Time-Series-Cross-Section Analysis

Vera E. Troeger*

February 1, 2019

*Department of Economics, University of Warwick, Coventry, CV4 7AL, United Kingdom, e-mail: v.e.troeger@warwick.ac.uk

1 Introduction

The use of pooled time series cross section (PTSCS) data has become ubiquitous in observational analyses across the social sciences. Even the identification revolution that swept through empirical social science in the last two decades finds some merit in using data that combines observations across units and over time, because it allows identification through differences-in-differences approaches exploiting within unit variation. In "Mostly Harmless Econometrics" (Angrist and Pischke 2009) - which has obtained cult-like status in applied empirical economics - the authors recommend the use of diff-in-diff or unit fixed effects approaches to identification, requiring pooled data, if an experiment is infeasible, a meaningful discontinuity, or exogenous variation in the form of an instrument cannot be found.

Pooled data analysis has become the standard for analysing observational data in quantitative political analysis. This is particularly true in sub-disciplines like International Relations, Comparative Politics and Comparative Political Economy, but it has even extended to fields that use micro data, such as Political behavior or American Politics. Because more survey data over time is available, these fields are using more PTSCS data. Panel data pool cross-sectional information (number of units N) with information over time (number of time points T), e.g. data on individuals or firms at different points in time, information on countries, regions over time etc. Thus, panel data consist of repeated observations on a number of units. We can distinguish between cross-sectional dominant data (Cross-Section Time-Series CSTS), time-series dominant data (Time-Series Cross-Section TSCS) or pooled data with a fixed number of units and time-points. The data structure has implications for the model choice, since asymptotic properties of estimators for pooled data are either derived for $N \to \infty$ or $T \to \infty$. In addition, violations of full ideal conditions and specification issues have more or less severe effects for bias and efficiency, depending on whether the number of units exceeds the number of observations over time, or vice versa. Below, we discuss the strengths and weaknesses of this method and various ways to cope with some of the inherent problems.

Some have argued that TSCS and CSTS data consist of observations at different points in time for fixed units of theoretical interest, such as countries or dyads. In contrast, in panel data the units, mostly individuals in surveys, are of no specific interest and are randomly sampled from an underlying population with all inferences dedicated to uncovering the relationships in the population. Text books and articles, however, use these terms quite loosely. This entry will follow this trend and discuss general estimation procedures

and specification issues with respect to different kinds of data pooling cross-sectional and time series information.

For each specification issue this entry briefly discusses the solutions presented in commonly used textbooks like Wooldridge (2010) and then turns to more recent discussions in the political methodology literature.

2 Advantages and Disadvantages of PTSCS Data Analysis

Panel data pool observations for units (i) and time periods (t). The typical data generating process can be characterized as:

$$y_i t = \alpha + \sum_{k=1}^{K} \beta_k x_{kit} + \epsilon_{it}, \;\; i = 1, ..., N; \;\; t = 1, ...T \tag{1}$$

with k independent variables x which have observations for N units (i) and T periods (t). The dependent variable y is continuous (though in principle it can be a limited dependent variable, which requires non-linear estimation procedures) and also observed for i and t. ϵ_{it} describes the error term for observations i, t and we can assume a NT x NT Variance-Covariance Matrix Ω of the error term with the typical element $E(\epsilon_{it}, \epsilon_{js})$. In case all Gauss-Markov assumptions are met (the error term is iid) this model can be straightforwardly estimated by OLS. Since PTSCS data combine time-series and cross-section information this is rarely the case. However, the analysis of PTSCS data offers significant advantages over the analysis of pure time series or pure cross-sectional data. First, using pooled data increases the number of observations and therefore the degrees of freedom which means that more complex arguments can be tested by employing more complex estimation procedures. More importantly, most theories in the social sciences generate predictions over space and time, and it seems imperative therefore to test these hypotheses by using data providing repeated information for theoretically interesting units. PTSCS data analysis can be used to model dynamics, which is impossible when examining pure cross-sections and may lead to spurious regression results. Finally, it is possible to control for unit heterogeneity when analyzing pooled data, beyond the inclusion of additional right-hand side (RHS) variables. Accordingly, we use pooled data to eliminate some kinds of omitted variable bias, make the best of the available information, test theories that predict changes and test theories that predict parameter heterogeneity.

The most obvious disadvantage of panel data analysis is that an econometrically sound model specification is typically hard to find, since the data structure combines all of the

problems of cross-sectional and time-series data, but these problems typically arise simultaneously. Specification problems in pooled data analysis can be summarized as follows:

1. The residuals are typically serially correlated and not independent of each other
2. The residuals have different variances for different units (panel heteroskedasticity)
3. The residuals of different units are contemporaneously correlated
4. The residuals of unit i co-varies with residuals of unit j for different points in time, and
5. The expected mean of the error term deviates from zero for different units.

While each single violation of the underlying model assumptions is often straightforwardly accounted for by existing econometric measures, combinations of problems might not be solved simultaneously in a satisfying manner. Econometric solutions are often incompatible with theories. Sometimes it is hard to find models that are at the very same time econometrically sound (unbiased, efficient) and provide an appropriate test of the theory. Weighing advantages and disadvantages of pooled data analysis, the positive aspects certainly prevail, especially because the analysis of pooled data allows testing complex arguments over space and time, which is characteristic for the social sciences. From this perspective the steep increase in the popularity of panel data analysis does not seem surprising. However, the big challenge using pooled data is how to deal with simultaneously occurring violations of the underlying assumptions.

First, I will discuss common violations of underlying assumptions and solutions for each misspecification separately, and will later turn to attempts to account for simultaneously occurring specification problems.

3 Heteroskedasticity and Contemporaneous Error Correlation in PTSCS Data

Heteroskedasticity in pooled data presents a more complex problem than in pure cross-sections since a) the error term can have unit-specific variances (panel heteroskedasticity), b) the error term can be contemporaneously correlated, i.e. the error term of unit *i* is correlated to that of unit *j* in the same year, and c) the error term of one unit *i* can be correlated with the error term of unit *j* at different points in time. In addition, the error term can have time dependent error variances (autoregressive conditional heteroskedasticity).

