7

Repeated Measures and Multilevel Modeling

Every year, the international data infrastructure comes to include data for more countries for more years. Previously existing data never disappear. The inexorable temporal and spatial expansion of the international data infrastructure has inevitably raised the question of what to do with all the extra years of data. The obvious answer is to use extra observations of the same variables for the same countries as additional cases for analysis, though other answers are possible (e.g., using additional years to produce ever-more-robust period averages and ever-longer time lags). With new observations forthcoming every year for pretty much every country for which data are already available, new analyses are always possible (and publishable). Science progresses and careers are built.

Model designs that make use of vertical data structures in which the same countries appear multiple times in the same database are known as repeated measures designs. With repeated measures designs it is possible to study multiple examples of change over time, contemporaneous (or lagged) movements in variables across time and geography, or (under certain conditions) simply more cases of the same underlying phenomena. There is a danger, however, that in using repeated measures to create additional cases for analysis, the additional observations of the same country at multiple points in time are not really additional cases in the sense of new, independent realizations of underlying quantitative macro-comparative research (QMCR) data-generating processes. Along these lines, Kittel (1999) makes a distinction between observations and cases because it might reasonably be questioned to what extent (say) France in 1996 was a different analytical case from France in 1997.

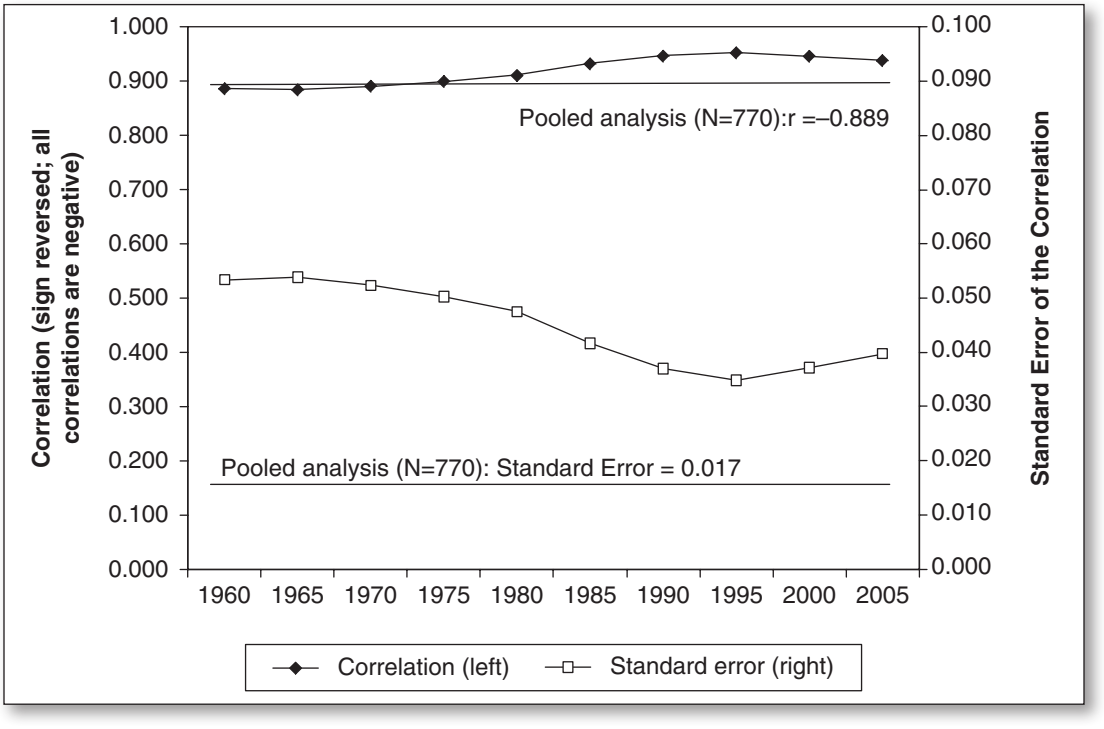
Unreflexively treating each additional time observation as a new case leads to the *reductio ad absurdum* that it is possible to generate an infinite number of cases just by slicing time into thinner and thinner units:

