

Double Machine Learning for Panel Data^{*}

Paul Clarke[†] Annalivia Polselli[‡]

December 19, 2023

Abstract

Machine Learning (ML) algorithms are powerful data-driven tools for approximating high-dimensional or non-linear nuisance functions which are useful in practice because the true functional form of the predictors is *ex-ante* unknown. In this paper, we develop estimators of policy interventions from panel data which allow for non-linear effects of the confounding regressors, and investigate the performance of these estimators using three well-known ML algorithms, specifically, LASSO, classification and regression trees, and random forests. We use Double Machine Learning (DML) (Chernozhukov et al., 2018) for the estimation of causal effects of homogeneous treatments with unobserved individual heterogeneity (fixed effects) and no unobserved confounding by extending Robinson (1988)’s partially linear regression model. We develop three alternative approaches for handling unobserved individual heterogeneity based on extending the within-group estimator, first-difference estimator, and correlated random effect estimator (Mundlak, 1978) for non-linear models. Using Monte Carlo simulations, we find that conventional least squares estimators can perform well even if the data generating process is non-linear, but there are substantial performance gains in terms of bias reduction under a process where the true effect of the regressors is non-linear and discontinuous. However, for the same scenarios, we also find – despite extensive hyperparameter tuning – inference to be problematic for both tree-based learners because these lead to highly non-normal estimator distributions and the estimator variance being severely under-estimated. This contradicts the performance of trees in other circumstances and requires further investigation. Finally, we provide an illustrative example of DML for observational panel data showing the impact of the introduction of the national minimum wage in the UK.

JEL codes: C14, C18, C33, C45.

Keywords: Homogeneous treatment effect, LASSO, tree-based approaches, hyperparameter tuning.

^{*}We thank Thomas Cornelissen, Riccardo Di Francesco, Damian Machlanski, Spyros Samothrakis, David Zentler-Munro, and the participants to the Annual MiSoC Workshop, RSS 2023, ISER seminar, MiSoC Advisory Board, IADS Meeting group. This research was funded by the UK Economic and Social Research Council award ES/S012486/1 (MiSoC). The program is implemented in the statistical software R; the package XTDML is available at <https://github.com/POLSEAN/XTDML>.

[†]Institute for Social and Economic Research, University of Essex. (e-mail: pclarke@essex.ac.uk)

[‡]Institute for Analytics and Data Science, University of Essex. (Corresponding author. e-mail: annalivia.polselli@essex.ac.uk. Postal address: Wivenhoe Park, Colchester CO4 3SQ, UK.)

1 Introduction

Machine Learning (ML) is a field at the interface of artificial intelligence and computer science concerned with developing algorithms for solving prediction and classification problems. In theory, the powerful algorithms of supervised ML allow researchers to fit, or *learn*, high-dimensional non-linear functions of predictor variables (including e.g. complex interaction structures) without having to specify the functional form of these relationships. Classical examples of supervised ML algorithms widely used across many fields include the Least Absolute Shrinkage and Selection Operator (LASSO), Classification and Regression Trees (CART) and Random Forests (RF). In its simplest form, ML requires randomly partitioning the sample data into training and testing samples, where the algorithm is learnt by fitting the training sample, and its performance assessed using the testing sample. A major practical challenge for ML is to avoid *overfitting*, that is, learning that incorporates excessive noise from the training sample and so results in an over-complicated algorithm that performs sub-optimally when applied to the testing sample. ML can avoid overfitting by a suitable choice of the *hyperparameters* for a particular algorithm, the values of which are tuned using the validation data (a further partition of the sample data) or cross-validation. For example, LASSO performs *regularisation* to penalize model complexity using the L_1 -norm, with validation data or cross-validation used to choose the appropriate value of the tuning hyperparameter (Hastie et al., 2009).