Panel heteroskedasticity mainly occurs if the model specification fits different units with a different degree of accuracy. Correlations of the errors across units are determined by unobserved features of one unit that are linked to another unit. Both features violate Gauss-Markov assumptions: while they leave the simple estimators consistent, such estimators are now inefficient and standard errors may be incorrect. More importantly, both heteroskedasticity and error correlation often signal omitted variables bias, since in both cases something that should have been included into the structural part of the equation was left out.

This problem can be solved in a substantive way by identifying the causes of the omitted variable bias and including these variables into the right hand side of the models. Often, this approach is not feasible because the sources for heteroscedasticity are not known or excluded factors cannot be measured. In this case, several econometric solutions have been proposed. Parks (1967) and Kmenta (1986) were the first to propose a Feasible Generalized Least Squares (FGLS) estimation which is characterized by a NT x NT block diagonal matrix with an N x N matrix Σ that contains contemporaneous covariances along the block diagonal. Parks and Kmenta also suggest an Ω matrix with panel specific AR1 error structure and contemporaneously correlated errors, but in principle FGLS can handle all different correlation structures. Because the true structure of Σ and Ω are unknown, this procedure requires estimating a very large number of parameters in order to obtain the error covariances, which in turn leads to very inefficient, and therefore unreliable, results. Beck and Katz (1995) show that the Parks' method highly underestimates standard errors and therefore induces overconfidence in estimation results. As a result, this estimation procedure has fallen into disuse in recent work using pooled data.

Beck and Katz (1995) suggest a different way of dealing with panel heteroskedasticity. They argue that coefficient estimates of OLS are consistent but inefficient in pooled data and that the degree of inefficiency depends on the data and the exact error process. They suggest using OLS and correcting the estimated standard errors by taking the specific panel structure of the data into account:

$$Var[\beta] = (X'X)^{-1}X'\Omega X(X'X)^{-1} \tag{2}$$

with

$$\Omega = (E'E/T) \otimes I_t \tag{3}$$

This method is dubbed panel corrected standard errors. Other violations of Gauss-Markov assumptions such as serial correlation of the error term have to be treated be-

forehand. Since this approach only manipulates the standard errors of an OLS model, the coefficients are biased whenever OLS is biased.

4 Dynamics in Pooled Data Analysis

As pooled data combine information across units and over time, another problem arises if dynamics are present and the error term is serially correlated. The error term in t is dependent on the error term in t-1:

$$\epsilon_{it} = \rho_i \epsilon_{it-1} + \xi_{it} \tag{4}$$

From a formal econometric point of view, violating the independence assumption only influences the efficiency of the estimation. Yet, since the residual of a regression model picks up the influences of those variables that have not been included, persistence in excluded variables is the most frequent cause of serial correlation. Several remedies for serial correlation are available, all of which have different consequences for the model specification and interpretation of the estimation results.

A substantive solution to the problem of serial correlation is the inclusion of a lagged dependent variable $y_i t$ (LDV) to the right hand side of the regression equation:

$$y_{it} = \beta_0 y_{it-1} + \beta_k x_{kit} + u_i + \epsilon_{it} \tag{5}$$

In many cases this is enough to eliminate serially correlated error terms. However, there are also many perils to adding a LDV to the list of regressors. One of the main problems arises because the inclusion of an LDV makes it very hard to interpret the effects of the substantial RHS variables directly and correctly, since the conditional effect of x on y is dynamic and aggregated over all periods. It can be described by the following polynomial:

$$y(x)_{t_1 \to t_p} = \beta_1 x_{it} + \sum_{p=1}^{t_p} (\beta_0^{t-p} \beta_1 x_{it}) \tag{6}$$

the long-term effect of x_k reduces to:

$$\frac{\hat{\beta_k}}{(1 - \hat{\beta_0})} \tag{7}$$

Unfortunately, the standard errors of the function in equations 6 and 7 cannot be easily calculated. Since including a LDV resembles a shortened distributed lag model, we

implicitly assume that all variables exert an equally strong one period lagged impact on the dependent variable. Therefore, finding a non-significant coefficient of a theoretically interesting explanatory variable in a LDV model does not necessarily mean that this variable has no effect. It only tells us that this variable does not affect the dependent variable contemporaneously, but it might still have a lagged effect. From this it follows that the coefficient of the LDV estimates at best the average dynamic effect of all substantive RHS variables, rather than the actual dynamic effect of each explanatory variable. When including a LDV into the specification it is very important to calculate or simulate the short- and long-term effects of all RHS variables to correctly interpret effects, their size, and significance. Recent work by Williams and Whitten (2011) shows how to simulate long-term dynamics for autoregressive series in pooled data.

Another problem occurs when combining a LDV with the estimation of unit-specific effects by a fixed effects specification or a least squares dummy variable model (see next section for a more detailed description). This leads to biased estimates since the LDV co-varies with the time-invariant part of the error term. This problem is called Nickell-bias (Nickell 1981). The Nickell-bias is sizable in panel data with very short time periods (Pickup 2018) but becomes negligible as T increases. The best known suggestions tackling the problem of Nickell-bias are the instrumental variable approach by Anderson-Hsiao (1981) (AH), the differenced GMM (Generalized Methods of Moments) model by Arellano-Bond (1991) (AB) and the Kiviet (1995) correction, which proposes a corrected within estimator that subtracts a consistent estimate of the bias from the original fixed effects estimator. The first two approaches solve the bias problem by taking the first difference of both sides of the regression equation and instrumentation of the LDV with higher order lags of the LDV. Therefore, AH is only using the two periods lagged LDV as an instrument while AB can make use of all possible lags of the LDV and all exogenous variables in the model. Both approaches generate asymptotically consistent estimation results, but AB produces more efficient estimates due to the exploitation of all moment conditions. In finite samples, however, both estimators are problematic with regard to efficiency, as recent Monte Carlo experiments examining the finite sample properties reveal (Pickup 2018). Higher lags of the LDV provide good instruments only in the case where y is highly persistent over time. Unfortunately in such a case, the probability that the instruments also co-vary with the error term remains high. From this perspective both estimators cannot solve the problem of Nickell-bias if y is highly persistent or solve the problem very inefficiently, in case of low persistence. More recent research by Pickup (2018) shows evidence that transformed-likelihood estimators e.g. the orthogonal reparameterization (OPM) fixed-effects approach proposed by Lancaster (2000) and the quasi-

maximum likelihood (QML) fixed-effects approach by Hsiao et al. (2002) outperform the so called dynamic panel models discussed above by far, especially for very short T.

