Nonlinear Correlated Random Effects Models with Endogeneity and Unbalanced Panels

Michael D. Bates*, Leslie E. Papke†, Jeffrey M. Wooldridge ‡

August 18, 2023

Abstract

We present simple procedures for estimating nonlinear panel data models in the presence of unobserved heterogeneity and possible endogeneity with respect to time-varying unobservables. We combine a correlated random effects approach with a control function approach while accounting for missing time periods for some units. We examine the performance of the approach in comparisons with standard estimators using Monte Carlo simulation. We apply the methods to estimate the effects of school spending on student pass rates on a standardized math exam. We find that a 10 percent increase in spending leads to an approximately two percentage point increase in math pass rates.

*Department of Economics, University of California at Riverside, Riverside, CA 92521 (email: mbates@ucr.edu).
†Department of Economics, Michigan State University, East Lansing, MI 48824-1038 (email: papke@msu.edu).
‡Department of Economics, Michigan State University, East Lansing, MI 48824-1038 (email: wooldri1@msu.edu).

1 Introduction

Unbalanced panel data – where some units do not have a complete set of observations in some time periods – are prevalent in empirical work. Researchers have documented unbalancedness stemming from both intermittent non-response and early attrition in panel surveys used in research in labor, public, and development economics. Aughinbaugh (2004) and Falaris and Peters (1998) note yearly non-response rates of three to four percent in the 1979 National Longitudinal Survey of Youth. Fitzgerald (2011) notes that only one-third of the children from the 1968 round of the Michigan Panel Survey of Income Dynamics remain in the data by 2007. Alderman et al. (1999) find significant annual attrition rates in household surveys from seven different developing countries that range from two to 20 percent.

As is well known, unbalanced panel data in a linear model context can be handled by fixed effects estimation provided the selection is based on observed variables or unobserved, time-constant heterogeneity; see, for example, Wooldridge (2019). When explanatory variables are endogenous with respect to time-varying unobservables, Joshi and Wooldridge (2019) show how linear fixed effects and control function methods can be applied to unbalanced panels for estimation and specification testing. But as pointed out in Wooldridge (2019), unbalanced panels cause significantly more difficulties in nonlinear panel data models. Wooldridge (2019) proposes a correlated random effects (CRE) approach to allow the heterogeneity to be correlated with time-constant functions of selection indicators for general nonlinear panel data models. The CRE approach allows explanatory variables also to be correlated with time-constant unobservables – so-called "unobserved heterogeneity." In some cases, however, one might be concerned that a key explanatory variable is correlated with unobserved time-varying variables. In the panel data literature, this is called a failure of the "strict exogeneity" assumption. Failure of strict exogeneity is often due to omitted time-varying variables, or feedback from shocks to future outcomes of the explanatory variables. Simultaneity and (time-varying) measurement error can also cause failure of strict exogeneity.

In this paper, we extend Wooldridge (2019) to the estimation of nonlinear models in the presence of unbalanced panel data when the covariates may be endogenous with respect to time-varying unobservables as well as time-constant heterogeneity. Our work can also be viewed as extending Joshi and Wooldridge (2019), who consider linear models with unbalanced panels, to a nonlinear context. When the outcome variable is limited in some way – such as being binary or a fractional response

– nonlinearity can be important along with endogeneity as a function of idiosyncratic shocks, as demonstrated in Papke and Wooldridge (2008) with balanced panel data. When explanatory variables are allowed to be endogenous with respect to idiosyncratic shocks, we require time-varying instrumental variables that are exogenous with respect to those shocks. As in Wooldridge (2019) and Joshi and Wooldridge (2019), our key assumption is that the missingness of data is not correlated with idiosyncratic shocks.[1] To summarize: we extend Wooldridge (2019) by allowing endogeneity with respect to time varying shocks, we extend Joshi and Wooldridge (2019) by allowing a nonlinear response function, and we extend Papke and Wooldridge (2008) to allow unbalanced panels.

The approach we take is to combine the CRE approach for unbalanced panels – which we refer to as "CREU" for shorthand – with the control function approach when strict exogeneity fails. In other words, we combine the best features from these two approaches to allow nonlinear models with unbalanced panels and endogenous explanatory variables. We consider different strategies for allowing correlation between unobserved heterogeneity and the selection indicators. We are specifically interested in comparing the CREU approach for nonlinear fractional response models, implemented using pooled quasi-maximum likelihood estimation (QMLE), with standard fixed effects estimation strategies for linear unobserved effects models. We find that the CREU approaches perform comparably to the CRE approach that ignores the unbalanced nature of the panel and linear fixed effects estimation in uncovering average partial effects (APEs), but the CREU approach provides efficiency gains in estimating APEs. Further, because the fractional response model is nonlinear, we are able to study partial effects at different values of the key explanatory variables.

We illustrate our approach with an empirical application in the economics of education literature: estimating the effects of school spending on school pass rates of fourth graders on the Michigan state mathematics standardized exam. Papke (2005) used unbalanced school-level data and linear models estimated by fixed effects and instrumental variables. Given the bounded nature of the pass rate, a linear model may not be the best way to estimate average effects or effects at different points in the spending distribution. Papke and Wooldridge (2008) showed how to adapt fractional

[1] While this assumption can be violated, all standard estimation methods – even in linear models – rule out the possibility that selection is correlated with shocks. As with the usual fixed effects and fixed effects IV estimators of linear models, we allow selection to be correlated with unobserved heterogeneity. Allowing selection to be correlated with the shocks is considerably more difficult. Wooldridge (1995) and Semykina and Wooldridge (2010) show how to allow this in the context of linear models. We leave to future research the possibility of extending selection methods to fractional responses.

response models to a panel data setting, but they assumed a balanced panel and applied a combined correlated random effects/control function approach to balanced district-level data. We reexamine the results from Papke (2005) while accounting for the unbalancedness of the school-level data and the bounded nature of pass rates. While we find some evidence of correlation between school spending and unbalancedness, our results largely uphold the evidence presented in Papke (2005): a 10% increase in spending leads to an approximately two percentage point increase in math pass rates, though our inference is sensitive to specification and level of clustering. These results are also similar to those in Papke (2008) using the balanced district-level data.

We organize the remainder of the paper as follows. In Section 2 we present the model and estimation methods, considering first the case where all explanatory variables are exogenous with respect to the time-varying unobservables. We then derive a method that combines an extended version of the Mundlak (1978) device and a control function method to allow some explanatory variables to be correlated with time-varying unobservables. We present our simulation evidence in Section 3. In our application in Section 4, we demonstrate estimation with and without requiring school spending to be strictly exogenous with respect to idiosyncratic shocks in determining the effects of spending on fourth-grade math pass rates. Section 5 concludes.

2 Model and Estimation

We begin with a population from which we draw a random sample of N cross-sectional units. For each random draw i from the cross section, there are potentially T observations across time, $t = 1, ..., T$, containing an outcome, y_{it}, and a vector of observed covariates, $\mathbf{x}_{it}$. Except for specific functional form and distributional assumptions, the approach proposed here applies to nonlinear models in general, but we focus on the case where y_{it} is a fractional response that may take values at the endpoints in $[0, 1]$. Along with the $\mathbf{x}_{it}$, we expect unobserved heterogeneity, c_i, to play a role in determining y_{it}. In non-experimental settings, it is likely that c_i is correlated with at least some components of $\mathbf{x}_{it}$. We use a correlated random effects strategy to allow all elements of $\mathbf{x}_{it}$ that vary somewhat across i and t to be correlated with c_i. When one or more elements of $\mathbf{x}_{it}$ is correlated with underlying idiosyncratic shocks to y_{it} – to be made precise shortly – we will assume the availability of some time-varying instrumental variables. Then, $\mathbf{z}_{it}$ will denote the vector of all variables strictly exogenous with respect to shocks. We still allow all elements of $\mathbf{z}_{it}$ to be correlated with c_i.

To account for the unbalanced nature of the panel data, we introduce a selection indicator – also known as a "complete cases" indicator, s_{it}. This indicator is one if we observe the outcome, all covariates, and any instrumental variables for unit i in time t. It is important in what follows that only the complete cases are used, as using incomplete cases generally requires more assumptions and more complications. The default of estimation methods in econometrics packages is to use a data point only if all necessary variables are observed, and that is what the definition of s_{it} captures. Therefore, $s_{it} = 1$ means we use observation (i,t) in the estimation and $s_{it} = 0$ means we do not. The series of selection indicators for unit i is $\{s_{i1}, ..., s_{iT}\}$.

2.1 Strict Exogeneity

We begin with the case where the explanatory variables are strictly exogenous conditional on the heterogeneity. The population model, written for a random draw i, is

$$E(y_{it}|\mathbf{x}_i, c_i) = E(y_{it}|\mathbf{x}_{it}, c_i) = \Phi(\mathbf{x}_{it}\boldsymbol{\beta} + c_i), t = 1, ..., T, \tag{1}$$

where $\mathbf{x}_i = (\mathbf{x}_{i1}, ..., \mathbf{x}_{iT})$ is the entire history of the covariates and $\Phi(\cdot)$ is the standard normal cumulative distribution function. Our use of Φ rather than some other cumulative distribution function leads to simple procedures in the presence of unobserved heterogeneity and easy calculation of average partial effects. It is also convenient when we have endogenous explanatory variables because mixing two normal distributions still gives a normal distribution.

To account for sample selection, let $\mathbf{s}_i = (s_{i1}, ..., s_{iT})$ be the entire history of selection. We assume that, conditional on $\mathbf{x}_i$ and the unobserved heterogeneity, selection is strictly exogenous in the following sense:

$$E(y_{it}|\mathbf{x}_i, \mathbf{s}_i, c_i) = E(y_{it}|\mathbf{x}_i, c_i),\ \ t = 1, ..., T \tag{2}$$

This assumption allows selection to be arbitrarily correlated with both the explanatory variables and unobserved heterogeneity – because we are conditioning on them – but rules out correlation between selection and unobserved idiosyncratic fluctuations in the outcome.

