The Unbiased Estimation of Heterogeneous Coefficients in Panel Data Models with Common Factors and Feedback Effects

By Timothy Neal [*]

This article introduces the Augmented Panel Dynamic OLS estimator ('AP-DOLS'), which is capable of consistently estimating heterogeneous coefficients in panel time series datasets (large N and T) with endogenous regressors that suffer from both cross-sectional dependence (i.e. common factors/shocks) and other forms of endogeneity. Monte Carlo simulations show that in certain environments AP-DOLS is unbiased where competing panel time series estimators are not. The estimator is then applied to the topic of a long run relationship between inequality and crime using U.S. state data. The results illustrate the significant impact that proper estimation in heterogeneous panels can have on empirical conclusions.

The use of panel time series data (large N and T) has become increasingly popular in recent years due in part to the expanding size and availability of international macroeconomic databases such as the Penn World Tables, the World Bank's World Development Indicators, the IMF's International Financial Statistics, and the World Top Incomes Database. Its use in applied econometric papers followed the development of several estimators for panel time series including Mean Group OLS, Pooled Mean Group (Pesaran, Shin and Smith (1999)), Panel Dynamic OLS (Pedroni (2001)), and Panel Fully Modified OLS (Pedroni (2000)). These estimators are all robust to slope heterogeneity in the panel units, and the latter two are also consistent in the presence of certain forms of endogeneity in the regressors (such as feedback effects).

However, the theoretical literature for panel time series estimation has since shown that these estimators are biased in the presence of cross-sectional dependence in the data, also known as common factors or common shocks. Unobserved common factors in the panel can lead to correlation in the residuals across panel units, as well as correlation between the residuals and the regressors themselves. Recent work that has focused on correcting for these unobservable common factors in the estimation procedure often assume slope homogeneity, including Bai (2009) (who introduced a principal components method) and Moon and Widner (2012). However, assuming slope homogeneity leads to inconsistent estimation when the panel contains heterogeneous slopes (see Pesaran, Shin and Smith (1999)

[*] University of New South Wales, timothy.neal@unsw.edu.au

for a discussion on pooled panel estimation, as well as the simulation results in Kapetanios, Pesaran and Yamagata (2011).

A number of panel time series estimators have been proposed that are robust to both cross-sectional dependence as well as slope heterogeneity. This includes the Principal Components method (Song (2013) extended the work of Bai (2009) to allow for slope hheterogeneity, Common Correlated Effects Mean Group (Pesaran (2006)), Augmented Mean Group (Bond and Eberhardt (2013)), and the CoVariance estimator (Li and Lina (2014)). Common factors can represent a severe form of endogeneity in the regressors, and accordingly it has garnered a great deal of attention in the theoretical literature. It is not, however, the only form of endogeneity that can be found in the regressors of panel time series studies. Another form of regressor endogeneity is feedback effects, where the idiosyncratic errors of the dependent variable feed into future observations of the regressors (or vice versa). This can arise in practice for several reasons, including reverse causation or simultaneity.

This article introduces an estimator, known as Augmented Panel Dynamic OLS (‘AP-DOLS’), which directly follows the work of both Pesaran (2006) and Pedroni (2001). It is designed to be robust to the presence of both forms of endogeneity in a panel time series structure with slope heterogeneity. A Monte Carlo simulation study demonstrates that this estimator is unbias and efficient in small samples under a range of different data structures, including regressor exogeneity, cross-sectional dependence, feedback effects, and most importantly data that contains both feedback effects and cross-sectional dependence. Compared to competing estimators outlined above, it is the only estimator in the simulations that was unbias and efficient with both forms of endogeneity, and also shows significantly less bias when more complex endogeneity structures are generated in the simulated data.

This estimator, along with others discussed in this paper, are then used to estimate the long run relationship between inequality and crime using a panel time series dataset of U.S. states. The results provide evidence of a positive relationship between inequality and violent crime, but no evidence of a relationship with property crime. It also illustrates the high degree of sensitivity that results can have depending on the type of estimator used. Accordingly, it is important to carefully consider the most appropriate estimator for the data when conducting an applied panel time series study.

The rest of the article is organised in the following way. Section 1 reviews a number of panel time series estimators that are robust to common shocks. Section 2 introduces the Augmented Panel Dynamic OLS (‘AP-DOLS’) estimator. Section 3 conducts Monte Carlo simulations with a range of estimators. Section

4 applies the estimators to the issue of a potential long run relationship between inequality and crime in US states. Section 5 provides a concluding discussion. Finally, the appendix reports and discusses supplementary simulation results.

