MODELING ERROR IN QUANTITATIVE MACRO-COMPARATIVE RESEARCH

Salvatore J. Babones
Department of Sociology and Social Policy
The University of Sydney, Australia
sbabones@inbox.com

ABSTRACT

Much quantitative macro-comparative research (QMCR) relies on a common set of published data sources to answer similar research questions using a limited number of statistical tools. Since all researchers have access to much the same data, one might expect quick convergence of opinion on most topics. In reality, of course, differences of opinion abound and persist. Many of these differences can be traced, implicitly or explicitly, to the different ways researchers choose to model error in their analyses. Much careful attention has been paid in the political science literature to the error structures characteristic of time series cross-sectional (TSCE) data, but much less attention has been paid to the modeling of error in broadly cross-national research involving large panels of countries observed at limited numbers of time points. Here, and especially in the sociology literature, multilevel modeling has become a hegemonic – but often poorly understood – research tool. I argue that widely-used types of multilevel models, commonly known as fixed effects models (FEMs) and random effects models (REMs), can produce wildly spurious results when applied to trended data due to mis-specification of error. I suggest that in most commonly-encountered scenarios, difference models are more appropriate for use in QMC.

INTRODUCTION

Quantitative macro-comparative research (QMCR) involves the statistical analysis of quantitative data about countries. Undertaking QMCR is at the same time both much simpler and much more difficult than engaging in other kinds of social research. It is simple because it often involves the application of off-the-shelf statistical techniques to widely-available published data; a typical study can be completed in a few months from start to finish by a lone researcher working in isolation on a desktop computer. It is difficult for the same reasons: low barriers to entry mean that almost every combination of variables that can be studied has been studied. Moreover, we generally only get one additional year of new data points every year, and worse, each new year's data points look much like those from the year before. Progress in most branches of social research relies on the constant generation of new data – both new cases and new variables – to drive forward both research and theory, but QMCR practitioners are doomed to pick over the same well-studied datasets again and again. New ideas arise all the time, but all too often there

Copyright ©2009, American Sociological Association, Volume XV, Number 1, Pages 86-114
ISSN 1076-156X

are no new data on which to test them. Instead of filling old bottles with new wine, we fill new bottles with the same old wine year after year.

A large and recurring area of controversy in QMCR is the appropriate modeling of error. Error in a statistical model is the sum total of all other factors not explicitly accounted for in the model. It is reflected in the degree to which the observed values of dependent variables differ from their predicted values. Every statistical model includes implicit or explicit assumptions about the sources, distributions, and structures – in short, the behavior – of error. Misspecification of the error in a statistical model can lead to reported coefficients that are lower or higher than the true effects they are intended measure (biases). It can also lead to reported standard errors for coefficients that are lower or higher than they should be (resulting in overconfidence or underconfidence in results). I suspect that most researchers spend much more time thinking about what variables to include in their models than about how to specify the error structures of their models. This is a mistake. Error assumptions can have enormous impacts on statistical results and their interpretation. Moreover, even the choice of variables for inclusion in a model can be thought of as a form of error modeling. We must pay greater attention to the role played by (statistical) error in QMCR.

Toward that end, this paper is meant to serve as a guide to (and critique of) the treatment of error in QMCR. Though the mathematics underlying the statistical models used in QMCR are well-understood (by the statistical software writers, if not by the statistical software users), the implications of applying these models in typical QMCR settings are not. Unfortunately, most methodological guidebooks rely heavily on mathematical, rather than verbal, explication, which leaves a major gap in most researchers' understandings of the implications of error modeling. After all, it is usually not the mathematics but the appropriate verbal expression of the mathematics that is at the heart of both substantive and methodological debates in QMCR. Accordingly, this paper focuses on methods, not mathematics.

Throughout the paper, concepts are illustrated using data on the relationship between national income per capita and infant mortality rates as an applied example. The national income and infant mortality data used are taken from the World Bank's World Development Indicators 2007 database. National income is operationalized as gross domestic product per capita evaluated at market foreign exchange rates. Both series have been logged to correct for positive skew. The correlation between national income and infant mortality for 167 countries for 2005 is r = -.89. As would be expected, there is a very close connection between the two variables: higher levels of national income per capita are associated with lower levels of infant mortality. It should be noted that even in the 2005 panel, 41 countries, or almost a fifth of the world's total, are missing data. For earlier years, levels of missing data are of course higher.

The remainder of this paper is divided into seven sections. I begin by discussing the role played by error in statistical models and the consequent relevance of significance testing in QMCR. I then review typical dependence structures in that error and ways that control variables can be used to model those structures. This discussion of control variables leads into a wider discussion of causality in QMCR. Appropriate error modeling is central to the credibility of causal claims. Model results are rarely if ever problematic in their own right; controversies only arise when coefficients are interpreted causally. An important class of models that I believe have regularly given rise to inappropriate causal claims is that of multilevel models (MLMs), which include what are often called fixed effects models (FEMs) and random effects models (REMs). I devote a full section to these models, then a second section to comparing and contrasting them to a competing class of models, difference models. I conclude with recommendations for sound methodology in QMCR, particularly concerning the treatment of time.

THE RELEVANCE OF SIGNIFICANCE TESTING

At the core of the modern practice of QMCR is the determination of the statistical significance of independent variables when used in modeling dependent variables of interest. Many of the hottest debates in QMCR hinge not on the magnitudes of the effects of independent variables, but on the question of whether or not such effects could have arisen by chance. Typically, QMCR operates within a Neyman-Pearson framework of testing null hypotheses that independent variables are unrelated to dependent variables; when these null hypotheses are rejected, the relationships between independent and dependent variables are inferred to be statistically significant.

Many scholars argue that it is inappropriate to make statistical inferences in QMCR settings, though few make such arguments in print (Berk 2004:51-56 is a notable exception). Occasionally, even macro-comparative researchers repeat this claim (e.g., Ebbinghaus 2005). Their argument, in a nutshell, is that since the data used in QMCR usually constitute entire populations of cases, rather than random samples from larger populations, there is nothing to make inferences about. They argue that the parameters estimated in QMCR are not sample estimates of population parameters but are themselves population parameters, and thus not subject to error. A regression line, in this view, represents nothing more than the mean of the dependent variable when conditioned on the independent variable. They argue that it is inappropriate to say that the slope of this line is "significantly" different from zero: the slope is what it is, but with no further implications. To claim otherwise, they argue, is to posit that there exists some "imaginary superpopulation" (Berk 2004:51) of countries that exhibits a true, population regression line, and that the observed population of countries is merely one of many possible samples of this superpopulation. The argument that statistical significance testing is inappropriate when the underlying data constitute a population is completely without merit.

Key to the argument that statistical inference is inappropriate when analyzing populations is the (mis-) conceptualization of regression coefficients as sample estimates of population parameters. They are not. Regression coefficients are random variables, but not sample estimates. Confusion arises because sample estimates of population parameters are random variables, but there are many other kinds of random variables. According to the *Cambridge Dictionary of Statistics*, a random variable is a "variable, the values of which occur according to some specified probability distribution" (Everitt 2002:313). So, for example, the sample mean of a variable randomly drawn from a larger population of cases for which that variable has been measured is known to follow a Normal distribution with mean equal to the population mean and variance equal to the population variance divided by the size of the sample. The sample mean of a variable won't always equal the population mean, but it will vary around it normally, regardless of the underlying distribution of the variable. It is a random variable – it is "a variable, the values of which occur according to some specified probability distribution" – in this case, a Normal distribution of known parameters.

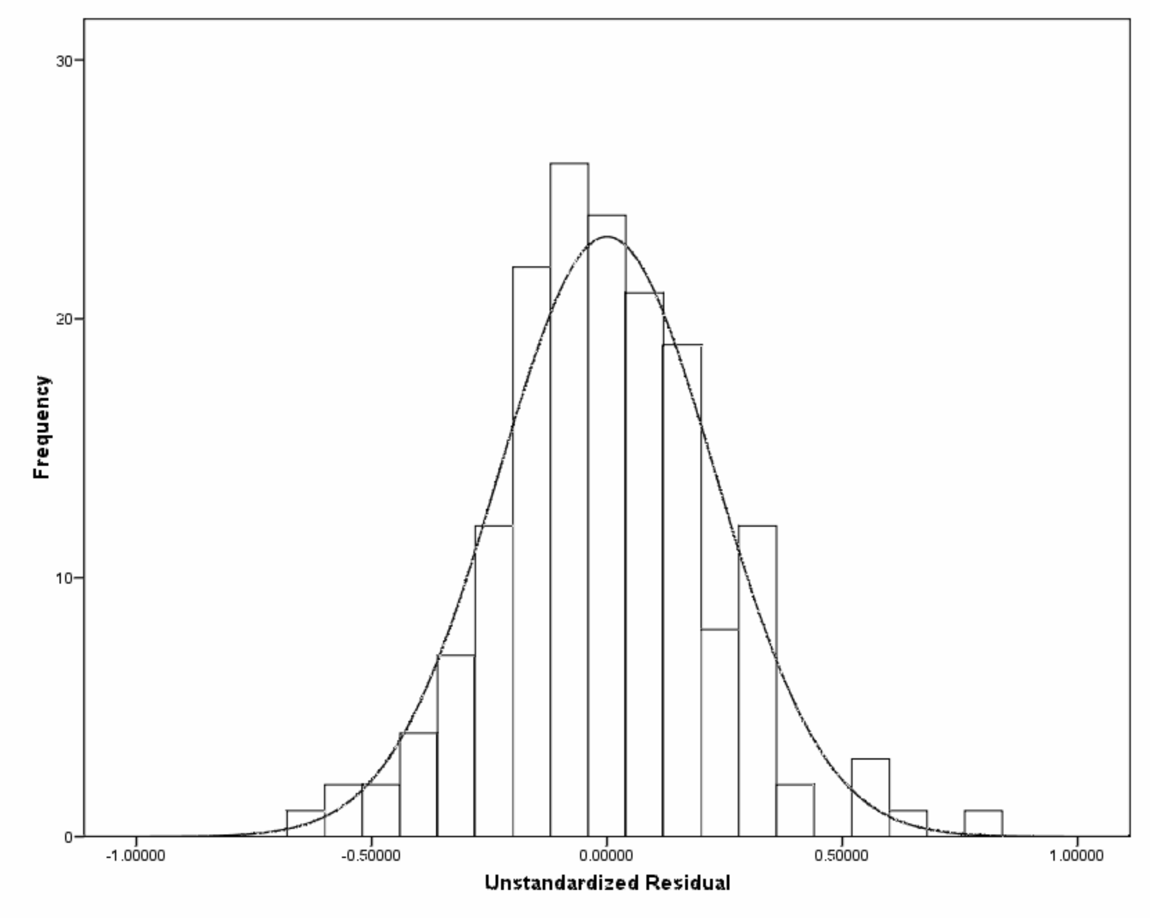
The dependent and independent variables in regression analyses are emphatically NOT random variables. In everyday language they may be labeled as such, and even some standard reference sources do not distinguish between the mathematical definition of a random variable and its common usage (e.g. Vogt 1999:235). Clearly, however, none of the variables used in QMCR (or, indeed, any social science research) follow known, pre-specified probability distributions. This is true whether the data on which a regression is estimated constitute a population or a sample from a larger population. In regression analyses, the key random variables are the regression error term and the regression coefficients. Other regression estimates, like R-squared, are also random variables. Regression error is a random variable by construction: remember that regression error is assumed to be a Normal random variable with mean zero.