Following Wooldridge (2019), we use a correlated random effects approach to specify a model for the following conditional distribution:

$$D(c_i|\{(s_{it}\mathbf{x}_{it}, s_{it}) : t = 1, ..., T\}), \tag{3}$$

where multiplying the covariates by the selection indicator reflects our usage of complete cases only. Generally, Wooldridge (2019) suggests modeling (3) as fairly simple time-constant functions, say $\mathbf{w}_i$, of $\{(s_{it}\mathbf{x}_{it}, s_{it}) : t = 1, ..., T\}$ that effectively act as sufficient statistics in the relationship between the covariates and selection. It is natural to extend the Mundlak (1978) device to the unbalanced case by using the time averages

$$\bar{\mathbf{x}}_i = T_i^{-1}\sum_{t=1}^{T_i}\mathbf{x}_{it},$$

where $T_i = \sum_{t=1}^{T} s_{it}$ is the number of complete cases for unit i. (If $T_i = 0$, then there are no complete time periods for unit i, and such units are not used in the estimation). To handle correlation between c_i and selection, we use a flexible mean specification where the intercept and slopes can depend on the number of complete cases, as given by the indicators $1\left[T_i = r\right]$, which are one if and only if unit i has r complete cases. Then,

$$E(c_i|\mathbf{w}_i) = \sum_{r=1}^{T}\psi_r 1[T_i = r] + \sum_{r=1}^{T}\left(1[T_i = r]\cdot\bar{\mathbf{x}}_i\right)\boldsymbol{\xi}_r. \tag{4}$$

If we also assume $D(c_i|\mathbf{w}_i)$ is a normal distribution, then we have

$$E(y_{it}|\mathbf{x}_{it}, \mathbf{w}_i, s_{it} = 1) = \Phi\left(\frac{\mathbf{x}_{it}\boldsymbol{\beta} + \sum_{r=1}^{T}\psi_r 1[T_i = r] + \sum_{r=1}^{T}1[T_i = r]\cdot\bar{\mathbf{x}}_i\boldsymbol{\xi}_r}{\{1 + Var(c_i|\mathbf{w}_i)\}^{\frac{1}{2}}}\right), \tag{5}$$

because a mixture of independent normal distributions is normal. Equation (5) extends Papke and Wooldridge (2008), who assumed $Var(c_i|\mathbf{w}_i)$ is constant, to the case of unbalanced panels. Rather than assume $Var(c_i|\mathbf{w}_i)$ is constant, it is natural to allow, at a minimum, the variance of c_i to vary with the number of complete cases. A simple way to do this is

$$Var(c_i|\mathbf{w}_i) = \exp\left(\tau + \sum_{r=1}^{T-1}1[T_i = r]\boldsymbol{\omega}_r\right), \tag{6}$$

where $\exp(\tau)$ is the variance for the complete-cases base group ($T_i = T$) and each ω_r captures the deviation from the base group.

Combining (5) and (6), we have

$$E(y_{it}|\mathbf{x}_{it}, \mathbf{w}_i, s_{it}=1) = \Phi\left(\frac{\mathbf{x}_{it}\boldsymbol{\beta} + \sum_{r=1}^{T}\psi_r 1[T_i=r] + \sum_{r=1}^{T} 1[T_i=r]\cdot\bar{\mathbf{x}}_i\boldsymbol{\xi}_r}{\left\{1+\exp\left(\tau+\sum_{r=1}^{T-1} 1[T_i=r]\boldsymbol{\omega}_r\right)\right\}^{\frac{1}{2}}}\right). \tag{7}$$

For all $r \geq 2$ these scaled coefficients are identified as long as there is some time variation in all elements of $\mathbf{x}_{it}$ and no perfect collinearity among the elements of $\mathbf{x}_{it}$.

Given the expression (7) for the conditional mean, we can follow Papke and Wooldridge (2008) in estimating the parameters using a pooled quasi-maximum likelihood approach with the log-likelihood being chosen to be that for the Bernoulli distribution. Given the functional form in (7), the pooled quasi-log-likelihood function is equivalent to that from a particular heteroskedastic probit model, where the heteroskedasticity function is $1+\exp\left(\tau+\sum_{r=1}^{T-1} 1[T_i=r]\boldsymbol{\omega}_r\right)$. As a practical matter, we can drop the "1+" term because we allow an intercept τ inside the exponential function. The resulting parameters in the "mean" function, $\mathbf{x}_{it}\boldsymbol{\beta}+\sum_{r=1}^{T}\psi_r 1[T_i=r]+\sum_{r=1}^{T} 1[T_i=r]\cdot\bar{\mathbf{x}}_i\boldsymbol{\xi}_r$, get rescaled, but this does not affect estimating the magnitudes of the effects. It is easy to use software, such as Stata, that has a command for estimating fractional response models with heteroskedasticity. In obtaining proper standard errors and inference, we obtain a cluster-robust variance-covariance matrix estimator that accounts for heteroskedasticity, serial correlation, and the fact that in the fractional response case, the variance $Var(y_{it}|\mathbf{x}_{it}, \mathbf{w}_i, s_{it}=1)$ does not have the same form as when y_{it} is a binary variable. Our approach applies immediately to the binary response case, as its mean is the same as the response probability, which we take here to be probit.

Estimating the average partial effects – the quantities typically of interest – requires some care if data are missing on the $\mathbf{x}_{it}$. At a minimum, we can plug in reasonable values of the covariates and average across the functions of $(\mathbf{x}_i, \mathbf{s}_i)$ that act as proxies for the heterogeneity. This leads to

$$\widehat{APE}_j(\mathbf{x}_t) = \hat{\beta}_j\left[N^{-1}\sum_{i=1}^{N}\phi\left(\frac{\mathbf{x}_t\hat{\boldsymbol{\beta}} + \sum_{r=1}^{T}\psi_r 1[T_i=r] + \sum_{r=1}^{T}\left(1[T_i=r]\cdot\bar{\mathbf{x}}_i\right)\hat{\boldsymbol{\xi}}_r}{1+\exp\left(\hat{\tau}+\sum_{r=1}^{T-1} 1[T_i=r]\hat{\boldsymbol{\omega}}_r\right)}\right)\right].$$

It is harder to obtain an effect averaged across the distribution of $\mathbf{x}_{it}$ because data may be missing as a systematic function of $\mathbf{x}_{it}$. The simplest approach is to average the APEs across the selected observations:

$$\widehat{APE}_j = \hat{\beta}_j \left[N^{-1} \sum_{i=1}^{N} T_i^{-1} \sum_{t=1}^{T} s_{it} \phi \left(\frac{\mathbf{x}_{it}\hat{\boldsymbol{\beta}} + \sum_{r=1}^{T} \psi_r 1[T_i = r] + \sum_{r=1}^{T} \left(1[T_i = r] \cdot \bar{\mathbf{x}}_i\right) \hat{\boldsymbol{\xi}}_r}{1 + \exp\left(\hat{\tau} + \sum_{r=1}^{T-1} 1[T_i = r]\hat{\boldsymbol{\omega}}_r\right)} \right) \right].$$

As an extension, $\bar{\mathbf{x}}_i$ can be added to the variance function along with interactions between the dummies $1[T_i = r]$ and $\bar{\mathbf{x}}_i$.

2.2 Endogenous Explanatory Variables

In many applications, researchers are hesitant to assume strict exogeneity of covariates. In our application, we worry that deviations in school spending may be linked with unobserved fluctuations in student performance. This may come from unobserved demands of cohorts or accountability pressure, as depicted in Chiang (2009).

Here we present a straightforward approach to handle endogeneity of an explanatory variable y_{it2} in nonlinear panel data models in the presence of unobserved heterogeneity and panel imbalance. We first assume the presence of instrumental variables $\mathbf{z}_{it2}$ that are both relevant to y_{it2} and otherwise exogenous. We more precisely state these assumptions below. We allow there to be additional exogenous covariates denoted as $\mathbf{z}_{it2}$ with $\mathbf{z}_{it} = (\mathbf{z}_{it1}, \mathbf{z}_{it2})$ denoting the complete vector of pertinent exogenous variables. We follow Papke and Wooldridge (2008) in modeling the conditional mean as

$$\begin{aligned} E\left(y_{it1}|y_{it2}, \mathbf{z}_i, \mathbf{s}_i, c_{i1}, v_{it1}\right) =& E\left(y_{it1}|y_{it2}, \mathbf{z}_i, c_{i1}, v_{it1}\right) = E\left(y_{it1}|y_{it2}, \mathbf{z}_{it1}, c_{i1}, v_{it1}\right) \\ =& \Phi\left(\beta_1 y_{it2} + \mathbf{z}_{it1}\delta_1 + c_{i1} + v_{it1}\right), \end{aligned} \tag{8}$$

where c_{i1} is time-invariant unobserved heterogeneity across units and v_{it1} is an omitted factor that varies over both units and time. Once we have already conditioned on the explanatory variables and the source of endogeneity, the conditional mean is unaffected by conditioning on $\mathbf{z}_{it2}$. Thus, $\mathbf{z}_{it2}$ is excluded from equation (8). Additionally, note that we continue to assume that selection is ignorable conditional on the observed variables and the unobservables, c_{i1} and v_{it1}.

In equation (8) the variable y_{it2} may now be endogenous with respect to v_{it1} as well as with respect to c_{i1}. We handle the latter endogeneity similarly to the strict exogeneity case by using a correlated random effects approach to specify a model for c_{i1}, following the unbalanced case in Wooldridge (2019). In particular, to account for the unbalanced panel, we allow the coefficients on the time averages to change with the number of time periods observed for each i, in addition to allowing separate intercepts for each T_i:

$$c_{i1} = \sum_{r=1}^{T} \psi_{r1} 1[T_i = r] + \sum_{r=1}^{T} \left(1[T_i = r] \cdot \bar{\mathbf{z}}_i\right) \boldsymbol{\xi}_{r1} + a_{i1},\ \ a_{i1}|\mathbf{z}_i \sim Normal(0, \sigma_{a1}^2), \tag{9}$$

where $\bar{\mathbf{z}}_i = T_i^{-1} \sum_{r=1}^{T_i} s_{it}\mathbf{z}_{it}$ is the time average over the complete cases and a_{i1} is an error term that we assume to be independent of $(\mathbf{z}_i, \mathbf{s}_i)$. As before, conditional normality leads to a relatively straightforward analysis.