I. Panel Time Series Estimation with Common Factors

Consider the following empirical set up:

$$\begin{aligned} y_{it} &= \beta_i x_{it} + u_{it} \\ u_{it} &= \alpha_i + \lambda_i f_t + \epsilon_{it} \\ x_{it} &= \pi_i + \gamma_i f_t + v_{it} \\ i = 1, 2, ..., N&; t = 1, 2, ..., T \end{aligned}$$

The dependent variable y_{it} and the kx1 vector of explanatory variables x_{it} are both partly determined by a vector of rx1 common factors f_t. The strength of this relation is driven by the rx1 factor loading vectors that varies by panel unit: λ_i and γ_i. It is assumed that the panel follows a panel time series structure, where both N and T are medium-to-large in number. In this structure we see a form of endogeneity in the regressors known as cross-sectional dependence, since the error term u_{it} is related to the explanatory variables x_{it} through the common factors f_t (assuming there exists $\lambda_i \neq 0$ and $\gamma_j \neq 0$ such that $i = j$). This has serious implications for the consistent estimation of β_i if f_t is unobserved as the regressors are endogenous.

The estimation of β in the presence of this common factor structure (i.e. data with cross-sectional dependence) has been a topic of great importance in the recent literature on panel time series. When the panel is hetereogeneous, the literature has thus far focused on mean group estimators. Mean group estimation is where regressions are run for each panel unit individually (thereby requiring medium-to-large T), and then in some way averaged into a final panel-wide estimate (thereby requiring medium-to-large N). The simplest formulation of mean group estimation is mean group OLS:

$$y_{it} = \beta_i x_{it} + e_{it}$$

Two mean group estimators that are robust to heterogeneous panels with cross-sectional dependence are Common Correlated Effects Mean Group ('CCE-MG') introduced in Pesaran (2006) and Augmented Mean Group ('AMG') introduced in Bond and Eberhardt (2013). The CCE estimation method, first introduced in Pesaran (2006), seeks to approximate the projection space of the unobserved factors f_t through the cross-sectional averages (i.e. the average across panel units over one period of time) of the dependent and explanatory variables. As such, it expands on mean group OLS as follows:

$$y_{it} = \beta_i x_{it} + \eta_i \bar{x}_t + \tau_i \bar{y}_t + e_{it}$$

where $\bar{x}_t = \sum_1^N x_t/N$ is the cross-sectional average of the regressor(s), and $\bar{y}_t = \sum_1^N y_t/N$ is the cross-sectional average of the dependent variable. This approach is computationally simple, and the asymptotic consistency and efficiency of this estimator under a common factor structure is shown in Pesaran (2006). However, as explained by Li and Lina (2014), rank conditions are needed for this approximation to be adequate, and bias may be introduced when the factor structure becomes sufficiently complex.

Augmented Mean Group, introduced in Bond and Eberhardt (2013), uses a two-step regression that includes a common dynamic effect to the individual panel unit regressions in the second stage. The dynamic effect is estimated through time dummies included in the first stage first-difference pooled regression. The chief difference between AMG and CCE is that the former provides an explicit estimate for f_t, which may be useful for some applications including cross-country production functions (which was the original intended use for the estimator). The set up is as follows:

Stage 1:

$$\Delta y_{it} = \beta \Delta x_{it} + \sum_{t=2}^{T} c_t \Delta D_t + e_{it}$$

In this pooled first-difference regression, D_t represents time dummies (starting from the second period as they are differenced). The coefficients to the time dummies, c_t, are turned into a variable shared across panel units $\hat{\mu}_t$, as a coefficient estimate will exist for each time period in the panel.

Stage 2:

$$y_{it} = a_i + \beta_i x_{it} + c_i t + d_i \hat{\mu}_t + e_{it}$$

The time dummy coefficient variable included in stage 2 approximates the unobserved common factors that are potentially driving the variables in each panel unit (note that it only accounts for the unobserved common factors as the regressors were included in the first stage regression as well).

II. The Augmented Panel Dynamic OLS Estimator

For the simple common factor structure outlined in the previous section, both CCE-MG and AMG have been shown to be unbiased in small samples. Now consider the addition of feedback effects into the data structure. The regressors x_{it} take the following form:

$$x_{it} = \pi_i + \gamma_i f_t + v_{it} + \tau \epsilon_{i,t-1}$$

In this modification to the earlier data structure, an idiosyncratic shock to the dependent variable impacts on the regressor(s) in the following period. There may be a number of reasons why this appears in the data, including reverse causation or simultaneity bias. Some of the earlier panel time series estimators were robust to this form of endogeneity. For instance, consider a mean group version of Panel Dynamic OLS, introduced in Pedroni (2001) and based on the time series estimator for cointegrated systems seen in Stock and Watson (1993):

$$y_{it} = \beta_i x_{it} + \sum_{s=-l}^{l} \Delta x_{is} + e_{it}$$

A regression is run for each individual panel unit with the regressor(s) x_{it} and the lags and leads of the first difference of each regressor. As with all other mean group panel estimators, the results for each unit are in some way averaged to give an overall estimate of β. The most common method is a simple average $\beta = N^{-1} \sum_{i=1}^{N} \beta_i$, but other options exist that will be explored later in the paper. The number of lags and leads used in each regression, l, is flexible. Researchers often use *a priori* rules for lag/lead selection in DOLS or PDOLS (e.g. two lags and leads). Kejriwal and Perron (2008) showed, however, that data dependent rules could be used to minimise the mean-squared error in DOLS regressions (see the appendix for further discussion about lag/lead selection). The t-statistics are calculated by the following:

$$t_{\beta_i} = (\beta_i - \beta_0) \left(\hat{\sigma}_i^{-2} \sum_{t=1}^{T} (x_{it} - \bar{x_i})^2 \right)^{1/2}$$

where β_0 is the null hypothesis, $\bar{x_i}$ is the unit average of the regressor, and σ_i^2 is the long run variance of the residuals (computed using HAC methods) defined by $\sigma_i^2 = \lim_{T \to \infty} E[T^{-1}(\sum_{t=1}^{T} \hat{e}_{it})^2]$. The group mean t-statistic can be averaged using a number of methods, with this study adopting a simple average as above $t_\beta = N^{-1} \sum_{i=1}^{N} t_{\beta_i}$.