> For example, consider investigating the effect of regime type on the provision of public goods with data on 20 countries. Suppose now that we obtain 20 years of data for these countries . . . is this data inflation really legitimate? Why not take monthly observations for each of these 20 countries, then we would have 4800 data points, and surely all our estimates would be statistically significant. . . . In short, if we have 20 countries in our data set, we have 20 countries, not 400 [20 countries times 20 years]. . . . This topic has received little attention in the literature. (Wilson and Butler 2007:108)

On the other hand, as Figure 1.4 illustrated, the same argument could be made for Guatemala 2000 versus Honduras 2000 as for France 1996 versus France 1997; in other words, the proliferation of cases by year is just a form of compositional interdependence. The problem, however, is more severe for repeated observations of the same country over time than for observations of different but related countries. All QMCR data are susceptible to suspicions of case inflation due to compositional interdependence, but repeated measures data are especially and explicitly susceptible.

This is well illustrated by the cross-national relationship between national income and infant mortality. Babones (2009c) reports a range of correlations between infant mortality and national income measured over a 45-year period (10 time points). These are replicated in Figure 7.1. Presumably due to improved measurement, the correlation has slowly increased in magnitude over the years, from r = –0.887 in 1960 to r = –0.939 in 2005 (the relationship between infant mortality and poverty is tightening). The standard error of the correlation has slowly declined from 0.053 to 0.040. Pooling all 10 time points together into a single analysis yields N = 770 observations (77 countries by 10 time points). This has no real effect on the correlation; the pooled correlation

Figure 7.1 Correlations and Standard Errors Between Infant Mortality (logged) and GDP per Capita (logged), 1960–2005



Source: After Babones (2009c:93).

Note: Constant Panel of **N** = 77 Countries.

(r = –0.889) falls within the range of the observed correlations for the 10 individual time points. On the other hand, it has a dramatic effect on the standard error; the pooled standard error (0.017) is much lower than any of the 10 original standard errors. As this example illustrates, parameters estimated using repeated measures data are highly susceptible to downward biases in their standard errors (and thus inflated statistical significance).

Scenarios like the one laid out in Figure 7.1 are easily handled by the statistical tools that have been developed for repeated measures models (described below), but researchers have to know about these tools and use them properly for them to make any difference. In a review of 195 papers from the political science literature, Wilson and Butler (2007:100) found that only seven met what they considered "basic criteria" for diagnosing and treating common problems with repeated measures designs. It might reasonably be argued that Wilson and Butler's "basic criteria" are very advanced indeed, but Wilson and Butler found that over 20% of the papers they studied did nothing whatsoever to address the kinds of errors portrayed in Figure 7.1. This result is especially shocking given the fact that Wilson and Butler's study universe consisted entirely of relatively sophisticated papers that had cited either Beck and Katz (1995) or Beck and Katz (1996), methodological contributions that explicitly warned of the necessity of correcting standard errors.

Although all repeated measures designs share some features in common, there are, broadly speaking, two common scenarios for the use of repeated measures data. Time series cross-sectional (TSCS) designs make use of relatively large numbers of time points (T) for relatively small numbers of countries (N) so that T is (usually) much greater than N. Multilevel modeling (MLM) designs—also called hierarchical linear model (HLM) designs—make use of relatively small numbers of time points (T) for relatively large numbers of countries (N) so that N is (usually) much greater than T. The set of countries included in a repeated measures database is known as the panel, so both methods (though MLM more often than TSCS designs) are referred to collectively as methods for the analysis of panel data, or panel designs.

The TSCS approach, as its name implies, puts greater emphasis on the cross-sectional variability between countries, while the MLM approach puts greater emphasis on the over-time variability within countries. That said, it must be emphasized that the two types of models share a common statistical toolkit. The difference between TSCS and MLM designs is methodological, not statistical: identical statistical tools are used in each, just with different frequency for different purposes, and there is no firm line between the two. The first section below lays out some of the common issues that arise from the use of repeated measures data in any model design. The second section focuses (briefly) on the methodological literature around TSCS designs in econometrics and political science, much of which applies equally well to the MLM designs used in the social sciences more broadly. The third section focuses in more depth on MLM designs, which are more common in QMCR proper, addressing in particular the issue of fixed versus random effects. The chapter concludes with an evaluation of the risks and benefits associated with the use of repeated measures data and suggests a new way forward, the slope-slope model.

