

Problems with and Solutions for Two-Dimensional Models of Continuous Dependent Variables

Ben Goodrich¹

October 8, 2004

¹Harvard University, Department of Government, Littauer Center (North Yard), 1875 Cambridge St., Cambridge, MA, 02163; email: goodrich@fas.harvard.edu

Abstract

This paper addresses the specification of hierarchical models with continuous dependent variables, such as time-series cross-section models. The pooled OLS model and the random effects estimator implicitly place unreasonable constraints on the coefficients and produce standard errors that are too small due to the structure of the dataset. A two-part estimator that restructures the data is recommended to overcome these problems and its usefulness is demonstrated with an empirical example.

1 Introduction

How should we specify linear regression models when the data vary across two dimensions, such as space and time? It is impossible to say what should be done in all situations, but it is easier to determine what should *not* be done. This paper gives some strong recommendations as to what should not be done, and offers some alternatives that can be used, at least when the dependent variable is continuous.

The pros and cons of fixed effects models, random effects models, and pooled models have been debated ad nauseam in the literature. While no consensus has emerged as to which model is better, a great many principles and rules of thumb have become generally accepted. For example, many believe that fixed effects models are always less efficient than random effects models and pooled models. Many also believe that it is necessary to use a random effects model or a pooled model to estimate the effects of variables that do not vary within units of observation. Many believe that fixed effects models are atheoretical, and so on. I disagree with all of these statements and offer a new perspective on such issues.

The stakes in this debate are enormous. As data have become easier to collect, models for two-dimensional data have become more prevalent. Many of the conclusions of this paper apply to all such models, not just the particular hierarchical models known as “time-series cross-section” models. Nevertheless, the paradigmatic examples come from comparative and international political economy, where each country (or country-pair) is observed over time (usually years). But scholars of American politics utilize two-dimensional models when counties are nested within states, justices within circuit courts, etc. The conclusions also apply throughout the social sciences – economic models where firms are nested within industries, policy models where students are nested within schools, and sociology models where individuals are nested within families to name a few.

The problems I identify stem from the structure of two-dimensional data, rather than the

estimators for two-dimensional data. However, some estimators structure the data properly, while most do not, which causes the coefficient estimates and the standard errors to be biased. Everyone, I am sure, has witnessed an undergraduate who upon discovering that a coefficient is insignificant, copies all the observations and appends them to the bottom of the dataset so that the coefficient “becomes” significant. The novel point of this paper is that researchers implicitly utilize the same trick when the data are two-dimensional. I focus on various least squares estimators, so I do not discuss solutions for models with discrete dependent variables, nor do I discuss models with three or more dimensions.

2 Variance in two dimensional models

This section makes an important distinction between two types of variance and establishes the majority of the notation. When there are two dimensions, there are multiple ways to calculate the expectation of a variable and thus multiple types of variance. For concreteness, I focus on a subsample of an international relations dataset that is used in Green, Kim and Yoon (2001). I refer to all datasets where micro-*observations* are nested within macro-*units* of observation as two-dimensional datasets or variables therein.

Green, Kim, and Yoon’s dataset contains dyadic data from 1952 to 1992. Dyads, not individual states, are the macro-units of observation, and dyad-years are the micro-observations. Dyadic data raise unique problems if the relationship under study is not symmetrical, but I ignore those issues in this paper. Let $i = 1, 2, \dots, N = 271$ index the dyads and let $t = 1, 2, \dots, T = 41$ index time. The dataset analyzed in Green, Kim and Yoon (2001) is actually much larger, but I restrict my analysis to the 271 dyads that have a full set of 41 observations each. Allowing for unbalanced data complicates the math and distracts attention from the overall points of this paper, so I relegate its discussion to the footnotes.¹

¹Any unbalanced dataset can be balanced using the technique of multiple imputation (see King et al., 2001; Little and Rubin, 2002). However, as far as I am aware, no one has written a paper that discusses the

Let a and b indicate the two states in a dyad. The dependent variable in Green, Kim, and Yoon (2001) is the logarithm of bilateral trade (*TRADE*) between a and b . The model is a “gravity model”, so the logarithm of the distance (*DISTANCE*) between a and b is included on the right-hand side among the following two-dimensional variables: the sum of logarithms of the two gross domestic products (*GDP*), the sum of the logarithms of the two populations (*POPULATION*), an indicator for the presence of a formal alliance (*ALLIANCE*), and the minimum level of democracy between a and b (*DEMOCRACY*).

The notation used in this paper is consistent with common practice wherever possible. I can say from the writing experience that it is easy to get confused if subscripts and other notational details are overlooked. Capital boldface letters represent matrices (\mathbf{X}_{it}). Lower-case boldface letters indicate column vectors (\mathbf{x}_{it}); individual elements of those vectors have no boldface (x_{it}). The sample mean of a variable, \bar{x} , is simply equal to $\frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T x_{it}$ but is distinct from the vector of *unit-specific* means, $\bar{\mathbf{x}}_i$, each element of which is equal to $\frac{1}{T} \sum_{t=1}^T x_{it}$. This vector of unit-specific means should be of length N . However, in two-dimensional datasets we usually copy each unit-specific mean T times, making the length of the vector NT .

The total variance of \mathbf{x}_{it} is simply $\frac{1}{NT-1} \sum_{i=1}^N \sum_{t=1}^T (x_{it} - \bar{x})^2$. However, total variance can be decomposed into within-unit variance and between-unit variance. Within-unit variance is calculated around the unit-specific means (rather than the sample mean) and is equal to $\frac{1}{NT-N} \sum_{i=1}^N \sum_{t=1}^T (x_{it} - \bar{x}_i - 0)^2$. The extra zero is included to emphasize that the expectation of $(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)$ is exactly zero since the expectation is zero for each unit.

From this point forward, I use the tilde notation to represent a two-dimensional variable that is expressed in terms of deviations from the unit-specific means. Thus, $\tilde{\mathbf{x}}_{it} = \mathbf{x}_{it} - \bar{\mathbf{x}}_i$ and the *total* variance of $\tilde{\mathbf{x}}_{it}$ equals the *within-unit* variance of \mathbf{x}_{it} . Within-unit variance should be normalized by $NT - N$ because N unit-specific means are estimated when within-unit issues that arise for between estimation in the context of a multiple imputed dataset.

Table 1: Descriptive statistics for a subsample of data used in Green, Kim, and Yoon (2001)

Variable	Sample Mean	Standard Deviation
Dependent variable: $\ln(Trade^{[ab]})_{it}$	18.207	2.952
$\ln(Distance^{[ab]})_i$	7.960	1.101
$\ln(GDP^{[a]} \times GDP^{[b]})_{it}$	48.439	3.045
$\ln(Population^{[a]} \times Population^{[b]})_{it}$	32.411	2.375
$(Alliance\ Dummy^{[ab]})_{it}$	0.279	0.449
$\min(Democracy^{[a]}, Democracy^{[b]})_{it}$	5.636	6.028

Note: The superscripts a and b denote the states in the dyad; they are not exponents. There are a total of 93,924 observations in Green, Kim, and Yoon’s dataset. This table and the rest of this paper are solely based on the 271 dyads that have 41 observations each, yielding a total for the subsample of 11,111 observations.

unit variance is calculated.² Between-unit variance is defined as $\frac{1}{N-1} \sum_{i=1}^N (\bar{x}_i - \bar{x})^2$, and is the variation in the unit-specific means. When time is one of the dimensions, within-unit variance is also called temporal variance or within variance, and between-unit variance is called cross-sectional variance or between variance.

Table 1 gives the sample means and standard deviations of the variables in the example dataset. The distance between a and b varies between dyads but does not vary within dyads. Time invariant variables, such as distance, have been the subject of much methodological consideration. In short, if unique intercepts (“fixed effects”) are included for each unit, then time invariant variables cannot be included in the model due to perfect collinearity. Since time invariant variables are sometimes substantively interesting, many have turned to alternatives to fixed effects models. The rest of this paper will emphasize a point made in Zorn (2001) that the “problem” of time invariant variables is a red herring.

²Paul Rathouz has written a correction to Stata’s procedure for calculating within and between variance (see <http://health.bsd.uchicago.edu/rathouz/HS333/xtsumcorr.ado>).

3 Coefficient estimates with two-dimensional data

This section discusses a general linear model for two-dimensional data, several models that derive from it, and the coefficient estimates that each model produces. Section 4 discusses the standard errors. The general model was introduced in the political science literature by Zorn (2001) but has been derived in other disciplines (see Neuhaus and Kalbfleisch, 1998; Gould, 2001). Zorn’s model is useful because it allows other models for two-dimensional data to be put into a common framework. Other two-dimensional models place different restrictions on Zorn’s model, but some restrictions are much better than others.

The pooled linear model,

$$\mathbf{y}_{it} = \alpha + \mathbf{X}_{it}\boldsymbol{\beta} + \boldsymbol{\epsilon}_{it}, \tag{1}$$

differs from the standard textbook regression model only in the sense that the data are two-dimensional. The $NT \times K$ matrix of explanatory variables, \mathbf{X}_{it} , must include at least one two-dimensional variable but can include variables that are time-invariant or variables that vary over time only.³ Such one-dimensional variables can nevertheless be considered two-dimensional variables that happen to have zero variance along one dimension. A typical two-dimensional dataset has the following form:

³Variables that vary over time occur only when the data are balanced. “Year effects” are one example, but these do have cross-sectional variation if the data are unbalanced, which suggests a problem for the many papers that use such year effects to control for an unobserved temporal variable in unbalanced samples.