This approach to panel time series estimation is robust to feedback effects of many kinds, as shown in Stock and Watson (1993). However, the estimator relies on cross-sectional independence for consistency, which prevents its use in the above data structure or many real panel time series datasets. Bias introduced from common factors can be quite severe (see the simulation results below) depending on the number of factors and the magnitude of the factor loadings. A potential solution proposed by Pedroni is to time-demean the data before estimation. This is, in essence, the same as adding time dummies to the regression. However, the simulation results presented below show how in certain common factor structures, time-demeaning the data can help with bias yet will introduce extreme inefficiency to the estimator.

The solution proposed in this article for the unbias estimation of panel time series where the regressors possess both common factors and feedback effects is to adopt a new estimation approach that this study calls Augmented Panel Dynamic OLS ('AP-DOLS'). AP-DOLS combines the augmentations of the CCE and PDOLS estimators to account for their respective form of endogeneity into one regression (thus maintaining computational simplicity):

$$y_{it} = \beta_i x_{it} + \sum_{s=-l}^{l} \Delta x_{is} + \tau_i \bar{x}_t + \gamma_i \bar{y}_t + e_{it}$$

Being a mean group estimator, this regression is run for each individual panel unit and the results are averaged (in one of several ways) to form the mean-group β estimate. See Pesaran (2006) and Stock and Watson (1993) for discussions on the asymptotic consistency and efficiency of the separate augmentations used in the regression. The next section will determine how this estimator performs relative to a number of others in small samples using Monte Carlo simulations. The results will determine whether it might be practical for applied use.

III. Monte Carlo Simulation Results

The Monte Carlo simulations undertaken in this article broadly follow the approaches of Bond and Eberhardt (2013), Coakley, Fuertes and Smith (2006), and Kapetanios, Pesaran and Yamagata (2011). Six scenarios were constructed that give different features to the data, and test the performance of the panel estimators under situations that they may experience in real datasets. In all of the scenarios, the generated dataset will be panel time series (with N and T between 30 and 100), and the variables nonstationary and cointegrated. The performance of each estimator will be measured using the mean and standard error of the estimated beta coefficients, as well as the root mean squared error:

$$\sqrt{1/S \sum_{s=1}^{S} (\hat{\beta}_s - \beta)^2}$$

Where s is the simulation number and S is the total number of simulation repetitions which was set to 1,000 for this study.

The regressor x takes varying forms across the five scenarios. Strict exogeneity is assumed in the first scenario, for benchmarking purposes. Scenarios 2, 4, 5, and 6 introduce cross-sectional dependence in the regressor term (i.e. it is correlated with the error term through common unobserved factors/shocks). Scenarios 3, 4, 5, and 6 also adds feedback effects into the regressor term (i.e. idiosyncratic shocks to the dependent variable feed back into future values of the regressor), with scenarios 5 and 6 containing severe forms of feedback effects. The subsection below details the data generation process used for each simulation scenario.

A. Data Generation Process

The basic framework for the data generation process closely follows the approach used in Bond and Eberhardt (2013). The dependent variable is defined by:

$$y_{it} = \beta_i x_{it} + u_{it}$$
$$i = 1, 2, ..., N; t = 1, 2, ..., T$$

In all scenarios, we generate $\beta_i = 1 + e_i^\beta$ with $e_i^\beta \sim U[-0.25, 0.25]$. Accordingly, an unbias estimate of β (β being the group mean of β_i) will be equal to 1. The regressor x and error term u is able to take two forms, depending on whether the scenario includes common factors or not.

Without common factors:

$$x_{it} = x_{i,t-1} + e_{it}$$

$$u_{it} \sim N(0, 1)$$

where $e_{it} \sim N(1, \sigma_x)$ and $\sigma_x \sim U[0.5, 1.5]$.