THE STRUCTURE OF REPEATED MEASURES DATA

Repeated measures QMCR data are organized into vertical database structures (Figure 4.3) in which each country appears at least once—but usually more than once—as a database row. Each observation of a country is usually keyed to a year, though it need not be: repeated measures can just as well be for months, quarters, 5- or 10-year intervals, and so on. In practical application, time units below 1 year are not used in QMCR outside economics due to a lack of sub-annual data for most variables of interest, while time units greater than 1 year are rarely used because period-averaging reduces the (nominal) number of observations available for analysis. As a result, the year is the standard unit. Observations, however, are not necessarily annual; sometimes (as in Figure 7.1) year-specific data are used in 5- or 10-year intervals.

Repeated measures databases may be balanced or unbalanced. Database searches of the QMCR literature suggest that unbalanced panels are actually more common than balanced panels in QMCR. Unfortunately, the properties of repeated measures models based on unbalanced panels are not well established, since nearly all methodological research on the properties of repeated measures models presumes that the panels are balanced. As a result, it is not obvious to what extent the results of studies based on unbalanced panels of countries are valid. Moreover, many of the tools available for analyzing repeated measures data are only applicable to balanced panels. This is not to say that unbalanced panels cannot be studied, but rather that unbalanced panel designs have not been adequately studied and so are (by comparison with balanced panel designs) poorly understood.

The fundamental challenge of working with repeated measures data, however, arises from their very nature as multiple observations of the same units. Classical statistical methods are built on the assumption that each observation varies independently of all other observations. With repeated measures data, this is clearly not the case. While it can be argued that cross-sectional data are riddled with complex dependence structures, with repeated measures data there is no argument: repeated observations of the same country over time are not statistically independent cases for analysis. When cases are not independent, regression results (both coefficients and their standard errors) can become seriously biased. Some typical problems of error dependence in repeated measures models and commonly used solutions are discussed in this section.

The Problem of Nonspherical Errors

The main difficulty with repeated measures data is that models based on them routinely violate the spherical errors assumption that underlies the estimation of regression coefficients using ordinary least squares (OLS). The spherical errors assumption in words roughly translates as “the errors look the same from every direction”: the error associated with every case used in a regression analysis is independent of the error associated with every other case (there is no clustering in the errors) and the error associated with every case has the same variance as the error associated with every other case (the errors are homoskedastic). The data underlying Figure 7.1 fail on both counts. First, there is severe clustering by country in the size of the regression errors. Out of 77 countries, 9 have errors that are positive for at least 9 out of 10 time points and 16 have errors that are negative for

at least 9 out of 10 time points. Second, there is noticeable clumping by year in the variability of the regression errors. The standard deviation of the errors ranges from a low of 0.145 for the 1990 observations to a high of 0.191 for the 2005 observations and varies widely across countries.

These two violations of error sphericity—clustering and heteroskedasticity—have the potential to wreak havoc on regression models. Error clustering can be thought of as error mean dependence: the mean level of the error is not independent across observations. Wilson and Butler (2007) delineate two kinds of error mean dependence that are especially common in QMCR: country unit heterogeneity (errors are similar for all observations of the same country) and dynamic dependence over time (errors for one year are related to errors for the next year for the same country). Though not included in Wilson and Butler's framework, both of these kinds of country-oriented mean dependence potentially have time-oriented counterparts: unit heterogeneity among years (errors are similar for all observations of the same year) and dynamic dependence among countries (errors for one country are related to errors for neighboring countries in the same year, i.e., spatial dependence). Other, more exotic forms of mean dependence are also possible.

Heteroskedasticity can be thought of as error variance dependence. It often arises in repeated measures models when observations for different countries or different time periods (or both) have different error variances. This is in addition to the heteroskedasticity that might exist, say, between rich and poor countries. Error variance dependence is much less damaging than mean dependence because it usually causes no biases in the coefficients themselves, though it does affect their standard errors. The general consensus is that it has to be severe before it becomes a concern. When the variables being used are reasonably well distributed or properly transformed (Chapter 3) heteroskedasticity on its own is unlikely to have a major effect on results.

