The Stata Journal (yyyy) **vv**, Number ii, pp. 1–30

Simplifying the Estimation of Correlated Random Effects Models

Fernando Rios-Avila
Universidad Privada Boliviana
La Paz, Bolivia
f.rios.a@gmail.com

Abstract.

This paper introduces `cre`, a Stata prefix command designed to simplify the implementation of Correlated Random Effects (CRE) models, following the Mundlak (1978) approach, for a wide range of linear and nonlinear estimation commands. Standard Fixed Effects (FE) estimators, while consistent under unobserved heterogeneity, cannot identify coefficients for time-invariant variables. Standard Random Effects (RE) estimators can identify such coefficients but rely on the strong, often violated, assumption that individual effects are uncorrelated with regressors. CRE models offer an empirical middle ground, providing FE-equivalent estimates for time-varying coefficients in linear models while still allowing identification of time-invariant effects. The `cre` command facilitates this by automatically generating the required group means of time-varying regressors and adding them to the specified model, handling both balanced and unbalanced panels. It also integrates seamlessly with Stata's factor variables and post-estimation commands like `margins`, enhancing its usefulness for applied researchers.

Keywords: st0001, Mundlak approach, correlated random effects, panel data, nonlinear models, Stata, prefix command

1 Introduction

Panel data offers significant advantages for empirical research by allowing researchers to control for unobserved individual heterogeneity that remains constant over time. The two dominant approaches for analyzing such data are fixed effects (FE) and random effects (RE) models. However, both have limitations. FE models provide consistent estimates by eliminating time-invariant unobserved factors, but consequently, cannot estimate the effects of observed time-invariant variables (e.g., gender, race, baseline characteristics), which are often of substantive interest. RE models estimate effects for both time-varying and time-invariant variables but require the strong assumption that the unobserved individual-specific effects are uncorrelated with the explanatory variables—an assumption frequently questioned in practice (Wooldridge 2019b). Violating this assumption leads to inconsistent estimates.

A third, less commonly implemented approach, offers a simple yet feasible alternative: Correlated Random Effects (CRE) models. First introduced by Mundlak (1978) and further developed by Chamberlain (1982), the CRE framework explicitly models the correlation between the unobserved individual effects and the explanatory variables. Specifically, the Mundlak (1978) approach, which is the focus here, assumes the individual effect can be written as a linear projection of the *individual-level-means* of the time-varying covariates, plus a random component uncorrelated with the covariates. Thus, to avoid the problems related to the RE assumption, the Mundlak approach simply augments the model specification by including the means of the time-varying covariates as additional regressors. 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.

Perhaps the most significant advantage of the Mundlak CRE approach lies in its applicability to **nonlinear models** (e.g., probit, logit, tobit, Poisson). For such models, FE estimation is either computationally intensive, suffers from the incidental parameter problem (leading to inconsistency as T remains small), or simply does not exist (Wooldridge 2019b). The CRE approach provides a practical and consistent estimation strategy that can be easily applied to such cases, typically with cluster-robust standard errors (Wooldridge 2010, 2019b) for statistical inference.

Despite these benefits, CRE models are not as widely used as FE or RE, partly due to perceived implementation hurdles, and lack of readily available, flexible software tools. While Stata recently introduced a `cre` option for `xtreg` (StataCorp 2025, `xtreg, cre`), its use is restricted to linear models estimated by `xtreg`. On the other hand, Community-contributed commands like `mundlak` (Perales 2013) and `xthybrid` (Schunck and Perales 2017) exist. While `mundlack` focuses primarly on linear model, `xthybrid` is the closest to the principles presented here in that it is a wrapper to `meglm`, allowing the application of CRE approach to any nonlinear model that are supported by `meglm`, extending the CRE estimator to the generalized mixed-effects framework.[1]

In contrast to `xthybrid`, `cre` is not a wrapper for any specific command. Instead, it is a **prefix command** that can be used with most Stata estimation commands (both official and community-contributed) to implement the Mundlak CRE approach. It does so by identifying all time-varying regressors in the model, as well as the relevant sample restrictions imposed by the user, to compute individual-specific means for all time varying covariates, and adding them to the original model specification. This flexibility provides the following advantages:

- Allows researchers to apply the Mundlak CRE approach to a wide range of models, including linear and nonlinear specifications, without being limited to specific commands or model types.
- Automatically generates the necessary constructed variables (mean variables) to be included in the model.
- It correctly handles unbalanced panels by calculating means using only the available observations for each individual within the estimation sample, following the approach described by Wooldridge (2019b).

[1] The command `xthybrid` goes beyond the Mundlak approach, focusing on the estimation of within and between effects, random-slope models, in addition to the standard Mundlak approach or CRE.

- It fully supports Stata's factor variables (`i.`, `c.` and `l.`).
- and last but not least, it integrates smoothly with post-estimation commands `margins`, if the original command allows, for the estimation average partial effects (APEs).

This paper proceeds as follows: Section 2 reviews the theoretical framework for CRE models in linear and nonlinear settings, discussing unbalanced panels and estimation of standard errors, and average partial effects estimation. Section 3 presents the `cre` command. Section 4 provides an empirical example. Section 5 provides Monte Carlo simulation evidence on the performance of the CRE approach for nonlinear models. Section 6 concludes.

2 Theoretical Framework

2.1 Correlated Random Effects Models - Linear Case

Consider a standard linear panel data model:

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

where $y_{i,t}$ is the outcome for individual i at time t, $x_{i,t}$ is a $1 \times K$ vector of time-varying explanatory variables, z_i is a $1 \times G$ vector of time-invariant variables, α_i is the unobserved time-invariant individual-specific effect, and $u_{i,t}$ is the idiosyncratic error term, assumed uncorrelated with $x_{i,t}$, z_i, and α_i for all t.

The standard RE estimator assumes $Cov(x_{i,t}, \alpha_i) = 0$ and $Cov(z_i, \alpha_i) = 0$. If this holds, RE is consistent and efficient. However, if $Cov(x_{i,t}, \alpha_i) \neq 0$, the RE estimator is inconsistent. The FE estimator addresses this by transforming the data (e.g., demeaning) to eliminate α_i, yielding consistent estimates for β_x. However, this transformation also eliminates z_i, making β_z unidentified.

The Mundlak (1978) CRE approach offers a solution to this problem by explicitly modeling the correlation between α_i and the time-varying covariates $x_{i,t}$. Specifically, it assumes that α_i can be written as a linear projection of individual means of $x_{i,t}$:[2]

$$E[\alpha_i|\bar{x}_i] = \gamma_0 + \bar{x}_i\gamma$$

where $\bar{x}_i = T_i^{-1}\sum_{t=1}^{T_i} x_{i,t}$ is the $1 \times K$ vector of individual-specific means of the time-varying variables for individual i over the T_i periods they are observed. This allows us to write α_i as:

[2] 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.

$$\alpha_i = \gamma_0 + \bar{x}_i \gamma + v_i \tag{2}$$

where v_i is a random component such that $E[v_i|\bar{x}_i] = 0$. More explicitly, the Mundlak approach assumes that once we control for $\bar{x}_i$, v_i is also uncorrelated with the *full history* of covariates $x_i = (x_{i,1}, ..., x_{i,T_i})$. Substituting Equation 2 into Equation 1 yields the CRE model specification:

$$y_{i,t} = (\beta_0 + \gamma_0) + x_{i,t}\beta_x + z_i\beta_z + \bar{x}_i\gamma + v_i + u_{i,t} \tag{3}$$

This augmented model can be estimated using pooled OLS or a RE estimator. The composite error term $\mu_{i,t} = v_i + u_{i,t}$ is uncorrelated with $x_{i,t}$, z_i, and $\bar{x}_i$ by construction (under the Mundlak assumptions), but will be correlated across time for the same individual. This correlation can be accounted for using cluster-robust standard errors at the individual level, or by applying the RE estimator.