With common factors:

$$x_{it} = a_i + \lambda_{x_1,i} f_{1t} + \lambda_{x_3,i} f_{3t} + \epsilon_{it}$$

$$\epsilon_{it} = \rho \epsilon_{i,t-1} + e_{it}$$

Once the expression for ϵ_{it} is integrated into the main equation, the regressor becomes:

$$x_{it} = (1-\rho)a_i + \rho x_{i,t-1} + \lambda_{x_1,i} f_{1t} - \rho \lambda_{x_1,i} f_{1,t-1} + \lambda_{x_3,i} f_{3t} - \rho \lambda_{x_3,i} f_{3,t-1} + e_{it}$$

$$u_{it} = \alpha_i + \lambda_{y_1,i} f_{1t} + \lambda_{y_2,i} f_{2t} + \eta_{it}$$

where $e_{it} \sim N(0, \sigma_e)$, $\sigma_e \sim U[0.003, 0.001]$, $\eta_{it} \sim N(0, 0.00125)$, and the common AR-coefficient is $\rho = 0.25$. The series begins with $x_{i,-49} = a_i$ and continues along the relevant process with $t = -48, ..., 0, 1, ..., T$. Before the monte carlo simulation begins the first 50 observations are discarded, to ensure that the results are not sensitive to the value chosen for a_i. Factor loadings $\lambda_{x_1,i}$ and $\lambda_{y_1,i}$ are *i.i.d.* $U[0,1]$, and $\lambda_{x_3,i}$ and $\lambda_{y_2,i}$ are *i.i.d.* $U[0.25,1.25]$.

The common factor series are generated under the following process:

$$f_{jt} = \mu_j + f_{j,t-1} + v_{fjt}$$
$$t = -48, ..., 0, 1, ..., T$$
$$f_{j,-49} = 0, v_{fjt} \sim N(0, \sigma^2_{fj})$$
$$\sigma^2_{fj} = 0.00125, \mu_j = (0.015, 0.012, 0.01)$$
$$j = 1, 2, 3$$

This article considers the following scenarios. All scenarios from 3 onwards include additional effects on the observed regressor, here called $\hat{x}$, to add endogeneity that goes beyond the presence of common factors (i.e. feedback effects).

- Scenario 1: As above without common factors.
- Scenario 2: As above with common factors.
- Scenario 3: Without common factors. Feedback effect: $\hat{x_{it}} = x_{it} + 0.25\epsilon_{i,t-1}$.
- Scenario 4: With common factors. Feedback effect: $\hat{x_{it}} = x_{it} + 0.25\epsilon_{i,t-1}$.
- Scenario 5: With common factors. Feedback effect: $\hat{x_{it}} = x_{it} + 0.25\epsilon_{i,t-1} + 0.25\epsilon_{i,t-2} + 0.25\epsilon_{i,t-3}$.
- Scenario 6: With common factors. Feedback effect: $\hat{x_{it}} = x_{it} + 0.25\epsilon_{i,t-1}$ if $\epsilon_{i,t-1} <= 0.002$, or $\hat{x_{it}} = x_{it} + 0.75\epsilon_{i,t-1}$ if $\epsilon_{i,t-1} > 0.002$

Scenario 1 serves as a benchmark for later scenarios, since unbiasness and efficiency is anticipated for all estimators under consideration. With the addition of common factors, scenario 2 serves as the first major test of a panel time series estimator. Scenario 3 and 4 introduce an additional form of endogeneity: feedback effects with one lag (with scenario 3 not containing common factors, and scenario 4 containing common factors). Scenario 5 extends the feedback effect to three lags, and finally scenario 6 adds a non-linear feedback effect.

B. Results

The purpose of this simulation study is to test estimators from the panel time series literature (including the one introduced in this article) under a range of scenarios involving endogeneity in the regressors. While cross-sectional dependence (i.e. common factors) has recently received a great deal of attention in the literature, this simulation study also focuses on feedback effects. The estimators tested in this study are: Pooled OLS (with fixed effects for time and panel unit), infeasible Mean Group (with the common factors added to the regression for benchmark purposes), Panel Dynamic OLS (with and without time dummies), Augmented Mean Group, Common Correlated Effects (mean group formulation), and finally Augmented Panel Dynamic OLS.

Table 1—Simulation Results - Scenario 1 & 2

Estimator	Mean β	Std. Err.	RMSE
Scenario 1			
POLS-FE	0.995	0.083	0.084
MG-inf	1.000	0.027	0.027
PDOLS	1.000	0.020	0.020
PDOLS-DUM	0.995	0.111	0.111
AMG	1.000	0.030	0.030
CCE-MG	1.000	0.025	0.025
AP-DOLS	1.000	0.029	0.029
Scenario 2			
POLS-FE	1.002	0.174	0.174
MG-inf	1.000	0.025	0.025
PDOLS	2.279	0.117	1.285
PDOLS-DUM	0.914	1.130	1.133
AMG	1.009	0.082	0.083
CCE-MG	1.000	0.025	0.025
AP-DOLS	1.000	0.032	0.032

Note: 1,000 Monte Carlo Simulations with N = 50 and T = 50.

Table 1 presents the results for Scenario 1 and 2. No estimators exhibit bias in the first scenario since the regressor is exogenous. The most efficient estimators in this case were found to be PDOLS, closely followed by CCE-MG and then AP-DOLS. While pooled OLS (with time and unit fixed effects) and PDOLS with time dummies (used in the past as a crude method to account for common factors) were found to be unbias, they were both highly inefficient.