Row	Unit	Time	y	Intercept	$\mathbf{x}^{[1]}$...	$\mathbf{x}^{[K]}$
1	1	1	y_{11}	1	$x_{11}^{[1]}$...	$x_{11}^{[K]}$
2	1	2	y_{12}	1	$x_{12}^{[1]}$...	$x_{12}^{[K]}$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	...	\vdots
T	1	T	y_{1T}	1	$x_{1T}^{[1]}$...	$x_{1T}^{[K]}$
$T + 1$	2	1	y_{21}	1	$x_{21}^{[1]}$...	$x_{21}^{[K]}$
$T + 2$	2	2	y_{22}	1	$x_{22}^{[1]}$...	$x_{22}^{[K]}$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	...	\vdots
NT	N	T	y_{NT}	1	$x_{NT}^{[1]}$...	$x_{NT}^{[K]}$

and is observationally equivalent to this dataset, once it is recognized that $\mathbf{y}_{it} = \bar{\mathbf{y}}_i + \tilde{\mathbf{y}}_{it}$ and $\mathbf{x}_{it} = \bar{\mathbf{x}}_i + \tilde{\mathbf{x}}_{it}$:

Row	Unit	Time	y	Intercept	$\mathbf{x}^{[1]}$...	$\mathbf{x}^{[K]}$
1	1	1	$\bar{y}_1 + \tilde{y}_{11}$	1	$\bar{x}_1^{[1]} + \tilde{x}_{11}^{[1]}$...	$\bar{x}_1^{[K]} + \tilde{x}_{11}^{[K]}$
2	1	2	$\bar{y}_1 + \tilde{y}_{12}$	1	$\bar{x}_1^{[1]} + \tilde{x}_{12}^{[1]}$...	$\bar{x}_1^{[K]} + \tilde{x}_{12}^{[K]}$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	...	\vdots
T	1	T	$\bar{y}_1 + \tilde{y}_{1T}$	1	$\bar{x}_1^{[1]} + \tilde{x}_{1T}^{[1]}$...	$\bar{x}_1^{[K]} + \tilde{x}_{1T}^{[K]}$
$T + 1$	2	1	$\bar{y}_2 + \tilde{y}_{21}$	1	$\bar{x}_2^{[2]} + \tilde{x}_{21}^{[1]}$...	$\bar{x}_2^{[K]} + \tilde{x}_{21}^{[K]}$
$T + 2$	2	2	$\bar{y}_2 + \tilde{y}_{22}$	1	$\bar{x}_2^{[2]} + \tilde{x}_{22}^{[1]}$...	$\bar{x}_2^{[K]} + \tilde{x}_{22}^{[K]}$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	...	\vdots
NT	N	T	$\bar{y}_N + \tilde{y}_{NT}$	1	$\bar{x}_N^{[N]} + \tilde{x}_{NT}^{[1]}$...	$\bar{x}_N^{[K]} + \tilde{x}_{NT}^{[K]}$

If you are skeptical of this reformulation, consider the vector $[1, 2, 3]$, which can be written as $[\bar{2} + (\tilde{-1}), \bar{2} + \tilde{0}, \bar{2} + \tilde{1}]$. Zorn (2001) shows that each two-dimensional variable in \mathbf{X}_{it} can be parsed into a component that has between variation only and a component that has

within variation only:

$$\mathbf{y}_{it} = \bar{y}_i + (\mathbf{y}_{it} - \bar{y}_i) = \alpha + (\bar{\mathbf{X}}_i + (\mathbf{X}_{it} - \bar{\mathbf{X}}_i)) \boldsymbol{\beta} + (\bar{\boldsymbol{\epsilon}}_i + (\boldsymbol{\epsilon}_{it} - \bar{\boldsymbol{\epsilon}}_i)), \quad (2a)$$

$$\mathbf{y}_{it} = \bar{y}_i + \tilde{\mathbf{y}}_{it} = \alpha + \bar{\mathbf{X}}_i \boldsymbol{\beta}_b + \tilde{\mathbf{X}}_{it} \boldsymbol{\beta}_w + \bar{\boldsymbol{\epsilon}}_i + \tilde{\boldsymbol{\epsilon}}_{it}. \quad (2b)$$

This generalization of the pooled linear model turns on the very simple observation that $\mathbf{X}_{it} = \bar{\mathbf{X}}_i + (\mathbf{X}_{it} - \bar{\mathbf{X}}_i) = \bar{\mathbf{X}}_i + \tilde{\mathbf{X}}_{it}$ and similarly for the other terms. The pooled OLS estimator imposes the constraint that $\boldsymbol{\beta}_b = \boldsymbol{\beta}_w$, or in other words that a given increase in $\bar{\mathbf{x}}^{[1]}$ has the same effect as an equivalent increase in $\tilde{\mathbf{x}}^{[1]}$ and similarly for the other two-dimensional covariates in the dataset. Zorn (2001) merely relaxes those constraints for any – or in this case all – of the independent variables.

The columns in $\bar{\mathbf{X}}_i$, which I call “meaned variables”, have no temporal variation, and $\boldsymbol{\beta}_b$ represents the (cross-sectional) “between” effects of the variables in \mathbf{X}_{it} . Conversely, the columns in $\tilde{\mathbf{X}}_{it}$, which I call “demeaned variables”, have no between variation because the between variation in \mathbf{X}_{it} is removed by subtracting $\bar{\mathbf{X}}_i$. Thus, $\boldsymbol{\beta}_w$ represents the (temporal) “within” effects of the variables in \mathbf{X}_{it} . After making the transformations suggested in Zorn (2001), the dataset has this form:

Row	Unit	Time	y	Intercept	$\bar{\mathbf{x}}^{[1]}$	$\tilde{\mathbf{x}}^{[1]}$...	$\bar{\mathbf{x}}^{[K]}$	$\tilde{\mathbf{x}}^{[K]}$
1	1	1	$\bar{y}_1 + \tilde{y}_{11}$	1	$\bar{x}_1^{[1]}$	$\tilde{x}_{11}^{[1]}$...	$\bar{x}_1^{[K]}$	$\tilde{x}_{11}^{[K]}$
2	1	2	$\bar{y}_1 + \tilde{y}_{12}$	1	$\bar{x}_1^{[1]}$	$\tilde{x}_{12}^{[1]}$...	$\bar{x}_1^{[K]}$	$\tilde{x}_{12}^{[K]}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	...	⋮	⋮
T	1	T	$\bar{y}_1 + \tilde{y}_{1T}$	1	$\bar{x}_1^{[1]}$	$\tilde{x}_{1T}^{[1]}$...	$\bar{x}_1^{[K]}$	$\tilde{x}_{1T}^{[K]}$
T + 1	2	1	$\bar{y}_2 + \tilde{y}_{21}$	1	$\bar{x}_2^{[1]}$	$\tilde{x}_{21}^{[1]}$...	$\bar{x}_2^{[K]}$	$\tilde{x}_{21}^{[K]}$
T + 2	2	2	$\bar{y}_2 + \tilde{y}_{22}$	1	$\bar{x}_2^{[1]}$	$\tilde{x}_{22}^{[1]}$...	$\bar{x}_2^{[K]}$	$\tilde{x}_{22}^{[K]}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
NT	N	T	$\bar{y}_N + \tilde{y}_{NT}$	1	$\bar{x}_N^{[1]}$	$\tilde{x}_{NT}^{[1]}$...	$\bar{x}_N^{[K]}$	$\tilde{x}_{NT}^{[K]}$

One important point is that meaned variables have no covariance with demeaned variables. Intuitively, meaned variables only have between variance while demeaned variables only have within variance, so there can be no cross-dimensional covariance. This intuition can be proven for the two variable case:

$$E [\tilde{\mathbf{x}}_{it}] = \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T \tilde{x}_{it} = 0, \quad (3)$$

$$E [\bar{\mathbf{x}}_i \times \tilde{\mathbf{x}}_{it}] = \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T (\bar{x}_i \times \tilde{x}_{it}) = 0, \quad (4)$$

$$Cov(\bar{\mathbf{x}}, \tilde{\mathbf{x}}) = E [\bar{\mathbf{x}}_i \times \tilde{\mathbf{x}}_{it}] - E [\bar{\mathbf{x}}_i] E [\tilde{\mathbf{x}}_{it}] = 0. \quad (5)$$

A demeaned variable, when summed over time, equals zero by definition. Thus, multiplying a demeaned variable by scalars that do not vary over time, such as \bar{x}_i , and then summing over time also results in a zero. Therefore, the pairwise covariance between a meaned variable and demeaned variable is zero. Proving the orthogonality of meaned and demeaned variables in the multivariate case is left as an exercise for the reader but can readily be observed from the example in section 5.

There are two important implications to orthogonality. The first is that meaned (demeaned) variables cannot explain the temporal (cross-sectional) variance in the dependent variable because there is no cross-dimensional covariance. Second, meaned (demeaned) variables can be dropped from equation 2b without affecting the temporal (cross-sectional) estimates.

The obvious question is, “Why would the between effect of explanatory variable k not be equal to the within effect of k ?”. That they should not differ is one of the few points on which Green, Kim and Yoon (2001) and Oneal and Russett (2001, p.481) agree. However, a better question is “Why not estimate a more general model and check?” Zorn (2001) and Ray (2003) provide several examples where one might expect between and within effects to

differ, either in sign or magnitude. To me, the most intuitive example comes from Gould (2001): Suppose the dependent variable is a sample of Americans’ wages over time, and the independent variables are regional dummies with the northeast excluded. Wages in southern states are lower than in the northeast, on average, and the cross-sectional effect of the “South” dummy variable is expected to be negative. However, if people *move* to the south from the northeast, they are likely taking better-paying jobs. Thus, the temporal effect of the “South” dummy variable is expected to be positive.

Gould’s example has a philosophical element. If we had a “perfect” model and “perfect” data – in the sense that we have all the relevant variables that determine wages, including all the competing job offers people receive – then maybe the temporal effect of the “South” dummy variable would not be positive. Although it is neither possible nor necessary to gather data on all relevant variables, we should always admit the possibility that a specification error could drive a wedge between $\hat{\beta}_b$ and $\hat{\beta}_w$ for at least *one* of the explanatory variables, which may be the explanatory variable(s) of interest or a control variable that is correlated with the explanatory variable(s) of interest. The rest of this paper justifies the claim that we should *always* estimate models that allow the between and within effects of variables to differ.

In so doing, I explicitly allow the right-hand side of the model to include lagged variables. The claims in Zorn (2001) apply to all two-dimensional models, but for the rest of section 3, I focus specifically on TSCS models. Beck and Katz (1996) recommends the following two-dimensional version of the “auto-regressive distributed lag” (ARDL) model as a general starting point from which to “test down”:

$$\mathbf{y}_{it} = \alpha + \phi \mathbf{y}_{it-1} + \mathbf{X}_{it} \boldsymbol{\beta} + \mathbf{X}_{it-1} \boldsymbol{\gamma} + \boldsymbol{\epsilon}_{it}, \quad |\phi| < 1, \quad (6)$$

The first number in ARDL(**1**,1) notation indicates that the right-hand side includes one lag of the dependent variable (\mathbf{y}_{it-1}). The second number indicates that the right-hand

side includes one lag of the exogenous variables (\mathbf{X}_{it-1}). This ARDL model also includes the contemporaneous exogenous variables (\mathbf{X}_{it}) and a *single* intercept (α). Beck and Katz (2001) elaborates that N intercepts should be estimated when they are “necessary” – provided that no time invariant variables are of substantive interest – and that the Bayesian Information Criterion, rather than a F test, should be used to judge necessity.