The CRE estimator has some important properties:

1. The estimator for β_x from Equation 3 is numerically identical to the FE estimator in the linear case (Mundlak 1978; Wooldridge 2010, chap 10) for all time varying variables.
2. The model allows estimation of β_z, the coefficients on time-invariant variables.[3]
3. A test of $H_0 : \gamma = 0$ provides a robust test for correlation between α_i and $x_{i,t}$, effectively a Hausman-type test comparing FE and RE (Wooldridge 2010; StataCorp 2025, pg 521,531).

As mentioned before, one caveat of the methodology is that, unless ν_i=0 for every observation, $\mu_{i,t}$ is correlated across time by construction. Because of this, standard errors of the estimated coefficients of interest have to be estimated with this into consideration. This could imply using clustered standard errors or using a GLS (RE estimator) (Wooldridge 2010, Chapter 10, pg 332).

2.2 Handling Unbalanced Panels

Traditional Approach

A significant advantage of the Mundlak approach is its straightforward application to unbalanced panels. As shown by Wooldridge (2019a), the individual means $\bar{x}_i$ are simply calculated using the available T_i observations for each individual i present in the estimation sample.

[3] Note that depending if the model is estimated via pooled OLS or RE, the coefficients of the time-invariant variables will be different.

More specifically, for the case of panel data, the appropriate within-individual average for variable w_{ij} should be estimated as:

$$\bar{w}_i = \frac{1}{T_i} \sum_{t=1}^{T} s_{i,t} w_{i,t}$$

where $s_{i,t}$ is an indicator variable that takes the value of 1 if all other the elements of the control variables are dependent variables are observed, and 0 if any of the variables are missing.

The estimation of Equation 3 then proceeds using the pooled data, and using the within group averages as controls. This contrasts with the Chamberlain (1982) approach, which requires conditioning on x_i values from all periods and becomes complex with unbalanced data (Abrevaya 2013).

Alternative Approach

While the above approach is the most common, it is much simpler to leverage the use of regression analysis and the equivalence of using a dummy inclusion approach. This makes use of the advances in the estimation of fixed effect models as described in Correia (2016), Rios-Avila (2015), and more recently Stata's multiway fixed effects estimator (StataCorp (2025)).

In this framework, for each variable w_{ij}, we simply estimate the following regression:

$$w_{i,t} = a_0^w + \epsilon_i^w + \mu_{i,t}^w \tag{4}$$

Constraining the sample to $s_{i,t} = 1$, and under the assumption that the $E(\epsilon_i^w) = 0$. Under this assumption, a_0^w is equal to the grand mean of w_{ij}.

As is in the case of linear regression with dummies, it follows that $\bar{w}_i = a_0 + \epsilon_i^w$. With this change, however, rather than including the within group mean $\bar{w}_i$ in the model, we include the individual fixed effect ϵ_i^w instead.[4] While less explored, this approach can be used to expand the Mundlak approach to cases with multiple dimensions, as shown in the appendix.

Other approaches

It is important to notice that the simple CRE approach may not be valid for other contexts, such as the one described in Albarran et al. (2019), for dynamic and autoregressive nonlinear models. In particular, the estimation of such problems is more

[4] This parallels a control function approach, where exogenous variables help isolate an exogenous component via regression, and residuals control for endogeneity. In the Mundlak CRE case, individual fixed effects serve a similar role by capturing the component potentially correlated with unobserved heterogeneity; thus, instead of residuals $\mu_{i,t}^2$, we include fixed effects ϵ_i^w in the main model

difficult due to the “initial conditions problem”, due to using a lag of the dependent variable as explanatory variable. In this case, the simple Mundlak framework is not valid. For those interested in this approach, the authors propose a methodology for the general nonlinear case, and a community contributed command `xtprobitunbal`, which implements the approach for probit models.

2.3 Correlated Random Effects Models - Multiple Dimensions

One potential advantage of CRE-Mundlak estimation is that it can be easily extended to accommodate for multiple fixed effects/dimensions, specially in the case of balanced panels. In the standard case of panel data, for example, one may be interested in controlling for both individual and time fixed effects. However, in cases with nested data, such as students in schools, or patients in hospitals, one may also be interested in controlling for multiple levels of fixed effects.

There are few papers discussing this extension. Baltagi (2023) focuses on formalizing the equivalence with two-way fixed effect estimation. More recently, Baltagi (2024) and Yang (2022), have extended this identity to the case of multidimensional fixed effects, albeit only for the case of balanced data (where there are the same number of observations for every combination of the fixed effects dimensions).

Consider the following data generating process:

$$y_{i,t} = \beta_0 + x_{i,t}\beta_x + z_i\beta_z + w_t\beta_w + \alpha_i + \tau_t + u_{i,t} \tag{5}$$

and that the data is balanced. That is that every combination of i and t is observed, and they are all the same size (1 for panel data).

In addition to the components from Equation 1, Equation 5 also considers individual-invariant variables z_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_t) effects are uncorrelated with the explanatory variables. Following the same principle, Baltagi (2024) and Yang (2022) suggest that CRE could also be expanded to multiple dimensions, and that, if the data is balanced, one simply needs to include the group means of every variable respect to each FE dimension. In other words, they expand the analogy propose by Mundlak (1978), but for every dimension:

$$\begin{aligned} \alpha_i &= \gamma_0 + \bar{x}_i\gamma_x + v_i \\ \theta_t &= \theta_0 + \bar{x}_t\theta_x + v_t \end{aligned} \tag{6}$$

As noted in Wooldridge (2019b), while this procedure is straighforward for the case of panel data, the same expression is not valid when the data is unbalanced, and units i are observed different number of periods. Because of this, simply including the group means of the variables in the model does not fully control for the correlation between the unobserved effects and the explanatory variables. This is similar to the problem of using the within transformation for the estimation of M-way fixed effects (Rios-Avila 2015; Correia 2016; StataCorp 2025, `areg`)[5].

In the case of unbalanced data, the alternative approach described in Section 2.2 is more appropriate. Thus, each independent variable is regressed on both sets of fixed effects:

$$x_{i,t} = a_0^x + \epsilon_i^x + \epsilon_t^x + \mu_{i,t}^x \tag{7}$$

Using the sample where all data is observed ($s_{i,t} = 1$), and impossing the condition $E(\epsilon_i^x) = 0$ and $E(\epsilon_t^x) = 0$. It follows that a_0^x is the overall mean of the variable $x_{i,t}$, and ϵ_i^x and ϵ_t^x are the individual and time fixed effects, respectively. $\mu_{i,t}^x$ is now uncorrelated with $x_{i,t}$.

Interestingly, if the panel data is balanced, $a_0^x + \epsilon_i^x$ and $a_0^x + \epsilon_t^x$ are equivalent to the group means of the variables. However, that is not the case when the data is unbalanced.