A Prais-Winsten (PW) transformation of the model offers another solution to serial correlation. The advantage of the Prais-Winsten approach lies in the transformation of both the left and the right hand side of the equation, which allows a direct interpretation of the regression coefficients. Prais-Winsten is estimated by GLS and is derived from the AR(1) model for the error term. First a standard linear regression is estimated:

$$y_{it} = x_{it}\beta + \epsilon_{it} \tag{8}$$

An estimate of the correlation in the residuals is then obtained by the following auxiliary regression:

$$\epsilon_{it} = \rho\epsilon_{it-1} + \xi_{it} \tag{9}$$

A Cochrane-Orcutt transformation is applied for observations $t = 2, ..., n$

$$y_{it} - \rho y_{it-1} = \beta(x_{it} - \rho x_{it-1}) + \zeta_{it} \tag{10}$$

And the transformation for t=1 is as follows:

$$\sqrt{1-\rho^2 y_1} = \beta(\sqrt{1-\rho^2 x_1}) + \sqrt{1-\rho^2 \zeta_1} \tag{11}$$

Equation 11 shows that another advantage of the Prais-Winsten transformation is the preservation of the first period. The differences between a PW and a LDV model might be substantial depending on ρ and the serial correlation in both y and x.

Dynamics in pooled data estimation presents one of the major challenges, because the sources of the dynamics often remain unknown to the researcher. Therefore, specifying dynamics in panel data is difficult. A researcher may believe that the data generating process (DGP) for their data contains a lag of the dependent variable, lags of the independent variables and/or serial correlation in the errors. If researchers could observed the DGP, they could select the corresponding model and in doing so account for the dynamics. Unfortunately, they cannot and if residuals show clear evidence of serial correlation this could indicate missing dynamics in either the mean or the error equation (or both). There is no straightforward statistical test that can determine which is one it is. This leaves the researcher uncertain how to proceed. Some scholars suggest always adding dynamics to the mean equation (Wilkins 2018) or at least ensuring the data model includes many lags of the dependent and independent variables as there are in the data generating process

(Wooldridge 2010, 194). However, the consequences of doing so when the DGP dynamics are in the error equation remain largely unknown. Pickup and Troeger (2019) explore theoretically and empirically whether the source of the dynamic process makes a difference for bias and efficiency of the estimation. They also show that the consequences for bias can be severe when dynamics are included in the mean equation when they should be in the error equation and vice versa.

Whether the dynamics in a particular set of panel data need to be accounted for in the error term or in the mean equation can be unclear. There are four common DGPs that might create uncertainty for the researcher as to whether they require a static model with a control for serial correlated errors or a dynamic model: 1) slow dissipation of effects due to covariates included or not included in the estimation equation; 2) lags of the covariates that are included in the estimation; 3) errors in the DGP that are serially correlated, often due to measurement artifacts; 4) autoregressive covariates omitted from the estimation.

The first two types of DGPs are dynamic in the mean equation. The first is dynamic in the dependent variable and the second is dynamic in the independent variable. The third and fourth types are static in the mean equation. The third is dynamic in the error equation. The fourth is static in the error equation but if the autoregressive covariate is excluded from the data model, the DGP for the residuals will be dynamic. Given that the DGP may be dynamic in the mean equation or in the error equation and given that the researcher typically does not observe the DGP, it is important to consider the consequences of: 1) using a model that is dynamic in the mean equation; or 2) static in the mean equation with a control for dynamics in the error equation: First, if a dynamic model is used (correctly or incorrectly), it must be dynamically complete. That is an autoregressive distributive lag (ADL(p,q)) model with p lags of the dependent and q lags of the independent variables needs to be estimated to avoid bias:

$$y_{i,t} = \alpha_1 y_{i,t-1} + ... + \alpha_p y_{i,t-p} + \beta_1 x_{i,t} + \beta_2 x_{i,t-1} + ... + \beta_q x_{i,t-q} + u_i + \epsilon_{i,t} \qquad (12)$$

Second, if the dynamic model is incorrectly used, it will either not lead to incorrect inference (in the case of AR errors and assuming dynamic completeness) or it will result in MA serially correlated errors. In the latter case, the estimation will likely generate some degree of bias. The magnitude of this bias will depend on T and the degree of serial correlation. Clearly, it is of utmost importance to specify the dynamic processes as closely to the true DGP as possible. Otherwise, biased estimates will occur. As a consequence testing for serial correlation and detecting misspecification in dynamic panel models is very important.

However in reality, applied researchers often perceive serially correlated errors as noise rather than information (DeBoef and Keele 2008). Yet, serially correlated errors clearly indicate a potentially severe model misspecification, which can result from various sources and occur either in the mean- or error equation. Perhaps most obviously, serially correlated errors are caused by: incompletely or incorrectly modelled persistency in the dependent variable; time-varying omitted variables or changes in the effect strengths of time-invariant variables, or misspecified lagged effects of explanatory variables. Conditionality makes modelling dynamics more complicated (Franzese 2003a,b, Franzese and Kam 2009). Few empirical analyses model all potential conditioning factors of the variables of interest. If, however, treatment effects are conditioned by unobserved time-varying factors then treatment effects vary over time, and the strength of these effects also changes over time as un-modelled conditioning factors change. Finally, serially correlated errors may result from misspecifications that at first sight have little to do with dynamics, for example from spatial dependence. Yet, spatial effects are certainly misunderstood if they are perceived as time-invariant: ignoring spatial dependence causes errors to be serially correlated (Franzese and Hays 2007).

In an ideal world these model misspecifications whould be avoided: dynamics should be directly modelled to obtain unbiased estimates. This proves difficult in reality. Since dynamic misspecifications are manifold and complex, econometric tests for dynamics at best reveal serially correlated errors, but they are usually unable to identify the underlying causes of autocorrelation. Often, these tests are also weak and do not reveal the true dynamic structure of the DGP, which may lead to overfitting of the data (Keele et al. 2016) Thus, empirical researchers often try to simplify their empirical model and to treat problems such as serially correlated errors with straightforward econometric textbook solutions, such as lagged dependent variables, period dummies, and simple homogeneous lag structures.