There is a growing interest in economics in the potential of ML for data modelling and for enhancing existing approaches to estimation, including for the effects of treatments or policy interventions on the population of interest. Notable developments include ML algorithms for causal analysis such as Honest Trees (Athey and Imbens, 2016), Causal Forests (Wager and Athey, 2018) and Generalised Random Forests (Athey et al., 2019). These set out the basis for further developments of tree-based approaches for estimation and inference (e.g., Lechner and Okasa, 2019; Lechner and Mareckova, 2022; Di Francesco, 2022, 2023). However, the key development, as far as this paper is concerned, is Double/Debiased Machine Learning (DML) for semi-parametric estimation problems with nuisance parameters that are non-linear or high-dimensional functions of the model variables (Chernozhukov et al., 2018). The motivation for DML is that the intuitive approach, to use ML to model the nuisance parameters needed to calculate the estimate and plug predictions based on the the learnt algorithm into the estimating equation, can lead to substantial bias. The bias comes because learners optimise mean square error loss rather than the bias, which can arise through either regularisation or overfitting, and results in estimators of the interest parameter that do not converge at the usual \sqrt{N} rate, which causes difficulties for conventional first-order asymptotic theory. Chernozhukov et al. (2018) propose DML to correct the resulting bias and allow conventional first-order inference by constructing an orthogonal version of the estimating equation and using *cross-fitting* to average out the learning biases.

In this paper, we develop novel DML procedures for the causal analysis of a repeatedly measured treatment (or exposure) using non-linear panel data models. We extend Robinson (1988)’s partially linear regression model to static panel data models with additive noise and fixed effects. The particular family of estimation problems we consider is for causal inference about the homoge-

neous effect of a repeatedly measured treatment with potentially many (irrelevant) control variables, non-linearities in the regressors (e.g., trigonometric and exponential functions, and complex interaction structures): hence, the need for ML tools. We focus on homogeneous effects but note that the method generalises to heterogeneous treatment effects provided the analyst is prepared to specify a parametric treatment effect model. However, the focus on homogeneous effects is justified by the contribution to practice because linear static panel models with homogeneous causal effects are so widely used in applied microeconomics and the wider social and health sciences.

We use synthetic data to assess our DML method in terms of bias, root mean squared error, variability, and sampling distribution of the estimated causal effect. We contrast the simulation results of conventional least squares (LS) estimator with those obtained using DML based on LASSO, CART and RF. We use the estimates of the Oracle estimator (i.e., when the true functional form is known *a priori*) as benchmark. We find there are gains from the use of DML with flexible learners when the data generating process involves a non-linear discontinuous function of the regressors, but LS estimates *can* outperform ML when the data generating process is linear (which is expected) but also if it is non-linear but smooth and excludes interactions. Finally, we illustrate the applicability of DML with observational panel data by reanalysing part of the study by [Fazio and Reggiani \(2023\)](#) on the effect of the introduction national minimum wage in the UK on voting for conservative parties.

The remainder of the paper is structured as follows. Section 2 provides an overview of the literature and our contribution to it. Section 3 introduces the reader to the panel data model for causal analysis and the estimators. In section 4, we discuss the estimation procedure based on [Chernozhukov et al. \(2018\)](#)’s DML procedure. In section 5, we describe the Monte Carlo simulation design and discuss the main results. Section 6 illustrate an empirical application of the procedure. Finally, Section 7 concludes.

2 Contribution to the Literature

There is a growing body of econometrics/statistical literature on causal inference with ML. One strand focuses on building or modifying existing learners to consistently estimate and make inferences about causal effects (e.g., [Athey and Imbens, 2016](#); [Wager and Athey, 2018](#); [Athey et al., 2019](#); [Künzel et al., 2019](#); [Lechner, 2019](#); [Lechner and Mareckova, 2022](#); [Di Francesco, 2022, 2023](#)). Another strand focuses on incorporating ML into traditional statistical estimators – e.g., LS, generalised method of moments (GMM), maximum likelihood – to estimate causal effects more accurately (e.g., [Belloni et al., 2016](#); [Chernozhukov et al., 2018](#); [Nie and Wager, 2021](#); [Chernozhukov et al., 2022](#)).

This paper falls into the second strand. [Belloni et al. \(2016\)](#) provides two-step procedures for panel data with additive individual-specific heterogeneity that first select the potential control variables to be included in the final model through LASSO, and then estimate the homogeneous treatment effect with LS. The authors rely on linear combinations of control variables to approximate the unknown nuisance functions, transforming [Robinson \(1988\)](#)’s partially linear regression model with additive noise and fixed effects into a conventional linear panel model with fixed effects and a

high-dimensional set of confounders. With the Post-Cluster-LASSO and the Post-Cluster-LASSO IV, [Belloni et al. \(2016\)](#) set the grounds for estimation and inference of using Double/Debiased ML (DML) ([Chernozhukov et al., 2018](#)).