The dependent variable in a regression analysis is modeled as a linear function of the independent variables plus this random variable. Since the values of the independent variables are fixed (not random), the Normal distribution of the regression error propagates through the model to the coefficients, which, due to the finite number of cases, are distributed t rather than Normal. (When the independent variables are not fixed, but are instead measured with error, attenuation adjustments have to be made to the estimated regression coefficients.) A simple t test can then be used to test each estimated regression coefficient against the null hypothesis that the mean of its sampling distribution is actually zero. If the null hypothesis is rejected, a coefficient is said to be "significant." What this means is that a model with coefficients equal to zero would have been unlikely to have produced the observed data, given the assumption that the regression errors are independent normally distributed random variables with zero mean and constant variance. Testing the significance of a coefficient does not in any way imply the existence of a larger population of cases to which the coefficient applies.

Other than its coincidental presence in the statistical term "sampling distributions" of the regression coefficients, nowhere in this process is sampling involved. Quite the contrary: an implicit assumption of regression modeling is that the data being analyzed constitute a population, not a sample from some larger population. In some cases (for example, when variables have been chosen for inclusion in the model based on a stepwise selection algorithm) regression coefficients estimated using sample data may not even be unbiased estimates of the corresponding population coefficients. In all cases, the sample r-squared statistic is a biased estimate of the population R-squared, due to the phenomenon of "fitting to the sample." This is why so-called "adjusted R-squared" statistics (Lucke and Whitely 1984) are used to estimate the percent of variance in the population dependent variable that would be achieved by applying sample-based estimates of the regression coefficients to the population independent variables. Note that such "adjusted R-squared" statistics are generally not applicable in most QMCR analyses, since QMCR data structures almost always constitute complete populations, not samples drawn from some larger population. Though often used inappropriately, "adjusted R-squared" is completely irrelevant in typical QMCR settings. Overfitting can be a problem in QMCR, but it should be evaluated using the F statistic, not by making an inappropriate adjustment to the reported R-squared.

If there's no sampling involved and regression errors are not in any sense sample estimates of population means, why should regression errors be expected to be normally distributed? As Berk acknowledges (2004:54) the Central Limit Theorem of probability theory predicts that that they should. As discussed above, the Central Limit Theorem ensures that any random variable that is the sum of a large number of other random variables will follow a Normal distribution. The number of summed variables need not even be very large if they are reasonably well-behaved. Empirically, when regression variables are properly specified (with appropriate transformations) the regression errors are almost always credible realizations of a Normal distribution. For example, the distribution of realized regression errors from the regression of infant mortality on national income is summarized in Figure 1. This realized distribution is very nearly Normal. In fact, a Kolmogorov-Smirnov test emphatically fails to reject the Normal distribution as the origin of these realizations (p = .940). It is clear that the assumption that regression errors are drawn from a Normal random variable is not only theoretically well-grounded, but empirically reasonable as well.

Figure 1: Distribution of Realized Regression Error in a Model for Infant Mortality



ERROR DEPENDENCE STRUCTURES

It is important to understand the role played by error in statistical modeling because different model designs either implicitly or explicitly contain assumptions about the behavior of error. In the simplest statistical models, the error associated with each case is assumed to be independent of the errors associated with every other case. Such independent errors exhibit no patterns across cases. There are many ways in which the peculiarities of QMCR data structures, however, lead to highly patterned forms of error. For starters, countries are not independent cases: the United

States and Canada, for example, are strongly linked in almost every way. Even worse, when the same countries are included multiple times in the same dataset, their multiple realizations (US 2000 versus US 2005) can hardly be considered to be independent of each other. For these and other reasons, regression models in QMCR often exhibit dependence in their regression error structures.

There are two broad classes of regression error dependence: mean dependence and variance dependence. Mean dependence occurs when the expected value of the regression error is not zero for a class of cases; for example, East Asian countries have systematically lower than expected levels of infant mortality, conditional on their income levels. Variance dependence occurs when the variability of the regression error is not constant across all classes of cases; for example, a dependent variable like infant mortality may be much more poorly measured in sub-Saharan Africa than in the rest of the world, resulting in a systematically higher error variance in sub-Saharan African countries than in others. In general, mean dependence is easier to identify and address than variance dependence, and is the more serious problem, since it directly affects the estimation of regression coefficients (both their levels and their standard errors). Variance dependence, on the other hand, typically affects only the standard errors of coefficients, not their levels.

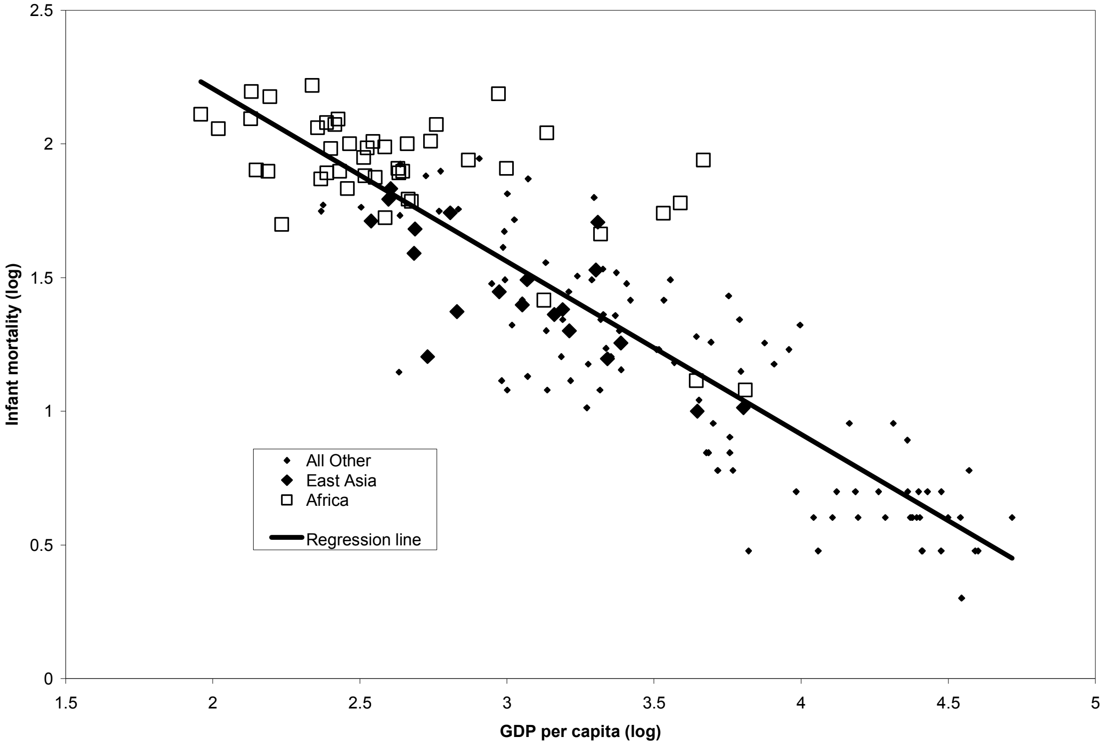
The simplest form of mean dependence results from the questionable treatment of countries as representing statistically independent cases, as discussed above. Entire blocs of countries may behave in statistically similar ways, despite their nominal independence as distinct countries. For example, all oil-dependent economies might be expected to depart from their modeled rates of economic growth in the same ways at the same times: higher than modeled when oil prices are high and lower than modeled when oil prices are low. As a result, controls for OPEC membership or oil dependence are common in growth regressions. Oil dependence, however, is simply an extreme example of a much more general phenomenon. There are probably as many ways for countries to covary as there are countries, and probably more, since the number of possible combinations of country covariation is far larger than the number of countries.

Ironically, the plethora of possible country dependence "clubs" is perhaps a blessing in disguise. Countries are members of so many potential dependence clubs that their overall influence might, in most cases, aggregate to a normally distributed contribution to general background error. Consider: the Europa World database lists 95 United Nations and "major" non-UN intergovernmental organizations that countries may belong to. Add to these continental clubs and clubs based on economic characteristics (natural resource dependence, susceptibility to agricultural shocks, participation in global commodity chains, etc.) and the typical country may be found to belong to dozens of potential dependence clubs. It may be necessary to adjust only for the club memberships that are most directly relevant to any given analysis. Luckily, this is easy to do: including a dummy variable for club membership will in most cases eliminate any associated error dependence. Even this simple adjustment is not always necessary. So long as countries representing many dependence clubs are included in the data on which a model is estimated, error dependence structures affecting just a few will have no measurable impact on the broader results. This is true even when such club dummies turn out to have statistically significant coefficients.

For example, an OLS regression of infant mortality in 2005 on national income in 2005 (N=167) yields a metric coefficient of b = -.646 for national income, with a standard error of SE_b = .026. Controlling for East Asian location as a dependence club results in no change at all in the coefficient (to three decimal places) and only a trivial reduction in its standard error (to .023).

This is despite the fact that the effect of East Asian status is highly significant (t = -4.798). East Asian countries have systematically lower than modeled infant mortality rates, but this does not substantively affect the overall evaluation of the effect of national income on life expectancy. The reason for this is illustrated in Figure 2. Although there is a clear East Asian error dependence (16 of 20 East Asian countries' infant mortality rates fall below their modeled values), its impact is spread evenly across the range of the regression. Consequently, it affects the intercept of the regression line, but not its slope. Sub-Saharan African countries, on the other hand. form an equally obvious dependence club, but one with a structure that does not fit so neatly into the overall pattern of the worldwide relationship between national income and infant mortality. Controlling for sub-Saharan African location does substantially affect the coefficient for national income, reducing the estimated magnitude of the slope to b = -.558 (SE_b = .029).