Substituting equation (9) into equation (8) gives

$$\begin{aligned} &E(y_{it1}|y_{it2}, \mathbf{z}_i, r_{it1}, s_{it} = 1) = \\ &\qquad \Phi\left(\beta_1 y_{it2} + \mathbf{z}_{it1}\delta_1 + \sum_{r=1}^{T} \psi_{r1} 1[T_i = r] + \sum_{r=1}^{T} \left(1[T_i = r] \cdot \bar{\mathbf{z}}_i\right) \boldsymbol{\xi}_{r1} + r_{it1}\right), \end{aligned} \tag{10}$$

where $r_{it1} = a_{i1} + v_{it1}$ is a composite error term. Researchers may also wish to follow Lin and Wooldridge (2019) by including $\bar{y}_{2i} = T^{-1} \sum_{r=1}^{T} y_{2ir}$ to clearly separate the endogeneity due to c_{i1} from the endogeneity due to v_{it1}. We omit it here to coincide with previous approaches in our application.

Secondly, we must deal with the endogeneity of y_{it2}. Following Mundlak (1978), we linearly model y_{it2} as a function of the exogenous explanatory variables, excluded instruments, and their time averages. As selection may be correlated with y_{it2}, we generally include indicators for the number of time-observations and interactions with time averages here as well. We present this first-stage equation below:

$$y_{it2} = \mathbf{z}_{it}\pi_2 + \sum_{r=1}^{T} \psi_{r2} 1[T_i = r] + \sum_{r=1}^{T} 1[T_i = r] \cdot \bar{\mathbf{z}}_i \boldsymbol{\xi}_{r2} + v_{it2}, \tag{11}$$

where v_{it2} represents time-varying unobserved elements of y_{it2} and we have included a full set of dummies and omitted an intercept. In equation (11) the endogeneity in y_{it2} is due to the correlation between r_{it1} and v_{it2}. Following Rivers and Vuong (1988) and Papke and Wooldridge (2008), we model r_{it1} as linear in v_{it2} and conditionally normal:

$$r_{it1} = \eta_1 v_{it2} + e_{it1},\ e_{it1}|\,(\mathbf{z}_i, \mathbf{s}_i, v_{it2}) \sim Normal(0, \sigma_e^2).$$

Note that e_{it1} is also independent of y_{it2}. Together with (11), the normality assumption effectively rules out discreteness in y_{it2}, and so our approach is applicable to cases where y_{it2} has a conditional distribution well approximated by a normal distribution. Sometimes one may need to use a transformation to make normality more plausible. In our empirical application, we use the logarithm of school spending. As in the balanced case in Papke and Wooldridge (2008), given the assumptions, we can replace $r_{it1} = \eta_1 v_{it2} + e_{it1}$ and then integrate out e_{it1} using the properties of the normal distribution. The resulting coefficients are scaled by $(1+\sigma_e^2)^{-\frac{1}{2}}$:

$$\begin{aligned} E(y_{it1}|y_{it2}, &\mathbf{z}_i, v_{it2}, s_{it}=1) = \\ &\Phi\left(\beta_{1e} y_{it2} + \mathbf{z}_{it1}\delta_1 + \sum_{r=1}^{T} \psi_{r1e} 1[T_i = r] + \sum_{r=1}^{T} (1[T_i = r] \cdot \bar{\mathbf{z}}_i)\, \boldsymbol{\xi}_{r1e} + \eta_{1e} v_{it2}\right), \end{aligned} \tag{12}$$

where subscript e denotes the scaling of the coefficients. The average partial effects – where we necessarily average over the selected sample – depend on the scaled coefficients, as discussed in Papke and Wooldridge (2008) in the balanced panel case. Therefore, in what follows, we drop the e subscript from the parameters. Equation (12) constitutes the primary estimating equation, and it was derived under several normality assumptions. It probably would do little harm to use the logistic function in (12) in place of Φ, but the logit functional form does not follow from the primitive assumptions.

We follow a two-step procedure to estimate equation (12). In the first step, we estimate (11) by regressing our endogenous explanatory variable, y_{it2}, on the exogenous variables, $\mathbf{z}_{it}$, that include the instruments and time indicators, indicators for the number of time observations per unit, and interactions between those indicators and time averages of the exogenous variables. We save the residuals from that regression, $\hat{v}_{it2}$, for the complete cases. In step two, we substitute these residuals for v_{it2} and estimate equation (12) using the complete cases and pooled probit QMLE of y_{iy1} on $\mathbf{z}_{it1}$; the indicators, $1[T_i = r]$; all interactions, $1[T_i = r] \cdot \bar{\mathbf{z}}_i$; and the first-stage residuals, $\hat{v}_{it2}$.

Due to the estimation of $\hat{v}_{it2}$ in the first step, the standard errors in the second stage should be adjusted. Bootstrapping the entire procedure by resampling individual units with replacement is one way to account for the first stage estimation. We adopt this approach in our empirical application.

We are mainly interested in the APEs of the endogenous explanatory variable, y_{t2}. We can compute APEs at different values of y_{t2} and $\mathbf{z}_{t1}$ using the average structural function (ASF), as defined by Blundell and Powell (2003). This is easily obtained from the estimating equation (12) as

$$ASF(y_{t2}, \mathbf{z}_{t1}) = \\ E_{(\overline{\mathbf{z}}_i, T_i, v_{it2})}\left[\Phi\left(\beta_e y_{t2} + \mathbf{z}_{t1}\boldsymbol{\delta}_e + \sum_{r=1}^{T}\psi_{re1}1[T_i = r] + \sum_{r=1}^{T}1[T_i = r]\cdot\overline{\mathbf{z}}_i\boldsymbol{\xi}_{re1} + \eta_e v_{it2}\right)\right], \tag{13}$$

where the notation means that we average across the joint distribution of $(\overline{\mathbf{z}}_i, T_i, v_{it2})$ with y_{t2} and $\mathbf{z}_{t1}$ fixed arguments of the ASF. The ASF can be estimated by replacing the expectation with a sample average and replacing the unknown parameters with their consistent estimators (including replacing v_{it2} with $\hat{v}_{it2}$):

$$\widehat{ASF}(y_{t2}, \mathbf{z}_{t1}) = N^{-1}\sum_{i=1}^{N}\left[\Phi\left(\hat{\beta}_e y_{t2} + \mathbf{z}_{t1}\hat{\boldsymbol{\delta}}_e + \sum_{r=1}^{T}\hat{\psi}_{re1}1[T_i = r] + \sum_{r=1}^{T}1[T_i = r]\cdot\overline{\mathbf{z}}_i\hat{\boldsymbol{\xi}}_{re1} + \hat{\eta}_e\hat{v}_{it2}\right)\right],$$

One can then compute the derivative with respect to y_{t2} to obtain APEs that can vary by $(y_{t2}, \mathbf{z}_{t1})$.

Often, we want to estimate APEs that measure the effect, on average, of y_{it2} on the mean outcome of y_{it1}. For a given t, this APE is consistently estimated as

$$\hat{\beta}_e\cdot\left[N^{-1}\sum_{i=1}^{N}\phi\left(\hat{\beta}_e y_{it2} + \mathbf{z}_{it1}\hat{\boldsymbol{\delta}}_e + \sum_{r=1}^{T}\hat{\psi}_{re1}1[T_i = r] + \sum_{r=1}^{T}1[T_i = r]\cdot\overline{\mathbf{z}}_i\hat{\boldsymbol{\xi}}_{re1} + \hat{\eta}_e\hat{v}_{it2}\right)\right], \tag{14}$$

where $\phi(\cdot)$ is the standard normal probability density function. Notice how all arguments are averaged out, including y_{it2} and $\mathbf{z}_{it1}$. To obtain a single APE, (14) can be averaged across the T time periods, and this average is what we report in the simulations and empirical application. Again, applying a clustered bootstrap is a convenient way to obtain valid standard errors.

To summarize, the approach we describe uses a Chamberlain (1980) and Mundlak (1978) approach to handle the time-invariant heterogeneity and a control function to handle endogeneity. We allow the unbalancedness of the panel to correlate with the unobserved heterogeneity by including indicators for the number of time-observations a cross-sectional unit appears in the data. We do not allow the data missingness to depend on the idiosyncratic error. Testing directly for selection based on the idiosyncratic fluctuations typically requires a second instrument and exclusion restriction, which we view as beyond the scope of this paper. Researchers may obtain indirect evidence of the plausibility of this assumption by testing whether we may predict missing values by including an indicator for whether the lead or lag observation is present. We demonstrate this approach when we apply these methods to study the effect of school spending on academic achievement.

3 Simulation Evidence

We conduct a simulation study to investigate the performance of our approaches that handle the panel unbalancedness against standard estimators that do not. In particular, we are interested in the bias that correlated unbalancedness may produce in estimated APEs and the relative efficiency of the estimators. For comparison, we first use POLS and FE as approximations of the APE from linear models. We then consider standard nonlinear approaches; namely, pooled fractional response probit QMLE (PFR), and PFR where we model the correlated random effects using the time averages of covariates (CRE).

In using CRE in the linear case, once $\bar{\mathbf{x}}_i$ has been included, interacting the indicators, $1\left[T_i = r\right]$, with $\bar{\mathbf{x}}_i$ does not change the estimates; they are the usual fixed effects estimates. This follows from Wooldridge (2019). To handle the imbalance of the panel in the linear model, we mimic fixed effects estimation using time averages of all explanatory variables and add time-observation indicators interacted with time averages of the explanatory variables (FEU). In nonlinear formulations, we add to CRE time-observations indicators (CREU). We next add to CREU interactions between time-observation indicators and covariate time averages (CREU1). We then add to CREU1 interactions between time-observation indicators and the covariates themselves (CREU2). Finally, we also include triple interactions between time-observation indicators, the covariates, and their time averages (CREU3), which may provide robustness to other forms of correlation between panel unbalance and the covariates.

3.1 Data Generating Process

We generate the data using a slight generalization of our single-stage model of interest. We consider two cases ($g = 0, 1$) and generate the outcome, y, according to the following:

$$y_{it} = \Phi[\alpha + (\beta_1 + u_{ig})x_{it1} + \beta_2 x_{it2} + c_i + v_{it}],\ v_{it} \sim Normal(0, 0.2), \tag{15}$$

where Φ is the standard normal CDF and $\beta_1 = \beta_2 = 1$. We first draw school-year variables $x_{it1} \sim Normal(0,\ 0.2)$ and $x_{it2} \sim Binomial(1,\ 0.3)$ over 500 "schools" across 5 "years," and generate time averages of these variables by school building. The standard deviation of x_{it1} is set to approximate the standard deviation of our spending variable in the empirical application. The generalization in this simulation is that unobserved school heterogeneity takes the form of fixed effects, c_i, and random coefficients, u_{ig}. Because we assume in estimation that the coefficient on x is constant we are adding an additional term that is a function of x to the error term inside the probit. This has the effect of misspecifying the functional form.