These various kinds of mean and variance dependence are summarized in Table 7.1, along with some of the tools that are commonly used to address them.

Table 7.1 Error Sphericity Violations and Associated Tools for Addressing Them

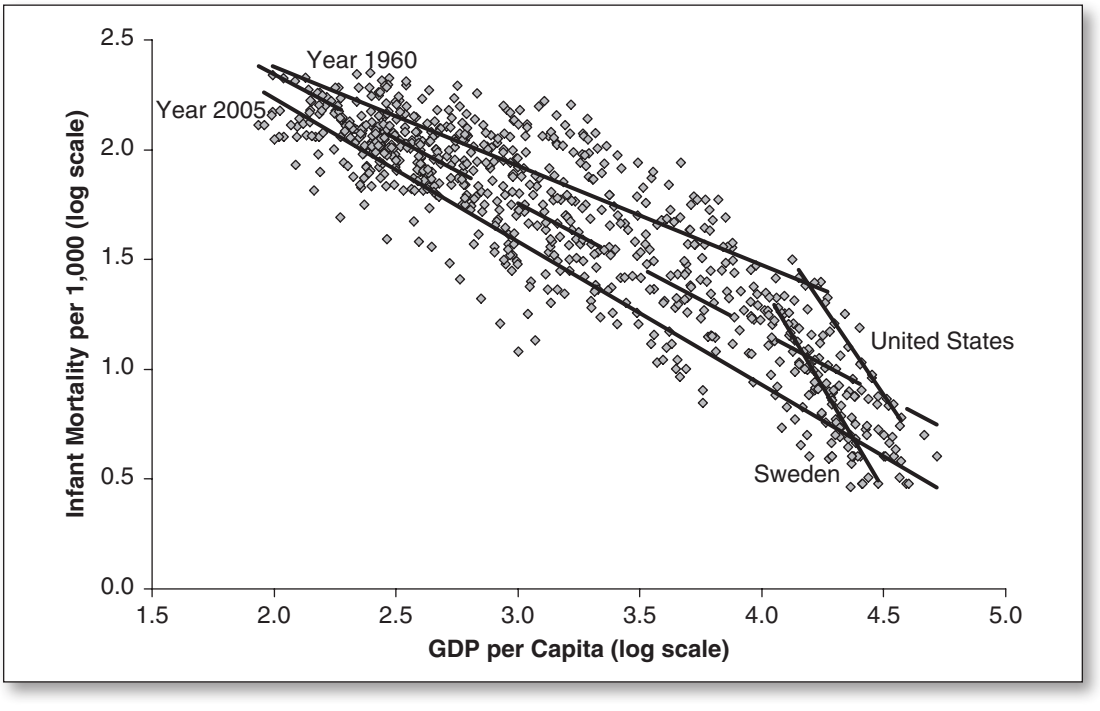
Sphericity violation	*Dependence structure*	*Associated tools*
Mean dependence	Unit heterogeneity across time units	Multilevel modeling with fixed effects for time
	Unit heterogeneity across countries	Multilevel modeling with fixed effects for country
	Dynamic dependence for adjacent time units	AR(1) error terms and/or lagged dependent vars.
	Dynamic dependence for adjacent countries	Panel corrected standard errors (PCSE)
Variance dependence (heteroskedasticity)	Differences in variance across time units	Generally not corrected—assumed not to exist
	Differences in variance across countries	Panel corrected standard errors (PCSE)

Correcting for Mean Dependence

Straightforward examples of unit heterogeneity in which the mean value of a dependent variable changes over time or across countries are easily addressed using dummy variables for countries and time points in a fixed effects regression model. Figure 7.2 illustrates the need for such adjustments using the underlying data from Figure 7.1. Figure 7.2 plots infant mortality against national income per capita for 77 countries at 10 five-year time increments, for a total of 770 observations. Regression lines have been plotted for the relationship between national income and infant mortality in 1960 versus 2005 and for the 10 observations of the United States over time versus the 10 observations for Sweden. The dashed line represents the OLS regression line for all 770 pooled observations. In the absence of unit heterogeneity, the lines for 1960, 2005, the United States, and Sweden should all (roughly) coincide with the overall pooled regression line. Clearly they do not.