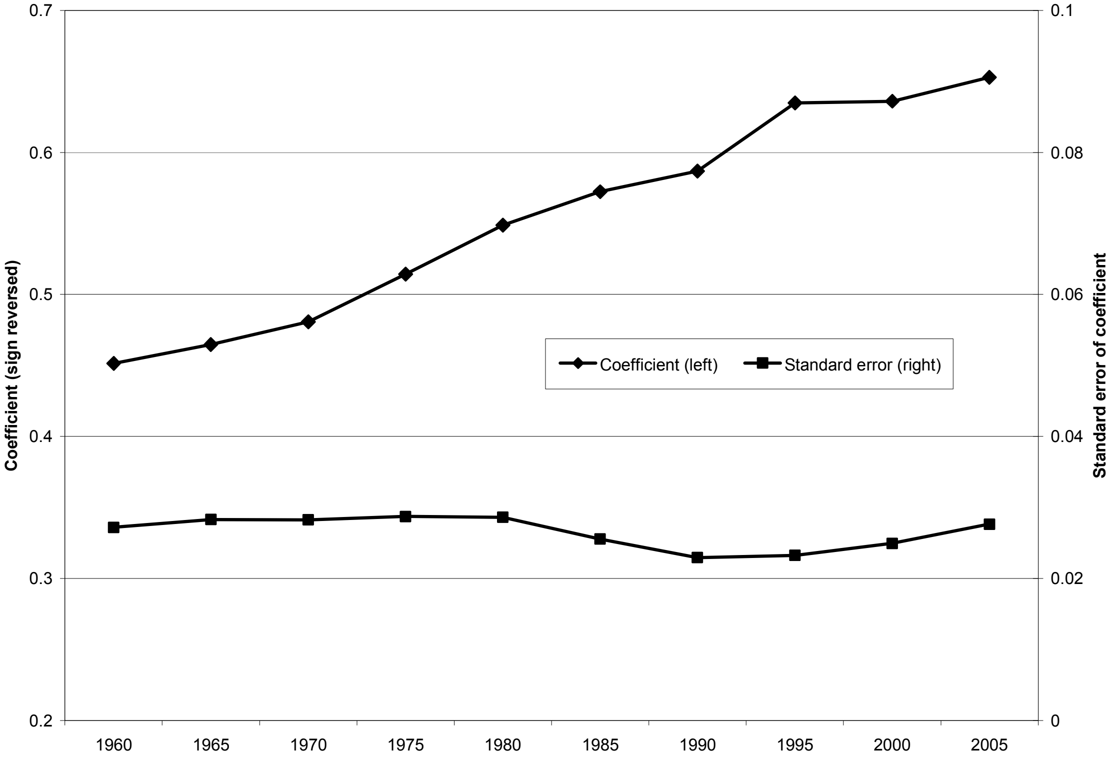
Figure 2: Illustration of East Asian and African Dependence Clubs, 2005



A similar type of mean dependence occurs when the same country appears multiple times as a case in the same regression analysis. This is, in a way, similar to the dependence club situation, only instead of several countries being expected to exhibit similar errors one country is expected to exhibit similar errors each time it is used as a case. This situation occurs when some attribute of a country, rather than the country itself, is the unit of analysis. For example, countries may be listed multiple times in the same data structure when each case represents an ethnic group within a country. In a study of how the average educational levels of ethnic groups affect their average income levels, it is important to remove overall country biases in income levels: all ethnic groups in rich country would be expected to have higher incomes than all ethnic groups in poor countries, regardless of their relative educational levels. This can be accomplished by including country dummy variables in the regression model. This is mathematically equivalent to a fixed effects model design (discussed below), though the data structure is distinct (since the multiple entries per country represent distinct cases, not repeated measures of the same case).

Country dummies are emphatically not appropriate for eliminating country-related mean dependence in regression error in repeated measures designs. In fact, mean dependence does not even bias regression coefficients in repeated measures designs when the panels are balanced – biases are only introduced when some countries are represented more times than others, which is really a case of sample selection bias due to missing data in unbalanced panels. This can be illustrated by extending the infant mortality – national income example. Coefficients from regressing infant mortality on national income for each of the ten five-year intervals 1960-2005 yields coefficients ranging from b = -.451 (1960) to b = -653 (2005). These coefficients and their standard errors, estimated on the constant panel of N=77 countries for which data are available for both variables for all years, are plotted in Figure 3.

Figure 3: Coefficients from Cross-Sectional Regressions of Infant Mortality on GDP per capita, 1960-2005 (constant panel N=77)



A pooled regression of infant mortality on national income, including data from all years in a balanced panel of N=770, yields a coefficient for national income of b = .587, well within the range of the ten cross-sectional coefficients. In fact, it is very near their simple arithmetic mean of -.554. Adding country dummies to the model, however, dramatically changes the estimated coefficient for national income; it nearly doubles to b = -1.034. Obviously, an adjustment meant to reduce country dependence in regression error should not have such a dramatic effect on a coefficient. Country-related mean dependence should not pull the slope of the national income – infant mortality relationship in the pooled data so far outside the range of slopes observed in the constituent cross-sections. What happens when country dummies are included in the pooled model is not a correction for potential country-related mean dependence but a complete change in the character (and meaning) of the model. The actual mechanics behind this are explored below in the discussion of multilevel models.

Returning to the regression of infant mortality on national income (without country dummies), the main difference between the pooled analysis and the ten cross-sectional analyses is not the magnitude of the coefficient but the size of its standard error. Standard errors for the national income coefficient in the ten cross-sectional models range narrowly from SE_b = .023 to SE_b = .029. The standard error for the national income coefficient in the pooled cross-sectional model, by contrast, is only SE_b = .011. This is because the sample size has been increased by a factor of ten without a corresponding increase in the total variability of the data (since the variability of the data is constrained by the fact that they are drawn from the same 77 cases, even measured if at different points in time). Were data for infant mortality available annually, instead of every five years, the number of cases could be further multiplied by a factor of five, again without any corresponding increase in the scope of coverage. Of course, there is no philosophical reason for stopping at annual increments. The reductio ad absurdum would be to include every country as a case every fraction of a second, to yield millions of "cases" for analysis each identical or nearly identical to the one before. This would drive standard errors for regression coefficients down toward zero as the number of "cases" rises toward infinity.

Repeated measures designs incorporate a serious country-related mean dependence in their regression error structures, but it is not as simple as a broad, country-wise mean bias. Repeated measures from the same country are not just correlated with each other; their dependence is highly structured. Each realization of country data is directly conditioned on the one immediately preceding it, but not directly conditioned on further prior realizations. The dependence structure is not universally mutual, but instead is structured into a directed chain running from earliest to latest realization. This kind of dependence structure is called "Markovian" dependence. Assuming the dependence between adjacent realizations is linear and of the same magnitude for all time points, it is not only Markovian but specifically autoregressive with order 1. Such AR(1) error structures are very common in QMCR but cannot be estimated using OLS regression. They can, however, be estimated using iterative MLE; procedures for doing so have existed for over fifty years. Since MLE techniques are approximations (as opposed to OLS estimates, which are exact), different statistical software can give slightly different solutions to regression models estimated using MLE.

Using SPSS PROC MIXED to estimate the MLE solution for the national income coefficient in a model for infant mortality with AR(1) errors within countries is b = -.609, with SE_b = .026 (balanced panel of N=770 cases). These figures are in line with the cross-sectional results of b = -.451 to b = -653 and SE_b = .023 to SE_b = .029. It turns out that including ten panels of very similar data for multiple years improves very little on simply analyzing a single cross-section, once appropriate specification has been made to the error term.

Other forms of sequentially organized dependence structures are possible, but are much less commonly encountered in QMCR. When annual data are used, however, more complex error modeling is required. The ordinary business cycle of 3-8 years introduces error dependence structures into many QMCR variables (national income, investment, international trade, etc.) that are not Markovian at the annual level. Periodic sinusoidal regression error structures correspond to order 2 autoregressive processes; asymmetrical cycles like the business cycle are best modeled using autoregressive - moving average (ARMA) models. Quarterly and monthly data that incorporate seasonality take "integrated" ARIMA models. Such complex econometric models are rarely encountered in QMCR, but models based on data structures that include annual observations must take them into account. Since fine-grained annual variability is rarely the focus of QMCR, a reasonable fudge is simply to work with more widely-spaced data. As Chase-Dunn (1989) points out, the substantively relevant "width of a time point" (321) in QMCR may not necessarily be one year.

Compounding these difficulties of mean dependence in QMCR regression error structures is the much more subtle problem of variance dependence. In a repeated measures panel consisting of multiple countries measured at multiple time points, it is possible that different countries will exhibit systematically greater or lesser regression error variance than others. In such cases, panel-weighted least squares (PWLS) estimation can be used to adjust for variance dependence. Beck and Katz (1996), however, show that PWLS is only effective when the number of repeated observations for each country is large (20 or more time points). Such time-series cross-sectional (TSCS) data structures based on annual observations of countries are not, however, typical of QMCR of the kind being examined here. Beck and Katz (1996) go on to show that OLS estimates with panel-corrected standard error (PCSE) adjustments, developed and discussed at length in Beck and Katz (1995), produces much better estimates than PWLS when the number of time points is "small," or, in their examples, as few as five.

Of course, in much QMCR having as many as five repeated measures per country is a rare luxury. Nonetheless, the PCSE approach is a major methodological advance, and should be applied whenever repeated measures of the same cases are included in QMCR. As Beck and Katz (1996) show, PCSE adjustments systematically reduce the underestimation of the standard errors of regression coefficients due to country-related variance dependence. They also have the gratuitous effect of reducing or eliminating biases due to mean dependence structures in which countries' regression errors are correlated to each other in a fixed pattern that is the same for all time points (Beck and Katz 1995). This may sound esoteric, but it is in fact a very common condition. As discussed above, spatial correlation (including neighbor correlation) among countries is almost certainly a feature of all QMCR data structures. In cross-sectional analyses, such correlation structures cannot be detected or adjusted for, but in panels with multiple repeated measures, the PCSE approach provides an effective correction.

The PCSE approach, though developed for use with OLS regression models, can also be applied to GLM designs. An effective strategy is to estimate a GLM regression with an AR(1) error structure, then correct the results for spatial correlation and variance dependence in the remaining regression error using a PCSE adjustment. This two-step approach corrects for most of the common complications that apply when using repeated measures of countries as cases in QMCR. An alternative strategy is to use OLS regression including lagged dependent variables

(Beck et al. 1993), but this strategy introduces downward biases when the autocorrelation of the dependent variable is high (Keele and Kelly 2005). Another alternative is to use OLS regression based on change scores (Beck and Katz 1995) to eliminate the autoregressive error structure within countries, but this strategy can also lead to large downward biases in the estimation of coefficients (Wawro 2002). See Wilson and Butler (2007) for a comparison of competing methods for dealing with TSCS data.