Plumper and Troeger (2018) show that econometric textbook solutions are not correct per se because they are usually not modelling the true dynamic process in the underlying DGP. In addition if dynamics occur in combination with other misspecifications, e.g. unit heterogeneity, treating one problem can render the effects (bias) of another violation worse.

One strategy that may reduce the size of the problem is to use less constrained econometric solutions. Distributed lag models, models with a unit-specific lagged dependent variable, panel co-integration models, models with heterogeneous lag structure (Plumper et al. 2005), more attention to periodization (Franzese 2003a), and better specified spatial models (Franzese and Hays 2007) may all reduce the size of the problem. However,

as the number of possible dynamic specifications increases, a higher order problem of model selection arises: since all of these different models generate different estimates and often demand different inferences, the next question is how empirical researchers select their preferred model. To eliminate or at least reduce the arbitrariness of model selection, DeBoef and Keele (2008: 187) suggest a testing down approach, starting with a full autoregressive distributive lag model and stepwise removing parameters according to predetermined criteria, often the significance of parameters. This procedure will result in a dynamic specification that maximizes the variance absorbed by the minimum number of parameters. As with all testing down approaches, this approach suffers from the arbitrariness in the choice of a starting model because we do not have an infinite number of degrees of freedom. Pickup (2018) similarly suggests a general-to-specific approach to modelling dynamics especially for panel data with small T to find a plausible dynamic specification before dealing with other misspecifications, such as unit heterogeneity.

The issue of specifying dynamic processes in pooled data becomes even more complicated if this problem is coupled with other potential misspecifications such as unit heterogeneity. The Nickell-bias discussed here is the least of the problems that occurs when addressing both issues separately, as is discussed below.

5 Heterogeneity

The identification revolution has not failed to impact the analysis of PCSTS data. Since Angrist and Pischke (2009) published "Mostly Harmless Econometrics" the so-called Fixed Effects Model has become the workhorse specification when using pooled data of any kind. Clearly, one of the advantages of analysing pooled data is the possibility of controlling for heterogeneity across units. When examining cross-sectional data, it is impossible to tell whether the estimated effects are contingent on unobserved effects that are specific to each unit, and therefore biased. PTSCS data analysis rests on the assumption that units are similar enough to be pooled together. If that is not the case we can still find appropriate specifications that allow accounting for differences across units, which might influence the estimation results. Textbooks usually discuss this problem under the header unit-heterogeneity and offer remedies such as fixed effects or random effects models. However, these models only deal with time invariant unit-specific effects: units can also be heterogeneous with respect to slope parameters, dynamics or lag structures. The next sections discuss different versions of unit heterogeneity, approaches to dealing with them and their advantages and disadvantages.

5.1 Unit Heterogeneity

When units have specific characteristics which cannot be measured and are time invariant, they offer different initial conditions which might bias the estimated coefficients. For example, geography is often considered time invariant - a country or region can be land-locked or on the European continent, cities have a certain distance to the next port etc. Other examples are inheritance, being a former colony, the sex of an individual or her genetic pool, are inherited specific to this unit and do not change over time. Especially if these time invariant unit-specific effects are correlated with any of the RHS variables, e.g. if the gender of a person determines specific behavior such as party identification or voting, coefficient estimates are distorted by omitted variable bias. If that is the case and we do not control for unit-specific effects the Gauss-Markov assumption of x being deterministic is violated:

$$y_{it} = \sum_{k=1}^{K} \beta_k x_{kit} + \sum_{m=1}^{M} \gamma_m z_{mi} + u_i + \epsilon_{it} \tag{13}$$

where u_i denotes the unit-specific effects and z other explanatory variables that are time invariant but can be measured and are of theoretical interest. If u_i is excluded from the estimation it becomes part of the overall error term, and will make the model less efficient in the case that it does not co-vary with any of the x or z but induces bias if u_i co-varies with any of the regressors. Econometrically, we can solve for correlated unit-specific effects by including a dummy variable for each unit into the right hand side of the model which generates unit-specific intercepts. This estimation procedure is called Least Squares Dummy Variable (LSDV) model.

$$y_{it} = \sum_{k=1}^{K} \beta_k x_{kit} + \gamma_{n-1} D_i + \epsilon_{it} \tag{14}$$

The unit-specific dummy variables (D_i) are multi-collinear to any time invariant variable z, the coefficients for z are therefore not identified. We also can employ fixed effects (FE) specification which is econometrically equivalent to a LSDV model. The fixed effects model first de-means all variables in the model by subtracting the unit-specific mean and then estimates the transformed equation by OLS.

$$y_{it} - \bar{y}_i = \sum_{k=1}^{K} \beta_k (x_{kit} - \bar{x}_{ki}) + \epsilon_{it} - \bar{\epsilon}_i + u_i - \bar{u}_i \tag{15}$$

$$\equiv \ddot{y}_{it} = \sum_{k=1}^{K} \beta_k \ddot{x}_{kit} + \ddot{\epsilon}_{it} \tag{16}$$

with

$$\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}, \;\; \bar{x}_i = \frac{1}{T}\sum_{t=1}^{T} x_{it}, \;\; \bar{\epsilon}_i = \frac{1}{T}\sum_{t=1}^{T} \epsilon_{it}$$

The fixed effects transformation eliminates the unit-specific effects, but also time invariant variables that might be of theoretical interest. FE can become highly inefficient because it does only use the within information of all variables. Yet, not controlling for unit-specific effects leads to biased estimates if unit effects exist and are correlated with any of the regressors.

If unit-specific effects do exist but do not co-vary with any of the RHS variables, not controlling for unit effects does not bias the estimates but increases the sampling variation of the OLS estimator, and therefore generates less efficient estimates. A straightforward remedy is a random effects (RE) specification which treats the u_i as a random unit-specific part of the error term. The random effects model only quasi-demeans the variables: rather than removing the time average from the explanatory and dependent variables at each t, RE removes a fraction of the time average. The RE estimator generates more efficient results than the FE estimator but the RE model produces biased estimates if RHS variables co-vary with the unobserved unit-specific effects. RE resembles a feasible GLS estimator where the Ω matrix (VC matrix of the error term) has a specific RE structure which only depends on two parameters σ_u^2 and σ_ϵ^2 . RE and FE estimates tend to grow similar if T gets large or the variance of the estimated unit effects increases, as compared to the error variance.