A common correction for unit heterogeneity is the use of country fixed effects. Country fixed effects effectively center the data on a country-wise basis, eliminating cross-national differences in the levels of variables while retaining their over-time variability. Figure 7.3 plots country-centered infant mortality rates versus country-centered national income

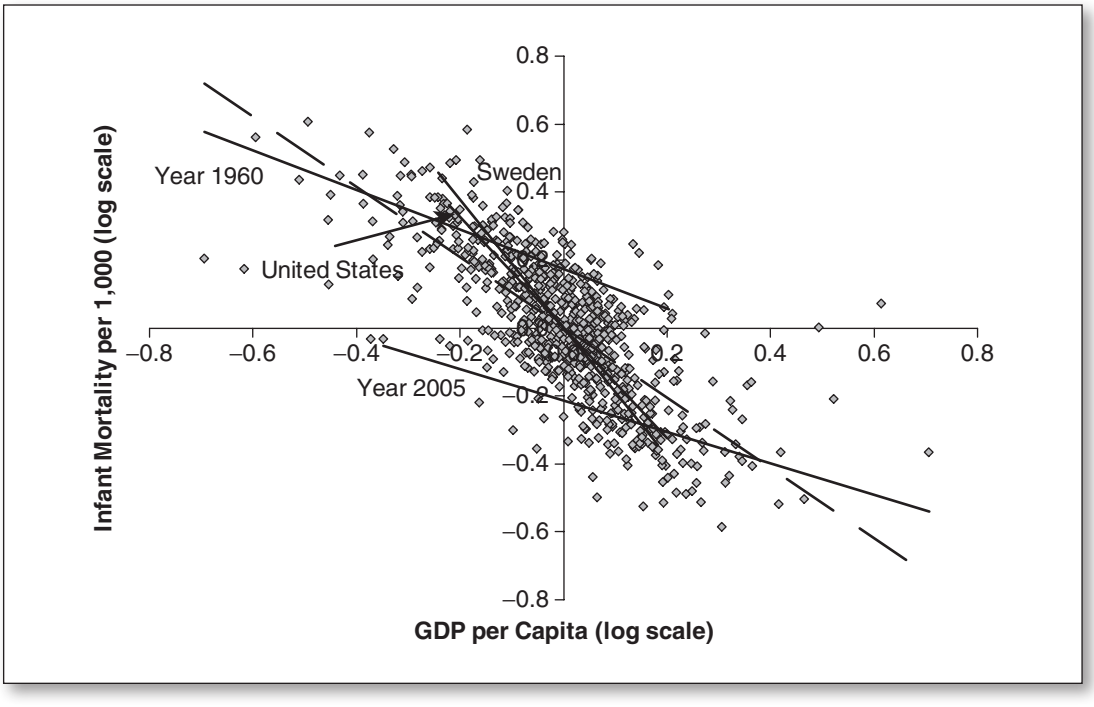
Figure 7.2 Relationship Between National Income and Infant Mortality



levels for the 770 observations from Figure 7.2. As in Figure 7.2, illustrative regression lines for 1960, 2005, U.S., and Swedish data are depicted. Adjusting for the fixed effects of country has removed essentially all of the country mean dependence, with the U.S. and Swedish regression lines falling very close to the pooled regression line and crossing it at the origin. It is also now very clear that mean dependence by year exists as well, with the 1960 and 2005 regression lines falling almost parallel, separated by a fixed distance between them. This could be corrected through the inclusion of period fixed effects alongside the country fixed effects.

The second form of mean dependence, dynamic dependence, is much more subtle than unit heterogeneity. Dynamic dependence occurs when regression errors are correlated in a daisy-chained manner instead of being correlated for fixed groupings of observations. The two most obvious forms of dynamic dependence are temporal autocorrelation and spatial autocorrelation. In temporal autocorrelation, the error for each time period depends on the error for the previous one; in spatial autocorrelation, the error for each country depends on the errors of its neighbors. In the simplest scenario, these autocorrelated errors are Markovian, meaning that the error for each observation depends only on that of the observation immediately preceding it. In more-complex scenarios, the errors may be structured in multistep patterns, as with business cycles that span multiple years.