The two cases differ depending on the construction of the unobserved heterogeneity and which form of heterogeneity is correlated with selection. In both cases, we generate the unobserved fixed effects according to the following equation:

$$c_i = \sqrt{T}\bar{x}_{i1}\gamma_1 + \eta_i,\ \eta_i \sim Normal(0, 0.14). \tag{16}$$

However, in the first case, $g = 0$, the school-level, random slope of x_1, u_{i0}, is defined as the time average of an independently distributed normal random variable, and thus, is not correlated with x_1 and selection into the panel. Consequently, it is not a far departure from the standard model introduced above. In the second case, $g = 1$, we extend the data-generating process to include a correlated random coefficient on x_1. This random slope, u_{i1}, is correlated with x_1 and selection into the panel. Specifically, the random slopes take the following form:

$$u_{ig} = \begin{cases} u_{i0} = T^{-1}\sum_{t=1}^{T} e_{it0},\ e_{it0} \sim Normal(0, 0.14) \text{ in simulation one,} \\ u_{i1} = \sqrt{T} \times \bar{x}_{i1}\gamma_2 + \gamma_3 c_i + e_{i1},\ e_{i1} \sim Normal(0,\ 0.14) \text{ in simulation two,} \end{cases} \tag{17}$$

where T represents the five possible time-observations, γ_1 and γ_2 are each set to 0.7, and γ_3 is set to 0.2, and η_i , e_{i0}, and e_{i1} are each drawn from independent, mean-zero, normal distributions.

We model selection depending on the unobserved effect, c, in simulation one and on the unobserved correlated random slope, u_1, in simulation two. In both cases, the selection of each time-observation is drawn from a binomial distribution with probability p_{ig} defined below.

$$p_{ig} = \begin{cases} \Phi(a_{it} + c_i) \text{ in simulation one,} \\ \Phi(a_{it} + u_{i1})\text{in simulation 2,} \end{cases} \tag{18}$$

where a_{it} is an independent normal distributed random variable with a mean of 0.75 and a standard deviation of 0.2.

3.2 Simulation Results

We present the resulting correlations from simulation one on the left and simulation two on the right of Table 1. In the first case, only the unobserved fixed heterogeneity is correlated with time-observation selection and with x_1. The resulting average number of time-observations across the 500 replications is 3.83, with a correlation of 0.299 between the number of time-observations and the unobserved fixed effect. In contrast, the correlation between the number of time-observations and the random slope is 0.001. The correlation between x_1 and c is 0.315, while the correlation between x_1 and u is 0.0006.

Due to the positive correlation between x_1 and c, we may expect POLS and PFR to exhibit an upward bias for β_1. Indeed, we see exactly this in Table 2. The first row of Table 2 provides the "true" APEs of x_1 when averaged over the population (as if the panel were balanced), the sample (where the number of time-observations is non-randomly unbalanced), and then disaggregated by each number of time-observations that the schools are present in the data. Over the population, the APE of x_1 is 0.2979, and averaged over the sample, the APE of x_1 is 0.2953. Both POLS and PFR overstate this effect by 0.087. The bias is easily statistically significant, as the standard deviations of the POLS and PFR estimates over the 500 replications are 0.0111 and 0.0102 respectively, .

Beyond POLS and PFR, all estimated APEs are quite close to the true APE, and none are more than a third of a standard deviation of the estimated APEs away from the true APE over the population. Still, the FE estimated APE is the furthest from the truth, with an estimated APE of 0.2939, 1.3 percent lower than the true effect. Adding time-observation indicators to the time-means in FE estimation in FEU increases the estimated APE to 0.2947, 1.1 percent lower than the true APE.

Even without accounting for the unbalancedness of the panel, with an estimated APE of 0.2947, the CRE estimates are remarkably close to those estimated by FEU. Neither adding indicators for the number of time-observations in CREU nor including time-observation indicators interacted with time averages in CREU1 alter the estimated APEs to the fourth decimal place. Further, the standard deviations of the APE estimates remain remarkably similar among CRE (0.0097), CREU (0.0098), and CREU1 (0.0098).

We examine additional specifications by adding interactions between covariates and time averages of covariates to the covariates used in CREU1 estimation (labeled CREU2), and second, by including the triple interactions between the covariates, their time averages, and the number of time-observations (labeled CREU3).[2] Including these additional interactions makes sense for a model that includes correlated random slopes, as shown in equation (15). Here, we model the heterogeneous school-level slopes by interacting the time averages of covariates with each covariate just as we model the fixed unobserved heterogeneity by inserting the time averages additively. Incorporating the triple interaction between the covariates, the time averages, and indicators for the number of time-observations in CREU3 addresses the potential that selection of time-observations is related to random slopes.

Both estimators provide very similar APE estimates to those from CRE. The CREU2 approach yields an estimated APE of 0.2949, while CREU3 estimates the APE of x_1 at 0.2951. The estimates become slightly less precise as we add covariates—the standard deviation of the CREU3 APE estimates increases to 0.0092.

We provide both the average standard error as well as the standard deviation of estimates across repetitions to show how well the estimated standard errors reflect the precision of each estimator. The standard errors of most estimators perform well, with the ratio of average standard errors to Monte Carlo standard deviations ranging from 0.93 to 1.02.

We turn next to the case where the selection of time-observations depends on the random x_1 coefficient, which is u_1. The average number of time-observation across the 500 replications is 3.82.

[2]We reran the simulation generating using data-generating processes from two non-normal distributions to investigate the performance of the estimator when the distribution is misspecified. The description and results appear in the online appendix and reveal that the estimators perform well in both cases.

The resulting correlation between the number of time-observations and the unobserved fixed effect is 0.2065, and the correlation between the number of time-observations and the random slopes is 0.3300. The correlation between x_1 and c_0 is 0.3152, while the correlation between x_1 and u_1 is 0.3404.

The results of the simulation with selection based upon correlated random slopes appear in Table 3. The "true" APE of x_1 is 0.2902, when averaged over the population with the panel balanced. The correlation between selection and the heterogeneous slopes is apparent looking at the top row across the true APEs averaged across schools with one, two, three, four, or five time-observations appearing in the data. The relationship is monotonically positive with the APE among those with only one time-observation being 0.2523, whereas the APE among those with five time-observations is 0.3034. The true APE over the unbalanced sample is about 1 percent higher than the true APE over the population. This positive correlation may be expected in many contexts where those with favorable numbers may be more likely to report their data. In our application, schools that report their data more frequently tend to be higher-performing.

In the presence of correlation between the heterogeneous slopes and selection of time-observations, all estimators overstate the average effect of the endogenous regressor. All estimates are greater than the true APE among the unbalanced sample. POLS and PFR overstate this effect most. POLS estimates the APE to be 0.3823 (standard deviation 0.0101). PFR probit estimates the APE to be 0.3827 (standard deviation 0.0094). Again, the true value lies far outside the 95% confidence interval of both estimators.

FE and CRE estimates lie significantly closer to the true estimates at 0.2954 and 0.2945, respectively. The CRE estimates (with a standard deviation of 0.0083) are more precise than the FE estimates (with a standard deviation of 0.0094). This drop of almost 12% in the Monte Carlo standard deviation in moving from the linear model to the fractional response model is not startling but it also is not trivial. Even in cross-sectional settings with binary response one rarely sees huge improvements in precision in moving from a linear to a nonlinear model. Adding indicators for the number of time-observations moves the FE estimates closer to the true APEs, though not statistically significantly so. The FEU estimate of the APE of x_1 is 0.2941 (standard deviation 0.0104).

Regarding the nonlinear estimators, adding indicators for the number of time-observations in

CREU only marginally affects the estimates of the APE (0.2946) and its precision (standard deviation of 0.0083). Adding interactions between time averages and time-observation indicators marginally increases the estimate of the APE to 0.2947 (standard deviation of 0.0095), though again not statistically significant. Only adding the triple interactions between the covariates, their time averages, and time-observation indicators in CREU3 makes a somewhat larger impact. With an estimated APE of 0.2937, CREU3 again provides the estimates closest to the true APE, though it is less precise with a standard deviation across simulations of 0.0096. Still, all CREU estimates fall within 0.3 percent of the estimated APE using the standard CRE approach.

Across specifications, the standard errors perform similarly to the standard deviations across repetitions. The ratio of mean standard errors to the standard deviation of the APEs across replications is 0.89 to 1.1 in this second simulation. CREU provides the most conservative standard errors relative to the standard deviation of estimates, and the ratio for CREU3 indicates that the standard errors perhaps overstate the estimator's precision. The nonlinear approaches mostly appear to be more efficient than the approaches using a linear specification. The mean standard errors are approximately 10 percent smaller using one of the fractional response probit specifications as opposed to an analogous linear specification. The relative precision of the nonlinear estimators makes sense given the nonlinearity of the estimated effects. Panel C of Table 3 shows the estimated partial effects at the tenth, thirtieth, fiftieth, seventieth, and ninetieth decile of x_1 using the CREU1 approach. The estimated partial effect at the tenth percentile is 11 percent larger than the partial effect at the median and 31 percent larger than the estimated partial effect at the ninetieth percentile.

To summarize, under both formulations of selection of time-observations into the sample, all estimators that account for unobserved heterogeneity do comparably well in avoiding bias. Nonlinear estimators have the additional advantage of the ability to detect nonlinear effects, particularly at the tails of the support. As is often the case, the nonlinear estimators also have somewhat smaller standard errors even when allowing for arbitrary correlations across observations within individual units. In the next section, we apply the methods to study the effect of school spending on student achievement.

4 Empirical Application

In revisiting Papke (2005), we initially treat spending as strictly exogenous and apply single-stage estimation of both linear and nonlinear models. Then, following Papke (2005), Chaudhary (2009), and Roy (2011), we use the 1994 centralization of school financing that occurred in Michigan under Proposal A to provide plausibly exogenous variation in school expenditures.[3] We use this policy to apply instrumental variables to spending and demonstrate these methods with an endogenous regressor. As in Papke (2005), we conduct our analysis at the school-building level over the time period of 1993-1998 and measure spending as the log of average real expenditures over the current and previous year. Our data contain 7,242 building-year observations from the 1,771 elementary schools in the state over the five-year period when funding equalization was most dramatic. We focus on the effects of spending on *math4* – the fraction of fourth-grade students who pass the mathematics section of the Michigan Education Assessment Program (MEAP).

