The Stata Journal (yyyy) **vv**, Number ii, pp. 1–10

Simplifying the Estimation of Correlated Random Effects Models

Fernando Rios-Avila
Levy Economics Institute of Bard College
Annandale-on-Hudson, NY 12504
friosavi@levy.org

Abstract.

This paper introduces the `cre` command, a prefix type command, that helps in the implementation of Correlated Random Effects (CRE) estimator for linear and nonlinear models. For the linear case, CRE models offer a simple approach that combines the advantages of both fixed effects and random effects estimators, providing consistent estimates identical to the Fixed Effect estimator, but allowing for the identification of coefficients for time-invariant variables. For the nonlinear case, it provides an alternative to fixed effects estimators that may be difficult to implement or simply non-existent. The `cre` command provides a user-friendly approach for estimating these models, supporting both balanced and unbalanced panels, which can be applied to most linear and nonlinear estimators.

Keywords: st0001, Mundlak approach, correlated random effects, panel data

1 Introduction

Panel data analysis has become increasingly important in empirical research in economics and social sciences, allowing researchers to control for unobserved individual heterogeneity that is believed to be fixed across time. The two main approaches that have been used to model relationships using this type of data have been fixed effects (FE) and random effects (RE) models. Each one of them, however, comes with limitations. On the one hand, while fixed effects models can be used to provide consistent estimates, while controlling for time-invariant unobserved factors, they cannot be used to estimate the effects of time-invariant variables, which may be relevant for some research questions. On the other hand, while random effects models allow you to identify effects of time-invariant variables, the estimation relies on the strong assumption that individual-specific effects are uncorrelated with other explanatory variables in the model, which is often violated in practice (Wooldridge 2019).

While less commonly used, there is a third option that shares some of the strengths of both FE and RE models: Correlated Random Effects (CRE) models. First introduced by Mundlak (1978) and further developed by Chamberlain (1982), CRE proposes a middle ground approach to address the limitations of RE models by explicitly allowing for correlation between the individual-specific effects and the time-varying explanatory variables. By doing so, CRE can estimate the effects of time-invariant variables while also providing consistent estimates for time-variant coefficients that are identical to the FE estimator in linear models. Despite these advantages, CRE models have seen limited use in applied research, partly due to the lack of readily available software implementations. [1]

This paper introduces the `cre` command for Stata, which aims to provide a straightforward and flexible tool for estimating CRE models for linear and non-linear models, supporting both balanced and unbalanced panels, as well as multiple fixed effects. This command is a prefix-type command that identifies all explanatory variables in a model, calculates the group means (or mean-like statistics) for each variable, and adds them to the model to provide a Mundlak (1978) type estimator. Because of this, there are few restrictions on the type of models that can be estimated using this command, making it a versatile tool for applied researchers. Furthermore, because it integrates seamlessly with Stata's existing estimation commands, all post-estimation commands and diagnostics can be used.

We begin in Section 2 reviewing the theoretical foundations of CRE models and their relationship to FE and RE models, extending our discussion to Multiple fixed effects, and application with non-linear models. Section 3 presents the implementation of CRE estimation in Stata, describing the syntax of the command. Finally, Section 4 presents a small Monte Carlo simulation to assess the performance of the CRE approach. Section 5 concludes.

2 Theoretical Framework

2.1 Correlated Random Effects Models - 1 Dimension

The Correlated Random Effects (CRE) is an alternative estimation approach for panel data models that was first introduced by Mundlak (1978) and further developed by Chamberlain (1982). In contrast with standard Fixed Effects estimator, CRE allows users to control and identify the effects of time-variant and time-invariant variables. And, in contrast with standard random effects estimator, CRE lifts the assumption that individual-specific effects are uncorrelated with other explanatory variables in the model. As pointed out in Wooldridge (2010), for the case of linear models, the CRE point estimates are identical to the Fixed Effects estimator.

To understand how CRE models work, let's consider the following data generating process:

$$y_{i,t} = \beta_0 + x_{i,t}\beta_x + z_i\beta_z + \alpha_i + u_{i,t} \tag{1}$$

where $y_{i,t}$ is the dependent variable for individual i at time t, $x_{i,t}$ is a vector of time-varying explanatory variables, z_i a set of time-invariant factors, α_i is the individual-specific effect, and $u_{i,t}$ is the idiosyncratic error term.