The theory for DML is very general but no panel examples are considered by [Chernozhukov et al. \(2018\)](#). In the DML framework, [Semenova et al. \(2023\)](#) and [Klosin and Vilgalys \(2022\)](#) are among the few to have extended the applicability of DML to different panel data settings, respectively, for dynamic panel data models and panel data models with fixed effects and continuous treatments. Both implementations consider inference on heterogeneous treatment effects and are restricted to the context of linear penalized regression (i.e., LASSO) requiring the researcher to specify a dictionary of non-linear terms. Our contribution to this literature is to implement DML for panel models with additive unobserved individual heterogeneity (or fixed effects) that are (a) widely used in applied research across applied economics and the quantitative social and health sciences and (b) develop an approach general enough to allow analysts to choose *any* suitable learner for the functional form of their regressors. In other words, we do not exclusively rely on an *ex-ante* choice of non-linear function transformations of variables or their interactions (i.e., by having to specify a sufficiently rich ‘dictionary’ of non-linear terms as required by LASSO), but allow the use of any learner (e.g., tree-based) to capture the unknown functional form.¹

In this context, both the fixed effects and non-linearity present a number of challenges. We propose two ways of handling these challenges. The first is an intuitive and pragmatic approach in which either the within-group (WG) or first-difference (FD) transformation are applied to the data to remove the fixed effect, and ML applied to the transformed data to learn the resulting nuisance function. We call this the *approximation* approach because it relies on being able to approximate the true function with non-linear function of the transformed regressors. We also propose ‘exact’ approaches based on the correlated random effects (CRE) model proposed by [Mundlak \(1978\)](#) in which the individual heterogeneity term is specified to follow a linear model: while less robust than correlated random effects estimators in the linear case, this approach forms the basis of both the exact WG and FD estimators, which we refer to as *hybrid* approaches, and a CRE estimator for the partially linear panel model.

Modern advances in ML algorithms for causal analysis have found empirical applicability in labour economics (e.g., [Davis and Heller, 2017](#); [Lechner, 2019](#); [Knaus et al., 2022](#); [Cengiz et al., 2022](#)), health economics (e.g., [Heiler and Knaus, 2021](#); [Di Francesco, 2022](#)), and environmental economics (e.g., [Klosin and Vilgalys, 2022](#); [Stetter et al., 2022](#)). We contribute to the causal literature by leveraging the power of ML for policy evaluation. The potential value of ML over conventional methods for causal analysis has already been explored for Difference-in-Differences (DID), randomised control trials (RCTs), and quantile regression by [Baiardi and Naghi \(2021\)](#), [Knaus \(2022\)](#), and [Strittmatter \(2023\)](#). [Baiardi and Naghi \(2021\)](#) revisit various empirical studies with causal machine learning methods (DML models for the average treatment effects, and Generalised Random Forests for heterogeneous treatment effects) to understand whether the researcher benefits from using new ML methods for causal analysis over traditional estimators. [Knaus \(2022\)](#) provides a review, extension and application of various DML-based methods from the perspective of

¹The R package allows the use of learners available in `mlr3`, `mlr3learners`, `mlr3extralearners`.

a researcher interested in standard programme evaluation under unconfoundedness. They provide a comprehensive investigation to estimate the effect of four programmes of the Swiss Active Labour Market Policy on employment. [Strittmatter \(2023\)](#) revisits the effects of Connecticut’s Jobs First welfare experiment on the labor supply by comparing conditional average treatment effects from DML with quantile treatment effects. Along this line, we aim to understand the applicability of DML with observational panel data over standard methods, such as DID. We hence re-evaluate [Fazio and Reggiani \(2023\)](#)’s study on the effect of NMW in UK comparing their DID results with DML, using the British Household Panel Survey (BHPS). Our main contribution here is to apply the DML method to investigate whether it can produce substantively different conclusions.

3 The Model and Estimators

3.1 Econometric Background

Observational panel data are longitudinal survey studies that collect repeated measures of the survey variables from randomly sampled units from a population (e.g., households, workers, firms) in more than one wave. Some examples are the Panel Study of Income Dynamics (PSID) for the US, the UK Household Longitudinal Study (UKHLS) for the UK, EU labour force survey (EULFS) for European countries. One of the main issues of observational panel data is that the sample is subject to attrition over time because the respondent may drop out the study due to refusal to participate, migration or death, which leads to non-random selection. Once these complications have been accounted for, panel data present researchers with opportunities for more robust identification strategies for causal effect estimation than offered by ‘cross-sectional’ studies (taken to include longitudinal studies involving an initial measure of the treatment and another of the outcome after a follow-up period) by exploiting within-individual variation over time.