Since the RE estimator is more efficient than FE if the unit effects are uncorrelated with the regressors, it is useful to determine which of the two specifications should be used. Textbooks typically suggest employing the Hausman test. The Hausman test (Hausman 1978) is based on the following logic: since the RE estimator is biased if unit-specific effects are correlated, differences between FE and RE estimates are interpreted as evidence against the random effects assumption of zero covariance between x and u_i. Econometricians attest the Hausman test good asymptotic properties. Nevertheless, in finite samples the test results are influenced by the trade-off between bias and efficiency. The Haus-

man test is only powerful in the limit: since FE is consistent, the difference of RE and FE estimates can only be caused by biased RE estimates. In finite samples, however, the differences can result from two sources: biased RE estimates and unreliable FE point estimates due to inefficient estimation of variables with low within variation. The Hausman test actually mirrors this trade-off since it divides the difference between RE and FE estimates by the difference in the asymptotic variances of the RE and FE estimates. From this it follows that the test results are especially unreliable if the estimation equation contains regressors which are both correlated with the unit-specific effects and rarely changing over time. Recent research (Plumper and Troeger 2018, Pickup and Troeger 2019) shows that the Hausman test is highly unreliable, especially when the estimation also suffers from dynamic misspecifications or uses econometric fixes to model dynamics. The Hausman test is generally biased towards a fixed effects specification. Pickup and Troeger (2019) also show that the Mundlak formulation of the Hausman test is always preferable (especially when the estimation is dynamically more complete) because it does not rely on estimating the differences in variance between the fixed effects and random effects estimates.

The estimation of time-invariant or nearly time-invariant variables is highly problematic in a FE specification. It is easy to see that including completely time-invariant variables would be a problem, but it is less obvious that estimating rarely changing variables would be problematic, because FE specifications generate an estimate. However, this estimate might be very inefficient since FE eliminates all cross-sectional variation and only the variance over time is used to compute the coefficient. If this within unit variation is very small, the sampling variation of FE estimates increases dramatically, which leads not only to large standard errors but also to very unreliable point estimates. In empirical analyses across the social sciences we are often interested in the effect of variables that only change once in a while, for example the level of democracy in a country, electoral rules, central bank independence, marital status, family income etc.

In the case of time invariant variables, applied researchers often resort to a simple pooled OLS or a RE model which permit the estimation of coefficients for time-invariant variables. These estimates are biased if the unit-specific effects co-vary with the regressors. Hausman and Taylor (1981) as well as Amemiya and MaCurdy (1986) propose estimators that use the uncorrelated RHS variables as instruments for the correlated regressors. The models are based on a correlated random effects model (see Mundlak 1978) and use instrumental variables for the endogenous RHS variables. The underlying assumption is that only some of the time-varying (x_{it}) and time-invariant (z_i) variables are correlated with the unit-specific effects u_i. The uncorrelated x_{it} and z_i therefore can be

used as instruments for the correlated RHS variables. The within transformed x_{it} serve as instruments for the correlated x_{it} (these are estimated by FE) and the unit means of the uncorrelated x_{it} ($\bar{x}_i$) as well as the uncorrelated z_i serve as instruments for the correlated z_i. However, if the instruments are poor, Hausman-Taylor produces highly inefficient parameter estimates.

Of course a fixed effects estimator generates clean estimates for the within effect. However, theories often do not tell us whether we should observe an effect between or within units and whether these effects should be the same. Given that the fixed effects estimator generates highly unreliable estimates if RHS variables are slow moving, and specification tests are highly unreliable, more recently there has been a lot of focus on the conditions under which one should use a fixed or random effects formulation. Plumper and Troeger (2007, 2011) show that it depends on the ratio of within to between variation of the RHS variables whether a fixed or random effects formulation produces better (that is, estimates with lower Root Mean Squared Error) results. Clark and Linzer (2015) focus on the relationship between the number of units and within unit time points. Bell and Jones (2015), in comparison, demonstrate that it is important for the choice of estimator whether a panel is balanced or not.

5.2 Parameter Heterogeneity

We observe parameter heterogeneity if the coefficient of an explanatory variable differs significantly across units or over time. If parameters change across time or units we are likely to deal with unobserved, and therefore excluded, interaction effects or we have assumed the wrong functional form of the statistical relationship. If the source of parameter heterogeneity is known or our theoretical model even predicts differences in parameters across units or time periods, we can straightforwardly specify the correct model by including interaction terms between time periods or groups of units and the specific right hand side variables.

In case the source of parameter heterogeneity is unknown, seemingly unrelated regressions (SUR) or random coefficients models (RCM) offer an econometric solution to the problem. SUR models estimate a single regression for every unit but exploit the panel structure of the data by assuming a joint error process for all units. This increases efficiency of estimation by borrowing strength. SUR models only generate acceptable parameter estimates for long time-series, that is when T largely exceeds N. SUR models employ a GLS type estimator for the VC matrix, which weights the standard errors by the cross-section specific mean squared errors. The random coefficients estimator (e.g.

Beck and Katz 2007) provides a compromise between estimating the fully pooled model and a fully unpooled estimate (separate OLS for each unit). Pooled OLS depends on the stark assumption of unit homogeneity whereas separate OLS estimation for each unit produces inefficient results. The RCM borrows strength by shrinking each of the individual unit OLS estimates back to the overall (pooled) estimate. It is, therefore, also a good test for poolability of the data. The RCM generalizes the RE estimator from the intercept to all parameters of interest

$$y_{it} = \alpha + \beta_{ki} \sum_{k=1}^{K} x_{kit} + \delta_m \sum_{m=1}^{M} w_{mit} + \epsilon_{it} \tag{17}$$

$$\beta_i \sim N(\beta, \sigma_\beta^2)$$

The RCM can be made more useful by allowing the β_i to be functions of other unit-specific variables.