Figure 7.3 Relationship Between National Income and Residual Infant Mortality After Adjusting for Country Fixed Effects



Straightforward temporal error autocorrelation that is both Markovian and linear can be addressed in one of three ways: through the use of a lagged dependent variable, through the use of differenced variables, or through the use of explicit autoregressive error modeling. The first two approaches can be implemented using OLS, while the last requires the use of generalized least squares (GLS) estimation techniques. More-complicated forms of error autocorrelation also exist that require the use of GLS techniques from econometric time series analysis that are rarely used in QMCR.

Beck and Katz (1996) strongly recommend the use of OLS regression with a lagged dependent variable to address error autocorrelation. In this approach, the dependent variable measured at period t is regressed on the other regressors (including any fixed effects) plus the value of the dependent variable for the same case for the period t–1. This is similar to the lagged dependent variable approach to establishing nonspuriousness discussed in Chapter 6, but in repeated measures settings the other independent variables are measured contemporaneously with the dependent variable, not with its earlier manifestation. That is to say, the form of the model is,

$$Y_t = B_0 + B_1 * X1_t + B_2 * X2_t \ldots + B_i * Xi_t + B_{i+1} * Y_{t-lag} + \text{Error},$$

where i represents the series of independent variables and t represents the period. The lagged dependent variable serves a different purpose in repeated measures designs than it does in cross-sectional designs. It is not there to eliminate common-cause variables but to eliminate the autocorrelation of errors among the repeated observations of the same country. Keele and Kelly (2006) caution, however, that the inclusion of a lagged dependent variable does not always eliminate error autocorrelation, and that explicit testing for residual error autocorrelation is necessary, while Maddala (1998) warns researchers off the use of lagged independent variables entirely. Finally, Keele and Kelly (2006) warn that lagged dependent variables should never be used with cyclical data, since lagged dependent variables fail to capture any non-Markovian dependence.

Regression based on differenced variables (change scores) should in principle remove error autocorrelation as well, but change scores are rarely used for this purpose outside economics (see Stuckler 2008 for a rare example). The change score approach to the analysis of repeated measures data is identical in logic to the difference model approach to controlling for time-invariant common-cause variables, with the distinction that in the change score approach differences are calculated for multiple time periods. Thus, instead of regressing a one-time change in the dependent variable on a one-time change in the independent variables, the dependent variable change in every period is regressed on the independent variables change in that period. As with the difference model approach to causality, one suspects that these models are rarely used because of their low statistical power. Difference and change score designs are highly robust, but due to low power they rarely produce significant results.

Explicit autoregressive modeling is also used, especially in the political science literature. In this approach, the dynamic dependence structures in the regression error are explicitly modeled instead of being adjusted away through creative model design. The most commonly used autoregressive error model is the AR(1), or 1-period autoregressive

model. Like the lagged dependent variable and change score approaches, the AR(1) model assumes that the error dependence structure is Markovian. Multi-period autoregressive models, however, can overcome this assumption. For example, in the AR(2) model each year's error term is assumed to depend on the prior 2 years' errors. The AR(2) model is very flexible, and can accommodate simple business cycles. Econometrics textbooks provide extensive material on autoregressive error modeling.

None of these methods addresses spatial autocorrelation: the correlation of errors for neighboring countries. The panel corrected standard errors (PCSE) technique developed by Beck and Katz (1995) to deal with differences in error variance across countries has the fortuitous side effect that it also adjusts for any pattern of mean dependence that is static across time periods.As a result, it corrects for the spatial autocorrelation of errors—as long as the patterns of spatial autocorrelation are stable over time. Beck et al. (2006), however, argue that instead of being adjusted away, spatial effects should be explicitly modeled. They take an expansive view of the spatial, including, in spatial modeling, not just geographical proximity but social and economic proximity as well (e.g., connection through trade networks). The spatial econometrics techniques they describe are not yet widely used in QMCR but may be in the future.