Suppose the panel study design is to collect information on each of N individuals at each of the T time periods, or waves. To simplify notation, we assume a balanced panel with observed data on every individual at all T waves.² Let $\{y_{it}, d_{it}, \mathbf{x}_{it} : t = 1, \dots, T\}$ be independent and identically distributed (*iid*) random vectors for individuals $i = 1, \dots, N$, where y_{it} is the outcome (or independent variable), d_{it} a continuous or binary treatment variable (or intervention), and \mathbf{x}_{it} a $1 \times p$ vector of regressors, usually including the constant term, used to adjust for non-random selection. For continuous d_{it} , if $d_{it} \geq 0$ we presume a dose-response relationship with $d_{it} = 0$ indicating null treatment; otherwise, d_{it} is taken to be centred around its mean μ_D such that $d_{it} \equiv d_{it} - \mu_D$. For binary $d_{it} \in \{0, 1\}$, $d_{it} = 0$ is taken to indicate the absence and $d_{it} = 1$ the presence of treatment.

The challenge is to use these data to estimate the causal/treatment effect of exposure/treatment d_{ti} on outcome y_{ti} using confounding variables \mathbf{x}_{it} to adjust for non-random treatment selection. We consider approaches based on constructing a consistent estimating equation, or score, $\boldsymbol{\psi}(\boldsymbol{\theta}, \boldsymbol{\eta}) = N^{-1} \sum_i \boldsymbol{\psi}_i(\boldsymbol{\theta}, \boldsymbol{\eta})$ satisfying $\mathbb{E}[\boldsymbol{\psi}_i(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0)] = \mathbf{0}$, where $\boldsymbol{\theta}_0$ and $\boldsymbol{\eta}_0$ are true values and we wish to make inferences about $\boldsymbol{\theta}_0$ given a suitable estimate of nuisance parameter $\boldsymbol{\eta}$. Generally, stage one involves obtaining an estimate $\hat{\boldsymbol{\eta}}$, and stage two solving $\boldsymbol{\psi}(\boldsymbol{\theta}, \hat{\boldsymbol{\eta}}) = \mathbf{0}$ to obtain $\hat{\boldsymbol{\theta}}_N = \hat{\boldsymbol{\theta}}(\hat{\boldsymbol{\eta}})$.

²The estimation problem and results hold with unbalanced panel with appropriate modifications in the notation.

We consider problems where $\boldsymbol{\eta}_0$ comprises distinct $\boldsymbol{\eta}_{0i} = \boldsymbol{\eta}_0(\mathbf{x}_{1i}, \dots, \mathbf{x}_{iT})$ for each individual so stage one requires the analyst to model $\boldsymbol{\eta}_0$.

Provided that a finite-dimensional family of parametric models for $\boldsymbol{\eta}_0$ can be found such that $\hat{\boldsymbol{\eta}}$ is consistent, we can (under regularity conditions) rely on $\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0 = o_p(N^{-1/2})$ and hence standard first-order asymptotic theory for inference about θ_0 . However, our interest lies in problems where this cannot be guaranteed because of either the high-dimensionality of \mathbf{x}_{it} or the nuisance parameter potentially having an unknown non-linear form (or both). ML is appropriate because there is no substantive interest in inference about the nuisance parameters. Hence, it is proposed to learn $\boldsymbol{\eta}_0$ using ML. Let $\mathbb{I} = \mathbb{I}(h)$ indicate the choice of learner with its hyperparameter h , and $\hat{\boldsymbol{\eta}}_n = \hat{\boldsymbol{\eta}}(\mathbb{I})$ the resulting prediction of $\boldsymbol{\eta}_0$, where $n = \mathcal{O}(N)$ is the size of the training sample. The premise is that, while we can rely on $\hat{\theta}_N(\hat{\boldsymbol{\eta}}_n) - \theta_0 = o_p(1)$ provided that \mathbb{I} and h are carefully chosen, *we cannot rely on $\hat{\theta}_N(\hat{\boldsymbol{\eta}}_n) - \theta_0 = o_p(N^{-1/2})$* so the usual first-order asymptotic results for $\hat{\theta}_N$ do not hold.