5.3 Heterogeneity of Dynamics and Lag Structures

In pooled data not only coefficient estimates but also dynamic effects can vary across units. In addition, different RHS variables might exert a differently lagged impact on the dependent variable and the lag length can differ across units. Different dynamics can be straightforwardly incorporated into a RCM or SUR models by including a LDV with unit-specific coefficients. Prais-Winsten specifications also allow for unit-specific autoregressive processes in the error term.

Since statistical tests for heterogeneous dynamics or lag structures are not readily available, specific dynamics should be defined on theoretical grounds. Unit-specific dynamics then can be more directly modelled by interacting the relevant regressors or the LDV with dummies for specific groups of units. Since different explanatory variables can have differently lagged effects across units as well, it is not plausible to just vary the estimates of the LDV, because the marginal effect of a RHS variable at time $t > 1$ partially depends on the estimate for the LDV.

In summary, different kinds of unit heterogeneity do not prevent pooling of information over time and across units. Theoretically, unit heterogeneity leads to interesting research questions which can be empirically analyzed with the appropriate model specification. The possibility of controlling for unit heterogeneity renders pooled data analysis more attractive than pure cross-section or time series analysis.

5.4 Dynamics and Unit Heterogeneity

Since PCSTS data by definition pool information over different dimensions, in this case space and time, the main challenge when using pooled data is dealing with simultaneously occurring violations of the underlying assumptions. Textbooks usually discuss the underlying assumptions of an estimator and, thus, solutions to violations of individual assumptions, separately. However, it is highly unlikely that only one assumption is violated, e.g. in pooled data we usually observe dynamics and unit heterogeneity at the same time. The main problem results from the expectation that solving one specification issue will reduce the overall bias. Unfortunately this is not the case, and treating one problem might make another problem worse, i.e. increase the overall bias.

Plumper and Troeger (2018) demonstrate this theoretically and with Monte Carlo experiments for the simultaneous occurrence of correlated unit-specific effects and dynamic misspecifications. They look at three common dynamic DGPs that all lead to serially correlated errors: omitted time-varying variables, omitted trends, and misspecified lag structures of the RHS variables. They then estimate six different dynamic specifications for each DGP (including a LDV, Prais-Winsten, Period fixed effects, ADL) and include unit fixed effects. They show that if the econometric solution does not match the DGP, a fixed effects approach increases the bias substantially, compared to not including fixed effects. This exercise shows that misspecification issues need to be addressed simultaneously and focussing on one issue (unit heterogeneity) but ignoring another (dynamics) can increase the overall bias.

Testing for dynamics turns out to be especially challenging in pooled data, because individual effects, autoregressive errors and moving average errors can all present as serial correlation, and the sequence of testing for different violations of Gauss-Markov severely affects the conclusions we draw. For example, as shown by Plumper and Troeger (2018) the Hausman test (1978) for correlated unit-specific effects performs very poorly under different dynamic DGPs and specifications.

As we know, a model with individual effects has composite errors that are serially correlated by definition. The presence of the time-invariant error component gives rise to serial correlation that does not die out over time. As a consequence, standard tests applied to pooled data always reject the null of spherical residuals. As discussed above dynamic processes in the mean equation that are not specifically modelled, and serial correlation in the error term, can both result in biased estimates. Researchers therefore need to test for these misspecifications. However, given the likelihood of individual effects in pooled data (whether correlated or not with RHS variables) testing for serial correlation in concurrence with individual effects is indispensable.

Pickup and Troeger (2019) analyze the performance of the most relevant specification tests for pooled data to determine under what conditions researchers can identify whether dynamic misspecifications and unit-specific effects (correlated or not) occur simultaneously. Since unit-specific effects generate serially correlated errors, the sequence of testing for fixed effects and serial correlation and/or the performance of joint tests becomes very important. They look at a large set of dynamic and static DGPs and estimation specifications for pooled data with (correlated) unit-specific effects and test the performance of the Wooldridge (2010) test for serial correlation, the Born-Breitung (2016) test for serial correlation under FE, the Inoue-Solon (2006) test for serial correlation under FE, the Baltagi-Li (1995) test for serial correlation under random effects and vice versa, and the Hausman (1978) and Mundlak (1978) tests for correlated unit-specific effects. None of these tests is designed to uncover the source of potential dynamics. The authors show that while most of the tests underperform for a large set of DGPs and estimation specifications, some guidance for applied researchers can be drawn from this exercise:

1. The Inoue-Solon test for serial correlation is seriously biased towards rejecting the Null of no serial correlation.

2. When the estimation includes a lagged dependent variable most tests for serial correlation in pooled data are oversized.

3. Tests for serial correlation in pooled data perform best when lags of the RHS variables but no LDV are included.

4. The Mundlak test always outperforms the Hausman test.

5. Both Hausman and Mundlak tests are biased toward rejecting the Null of uncorrelated unit effects when an LDV is included.

6. The Mundlak test performs relatively well when the estimation includes lags of the RHS variables and is more dynamically complete.

7. The Baltagi-Li test for random effects works best with more dynamically complete specifications.

In general, as has been discussed previously, a careful dynamically complete specification is of utmost importance before dealing with other issues such as unit heterogeneity.

6 PCSTS data with limited dependent variables

In principle all typical limited dependent variable models (Binary choice, Count, Tobit etc.) are also applicable to PCST Data. As for linear models, the pooled data structure adds problems for specification, that are sometimes hard to solve. I will briefly discuss specification issues in binary choice models with pooled data. The implications for other limited dependent variable models are similar.

In binary choice models the presence of unit-specific effects complicates matters significantly, whether they are correlated with explanatory variables or not. As in the linear case, it is possible to add $N-1$ unit dummies to the RHS of the estimation equation to control for correlated unit-specific effects, but only when estimating a logistic or logit model. This is called the unconditional fixed effects logit model. Since the normal CDF used in the probit binary choice model has no closed form representation, adding $N-1$ dummies will not allow for the model to be identified. The unconditional FE logit estimator is generally inconsistent but as Katz (2001) demonstrates, the bias is negligible for $T >= 20$.