We note significant imbalance in the school-building-by-year panels. The bottom row of Table 4 shows 37 percent of schools are missing at least one of the five possible observations, and 23 percent are missing at least two observations. Further, there is significant variation in fourth-grade math pass rates, enrollment, student composition, and spending across schools that appear in the data for each number of years, suggesting that the imbalance may be consequential for estimating the APE of spending. In directly addressing this unbalancedness, we move beyond Papke and Wooldridge (2008), who conduct analysis at the district level due to this issue.

4.1 Treating spending as strictly exogenous

The population linear model estimated in Papke (2005) can be written as the following:

$$\begin{aligned} math4_{it} = {} & \theta_t + \beta_1 log(avgrexp_{it}) + \beta_2 lunch_{it} + \beta_3 lunch_{it}^2 \\ & + \beta_4 log(enroll_{it}) + \beta_5 log(enroll_{it})^2 + c_i + e_{it} \end{aligned} \tag{19}$$

We control for year indicators and quadratics of both the percent of free and reduced-price lunch students and the log of student enrollment. This model may be estimated using pooled ordinary least squares (POLS). However, the estimated coefficients would be inconsistent if the unobserved heterogeneity is correlated with any of the explanatory variables. Consequently, researchers may

[3] Papke (2005), Chaudhary (2009), and Roy (2011) provide fuller discussion of this school finance reform.

opt to use fixed effects (FE) estimation or equivalently include time averages of all explanatory variables. Further, the estimated coefficients may provide good approximations to the APEs when the actual model is nonlinear, but there is no general result that says so.

Since we do not observe all time-observations for each school building, let s_{it} represent an indicator for whether school i appears in the data in year t. Equation (20) represents this linear model in the presence of unbalancedness.

$$s_{it}math4_{it} = s_{it}\theta_t + s_{it}\mathbf{x}_{it}\boldsymbol{\beta_a} + s_{it}c_i + s_{it}e_{it}, \tag{20}$$

where $\mathbf{x}_{it}$ includes $log(avgrexp_{it})$, $lunch_{it}$, $lunch^2_{it}$, $log(enroll_{it})$, and $log(enroll_{it})^2$. In order for fixed effects estimation to be consistent in the presence of such unbalancedness, we must assume strict exogeneity of the covariates *and* selection, conditional on the unobserved heterogeneity. Put more formally,

$$E(u_{it}|\mathbf{x}_i,\ \mathbf{s}_i,\ c_i) = 0, \tag{21}$$

where $\mathbf{x}_i = (\mathbf{x}_{i1}, \mathbf{x}_{i2}, ..., \mathbf{x}_{iT})$ and $\mathbf{s}_i = (s_{i1}, s_{i2}, ..., s_{iT})$. As an example, an idiosycratic low pass rate in year $t-1$ affecting selection in year t would violate this condition.

The dependent variable, $math4$, is bounded between zero and one. Papke and Wooldridge (2008) estimate a fractional response probit unobserved effects model, where the unobserved heterogeneity is modeled using time averages of each covariate as in Chamberlain (1980) and Mundlak (1978). Their correlated random effects (CRE) estimation equation is:

$$E[math4_{it}|x_{i1}, x_{i2}, ..., x_{iT}] = \Phi(\psi_t + \mathbf{x}_{it}\boldsymbol{\beta} + \overline{\mathbf{x}}_i\boldsymbol{\xi}) \tag{22}$$

where $\overline{\mathbf{x}}_i$ includes the time averages of each covariate, $\Phi(\cdot)$ represents the normal CDF, and ψ_t allows for year-specific intercepts.[4]

We use indicators for each number of times a particular school appears in the data, $\mathbf{T}_i = T_{1i}, T_{2i}, ..., T_{5i}$, as sufficient statistics for the dependence between the unobserved, school-level heterogeneity and the selection of time-observations into our data. Accounting for the imbalance of

[4] As discussed above, the coefficients remain scaled by the variance of the unobserved heterogeneity.

the school-level data, we estimate:

$$E[math4_{it}|\mathbf{x}_{it}, \overline{\mathbf{x}}_i, \mathbf{T}_i] = \Phi(\psi_t + \mathbf{x}_{it}\boldsymbol{\beta} + \mathbf{T}_i\boldsymbol{\gamma} + \overline{\mathbf{x}}_i\boldsymbol{\xi}) \tag{23}$$

We term the above approach Correlated Random Effects for Unbalancedness (CREU).

In another specification, we interact time averages with indicators for the number of time-observations in an approach we label CREU1 below:

$$E[math4_{it}|\mathbf{x}_{it}, \overline{\mathbf{x}}_i, \mathbf{T}_i] = \Phi(\psi_t + \mathbf{x}_{it}\boldsymbol{\beta} + \mathbf{T}_i\boldsymbol{\gamma} + \overline{\mathbf{x}}_i\boldsymbol{\xi} + (\mathbf{T}_i \otimes \overline{\mathbf{x}}_i\boldsymbol{\delta})). \tag{24}$$

Table 5 provides estimates of these linear and nonlinear models using each methodology. The first two columns report estimates from POLS and FE linear regressions. The third and fourth columns report estimates from the analogous pooled fractional response (PFR) and CRE estimation. Columns five and six report the estimates from correlated random effects estimation accounting for unbalancedness without (CREU) and with (CREU1) interactions between time averages and time-observation indicators.

It is unclear how we should compute standard errors in this application. The observations are school buildings appearing (or sometimes not appearing) over time, making school-level clustering an obvious choice. However, the policy variation we next use to instrument for spending occurs at the district level, similar to a situation studied in Abadie et al. (2023), where a policy intervention is correlated within clusters. The Abadie et al. (2023) paper does not cover the panel data setting or instrumental variables, but it seems reasonable to conclude here that clustering at the district level is either correct or somewhat conservative. Consequently, we present school-building-clustered standard errors in parentheses and district-clustered standard errors in brackets, and use the generally more conservative, district-clustered standard errors for inference in our discussion.

Across all estimators, the estimated effect of expenditures on fourth-graders' achievement in math is positive, but accounting for unobserved heterogeneity leads to a decrease in the estimated effect. From the first row of column one using POLS, a 10% increase in average spending leads to an 0.84 percentage point (p-value < 0.001) increase in fourth-grade math pass rates. From column 2, the fixed-effects estimated effect of the same increase in spending is 0.72 percentage points (p-value = 0.034). Thus, using fully-robust, district-clustered standard errors, the estimated effect is statistically significantly positive using either linear approach.

PFR probit estimation yields similar results to POLS. Using this nonlinear approach, a 10% increase in average spending leads to an 0.87 percentage point (p-value < 0.001) increase in fourth-grade math pass rates. However, including time averages in the PFR probit causes the estimated effect to fall to 0.49 percentage points (p-value = 0.136) under the CRE approach.

Accounting for panel imbalance in columns five and six does little to change the CRE estimates. The estimated average effect of spending remains between 0.48 and 0.5 percentage points with p-values between 0.13 and 0.15. The similarity of these CREU-estimated coefficients to those estimated ignoring the selection of time-observations suggests that the unbalancedness of the panel is not driving the estimates in this application. It appears that once spending is allowed to be correlated with heterogeneity, allowing sample selection to be correlated with the same unobserved heterogeneity has little effect. Naturally, one cannot know this without having a method that explicitly allows for unbalanced panels – as we have provided in this paper.

Despite the similarity of the estimates, diagnostics provided in Table 6 support our interest in data imbalance. Using Wald tests between nested models and clustering at the two levels, we reject the hypothesis that the coefficients on the time averages are zero for the linear model (column one) and the nonlinear model (column two). In the third column, we test the coefficient estimates on the four indicators for the number of time-observations. We reject the null hypothesis that all four coefficients are zero at the 5% confidence level when we cluster the data at the school-level. However, clustering at the district-level, we fail to reject the null with a chi-squared test statistic of only 6.2 corresponding to a p-value of 0.18. When testing the 30 interactions between indicators for the number of time-observations and the time averages, we reject the null hypothesis that the coefficient on each is zero.[5] Thus, there appears to be significant unbalancedness and unobserved heterogeneity, but the estimated effects of spending on pass rates remain robust to controlling for panel imbalance when treating spending as exogenous.

[5]Note that due to collinearity, interactions between the indicator for 5 time-observations and time averages for indicators for years 1995, 1996, 1997, and 1998 are omitted as is the interaction between the indicator for four time-observations and the time average of the indicator for year 1997 and 1998.

4.2 Allowing spending to be endogenous

The FE, CRE, and CREU results are robust to unobserved heterogeneity that may be correlated with spending and math pass rates, but in this section, we use instrumental variables to address the possible endogeneity of spending and violations of the strict exogeneity assumption. Michigan's Proposal A began to compress the variation in revenue to districts according to a non-smooth function determined by district spending in 1994, and we use the log of the foundation grants ($lfound_{it}$) as a time-varying instrument for spending ($lrexppp_{it}$). Further, we control for the log of real per-pupil expenditures in 1994 ($lrexppp_{i94}$) to capture this initial heterogeneity in spending, and model cumulative spending according to equation (25) below.

$$\begin{aligned} log(avgrexp_{it}) =& \eta_t + \pi_1 lfound_{it} + \pi_2 lrexppp_{i94} + \pi_3 lunch_{it} + \pi_4 lunch_{it}^2 \\ & + \pi_5 log(enroll_{it}) + \pi_6 log(enroll_{it})^2 + c_i + v_{it2} \end{aligned} \tag{25}$$

As a benchmark, we estimate equation (19) using pooled two-stage least squares (P2SLS) with equation (25) serving as the first stage. The F-statistic on $lfound_{it}$ from the first stage POLS regression is 180.95 demonstrating that the state foundation grants are indeed predictive of spending. We add the residuals from POLS estimation of equation (25), $\widehat{v}_{it2}$, to the pooled fractional probit model to accommodate the nonlinear functional form of fourth-grade math pass rates (PFR CF).