Adding cross-sectional dependence (i.e. common factors) in the second scenario introduces severe bias in Panel Dynamic OLS, as expected. Using time dummies in the estimation goes a long way in correcting the mean bias. However, the estimate becomes extremely inefficient (with a RMSE almost as high as regular PDOLS), and remains more biased than other alternatives. Although pooled OLS does not exhibit positive or negative bias in scenario 2 (perhaps surprisingly), it is highly inefficient. It is also worth noting that bias will be introduced into POLS-

FE whenever the coefficient is not randomly distributed around a fixed number as it is in this simulation study (see Bond & Eberhardt (2013) for a set of simulation results that show this).

AP-DOLS successfully removes the bias found in regular PDOLS, joining CCE-MG and AMG as the three estimators considered in this study that are robust to common factors. CCE-MG was found to be the most efficient in this scenario, followed closely by AP-DOLS and finally AMG. The results from the second scenario emphasise the importance of considering a second generation panel time series estimator (one that corrects for common factors) in any real empirical application where common factors are potentially present.

Table 2 presents the simulation results for Scenarios 3, 4, 5, and 6. Scenario 3 adds a feedback effect between the dependent variable and the regressor (a common form of endogeneity in certain economic variables), but removes the common factor structure. Unsurprisingly, with the presence of feedback effects PDOLS and AP-DOLS are the only unbias and efficient estimators. POLS-FE possesses high inefficiency, and CCE-MG and AMG display a degree of underestimation in the mean coefficient. Scenario 4 retains the feedback effect and returns the common factor structure to the data. AP-DOLS remains the only unbias and efficient estimator. In terms of RMSE, AP-DOLS is followed by CCE-MG and then AMG which features a relatively high standard error.

Scenario 5 retains the common factor structure and expands the feedback effect to three lags. With a complex feedback effect, all of the estimates possess bias (save for POLS-FE, yet once again it demonstrates very high inefficiency), yet AP-DOLS has the least bias and also maintains efficiency. The bias in CCE-MG is severe, and while AMG possesses les bias it is less efficient relative to the other two estimators. Lastly in the table, scenario 6 adds a non-linear feedback effect on the errors. The bias in AMG and CCE-MG become more severe in this scenario, yet the results for AP-DOLS improve slightly relative to the previous scenario.

To summarise the simulation results, AP-DOLS is the only estimator in the test to be unbias (or least bias as in scenario 5 and 6) and efficient across all six scenarios. While it possesses a slightly lower level of efficiency relative to CCE-MG, it offers substantially less bias (or no bias) in any data structure featuring feedback effects and/or common factors. The appendix provides supplementary simulation results that use different sample sizes, as well as a test on the selection of lag and lead length in AP-DOLS. The next section will test the various estimators used here on an empirical panel time series dataset.

Table 2—Simulation Results - Scenario 3, 4, 5, & 7

Estimator	Mean β	Std. Err.	RMSE
Scenario 3			
POLS-FE	0.996	0.080	0.080
MG-inf	0.953	0.025	0.053
PDOLS	1.000	0.020	0.020
PDOLS-DUM	1.005	0.107	0.107
AMG	0.940	0.023	0.064
CCE-MG	0.962	0.023	0.044
AP-DOLS	1.005	0.026	0.026
Scenario 4			
POLS-FE	1.002	0.174	0.174
MG-inf	0.967	0.024	0.041
PDOLS	2.279	0.117	1.285
PDOLS-DUM	0.895	1.120	1.124
AMG	0.986	0.080	0.081
CCE-MG	0.970	0.024	0.039
AP-DOLS	1.008	0.029	0.030
Scenario 5			
POLS-FE	1.002	0.174	0.174
MG-inf	0.920	0.025	0.083
PDOLS	2.279	0.117	1.285
PDOLS-DUM	0.905	1.122	1.126
AMG	0.965	0.082	0.089
CCE-MG	0.924	0.024	0.080
AP-DOLS	0.974	0.029	0.039
Scenario 6			
POLS-FE	1.002	0.174	0.174
MG-inf	0.880	0.024	0.122
PDOLS	2.279	0.117	1.285
PDOLS-DUM	0.894	1.102	1.107
AMG	0.928	0.077	0.105
CCE-MG	0.886	0.024	0.116
AP-DOLS	0.977	0.027	0.036

Note: 1,000 Monte Carlo Simulations with N = 50 and T = 50.

IV. Empirical Application: Inequality and Crime

The topic of inequality and crime is a fitting empirical application for panel time series econometrics, as it demonstrates the degree to which accounting for endogeneity can impact on the conclusion of the results. Empirical papers that have investigated the impact that inequality has on crime has a long history (e.g. Ehrlich (1973)), Kelly (2000), and Fajnzylber, Lederman and Loayza (2002)). Papers finding a positive relation, usually using cross-country statistics, agree with the theoretical literature (e.g. Becker (1968)) that high inequality increases the economic incentives for criminal activity.

