Station, TX: Stata Press.

Chamberlain, Gary (1984), “Panel Data,” in Z. Griliches and M. Intriligator, eds., *Handbook of Econometrics*. Amsterdam: North Holland, 1247–1318.

Greene, William (2017), Econometric Analysis. Lecture notes. http://people.stern.nyu.edu/wgreene/Econometrics/Econometrics-I-13.pdf

Griliches, Zvi and Jerry A. Hausman (1986), “Errors in Variables in Panel Data,” *Journal of Econometrics*, 31(1), 93–118.

Hansen, Lars Peter (1982), “Large Sample Properties of Generalized Method of Moments Estimators,” *Econometrica: Journal of the Econometric Society*, 50(4), 1029–1054.

Hausman, Jerry A. (1978), “Specification Tests in Econometrics,” *Econometrica* 46 (November), 1251–1271.

Hausman, Jerry A. and William E. Taylor (1981), “Panel Data and Unobservable Individual Effects,” *Econometrica*, 49(6), 1377–1398.

Hsiao, Cheng (2014), *Analysis of Panel Data*, Cambridge: Cambridge University Press. 3rd edition.

Jacobson, Robert (1990), “Unobservable Effects and Business Performance,” *Marketing Science*, 9 (Winter), 74–85, 92–95.

Kirzner, Israel M. (1976), “On the Method of Austrian Economics,” in E.G. Dolan, ed., *The Foundations of Modern Austrian Economics,* Kansas City: Sheed and Ward, 40–51.

Mizik, Natalie and Robert Jacobson (2004), “Are Physicians ‘Easy Marks’? Quantifying the Effects of Detailing and Sampling on New Prescriptions,” *Management Science*, 1704–1715.

Mundlak, Yair (1978), “On the Pooling of Time Series and Cross-Sectional Data,” *Econometrica*, 46 (January), 69–86.

Nickell, Stephen (1981), “Biases in Dynamic Models with Fixed Effects,” *Econometrica*, 1417–1426.

Pischke, Jörn-Steffen (2007), Lecture notes on measurement error. London School of Economics. http://econ.lse.ac.uk/staff/spischke/ec524/Merr_new.pdf.

Roodman, David (2009), “How to do xtabond2: An introduction to difference and system GMM in Stata,” *Stata Journal*, 9 (1), 86–136.

Rumelt, Richard (1984), “Towards a Strategic Theory of the Firm,” in B. Lamb (ed.), *Competitive Strategic Management*, Englewood Cliffs, NJ: Prentice Hall, 556–570.

Sargan, John D. (1958), “The estimation of economic relationships using instrumental variables,” *Econometrica: Journal of the Econometric Society*, 393–415.

Trognon, Alain (1978), “Miscellaneous Asymptotic Properties of Ordinary Least Squares and Maximum Likelihood Estimators in Dynamic Error Components Models,” *Annales de l'INSEE*. Institut National de la Statistique et des Études Économiques, p. 631–657

Wernerfelt, Birger (1984), “A Resource-based View of the Firm,” *Strategic Management Journal*, 5 (April–June), 171–180.

Windmeijer, Frank (2005), "A Finite Sample Correction for the Variance of Linear Efficient Two-step GMM Estimators", *Journal of Econometrics*, 126(1), 25-51.

Wooldridge, Jeffrey (2002), *Econometric Analysis of Cross Section and Panel Data*, Cambridge, MA: MIT Press.

Wooldridge, Jeffrey (2006), *Introductory Econometrics: A Modern Approach*, Mason, OH: Thomson/South-Western.