With this, we can write the CRE model as:

$$y_{i,t} = \beta_0 + x_{i,t}\beta_x + z_i\beta_z + w_t\beta_w + \gamma_0 + \hat{\epsilon}_i^x\gamma + \hat{\epsilon}_t^x\delta + v_i + v_t + u_{i,t} \tag{8}$$

where v_i and v_t are the individual and time error term that are uncorrelated with $x_{i,t}$. As shown by Baltagi (2024) and Yang (2022), this can be estimated via pooled OLS or GLS (RE estimator), with the estimates for β_x being equivalent to the MW-FE estimates.

Thus, in general, it is feasible to extend the Mundlak approach to multiple dimensions, following the next steps:

1. Define the sample that is common to all dimensions and variables $s_{i,t}$.
2. Estimate the group pseudo-averages for each dimension ϵ_k^w, and for all variables.
3. Include the new variables in the main model, and estimate that model using OLS or GLS.

2.4 Nonlinear Models and CRE

The CRE approach becomes particularly valuable for nonlinear models where FE estimation faces challenges, as has been shown in Wooldridge (2019b) and Schunck and Perales (2017). Wooldridge (2019b) discusses how the Mundlak CRE approach extends naturally to nonlinear models, exploring more flexible cases, such as including interactions of the mean variables with time dummies, as well as suggesting the specifications for approximations to Random coefficients models. Schunck and Perales (2017) also

[5] 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.

discusses the Mundlak approach, implementing it in the `xthybrid` command, which allows for the estimation of CRE models, among others, in nonlinear settings withing the generalized linear mixed-effects framework.

Consider a general nonlinear model where the relevant latent variable depends on individual effects:

$$y^*_{i,t} = x_{i,t}\beta_x + z_i\beta_z + \alpha_i \tag{9}$$

where $y_{i,t}$ is observed based on $y^*_{i,t}$ (e.g., $y_{i,t} = 1(y^*_{i,t} + u_{i,t} > 0)$ for binomial models).

Including dummy variables for α_i in nonlinear models generally leads to inconsistent estimates for β_x and β_z due to the incidental parameter problem (Lancaster 2000; Neyman and Scott 1948). While consistent FE estimators exist for some specific models (e.g., conditional logit, FE Poisson), they are unavailable for many others (e.g., probit, tobit, ordered models).

Wooldridge (2019b) and Wooldridge (2010, chap 15.8) shows that the Mundlak CRE approach extends naturally to these cases. We maintain the assumption from Equation 2 that $\alpha_i = \gamma_0 + \bar{x}_i\gamma_x + v_i$, which leads to the following augmented model:

$$y^*_{i,t} = (\beta_0 + \gamma_0) + x_{i,t}\beta_x + z_i\beta_z + \bar{x}_i\gamma_x + v_i$$

This specification of the latent variable can then be used in the nonlinear model of interest. The model can then be estimated by applying the standard pooled estimator (e.g., pooled probit, pooled Tobit) to the augmented specification. The parameters β_x, β_z, and γ_x are consistently estimated under appropriate assumptions for the specific nonlinear model, provided the conditional expectation or density is correctly specified (Wooldridge 2019b). Cluster-robust standard errors at the individual level are required for statistical inference, although accounting for the the within correlation using GLS may also be appropriate.[6]

This approach provides consistent estimates of the parameters for both time-varying and time-invariant variables, even when traditional FE methods fail or are unavailable. A recent example of its application is the work of Wooldridge (2023), where this strategy is used for estimating treatment effects with staggered adoption in nonlinear settings. The flexibility can be further increased by including interactions between $\bar{x}_i$ and time dummies, or other variables, in the specification (Wooldridge 2019b). In my implementation of `cre`, I focus on the simplest Mundlack approach, however, more complex models can be estimated following Wooldridge (2019b).

2.5 Standard Errors and Hypothesis Testing

As described above, estimating a model via CRE, as in Equation 3, involves using generated regressors ($\bar{x}_i$). While the point estimates for β_x in the linear model are

[6] This is, for example, what `xthybrid` does in the framework of Generalized linear mixed models.

identical to FE, the standard errors are not necessarily the same, even asymptotically, because the inclusion of $\bar{x}_i$ changes the model structure compared to the demeaning process of FE, affecting the new compounded error and degrees of freedom of the model. Specifically, unless $\bar{x}_i$ can fully explain the unobserved component α_i (thus making $\nu_i = 0$), the procedure introduces intra group correlation in the model.

For **linear models**, Wooldridge (2010, chap 10.5.3) shows that using the standard RE estimator (GLS) on the augmented equation Equation 3 yields standard errors for β_x that are asymptotically equivalent to the usual FE standard errors. Alternatively, estimating Equation 3 by pooled OLS and using cluster-robust standard errors (clustered at the individual level i) also yields asymptotically valid standard errors, which are equivalent to the clustered FE standard errors.

For **nonlinear models**, the standard approach is to estimate the augmented model using the pooled maximum likelihood estimator (e.g., `probit`, `logit`, `poisson`) and compute standard errors clustered at the individual level (`vce(cluster id)`). This accounts for the within-individual correlation induced by v_i (and potentially $u_{i,t}$ if serially correlated) (Wooldridge 2010, chap 15.8). Alternatively, RE estimators could also be applied to the augmented model, although this is less common in practice.

A more robust, though computationally intensive, alternative for obtaining standard errors, especially if the distributional assumptions about v_i or $u_{i,t}$ are uncertain, is to **bootstrap** the entire estimation process. This involves resampling individuals (clusters) with replacement, recalculating the $\bar{x}_i$ within each bootstrap sample, estimating the augmented model, and obtaining the distribution of the coefficients across bootstrap replications.

With multiple dimensions, the literature says very little regarding the correct estimaton of standard errors. Because using multiple dimensions can also introduce within-groups correlation, it may be that one needs to control for multiple levels of clustering, or use Bootstrap methods to at least account for one dimension adjustments. This is an area that requires further research, but the implemention is left as an experimental feature.

2.6 Calculating Average Partial Effects (APEs)

The calculation of average partial effects in linear models is straight forward. Because the CRE approach provides consistent estimates that are equivalent to the FE estimator, the partial effects can be calculated directly from the estimated coefficients as is usual. In nonlinear models, the estimated coefficients typically do not directly represent the partial effect of a covariate on the outcome variable, specially in the presence of interactions of polynomials. In this case, Average Partial Effects (APEs) or Marginal Effects at Means (MEMs) are necessary to better understand the results of a model. A significant practical advantage of the CRE approach implemented via `cre` is the relative ease of calculating APEs.

Based on Wooldridge (2019b) (section 5), the APE for a variable $x_{i,t}$ in a nonlinear model with unobserved heterogneity can be calculated as follows:

1. Define the outcome Average Structural Function (ASF) for the model, as:

$$ASF(x_{i,t}) = E_{c_i}[m_t(x_{i,t}, c_i)]$$

which eliminates the unobserved heterogeneity c_i from the model, by averaging it out.

2. When estimating nonlinear models using the CRE approach, rather averaging over the unobserved heterogeneity c_i, one averages over the auxiliary terms $\bar{x}_i$. Specifically we would have:

$$ASF(x_{i,t}) = \frac{1}{N_s} \sum_{i,t} [m_t(x_{i,t}, \bar{x}_i)]$$

3. The APE for a variable $x_{i,t}$ is then calculated as the derivative of the ASF with respect to $x_{i,t}$:

$$APE(x_{i,t}) = \frac{\partial ASF(x_{i,t})}{\partial x_{i,t}} = \frac{1}{N_s} \sum_{i,t} \frac{\partial m_t(x_{i,t}, \bar{x}_i)}{\partial x_{i,t}}$$

where N_s is the number of observations in the estimation sample.