Chernozhukov et al. (2018) proposed DML for constructing estimators for θ_0 that converge at the required \sqrt{N} rate even if ML is used to estimate the nuisance parameters. There are two key components of DML: first is the construction of a consistent *Neyman orthogonal* score $\boldsymbol{\psi}^\perp(\theta, \boldsymbol{\eta})$ in the sense that its derivative (however defined) with respect to $\boldsymbol{\eta}$ at the truth is zero (Chernozhukov et al., 2018, Definition 2.1). This property implies that the covariance matrix derived assuming $\boldsymbol{\eta}_0$ is known is correct for situations where it has to be estimated, that is, we can simply plug in $\hat{\boldsymbol{\eta}}_n$. Neyman orthogonality facilitates the effectiveness of the second key component of DML: K -fold data splitting to control the impact of finite-sample bias in $\hat{\boldsymbol{\eta}}_n$ by averaging over the parameter estimates obtained from using each split. The resulting DML estimator is consistent provided the learner(s) \mathbb{I} provides a good approximation of the true function, which in regular problems boils down to requiring that the predictions from learner(s) \mathbb{I} converge at a rate at least $N^{1/4}$, although this is not straightforward to justify for all ML algorithms. Standard errors can then be consistently estimated using the sandwich estimator proposed by Chernozhukov et al. (2018, Theorems 3.1 and 3.2).

Before proceeding further with the discussion, we clarify the key vocabulary used herein. We say that the causal parameter θ_0 is *estimated* because we use the method of moments to retrieve its effect and conduct statistical inference. Conversely, the nuisance parameters $\boldsymbol{\eta}_0 = (l_0, m_0)$ defined in the next section are *learnt* because no statistical inference is conducted and ML tools are used only to capture complex functional structures in the data.

3.2 Model under the Fixed-effects Assumption

In the context of causal analysis, the repeated measures available from panel data potentially allow the analyst to relax the *selection on observables* assumption if non-random exposure selection depends on latent individual heterogeneity α_i taken to be fixed for the duration of the panel. Before describing the panel estimation problem, we set out the following assumptions which must be satisfied by the underlying data generating process for the target parameter to have a causal interpretation:

ASM.1 *Strict exogeneity* such that $\mathbb{E}(y_{it}|\mathbf{d}_i, X_i, \alpha_i) = \mathbb{E}(y_{it}|d_{it}, \mathbf{x}_{it}, \alpha_i)$

ASM.2 *Selection on observables and individual heterogeneity:* $y_{it}(\cdot) \perp\!\!\!\perp d_{it} \mid \mathbf{x}_{it}, \alpha_i$.

ASM.3 *Linearity and homogeneity of the causal effects:* $\mathbb{E}\{y_{it}(d) - y_{it}(0) \mid \mathbf{x}_{it}, \alpha_i\} = d\theta_0$

ASM.4 *Nuisance parameters:* $\mathbb{E}(y_{it} \mid \mathbf{x}_{it}, \alpha_i) = l_0(\mathbf{x}_{it}, \alpha_i)$ and $\mathbb{E}(d_{it} \mid \mathbf{x}_{it}, \alpha_i) = m_0(\mathbf{x}_{it}, \alpha_i)$

ASM.5 *Fixed effects α_i :* $\mathbb{E}(\alpha_i \mid d_{it}, \mathbf{x}_{it}) \neq 0$,

where $\mathbf{d}_i = (d_{i1}, \dots, d_{iT})$, $X_i = (\mathbf{x}'_{i1}, \dots, \mathbf{x}'_{iT})$ and $y_{it}(\cdot) = \{y_{it}(d) : d \in \Omega_D\}$ is set of potential outcomes $y_{it}(d)$, that is, the realisation of the outcome for individual i at wave t were the treatment level set to d .