COMPLEMENTARY, COMPETING, AND ORTHOGONAL CONTROLS

In substantive terms, it is reasonable to think of the regression error as the effect on the dependent variable of "all other factors" not included in the model. The vast majority of possible QMCR variables, of course, are orthogonal (not linearly related) to the dependent variable of interest in any particular analysis, and thus can safely be ignored. They do not contribute to regression error as "other factors." Many variables, however, are related to the dependent variable, but are also colinear with the independent variables of interest, competing with them for explanatory power in the regression model. To the extent that they compete with or "partial" variables that are already in the model, they are not "other factors" that contribute to regression error but more like alternative operationalizations of the independent variables. A third class of variables that are independently related to the dependent variable can, however, be often identified. When not explicitly included in the regression model, these complementary variables are clear examples of the "other factors" that are subsumed into the regression residual. Including them directly reduces the variance of the regression residual, illustrating the reasonableness of its interpretation as a sum of "all other factors" not included in the model.

These three classes of potential control variables – complementary, competing, and orthogonal – affect the results of statistical models in distinctive ways and are subject to different rationales for inclusion as controls. All social scientists struggle with the question of what variables to include in (and, implicitly, what variables to exclude from) their statistical models. This question is made particularly difficult in QMCR by the fact that the number of variables available in published data compilations far exceeds the number of countries available for analysis as cases. Parsimony is thus at a premium. In research based on sample surveys with thousands of respondents, the effects of dozens of independent variables can be estimated simultaneously, and though this may present serious problems of interpretation, it is typically not a problem from the standpoint of estimation. In QMCR, with relatively few countries available as cases, sufficient degrees of freedom usually exist for estimating the effects of no more than a dozen or so variables, and often far fewer. As a result, practitioners of QMCR typically must show far greater care than other social scientists in their choices of variables to include in their statistical models.

Complementary controls, though hard to identify, are almost always desirable in a model, since they serve mainly to "soak up" error that would otherwise tend to obscure the relationships between the independent variables of interest and the dependent variable. An illustration of the effective use of a complementary control is given in Table 1, Models 1 and 2. One might reasonably surmise that countries with greater female labor force participation (LFP) would tend to have lower levels of infant mortality. Infant mortality is regressed on female LFP in Model 1. The coefficient, as expected, is negative, but it is not statistically significant. When urbanization is introduced as a control, however, the coefficient for female LFP becomes highly significant (Model 2). From the standpoint of female LFP, urbanization is a complementary control. Its inclusion in the model dramatically reduces the model's residual error variance, thus clarifying the (relatively weak) effect of female LFP on infant mortality. The inclusion of urbanization increases the signal-to-noise ratio in the relationship between female LFP and infant mortality not by increasing the strength of the signal but by reducing the volume of the noise. Complementary controls can be thought of as very useful error filters.

Table 1: Illustration of Complementary, Competing, and Orthogonal Controls – Models for Infant Mortality (log), 2005

	Model 1	**Model 2**	**Model 3**	**Model 4**
[Constant]	1.556 (0.213)	2.575 (0.182)	3.797 (0.132)	3.816 (0.211)
Female LFP (%)	-0.005 (0.005)	-0.010 (0.004)	-0.007 (0.002)	-0.007 (0.002)
Urban population (%)		-0.014 (0.001)	0.000 (0.001)	0.000 (0.001)
GDP/capita - F/X (log)			-0.662 (0.040)	-0.662 (0.040)
Population (log)				-0.003 (0.024)
R-squared	.005	.450	.801	.801
N	162	162	162	162

Note: Entries in table are metric coefficients (standard errors in parentheses)

Competing controls are not always so obviously desirable. What if we were interested in the relationship between urbanization and infant mortality? Model 3 reveals an incredibly significant negative relationship between the two variables. This makes sense, since we would expect highly urbanized countries to have lower infant mortality rates than predominantly rural countries. So far so good. But controlling for national income (as nearly all GMCR studies do) brings the estimated effect of urbanization on infant mortality down to zero. National income is clearly a competing control vis-à-vis urbanization (though not vis-à-vis women's LFP). Urbanization is completely unrelated to infant mortality at any given level of national income, but it seems highly unlikely that urbanization does not reduce infant mortality. This is a difficult conundrum. The use and interpretation of competing controls are tied up with assumptions about concepts, causality, and causal order.

If urbanization is hypothesized to be a proximate cause of infant mortality, and if national income is thought to be a concept completely distinct from urbanization, then it makes sense to control for national income and to conclude that urbanization has no detectable effect. On the other hand, it could be that urbanization affects infant mortality through a range of intermediating variables (such as the availability of medical staff, electricity, and clean water) and that national income acts as a proxy for these very variables, in which case it might be inappropriate to control for national income (despite the fact that it has such a strong effect on the dependent variable). In this case it is probably appropriate to control for national income, but in many cases involving potential competing controls it might not be appropriate to include them.

The third class of potential control variables, orthogonal controls, have little effect on the coefficients of the rest of the variables in a model. An example of an orthogonal control is given in Model 4. Here, population size has virtually no effect on infant mortality and virtually no effect on the coefficients of the other variables in the model. The only reason to include an orthogonal control is to demonstrate that it is, in fact, orthogonal. Once this has been demonstrated, Occam's razor suggests they be eliminated from the model. The judicious use of competing controls combined with the elimination of orthogonal controls would lead to much simpler, more easily grasped models in QMCR.

CAUSALITY AND MODEL DESIGN

The most basic model design and still the one most widely used in QMCR is the cross-sectional multiple linear regression model: a single dependent variable is regressed on one or (usually) more independent variables. Models 1-4 are all examples of cross-sectional models. Cross-sectional models are easy to interpret largely because their error structures are so straightforward: errors are typically assumed to be independent across cases and normally distributed with constant variance. Because of this simplicity, the cross-sectional model is probably the model design best suited to answering the simple question "does the dependent variable covary across countries with the independent variable?" Most early QMCR (Bornschier, Chase-Dunn and Rubinson 1978) and many QMCR studies today use a simple cross-sectional design. As new data come available for such QMCR topics as state structure (Evans and Rauch 1999), the environment (Marquart-Pyatt 2004), and political corruption (Sandholtz and Taagepera 2005), cross-sectional models are usually the first kind of model analyzed because initially only cross-sectional data are available.

As fields mature and over-time data begins to accumulate, more complex model designs typically follow. Cross-sectional designs are effective for establishing the existence of a relationship between two variables, but they are almost useless for establishing its causality. In fact, cross-sectional designs are highly vulnerable to reverse causality and endogeneity biases.

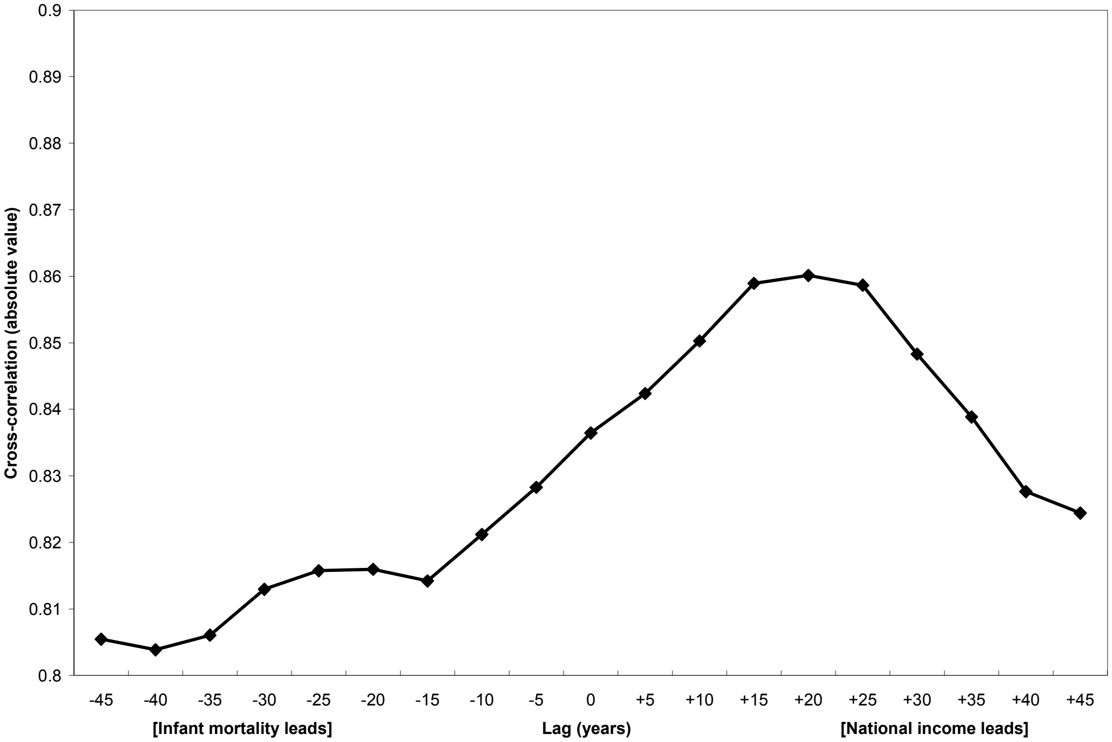
It has been broadly accepted for at least three decades that three conditions must be met to establish the causality of a relationship between two variables: correlation, temporal precedence, and non-spuriousness (Kenny 1979). Establishing correlation is usually not a problem in cross-sectional models. Temporal precedence and non-spuriousness, however, can be much more difficult to establish.

The temporal order governing the relationship between two variables is sometimes obvious (as with national income and infant mortality). At other times it can be established empirically through investigation of the lag structure connecting them. As an illustration, Figure 4 plots the lag structure of the correlation between national income and infant mortality. The maximum correlation between the two variables occurs when national income is compared with infant mortality rates 20 years later. This is strong circumstantial evidence that national income causally precedes infant mortality, though it does not eliminate the possibility that reverse causality occurs as well.

Note that simply lagging the dependent variable by a year (or more) is not sufficient to establish the temporal precedence of the independent variable, since most QMCR variables are highly autocorrelated within countries. Thus, for example, the regression of national income in

2000 on infant mortality in 1995 would not ensure the temporal precedence of infant mortality in the analysis, since infant mortality in 1995 is also a very good proxy for infant mortality in, say, 2005. This is obvious from Figure 4. In fact, infant mortality is strongly correlated (r > .80) with national income over 40 years later! A regression of national income in 2000 on infant mortality in 1960 would produce highly significant results without in any way establishing the causal precedence of infant mortality. Nonetheless, short- and long-term lags (sometimes as short at one year) have long been (incorrectly) claimed to establish temporal precedence in QMCR.