Another alternative is the conditional logit fixed effects model (Chamberlain 1980). This estimator uses a conditional likelihood where only units that switched from zero to one or vice versa are used for estimation. This eliminates the unit-specific effect u_i. Chamberlain derives this procedure for $T = 2$. There has been some discussion (Greene 2004) whether this model can easily be applied to cases where $T > 2$. Chamberlain (1980) proposes a solution in the form of maximizing a conditional version of the likelihood function. The intuition is that the u_i disappear from the likelihood if the likelihood of a given period of peace (i.e. of a given country) is calculated conditioning on the total number of periods (years) of peace for that country. The unit fixed effects are eliminated from the conditional logit likelihood via a transformation that is analogous to first differencing in linear pooled data models. The units for which the outcome is always 0 or always 1 do not contribute to the likelihood. In other words, the information that they provide is not used to estimate β. These units are unaffected by the explanatory factors. However, if for example 99 percent of the sample is in this situation, we may still estimate a significant β using the 1 percent of the sample which changed outcome during the observation period. No weight would be given to the fact that for the vast majority of the sample the explanatory factors do not affect the outcome. This is similar to the linear case with slowly changing explanatory factors. Finally, this conditional likelihood approach cannot be adopted in the presence of lagged dependent variables.

Conditional fixed effects probit estimation is not feasible because the conditional likelihood approach does not yield computational simplifications for the FE probit. The u_i cannot be swept away, and maximizing the likelihood over all the parameters including the fixed effects will in general lead to inconsistent estimates for large N and fixed T.

In case the unit-specific effects remain uncorrelated with the RHS variables, a Random Effects Probit model can be estimated. MLE in this case yields a consistent and efficient estimator of β, but MLE is computationally costly since one has to compute the joint probabilities of a T-variate normal distribution, which involves T-dimensional integrals. This however becomes infeasible if T grows very large. However, it is possible to reduce the computation to a single integral.

There has been a lot of recent research focussing on fixed effects in binary choice and rare events data. Most methodologists dealing with pooled binary TSCS data agree that unobserved unit-specific effects generate a problem for estimation but there is no clear consensus on how to deal with this problem. As discussed above, it is commonly believed that one of the major problems with rare events data is the fact that estimating a conditional fixed effects model generates inefficient estimates. Cook et al. (2018) revisit this issue and demonstrate that the main concern with fixed effects models of rare events data is not inefficiency, but rather biased estimation of marginal effects. They argue that only evaluating switching units generates a biased estimate of the baseline risk and thus incorrect estimates of the effects. The authors propose a penalized maximum likelihood fixed effects (PML-FE) estimator, which retains the complete sample by providing finite estimates of the fixed effects for each unit.

As with linear models, serially correlated errors violate the independence assumption of most MLE, Logit and Probit models. Therefore, serially correlated errors encounter similar problems as in the linear case. Chamberlain (1993) shows that Binary Choice Models using a Lagged Dependent Variable are not identified. Beck, Katz and Tucker (1998) argue that Binary Choice PCSTS data are grouped duration data and suggest adding a series of period dummies to account for time dependency. These dummies account for the time elapsed since the last failure (1): $k_t - t_0$. This is equal to assuming duration dependence in a hazard model, measuring the length of non-eventful binary spells. Thus corrected, the logit resembles a BTSCS event history model. One can use a Likelihood Ratio test to check whether all included period dummies are jointly zero. This means that if T grows large there will be many dummy variables, which could pose a degrees of freedom problem. In this case period dummies can be transformed into cubic splines instead. The researcher has to specify a number of knots that define what segments of the time variable will have a cubic polynomial fit to it. These splines can be interpreted as

hazard rates: the estimated coefficients measure the effect of the calculated base vector on the probability of and outcome e.g. war. Carter and Signorino (2010) advocates the use of the exponential of the time variable instead (t, t^2, t^3). They argue that important elements of the splines are ad hoc and have no theoretical justification, which could lead to bias. Exponentials of t are more readily interpretable.

7 Conclusion

This entry gives a short overview of basic estimation procedures and specification issues in PTSCS data. Due to space constraints, other important topics such as spatial effects in pooled data, non-stationarity and the usefulness of error correction models, as well as a detailed discussion of estimation procedures and specification issues in limited dependent variable models for pooled data, have to remain untouched. While estimators and statistical tests are discussed in most econometric textbooks, a thorough discussion of specification problems in pooled and panel data remains important. This is also true for the question of how asymptotic properties of estimators translate to finite sample analysis. There have been many recent exciting developments in PTSCS data analysis in the social sciences. For example, much has been done to properly model spatial effects and spatio-temporal effects both in linear (Franzese and Hays 2007, Franzese et. al 2010) and limited dependent variable models (Fanzese et. al 2016) for pooled data. In addition, the identification revolution in empirical social sciences has influenced the analysis of pooled data. Blackwell and Glynn (2018), for example, use potential outcomes to define causal quantities of interest in dynamic PTSCS models, and clarify how standard models like the ADL can generate biased estimates of these quantities because of post-treatment conditioning. They propose inverse probability weighting and structrual nested mean models to deal with these post-treatment biases. The major challenge to the analysis of data that combines different dimensions, e.g. pooling cross-sectional and time-series information, remains the treatment of simultaneously occurring misspecifications, because existing tests cannot discriminate effectively between different sources of misspecifications.

References

Amemiya, Takeshi and Thomas E. MaCurdy 1986. Instrumental-variable estimation of an error-components model. *Econometrica* 54: 869-881.

Anderson, T.W. and Cheng Hsiao 1981. Estimation of Dynamic Models with Error Components. *Journal of the American Statistical Association* 76: 598-606.

Angrist, Joshua D. and Joern-Steffen Pischke 2009. Mostly Harmless Econometrics. Princeton: Princeton Universtiy Press.

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

Baltagi BH, Li Q 1995. Testing AR(1) Against MA(1) Disturbances in an Error Component Model. *Journal of Econometrics* 68: 133-151.

Beck, Nathaniel 2001. Time-Series-Cross-Section Data: What Have We Learned in the Past Few Years? *Annual Review of Political Science* 4: 271-293.

Beck, Nathaniel and Jonathan 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 2007. Random Coefficient Models for Time-Series-Cross-Section Data: Monte Carlo Experiments. *Political Analysis* 15: 182-195.

Beck, Nathaniel, Jonathan N. Katz and Richard Tucker 1998. Taking Time Seriously: Time-Series-Cross-Section Analysis with a Binary Dependent Variable. *American Journal of Political Science* 42 (4): 1260-1288.