[1] StataNow released an option for the estimation of CRE models as part of the panel data estimators on June 25, 2024. There are also the community-contributed commands `xthybrid`(Schunck and Perales 2017) and `mundlak`(Perales 2013)

Under the assumption that α_i is uncorrelated with $x_{i,t}$, in addition to the standard assumption of exogeneity of u_{it}, Equation 1 could be consistently estimated using ordinary least squares (OLS), Random effects estimator, or fixed effects estimator.

In the case of using OLS, standard errors would need to be adjusted to account for the fact that α_i is an effect that is clustered within individuals. In the case of Fixed effects, if the panel data is balanced, one could simply demean all the variables with group individual means and estimate the model with the transformed data. This demeaning process would eliminate the individual-specific effect α_i from the model, but would also make it impossible to estimate the effects of time-invariant variables. In the case of random effects, one could quasi-demean the data, before estimating the model. This transformation eliminates the within-individual autocorrelation, allowing for the estimation of coefficients of time-invariant variables. However, this approach is not consistent if the assumption that α_i is uncorrelated with $x_{i,t}$ is violated.

The solution proposed by Mundlak (1978) and Chamberlain (1982) was to explicitly allow for correlation between the individual-specific effects and the time-varying explanatory variables, by assuming that the individual-specific effect can be expressed as a projection of (mean) time-varying variables plus an uncorrelated disturbance. Specifically:

$$\begin{aligned} Mundlak: &\quad \alpha_i = \gamma_0 + \bar{x}_i\gamma + v_i \\ Chamberlain: &\quad \alpha_i = \gamma_0 + x_{i,1}\gamma_1 + x_{i,2}\gamma_2 + \cdots + x_{i,T}\gamma_T + v_i \end{aligned} \tag{2}$$

where $\bar{x}_i$ is the individual specific mean of the time-varying variables, $x_{i,t}$ is the realization of x for individual i at time t, and v_i is an uncorrelated disturbance. The main difference between both approaches was that Chamberlain (1982) allowed for a more flexible specification of the correlation between the individual-specific effect and the time-varying variables. Mundlak (1978), on the other hand, assumed that the correlation was constant, only depending on the individual average. If we substitute Equation 2 into Equation 1, the final model can be written as:

$$y_{i,t} = \beta_0 + x_{i,t}\beta_x + z_i\beta_z + \gamma_0 + f(x_{i,t})\Gamma + v_i + u_{i,t} \tag{3}$$

where $f(x_{i,t})$ can be the full set of time-varying variables or just the average of them. Notice that in this specification, β_0 and γ_0 cannot be independently identified, and that the new model now has a compound error $v_i + u_{i,t} = \mu_{i,t}$, which is uncorrelated with $x_{i,t}$ by construction.

While this model could now be estimated using OLS, to account for the within-individual correlation driven by v_i, the model should be estimated using either random effects estimator, or clustering standard errors at the individual level (see Wooldridge (2010) for a discussion). Interestingly, both methods provide the same results for time-varying covariates if the panel data is balanced, and all covariates are strictly exogenous.[2] However, this identity breaks down in other cases (see Abrevaya (2013)).

[2] For time-invariant covariates, the RE estimator will be identical to Chamberlain (1982) approach

While both approaches will produce consistent estimates for the time-varying covariates, the implementation of Chamberlain (1982) is more difficult when the panel data is unbalanced (see Abrevaya (2013)). On the other hand, Mundlak (1978) approach is easier to implement with unbalanced panels, because it only requires the calculation of the individual means for the observed data for each individual, as it has been shown by Wooldridge (2019). Furthermore, Mundlak (1978) only requires adding a few covariates to the model regardless of the number of periods in the panel, compared to an increasing number of covariates in Chamberlain (1982) approach.

2.2 Correlated Random Effects Models - Multiple Dimensions

One potential advantage of CRE-Mundlak estimation that has been less discussed in the literature is that it can be easily extended to accommodate for multiple fixed effects/dimensions. In the standard case of panel data, for example, one may be interested in controlling for both individual and time fixed effects. Among the few papers discussing this extension, Baltagi (2023) focuses on formalizing the equivalence with two-way fixed effect estimation, while Wooldridge (2021) have discussed the advantages of CRE-Mundlak estimation for the identification of treatment effects in setups of staggered adoption of treatments. Both authors discuss the CRE-Mundlak approach in the context of two fixed effects, however, the extension to more than two fixed effects is straightforward.