The ARDL model has a long history in the econometrics literature for single time-series, but is not immune from criticism. The ARDL model assumes that the error term is uncorrelated with current values of \mathbf{X}_{it} , past values of \mathbf{X}_{it} , and *future* values of \mathbf{X}_{it} . This is a very strong assumption, but is arguably more plausible in political science where the units of analysis are presumed less capable than economic agents to anticipate the future rationally (although I cannot recall a political science paper that justified this assumption before using an ARDL model). I focus on the ARDL model because it has been specifically recommended for political scientists and use it as an example of what can go wrong when the data are two-dimensional – even when the strong assumptions are met, T is relatively large, etc. However, the problems that arise due to the two-dimensional structure of the data are likely to crop up even if the ARDL model is abandoned in favor of an alternative model.

Wilson and Butler (2003, table 1) claims that 135 published papers for linear TSCS models have cited Beck and Katz (1995) or Beck and Katz (1996) as of May 31, 2003. Of those, 59 use a specification like that in equation 6 (possibly without the \mathbf{X}_{it-1} term), 32 consider some other dynamic structure, while 44 use a specification like equation 1, is a constrained version of equation 6. Thus, it is important to determine if these procedures are sound. To be fair, 47 papers report fixed effects estimates, but my hunch is that pooled specifications were usually given more emphasis.

There are two strategies to further generalize Beck and Katz’s recommended model. The first is to include variables that have been lagged two or more periods, a specification that is not specifically addressed in this paper but raises few methodological issues beyond those

discussed here. The second is to relax the “parameter constancy assumption”, which can be done in a variety of ways, but I focus on the way Zorn (2001) relaxes the parameter constancy assumption.

Zorn’s model lacks a name, so I will call it the simultaneous parsed model (SPM) – *parsed* because the effects of the covariates are split into their between and within components and *simultaneous* because the between and within estimates are obtained at the same time (rather than consecutively, which is discussed in section 4.2). The ARDL(1,1) model imposes the constraints that $\phi_b = \phi_w$, $\beta_b = \beta_w$, and $\gamma_b = \gamma_w$ on this SPM:

$$\mathbf{y}_{it} = \bar{\mathbf{y}}_i + \tilde{\mathbf{y}}_{it} = \begin{array}{l} \alpha + \phi_b \bar{\mathbf{y}}_{i[t-1]} + \bar{\mathbf{X}}_i \beta_b + \bar{\mathbf{X}}_{i[t-1]} \gamma_b + \bar{\boldsymbol{\epsilon}}_i \\ + \phi_w \tilde{\mathbf{y}}_{it-1} + \tilde{\mathbf{X}}_{it} \beta_w + \tilde{\mathbf{X}}_{it-1} \gamma_w + \tilde{\boldsymbol{\epsilon}}_{it}. \end{array} \quad (7)$$

When lagged variables are used, data are lost. The example dataset actually starts in 1951 but when one-period lags are included in the model, the viable observations start in 1952. It is important to note that $\bar{\mathbf{y}}_{i[t-1]}$ and $\bar{\mathbf{X}}_{i[t-1]}$ represent the “unit-specific means of the lag” not the “lag of the unit-specific means” (which would not make sense). Also, $\bar{\mathbf{y}}_{i[t-1]} \neq \bar{\mathbf{y}}_i$ and $\bar{\mathbf{X}}_{i[t-1]} \neq \bar{\mathbf{X}}_i$ in general because the indices of summation are different. In this paper, $t = 0$ indicates data from 1951 and

$$\bar{y}_{i[t-1]} = \frac{1}{T} \sum_{t=0}^{T-1} y_{it} \neq \frac{1}{T} \sum_{t=1}^T y_{it} = \bar{y}_i, \quad (8)$$

$$\bar{x}_{i[t-1]} = \frac{1}{T} \sum_{t=0}^{T-1} x_{it} \neq \frac{1}{T} \sum_{t=1}^T x_{it} = \bar{x}_i. \quad (9)$$

Analogously, $\tilde{\mathbf{y}}_{it-1}$ and $\tilde{\mathbf{X}}_{it-1}$ represent the deviations of the lagged variables from the unit-specific means of the lagged variables.

Unfortunately, the literature often conflates models with estimators. In truth, all the

models discussed in this paper are usually estimated by least squares. But to adhere with the terminology of the literature, from this point forward, I use the terms model and estimator interchangeably. For example, the SPM can be seen as the sum of the between estimator (B-E) and the within estimator (W-E) respectively once it is recognized that $\mathbf{y}_{it} = \bar{\mathbf{y}}_i + \tilde{\mathbf{y}}_{it}$:

$$\bar{\mathbf{y}}_i = \alpha + \phi_b \bar{\mathbf{y}}_{i[t-1]} + \bar{\mathbf{X}}_i \boldsymbol{\beta}_b + \bar{\mathbf{X}}_{i[t-1]} \boldsymbol{\gamma}_b + \bar{\boldsymbol{\epsilon}}_i, \quad (10)$$

$$\tilde{\mathbf{y}}_{it} = 0 + \phi_w \tilde{\mathbf{y}}_{it-1} + \tilde{\mathbf{X}}_{it} \boldsymbol{\beta}_w + \tilde{\mathbf{X}}_{it-1} \boldsymbol{\gamma}_w + \tilde{\boldsymbol{\epsilon}}_{it}. \quad (11)$$

Due to the orthogonality of meaned variables and demeaned variables, *estimating the SPM will produce the same estimates as would estimating the B-E and the W-E separately* (although the standard errors differ). In textbooks, the B-E never includes the unit-specific means of lagged variables for good reason. As is discussed in section 3.1, we should drop $\bar{\mathbf{y}}_{i[t-1]}$ and $\bar{\mathbf{X}}_{i[t-1]}$ from the B-E, but I include them to demonstrate methodological points.

The W-E is substantively equivalent to the “least squares dummy variable estimator” (LSDVE) and both are referred to as fixed effects models. The LSDVE,

$$\mathbf{y}_{it} = \boldsymbol{\alpha}_i + \phi_w \mathbf{y}_{it-1} + \mathbf{X}_{it} \boldsymbol{\beta}_w + \mathbf{X}_{it-1} \boldsymbol{\gamma}_w + \tilde{\boldsymbol{\epsilon}}_{it}, \quad (12)$$

looks like an ARDL(1,1), except for the fact that the LSDVE estimates N intercepts ($\boldsymbol{\alpha}_i$) rather than one intercept (α). The coefficient estimates from the LSDVE should be interpreted as within estimates since they are identical to those produced by the W-E.

The error term in a LSDVE does not have a unit-specific component because the N intercepts control for the omitted variables that would otherwise constitute $\bar{\boldsymbol{\epsilon}}_i$. The random effects estimator (REE) does not include N intercepts and assumes that $\bar{\boldsymbol{\epsilon}}_i$ has a normal distribution and is uncorrelated with the independent variables. A REE,

$$\tilde{\mathbf{y}}_{it}^* = \alpha + (0) \tilde{\mathbf{y}}_{it-1}^* + \tilde{\mathbf{X}}_{it}^* \boldsymbol{\beta} + \tilde{\mathbf{X}}_{it-1}^* \boldsymbol{\gamma} + \bar{\boldsymbol{\epsilon}}_i + \tilde{\boldsymbol{\epsilon}}_{it}, \quad (13)$$

usually does not include a lagged dependent variable because it would be correlated with $\bar{\epsilon}_i$ by definition. In equation 13, $\tilde{\mathbf{y}}_{it}^* = \mathbf{y}_{it} - \hat{\omega}\bar{\mathbf{y}}_i$, and the other variables are defined similarly. The fractional parameter, $\hat{\omega}$, equals $1 - \sqrt{\frac{\widehat{Var}(\tilde{\epsilon}_{it})}{\widehat{Var}(\bar{\epsilon}_i)}}$, where the variances of the error terms are estimated to be the variances of the residuals in a W-E and a B-E. If $\hat{\omega} = 1$, the REE reduces to the W-E. If $\hat{\omega} = 0$, the REE reduces to a pooled OLS estimator (POLSE).

All of the models that will be discussed stem from the ARDL(1,1) model suggested in Beck and Katz (1996). The SPM is the most general and is the sum of the B-E and the W-E – the LSDVE being equivalent to the W-E. The REE is a quasi-W-E, but how far the REE departs from the POLSE depends on the ratio of the residual variances in the W-E and B-E. The POLSE has the strongest assumptions since it is a SPM where the between effects of all the independent variables are constrained to equal the corresponding within effects.

3.1 Problems with the pooled estimator

There are three ways to conceptualize estimating the ARDL(1,1) model with the POLSE. The common view, as is typified by Green, Kim and Yoon (2001), is to claim that the ARDL(1,1) model is a constrained version of the LSDVE, where all the unit-specific intercepts are constrained to be equal, and these constraints can be tested with an F test.⁴ Another perspective is that a pooled model is a limiting case of the REE as $\hat{\omega}$ approaches zero, which is tantamount to the null hypothesis of the Breusch-Pagan test for unique intercepts.⁵ Whether the unit effects are “random” or “fixed” is often assessed by a Hausman test.⁶ The third view, which I would like to emphasize, is that the POLSE imposes J re-

⁴An F test is a joint test of the null hypothesis that all the intercepts are equal. It plugs the sum of the squared t statistics for all N intercepts into a F distribution with numerator degrees of freedom equal to $N - 1$ and denominator degrees of freedom equal to $NT - N - K - 1$.

⁵The Breusch-Pagan test is usually couched in terms of the variance of the individual effects being zero, but there are multiple valid ways to motivate it. According to Greene (2000, p.572), this Lagrange multiplier statistic is equal to $\frac{NT}{2(T-1)} \left[\frac{\sum_{i=1}^N (T\bar{e}_i)^2}{\sum_{i=1}^N \sum_{t=1}^T (e_{it})^2} - 1 \right]^2$ and has a χ_1^2 distribution.

⁶Let \mathbf{D} be a vector that contains the difference between the estimates from the W-E and the estimates from the REE and let \mathbf{Z} be a matrix that includes all the variables in the model. The tilde notation indicates

restrictions on the SPM that $\phi_w = \phi_b$, $\beta_w = \beta_b$, and $\gamma_w = \gamma_b$. The basic problem with the two-dimensional ARDL(1,1) model is that these J constraints are not justifiable.