Figure 4: Lag Structure of the Relationship between National Income and Infant Mortality, Demonstrating the Temporal Precedence of National Income



An alternative use of lags to establish temporal precedence is implemented in the confusingly-monikered "panel" or lagged dependent variable (LDV) design. In LDV models, the dependent variable is measured at a long time lag (often decades) after the independent variables, which include among them an early realization of the dependent variable itself. The dependent variable is thus regressed on both itself and a set of independent variables all measured at an earlier time period. Note that this is a very different usage from short-term LDV designs used in the political science literature to control for error autocorrelation. The LDV design "provides an estimate of the effect of the independent variable which is 'independent' of variance in the dependent variable" (Chase-Dunn 1975:726-727). Though the LDV design has been strongly criticized for not eliminating the possibility of spurious causation (Firebaugh and Beck 1994), it is effective at establishing precedence, since any (contemporaneous) reverse causality of the independent variable on the dependent variables is partialled out. The main pitfall of LDV models is that they are often misinterpreted.

For example, when national income is the dependent variable, the LDV model is often mis-read as a model for growth, when it is in reality it is only a model for national income. Since income, not growth, is the dependent variable, the model only establishes the temporal precedence of the independent variables vis-à-vis national income, not vis-à-vis growth. This may seem a pedantic distinction, but it is important to keep in mind. Imagine the existence of a panel of countries starting with identical national income and foreign capital penetration levels at Time 0. Some subsequently grow at a fast rate, while others grow at a slow rate, with growth rates highly stable over time. Assume that fast growth leads causally to reductions in foreign capital penetration. At each subsequent time point (1, 2, ...) national income would come to be ever more (negatively) correlated with foreign capital penetration. Regressing national income at Time 2 on foreign capital penetration at Time 1 would give a negative coefficient for penetration, even controlling for national income at Time 1. Though loading negatively on income, however, foreign capital penetration in such a scenario would have no causal effect on growth. Temporally preceding income is not the same as temporally preceding growth.

Another approach to establishing temporal precedence is the use of instrumental variables. Instrumental variables are clearly exogenous variables that are correlated with the dependent variable at least in part through their relationship with the independent variable, with no possibility of reverse causation. Thus, the correlation between the instrumental variables and the dependent variable can be used as evidence of the directionality of the relationship between the independent variable and the dependent variable. Instrumental variables can be used either in a structural equation modeling setting (for example, Kentor 2001) or a two-stage least-squares setting (for example, You and Khagram 2005). Instrumental variables usually pertain to time periods well before the study period (to ensure their exogeneity). In a famously creative (perhaps even notorious) example, Acemoglu, Johnson and Robinson (2001) use nineteenth century colonial settler mortality as an instrument for the strength of civil society today. The main shortcoming with the instrumental approach is the difficulty of finding good instruments.

Ensuring non-spuriousness, however, presents far greater difficulties. It is often claimed that instrumental variables can be used to establish non-spuriousness, but this is only true under the assumption that instrumental variables can only affect the dependent variable through the independent variable, not via any other causal path through any potential (and potentially unmeasured) common-cause variable (Angrist and Krueger 2001). This assumption is so heroic as to be meaningless. The root problem is that any observed correlation between two variables A and B could actually be the result of their common correlation with an omitted variable C. Such a situation results in "omitted variable bias": the observed correlation between A and B becomes larger than the true effect of A on B. Of course, it is more likely that instead of one there are many omitted common-cause variables that are responsible for the relationships observed in the data. It can even be argued that nearly all observed relationships are spurious at their root, and only approximately causal. For example, national income as such almost certainly doesn't "cause" infant mortality. Instead, the individual and societal wealth and productivity that lead to national income also lead to infant mortality. It is the near-identification of national income with "wealth," "productivity," "development," and the like that makes the statement "national income causes infant mortality" reasonable, though only as an approximation.

There are two general methods of establishing non-spuriousness in cross-sectional models. One is to control for all potential common-cause variables. This is difficult to accomplish in practice, for several reasons. First, actual data for the necessary variables may not exist. For example, human capital may be a common cause of national income and infant mortality, but the closest we come to measuring human capital is education data, which really don't adequately capture the concept. Second, there may be too many potential common-cause variables to test them all with the limited number of cases (and thus degrees of freedom) available. Third, potential common cause variables may simply go unnoticed. These difficulties make it impossible to guarantee non-spuriousness through the use of appropriate statistical controls, though a credible case for non-spuriousness might still be made. For example, if controlling for the most highly suspect common-cause variables has little effect on the proposed causal relationship, it can be argued that other, unconsidered potential common-cause variables would be unlikely to account for the observed relationship, even though they are not tested.

The other is to use fully longitudinal data (for both the dependent and independent variables) to eliminate at least some pathways through which a spurious relationship may arise between the variables of interest. Two classes of longitudinal models used in QMCR are multilevel models and difference models. Both are designed to account specifically for spuriousness due to the presence of time-invariant omitted variables. This is a very large, though not exhaustive class of potential common-cause variables, including any factor relating to a case that is constant across the study period. For example, countries' political geographies, topographies, climates, cultures, economic systems, and forms of government are all typical time-invariant (or nearly time-invariant) variables. Multilevel models and difference models both control for time-invariant omitted variables by focusing their statistical power on changes over time in the variables of interest within countries. This eliminates the effects of all time-invariant factors, since within countries they, by definition, do not change over time.

The cost in statistical power of doing this is, however, very steep, since multilevel models and difference models sacrifice the power to make inferences based on cross-national variation in the overall levels of the variables of interest. Still, the elimination of time-invariant common-cause alternatives makes for a dramatic advance in the causal credibility of purported relationships, which is often worth the accompanying sacrifice in statistical power. Moreover, when relationships can still be shown to be significant even in such low-power models as multilevel models and difference models, they are much more likely to be accepted as robust and important phenomena. Unfortunately, even so they must always remain subject to some degree of skepticism, since there are no off-the-shelf statistical models for ruling out spuriousness arising from omitted common-cause variables that do vary over time.

MULTILEVEL MODELS

In 1994, Firebaugh and Beck bemoaned the fact that "cross-national research in sociology currently is dominated" by lagged dependent variable models, which "are so common in cross-national research in sociology that practitioners refer to them as 'panel analyses'" (637-638), by which they meant that LDV designs had become synonymous with the analysis of panel data. Today, the same might be said for multilevel models (MLMs), sometimes also called hierarchical or (now) panel models, which have become so ubiquitous in QMCR over the past ten years that it could be argued that there exists an a priori assumption that they should be used in all cases where it is possible to do so. Critics and reviewers often demand MLM evidence even where sufficient longitudinal data do not exist with which to estimate a multilevel model. Perhaps partially as a result, MLMs are very often applied in ways that are entirely inappropriate to the data and research questions at hand. Statistical handbooks often serve to compound this problem, since they are generally written from the standpoint of users with very different analytical objectives than those found in QMCR.

The MLM design was initially developed for use in experimental settings, and is essentially an ANOVA (analysis of variance) model with covariates. In a standard ANOVA model, subjects are divided into groups, with each group receiving a different treatment. The ANOVA F test indicates whether or not outcomes for the subjects as a whole differ significantly across the treatment groups. Even in cases where there are no significant differences in response between specific pairs of groups, the overall ANOVA F test may detect significant differences among all groups analyzed collectively. The garden-variety ANOVA model is a multilevel design because error is introduced to the model at two distinct levels. There is a Level 1 sub-model, in which the subject's outcome is influenced both by participation in an experimental group and by random error idiosyncratic to the subject, and a Level 2 sub-model for the effect of participation in a group (which in this trivial case is an array of constants). This simple ANOVA model is an example of a fixed effects model (FEM), since the treatment effects are fixed (modeled without error). In a random effects model (REM), both the treatment effects and the individual subject outcomes are subject to random error.

The fundamental difference between FEMs and REMs is in how the error is apportioned, to the subject or to the treatment. In the simple one-way ANOVA setting these two sources of error cannot be distinguished, so the FEM and REM specifications yield identical estimates of the treatment effects. In less trivial models, however, the question of fixed versus random effects becomes important. For example, if the Level 1 sub-model includes covariates in addition to the treatment effects, the apportionment of error between Level 1 and Level 2 of the model helps determine the standard errors of the coefficients of these variables. From the standpoint of the significance of the Level 1 covariates, FEMs are generally more conservative than REMs. Conversely, from the standpoint of the significance of the Level 2 treatment effects, REMs are generally more conservative. In fact, REMs were developed for precisely this reason: in many situations, FEMs produce upwardly biased estimates of the effectiveness of Level 2 treatments. Where group treatment effects are the primary interest of modeling (as in drug trials), the REM design is preferred because it is both better-specified and more conservative than the FEM. In drug trials the primary interest is in the sizes of the group effects, not in the covariates, which are included simply as controls.

An early social science application for the REM was school effectiveness research (Rumberger and Palardy 2004). In a typical school effectiveness research setting, students (subjects) are arranged into schools (treatments) to study the effect of these treatments on the students' standardized test scores (outcomes). A simple one-way ANOVA of school means overstates the effect of schools as treatments, since their student populations may be heterogeneous on the dimensions that affect student performance (for example, student ability, family resources, family structure, etc.). Estimates of school effects are reduced by including student and family variables as covariates in a Level 1 equation for individual student performance and estimating an FEM instead of a simple ANOVA model. In the FEM design, schools take credit for all of the cross-school differences in student performance remaining after controlling for individual student attributes. This is a problem, though, because some of the school differences are themselves due to individual student attributes: after all, students (or their families) can, to some extent, choose what schools they attend (for example, though choosing to live in expensive neighborhoods). In any situation in which the subjects choose their own treatments, the effectiveness of those treatments may be overstated. The REM corrects for this bias by making the choice of treatment – in this case, which school is attended by the student – endogenous to the model.

Both FEM and REM variants of MLM designs are used in QMCR models. Although the mathematics for these models are identical to the mathematics used in experimental and quasi-experimental research, the focus in QMCR is entirely different. In QMCR, the notional "treatment" groups are countries and the "subjects" are country-years of observation. For example, in the illustrative data used below on infant mortality and national income, there are a total of 1320 distinct observations arising from 208 countries observed over ten time points (5-year intervals 1960-2005; not all countries report data for all time periods). There are thus 208 possible "treatments" (countries), each of which is experienced by up to ten "subjects" (country-years) per country. In experimental designs and quasi-experimental settings like school effectiveness research, the researchers' primary interest is in the group treatment effects. In QMCR, however, the treatment (country) effects are unimportant; only the covariates (independent variables) matter.

The appeal of MLM designs for QMCR is that in studying the effects of the covariates (independent variables), the effects of the treatments (countries) are controlled for. In controlling for country, the MLM implicitly controls for any factor that does not differ by country: i.e., all variables that are time-invariant within countries over the course of the study period. Thus, all time-invariant common-cause variables that might give rise to a spurious correlation between the dependent variable and the independent variables of interest are implicitly accounted for. This doesn't entirely establish the non-spuriousness of the relationship of interest, but it does go a long way toward eliminating plausible alternatives.