Consider the following data generating process:

$$y_{i,t} = \beta_0 + x_{i,t}\beta_x + z_i\beta_z + w_t\beta_w + \alpha_i + \tau_t + u_{i,t} \tag{4}$$

In addition to the components from Equation 1, Equation 4 also considers individual-invariant variables w_t, as well as effects that only vary across time, but not individuals τ_t. As before, pool OLS or random effects estimators are only consistent if the individual-specific (α_i) and time-specific (τ_i) effects are uncorrelated with the explanatory variables. Without loss of generality let's assume that all variables have an overall mean of zero.

Extending the analogy from Equation 2, we can project the sum of individual-specific and time-specific effect as a function of the individual and time averages of $X's$. Other variables are not included because they already are invariant in one of the dimensions:

$$\alpha_i + \alpha_t = \gamma_0 + \check{x}_i\gamma + \check{x}_t\delta + v_{i,t} \tag{5}$$

Interestingly, if the panel data is balanced, $\check{x}_i$ and $\check{x}_t$ can be estimated simply as the individual or period specific average. Furthermore, they would be orthogonal and Equation 5 could be expressed using two equations like Equation 2, one for each dimension. In either case, the final model would be:

only

$$y_{i,t} = \beta_0 + x_{i,t}\beta_x + z_i\beta_z + w_t\beta_w + \gamma_0 + \check{x}_i\gamma + \check{x}_t\delta + v_{i,t} + u_{i,t} \tag{6}$$

where $v_{i,t}$ is the compound error term $(v_i + v_t)$ that is uncorrelated with $x_{i,t}$, which could be estimated using OLS. However, balanced panel data is not the norm.

When the panel data is unbalanced, $\check{x}_i$ and $\check{x}_t$ cannot be estimated as simple group averages. This is similar to the problem of using the within transformation for the estimation of M-way fixed effects (Rios-Avila 2015; Correia 2016).[3] In this case, instead of estimating $\check{x}$ as individual or group averages, one should estimate them as the solution to the following model:

$$x_{i,t} = \check{x}_i + \check{x}_t + \epsilon_{i,t} \tag{7}$$

Notice that we assume the constant to be zero, given the zero mean assumption. $\check{x}_i$ and $\check{x}_t$ in this model can be estimated using an iterative demeaning, as long as the sample used is the same one as in the original model (Equation 4). In addition, one should only concentrate on variables that show variation in both dimensions. Once these are estimated, the final model can be estimated using Equation 6.

Extending this analogy to three or more dimensions is straightforward. One simply requires to:

1. Define the sample that is common to all dimensions.
2. Estimate the group pseudo-averages for each dimension, and for all variables that show variation in all dimensions.
3. Include the new variables in the main model, and estimate that model using OLS.

2.3 Nonlinear Models and CRE

While the CRE approach has some advantages over FE and RE in linear models, including the option for a robust test to choose between RE and FE models, unless one is interested in the effects of time-invariant variables, the incentives to use CRE over FE are minimal in the framework of linear models. However, as discussed in Wooldridge (2019) and Wooldridge (2023), CRE can be particularly important to provide an alternative to fixed effect estimation in non-linear models, where the simple inclusion of dummies is not possible due to the incidental parameter problem, and a fixed effect estimator is not available. In fact, Wooldridge (2010) shows that the CRE approach applies to commonly used models, such as probit, tobit, and count models, among others.

Consider a data generating process for a non-linear model, where the latent linear index is given by:

[3] As described in Rios-Avila (2015), it is possible to implement a within-transformation using an iterative demeaning process until convergence. More recently, StataNow 18.5 also released a command that allows for the estimation of M-way fixed effects using a similar (yet more efficient) approach.

$$y^*_{i,t} = \beta_0 + x_{i,t}\beta_x + z_i\beta_z + \alpha_i \tag{8}$$

In this case, y^* is the expected latent variable that depends on time varying and time invariant variables, as well as an individual-specific effect. The observed variable $y_{i,t}$ can then be generated as a random draw from a distribution that depends on $y^*_{i,t}$. For example:

$$\begin{aligned} \text{probit:} \quad & y_{i,t} = 1\{y^*_{i,t} + u_{i,t} > 0\} \text{ with } u_{i,t} \sim N(0,1) \\ \text{poisson:} \quad & y_{i,t} \sim poisson(exp(y^*_{i,t})) \\ \text{tobit:} \quad & y_{i,t} = max(0, y^*_{i,t} + u_{i,t}) \text{ with } u_{i,t} \sim N(0,\sigma) \\ etc \quad & \end{aligned}$$

In all these cases, there is an explicit individual fixed component α_i, which may cause a bias on estimated coefficients if not accounted for. Furthermore, while estimating models with explicit fixed effects is possible for some models, like the poisson or logit, explicit fixed effects estimators are not available for probit or tobit models, among others, and simply including dummies for individual fixed effects does not yield consistent estimates due to the incidental parameter problem.