The J restrictions can be “tested” with a Wald test, but Bartels (1996), among others, has two valid criticisms of the entire “test down” philosophy.⁷ First, the significance level for the test is arbitrary. Bayesians tend to believe that restrictions should be held to higher standards as the size of the sample increases, which is an important issue because many two-dimensional datasets are very large. Second, the uncertainty in the published estimates will not reflect the pretest uncertainty over whether the constraints should be imposed.

These criticisms can be mitigated by using the Bayesian Information Criterion (BIC) to “evaluate” restrictions. The BIC can be defined as $-2\ell^* + d \ln (NT)$ where ℓ^* is the log-likelihood at the maximum likelihood (ML) estimates and d is the dimension of the Hessian. Alternatively, the BIC can be defined as $(NT) \ln (1 - R^2) + (d - 1) \ln (NT)$, where R^2 is the unadjusted proportion of explained variance. The BIC makes inference a function of the sample size and can provide probabilities for model averaging (see Raftery, 1995, for additional details). Under this view, restrictions may be imposed if the restricted model has a smaller BIC.

Let the J constraints the POLSE imposes be written as a system of equations, such as:

$$\begin{array}{cccccccccccc} (0) \alpha & + & (1) \phi_w & + & (-1) \phi_b & + & \dots & (0) \gamma_w^{[K]} & + & (0) \gamma_b^{[K]} & = & 0, \\ \vdots & & \vdots & & \vdots & & \vdots & \vdots & & \vdots & & \vdots \\ (0) \alpha & + & (0) \phi_w & + & (0) \phi_b & + & \dots & (1) \gamma_w^{[K]} & + & (-1) \gamma_b^{[K]} & = & 0. \end{array}$$

This system of J equations can be written as $\mathbf{R}\boldsymbol{\theta}_{SPM} = \mathbf{0}$ where \mathbf{R} is a matrix of restrictions that the model is a W-E and the asterisk notation indicates that the model is a REE. Then, according to Greene (2000, p.577), the test statistic, W , equals $\mathbf{D}' [\tilde{\mathbf{Z}}' \tilde{\mathbf{Z}} - (\mathbf{Z}^*)' \mathbf{Z}^*]^{-1} \mathbf{D}$ and has a χ_K^2 distribution where K is the length of \mathbf{D} .

⁷Gould (2001) notes that the Hausman test is asymptotically equivalent to an F test that the between effects of all two-dimensional variables equal their corresponding within effects in a SPM, which avoids a problem with the Hausman test that $[\tilde{\mathbf{Z}}' \tilde{\mathbf{Z}} - (\mathbf{Z}^*)' \mathbf{Z}^*]$ can fail to be positive definite. However, in section 4, I show that the variance-covariance matrix of the SPM is flawed in a way that overstates the precision with which the cross-sectional effects are estimated.

that contains the numbers in parentheses, and $\boldsymbol{\theta}_{SPM}$ is a column vector that stacks the K parameters in the SPM (α , ϕ_b , $\boldsymbol{\beta}_b$, $\boldsymbol{\gamma}_b$, ϕ_w , $\boldsymbol{\beta}_w$, and $\boldsymbol{\gamma}_w$). Let \mathbf{Z}_* be a matrix of covariates where the subscript indicates the model. Under this framework,

$$\widehat{\boldsymbol{\theta}}_{POLSE} = \widehat{\boldsymbol{\theta}}_{SPM} - (\mathbf{Z}'\mathbf{Z})_{SPM}^{-1} \mathbf{R}'\boldsymbol{\lambda}, \quad (14)$$

$$\boldsymbol{\lambda} = \left[\mathbf{R} (\mathbf{Z}'\mathbf{Z})_{SPM}^{-1} \mathbf{R}' \right]^{-1} \left(\mathbf{R}\widehat{\boldsymbol{\theta}}_{SPM} - \mathbf{0} \right), \quad (15)$$

$$\begin{aligned} \widehat{\sigma}_{POLSE}^2 (\mathbf{Z}'\mathbf{Z})_{POLSE}^{-1} &= \widehat{\sigma}_{SPM}^2 (\mathbf{Z}'\mathbf{Z})_{SPM}^{-1} \\ &\quad - \widehat{\sigma}_{SPM}^2 (\mathbf{Z}'\mathbf{Z})_{SPM}^{-1} \mathbf{R}' \left[\mathbf{R} (\mathbf{Z}'\mathbf{Z})_{SPM}^{-1} \mathbf{R}' \right]^{-1} \mathbf{R} (\mathbf{Z}'\mathbf{Z})_{SPM}^{-1}. \end{aligned} \quad (16)$$

This technique is called the restricted least squares estimator (RLSE) (see Greene, 2000, p.281 for a less brief treatment).⁸ The upshot is that it is mathematically possible to use the output of the SPM to recover the estimates from the POLSE using equations 14 and 15 and recover the *scaled* variance-covariance matrix of the POLSE using equation 16.⁹

The elasticity of the pooled estimates with respect to the constraints is $\boldsymbol{\lambda}$, and equation 14 shows that the difference between $\widehat{\boldsymbol{\theta}}_{POLSE}$ and $\widehat{\boldsymbol{\theta}}_{SPM}$ is driven primarily by the unscaled variance-covariance matrix of the SPM. Equation 16 is less intuitive, but the second term is positive semi-definite, indicating that the constraints cannot decrease the precision of the estimates.

We should always think of the POLSE as a restricted version of a SPM, and the RLSE depicts the POLSE as a post-processor of the SPM output.¹⁰ Thus, there are two related questions: *Is the output of the SPM sound and does the POLSE post-process this output*

⁸This particular technique holds only if $(\mathbf{Z}'\mathbf{Z})^{-1}$ is non-singular.

⁹One can also use the RLSE to map from the LSDVE to the POLSE. However, doing so is more cumbersome because the POLSE imposes $N - 1$ constraints on the LSDVE that $\alpha_i = \alpha \forall i$.

¹⁰In order to make the coefficients from the RLSE come out exactly the same as the coefficients from the pooled OLS model, the unit-specific means must be calculated from the same sample of observations that are used in the pooled OLS model. When variables are lagged, the first observation on each unit is thrown out of the sample before the pooled OLS model is estimated. Thus, the unit-specific means must be calculated excluding the first observations. This issue is much more acute when some variables have missing data scattered throughout the dataset. Also, the standard errors are subject to rounding error.

well? The answer to both questions is negative. In order to constrain two coefficients to be equal, the two corresponding variables must be in the model, and the post-processing of the POLSE requires that the SPM include counterproductive variables merely to make the constraints possible. A much more sensible approach would exclude the counterproductive variables from the SPM and refrain from post-processing the output.

The constraints that $\beta_b = \beta_w$ lack a theoretical foundation. When $\tilde{\mathbf{y}}_{i[t-1]}$ is included in the SPM, β_w is a vector of short-term effects. Since “short-term cross-sectional effects” are oxymoronic, the constraints that $\beta_b = \beta_w$ never have a theoretical justification when a lagged dependent variable is included in the SPM. Thus, although the POLSE is often thought to be more theoretically grounded than the W-E and other estimators, the POLSE is actually theoretically incoherent when a lagged dependent variable is included. There might be a theoretical reason to constrain the long-run temporal effects $\left(\frac{\beta_w + \gamma_w}{1 - \phi_w}\right)$ to equal the cross-sectional effects but that is not what the POLSE does.

The worst constraint the two-dimensional ARDL model imposes on the SPM is the constraint that $\phi_b = \phi_w$. Indeed, it is a good candidate for the worst constraint ever, since it is the only constraint I have encountered that invalidates the model when true and yet it is ubiquitous in the political economy literature. First note that $\bar{\mathbf{y}}_{i[t-1]}$ is a consequence of $\bar{\mathbf{X}}_i$ and $\bar{\mathbf{X}}_{i[t-1]}$ and conditional on $\bar{\mathbf{y}}_{i[t-1]}$, $\bar{\mathbf{X}}_i$ and $\bar{\mathbf{X}}_{i[t-1]}$ have virtually no net effect on \mathbf{y}_{it} .

Thus, the only variable “explaining” the cross-sectional variation in \mathbf{y}_{it} is $\bar{\mathbf{y}}_{i[t-1]}$ since the demeaned variables – which are orthogonal to $\bar{\mathbf{y}}_{i[t-1]}$ – only explain the temporal variation in \mathbf{y}_{it} . Whatever the demeaned variables fail to explain is really just noise around $\bar{\mathbf{y}}_i$. Since $\bar{\mathbf{y}}_i$ and $\bar{\mathbf{y}}_{i[t-1]}$ are virtually identical, $\hat{\phi}_b \approx 1.0$ almost without fail in real datasets.¹¹

Given that $\hat{\phi}_b \approx 1$, the constraint the ARDL model imposes on the SPM that $\phi_b = \phi_w$

¹¹Hypothetical data can be constructed where the effect of $\mathbf{y}_{i[t-1]}$ is almost zero. However, if $\bar{\epsilon}_i$ has even moderate variance, $\hat{\phi}_b \approx 1$. Either way, $\mathbf{y}_{i[t-1]}$ should be excluded from the model. In actual datasets, I would be shocked to find a case where it would be sensible to obtain a pooled estimate of the effect of $\mathbf{y}_{i[t-1]}$. And if we cannot even specify what the unmeasured variables are, there is no theoretical reason to believe that their between effects equal their within effects.

is valid only if $\phi_w = 1$. Either way, the POLSE is not a good estimator. If $\phi_b = \phi_w = 1$, the model is explosive, the long-run effects of the covariates are infinite, the error variance is undefined, and the test statistics are meaningless. For all these reasons, ARDL models require that $|\phi| < 1$, but to assume that $\phi_b = \phi_w$ is to contravene this requirement. If the constraint that $\phi_b = \phi_w = 1$ is invalid, $\hat{\phi} > \phi_w$, and the estimates for β and γ are attenuated because \mathbf{X}_{it} and \mathbf{X}_{it-1} are presumably correlated with \mathbf{y}_{it-1} .¹²

Furthermore, it does not make sense to think about the effects of $\bar{\mathbf{X}}_i$ conditional on $\bar{\mathbf{X}}_{i[t-1]}$ (and vice versa) because the two matrices are conceptually the same and are almost perfectly collinear. However, it does make sense to think about $\tilde{\mathbf{X}}_{it}$ conditional on $\tilde{\mathbf{X}}_{it-1}$ if we think that lagged effects are possible. Thus, imposing the constraints that $\gamma_w = \gamma_b$ and that $\beta_w = \beta_b$ when γ_b and β_b are non-sensible undermines the estimates for γ_w and β_w .