The FEM design absolutely eliminates any time-invariant alternative explanations of the dependent variable, period. The REM design, however, does not. In the REM design, only a portion of the treatment (country) effect is assigned to the country itself, since the REM design assumes that selection into treatments (countries) is endogenous, not assigned exogenously. Just how much of the country-level variability remains is a bit of a mystery, since REMs were not designed for this purpose. Empirically, it seems that very little is left over: REMs usually give very similar results to the equivalent FEMs. Why, then, does anyone use them?

The REM design is used because REMs allow the estimation of the effects of covariates even if those covariates do not change over time. To my knowledge, this is not done in drugs or schools research using REM designs, where the focus of the research is on the group treatment effect, not on the covariates. In QMCR, however, we are primarily interested in the covariates, not the group effects, and often the covariates do not vary for particular treatment groups. For example, countries' latitudes are constant over time. Estimates of the effect of latitude on QMCR outcomes cannot be estimated in an FEM design because the fixed effect of "country" subsumes everything unchanging about a country. In a REM, however, not all of the country effects are assigned to the countries themselves, leaving some variance left over that can be assigned to time-invariant variables like latitude. Exactly how much of the between-group variance in REMs is assigned to country effects and how much to time-invariant covariates is, to my knowledge, unknown. It is not, in principle, unknowable, but REMs simply weren't designed for the eventuality that users might try to estimate the effects of time-invariant covariates in models treating countries as if they were endogenously chosen as treatments by country-years. Much further research involving Monte Carlo simulation would be necessary before we could understand the meaning and significance of the coefficients in REMs used in QMCR.

From this standpoint, the use of REMs is a case of having one's cake and eating it too (Halaby 2004). If the objective is to eliminate the possibility of spurious causality due to unmeasured time-invariant variables, the FEM design can used. If the objective is to estimate the effects of measured time-invariant variables, a simple cross-sectional model is the most effective design. The REM design notionally allows both to be done in the same model, but only partially and to unknown degrees. The REM corrects to some extent for unobserved time-invariant variables and allows the estimation to some extent of the effects of measured variables that are time-invariant, with the extent of each determined by the hypothetical degree of endogeneity of the treatment (country) choices exercised by the subjects (country-years), were such choices theoretically possible, which they aren't. Were it not for the convenience that REMs produce some kind of estimate of the effects of time-invariant covariates, it seems doubtful that they would ever be used.

Both kinds of MLM, however, present a much greater challenge where data exhibit trends over time. Analyses using MLMs are highly sensitive to trended data. The MLM design is fantastic for eliminating time-invariant variables, but time itself is, of course, not time-invariant. This fact seems to be poorly understood in the QMCR literature.

The basic logic of MLM designs is that they draw their statistical power from correlating (within countries) the deviations of both the dependent and the independent variables in each period from their overall country means. If the dependent variable tends to be (relatively) high when the independent variable is (relatively), and low with the independent variable is low, this is evidence of a relationship between the two. This is in itself a problem: it assumes that there is no lag between changes in an independent variable and consequent effects on the dependent variable. It is highly likely in most QMCR settings that lags are substantial. As a case-in-point, consider the roughly 20-year lag in the effects of national income on infant mortality suggested by Figure 4. Such lags could be incorporated in MLM designs but rarely are. I myself have never seen a published example of QMCR using MLMs that systematically investigated the appropriate lag to be used between the timing of the independent and dependent variables, though some have used pro forma 1 or 5 year lags.

A much bigger problem, however, is that many QMCR data series are strongly time-trended. National income, for example, can be correlated up to r = .98 with time (Babones 2007). Infant mortality, on the other hand, has generally fallen over time. In an MLM design without time adjustments, national income and infant mortality appear to be closely related because in years when national income is higher than average (the later years in a four-decade study), infant mortality will tend to be low, and vice versa. The two variables may in fact be related (and in this case almost certainly are), but the point is that any two time-trended variables will appear to be related in MLMs, even when they're not. Even worse, QMCR variables are time-trended at different rates in different countries. National income rises strongly over time in South Korea (which has been growing rapidly since 1960), moderately in the United States (which has been growing less rapidly), and slowly in Ghana (which has hardly grown at all). Moreover these time trends are not even constant over long periods within countries: China's national income growth was low in the period 1960-1980, moderately high in the period 1980-1995, and very high in the period 1995-2005. Similarly, most developed country national income series inflect in the mid-1970s, and formerly Communist country series around 1990. On top of all this, there are short-term business cycles. Time effects are everywhere.

Such time trends wreck havoc on the estimation of MLM coefficients, and are very difficult to account for through the explicit modeling of error structures. Period effects, time covariates, and autoregressive error structure corrections are all inadequate to account for the kinds of time trends inherent in QMCR data structures. The fact that QMCR variables are trended at different rates in different countries (and at different times) means that time trends must be dealt with on a country-specific basis, and sometimes in complex ways even within countries. One-size-fits-all controls of the kind typically found in QMCR completely fail to adequately adjust for the effects of trended data. As a result, the coefficients on independent variables estimated based on QMCR data are, in almost all cases, heavily biased. It is quite possible that most QMCR studies based on cointegrated or highly trended variables that have used MLM designs have done nothing more than model time. This possibility can be illustrated with an applied example.

Infant mortality is well-known to be closely related to national income per capita (Ross 2006; Babones 2008). They are certainly highly correlated (r = -.84). There is every reason to believe that this relationship is causal: higher national incomes allow countries to purchase improved maternal and newborn health. The cross-sectional estimate derived above in Models 3 and 4 for the relationship between national income on infant mortality was b = -.662, with SE_b = .040. The statistical significance of this relationship is astronomical. National income and infant mortality are about as closely related as any QMCR variables can be.

Nonetheless, a typical MLM of the relationship between national income and infant mortality should yield a non-significant or marginally-significant result. Why? Because we know logically (and, based on Figure 4, empirically) that any causal relationship between national income and infant mortality is not primarily contemporaneous. It is highly unlikely that changes in national income could instantaneously reduce infant mortality. Therefore, wherever national income exhibits a strongly significant contemporaneous effect, we can be reasonably sure that it is due to the strong correlation of both national income and infant mortality with time (or some other common-cause variable), not due to any contemporaneous causal relationship between the variables themselves. In short, the contemporaneous correlation between national income and infant mortality is spurious.

The results of a series of MLM estimates of the contemporaneous relationship between national income and infant mortality are reported in Table 2. A series of commonly used MLM time-trend corrections have been implemented. In particular, five distinct model configurations are considered:

Model 5: GDP is a covariate;
Model 6: GDP covariate plus fixed effect period dummies (early/late);
Model 7: GDP covariate plus fixed effects for each of the 10 time points;
Model 8: GDP covariate plus a continuous time covariate;
Model 9: GDP covariate, both GDP and IM detrended for each country.

Each of these model configurations is estimated in four variants:

(A) a model with fixed country effects with unstructured errors;

(B) a model with fixed country effects with an AR(1) error structure;
(C) a model with random country effects with unstructured errors;
(D) a model with random country effects with an AR(1) error structure.

Only the coefficient of interest (that for national income) and its standard error are reported for each model. All models are estimated on the same unbalanced panel of 1320 cases (country-years spanning the period 1960-2005 at five-year intervals).

Table 2: Comparison of MLMs for Infant Mortality (log), 1960-2005

		Fixed effects for country		Random effects for country	
Model	Configuration	Unstructured (A)	AR(1) structure (B)	Unstructured (C)	AR(1) structure (D)
(5)	GDP	-0.917 (0.027)	-0.452 (0.024)	-0.723 (0.021)	-0.454 (0.020)
(6)	GDP, Period dummy (early/late FE)	-0.586 (0.022)	-0.465 (0.022)	-0.571 (0.017)	-0.456 (0.018)
(7)	GDP, Time (FE)	-0.392 (0.020)	-0.189 (0.018)	-0.465 (0.016)	-0.238 (0.016)
(8)	GDP, Time (covariate)	-0.367 (0.020)	-0.162 (0.017)	-0.449 (0.016)	-0.206 (0.016)
(9)	GDP (both GDP and IM detrended)	-0.026 (0.019)	-0.034 (0.016)	-0.026 (0.018)	-0.035 (0.016)

Notes: Entries in table are metric coefficients (standard errors in parentheses); N=1320

As would be expected, the coefficients for national income are in all cases more significant in the REM variant than in the corresponding FEM, though they are broadly similar in magnitude. This is because the REM design assigns less of the total variability in infant mortality to the country (treatment group) effects and thus more to the variable of interest. This is an example of how the use of REMs can lead to biases in the estimated effects of covariates. The discussion of trend effects to follow focuses on the FEM variants, but the same patterns apply in the REM variants.

The initial model, Model 5(A), shows a fantastically strong relationship between national income and infant mortality (t = -33.599); this is not surprising, since Model 5(A) includes no trend adjustment whatsoever. Allowing for just an AR(1) error autocorrelation in Model 5(B) cuts the estimated effect of national income in half, but still leaves it extraordinarily highly significant (t = -18.875). A popular method of adjusting for trends in QMCR data is to use a period dummy; Models 6(A) and 6(B) show that splitting the observations here into two groups (1960-1975, 1980-2005) has virtually no effect on the estimated coefficient for national income. Even using ten period dummies (one for each time period) has only a minor effect, marginally reducing the coefficients for national income in Models 7(A) and 7(B) but leaving them nonetheless highly significant.

Model 8(A) introduces a linear time covariate; combined with an AR(1) error structure as in Model 8(B) this is a more aggressive trend adjustment than any typically found in the QMCR literature. Even though coefficient for national income reaches its minimum significance yet in Model 8(B), it is still very highly significant. The t-statistic for this model (t = -9.277) corresponds to a significance level of p < 0.0000000000000000001. All of these models would lead one to conclude that there is a strong contemporaneous relationship between national income and infant mortality. Were they published in the literature, it is unlikely that anyone would question the results, since it seems on the face of it that the two variables should be related. In fact, however, as I have argued above they are not -- at least, not contemporaneously.

Model 9(A) demonstrates this. In Model 9(A) the linear time trends in national income and infant mortality have been removed on a country-by-country basis before analysis. Thus, in Model 9(A) we're truly asking whether infant mortality has been higher than average in years that national income has been higher than average, leaving aside the secular trends in each. In Model 9(A) the FEM estimate of the contemporaneous effect of national income on infant mortality is negative, but weak and non-significant (t = -1.403). Model 9(B) shows the corresponding estimate when allowing for an AR(1) error structure. It is slightly larger, and weakly significant (t = -2.085). The staggeringly significant cross-sectional correlation between national income and infant mortality all but disappears in a MLM framework when time trends are eliminated.