There is no firm consensus in the methodological literature around which of these techniques is superior, nor around which should be used in what particular situations. Table 7.1, however, does present some general guidance. Unit heterogeneity can be addressed by fixed effects for country (and, if necessary, for time), while dynamic dependence can be addressed using AR(1) error models. As a result, many studies use fixed effects in combination with AR(1) error models. Other studies instead use lagged dependent variables to address both unit heterogeneity and dynamic dependence at the same time. Nearly all TSCS studies use PCSE adjustments. It is not clear, however, that these fixes solve the underlying problems. For example, Babones (2009c) demonstrates extensive residual unit heterogeneity in models that implement both fixed effects and AR(1) error models. The problem is that dependence structures tend to be much more complicated than these straightforward fixes assume.

Correcting for Variance Dependence

Like mean dependence, variance dependence can be structured either along country or time dimensions—or both. The main danger in practice, though, is that error variances may be distinctive of particular countries. If each country has its own error variance, which is different from that of other countries, the repeated observations of each country will form a cluster. When T is much greater than N, such country-wise clustering can become the dominant feature of the error variance.

For many years, the standard treatment for country-wise heteroskedasticity was a method attributed to Parks (1967) called feasible generalized least squares (FGLS). Beck and Katz (1995), however, famously showed that in typical research settings the FGLS method deflated the estimated standard errors of coefficients by anywhere from 30% to 75%. Since the publication of Beck and Katz (1995), most TSCS studies have used the alternative method promoted by Beck and Katz, OLS regression with PCSE adjustments.

The PCSE technique is based on an assumption that any patterns of error dependence that exist in the data are the same for every time point. This allows the pooling of the errors from all time points to estimate the true variance-covariance matrix of all the data points and, ultimately, correct standard errors for the regression coefficients.

Since it relies on pooling over T, the PCSE technique becomes more efficient as the number of time points rises. Beck and Katz (1996) show that it can be used effectively even when the number of time points is "small"—in their examples, as few as five. With such few repeated observations, however, the problems that the PCSE technique is designed to correct are usually not serious, and PCSE adjustments often produce no change in the interpretation of results. In practice, PCSE adjustments are usually made only in TSCS settings in which T>N and are rarely used in MLM settings in which N>T.

One great advantage of the PCSE technique is that it not only adjusts for country-wise variance dependence, but also for country-wise mean dependence (as long as both are constant over time and thus poolable). As discussed above, this includes all kinds of neighbor effects, which are surely ubiquitous in QMCR data. The great shortcoming of the PCSE technique is that it does nothing to correct for time-wise mean and variance dependence. Moreover, as Beck and Katz caution throughout their work, the PCSE technique can only be applied to data that do not exhibit error autocorrelation over time, since it is based on the assumption that the collection of data for each time period represents independent realizations of a single underlying error process. In their review of the political science literature citing the work of Beck and Katz, Wilson and Butler (2007) find that most authors do not even consider the possibility that this assumption may not be met. Though crediting Beck and Katz with important contributions, they argue that "the problems researchers tend to ignore are far more serious than the problems corrected with the PCSEs" (Wilson and Butler 2007:110).

One such problem (not even considered by Wilson and Butler) is time-wise variance dependence. This is illustrated in Figure 7.4. Figure 7.4 plots the residuals from the regression of infant mortality on national income, controlling for year and country fixed effects. Obviously, the error variance is lowest in the middle years and highest at the beginning and end of the time period. The temporal heteroskedasticity displayed in Figure 7.4 becomes less systematic, but does not disappear entirely, when a lagged dependent variable is included in the model. With the inclusion of the lagged dependent variable, however, the coefficient for national income in the regression model is reduced to near zero. In other words, once most (but not all) of the potential violations of error sphericity have been addressed, there is no relationship left to study, despite the fact that the cross-sectional correlation between national income and infant mortality is on the order of $r = 0.9$.

TIME SERIES CROSS-SECTIONAL MODELS

Model designs that use repeated observations of QMCR data to focus on the cross-sectional aspects of data observed at multiple points in time are called time-series cross-sectional (TSCS) models. In a typical TSCS design, a regression model is estimated using data from a small number of countries for which data are available at annual increments for