In other words, when we apply the CRE approach, the auxiliary terms $\bar{x}_i$ are treated as fixed, and independent of $x_{i,t}$. This is the same as assuming other control variables remain fixed when focusing on the impact of a single variable. On this point, Wooldridge (2019b) argues that standard errors for this expression can be difficult to obtain. However, Stata's `margins` command provides a convenient way to calculate APEs and their standard errors based on the delta method.

3 `cre` Command: Implementation in Stata

3.1 Overview: how `cre` works

Up to this point, we have discussed the theoretical framework for the CRE approach, its advantages, and how it can be applied to both linear and nonlinear models. The `cre` command is designed to facilitate the implementation of this approach in Stata, allowing users to easily apply the Mundlak CRE method to a wide range of estimation commands. The process behind `cre` involves several key steps:

1. Identification of dependent and independent variables: `cre`, as a prefix command, intercepts the estimation command to identify the dependent variable, independent variables, and any sample restrictions (e.g., `if`, `in`, `weight`). On top of that, it also identifies the variables specified in the `abs()` option, which are used to calculate the individual means, and may also modify the working sample.

2. Calculation of psudo-means: For each independent variable in the model, `cre` centers the variable around its "grand mean", so that the new auxiliary variable has a mean of zero. This is then plugged in into a regression model that only includes the fixed effects (Equation 7). The estimated fixed effects are used as the psudo-means/auxiliary variables. If there is no variation left in the variable, after the fixed effects are absorbed, the psudo-mean are excluded from the model. This allows for the original variable to remain in the model. If the variable list includes interaction terms or other type of factor notation (e.g., `i.`, `c.`, etc), the expression is expanded and a temporary variable created to estimate their psudo-means.

3. Estimation of the augmented model: After calculating the individual means are finalized, `cre` augments the original model specification by adding the calculated means as additional independent variables. It then executes the original estimation command with the modified variable list, and the originally specified options (e.g., `vce(cluster id)` for clustered standard errors).

4. Post-estimation: Because there is no modification to the original estimation command, any post-estimation commands can be used as usual. This includes commands like `margins` for calculating average partial effects (APEs).

In other words, `cre` acts as a flexible prefix command that allows users to apply the Mundlak CRE approach to a wide range of estimation commands in Stata, without requiring significant changes to their existing workflows. It automates the process of calculating individual means and augmenting the model specification, while still allowing for the use of standard Stata estimation and post-estimation commands.

3.2 Syntax and Usage of `cre`

The syntax of the command is:

```
    cre, abs(varlist) [options] : estimation_command depvar [indepvars] [if]
[in] [weight] [, est_options]
```

Required Option:

- `abs(varlist)`: Specifies the variable(s) identifying the groups (individuals) for which means should be calculated. Typically, this is the panel identifier variable (e.g., `abs(personid)`). Multiple variables can also be specified.[^ We provide this as an experimental option to allow for a psudo-Multi-way Mundlak specification. While this is numerically equivalent to specifying multiple fixed effects in linear models, the theoretical justification for this is limited, specially for non-linear models.]

Optional Options:

- `prefix(str)`: Sets the prefix for the generated mean variables. The default is `m1_`. For a variable `x`, the mean variable will be named `m1_x`. If multiple variables are specified in `abs()`, prefixes like `m1_`, `m2_` might be used. Check `e(m_list)` for the names of the generated variables.

- `hdfe(options)`: These options are passed directly to the `reghdfe` command (Correia 2016), which `cre` uses internally to efficiently compute the group means, especially useful for large datasets or complex fixed effects structures. This is mainly for performance tuning.

- `dropsingletons`: By default observations belonging to singleton groups (individuals observed only once) are not excluded from the specification. This is in contrast with `reghdfe`'s typical behavior in absorbing effects. However, one can use `dropsingletons` to drop these observations from the estimation sample, as its done in `reghdfe`.

  Note: Singletons provide no within-individual variation, so their impact on FE-equivalent estimates is null, but they might influence estimates of time-invariant variables or overall sample size in nonlinear models.

- `drop`: If specified, the generated mean variables are dropped from the dataset after the estimation command completes. The default is to keep them for potential inspection or use in post-estimation.

Stored Results:

In addition to the results stored by the `estimation_command`, `cre` adds the following to `e()`:

- `e(m_list)`: A list of the names of the generated mean variables added to the model.
- `e(abs_vars)`: The variable(s) specified in `abs()`.

Dependencies:

The `cre` command requires the `reghdfe` (Correia 2016) and `ftools` packages to be installed. These packages are used internally to efficiently handle the fixed effects structure.

4 Empirical Application

To illustrate the use of `cre` and compare it with alternative estimators, we use the `nlswork.dta` dataset bundled with Stata. This dataset contains panel data on young women from 1968-1988. We will estimate models for the log wage (`ln_wage`) using a linear model, Wages (Transformation from Log Wage) using a poisson mmodel, and union membership (`union`), based on a logit regression. The panel identifier is `idcode`.

For this example, I will use a simple set of variables and interactions, but the approach can be easily extended to more complex specifications.

```
// Load and setup data
webuse nlswork, clear
xtset idcode year
gen wage = exp(ln_wage) // Transform log wage to wage for poisson model
gen white = race==1 & race!=.
global depvar1 ln_wage
global depvar2 wage
global depvar3 union
global indep "age c.age##c.age c.tenure##c.tenure i.white i.south"
```

4.1 Linear Model: Log Wage

I start with a simple Log-Linear model using a linear regression approach, estimating the model using different methods: Fixed Effects (FE), Random Effects (RE), Stata's built-in CRE (`xtreg, cre`), and the `cre` prefix with `xtreg, re`, and with `regress, cluster()`.

```
// 1. Fixed Effects (xtreg, fe)
xtreg $depvar1 $indep, fe
est sto m1
// 2. Random Effects (xtreg, re)
xtreg $depvar1 $indep, re
est sto m2
// 3. Stata's built-in CRE (xtreg, cre)
xtreg $depvar1 $indep, cre
est sto m3
// CRE prefix - xtreg
cre, abs(idcode): xtreg $depvar1 $indep, re
est sto m4
// CRE prefix - regress
cre, abs(idcode): reg $depvar1 $indep, cluster(idcode)
est sto m5
```