Chintrakarn and Herzer (2012), to our knowledge, was the first paper in the literature to estimate the relationship using a panel cointegration approach. Their conclusion, quite surprisingly, was that the top 10% income share and the Gini coefficient had a negative relationship with the violent crime rate (an elasticity of -0.9 to -1.0). The estimator they used was Panel Dynamic OLS (both a pooled and mean group version), which they argued was robust to serial correlation and endogeneity. However, as shown in the simulation results the unbiasness of this estimator relies on cross-sectional independence. The assumption that inequality and crime do not share any common factors is extreme, considering the amount of variables that could easily be related across both variables and across panel units (e.g. wage growth, unemployment, government policy, and other national trends).

In order to determine whether there is a long run relationship between inequality and crime, this application will test a variety of estimators on the following basic regressions:

$$log(ViolentCrimeRate_{it}) = \beta_0 + \beta_1 log(Top10\%incomeshare_{it}) + e_{it}$$

$$log(PropertyCrimeRate_{it}) = \beta_0 + \beta_1 log(Top10\%incomeshare_{it}) + e_{it}$$

This section will use the same dataset seen in Chintrakarn and Herzer (2012), which according to the results in the paper are nonstationary and cointegrated. Crime data is taken from the FBI Uniform Crime Reports, and measures for income inequality on a US state level is taken from Frank (2009). The data ranges from 1960-2012 across all U.S States (including DC), and accordingly $T = 52$ and $N = 51$. This is a similarly dimensioned dataset to what was used in the Monte Carlo simulations. All variables will be transformed into logs before estimation in order to estimate the elasticity.

To determine if common factors are inherent in the data, a formal test for cross section dependence is undertaken. The specific test was formulated by Pesaran (2004), under a null hypothesis of no cross section dependence. For the purposes of the test, each of the two variables crime will be tested, as well as the top 10% income share and the residuals to the pooled OLS regressions. All variables (save the residuals) receive a log transformation before testing. Table 3 presents the results. Given the value of the test statistics, it is clear that cross section dependence pervades both the variables and the residuals of this dataset, and therefore using a panel time series estimator that is not robust to the existence of common factors will yield incorrect inference.

Table 3—Pesaran (2004) Test for Cross Section Dependence

Variable	Test Statistic	p-value
log(Top 10% Income Share)	225.32	0.000
log(Violent Crime Rate)	226.84	0.000
log(Property Crime Rate)	222.79	0.000
Residuals from Violent Crime Regression	199.65	0.000
Residuals from Property Crime Regression	217.96	0.000

Table 4 presents the estimation results from a variety of estimators on the two basic equations outlined above. The results are highly inconsistent across estimators. Pooled OLS with fixed effects report a negative elasticity between the top 10% income share and both the violent crime rate as well as the property crime rate. Panel Dynamic OLS offers a different prediction, depending on whether time dummies are used or not used. Without time dummies, PDOLS reports a positive average elasticity with both the violent crime rate and the property crime rate, although only the former is statistically significant. With time dummies, the elasticity is negative and statistically significant for both violent crime and property crime. The CCE-MG and AP-DOLS estimators, both robust to common factors and therefore more desirable for this analysis, reports a positive and statistically significant elasticity for the violent crime rate, and an insignificant elasticity for the property crime rate. With the former, the elasticity reported with AP-DOLS is significantly higher, potentially suggesting that feedback effects or other forms of endogeneity are also present in the data.

What information can be gained from these results? First of all, it demonstrates that with real world data the results can be highly sensitive to the type of panel time series estimator used, and therefore it is very important to ensure the best estimator for the data is being utilised. It is also clear that the negative elasticity reported in Chintrakarn and Herzer (2012) relies on the use of an inappropriate estimator. In these results we find limited evidence for the existence of a positive relationship between violent crime and inequality in US states, but none for property crime. Further analysis is required in order for the relationship between inequality and crime to be thoroughly understood.

Table 4—Regression Table - Inequality and Crime

Dep. Var.: log(Violent Crime Rate)		
Estimator	β	t-stat
POLS-FE	-0.880	-7.40***
MG	1.962	11.96***
PDOLS	1.608	17.42***
PDOLS-DUM	-0.518	-4.2***
CCE-MG	0.677	2.32**
AP-DOLS	0.949	3.958***
Dep. Var.: log(Property Crime Rate)		
Estimator	β	t-stat
POLS-FE	-1.368	-16.69***
MG	0.512	3.83***
PDOLS	0.186	0.124
PDOLS-DUM	-1.926	-16.62***
CCE-MG	-0.13	-1.00
AP-DOLS	0.037	0.949

Note: Inequality is measured by the top 10% income share. Data provided by Frank (2009) and FBI.

V. Conclusion

Panel time series estimators exist that account for endogeneity from feedback effects and endogeneity from common factors (i.e. cross-sectional dependence), but not both. Augmented Panel Dynamic OLS, introduced in this paper, is designed to be unbias and efficient in data structures with one, both, or neither forms of endogeneity. It takes the mean group Panel Dynamic OLS approach in Pedroni (2001) and augments the regressions with cross-section averages as seen in Pesaran (2006). Monte Carlo simulation studies show that this estimator is unbias and efficient in certain situations where competing estimators are not, and exhibits less bias when the structure of the feedback effects are highly complex. Since it also retains the computational simplicity of other popular estimators, it provides a clear improvement for data structures that contain both common shocks and feedback effects.