Bell, Andrew and Kelvyn Jones 2015. Explaining Fixed Effects: Random Effects Modeling of Time-Series Cross-Sectional and Panel Data. *Political Science Research and Methods* 3 (1): 133-153.

Born, Benjamin and Joerg Breitung 2016. Testing for Serial Correlation in Fixed-Effects Panel Data Models. *Econometric Reviews* 35 (7): 1290-1316.

Chamberlain, G 1980. Analysis of Covariance with Qualitative Data. *The Review of Economic Studies* 47: 225-238.

Carter, David B. and Curtis S. Signorino 2010. Back to the Future: Modeling Time Dependence in Binary Data. *Political Analysis* 18 (3): 271-292.

Chamberlain, G. 1993. Feedback in Panel Data Models. Harvard Institute of Economic Research Working Papers 1656, Harvard - Institute of Economic Research.

Clark Tom S. and Drew A. Linzer 2015. Schould I Use Fixed or Random Effects? *Political Science Research and Methods* 3 (2): 399-408.

Cook, Scott J., Jude C. Hays and Robert J. Franzese 2018. Fixed effects in rare events data: a penalized maximum likelihood solution. *Political Science Research and Methods*, forthcoming.

DeBoef, Suzanna. and Luke J. Keele 2008. Taking Time Seriously: Dynamic Regression. *American Journal of Political Science* 52 (1): 184-200.

Franzese, Robert. J. 2003a. Multiple hands on the wheel: empirically modeling partial delegation and shared policy control in the open and institutionalized economy. *Political Analysis*: 445-474.

Franzese, Robert.J., 2003b. Quantitative Empirical Methods and the Context-Conditionality of Classic and Modern Comparative Politics. *CP: Newsletter of the Comparative Politics Organized Section of the American Political Science Association* 14 (1): 20-24.

Franzese, Robert J., Jude C. Hays and Scott J. Cook 2016. Spatial- and Spatiotemporal-Autoregressive Probit Models of Interdependent Binary Outcomes. *Political Science Research and Methods* 4 (1): 151-173.

Franzese, Robert J. and Cindy Kam, 2009. Modeling and interpreting interactive hypotheses in regression analysis. University of Michigan Press.

Franzese, Robert J., Jude C. Hays and Aya Kachi 2010. A Spatial Model Incorporating Dynamic, Endogenous Network Interdependence: A Political Science Application. *Statistical Methodology* 7 (3): 406-428.

Franzese, Robert J. and Jude C. Hays 2007. Spatial-Econometric Models of Cross-Sectional Interdependence in Political-Science Panel and Time-Series-Cross-Section Data. *Political Analysis* 15 (2): 140-164.

Greene, William 2004. The Behaviour of the Maximum Likelihood Estimator of Limited Dependent Variable Models in the Presence of Fixed Effects. *The Econometrics Journal* 7: 98-119.

Hausman, Jerry A. 1978. Specification Tests in Econometrics. *Econometrica* 46 (6): 1251-71.

Hausman, Jerry A., and William E. Taylor 1981. Panel data and unobservable individual effects. *Econometrica* 49: 137798.

Hsiao, C., M. H. Pesaran, and A. K. Tahmiscioglu 2002. Maximum likelihood estimation of fixed effects dynamic panel data models covering short time periods. *Journal of Econometrics* 109:107-150.

Inoue, Atsushi and Gary Solon 2006. A Portmanteau Test for Serially Correlated Errors in Fixed Effects Models.*Econometric Theory* 22 (5): 835-851.

Katz, Ethan 2001. Bias in Conditional and Unconditional Fixed Effects Logit Estimation. *Political Analysis* 9 (4): 379-384.

Keele, L., Linn, S. and C. M. Webb 2016. Treating time with all due seriousness. *Political Analysis* 24 (1): 31-41.

Kiviet, J.F. 1995. On Bias, Inconsistency, and Efficiency of Various Estimators in Dynamic Panel Data Models. *Journal of Econometrics* 68: 53-78.

Kmenta, J. 1986. Elements of Econometrics. Macmillan Publishing Company, New York.

Lancaster, T. 2000. The incidental parameter problem since 1948. *Journal of Econometrics* 95: 391-413.

Mundlak, Yair. 1978. On the pooling of time series and cross section data. *Econometrica* 46: 6985.

Nickell, Stephen 1981. Biases in Dynamic Models with Fixed Effects. *Econometrica* 49 (6): 1417-1426.

Parks, Richard W. 1967. Efficient Estimation of a System of Regression Equations when Disturbances are Both Serially and Contemporaneously Correlated. *Journal of the American Statistical Association* 62 (318): 500-509.

Pickup, Mark 2018. A General-to-Specific Approach to Dynamic Panel Models with a Very Small T. Presented to the 2017 Meeting of the Midwest Political Science Association, Chicago Illinois.

Pickup, Mark and Vera E. Troeger 2019. Specifying Dynamic Processes in Panel Data. Presented to the 2018 Meeting of the America Political Science Association, Boston, Massachusetts.

Plumper, Thomas and Vera E. Troeger 2007. Efficient Estimation of Time-Invariant and Rarely Changing Variables in Finite Sample Panel Analyses with Unit Fixed Effects. *Political Analysis* 15: 124-139.

Plumper, Thomas and Vera E. Troeger 2011. Fixed-effects vector decomposition: properties, reliability, and instruments. *Political Analysis* 19 (2): 147-164.

Plumper, Thomas and Vera E. Troeger 2018. Not so Harmless After All: The Fixed-Effects Model. *Political Analysis*, forthcoming.

Plumper, Thomas, Troeger, Vera E. and Philip Manow 2005. Panel Data Analysis in Comparative Politics. Linking Method to Theory. *European Journal of Political Research* 44: 327-354.

Wilkins, Arjun S. 2018. To Lag or Not to Lag?: Re-Evaluating the Use of Lagged Dependent Variables in Regression Analysis. *Political Science Research and Methods* 6(2): 393-411.

Williams, Laron K. and Guy D. Whitten 2011. Dynamic Simulation of Autoregressive Relationships. *The Stata Journal* 11(4): 577-588.

Wooldridge, Jeffrey M. 2010. Econometric Analysis of Cross Section and Panel Data. Cambridge, MA: MIT Press.