The coefficients in Model 9 aren't weak because of some legerdemain in the detrending process; they are weak because the real contemporaneous effect of national income on infant mortality is so weak as to be almost undetectable. Even the small effects that do remain in Model 9 are probably due to imperfect detrending resulting from the fact that the actual time trends in national income and infant mortality may not be linear. The fact that the AR(1) error model improves the significance of the relationship is evidence for this interpretation, since the AR(1) model would allow for any residual trend to decay over time. That is to say: if both series trended until 1990 then flattened out, the linear detrending of the data would actually create a small trend for the period after 1990. An AR(1) error model would better capture this change in trend than would an unstructured error model.

If such a strong effect as that of national income on infant mortality (t = -33.599) disappears when time is removed from the analysis, what of the robustness of other, far more tenuous QMCR relationships? Nearly all QMCR variables trend over time, at least within countries. National income, income inequality, carbon emissions, population health, birth rates, educational levels, industrial output, agricultural employment, unionization, labor productivity, portfolio investment, investment dependence, political freedoms, military spending, tax efficiency, memberships in international organizations, frequencies of protest events, and the like all exhibit secular time trends within countries. As with national income and infant mortality, it is likely that their observed relationships in MLMs are partially or largely the result of inadequate controls for time.

The simple fact is that MLM designs are generally inappropriate for answering the questions that most QMCR practitioners want to ask. They are much better suited to studying the kinds of time-series data structures for which they are commonly used in the economics literature. The basic question asked by MLMs applied to repeated-measures panel data – do X and Y rise and fall together at the same time? – is not a question typically asked by scholars working in the QMCR tradition. Most QMCR is concerned with long-term changes in structural relationships, not short-term fluctuations in annual data. Accordingly, models for QMCR should focus on broad changes over long time periods, not on period-to-period variability, as in MLM designs. It is difficult to imagine a scenario in which QMCR practitioners using MLMs are actually interested in the annual variability in their data. I have never seen such an example published in the QMCR literature.

DIFFERENCE MODELS

Rather than MLMs, Firebaugh and Beck (1994) promoted the use of difference models as an alternative to "panel" models for eliminating spurious causality. In a difference model change in the dependent variable over time is regressed on change in the independent variables. The reasoning is that if variables systematically rise together and fall together, they are related in some way. Any time-invariant covariates of the dependent and independent variables can be safely omitted, since their own difference scores over time will be zero (by construction). Time itself can also be omitted as a confounding influence, since the time difference between the two periods being studied is a constant for all countries in the analysis. The difference model thus at a stroke solves both the time-invariant omitted variable problem inherent in panel designs and the time trend problem inherent in MLM designs.

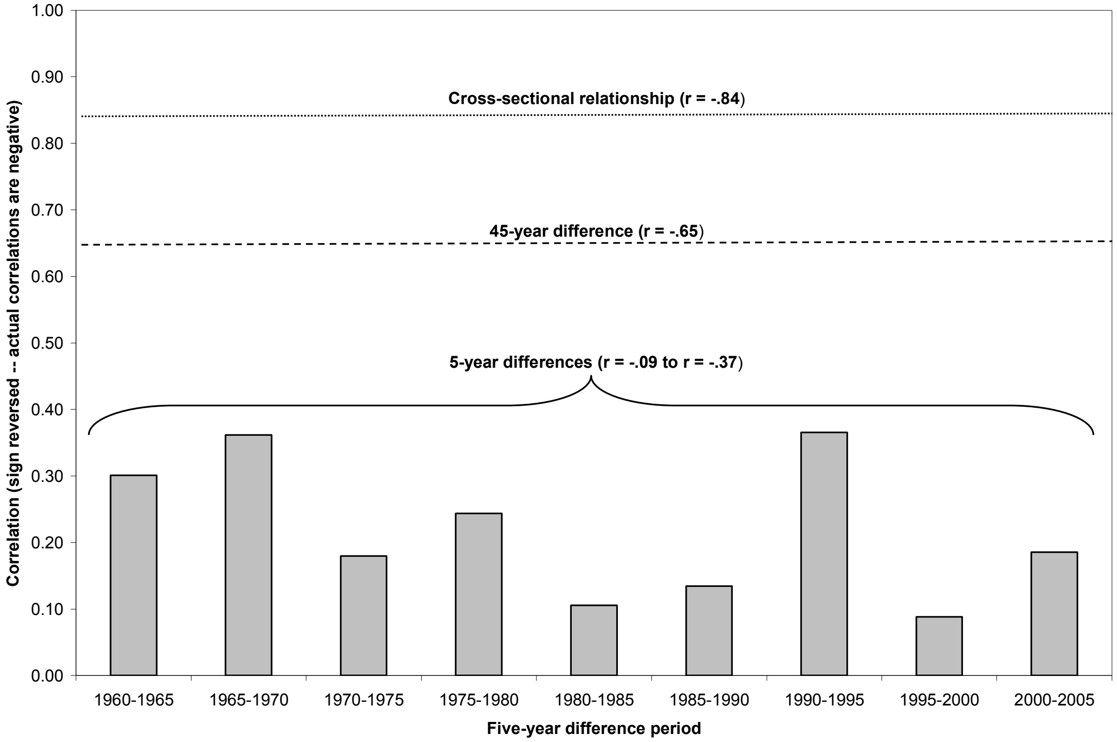
Since variables in the difference model are differenced at the country level, it doesn't (much) matter whether or not variables trend at different rates in different countries. When variables trend at different rates in different countries in the same way across variables (e.g, both slow in Country A but both fast in Country B), trends have no effect on the results. When variables trend in different ways across variables (e.g., the first variable fast and the second variable slow in Country A, but the first variable slow and the second variable fast in Country B), the pattern of trends will reduce the observed relationship between the variables. Thus, difference models are generally conservative from the perspective of time trends.

The price of these advantages is low power and poor data availability. The difference model approach is only effective when the differences can be computed over relatively long time periods (i.e., long enough for meaningful changes to occur in the variables being studied), meaning that often relatively small numbers of cases are available for analysis. The long time periods studied, however, ensure that even lagged relationships can be captured in difference models, since a study period of several decades will encompass the lag periods of most relationships. It is, however, possible to study differences of shorter time periods if desired. It is even possible to construct MLMs out of many short difference periods, though such designs would require extreme care, given their complexity.

A widely quoted myth about the difference model is that it is equivalent to a FEM design with two time periods. As correctly noted by Halaby (2004:515), this is only true when the independent variable in the difference model is a binary (0/1) variable. In nearly all QMCR settings of interest, both independent and dependent variables are continuous. In such cases, the difference model and the FEM can give widely varying results, especially when both the independent and dependent variables are trended over time.

Applying a 45-year difference model (1960-2005) to the regression of infant mortality on national income yields a highly significant effect of b = -0.589 (SE_b = 0.074) based on N=88 cases. This means that long-term increases in national income are strongly related to long-term decreases in infant mortality. This coefficient is not far off that reported in Model 6(A), but the similarity is purely coincidental. The coefficient in Model 6(A) has been shown to have been generated by the time trends in the data; the coefficient for the difference model is unaffected by such trends. In fact, a difference model based on the detrended data differs only in the constant term; the slopes is mathematically identical. The difference model thus clearly and robustly indicates a non-artefactual relationship between national income and infant mortality, independent of potential time-invariant common-cause factors and independent of time trends in the variables. The rate of improvement in national income is strongly related to the rate of improvement in infant mortality across countries, independent of constant country characteristics and independent of the fact that both have generally improved over time.

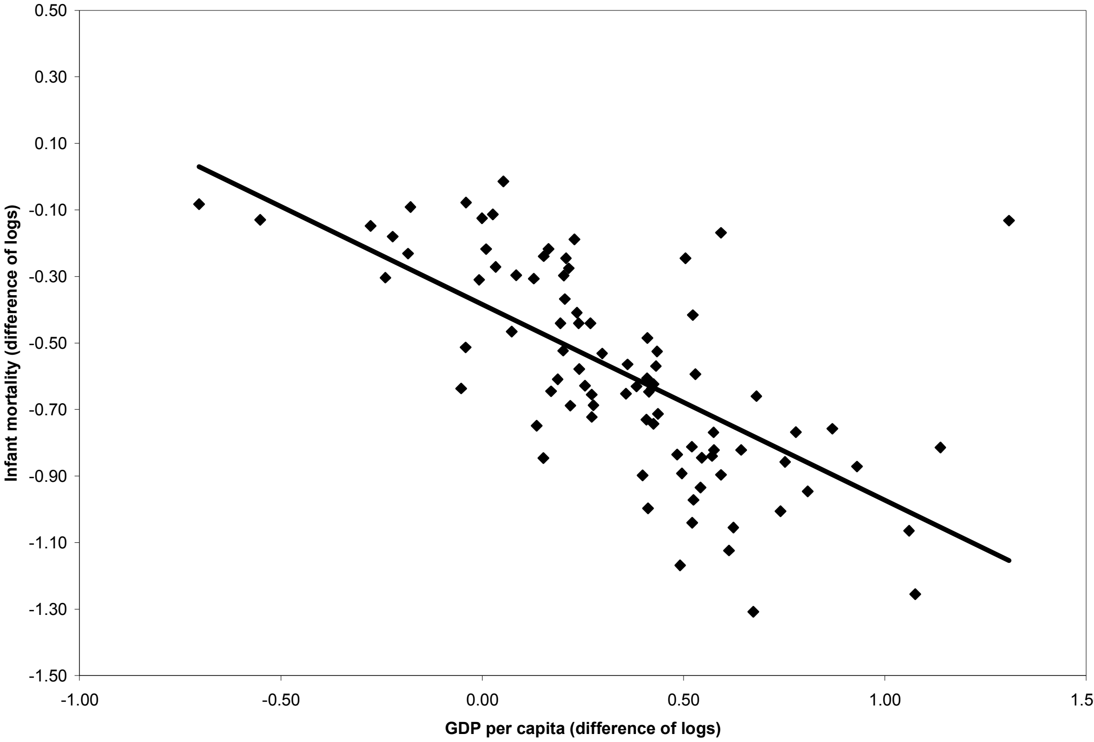
Figure 5: Attenuation and Variability of Difference Model Results using Short Time Intervals



Difference models are not common in sociology, though Firebaugh and Beck (1994) make a strong case for their use. As Firebaugh and Beck argue, the difference model is the direct longitudinal analog of the straightforward cross-sectional model. In a properly-specified cross-sectional model, a dependent variable is regressed on dependent variables that are thought to influence the dependent variable. This model implies that changes in the independent variables should be reflected in changes in the dependent variable. The difference model captures this expected relationship. The key is that the difference model must be estimated over a sufficiently long time period for changes in the independent variables to have an opportunity to manifest their effects on the dependent variable. If the time period over which the variables are differenced is long enough, causal relationships can be detected even when the causal lags are long, as with the relationship between national income and infant mortality.

