Simplifying the Application of Correlated Random Effects

Fernando Rios-Avila Aashima Sinha

2024-09-03

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 cae, 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. The `cre` command provides a user-friendly command for estimating these models, supporting both balanced and unbalanced panels, which can be applied to most linear and nonlinear estimators. We demonstrate the command's functionality through Monte Carlo simulations and an empirical application.

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. One the one hand, while fixed effects models can be used to provide concistent estimates, while controlling for time-invariant unobserved factors, they cannot be used to estimate the effects of time-invariant variables, which is may be relevant for some research questions. On the other hand, while random effects models allows you to identify effects of time invariant variables, the estimation relies on the strong assumptions that individual-specific effects are uncorrelated with other explanatory variables in the model, which is often violated in practice (Jeffrey M. 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 that to address the limita-

tions 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 provide 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 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 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 present the implementation of CRE estimation in Stata, describing the syntax of the command. Finally, Section presents an empirical example using showing the application of these methods in both linear and nonlinear contexts, followed by a Monte Carlo simulation in Section to assess their performance under various conditions. **?@sec-6** concludes.

Theoretical Framework

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-invairant variables. And, in contrast with standard random effects estamator, CRE lifts the assumption that individual-specific effects are uncorrelated with other explanatory variables in the model. As pointed out in Jeffrey M. Wooldridge (2010), for the case of linear models, the CRE point estimates are identical to the Fixed Effects estimator.

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

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.

Under the assumption that α_i is uncorrelated with $x_{i,t}$, in addition to the standard assumption of exogeneity of $u_i t$, Equation 1 could be consitently estimated using ordinarity 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 invidivuals. 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 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 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 Jeffrey M. 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)).

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 Jeffrey M. Wooldridge (2019). Furthermore, Mundlak (1978) only requires adding only 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.

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 accomodate 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 Jeffrey M. 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

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

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-specif 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 ortogonal and Equation 5 could be expressed using two equations like Equation 2, one for each dimension. In either case, the final model would be:

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

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 dimensions is straightforward. One simply requires to:

1. Define the sample that is common to all dimensions.

[3]As described in Rios-Avila (2015), it is possible to implement a within-transformation using an interative 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.

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.

Nonlinear Models and CRE

While the CRE approach has some advantages over FE and RE in linear 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 been discussed in Jeffrey M. Wooldridge (2019) and Jeffrey M. 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.

Consider the following data generating process for a non-linear model:

$$
\begin{aligned}
y_{i,t}^{*} &= \beta_0 + x_{i,t}\beta_x + z_i\beta_z + \alpha_i \\
y_{i,t} &= 1\{y_{i,t}^{*} > 0\}
\end{aligned}
$$

`cre` Command: Implementation in Stata

Empirical Application

Monte Carlo Simulations

Conclusion

Abrevaya, Jason. 2013. "The Projection Approach for Unbalanced Panel Data." *The Econometrics Journal* 16 (2): 161–78. https://doi.org/10.1111/j.1368-423x.2012.00389.x.

Baltagi, Badi H. 2023. "The Two-Way Mundlak Estimator." *Econometric Reviews* 42 (2): 240–46. https://doi.org/10.1080/07474938.2023.2178139.

Chamberlain, Gary. 1982. "Multivariate Regression Models for Panel Data." *Journal of Econometrics* 18 (1): 5–46.

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

Mundlak, Yair. 1978. "On the Pooling of Time Series and Cross Section Data." *Econometrica: Journal of the Econometric Society*, 69–85.

Perales, Francisco. 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, Fernando. 2015. "Feasible Fitting of Linear Models with N Fixed Effects." *The Stata Journal* 15 (3): 881–98.

Schunck, Reinhard, and Francisco 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.

Wooldridge, Jeffrey 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): 137–50.

———. 2023. "Simple approaches to nonlinear difference-in-differences with panel data." *The Econometrics Journal* 26 (3): C31–66. https://doi.org/10.1093/ectj/utad016.

Wooldridge, Jeffrey M. 2021. "Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators." Working Paper.