We address the unobserved heterogeneity, c_i, by modeling it as a function of the building-level time averages, and including them as regressors as in Chamberlain (1980) and Mundlak (1978). Equation (26) reflects the first stage of this approach.[6]

$$\begin{aligned} log(avgrexp_{it}) = & \eta_t + \pi_1 lfound_{it} + \pi_2 lrexppp_{i94} + \pi_3 lunch_{it} + \pi_4 lunch_{it}^2 \\ & + \pi_5 log(enroll_{it}) + \pi_6 log(enroll_{it})^2 + \pi_7 \overline{lunch_i} + \pi_8 \overline{lunch_i^2} + \pi_9 \overline{log(enroll_i)} \\ & + \pi_{10} \overline{log(enroll_i)^2} + \pi_{11} \overline{y96}_i + \pi_{12} \overline{y97}_i + \pi_{13} \overline{y98}_i + v_{it2}. \end{aligned} \tag{26}$$

Equation (26) is akin to the first stage in fixed effects instrumental variables (FEIV) except that we use the base expenditures ($lrexppp_{i94}$) to proxy for time-invariant spending as opposed to the time average of spending. To allow for nonlinearities, we incorporate the estimated residuals, $\widehat{v}_{it2}$, from

[6] Note that due to the unbalancedness of the data, we also include time averages of the year indicators – $\overline{y96}_i$, $\overline{y97}_i$, and $\overline{y98}_i$.

the same first-stage regression into equation (22). This correlated random effects control function (CRE CF) approach allows us to handle the endogeneity of spending while accommodating the nonlinear functional form.

We handle the possibly endogenous unbalancedness of the panel by incorporating the number of time-observations into the estimation of the model. We do this by first including indicators for the number of time-observations to the CRE CF in both estimation stages in what we term a correlated random effects unbalancedness control function (CREU CF) approach. Secondly, we incorporate interactions between time averages (and $lrexppp_{i94}$) and the number of time-observations to more fully account for the unbalancedness of the panel (CREU1 CF).[7]

Across all instrumental variables approaches the point estimates range from 0.19 to 0.25. These findings all lie within the P2SLS confidence interval reported in Papke (2005). The P2SLS estimated average partial effect (APE) of a 10% increase in $log(avgrexp_{it})$ is a 2.07 percentage point increase in fourth-grade math pass rates (p-value = 0.013). Once we account for the endogeneity, modeling the unobserved heterogeneity does little to change the point estimates.

The nonlinear estimators uncover slightly larger effects of spending than do the linear instrumental variables approaches. The coefficient estimates range from 0.2 to 0.24. Accounting for unobserved heterogeneity and adjusting for panel imbalance leads to smaller effects, though these differences are far from statistically significant. Including interactions between indicators for the number of time-observations and the time averages of covariates (CREU1 CF) drops the CRE CF estimated APE of spending by 9.5 percent, roughly the same amount as accounting for unobserved heterogeneity (comparing PFR CF to CRE CF).

There is little loss of efficiency when accounting for panel imbalance. Due to the estimation of the residuals in the first stage, we cluster-bootstrap the standard errors over 500 repetitions to account for the estimation error. The standard errors hardly change between CRE CF and CREU CF and the CREU1 CF standard errors are only 11 to 16 percent larger than those from CRE CF.

The test statistics on $\widehat{v}_{it2}$ from these control function approaches conveniently provide evidence regarding the prevalence of endogeneity of $log(avgrexp_{it})$. The t-statistics for PFR CF, CRE CF, and CREU CF range from 2.04 to 2.28 with CREU1 CF having a t-statistic of 1.66. These results

[7] We provide additional robustness checks in the appendix where we include interactions between time averages and time-observation indicators in the linear estimation (FEIVU), and standard differencing in the first-stage estimation of our nonlinear approaches (CRE FECF, CREU FECF, CREU1 FECF).

provide some evidence against the hypothesis that spending is strictly exogenous, and it is important to note that the estimated APEs from the instrumental variables approaches in Table 7 are two to four times larger than the estimated APEs in Table 5 assuming strict exogeneity of spending.

Table 8 provides the results from Wald tests between nested two-stage models. We test the constraints that the seven time averages have zero effect on 4th-grade math pass rates in the first two columns. Despite the closeness of the estimated APEs, in both cases, we reject the hypothesis that the coefficient estimates on the time averages are zero with p-values less than 0.001. In column four, we test the coefficient estimates on the four indicators for the number of time-observations. We are unable to reject the hypothesis that the coefficients on the four time-observation indicators are zero at the 95% level. When testing the 27 interactions between indicators for the number of time-observations and the time averages (and $lrexppp_{i94}$), however, the results reject the null hypothesis with p-values smaller than 0.001.[8] While these Wald tests inform whether or not the unobserved heterogeneity and unbalancedness affect test scores conditional on the covariates and foundation grants, the stability of the estimated APEs of spending is reassuring. The effects of spending on fourth-grade math pass rates are not driven by the unbalancedness of the panel.

The methods used here are robust to panel unbalance related to unobserved time-invariant heterogeneity, though they are not robust to selection based on idiosyncratic shocks. In order to gain indirect evidence of selection based on shocks, we attempt to predict missingness by including an indicator for whether the lead observation is present. Conducting a standard t-test on the coefficient of the lead of selection reveals whether there is a systematic relationship between future missingness and math test pass rates. Table 9 reveals no systematic relationship between the lead selection indicator and math test pass rates. While the indicator is statistically significantly positive in the first two columns with school-clustered standard errors, controlling for the number of time-observations causes the coefficient to fall in magnitude and lose statistical significance. The coefficient is never statistically significant with our preferred district-clustered standard errors.

In summary, there is significant evidence of unobserved heterogeneity across schools and of endogeneity in school spending. While the bulk of the evidence points to significant relationships between panel imbalance and fourth-grade math pass rates, the estimated effects of spending on

[8]Note that due to collinearity, interactions between the indicator for five time-observations and time averages for indicators for years 1996, 1997, and 1998 are omitted as are the interactions between the indicator for four time-observations and the time averages of the indicators for years 1996 and 1997.

pass rates remain robust to controlling for panel imbalance.

5 Discussion

This paper considers estimation of nonlinear panel data models when the panel is unbalanced in the presence of endogeneity. We allow the selection of time-observations to be correlated with both unobserved heterogeneity as well as explanatory variables. We take a correlated random effects approach and model the unobserved heterogeneity while controlling for selection of time-observations. We incorporate a control function approach to handle endogeneity of explanatory variables. The approach is easily extended to other nonlinear panel data models, such as ordered probit and Tobit. Pooled (quasi-) MLE can be applied by combining the CRE and control function approach we propose here. In estimating average partial effects we adopt quasi-MLE such that consistency does not require knowledge of the specific distribution. Our approach is straightforward to implement with standard statistical software and may be used when the outcome is binary, fractional response, or otherwise bounded with known upper and lower bounds. As cases of unbalanced panels are common in many applied fields of economics as well as in other disciplines such as quantitative sociology and political science, there is wide potential for application across the social sciences.

The approach we have proposed in this paper requires the existence of a time-varying instrumental variable for each explanatory variable allowed to be endogenous with respect to idiosyncratic shocks. In our application, we exploit a change in the way schools were funded. Many other examples exist. Just a few include Levitt (1996), who exploits variation in litigation timing, Black et al. (2002) who exploit coal price shocks, and Dahl and Lochner (2012), who exploit changes in the earned income tax credit.

In estimating the effect of school spending on fourth-grade pass rates on state mathematics exams, we see significant unbalancedness in the underlying data. Estimation is additionally complicated by likely unobserved heterogeneity across schools and the potential of contemporaneous endogeneity of school spending. Indeed, we find evidence supporting the existence of both. Using instrumental variables to identify the effect of spending off of plausibly exogenous changes in the funding structure is consequential. Whereas we estimate that a 10 percent increase in spending leads to a 0.5 percentage point increase in pass rates when we ignore the potential of endogeneity in school spending decisions, we estimate the same spending change to increase pass rates by around

2 percentage points with a variety of instrumental variables approaches.

Once we address the contemporaneous endogeneity of school spending, we find our results to be quite robust. Despite our Wald tests rejecting the null of no unobserved fixed heterogeneity, our estimates remain relatively stable regardless of our approach to address the unbalancedness of our panel. This stability may not carry over to other contexts, such as the effect of development on inequality (Nielsen and Alderson, 1995), the effect of private school vouchers on academic achievement (Rouse, 1998), the effect of income on armed conflict (Miguel et al., 2004), or the effect of voting history on turnout (Denny and Doyle, 2009), where researchers document panel unbalance with bounded dependent variables. In our application, our estimates provide further evidence of the positive effects of school spending on students' academic achievement. This result has found additional support in recent work, such as Jackson et al. (2016); Hyman (2017); and Lafortune et al. (2018) – perhaps finally turning the prevailing narrative to more positively depicting the efficacy of expenditures on public schooling.

Disclosures

Michael Bates, Leslie Papke, and Jeffrey Wooldridge declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article. The authors disclose receipt of the following financial support for the research, authorship, and/or publication of this article: The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through R305B090011 to Michigan State University. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education. The data used for this project lie outside the scope of IRB approval and no IRB approval was sought.

References

Abadie, A., S. Athey, G. W. Imbens, and J. M. Wooldridge (2023). When should you adjust standard errors for clustering? *The Quarterly Journal of Economics 138*(1), 1–35.

Alderman, H., J. R. Behrman, H.-P. Kohler, J. A. Maluccio, and C. S. Watkins (1999). *Attrition in longitudinal household survey data: some tests for three developing-country samples*. The World Bank.

Aughinbaugh, A. (2004). The impact of attrition on the children of the nlsy79. *Journal of human resources 39*(2), 536–563.

Black, D., K. Daniel, and S. Sanders (2002). The impact of economic conditions on participation in disability programs: Evidence from the coal boom and bust. *American Economic Review 92*(1), 27–50.

Blundell, R. and J. L. Powell (2003). Endogeneity in nonparametric and semiparametric regression models.

Chamberlain, G. (1980). Analysis of covariance with qualitative data. *Review of Economic Studies 47*, 225–238.

Chaudhary, L. (2009, February). Education inputs, student performance and school finance reform in Michigan. *Economics of Education Review 28*(1), 90–98.

Chiang, H. (2009). How accountability pressure on failing schools affects student achievement. *Journal of Public Economics 93*(9-10), 1045–1057.

Dahl, G. B. and L. Lochner (2012). The impact of family income on child achievement: Evidence from the earned income tax credit. *American Economic Review 102*(5), 1927–1956.

Denny, K. and O. Doyle (2009). Does voting history matter? analysing persistence in turnout. *American journal of political science 53*(1), 17–35.

Falaris, E. M. and H. E. Peters (1998). Survey attrition and schooling choices. *Journal of Human Resources*, 531–554.