Wooldridge (2019) and Wooldridge (2010) show that the CRE approach can be used to estimate these models consistently, by following the same logic as in the linear case. That is, estimate the group specific means for the time-varying variables and include them in the model specification, before the estimation of the non-linear model. The linear model could be directly estimated using pooled cross-section, and clustering standard errors at the panel level. In fact, this approach has been used in Wooldridge (2023) to estimate treatment effects of staggered adoption of treatments in non-linear models, albeit concentrating on a single dimension (panel).

In this framework, `cre` has the additional advantage over other methods. Because one has explicit access to all the variables in the model, including the constructed variables, the estimation of partial effects can be done directly using available post-estimation commands such as `margins`. There are no other changes in the estimation process, and the strategy is easily extended to multiple fixed effects, as discussed in the previous section. Furthermore, as discussed in Wooldridge (2019), one can consider even more flexible specifications by allowing for interactions between the group-specific means and other variables in the model, including time dummy variables.

3 `cre` Command: Implementation in Stata

The `cre` command is a prefix command that estimates the CRE model for most linear and non-linear models in Stata. The syntax of the command is as follows:

```
    cre, abs(varlist) [options]: [estimation command] [variables] [if in] [weights],
[options]
```

Where `estimation command` is the command that will be used to estimate the model. In this case, the `cre` command will identify all the dependent variables in the model, along with other sample restrictions, and calculate conditional group means based on the groups defined by the `abs(varlist)`. The group means are then added to the model before estimation. The `cre` command supports both balanced and unbalanced panels, as well as multiple fixed effects relying on `reghdfe` for the estimation of the pseudo conditional means.

A detailed list of other options include:

- `abs(varlist)`: a list of variables that define the groups for which the conditional means will be calculated. This can be a single variable or a combination of variables.
- `drop`: if specified, the command will drop the group means variables from the dataset after estimation.
- `dropsingletons`: if specified, the command will drop observations that have a single observation in any given group. Note that `reghdfe` default is to drop these observations, rather than keep them as in this case.
- `prefix`: provides a prefix to be used to name the group means variables. The default uses "`m`" as a prefix. For every variable that has variation across all groups defined in `abs(varlist)`, the command will create a new variable with the prefix followed by the variable name. `m1_x`, `m2_x`, etc, corresponding to the group means for variable `x` for the first, second, etc group defined in `abs(varlist)`. Note that when trying to estimate means for interaction terms, the command will attempt to name the new variable using Stata guidelines for variable names. If the variable name is too long, the command will create a generic name `_v#`.
- `hdfe(options)`: options to be passed to `reghdfe` command. This can be used to make better use of alternative speed-up options in `reghdfe`.

In addition to the standard information left by the estimation command, the `cre` will add the list of group mean variables created for the estimation in `e(m_list)`.

The command has been tested for most 1 and 2 word commands, and after `ivreg` and `ivregress` commands. However, for unsupported commands, an alternative approach would be to use `cre` for a fully specified model, and then use the command of interest including the group means variables manually in the model.

4 Monte Carlo Simulations

To assess the performance of the `cre` command, we conducted a Monte Carlo simulation study. Instead of focusing on a model with a single fixed effect, we considered a case with two unobserved fixed effects, which allows for a structure of unbalanced panels. Without loss of generality, we also consider a model with 2 explanatory variables that are correlated with the unobserved fixed effects, but are otherwise exogenous. Finally, we consider that the expected latent variable is a linear function of the explanatory variables and fixed effects. The data generating process is as follows:

```
set obs 1000
// Generates indicators for the two fixed effects
gen id1 = runiformint(1,100)
gen id2 = runiformint(1,100)
// fixed effects are assumed follow a uniform distribution
gen c1 = runiform(-.5,.5)
gen c2 = runiform(-.5,.5)
bysort id1:replace c1 = c1[1]
bysort id2:replace c2 = c2[1]
// explanatory variables are correlated with the fixed effects,
// thus correlated with each other
gen x1 = runiform(-1,1)+invnormal(c1+.5)+invnormal(c2+.5)
gen x2 = runiform(-1,1)-invnormal(c1+.5)-invnormal(c2+.5)
// and the expected latent ey_star is a linear combination x1, x2, c1, and c2
gen y_star = 1 + x1 + x2 + c1 + c2
```

Since the equivalence between FE and CRE has been shown for linear models, we focus on the performance of the `cre` in the context of non-linear models. For this, we consider 4 different models: probit, fractional regression probit, tobit, and poisson.