Table 1: Comparison of Linear Models

		xtreg		CRE	
	FE	RE	CRE	xtreg RE	regress
main					
age	0.042	0.047	0.042	0.042	0.042
	(9.74)	(11.47)	(14.78)	(9.74)	(9.74)
c.age#c.age	-0.001	-0.001	-0.001	-0.001	-0.001
	(-6.89)	(-8.57)	(-10.69)	(-6.89)	(-6.89)
tenure	0.041	0.048	0.041	0.041	0.041
	(18.99)	(23.19)	(25.19)	(18.99)	(18.99)
c.tenure#c.tenure	-0.001	-0.002	-0.001	-0.001	-0.001
	(-10.23)	(-11.94)	(-13.68)	(-10.23)	(-10.23)
1.white	0.000	0.090	0.081	0.081	0.093
	(.)	(7.79)	(6.93)	(7.27)	(8.02)
1.south	-0.071	-0.121	-0.071	-0.071	-0.071
	(-4.08)	(-11.56)	(-6.41)	(-4.08)	(-4.08)
m1_age				0.051	0.076
				(4.38)	(5.95)
m1_c_age_c_age				-0.001	-0.001
				(-4.47)	(-5.94)
m1_tenure				0.074	0.067
				(11.10)	(9.11)
m1_c_tenure_c_tenure				-0.003	-0.003
				(-6.09)	(-5.01)
m1___1_south				-0.105	-0.111
				(-5.14)	(-5.36)
_cons	0.835	0.715	-0.031	0.792	0.769
	(13.34)	(12.11)	(-0.21)	(12.51)	(12.16)
xt_means					
age			0.051		
			(4.82)		
c.age#c.age			-0.001		
			(-4.99)		
tenure			0.074		
			(11.88)		
c.tenure#c.tenure			-0.003		
			(-6.66)		
1.white			0.000		
			(.)		
1.south			-0.105		
			(-6.66)		
N	28093	28093	28093	28093	28093

t statistics in parentheses

Table 1 shows the comparison of the different estimation methods for the linear model with one-way fixed effects. Comparing the Fixed effect and Random Effects models, we see that the coefficients across specification is slightly different, as expected. The coefficient for "white" is drop in the FE model, as it is a time-invariant variable. We provide t-statistics as statistic of reference, because standard errors are too small to be displayed in the table. Perhaps the largest difference between the FE and RE models is the coefficient for `south`, which suggest a larger negative relationship with wages in the RE model.

Turing to the `xtreg, CRE` (official implementation) as expected, provides the same coefficients, and almost identical standard errors as the `xtreg, fe` for time varying variables, with a coefficient for `white` that is very close to the one in the RE model. Notice that the `xtreg, cre` also includes in the output the coefficients for the variables averages, but as part of a different equation. The standard errors difference between the `xtreg, fe` and `xtreg, cre` is due to the different degrees of freedom used in the estimation.

The last two columns show the results of the `cre` prefix, using `xtreg, re` and `regress, cluster(idcode)` as the estimation commands. First of all, `cre` produces idential results as Stata's `xtreg, cre`, for both main variables and the auxiliary variables. The last column shows the results that combines `cre` with `regress, cluster(idcode)`. Again, the coefficients are identical for time-varying variables, but a small difference for the time constant variables and the auxiliary variables. This happens because `xtreg, re` introduces a transformation of the data (quasi-differentiation) that is not done by `regress`. Also notice that the t-statistics, thus standard errors, are very different to the previous models. This happens because, in contrast with models 1-4, using `cluster(idcode)` with `regress` not only addresses within individual correlation, but also provides heteroskedasticity-robust standard errors. In the appendix, I also provide results for all other models using `robust` standard errors, which provide results identical to `regress`.[7]

4.2 Nonlinear Model: Poisson and Logit Models

To show how `cre` can be used with nonlinear models, I estimate two-nonlinear models: a Poisson model for the wage variable and a logit model for union membership, with similar specifications as before, but with some differences.

1. `xtpoisson, fe` automatically excludes groups with a single observation. Thus we also impose this restriction in the `cre` prefix command(`dropsingletons` option). I report results using fe and re options.
2. For `xthybrid`, I use the `family(poisson)` and `link(log)` options to specify the Poisson model, in addition to `cre` option. I also manually create the polynomial terms for `age` and `tenure`, since `xthybrid` does not support factor notation.

[7] In addition to the above mentioned cases, while experimental, I also provide in the appendix, results for two-way fixed effects: idcode and year.

3. For comparison, I use `xtpoisson, re`, `poisson` regression with clutered standard errors, as well as `meglm` with `family(poisson)` and `link(log)`, and allowing for random effects at the individual level. For this last model, because of the unique structure of `meglm`, `cre` cannot be used on it directly. Instead, I used the variables created when calling `poisson` to estimate the model with `meglm`.
4. In all cases, I use robust standard errors.

```
* ssc install xthybrid, replace // If needed

webuse nlswork, clear
xtset idcode year
gen wage = exp(ln_wage) // Transform log wage to wage for poisson model
gen white = race==1 & race!=.
global depvar1 ln_wage
global depvar2 wage
global depvar3 union
global indep "c.age##c.age c.tenure##c.tenure i.white i.south"

drop if white==. | union==. | age==. | tenure==. | south==.
bysort idcode:gen nobs=_N
drop if nobs==1

gen age_sqr = age*age
gen tenure_sqr  = tenure*tenure
global indep2 "age age_sqr tenure tenure_sqr white south"

xtpoisson $depvar2 $indep, fe vce(robust)
est sto m1
xtpoisson $depvar2 $indep, re vce(robust)
est sto m2
xthybrid $depvar2 $indep2 , clusterid(idcode) ///
    vce(robust) family(poisson) link(log)  cre se
est sto m3
cre, abs(idcode): poisson $depvar2 $indep, cluster(idcode)
est sto m4
cre, abs(idcode): xtpoisson $depvar2 $indep, re vce(robust)
est sto m5
meglm  $depvar2 $indep m1_* || idcode:, ///
    family(poisson) link(log) vce(robust)
est sto m6
```

Table 2 presents the results using poisson models. A before columns 1 and 2, represent Stata's official implementation. These results are very similar to what we observed in Table 1. One noticible difference is that the coefficient for south is more than twice as large in the RE compared to the FE model. On Column 3, we present the results using `xthybrid` (community contributed command). The coefficients are almost identical to those in the FE model, and slighly larger Standard errors. Note that the `xthybrid` command produces results that are somewhat different from standard Stata estimation commands. However, those coefficients were renamed to match the output from `cre` in columns 4 to 6.

The last three columns show the results using the `cre` prefix with `poisson`, `xtpoisson, re`, and `meglm`. Numerically, all three approaches produce very similar, but not identical, results to the FE model, with `xtpoisson, re` and `meglm` producing the closest estimates in terms of coefficients and standard errors.

Table 2: Comparison of Poisson models.

	xtpoisson				CRE	
	FE	RE	xthybrid	Poisson	xtpoisson	meglm
wage						
age	0.029	0.040	0.028	0.027	0.028	0.028
	(4.14)	(5.76)	(4.07)	(3.92)	(4.07)	(4.07)
c.age#c.age	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000
	(-1.35)	(-2.97)	(-1.30)	(-1.23)	(-1.30)	(-1.30)
tenure	0.037	0.045	0.037	0.036	0.037	0.037
	(10.89)	(13.24)	(10.80)	(10.07)	(10.79)	(10.80)
c.tenure#c.tenure	-0.002	-0.002	-0.002	-0.002	-0.002	-0.002
	(-4.94)	(-6.64)	(-4.90)	(-4.57)	(-4.89)	(-4.90)
1.white	0.000	0.095	0.093	0.104	0.097	0.093
	(.)	(6.82)	(7.31)	(6.78)	(6.80)	(7.31)
1.south	-0.093	-0.142	-0.094	-0.100	-0.094	-0.094
	(-3.56)	(-9.33)	(-3.52)	(-3.29)	(-3.50)	(-3.52)
m1_age			0.108	0.152	0.123	0.108
			(6.95)	(7.47)	(7.37)	(6.95)
m1_c_age_c_age			-0.002	-0.003	-0.002	-0.002
			(-7.01)	(-7.35)	(-7.31)	(-7.01)
m1_tenure			0.063	0.047	0.059	0.063
			(6.17)	(3.10)	(4.82)	(6.17)
m1_c_tenure_c_tenure			-0.002	-0.001	-0.002	-0.002
			(-2.37)	(-0.95)	(-1.80)	(-2.37)
m1__1_south			-0.087	-0.081	-0.087	-0.087
			(-2.93)	(-2.33)	(-2.85)	(-2.93)
_cons		0.837	-0.646	1.007	1.001	0.953
		(8.77)	(-3.35)	(10.42)	(10.31)	(9.81)
/						
lnalpha		-2.238			-2.324	
		(-7.40)			(-7.27)	
var(_cons[idcode])						0.095
						(24.88)
var(_cons[idcode])						
_cons			0.095			
			(24.88)			
N	27541	27541	27541	27541	27541	27541