The sum of $\hat{\beta}_b$ and $\hat{\gamma}_b$ is substantively meaningful even when $\bar{\mathbf{X}}_i$ and $\bar{\mathbf{X}}_{i[t-1]}$ are almost collinear, but when $\bar{\mathbf{y}}_{i[t-1]}$ is included in the SPM, $\hat{\beta}_b$ and $\hat{\gamma}_b$ wash out. As was noted above, $\bar{\mathbf{y}}_{i[t-1]}$ is a consequence of $\bar{\mathbf{X}}_i$ and $\bar{\mathbf{X}}_{i[t-1]}$, and $\hat{\beta}_b$ and $\hat{\gamma}_b$ suffer from an extreme form of what has been called “post-treatment bias” or “included variable bias” (see King, 1991).

It is tempting to say that $\bar{\mathbf{y}}_{i[t-1]}$ and $\bar{\mathbf{X}}_{i[t-1]}$ not are included in the ARDL(1,1) model but are included in the SPM, so all these criticisms should pertain to the SPM rather than the ARDL(1,1) model. However, the RLSE shows that $\bar{\mathbf{y}}_{i[t-1]}$ and $\bar{\mathbf{X}}_{i[t-1]}$ really *are* in the ARDL(1,1) model, which is just a SPM with the constraints that $\phi_b = \phi_w$, $\beta_b = \beta_w$, and $\gamma_b = \gamma_w$. But if the SPM is estimated instead, we can impose the alternative constraints that $\phi_b = 0$ and that $\gamma_b = \mathbf{0}$, which excludes $\bar{\mathbf{y}}_{i[t-1]}$ and $\bar{\mathbf{X}}_{i[t-1]}$ from the SPM and circumvents all the problems just discussed.

Beck and Katz (1996) recommends the ARDL(1,1) model in order to test hypotheses

¹²Both Green, Kim and Yoon (2001, p.453) and Kristensen and Wawro (2003, note 18), among others, recognize in passing that the pooled estimate of a lagged dependent variable is biased upward and blame heterogeneity in the units. However, the same phenomenon could theoretically occur with homogenous units that experience balanced, but unobserved, shocks. Thus, I think it is useful to conceptualize this problem in terms of bad constraints that are placed on the SPM.

about γ . If $\gamma = \mathbf{0}$, then \mathbf{X}_{it-1} can be excluded, resulting in a ARDL(1,0) model, which is also known as the “partial adjustment model” or what Beck and Katz call the “lagged dependent variable model”. It seems that many in political science skip the step where the hypothesis that $\gamma = \mathbf{0}$ is tested and proceed directly to the ARDL(1,0) model. This practice is not good, and the constraints that the ARDL(1,0) model impose on the SPM that $\phi_b = \phi_w$ and that $\beta_b = \beta_w$ are no less problematic in a partial adjustment model. Also, it may be the case that $\gamma = -\phi\beta$, in which case the error term has an AR(1) structure and a constrained model can be estimated. However, assuming that the ARDL(1,1) model includes a single intercept, neither of these tests is fruitful because all the coefficients are biased to the extent that the pooling constraints are invalid.

We can instead use the SPM to investigate whether $\gamma_w = \mathbf{0}$ or whether $\gamma_w = -\phi_w\beta_w$ and possibly impose further constraints based on the result. Temporal variation determines whether $\gamma_w = \mathbf{0}$ or whether $\gamma_w = -\phi_w\beta_w$ (or neither), so the SPM approach is actually more consistent with econometric practice in the single-time series literature, which is what Beck and Katz (1996) draws upon to justify its recommendations.

There is another problem with the POLSE, which incorrectly assumes that the “unit effects” ($\bar{\epsilon}_i$) do not exist. The existence of the unit effects implies that each residual is highly correlated with every other residual for that unit, which undermines the consistency of “panel-corrected standard errors” (PCSEs). Beck and Katz (1996) recommends – but does not derive – a Lagrange Multiplier (LM) test to verify that there is no correlation in the residuals.¹³ This LM test takes the form of an auxiliary regression of the residuals on their lags and every independent variable in the original model:

$$\hat{\epsilon}_{it} = \phi\hat{\epsilon}_{it-1} + \psi\mathbf{y}_{it-1} + \mathbf{X}_{it}\beta + \mathbf{X}_{it-1}\gamma + \nu_{it}. \quad (17)$$

¹³LM tests that are appropriate for two-dimensional data are summarized in Baltagi (2001, chapter 5.2.7).

Beck and Katz (1996) recommends looking primarily at the magnitude of $\hat{\phi}$ in equation 17 to determine if there is any autocorrelation left in the residuals, but a pooled auxiliary regression has all the same problems that plague the original pooled model. In particular, $\hat{\phi} > \phi_w$ unless $\phi_b = \phi_w = 1$ (which would be very bad). When a lagged dependent variable is included in the original model, ϕ_w is often negative, but there is so little cross-sectional variation in the residuals that ϕ_b has very little weight in determining $\hat{\phi}$. We do not exactly know how PCSEs fare when the small cross-sectional component of the residuals is almost perfectly predictable, but the Monte Carlo evidence in Kristensen and Wawro (2003) is not particularly encouraging for PCSEs.

3.2 Problems with the random effects estimator

The REE is a POLSE but with variables that have undergone a generalized least squares transformation. As $\hat{\omega}$ approaches one, the problems with the POLSE diminish, but they do not disappear unless $\hat{\omega} = 1$ and all the between variation is removed. Two proposed advantages of the REE are that the REE will estimate the effects of time-invariant variables and that the REE is efficient. Questions of efficiency will be postponed until section 4 so that this subsection can focus on coefficient estimation.

The REE estimates the effects of time-invariant variables whereas a fixed effects model cannot. But the SPM can also estimate the effects of time-invariant variables while still producing within estimates for two-dimensional variables. It is certainly possible to include time-invariant variables in $\bar{\mathbf{X}}_i$ because the unit-specific mean of a time invariant variable is the time invariant variable itself. Provided that $\bar{\mathbf{y}}_{i[t-1]}$ and $\bar{\mathbf{X}}_{i[t-1]}$ are excluded from the SPM, the estimated effects of time-invariant variables are unbiased, assuming that $\bar{\epsilon}_i$ is uncorrelated with the meaned variables, which is just the “no omitted variable bias” assumption.

Thus, the widespread belief that a REE (or POLSE) is necessary to estimate the effects

of time invariant variables is false. The SPM and the B-E not only provide estimates of time invariant variables, they are the only estimators that have the potential to estimate the effects of time invariant variables without bias. If the constraints that the POLSE and the REE impose are invalid (and they are never literally true), and the time invariant variables are correlated with variables whose effects have been improperly constrained (and there is always some correlation), then the estimated effects of the time invariant variables will be biased. The worse the constraints and the greater the correlation, the worse the bias, but this form of bias cannot arise in a SPM or B-E.

Another problem with the REE is that it is very difficult to include a lagged dependent variable without causing bias.¹⁴ If a lagged dependent variable is included in the REE, there will be correlation between the random effects and the lagged dependent variable, which produces a biased estimate of ω and may produce a $\hat{\omega}$ that is negative. The SPM should avoid this problem by excluding $\bar{\mathbf{y}}_{i[t-1]}$. Of course, $\tilde{\mathbf{y}}_{it-1}$ is also correlated with the error term in a SPM, but the bias is small if T is large (see Beck and Katz, 2004).

The RLSE shows that estimates from the POLSE and the REE are (different) matrix-weighted averages of between and within estimates. Fundamentally, all models for two-dimensional data produce between estimates or produce within estimates or produce matrix-weighted averages of the two. I do not believe that a matrix-weighted average of two components is a more valuable contribution to the literature than the two components themselves. The weights are not particularly theoretical; more weight is given to the dimension that has more variance but that may not be the dimension that is less biased.

Moreover, we do not have a substantive interpretation for a matrix-weighted average of between estimates and within estimates. Between estimates have a limited but valid cross-sectional interpretation: The dependent variable for two otherwise identical units is expected to differ by a between estimate if the two units differ by one on a particular independent

¹⁴Greene (2000, p.576, note 21) has references for a REE with a lagged dependent variable.

variable. Within estimates are the “least bad” estimates for interpreting the coefficients causally: The within estimate of a variable is the expected *change* in a unit’s dependent variable when that independent variable increases by one, holding everything else constant. It is virtually impossible to elaborate on what a matrix-weighted average of between and within effects is since their respective interpretations are different. However, in section 4, I explain how it is possible, although not necessarily advisable, to obtain averaged estimates without incurring the costs of the REE.

In summary, this section has raised serious questions about coefficient estimates from the POLSE and the REE. Most of the problems turn on the presence of a lagged dependent variable. A REE has a hard time including a lagged dependent variable. A POLSE can include a lagged dependent variable but the constraint that $\phi_b = \phi_w$ is fatal for the POLSE even if the constraint is true. The SPM, however, can avoid these problems by excluding the cross-sectional component of the lagged variables. The SPM produces coefficients that have clear interpretations whereas the REE and the POLSE produces estimates that are matrix-weighted averages and have no substantive interpretation.

4 Uncertainty in two-dimensional models

So far, I have merely elaborated upon the points made in Zorn (2001), although that paper does not specifically discuss lagged variables. However, neither Zorn (2001) nor any other paper that has discussed the SPM has addressed problems with the standard errors. It turns out that the standard errors produced by the POLSE and the REE suffer from the same problems that plague the SPM and give the appearance that the POLSE and the REE are more efficient than other estimators. In this section, I show that the estimates from the POLSE and the REE are not necessarily more precise than competing models once these problems are corrected.

The problems identified in this section apply to all two-dimensional models, not just TSCS models. In a TSCS context, these problems are not fixed by PCSEs (see Beck and Katz, 1995) or clustered standard errors (see Kristensen and Wawro, 2003), but the problems those corrections are intended to fix presumably still exist. PCSEs and clustered standard errors attempt to overcome problems with the residuals. The problems that I will discuss pertain to the structure of the data – not the residuals or the least squares estimator. Thus, it is important to come up with an estimation strategy and data structure that are compatible with PCSEs and clustered standard errors.