For each one, the data generating process is as follows:

```
// probit
gen y_probit = 1*(y_star-1+rnormal()>0)
// fractional probit
gen y_fprobit = normal(y_star-1+rnormal())
// tobit
gen y_tobit = max( y_star + rnormal(),0)
// poisson
gen y_poisson = rpoisson(exp(y_star))
```

We then estimate the models under three assumptions: one where the fixed effects are observed including `c1` and `c2` in the specification (benchmark), one where they are not considered, and one where we use the CRE approach using the `cre` command. This exercise is repeated 1000 times, and the results are compared in terms of the distributions of the marginal effects for the probit and fractional probit model, and coefficients for the tobit and poisson models, under the three scenarios. We use `parallel`(Vega Yon and Quistorff 2019) to speed up the process.

The results of the simulation are presented in Figure 1 and Table 1. The figure presents the densities of the estimated coefficients across all simulations, whereas the table provides a brief summary of the bias and mean absolute error of the estimated coefficients.

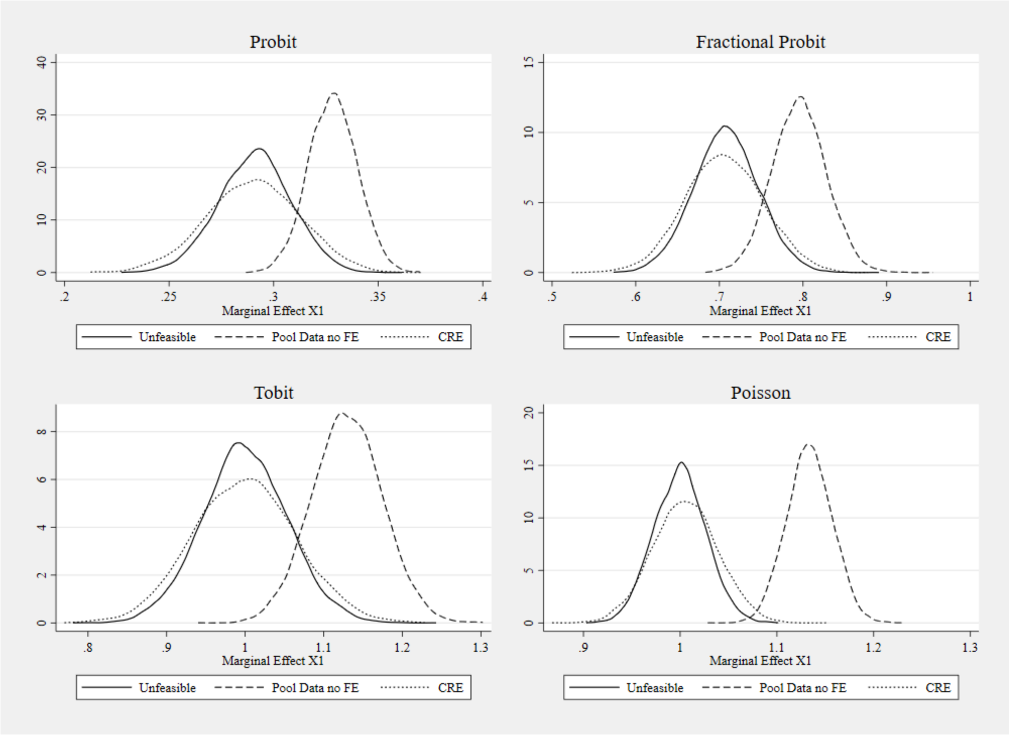

Figure 1: Estimated marginal effects/Coefficient densities for non-linear models

As expected, controlling for the unobserved effects, the unfeasible estimator provides what we would consider to be the benchmark/true estimates of coefficients/marginal effects given the sample and data generating process. For Table 1, we use the mean of the unfeasible estimator to represent the true point estimate and construct the bias and mean absolute error for the other estimators.

Because the individual effects are constructed to be correlated with the explanatory variables, the estimated coefficients are biased when they are ignored, and coefficients are estimated using pooled estimators. The magnitude of the bias is approximately 12-13% with respect to the point estimate of the coefficient or marginal effect.