t statistics in parentheses

Because nonlinear models coefficients do not represent APE on the outcome variable, we also estimate them using the `margins` command for **`age`** and **`tenure`**. Unfortunately, `xtpoisson, fe` does not support using `margins` for the variable in levels, and using `xtpoisson, re` can only be used with `margins` if the option `normal` is specified at the moment of estimation. Similarly, xthybrid does not support margins. As alternative, I use ppmlhdfe(Correia et al. 2020) to estimate the fixed effects poisson model. Similarly, for cre, we concentrate on the cases that use poisson or meglm.[8] In all cases, I only present the APE for age and tenure.

```
qui:ppmlhdfe $depvar2 $indep, abs(idcode) vce(robust) d
margins, dydx(age tenure) post
est sto m1mfx
qui:cre, abs(idcode): poisson $depvar2 $indep, cluster(idcode)
margins, dydx(age tenure) post
est sto m2mfx
qui:meglm  $depvar2 $indep m1_* || idcode:, ///
    family(poisson) link(log) vce(robust)
margins, dydx(age tenure) post
est sto m3mfx
```

The results for APE are presented in Table 3. In this table, we will consder ppmlhdfe as the benchmark, and compare with cre estimation. However, it is important to note that the ppmlhdfe command does not account the error contribution of the individual fixed effects, which may explain the considerable differences in t-statistics across implementations. Other than that, both cre models that provide results within 3 decimal places of the ppmlhdfe estimates.

Table 3: Comparison of Poisson models: APE

	(1)	(2)	(3)
	ppmlhdfe	CRE poisson	CRE meglm
age	0.101	0.100	0.101
	(12.11)	(8.85)	(9.30)
tenure	0.132	0.129	0.131
	(15.63)	(11.24)	(11.68)
N	18334	18334	18334

t statistics in parentheses

Lastly, I also I provide a similar excercise for a logit model, using union membership as dependent variable. In addition to the previous restrictions, the sample only considers those with non-missing values for the union variable, and with some variation in union membership over time. Given the results are qualitative similar to the previous case, I only present results for APE of age and tenure.

```
webuse nlswork, clear
```

[8] Note that ppmlhdfe produces point estimates that are identical to xtpoisson, fe, but with slightly different standard errors.

```
xtset idcode year
gen wage = exp(ln_wage) // Transform log wage to wage for poisson model
gen white = race==1 & race!=.
global depvar1 ln_wage
global depvar2 wage
global depvar3 union
global indep "c.age##c.age c.tenure##c.tenure i.white i.south"

drop if white==. | union==. | age==. | tenure==. | south==.
bysort idcode:gen nobs=_N
drop if nobs==1
bysort idcode:egen sig_union=sd(union)
drop if sig_union==0 | sig_union==.

xtlogit $depvar3 $indep, fe
margins, dydx(age tenure) post
est sto m1mfx
xtlogit $depvar3 $indep, re vce(robust)
margins, dydx(age tenure) post
est sto m2mfx
qui:cre, abs(idcode): logit $depvar3 $indep, cluster(idcode)
margins, dydx(age tenure) post
est sto m3mfx
qui:cre, abs(idcode): xtlogit $depvar3 $indep, re vce(robust)
margins, dydx(age tenure) post
est sto m4mfx
```

Table 4: Comparison of logit models: APE

	(1)	(2)	(3)	(4)
	xtlogit FE	xtlogit RE	CRE logit	CRE xtlogit RE
age	-0.000	0.001	-0.000	-0.000
	(-0.09)	(0.52)	(-0.12)	(-0.10)
tenure	0.023	0.018	0.020	0.021
	(5.71)	(7.07)	(6.44)	(6.46)
N	7522	7522	7522	7522

t statistics in parentheses

The restrictions imposed on this last model reduce the sample to just over 7,500 observations. The estimated APE between `xtlogit, fe` and `xtlogit, re` are similar. For age, `xtlogit, fe` produces a APE of almost zero, with a small possitive effect of tenure (0.023). For the `xtlogit, re` model, the estimates for age are slightly larger (0.001), but still non-significant, with a slighly smaller effect for tenure (0.018). finally, looking at the `cre` results, the APE's are closer to the `xtlogit, fe` estimates. In this case, in average, an additional year of tenure is associated with a 0.02pp increase in the probability of being a union member.

5 Monte Carlo Simulations

To assess the performance of the `cre` command, we conducted two Monte Carlo simulation study. The first one, shown here, considers a single unobserved fixed effect, since this is the most case. The second one, presented in the appendix, considers two unobserved fixed effects, which allows for a more complex structure of unbalanced panels. The data generating process is as follows:

```
// Setup
clear
set seed 12345
set obs 1000
// Generates indicators for one-way fixed effects
gen id1 = runiformint(1,100)
// fixed effect is assumed follow a uniform distribution
gen c1 = runiform(-.5,.5)
bysort id1:replace c1 = c1[1]
// explanatory variables are correlated with the fixed effects,
// thus correlated with each other
gen x1 = runiform(-1,1)+invnormal(c1+.5)
gen x2 = runiform(-1,1)-invnormal(c1+.5)
// and the expected latent ey_star is a linear combination x1, x2 and c1
gen y_star = 1 + x1 + x2 + c1
```

We focus on the performance of `cre` for four nonlinear models derived from `y_star`: probit, fractional probit, tobit, and poisson. The outcomes are generated as follows:

```
// Generate Observed Outcomes
// 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))
```

For each model type, we estimate three specifications over 10000 Monte Carlo replications:

1. **Unfeasible Benchmark:** Model estimated including the true fixed effect `c1` as a regressor.

2. **Pooled Estimator:** Model estimated ignoring `c1`
3. **CRE Estimator:** Model estimated using `cre, abs(id1): ...` including `x1`, `x2` but not `c1`.

We compare the distribution of estimated coefficients (or APEs for probit/fractional) for `x1` and `x2` across the methods, focusing on bias and Mean Absolute Error (MAE) relative to the unfeasible benchmark average. We use the `parallel` command (Vega Yon and Quistorff 2019) for efficiency.

The results of the simulation are presented in Figure 1 and Table 5. Figure 1 shows the densities of the estimated coefficients (or APEs) for the key parameters across simulations, while Table 5 summarizes the bias and MAE.