I first discuss the necessary degrees of freedom correction for the W-E because it is well-known. In section 3, I claimed that the LSDVE and the W-E were equivalent, but they produce the same standard errors only if a degrees of freedom correction is made to the W-E. The degrees of freedom correction for the W-E entails subtracting N because N unit-specific means are estimated for each covariate when demeaning the independent variables. This correction is described in any intermediate textbook. Thus, the W-E estimates the same number of parameters as a LSDVE, although the LSDVE estimates N intercepts and K coefficients, while the W-E estimates N unit-specific means and K coefficients. The degrees of freedom for both the LSDVE and the W-E are $NT - N - K$.

Turning now to the correct number of degrees of freedom for a B-E, we only have N genuine observations on the meaned variables, regardless of how many rows there are in the dataset. These N observations are copied $N(T - 1)$ times to facilitate demeaning the raw variables, but this is merely a procedural step. There is no dispute in the literature that a B-E has only N available observations because there are only N means.

The SPM is the sum of the W-E and the B-E and thus must take the idiosyncracies of both into account. Once again, the dataset for a SPM looks like this:

Row	Unit	Time	y	Intercept	$\bar{y}_{[t-1]}$	\tilde{y}_{t-1}	$\bar{x}^{[1]}$	$\tilde{x}^{[1]}$...	$\tilde{x}^{[K]}$
1	1	1	$\bar{y}_1 + \tilde{y}_{11}$	1	$\bar{y}_{1[t-1]}$	\tilde{y}_{10}	$\bar{x}_1^{[1]}$	$\tilde{x}_{11}^{[1]}$...	$\tilde{x}_{11}^{[K]}$
2	1	2	$\bar{y}_1 + \tilde{y}_{12}$	1	$\bar{y}_{1[t-1]}$	\tilde{y}_{11}	$\bar{x}_1^{[1]}$	$\tilde{x}_{12}^{[1]}$...	$\tilde{x}_{12}^{[K]}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	...	⋮
T	1	T	$\bar{y}_1 + \tilde{y}_{1T}$	1	$\bar{y}_{1[t-1]}$	\tilde{y}_{1T-1}	$\bar{x}_1^{[1]}$	$\tilde{x}_{1T}^{[1]}$...	$\tilde{x}_{1T}^{[K]}$
$T + 1$	2	1	$\bar{y}_2 + \tilde{y}_{21}$	1	$\bar{y}_{2[t-1]}$	\tilde{y}_{20}	$\bar{x}_{21}^{[1]}$	$\tilde{x}_{21}^{[1]}$...	$\tilde{x}_{21}^{[K]}$
$T + 2$	2	2	$\bar{y}_2 + \tilde{y}_{22}$	1	$\bar{y}_{2[t-1]}$	\tilde{y}_{21}	$\bar{x}_{22}^{[1]}$	$\tilde{x}_{22}^{[1]}$...	$\tilde{x}_{22}^{[K]}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
NT	N	T	$\bar{y}_N + \tilde{y}_{NT}$	1	$\bar{y}_{N[t-1]}$	\tilde{y}_{NT-1}	$\bar{x}_N^{[1]}$	$\tilde{x}_{NT}^{[1]}$...	$\tilde{x}_{NT}^{[K]}$

Clearly, N unit-specific means are estimated before the variables are demeaned, so we must subtract N from the degrees of freedom as if the SPM were a W-E. Failure to do so will make the standard errors of temporal estimates too small, but the correction is very minor. However, the cross-sectional dimension is more complicated. There are only N genuine observations on the unit-specific means of the independent variables, but the observations on the unit-specific means appear T times each. Substantively, this copying of observations is no different than when the undergraduate copies observations and appends them to the bottom of the dataset to make estimates significant. Thus, the standard errors of the cross-sectional estimates will be too small.

We could assert that there are only N “cross-sectional” observations but then the degrees of freedom for the cross-sectional estimates in a SPM would be $N - N - K!$ Thus, we would also have to assert that subtracting N from the degrees of freedom is only necessary for the demeaned variables. Although some people may be uncomfortable with the implication that different variables in the same model have different numbers of degrees of freedom and different rules for calculating them, the issue will ultimately become moot.

4.1 The standard errors produced by pooled models are wrong

It is important to think through the complications with the standard errors in a SPM because a POLSE is just a constrained SPM. If the redundant variance makes the standard errors of the SPM too small, constraining the cross-sectional estimates to equal the temporal estimates will pass this problem with the standard errors along to the pooled estimates.

The RLSE (equation 16) provides a way to use the SPM output to recover standard errors for pooled estimates that are the same as the standard errors produced by the POLSE. It is tempting to try to correct the degrees of freedom for the SPM and then use the RLSE to get correct standard errors for the pooled estimates. However, equation 16 assumes that $\hat{\sigma}^2$ is uniform across all the cells of the SPM's variance-covariance matrix, and this assumption is invalid if one asserts that the cross-sectional degrees of freedom are different from the temporal degrees of freedom. Thus, any fix to the standard errors of the SPM or the POLSE must go through the variance-covariance matrix rather than $\hat{\sigma}^2$.

Thus, we would prefer that these redundant variance components not count T times. This effect can be achieved by taking a dataset of demeaned variables and appending a meaned dataset to the bottom – provided that the meaned dataset has only N rows in it. This procedure will create holes in the dataset, which should be filled in with zeroes. The joint dataset has $NT + N$ rows, and, has two data-generating processes, which are called “regimes” in Bartels (1996) and all the points made in that paper are now relevant. The first NT observations belong to the temporal regime and the subsequent N observations belong to the cross-sectional regime. The joint dataset should take this form:

Row	Unit	Time	“y”	$\tilde{\mathbf{y}}_{t-1}$	$\tilde{\mathbf{x}}^{[1]}$...	$\tilde{\mathbf{x}}^{[K]}$	Intercept	$\bar{\mathbf{y}}_{[t-1]}$	$\bar{\mathbf{x}}^{[1]}$...	$\bar{\mathbf{x}}^{[K]}$
1	1	1	\tilde{y}_{11}	\tilde{y}_{10}	$\tilde{x}_{11}^{[1]}$...	$\tilde{x}_{10}^{[K]}$	0	0	0	...	0
2	1	2	\tilde{y}_{12}	\tilde{y}_{11}	$\tilde{x}_{12}^{[1]}$...	$\tilde{x}_{11}^{[K]}$	0	0	0	...	0
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	...	\vdots	\vdots	\vdots	\vdots	...	\vdots
NT	N	T	\tilde{y}_{NT}	\tilde{y}_{NT-1}	$\tilde{x}_{NT}^{[1]}$...	$\tilde{x}_{NT-1}^{[K]}$	0	0	0	...	0
$NT+1$	1	–	\bar{y}_1	0	0	...	0	1	$\bar{y}_{1[t-1]}^{[1]}$	$\bar{x}_1^{[1]}$...	$\bar{x}_{1[t-1]}^{[K]}$
$NT+2$	2	–	\bar{y}_2	0	0	...	0	1	$\bar{y}_{2[t-1]}^{[1]}$	$\bar{x}_2^{[1]}$...	$\bar{x}_{2[t-1]}^{[K]}$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	...	\vdots	\vdots	\vdots	\vdots	...	\vdots
$NT+N$	N	–	\bar{y}_N	0	0	...	0	1	$\bar{y}_{N[t-1]}^{[1]}$	$\bar{x}_N^{[1]}$...	$\bar{x}_{N[t-1]}^{[K]}$

If we do not impose any constraints on the coefficients, a regression of “y” on the rest of the covariates is still a SPM (which I call SPM2). The estimated coefficients are the same as for the previously described SPM, but the standard errors, especially the standard errors for the cross-sectional estimates, differ for three reasons. First, the redundant variance component is removed from the SPM2. Second, the SPM2 constrains the variance of the errors to be the same across the two regimes. This constraint is not a good one, although the POLSE imposes the same constraint. Recall that the POLSE can be seen as a REE where the ratio of within variance in the residuals to the between variance is unity. Third, there are now $NT + N$ observations whereas there were only NT before. But N unit-specific means are estimated so the degrees of freedom need to be adjusted to $NT + N - N - K$.

If we constrain all the temporal effects to be equal to the corresponding cross-sectional effects, then we have a modified POLSE (which I call the POLSE2). The constrained estimates from the POLSE2 will be somewhat different than those of a POLSE because the downweighting of the cross-sectional variance is not uniform across variables. Since the cross-sectional variance is downweighted, the standard errors differ as well.

The POLSE2 is somewhat similar to a REE in that both downweight the cross-sectional variance before estimating a pooled model. However, the method for downweighting is

different. Whereas the REE subtracts a fraction of the unit means from the two-dimensional variables, where the fractional parameter, $\hat{\omega}$, equals $1 - \sqrt{\frac{\widehat{Var}(\tilde{\epsilon}_{it})}{\widehat{Var}(\tilde{\epsilon}_i)}}$. In contrast, the POLSE2 sets ω equal to 1 for the observations from the temporal regime and tacks the meaned variables onto the end of the dataset. As $T \rightarrow \infty$ holding N constant, both procedures amount to a W-E. For finite T , which procedure retains the most cross-sectional variance depends on the initial cross-sectional variance, the size of N relative to T , and the accuracy of the in-sample predictions for the W-E and B-E. I discuss the POLSE2 not because it is a good estimator but because it is better than a POLSE. Circumventing the redundant variance component is a good thing, but the pooling constraints are probably not justified in either case, especially if a lagged dependent variable is included in the model.

The SPM2 has the virtue that it is good for evaluating potential constraints with the BIC. The SPM will produce the wrong BIC because the goodness-of-fit measure is inflated by the redundant cross-sectional variance. The SPM2 avoids this problem, and one can compare the BIC from a SPM2 to the BIC from a POLSE2. One downside of the SPM2 (and the POLSE2) is that one cannot use PCSEs or clustered standard errors, or at least not without substantially rewriting the code and redoing the Monte Carlo experiments.

4.2 The consecutive parsed estimator

However, there is an estimator that is superior to the SPM2. The SPM2 is essentially two regressions: a W-E for the temporal regime and a B-E for the cross-sectional regime. I will now demonstrate that estimating these two regressions separately, which I call the *consecutive* parsed estimator (CPE), is better than any other estimator discussed here.

Since there is no cross-dimensional covariance in a SPM or SPM2, the demeaned (meaned) variables can be excluded without affecting the cross-sectional (temporal) estimates. Thus, the SPM and SPM2 produce the same estimates as the CPE does. The only question to resolve is which method produces better standard errors. The standard errors of the SPM2

are superior to those of the SPM, but inferior to those of the CPE.