This estimator, along with others, was then applied to the empirical issue of the existence of a long run relationship between inequality and crime. Using a panel time series dataset of U.S. states over the last fifty years, the results found evidence for a positive relationship between inequality and violent crime, but not inequality and property crime. This is contrary to recent research that concluded inappropriately from the same dataset that there is evidence for a negative relationship. The results also illustrated the potential for results that are highly sensitive to the type of estimator used. Accordingly, it is important to carefully consider the most appropriate estimator for the data when conducting an applied panel time series study.

REFERENCES

Bai, J. 2009. “Panel data models with interactive fixed effects.” *Econometrica*, 77(4): 1229–1279.

Becker, G.S. 1968. “Crime and punishment: an economic approach.” *Journal of Political Economy*, 76: 169–217.

Bond, S., and M. Eberhardt. 2013. “Accounting for unobserved heterogeneity in panel time series models.” *Working Paper.*

Chintrakarn, P., and D. Herzer. 2012. “More inequality, more crime? A panel cointegration analysis for the United States.” *Economic Letters*, 116(3): 389–391.

Coakley, J., A.M. Fuertes, and R.P. Smith. 2006. “Unobserved heterogeneity in panel time series models.” *Computational Statistics & Data Analysis*, 50(9): 2361–2380.

Ehrlich, I. 1973. “Participation in illegitimate activities: a theoretical and empirical investigation.” *Journal of Political Economy*, 81: 521–565.

Fajnzylber, P., D. Lederman, and N. Loayza. 2002. “Inequality and violent crime.” *Journal of Law and Economics*, 45: 1–40.

Frank, M. 2009. “Inequality and Growth in the United States: Evidence from a New State-Level Panel of Income Inequality Measure.” *Economic Inquiry*, 47(1): 55–68.

Hayakawa, K., and E. Kurozumi. 2008. “The Role of ’Leads’ in the Dynamic OLS Estimation of Cointegrating Regression Models.” *Mathematics and Computers in Simulation*, 79(3): 555–560.

Kapetanios, G., M. Pesaran, and T. Yamagata. 2011. “Panels with Nonstationary Multifactor Error Structures.” *Journal of Econometrics*, 160(2): 326–348.

Kejriwal, M., and P. Perron. 2008. “Data Dependent Rules for Selection of the Number of Leads and Lags in the Dynamic OLS Cointegrating Regression.” *Econometric Theory*, 24(5): 1425–1441.

Kelly, M. 2000. “Inequality and Crime.” *Review of Economics and Statistics*, 82: 530–539.

Li, K., and L. Lina. 2014. “Efficient estimation of heterogeneous coefficients in panel data models with common shock.” *MPRA Working Paper*, 94(446): 621–634.

Moon, H., and M. Widner. 2012. “Linear regression for panel with unknown number of factors as interactive fixed effect.” *Working Paper, USC.*

Pedroni, P. 2000. "Fully Modified OLS for Heterogeneous Cointegrated Panels." *Advances in Econometrics*, 15: 93–130.

Pedroni, P. 2001. "Purchasing Power Parity Tests in Cointegrated Panels." *The Review of Economics and Statistics*, 83(4): 727–731.

Pesaran, M. 2004. "General Diagnostic Tests for Cross Section Dependence in Panels." *CESIFO WORKING PAPER NO. 1229.*

Pesaran, M. 2006. "Estimation and inference in large heterogeneous panels with a multifactor error structure." *Econometrica*, 74(4): 967–1012.

Pesaran, M., Y. Shin, and R. Smith. 1999. "Pooled Mean Group Estimation of Dynamic Heterogeneous Panels." *Journal of the American Statistical Association*, 94(446): 621–634.

Song, M. 2013. "Asymptotic theory for dynamic heterogeneous panels with cross-sectional dependence and its applications." *Working Paper, Columbia University.*

Stock, J., and M. Watson. 1993. "A Simple Estimator of Cointegrating Vectors in Higher Order Integrated Systems." *Econometrica*, 61(4): 783–820.

Supplementary Simulation Results

A1. Different Sample Sizes

Table A1 presents the Root Mean Squared Error ('RMSE') across a range of different NxT panel dimensions. The scenarios referenced here, and their underlying data generation process, are identical to the ones outlined in the main article. Various sample sizes were tested, including 30x50, 50x30, 30x30, and finally 80x80. Testing different sample sizes is important for simulation studies as it indicates how the performance of a specific estimator can change as you increase/decrease the width or the length of the panel. It also determines whether the relative comparisons between estimators outlined in the main part of the article are sensitive to the dimensionality of the data. The results will focus on the three second generation panel estimators studied in this article: CCE-MG, AMG, and of course AP-DOLS.