Fitzgerald, J. M. (2011). Attrition in models of intergenerational links using the psid with extensions to health and to sibling models. *The BE journal of economic analysis & policy 11*(3).

Hyman, J. (2017). Does money matter in the long run? effects of school spending on educational attainment. *American Economic Journal: Economic Policy 9*(4), 256–80.

Jackson, C. K., R. C. Johnson, and C. Persico (2016). The effects of school spending on educational and economic outcomes: Evidence from school finance reforms. *The Quarterly Journal of Economics 131*(1), 157–218.

Joshi, R. and J. M. Wooldridge (2019). Correlated random effects models with endogenous explanatory variables and unbalanced panels. *Annals of Economics and Statistics* (134), 243–268.

Lafortune, J., J. Rothstein, and D. W. Schanzenbach (2018). School finance reform and the distribution of student achievement. *American Economic Journal: Applied Economics 10*(2), 1–26.

Levitt, S. D. (1996). The effect of prison population size on crime rates: Evidence from prison overcrowding litigation. *The quarterly journal of economics 111*(2), 319–351.

Lin, W. and J. M. Wooldridge (2019). Testing and correcting for endogeneity in nonlinear unobserved effects models. In *Panel data econometrics*, pp. 21–43. Elsevier.

Miguel, E., S. Satyanath, and E. Sergenti (2004, August). Economic Shocks and Civil Conflict: An Instrumental Variables Approach. *Journal of Political Economy 112*(4), 725–753.

Mundlak, Y. (1978). On the pooling of time series and cross-sectional data. *Econometrica 46*, 69–86.

Nielsen, F. and A. S. Alderson (1995). Income inequality, development, and dualism: Results from an unbalanced cross-national panel. *American sociological review*, 674–701.

Papke, L. E. (2005, June). The effects of spending on test pass rates: evidence from Michigan. *Journal of Public Economics 89*(5–6), 821–839.

Papke, L. E. (2008, July). The Effects of Changes in Michigan's School Finance System. *Public Finance Review 36*(4), 456–474.

Papke, L. E. and J. M. Wooldridge (2008, July). Panel data methods for fractional response variables with an application to test pass rates. *Journal of Econometrics 145*(1–2), 121–133.

Rivers, D. and Q. H. Vuong (1988). Limited information estimators and exogeneity tests for simultaneous probit models. *Journal of econometrics 39*(3), 347–366.

Rouse, C. E. (1998). Private school vouchers and student achievement: An evaluation of the milwaukee parental choice program. *The Quarterly journal of economics 113*(2), 553–602.

Roy, J. (2011, February). Impact of School Finance Reform on Resource Equalization and Academic Performance: Evidence from Michigan. *Education Finance and Policy 6*(2), 137–167.

Semykina, A. and J. M. Wooldridge (2010). Estimating panel data models in the presence of endogeneity and selection. *Journal of Econometrics 157*(2), 375–380.

Wooldridge, J. M. (1995). Selection corrections for panel data models under conditional mean independence assumptions. *Journal of econometrics 68*(1), 115–132.

Wooldridge, J. M. (2019). Correlated random effects models with unbalanced panels. *Journal of Econometrics 211*(1), 137–150.

Table 1: Average correlations across simulation repetitions for both correlated fixed effects and correlated random-coefficient data generating processes (DPGs)

	Correlated Fixed Effect DGP					Correlated Random Coefficient DGP			
	x_1	c	u_0	T		x_1	c	u_1	T
x_1	1				x_1	1			
c	0.3152	1			c	0.3152	1		
u_0	0.0006	0.0009	1		u_1	0.3404	0.6294	1	
T	0.0941	0.2986	0.0013	1	T	0.1119	0.2065	0.3300	1

Notes: Average correlations over 500 simulation repetitions. x_1 is the primary variable of interest. c represents the unobserved fixed effects, and u_0 and u_1 are random slopes. T represents the number of time-observations for a given "school." Correlated fixed effect DGP used in simulations appearing in Table 2. Correlated random-coefficient DGP used in simulations appearing in Table 3.

Table 2: Simulation evidence with selection based on unobserved fixed effects

True mean APEs of x1 over:				Population	Sample	T =1	T =2	T =3	T =4	T =5
Mean APE				0.2979	0.2953	0.3252	0.3163	0.3068	0.2965	0.2851
Estimates	POLS	FE	FEU	PFR	CRE	CREU	CREU1	CREU2	CREU3	
Mean APE	0.3850	0.2939	0.2947	0.3850	0.2947	0.2947	0.2947	0.2949	0.2951	
Mean SE	0.0112	0.0099	0.0099	0.0101	0.0086	0.0086	0.0086	0.0086	0.0086	
SD	0.0111	0.0098	0.0098	0.0102	0.0087	0.0087	0.0086	0.0089	0.0092	
Mean SE/SD	1.0096	1.0108	1.0055	0.9915	0.9950	0.9924	0.9919	0.9611	0.9298	
Partial effects (PEs) from CREU1 at selected deciles						10	30	50	70	90
PE at decile						0.3314	0.3145	0.3001	0.2837	0.2576
SD of PE at decile						0.0104	0.0097	0.0091	0.0082	0.0069

Notes: Simulated over 500 repetition with 500 individual "buildings" and a mean of 3.83 out of 5 possible time-observations (T). All standard errors (SE) clustered at the building level. FE regressions are estimated as POLS including time averages. PFR includes the same covariates as POLS with a nonlinear functional form (normal GLM). CRE incorporates individual time averages into PFR. CREU incorporates indicators for the number of individual Tobs into CRE. CREU1 adds interactions between Tobs indicators and covariate time averages to CREU. CREU2 incorporates interactions between covariates and the time averages into CREU1. CREU3 adds triple interactions between covariates, their time averages, and indicators for Tobs.

Table 3: Simulation evidence with selection based on correlated heterogeneous slopes

True mean APEs of x1 over:	Population	Sample	T =1	T =2	T =3	T =4	T =5
Mean APE	0.2902	0.2933	0.2523	0.2670	0.2812	0.2931	0.3034

Estimates	POLS	FE	FEU	PFR	CRE	CREU	CREU1	CREU2	CREU3
Mean APE	0.3823	0.2954	0.2941	0.3827	0.2945	0.2946	0.2947	0.2945	0.2937
Mean SE	0.0109	0.0099	0.0099	0.0099	0.0090	0.0090	0.0090	0.0089	0.0086
SD	0.0101	0.0094	0.0091	0.0094	0.0083	0.0083	0.0083	0.0085	0.0096
Mean SE/SD	1.0756	1.0494	1.0840	1.0584	1.0795	1.0798	1.0794	1.0458	0.8941

Partial effects (PEs) from CREU1 at selected deciles	10	30	50	70	90
PE at decile	0.3321	0.3147	0.3000	0.2833	0.2568
SD of PE at decile	0.0103	0.0095	0.0087	0.0078	0.0064

Notes: Simulated over 500 repetition with 500 individual "buildings" and a mean of 3.82 out of 5 possible time-observations (T). All standard errors (SE) clustered at the building level. FE regressions are estimated as POLS including time averages. PFR includes the same covariates as POLS with a nonlinear functional form (normal GLM). CRE incorporates individual time averages into PFR. CREU incorporates indicators for the number of individual Tobs into CRE. CREU1 adds interactions between Tobs indicators and covariate time averages to CREU. CREU2 incorporates interactions between covariates and the time averages into CREU1. CREU3 adds triple interactions between covariates, their time averages, and indicators for Tobs.

Table 4: Summary statistics by number of time-observations per school

Number of time-observations	*Total*	1	2	3	4	5
Pass rate on fourth-grade math test	0.64	0.72	0.60	0.64	0.57	0.65
	(0.20)	(0.23)	(0.21)	(0.20)	(0.23)	(0.19)
Average real per-pupil expenditures	3,897.43	3,949.25	4,081.85	3,968.45	3,875.25	3,874.76
	(626.21)	(1,181.29)	(1,051.42)	(663.45)	(519.66)	(613.66)
Percent FRL eligible	0.37	0.19	0.49	0.40	0.57	0.31
	(0.25)	(0.27)	(0.25)	(0.26)	(0.26)	(0.22)
Number of enrolled students	419.91	282.13	366.40	411.40	540.24	397.97
	(164.77)	(172.61)	(222.63)	(153.23)	(222.59)	(137.03)
Number of schools	1,771	54	42	506	259	910

Notes: Sample means with standard deviations appearing in parentheses.

Table 5: APE estimates assuming spending to be strictly exogenous

	Linear		Fractional Probit			
VARIABLES	POLS	FE	PFR	CRE	CREU	CREU1
lavgrexpp	0.084	0.072	0.087	0.049	0.049	0.048
	(0.017)	(0.026)	(0.018)	(0.025)	(0.025)	(0.025)
	[0.023]	[0.034]	[0.025]	[0.033]	[0.033]	[0.033]
lunch	−0.447	−0.067	−0.439	−0.071	−0.070	−0.070
	(0.012)	(0.044)	(0.012)	(0.043)	(0.043)	(0.042)
	[0.026]	[0.047]	[0.026]	[0.046]	[0.046]	[0.045]
lenrol	−0.015	−0.022	−0.014	−0.019	−0.019	−0.019
	(0.009)	(0.021)	(0.008)	(0.021)	(0.021)	(0.021)
	[0.010]	[0.023]	[0.010]	[0.022]	[0.022]	[0.022]
Observations	7,242	7,242	7,242	7,242	7,242	7,242

Notes: School-clustered standard errors appear in parentheses. District-clustered standard errors are in brackets. CREU estimation includes indicators for each number of time-observations. CREU1 includes indicators for time-observations as well as interactions between time-observations and time averages of covariates. All regressions include year indicators. Regressions including time averages also include time averages of year indicators.