Differences over short time periods do not adequately capture the over-time relationship between changes in national income and changes in infant mortality, despite the very strong cross-sectional relationship between the variables. For example, the correlation between changes in national income and changes in infant mortality over the 45-year interval examined above is r = -.65. This does not quite reach the r = .84 recorded in the pooled cross-sectional design, but it is still reasonably strong and highly significant statistically. When national income and infant mortality are differenced over 5-year intervals, however, their correlation drops substantially. Their correlations over the nine five-year intervals 1960-2005 are plotted in Figure 5. They range from a high of r = -.37 for the 1990-1995 difference model to a low of r = -.09 in the period immediately following. Out of the nine time intervals studied, the correlations in four are not statistically significant at the p < .05 level. Three are not significant even at the more relaxed p < .10 level.

Figure 6: Difference Model (1960-2005) Regressing Changes in Infant Mortality (log) on Changes in National Income (log)



Even for such a powerful relationship as that between national income and infant mortality, reasonably long time intervals are called for. Note that simply pooling all available 5-year intervals does not have the same effect as using a 40-year interval: the correlation of all available 5-year differences of national income with all available 5-year differences of infant mortality is just $r = -.16$ (N=974). This is a relatively small correlation, though highly significant due to the large number of intervals aggregated.

Some analysts argue that difference models are wasteful in that they ignore large amounts of data that could be analyzed. The data thrown out in difference models, though, are for the most part data that are analytically irrelevant to the problem at hand. Viewed in this way, difference models are not wasteful but parsimonious. The one major practical shortcoming of difference models is that the differencing process often generates outliers and leverage points. The 45-year difference in logged infant mortality is plotted against the 45-year difference in logged national income per capita in Figure 6. There is one obvious leverage point in the data, Botswana, in which national income grew rapidly over the study period but infant mortality declined only slowly. Deleting Botswana from the model yields an even stronger estimate of the effect of national income of $b = -0.695$ ($SE_b = 0.067$). Dealing with leverage points in difference models is no more nor less difficult than dealing with leverage points generally.

RECOMMENDATIONS AND CONCLUSIONS

I conclude with a plea for simplicity. Quantitative macro-comparative research should focus on the effects of as few variables as possible in any one study. Control variables should be used judiciously, with explicit attention given to their actual roles in the models being estimated. It may be possible to estimate the coefficients of twelve variables using data on just twenty countries observed at five time points each, but this doesn't mean that it's advisable. This paper has paid less attention to the virtues parsimony than I might have liked, but hopefully it will be clear from the challenges examined here of correctly understanding the relationship between just two variables that the challenges of correctly understanding the relationships among dozens of variables might be near-insurmountable.

The data used in quantitative macro-comparative research are highly structured in idiosyncratic ways that create many pitfalls for those who analyze them. Obvious and hidden temporal effects are embedded everywhere in QMCR data structures. From this perspective, the difference model is an extremely effective design for QMCR because it eliminates the problem of (linear) time trends in variables and returns analysis full-circle back to a simple cross-sectional model (albeit a cross-sectional model of changes over time). Consequently, it is hard to mess up a difference model. Simple cross-sectional models, including appropriate control variables, should be estimated first, then difference models run as back-ups to help substantiate any claims of causality.

Multilevel models, on the other hand, should be approached with extreme caution. It is my considered opinion that it is quite possible that all reported multilevel model results reported to date in the quantitative macro-comparative research literature are nothing more than spuriously attributed effects of time. I have not explicitly replicated existing MLMs from the published literature, and so I cannot comment on them directly. Published results represent only the tip of the iceberg of the generally careful and comprehensive analyses that underlie any published work.

Researchers who have used these models in the past may in fact have privately tested detrended versions of their analyses, found that the results simply confirmed what they had found in other models, and as a result not considered them publication-worthy. In light of the evidence presented in this paper, however, anyone using such models in the future should certainly examine the possibility that their results are driven by nothing more than trends in their data.

The MLM design, especially in its REM variant, is extremely seductive because it offers boxes with labels for almost everything anyone might want to do in QMCR: control for omitted variables, estimate the effects of time-invariant variables, estimate interaction effects, account for the effects of time, etc. The problem is that the labels on the boxes often bear no intuitive relationship to what the boxes actually do. Researchers should be extraordinarily careful when using MLM designs to assure both themselves and their readers that the coefficients attached to the variables of interest actually mean what they purport to mean. An productive avenue for future research might be the application of Monte Carlo techniques to simulate the behavior of the coefficients of covariates and their standard errors in QMCR using REM designs.

The methodological challenge of undertaking quantitative macro-comparative research is a large part of what makes it exciting and intellectually stimulating. The ultimate reward for most QMCR practitioners, though, is the possibility of changing the world through better understanding how it works. Fortunately or unfortunately, we only have one world to work with, so the burden of better understanding that world largely falls back on methodology. Exculpatory data are rarely forthcoming, so we are generally constrained to argue, rather than experiment or survey, our way out of our problems. Nonetheless, it is important to remember that methodological virtuosity should be exhibited only in the service of improved substantive understanding, and not for its own sake. There is no need for an extensive toolbox when a hammer will do just fine.

Critics and reviewers especially should keep this in mind. A paper that effectively makes its substantive case with a minimum of complexity should be preferred over one that makes the same case with superfluous virtuosity. Very few practitioners of quantitative macro-comparative research fully grasp the mathematical properties of the error models they implicitly assume in their research. It's even less likely that without hands-on access to the data reviewers are equipped to pass judgment on the suitability of highly complicated models. As the examples presented in this paper illustrate, it's probably just as true of quantitative macro-comparative research as it is of the transwarp drive that "the more they overthink the plumbing, the easier it is to stop up the drain."

REFERENCES

Acemoglu, Daron, Simon Johnson, and James A. Robinson. 2001. "The Colonial Origins of Comparative Development: An Empirical Investigation." *American Economic Review* 91:1369-1401.

Angrist, Joshua D., and Alan B. Krueger. 2001. "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments." *Journal of Economic Perspectives* 15(4): 69-85.

Babones, Salvatore J. 2007. "Studying Globalization: Methodological Issues." Pp. 144-161 in *The Blackwell Companion to Globalization*, edited by George Ritzer. Oxford: Blackwell Publishers.

Babones, Salvatore J. 2008. "Income Inequality and Population Health: Correlation and Causality." *Social Science & Medicine* 66:1614-1626.

Beck, Nathaniel, and Jonathan N. Katz. 1995. "What to Do (and not to Do) with Time-Series Cross-Section Data." *American Political Science Review* 89:634-647.

Beck, Nathaniel, and Jonathan N. Katz. 1996. "Nuisance vs. Substance: Specifying and Estimating Time-Series-Cross-Section Models." *Political Analysis* 6:1-36.

Beck, Nathaniel, Jonathan N. Katz, R. Michael Alvarez, Geoffrey Garrett, and Peter Lange. 1993. "Government Partisanship, Labor Organization, and Macroeconomic Performance: A Comgendum." *American Political Science Review* 87:945-948.

Berk, Richard A. 2004. *Regression Analysis: A Constructive Critique*. Thousand Oaks, CA: Sage Publications.

Bornschier, Volker, Christopher K. Chase-Dunn, and Richard Rubinson. 1978. "Cross-National Evidence of the Effects of Foreign Investment and Aid on Economic Growth and Inequality: A Survey of Findings and a Reanalysis." *American Journal of Sociology* 84:651-683.

Chase-Dunn, Christopher K. 1975. "The Effects of International Economic Dependence on Development and Inequality: A Cross-National Study." *American Sociological Review* 40:720-738.

Chase-Dunn, Christopher K. 1989. *Global Formation: Structures of the World-Economy*. Cambridge, MA: Basil Blackwell.

Ebbinghaus, Bernhard. 2005. "When Less is More: Selection Problems in Large-N and Small-N Cross-National Comparisons." *International Sociology* 20:133-152.

Evans, Peter, and James E. Rauch. 1999. "Bureaucracy and Growth: A Cross-National Analysis of the Effects of 'Weberian' State Structures on Economic Growth." *American Sociological Review* 64:748-765.

Everitt, B.S. (ed.). 2002. *The Cambridge Dictionary of Statistics*, 2nd ed. Cambridge: Cambridge University Press.

Firebaugh, Glenn, and Frank D. Beck. 1994. "Does Economic Growth Benefit the Masses? Growth, Dependence, and Welfare in the Third World." *American Sociological Review* 59:631-653.

Halaby, Charles N. 2004. "Panel Models in Sociological Research: Theory into Practice." *Annual Review of Sociology* 30:507-544.

Keele, Luke, and Nathan J. Kelly. 2005. "Dynamic Models for Dynamic Theories: The Ins and Ours of Lagged Dependent Variables." *Political Analysis* 14:186-205.

Kenny, David A. 1979. *Correlation and Causality*. New York: John Wiley & Sons.

Kentor, Jeffrey. 2001. "The Long Term Effects of Globalization on Income Inequality, Population Growth, and Economic Development." Social Problems 48:435-455.

Lucke, Joseph F., and Susan Embretson Whitely. 1984. "The Biases and Mean Squared Errors of Estimators of Multinormal Squared Multiple Correlation." *Journal of Educational Statistics* 9:183-192.

Marquart-Pyatt, Sandra. 2004. "A Cross-National Investigation of Deforestation, Debt, State Fiscal Capacity, and the Environmental Kuznets Curve." *International Journal of Sociology* 34:33–51.

Ross, Michael. 2006. "Is Democracy Good for the Poor?" *American Journal of Political Science* 50:860-874.

Rumberger, R. W. and G. J. Palardy. 2004. "Multilevel Models for School Effectiveness Research." Pp 235-258 in *Handbook on Quantitative Methodology for the Social Sciences*, edited by D. Kaplan. Thousand Oaks, CA: Sage.

Sandholtz, Wayne and Rein Taagepera. 2005. "Corruption, Culture, and Communism." *International Review of Sociology* 15:109-131.

Vogt, W. Paul. 1999. *Dictionary of Statistics and Methodology: A Nontechnical Guide for the Social Sciences*, 2nd ed. Thousand Oaks, CA: Sage Publications.

Wawro, Gregory. 2002. "Estimating Dynamic Panel Data Models in Political Science." *Political Analysis* 10:25-48.

Wilson, Sven E., and Daniel M. Butler. 2007. "A Lot More to Do: The Sensitivity of Time-Series Cross-Section Analyses to Simple Alternative Specifications." *Political Analysis* 15:101-123.

You, Jong-sung, and Sanjeev Khagram. 2005. "A Comparative Study of Inequality and Corruption." *American Sociological Review* 70:136-157.