As seen in the table, CCE-MG remains the most efficient of the three estimators in Scenario 1 & 2 no matter the sample size, followed by AP-DOLS and then AMG. The efficiency of all three estimators appear to be more sensitive to a reduction in the time dimension as opposed to the number of panel units. Due to the feedback effect in scenario 3, both AMG and CCE-MG have a degree of bias in the β estimates, and again it appears that the time dimension is more important than the number of panel units in partially mitigating this bias. AP-DOLS is slightly more efficient in Scenario 3 than in Scenario 1 across all sample sizes. AP-DOLS, as before, provides an improvement over CCE-MG and AMG in Scenario 4, no matter the sample size, and a very significant improvement in bias in Scenario 5 & 6. Indeed, in Scenario 6 AP-DOLS contains around 33% - 50% of the RMSE of CCE-MG depending on the sample size. The main conclusions of the simulation study are not sensitive to changes in the dimensionality of the data. Firstly, AMG appears to be less efficient (or more bias) than CCE-MG in most of the scenarios outlined in this study. Secondly, AP-DOLS offers a tangible improvement to CCE-MG and AMG in any of the setups containing endogeneity from feedback effects.

Table A1—Simulation Results - Different Sample Sizes

Estimator	30x30	30x50	50x30	80x80
Scenario 1				
AMG	0.047	0.040	0.036	0.022
CCE-MG	0.040	0.032	0.033	0.018
AP-DOLS	0.066	0.037	0.053	0.019
Scenario 2				
AMG	0.102	0.108	0.077	0.070
CCE-MG	0.036	0.032	0.028	0.019
AP-DOLS	0.052	0.041	0.041	0.025
Scenario 3				
AMG	0.111	0.079	0.095	0.038
CCE-MG	0.075	0.049	0.070	0.029
AP-DOLS	0.056	0.034	0.042	0.018
Scenario 4				
AMG	0.099	0.104	0.075	0.070
CCE-MG	0.050	0.044	0.044	0.033
AP-DOLS	0.048	0.038	0.037	0.026
Scenario 5				
AMG	0.105	0.108	0.083	0.076
CCE-MG	0.089	0.082	0.085	0.073
AP-DOLS	0.085	0.057	0.081	0.040
Scenario 6				
AMG	0.119	0.117	0.101	0.094
CCE-MG	0.128	0.118	0.125	0.107
AP-DOLS	0.067	0.044	0.058	0.030

Note: RMSE results from 1,000 Monte Carlo Simulations across a range of n x t sample sizes.

A2. Lag/Lead structure with DOLS

As noted in Kejriwal and Perron (2008), there is no established methodology in selecting the number of lags and leads in a DOLS regression, much less in a panel time series framework. This section of the appendix investigates the issue with a

Monte Carlo simulation study. Using the same data generation process as above, the performance of AP-DOLS will be tested under the same six scenarios using a range of lag lengths in the DOLS regressions. Specific focus will be placed on the effect of adding more lags. It is expected that it might lead to a minor loss in efficiency in the simpler scenarios, due to a shortening of the time dimension. However, it may improve bias in the scenarios that contain more extreme forms of feedback effects. Additionally, motivated by the findings in Hayakawa and Kurozumi (2008), who argued that removing the leads in a DOLS regression will often result in a large efficiency gain, we also test AP-DOLS under the six scenarios without leads in the regression (while having the standard two lags).

Table A2 presents the RMSE from 1,000 Monte Carlo Simulations across the six scenarios under a range of lag and lead lengths in the AP-DOLS estimator as well as the removal of leads. Under Scenario 1 & 2, which lack any feedback effects, the table shows that efficiency declines non-linearly with the number of lags and leads. However, once a feedback effect is added to Scenarios 3 to 6, one lag becomes bias and has a significantly higher RMSE than all higher lag length selections. For a feedback effect of one lag found in Scenarios 3, 4, and 6, all lag and lead lengths above one are unbias yet lose efficiency beyond two lag and leads. For a feedback effect of three lags as found in Scenario 5, all estimators are slightly bias yet the use of 3 lag and leads in the regressions (unsurprisingly) offers a significantly improvement in RMSE. Furthermore, removing leads in data structures that do not contain feedback effects (Scenario 1 & 2) predictably leads to a slight efficiency gain, however this disappears and is radically reversed in all later scenarios. Removing leads adds a great deal to the RMSE in the last four scenarios. In conclusion, the use of at least two lags is recommended in any panel DOLS estimation. The choice of lags above that amount will depend on the expected structure of the feedback effect. Furthermore, based on these results removing the leads is not recommended in most applications.

Table A2—Simulation Results - Lag Length Selection for AP-DOLS

Scenario	1 Lag	2 Lags	3 Lags	4 Lags	5 Lags	6 Lags	No Leads
Scenario 1	0.028	0.030	0.033	0.036	0.042	0.048	0.026
Scenario 2	0.035	0.038	0.040	0.042	0.042	0.043	0.028
Scenario 3	0.066	0.027	0.029	0.031	0.035	0.039	0.036
Scenario 4	0.047	0.038	0.038	0.040	0.040	0.040	0.043
Scenario 5	0.066	0.051	0.040	0.051	0.055	0.058	0.104
Scenario 6	0.038	0.036	0.038	0.040	0.041	0.042	0.119

Note: RMSE results from 1,000 Monte Carlo Simulations of AP-DOLS with N = 50 and T = 50.