On the other hand, the CRE approach seems to provide consistent estimates for the coefficients/marginal effects, with a distribution that is centered around the true point estimate, with slightly higher variance than the benchmark. Based on the results from Table 1, the bias of the CRE approach is negligible, with a MAE that is about 20% to 30% larger than the benchmark.

Table 1: Bias and MAE for the estimated marginal effects/Coefficients for non-linear models

	Probit	FProbit	Tobit	Poisson
True:Bias	-0.000	-0.000	-0.000	-0.000
True:MAE	0.014	0.031	0.043	0.022
Pool:Bias	0.037	0.087	0.130	0.134
Pool:MAE	0.037	0.087	0.130	0.134
CRE:Bias	-0.001	-0.002	-0.001	0.005
CRE:MAE	0.018	0.037	0.052	0.027
N	10000	10000	10000	10000

5 Conclusion

This paper introduces the `cre` command, a prefix-type command that facilitates the implementation of Correlated Random Effects (CRE) models with a wide range of official and user-written Stata estimation commands. The CRE approach offers a middle ground between fixed effects and random effects models, addressing some of their limitations, particularly in the context of nonlinear model estimation.

As previously discussed in the literature, our Monte Carlo simulations show that the CRE approach, implemented through the `cre` command, consistently estimates coefficients and marginal effects, performing comparably to the unfeasible estimators that directly control for unobserved factors. The simulations reveal negligible bias, with an increase in the variance of the estimates, which is consistent with theoretical expectations.

The `cre` command addresses a significant gap in Stata's econometric toolkit, providing a user-friendly implementation of CRE models. This tool may prove valuable for researchers working with panel data or nested data structures, where standard fixed effects approaches may be challenging or nonexistent, and random effects assumptions are not appropriate.

6 Acknowledgments

Thanks to Aashima Sinha for her help in the preparation and comments for this paper, and Enrique Pinzon for his encouragement on pushing this project forward.

7 References

Abrevaya, J. 2013. The projection approach for unbalanced panel data. *The Econometrics Journal* 16(2): 161–178. http://dx.doi.org/10.1111/j.1368-423X.2012.00389.x.

Baltagi, B. H. 2023. The two-way Mundlak estimator. *Econometric Reviews* 42(2): 240–246. https://www.tandfonline.com/doi/full/10.1080/07474938.2023.2178139.

Chamberlain, G. 1982. Multivariate regression models for panel data. *Journal of econometrics* 18(1): 5–46.

Correia, S. 2016. A Feasible Estimator for Linear Models with Multi-Way Fixed Effects. *Unpublished Manuscript* .

Mundlak, Y. 1978. On the pooling of time series and cross section data. *Econometrica: journal of the Econometric Society* 69–85.

Perales, F. 2013. MUNDLAK: Stata module to estimate random-effects regressions adding group-means of independent variables to the model. Statistical Software Components, Boston College Department of Economics. https://ideas.repec.org/c/boc/bocode/s457601.html.

Rios-Avila, F. 2015. Feasible Fitting of Linear Models with N Fixed Effects. *The Stata Journal* 15(3): 881–898. Publisher: SAGE Publications.

Schunck, R., and F. Perales. 2017. Within- and Between-cluster Effects in Generalized Linear Mixed Models: A Discussion of Approaches and the Xthybrid command. *The Stata Journal* 17(1): 89–115. https://doi.org/10.1177/1536867X1701700106.

Vega Yon, G. G., and B. Quistorff. 2019. parallel: A command for parallel computing. *The Stata Journal* 19(3): 667–684. https://doi.org/10.1177/1536867X19874242.

Wooldridge, J. M. 2010. *Econometric analysis of cross section and panel data*. MIT press.

———. 2019. Correlated random effects models with unbalanced panels. *Journal of Econometrics* 211(1): 137–150.

———. 2021. Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators. Working paper. Available at SSRN: https://ssrn.com/abstract=3906345.

———. 2023. Simple approaches to nonlinear difference-in-differences with panel data. *The Econometrics Journal* 26(3): C31–C66. https://doi.org/10.1093/ectj/utad016.

About the authors

Fernando Rios-Avila is a research scholar at the Levy Economics Institute of Bard College under the Distribution of Income and Wealth program. His research interests include applied econometrics, labor economics, and poverty and inequality. He has contributed many commands to Statistical Software Components and written articles for the Stata Journal.