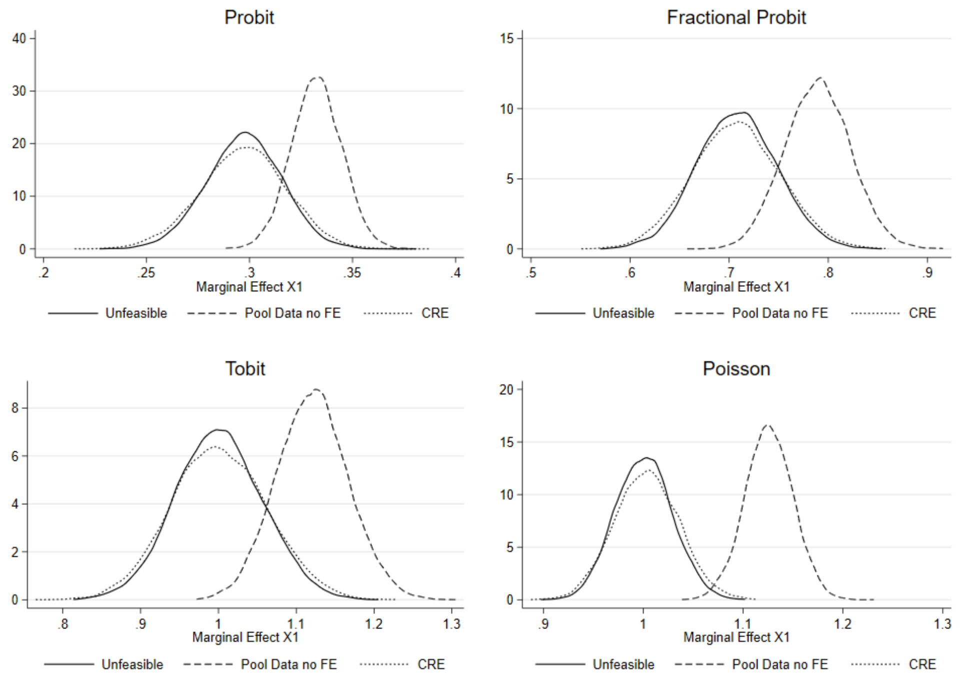

Figure 1: Estimated APE/Coefficient densities for non-linear models

As expected, the unfeasible estimator, controlling directly for the unobserved effect `c1`, provides the benchmark estimates. Since the unobserved effect `c1` is correlated with `x1` and `x2` by construction, the pooled estimators (which ignore `c1`) exhibit significant bias, as seen in both the density plots and the summary table.

In contrast, the CRE approach, implemented using the `cre` prefix, yields estimates whose distributions are centered close to the benchmark estimates, indicating negligible bias. While the CRE estimates show slightly higher variance compared to the unfeasible benchmark (as reflected in slightly larger MAE in Table 5), they effectively mitigate the bias caused by the omitted correlated fixed effect. This demonstrates the utility of the CRE method for obtaining consistent estimates in nonlinear panel models where FE is not viable and RE assumptions are violated.

Table 5: Bias and MAE for the estimated APE/Coefficients for non-linear models (Coefficients for `x1`)

	Probit	FProbit	Tobit	Poisson
True:Bias	0.000	-0.000	-0.000	-0.000
True:MAE	0.015	0.032	0.045	0.023
Pool:Bias	0.035	0.081	0.120	0.125
Pool:MAE	0.035	0.081	0.120	0.125
CRE:Bias	-0.000	-0.001	0.000	0.003
CRE:MAE	0.016	0.035	0.049	0.026
N	10000	10000	10000	10000

In the appendix, we extend the simulation to a more complex setting with two unobserved fixed effects, which further illustrates the use of the CRE approach in handling unbalanced panels and multiple dimensions of unobserved heterogeneity. As mentioned before, however, when multiple fixed effects are used, statistical inference remains challenging, and the results should be interpreted with caution.

6 Conclusion

This paper introduced `cre`, a versatile Stata prefix command designed to simplify the estimation of Correlated Random Effects (CRE) models based on the Mundlak (1978) specification. The CRE approach provides a valuable bridge between standard Fixed Effects and Random Effects models, offering several advantages: it allows for the estimation of time-invariant variable effects (unlike FE) while providing consistent estimates for time-varying coefficients even when the strict RE exogeneity assumption fails (matching FE estimates in linear models).

The primary contribution of the `cre` command lies in its flexibility and ease of use. As a prefix command, it can be applied to a wide array of standard Stata estimation commands, including user-written ones. It automatically handles the generation of individual means for time-varying covariates, supports both balanced and unbalanced panels, and integrates seamlessly with factor variables and post-estimation tools like `margins` for calculating Average Partial Effects, which is particularly crucial for interpreting nonlinear models.

While `cre` offers a significant simplification for estimating static, and potentially dynamic, CRE models, it does not address dynamic auto-regressive panel models. The inclusion of lagged dependent variables introduces further econometric challenges (e.g., the initial conditions problem) that require specialized estimators beyond the scope of this command.

In summary, the `cre` command provides applied researchers with a user-friendly and powerful tool for leveraging the benefits of the Correlated Random Effects approach in Stata, making it easier to estimate models that account for unobserved heterogeneity while retaining the ability to analyze the effects of time-invariant characteristics, especially in nonlinear settings.

7 Acknowledgments

Thanks to Aashima Sinha for her help in the preparation and providing feedback for this paper, and Enrique Pinzon for his encouragement on pushing this project forward. I am also grateful to the Editor and an anonymous referee for their constructive comments on a previous version, which significantly improved the focus and clarity of this paper.

8 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.

Albarran, P., R. Carrasco, and J. M. Carro. 2019. Estimation of Dynamic Nonlinear Random Effects Models with Unbalanced Panels. *Oxford Bulletin of Economics and Statistics* 81(6): 1424–1441. http://dx.doi.org/10.1111/obes.12308.

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.

———. 2024. The multidimensional Mundlak estimator. *Economics Letters* 236: 111607. http://dx.doi.org/10.1016/j.econlet.2024.111607.

Chamberlain, G. 1982. Multivariate regression models for panel data. *Journal of econometrics* 18(1): 5–46. https://doi.org/10.1016/0304-4076(82)90094-X.

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

Correia, S., P. Guimarães, and T. Zylkin. 2020. Fast Poisson estimation with high-dimensional fixed effects. *The Stata Journal: Promoting communications on statistics and Stata* 20(1): 95–115. http://dx.doi.org/10.1177/1536867X20909691.

Lancaster, T. 2000. The incidental parameter problem since 1948. *Journal of Econometrics* 95(2): 391–413. http://dx.doi.org/10.1016/S0304-4076(99)00044-5.

Mundlak, Y. 1978. On the pooling of time series and cross section data. *Econometrica: journal of the Econometric Society* 46(1): 69–85. https://doi.org/10.2307/1913646.

Neyman, J., and E. L. Scott. 1948. Consistent Estimates Based on Partially Consistent Observations. *Econometrica* 16(1): 1. http://dx.doi.org/10.2307/1914288.

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. https://doi.org/10.1177/1536867X1501500318.

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.

StataCorp. 2025. *Stata 19 Base Reference Manual*. StataCorp LLC, College Station, TX. https://www.stata.com/manuals/rareg.pdf.

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.

———. 2019a. *Introductory Econometrics: A Modern Approach*. 7th ed. Boston: Cengage Learning.

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

———. 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.

Yang, Y. 2022. A correlated random effects approach to the estimation of models with multiple fixed effects. *Economics Letters* 213: 110408. http://dx.doi.org/10.1016/j.econlet.2022.110408.

Appendix

Empirical ilustration: extended results

In Section 4, I presented the empirical application of the `cre` command using the `nlswork.dta` dataset, using default standard errors. In this appendix, I provide additional results using robust and clustered standard errors, as well as results for two-way fixed effects models.

As it can be seen in Table 6, once robust standard errors are used, the results for the linear model are almost identical to the benchmark FE model, in terms of coefficients and standard errors (proxied by t-statistics).