First, when the W-E and B-E are estimated separately, the error variances are allowed to differ across the two regimes, which relaxes the problematic homoskedasticity assumption of the SPM2. Second, it is easy to incorporate PCSEs or clustered standard errors into the W-E and White standard errors into the B-E. Third, the SPM2 estimates more parameters than the W-E and the B-E do individually. Fourth, in a SPM2 the precision of the temporal (cross-sectional) estimates is adversely affected by the inability of the meaned (demeaned) variables to explain all the variation in the dependent variable along the cross-sectional (temporal) dimension. In other words, misspecifications increase the variance of the errors uniformly, regardless of which dimension is misspecified. In contrast, the error variance of the W-E (B-E) is affected by misspecifications on the temporal (cross-sectional) dimension only. The last two arguments imply that the CPE is more precise, while the first two imply that the standard errors of the CPE are less likely to be biased than those of the SPM2.

Moreover, if desired, we can also average the estimate for variable k from the B-E with the estimate for variable k from the W-E using the following formulas:

$$\bar{\beta}^{[k]} = \frac{1}{2} \times \left(\widehat{\beta}_w^{[k]} + \widehat{\beta}_b^{[k]} \right), \quad (18)$$

$$SE \left(\bar{\beta}^{[k]} \right) = \frac{1}{2} \times \sqrt{\left[SE \left(\widehat{\beta}_w^{[k]} \right) \right]^2 + \left[SE \left(\widehat{\beta}_b^{[k]} \right) \right]^2 - 2 \times Cov \left(\widehat{\beta}_w^{[k]}, \widehat{\beta}_b^{[k]} \right)}. \quad (19)$$

Since $Cov \left(\widehat{\beta}_w^{[k]}, \widehat{\beta}_b^{[k]} \right) = 0$, the last term under the radical in equation 19 drops out and it is possible to average information from non-nested models. The REE also produces estimates that are averages of between estimates and within estimates. However, the averaging procedure in the CPE is superior to that of the REE for several reasons. First, it is optional, and we can choose which sets of coefficients to average. Second, it is easy to explain (although no easier to interpret), since it is just a scalar average of estimates. Third, lagged dependent variables pose serious problems for the REE, but the CPE should circumvent this problem

by excluding $\bar{y}_{i[t-1]}$ from the B-E.

However, if \tilde{y}_{it-1} is included in the W-E, then $\hat{\beta}_w$ represents *short-term* effects, and there is no basis for averaging a short-term within effect with a between effect. To obtain a plausible average, one must first calculate long-run within estimates. The formula for the long-run effect of variable k is $\frac{\hat{\beta}_w + \hat{\gamma}_w}{1 - \phi_w}$, and its standard error can be approximated by simulation or by the delta method. The long-run effect and its standard error can be used in place of $\hat{\beta}_w^{[k]}$ in equations 18 and 19 when averaging with the between estimate. But it is worth stressing again that there is no substantive interpretation for this or any other average of between and within estimates, except in the limiting case that the two effects are equal.

Finally, estimating the W-E before estimating the B-E offers an opportunity to address spatial correlation. It might appear that a B-E does not suffer from contemporaneous correlation, but \bar{y}_i is a function of the contemporaneous correlation in ϵ_{it} before the data are meaned. As Beck (2001) and Franzese and Hays (2004) note, the best solution is to model the contemporaneous correlation if possible.

The contemporaneous correlation in the residuals of a W-E can be used to form a covariate for the B-E. In order to make corrections to the standard errors, PCSEs calculate Σ , which is a $N \times N$ variance-covariance matrix of the residuals where the off-diagonal elements are estimates of the covariance in the residuals for two units (see Beck and Katz, 1995). Thus, one can create a vector \mathbf{w}_i that is a weighted sum of the other unit-specific means of the dependent variable where the weights are given by the off-diagonal elements of Σ :

$$\mathbf{w}_i = \Sigma \bar{\mathbf{y}}_i. \tag{20}$$

There are several additional considerations about \mathbf{w}_i that merit Monte Carlo experiments. First, the diagonal elements of Σ should probably be replaced by zeroes to preclude \mathbf{w}_i from being influenced by a unit's *own* dependent variable. Second, it is possible that Σ should

be rescaled to a correlation matrix before using equation 20. Third, the W-E may include a lagged dependent variable and unique intercepts for each year, both of which will reduce the contemporaneous correlation in the residuals of the W-E. It is possible that a underspecified W-E should be estimated (after the properly specified W-E) to create a Σ that is not affected by any variables that are included in the W-E but not the B-E. However, the CPE is the best estimation strategy discussed in this paper even if \mathbf{w}_i is omitted from the B-E.

5 Empirical illustration

This section uses a subsample of the data from Green, Kim and Yoon (2001) to illustrate the issues involving the estimation of two-dimensional models. The results in Green, Kim and Yoon (2001) are easy to replicate, but I do not present a replication in this paper.¹⁵ The three differences between the approaches are that I use a subsample of their data, include lagged values of the exogenous variables, and include separate intercepts for each year in the sample. If one wanted to interpret the results substantively, the subsample of data I use is not really suitable but is sufficient to illustrate the methodological points.

The first column in table 2 faithfully implements the ARDL(1,1) model recommended in Beck and Katz (1996), and the BIC suggests that the restriction that $\gamma = \mathbf{0}$ is justified but not the restriction that $\gamma = -\phi\beta$.¹⁶ However, this POLSE is inferior to the POLSE2 in the next column, which also constrain the temporal effects to equal to the corresponding cross-sectional effects for all the covariates. Most of the coefficients and their standard errors do not change much. However, the one estimate that does change substantially is that for the lagged dependent variable, which changes our belief about the reequilibrium rate. Thus, the hypothesis that $\gamma = -\phi\beta$ seems even more unlikely based on the BIC.

¹⁵The data and Stata commands to replicate the results in Green, Kim and Yoon (2001) can be obtained from <http://polmeth.wustl.edu/misc/dirtypool.zip>.

¹⁶I calculate the BIC under the constraints that $\gamma = -\phi\beta$ by estimating the model via ML. A similar procedure is used to evaluate the constraints that $\gamma_w = -\phi_w\beta_w$ for the other models in table 2.

Table 2: Comparison of various least squares estimators for a two-dimensional model of bilateral trade

Regressor	<u>ARDL</u>		<u>Bad SPM2</u>		<u>Good SPM2</u>		<u>CPE (best SEs)</u>		
	POLSE	POLSE2	Temporal	Cross-sect.	Temporal	Cross-sect.	W-E	B-E	Avg
DISTANCE	-0.142 (0.012)	-0.347 (0.059)		-0.023 (0.081)		-0.707 (0.066)		-0.707 (0.087)	
GDP	1.135 (0.211)	1.109 (0.196)	1.149 (0.198)	0.572 (9.952)	1.149 (0.202)	1.439 (0.092)	1.149 (0.200)	1.439 (0.121)	1.082 (0.067)
GDP _{t-1}	-0.844 (0.211)	-0.810 (0.198)	-0.833 (0.200)	-0.550 (9.942)	-0.833 (0.204)		-0.833 (0.202)		
POPULATION	-4.683 (1.112)	-3.778 (1.461)	-3.644 (1.509)	0.421 (11.17)	-3.644 (1.541)	-0.852 (0.088)	-3.644 (1.527)	-0.852 (0.116)	-0.815 (0.109)
POPULATION _{t-1}	4.513 (1.116)	3.661 (1.463)	3.305 (1.504)	-0.437 (11.197)	3.305 (1.536)		3.305 (1.522)		
ALLIANCE	0.159 (0.275)	0.281 (0.258)	0.236 (0.263)	0.013 (0.165)	0.236 (0.268)	0.050 (0.162)	0.236 (0.266)	0.050 (0.213)	-0.011 (0.172)
ALLIANCE _{t-1}	-0.123 (0.275)	-0.220 (0.258)	-0.269 (0.262)	Dropped to identify	-0.269 (0.268)		-0.269 (0.266)		
DEMOCRACY	0.004 (0.007)	0.008 (0.007)	0.008 (0.007)	0.051 (0.67)	0.008 (0.007)	0.015 (0.015)	0.008 (0.007)	0.015 (0.020)	0.012 (0.011)
DEMOCRACY _{t-1}	-0.004 (0.007)	-0.004 (0.007)	-0.004 (0.007)	-0.051 (0.666)	-0.004 (0.007)		-0.004 (0.007)		
TRADE _{t-1}	0.782 (0.006)	0.578 (0.007)	0.564 (0.007)	0.980 (0.047)	0.564 (0.008)		0.564 (0.008)		
Change in the Bayesian Information Criterion (BIC) When the Constraints on the Left Are Imposed									
Within = Between	NA	NA	114.518		NA	NA	NA	NA	NA
$\gamma_w = 0$	-7.977	-14.945	-14.423		-15.393	-14.891	NA	NA	NA
$\gamma_w = -\phi_w/\beta_w$	42.247	106.167	199.611		141.878	-19.470	NA	NA	NA

Notes: ARDL = Auto-Regressive Distributed Lag model; SPM2 = Simultaneous Parsed Model with plausible standard errors; CPE = Consecutive Parsed Estimator; POLSE = Pooled OLS Estimator; POLSE2 = Modified Pooled OLS Estimator. The data are a subsample of the data used in Green, Kim and Yoon (2001) and are described in section 2. Intercepts, including year effects, are included but not reported. The dependent variable is the logarithm of bilateral trade in the POLSE, but is transformed in various ways for the other estimators. Standard errors are in parentheses. The average estimates from the consecutive parsed estimator are obtained by first calculating the long-run within estimate and then averaging it with the between estimate.

In the POLSE2, there are 11111 demeaned observations but only 271 meaned observations. Thus, the POLSE2 gives very little weight to the cross-sectional dimension, but this claim would be not be true if T were short. In this case, the estimates from the POLSE2 are essentially within estimates, as can be seen by comparing them to the temporal estimates from the overspecified SPM2 in the next column.

The standard errors for the temporal estimates in the SPM are a little small due to the assumption that the error variance is the same under both the temporal and the cross-sectional regime. However, the primary problem with this SPM2 is that the cross-sectional dimension is overspecified in exactly the ways that were predicted in section 3.1. The cross-sectional effect of the lagged dependent variable is estimated to be 0.980. Given that the lagged dependent variable accounts for essentially all of the cross-sectional variation in the dependent variable, all the other cross-sectional effects are insignificant, including the effect of distance, which is significant in all the other specifications. Worse, $\hat{\beta}_b = -\hat{\gamma}_b$ for GDP, POPULATION, and DEMOCRACY and each estimate has huge standard errors.