Table 6: Wald testing between nested single-stage models

Model Comparison	POLS vs FE	PFR vs CRE	CRE vs CREU	CREU vs CREU1
Panel A: School-level clustered standard errors				
χ^2 (constraints)	13.7(9)	114.7(9)	9.7(4)	93.7(30)
Prob > χ^2	< 0.001	< 0.001	0.046	< 0.001
Panel B: District-level clustered standard errors				
χ^2 (constraints)	13.2(9)	107.2(9)	6.2(4)	145.1(30)
Prob > χ^2	< 0.001	< 0.001	0.184	< 0.001
Variables tested				
time averages	X	X		
time-observation			X	
time-observation interactions				X

Table 7: APEs allowing spending to be contemporaneously endogenous

	Linear model		Fractional response probit			
VARIABLES	P2SLS	FEIV	PFR CF	CRE CF	CREU CF	CREU1 CF
lavgrexpp	0.207	0.191	0.24	0.223	0.222	0.201
	(0.055)	(0.056)	(0.061)	(0.064)	(0.064)	(0.074)
	[0.083]	[0.083]	[0.093]	[0.093]	[0.095]	[0.107]
residuals			−0.219	−0.197	−0.194	−0.178
			(0.069)	(0.073)	(0.074)	(0.082)
			[0.096]	[0.095]	[0.097]	[0.107]
lunch	−0.454	−0.022	−0.655	−0.037	−0.043	−0.053
	(0.015)	(0.062)	(0.014)	(0.062)	(0.062)	(0.06)
	[0.030]	[0.066]	[0.048]	[0.065]	[0.066]	[0.064]
lenrol	−0.005	−0.037	−0.197	0.018	0.035	−0.009
	(0.012)	(0.033)	(0.011)	(0.033)	(0.033)	(0.034)
	[0.015]	[0.042]	[0.015]	[0.042]	[0.045]	[0.046]
Observations	4,853	4,853	4,853	4,853	4,853	4,853

Notes: School-clustered standard errors appear in parentheses. District-clustered standard errors are in brackets. Control function standard errors are from 500 cluster-bootstrap repetitions to handle residuals' estimation error. CREU estimation includes indicators for each number of time-observations. CREU1 includes indicators for time-observations as well as interactions between time-observations and time averages of covariates. All regressions include year indicators. Regressions including time averages also include time average of year indicators.

Table 8: Wald testing between nested two-stage models

Model Comparison	IV vs FEIV	PFR CF vs CRE CF	CRE CF vs CREU CF	CREU CF vs CREU1 CF
Panel A: School level clustered standard errors				
χ^2 (constraints)	78.7(7)	85.9(7)	8.2(4)	80.1(27)
Prob $> \chi^2$	< 0.001	< 0.001	0.083	< 0.001
Panel B: District level clustered standard errors				
χ^2 (constraints)	73.1(7)	55.7(7)	2.5(4)	59.4(27)
Prob $> \chi^2$	< 0.001	< 0.001	0.644	< 0.001
Variables tested				
time averages	X	X		
time-observation indicators			X	
time-observation interactions				X

Table 9: Evaluating selection based on shocks

VARIABLES	PFR CF	CRE CF	CREU CF	CREU1 CF
lavgrexpp	0.252	0.199	0.237	0.202
	(0.068)	(0.069)	(0.069)	(0.075)
	[0.107]	[0.107]	[0.109]	[0.121]
residuals	−0.238	−0.189	−0.227	−0.192
	(0.079)	(0.081)	(0.081)	(0.088)
	[0.114]	[0.115]	[0.117]	[0.125]
lunch	−0.473	0.014	0.018	0.005
	(0.016)	(0.069)	(0.070)	(0.068)
	[0.049]	[0.076]	[0.078]	[0.075]
lenrol	−0.008	0.006	0.014	0.009
	(0.013	(0.040)	(0.040)	(0.040)
	[0.015]	[0.048]	[0.048]	[0.049]
$1[S_{it+1} = 1]$	0.033	0.042	0.029	0.015
	(0.020)	(0.021)	(0.021)	(0.019)
	[0.033]	[0.027]	[0.021]	[0.017]
Observations	3,602	3,602	3,602	3,602

Notes: School-clustered standard errors appear in parentheses. District-clustered standard errors are in brackets. Control function standard errors are from 500 cluster-bootstrap repetitions to handle residuals' estimation error. CREU estimation includes indicators for each number of time-observations. CREU1 includes indicators for time-observations as well as interactions between time-observations and time averages of covariates. All regressions include year indicators. Regressions including time averages also include time average of year indicators.

6 Online appendix

6.1 Simulated distributional misspecification

To investigate the performance of the estimator when the distribution is not normal we reran the simulation using data-generating processes from two other distributions. In both,

$$y_{it} = \frac{exp\left[\alpha + (\beta_1 + u_{ig})x_{it1} + \beta_2 x_{it2} + c_i + v_{it}\right]}{1 + exp\left[\alpha + (\beta_1 + u_{ig})x_{it1} + \beta_2 x_{it2} + c_i + v_{it}\right]}. \tag{27}$$

We further add complication, by drawing v_{it} from a logistic distribution in the first case. In the second case, we draw v_{it} from a uniform [-1,1] distribution.

We present the results from these simulations in Table 10 and Table 11 respectively. Our estimators perform well in both cases. This is perhaps unsurprising. Given our expression for the conditional mean, we can estimate the parameters using a pooled quasi-maximum likelihood approach with the log-likelihood being chosen to be that for the Bernoulli distribution. This QMLE approach in estimating average partial effects allows for consistency without requiring knowledge of the specific distribution.

6.2 Application robustness

We also perform several robustness checks. First, we incorporate the interactions between time averages of each covariate and indicators for the number of time-observations of each school in the linear fixed effects model. We term this approach FEIVU, which serves as the linear analogous methodology to the CREU1 CF approach. Second, we treat the time-constant heterogeneity in schools by demeaning each covariate in the first stage to generate the residuals used in each of the three control function approaches. Naturally, the differencing eliminates time-constant covariates, such as the number of time-observations and the level of expenditures in 1994, from the first stage estimation. These three approaches (CRE FECF, CREU FECF, CREU1 FECF) appear in Table 12.

The point estimates across specifications are close to those presented in the main text. In the linear model, accounting for the unbalancedness of the panel drops the point estimate of a 10% increase in spending to 1.69 percentage points (p-value = 0.069) from 1.94 (FEIV), though it remains economically significant. The APEs from approaches using fixed effects estimation in the first stage range from 0.212 to 0.25, making them only slightly larger in magnitude than those previously discussed. The primary difference in these estimates is in their precision. Demeaning the

Table 10: Simulation evidence with selection based on correlated heterogeneous slopes with logistic DPG

True APEs of x1 over:	Population	Sample	T =1	T =2	T =3	T =4	T =5
Mean APE	0.2105	0.2135	0.1763	0.1886	0.2011	0.2128	0.2241

Estimated APEs of x1	POLS	FE	FEU	PFR	CRE	CREU	CREU1	CREU2	CREU3
Mean APE	0.2786	0.2153	0.2141	0.2785	0.2140	0.2140	0.2141	0.2140	0.2137
Mean SE	0.0074	0.0064	0.0064	0.0073	0.0063	0.0063	0.0063	0.0062	0.0060
SD	0.0069	0.0062	0.0060	0.0067	0.0058	0.0058	0.0058	0.0060	0.0067
Mean SE/SD	1.0795	1.0371	1.0741	1.0806	1.0753	1.0758	1.0759	1.0407	0.8933

Notes: Simulated over 500 repetition with 500 individual "buildings" and a mean of 3.83 out of 5 possible time-observations (T). All standard errors (SE) clustered at the building level. FE regressions are estimated as POLS including time averages. PFR includes the same covariates as POLS with a nonlinear functional form (normal GLM). CRE incorporates individual time averages into PFR. CREU incorporates indicators for the number of individual Tobs into CRE. CREU1 adds interactions between Tobs indicators and covariate time averages to CREU. CREU2 incorporates interactions between covariates and the time averages into CREU1. CREU3 adds triple interactions between covariates, their time averages, and indicators for Tobs.

foundation grants in the first-stage fixed-effects estimation rather than incorporating base-period expenditures in the first-stage increases the magnitude of the district-cluster-bootstrapped standard errors by roughly 20 to 30%.

Table 11: Simulation evidence with selection based on correlated heterogeneous slopes with mixture DPG

True APEs of x1 over:	Population	Sample	T =1	T =2	T =3	T =4	T =5
Mean APE	0.2010	0.2040	0.1664	0.1792	0.1917	0.2034	0.2144

Estimated APEs of x1	POLS	FE	FEU	PFR	CRE	CREU	CREU1	CREU2	CREU3
Mean APE	0.2668	0.2060	0.2052	0.2668	0.2052	0.2052	0.2053	0.2051	0.2056
Mean SE	0.0146	0.0163	0.0158	0.0145	0.0156	0.0156	0.0156	0.0156	0.0154
SD	0.0143	0.0161	0.0157	0.0141	0.0155	0.0155	0.0155	0.0186	0.0269
Mean SE/SD	1.0265	1.0099	1.0084	1.0275	1.0063	1.0079	1.0081	0.8360	0.5710

Notes: Simulated over 500 repetitions with 500 individual "buildings" and a mean of 3.83 out of 5 possible time-observations (T). All standard errors (SE) clustered at the building level. FE regressions are estimated as POLS including time averages. PFR includes the same covariates as POLS with a nonlinear functional form (normal GLM). CRE incorporates individual time averages into PFR. CREU incorporates indicators for the number of individual Tobs into CRE. CREU1 adds interactions between Tobs indicators and covariate time averages to CREU. CREU2 incorporates interactions between covariates and the time averages into CREU1. CREU3 adds triple interactions between covariates, their time averages, and indicators for Tobs.

Table 12: Robustness: APEs allowing spending to be contemporaneously endogenous

	Linear		Nonlinear	
VARIABLES	FEIVU	CRE	CREU	CREU1
		FECF	FECF	FECF
lavgrexpp	0.168	0.25	0.248	0.212
	(0.063)	(0.073)	(0.073)	(0.076)
	[0.092]	[0.127]	[0.126]	[0.129]
residuals		−0.223	−0.22	−0.19
		(0.081)	(0.081)	(0.083)
		[0.126]	[0.126]	[0.127]
lunch	−0.014	−0.039	−0.044	−0.056
	(0.061)	(0.065)	(0.065)	(0.063)
	[0.064]	[0.070]	[0.070]	[0.068]
lenrol	0.025	0.01	0.023	−0.032
	(0.034	(0.03)	(0.03)	(0.03)
	[0.042]	[0.036]	[0.036]	[0.035]
Observations	4,853	4,853	4,853	4,853

Notes: School-clustered standard errors appear in parentheses. District-clustered standard errors are in brackets. Control function standard errors are from 500 cluster-bootstrap repetitions to handle residuals' estimation error. CREU estimation includes indicators for each number of time-observations. CREU1 includes indicators for time-observations as well as interactions between time-observations and time averages of covariates. All regressions include year indicators. Regressions including time averages also include time average of year indicators.