Table 6: Comparison of Linear Models: Robust/Clustered Standard Errors

				CRE	
	xtreg FE	xtreg RE	xtreg CRE	xtreg RE	regress
main					
age	0.042	0.047	0.042	0.042	0.042
	(9.74)	(11.47)	(9.74)	(9.74)	(9.74)
c.age#c.age	-0.001	-0.001	-0.001	-0.001	-0.001
	(-6.89)	(-8.57)	(-6.89)	(-6.89)	(-6.89)
tenure	0.041	0.048	0.041	0.041	0.041
	(18.99)	(23.19)	(18.99)	(18.99)	(18.99)
c.tenure#c.tenure	-0.001	-0.002	-0.001	-0.001	-0.001
	(-10.23)	(-11.94)	(-10.23)	(-10.23)	(-10.23)
1.white	0.000	0.090	0.081	0.081	0.093
	(.)	(7.79)	(7.27)	(7.27)	(8.02)
1.south	-0.071	-0.121	-0.071	-0.071	-0.071
	(-4.08)	(-11.56)	(-4.08)	(-4.08)	(-4.08)
m1_age				0.051	0.076
				(4.38)	(5.95)
m1_c_age_c_age				-0.001	-0.001
				(-4.47)	(-5.94)
m1_tenure				0.074	0.067
				(11.10)	(9.11)
m1_c_tenure_c_tenure				-0.003	-0.003
				(-6.09)	(-5.01)
m1__1_south				-0.105	-0.111
				(-5.14)	(-5.36)
_cons	0.835	0.715	-0.031	0.792	0.769
	(13.34)	(12.11)	(-0.21)	(12.51)	(12.16)
xt_means					
age			0.051		
			(4.38)		
c.age#c.age			-0.001		
			(-4.47)		
tenure			0.074		
			(11.10)		
c.tenure#c.tenure			-0.003		
			(-6.09)		
1.white			0.000		
			(.)		
1.south			-0.105		
			(-5.14)		
N	28093	28093	28093	28093	28093

t statistics in parentheses

In Table 7, I present the results for a two-way fixed effects model, using the `cre`, absorbing both `idcode` and `year` as fixed effects. As benchmark I use the `reghdfe` command for the linear model, and `ppmlhdfe` for the Poisson model. Three things to notice:

1. `cre` produces the same point estimates as `reghdfe`, showing the potential of the command.
2. `cre` produces very similar point estimates to `ppmlhdfe`. This is similar to what we observed in the main text.
3. Standard errors are different, and no correction was applied for `cre` other than providing `robust` standard errors.

Table 7: Comparison of Linear Models: Two-way Fixed Effects

	reghdfe	ppmlhdfe	CRE regress	CRE poisson
main				
age	0.067	0.069	0.067	0.067
	(5.59)	(3.24)	(4.08)	(2.38)
c.age#c.age	-0.001	-0.001	-0.001	-0.001
	(-13.11)	(-7.71)	(-10.07)	(-5.89)
tenure	0.041	0.037	0.041	0.037
	(24.07)	(13.66)	(17.40)	(9.96)
c.tenure#c.tenure	-0.001	-0.002	-0.001	-0.002
	(-13.27)	(-6.22)	(-9.37)	(-4.84)
1.white	0.000	0.000	0.092	0.099
	(.)	(.)	(16.58)	(11.97)
1.south	-0.070	-0.090	-0.070	-0.099
	(-5.33)	(-4.64)	(-4.14)	(-3.48)
m1_age			0.068	0.093
			(3.69)	(2.94)
m2_age			-0.131	-0.174
			(-3.88)	(-3.67)
m1_c_age_c_age			-0.001	-0.002
			(-7.27)	(-6.54)
m2_c_age_c_age			0.003	0.003
			(3.98)	(3.83)
m1_tenure			0.068	0.046
			(14.58)	(3.69)
m2_tenure			0.068	0.149
			(3.09)	(5.16)
m1_c_tenure_c_tenure			-0.003	-0.001
			(-8.04)	(-0.96)
m2_c_tenure_c_tenure			-0.010	-0.013
			(-3.09)	(-3.18)
m1___1_south			-0.108	-0.076
			(-6.04)	(-2.49)
m2___1_south			0.313	0.065
			(0.54)	(0.08)
_cons	0.506	0.581	0.440	0.455
	(1.52)	(0.90)	(0.96)	(0.55)
N	27541	27541	28093	28093

t statistics in parentheses

Two-way Fixed Effects

In this appendix, I provide results for the two-way fixed effects model, using the `cre` command with two variables in the `abs()` option. The results are similar to those presented in the main text.

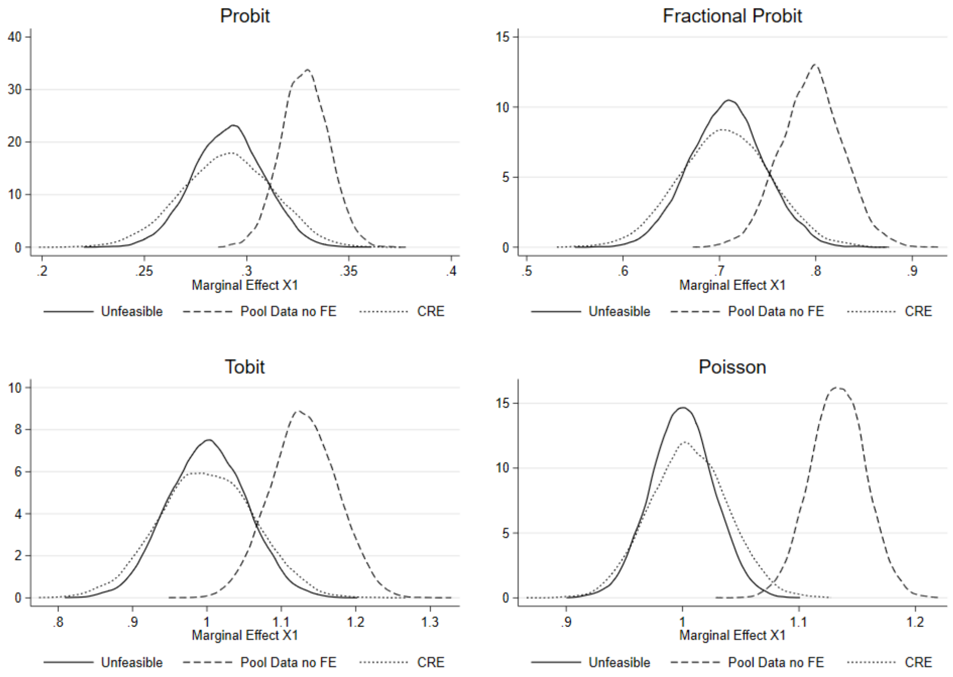

Figure 2: Estimated APE/Coefficient densities for non-linear models: Two-way fixed effects

Table 8: Bias and MAE for the estimated APE/Coefficients for non-linear models TWFE: Coefficients for `x1`

	Probit	FProbit	Tobit	Poisson
True:Bias	-0.000	0.000	0.000	-0.000
True:MAE	0.014	0.031	0.042	0.022
Pool:Bias	0.037	0.088	0.130	0.134
Pool:MAE	0.037	0.088	0.130	0.134
CRE:Bias	-0.001	-0.002	-0.001	0.006
CRE:MAE	0.018	0.038	0.051	0.027
N	10000	10000	10000	10000

About the authors

Fernando Rios-Avila is an applied econometrician with passion for econometrics and programming. 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.