It cannot be emphasized enough that the overspecified SPM2 is what a POLSE2 looks like before the modeler arbitrarily would add the *false* constraint that $\phi_w = \phi_b$ to the mix and publish the results. When the overspecified SPM2 is constrained to be a POLSE2, the BIC increases by 114.5, implying that the odds in favor of the SPM2 are 7.367×10^{24} to one.

But the overspecified SPM2 is not a good model in any substantive sense, at least for the cross-sectional dimension. The next two columns represent a good SPM2, which excludes the lagged variables from the cross-sectional dimension. The temporal estimates do not change at all, because the 5 variables that are excluded have no covariance with the demeaned variables. Once the included variable bias is avoided, the estimated effects of DISTANCE, GDP, and POPULATION bounce back to levels that are consistent with the literature and are statistically significant.

The standard errors of the temporal estimates do increase slightly when the lagged vari-

ables are excluded from the cross-sectional dimension, but this problem can be avoided by estimating the W-E and the B-E separately, the results of which are shown in the next two columns. These standard errors are closer to “correct” than any other set of standard errors in table 2. They are more efficient because the cross-sectional variance in the errors is kept separate from the temporal variance. Also, the CPE relaxes the constraint of the SPM2 that the error variance is the same for both regimes. The standard errors could easily be “further corrected” by calculating White standard errors for the B-E and PCSEs or clustered standard errors for the W-E. However, one should include a second lag of the dependent variable to eliminate the autocorrelation from the residuals of the W-E before using PCSEs or clustered standard errors. It is also important to note that the coefficient estimates from the CPE are identical to those of the good SPM2, again due to the orthogonality property of demeaned and meaned data.

We can always average within estimates and between estimates to produce something roughly similar to a REE, although there is no substantive reason to do so. When the W-E includes a lagged dependent variable and / or lagged exogenous variables, we first have to calculate an estimate for the long-run effect before averaging it with the cross-sectional effect. In this case, the averaged estimates are fairly similar to their components.

To recap, both the POLSE and the POLSE2 impose the constraints that the within effects of all covariates equal their corresponding between effects. The POLSE2 produces better standard errors, but the pooling constraints are bad in either case. By relaxing these constraints and excluding the lagged variables from the cross-sectional dimension of the SPM2, we were able to obtain reasonable estimates for all coefficients. The CPE produces the same coefficient estimates as the SPM2, and produces better standard errors, which are amenable to further post hoc corrections.

Finally, compare the different estimates for the distance variable. In the POLSE, the effect of distance is small but becomes larger in magnitude in the POLSE2 because the cross-

sectional variance in the lagged dependent variable is downweighted. However, the effect of distance in the POLSE2 is an artifact of the pooling constraints, because the overspecified SPM2 demonstrates that DISTANCE has no effect when the pooling constraints are relaxed. But when the harmful variables are excluded from the cross-sectional dimension, the effect of DISTANCE jumps in magnitude and is statistically significant. The B-E produces a slightly more appropriate standard error for the distance variable, but that is a minor point. The major point is that the B-E, not the POLSE or the POLSE2, is best for estimating the effects of unit-constant or slowly changing variables both on bias and on efficiency grounds.¹⁷

6 Conclusions

In this paper, I have demonstrated several things. The criticisms of lag specifications only apply to TSCS or panel models where time is one of the dimensions. However, the general point made in Zorn (2001) that the between effects of a covariate can differ from the within effects applies to all two-dimensional models, as does the point that the model recommended in Zorn (2001) produces the wrong standard errors – especially for the between estimates.

The first major conclusion is that there are probably no circumstances in which it is optimal to use a POLSE. If a lagged dependent variable is included in the POLSE, the constraint that $\phi_b = \phi_w$ is almost certainly invalid but if it were valid, the POLSE is explosive. If a lagged dependent variable is not included in the pooled model, omitted variable bias will probably invalidate the constraints that $\beta_b = \beta_w$. The standard errors of the POLSE are wrong due to redundant cross-sectional variation in the variables, and this problem cannot be fixed by PCSEs or clustered standard errors.

The REE suffers from the same problems as the POLSE but to a lesser degree. However,

¹⁷One might note that if the effect of distance in the pooled ARDL model were divided by the complement of the coefficient of the lagged dependent variable, then the “long run” effect of distance would be about the same as in the CPE, but it is weird to think about a long-run effect of a time-invariant variable.

the REE adds additional concerns because the assumption that the random effect is uncorrelated with the independent variables requires that the lagged dependent variable be excluded from the REE. The REE is conceptually similar to the POLSE2 in that both downweight the cross-sectional variance and then estimate a pooled model.

The POLSE2 more or less fixes the problem of redundant variance, but it is not immediately obvious how to make corrections like PCSEs or clustered standard errors for a POLSE2. Moreover, the pooling constraints the POLSE2 imposes are the same – and no less invalid – than those of the POLSE. The SPM2 relaxes these pooling constraints and makes it possible to exclude the problematic variables from the cross-sectional dimension.

But the SPM2 still has some problems with the standard errors, which are based on the assumptions of no autocorrelation, no contemporaneous correlation, and homoskedastic error variance both across units and across the two regimes. It is easy to relax these assumptions when the W-E is estimated separately from the B-E. This technique, which I call the CPE, is the best estimation strategy that is presented in this paper. Under the standard assumptions, the CPE produces unbiased estimates, including estimates of time-invariant variables, and lends itself to the standard post hoc corrections to the standard errors.

However, “the standard assumptions” particularly the assumption that the errors are uncorrelated with future values of the exogenous variables when a lagged dependent variable is included in the W-E may very well not hold. To be sure, there are other estimators for two-dimensional data that could be superior. But regardless of the estimator, researchers should consider that the structure of the dataset and the constraints that the estimator places on the coefficients. Hopefully this paper has provided a sufficiently detailed illustration of how to recognize and avoid these problems that can be drawn upon when evaluating the properties of any estimator for two-dimensional data.

References

- Baltagi, Badi H. 2001. Econometric Analysis of Panel Data. Second ed. New York: John Wiley & Sons, LTD. 18
- Bartels, Larry M. 1996. "Pooling Disparate Observations." American Journal of Political Science 40(3):905–942. 14, 24
- Beck, Nathaniel. 2001. "Time-Series–Cross-Section Data: What Have We Learned in the Past Few Years?" Annual Review of Political Science 4:271–93. 28
- Beck, Nathaniel and Jonathon Katz. 2004. "Time Series Cross Section Issues: Dynamics, 2004." Paper presented at the 2004 Political Methodolgy Conference and available from <http://polmeth.wustl.edu/retrieve.php?id=36>. 20
- Beck, Nathaniel and Jonathon N. Katz. 1995. "What to Do (and Not to Do) with Time-Series–Cross-Section Data." American Political Science Review 89(3):634–647. 10, 22, 28
- Beck, Nathaniel and Jonathon N. Katz. 1996. "Nuisance vs. Substance: Specifying and Estimating Time-Series–Cross-Section Models." Political Analysis 8(3):1–36. 9, 10, 13, 17, 18, 29
- Beck, Nathaniel and Jonathon N. Katz. 2001. "Throwing the Baby Out with the Bathwater: A Comment on Green, Kim, and Yoon." International Organization 55(2):487–495. 10
- Franzese, Robert J. and Jude C. Hays. 2004. "Empirical Modeling Strategies for Spatial Interdependence: Omitted-Variable vs. Simultaneity Biases." Paper presented at the 2004 Political Methodolgy Conference and is avaialbe from http://sitemaker.umich.edu/jchays/files/franzesehays_1_.polmeth.2004.pdf. 28
- Gould, William. 2001. "What is the Between Estimator?" STATA FAQ: <http://www.stata.com/support/faqs/stat/xt.html>. 5, 9, 14
- Green, Donald P., Soo Yeon H. Kim and David Yoon. 2001. "Dirty Pool." International Organization 55(2):441–468. 2, 8, 13, 17, 29, 30
- Greene, William H. 2000. Econometric Analysis. Fourth ed. Upper Saddle River, NJ: Prentice Hall. 13, 14, 15, 20
- King, Gary. 1991. "'Truth' is Stranger than Prediction, More Questionable Than Causal Inference." American Journal of Political Science 35(4):1047–1053. 17
- King, Gary, James Honaker, Anne Joseph and Kenneth Scheve. 2001. "Analyzing Incomplete Political Science Data." American Political Science Review 95(1):49–69. 2

- Kristensen, Ida Pagter and Gregory Wawro. 2003. "Lagging the Dog? The Robustness of Panel Corrected Standard Errors in the Presence of Serial Correlation and Observation Specific Effects." Paper presented at the 2003 Political Methodology Conference. Preliminary version available from: <http://polmeth.wustl.edu/papers/03/krist03.pdf>. 17, 19, 22
- Little, Roderick J.A. and Donald B. Rubin. 2002. Statistical Analysis with Missing Data. Second ed. Hoboken, New Jersey: John Wiley & Sons, Inc. 2
- Neuhaus, J. M. and J. D. Kalbfleisch. 1998. "Between- and Within-Cluster Covariate Effects in the Analysis of Clustered Data." Biometrics 54:638–645. 5
- Oneal, John R. and Bruce Russett. 2001. "Clear and Clean: The Fixed Effects of the Liberal Peace." International Organization 55(2):469–485. 8
- Raftery, Adrian E. 1995. Bayesian Model Selection in Social Research. In Sociological Methodology, ed. Peter Marsden. Vol. 25 Oxford: Basil Blackwell. 14
- Ray, James L. 2003. "Explaining Interstate Conflict and War: What Should be Controlled for?" Conflict Management and Peace Science 20(2):1–31. 8
- Wilson, Sven E. and Daniel M. Butler. 2003. "Too Good to Be True? The Promise and Peril of Panel Data in Political Science." Working Paper. Preliminary version available from <http://fhss.byu.edu/POLSCI/Wilson/papers/>. 10
- Zorn, Christopher. 2001. "Estimating Between- and Within-Cluster Covariate Effects, with an Application to Models of International Disputes." International Interactions 27(4):433–45. 4, 5, 6, 7, 8, 9, 11, 21, 33