Under assumptions, [ASM.1-ASM.5](#), standard treatment effect arguments lead to the non-linear additive noise model

$$\begin{aligned} y_{it} &= v_{it}\theta_0 + l_0(\mathbf{x}_{it}, \alpha_i) + u_{it} \\ v_{it} &= d_{it} - m_0(\mathbf{x}_{it}, \alpha_i) \end{aligned} \tag{1}$$

where $\mathbb{E}(u_{it} \mid v_{it}, \mathbf{x}_{it}, \alpha_i) = \mathbb{E}(v_{it} \mid \mathbf{x}_{it}, \alpha_i) = 0$. This is an extension of the ‘partialling out’ (PO) *partially linear regression* (PLR) model to panel data with treatment d_{it} replaced by treatment equation residual v_{it} . There is an alternative formulation based on $y_{it} = d_{it}\theta_0 + g_0(\mathbf{x}_{it}, \alpha_i) + u_{it}$, where $g_0(\mathbf{x}_{it}, \alpha_i) = E\{y_{it}(0) \mid \mathbf{x}_{it}, \alpha_i\}$, called the ‘instrumental variable’ (IV) PLR because, in the cross-sectional case where $\alpha_i = 0$, it leads to an IV-style estimator for θ_0 . The first equation in model (1) is the structural (or output) equation, and the second treatment equation is the residual of a linear model for treatment selection. Both components are required to construct a Neyman orthogonal score function, but this PLR model presents an unfeasible learning problem because α_i is unobserved.

To derive a feasible final model, we must make the following assumption concerning the unobserved heterogeneity:

ASM.6 *Additive separability:* $l_0(\mathbf{x}_{it}, \alpha_i) = l_0(\mathbf{x}_{it}) + \alpha_i$ and $m_0(\mathbf{x}_{it}, \alpha_i) = m_0(\mathbf{x}_{it}) + c_i$

Note that α_i and $c_i = c(\alpha_i)$ are generally correlated because $\mathbb{E}(\alpha_i) = \mathbb{E}(c_i) = 0$ but $E(\alpha_i c_i) \neq 0$. Then combining assumptions [ASM.1-ASM.6](#) leads finally to the PO PLR panel model

$$\begin{aligned} y_{it} &= v_{it}\theta_0 + l_0(\mathbf{x}_{it}) + \alpha_i + u_{it} \\ v_{it} &= d_{it} - m_0(\mathbf{x}_{it}) - c_i, \end{aligned} \tag{2}$$

which is a feasible learning problem. We note that it is possible to extend this model to relax assumption [ASM.3](#) and estimate heterogeneous treatment effects,³ but we do not explore the performance of such an estimator here. However, the Neyman orthogonal score for such a model is outlined in [Appendix C](#), and we discuss effect heterogeneity further on in [Section 7](#).

Model (2) has been considered at length for cross-sectional cases where a major advantage of the PO formulation is that l_0 and m_0 can be learnt directly from the observed data ([Chernozhukov](#)

³Heterogeneous causal effects can be estimated if the analyst is prepared to specify f to index dose-response d and interactions (d, \mathbf{x}_{it}) and (d, t) such that $\mathbb{E}\{y_{it}(d) - y_{it}(0) \mid \mathbf{x}_{it}, \alpha_i\} = f_{\theta_0}(d; \mathbf{x}_{it}, t)$. A PO PLR would also require evaluating $v_{it} = f_{\theta_0}(d_{it}; \mathbf{x}_{it}, t) - E\{f_{\theta_0}(d_{it}; \mathbf{x}_{it}, t) \mid \mathbf{x}_{it}, \alpha_i\}$, but this would be based on the same treatment equation residual in (2) were f curvilinear or otherwise such that $\mathbb{E}\{f_{\theta_0}(d_{it}; \mathbf{x}_{it}, t) \mid \mathbf{x}_{it}, \alpha_i\} = f_{\theta_0}\{m_0(\mathbf{x}_{it}, \alpha_i); \mathbf{x}_{it}, t\}$.

et al., 2018, p. C33). However, for panel data under assumption [ASM.5](#), this is not generally true because only $l_0^*(\mathbf{x}_{it}) = l_0(\mathbf{x}_{it}) + \mathbb{E}(\alpha_i|\mathbf{x}_{it})$ and $m_0^*(\mathbf{x}_{it}) = m_0(\mathbf{x}_{it}) + \mathbb{E}(c_i|\mathbf{x}_{it})$ can be learnt from the available data. Parameter estimation is therefore relatively straightforward under the random effects assumption $\mathbb{E}(\alpha_i|\mathbf{x}_{it}) = 0$, but we presume analysts generally do not believe causal inference is credible under it.⁴ As such, the potential presence of fixed effects unobserved heterogeneity correlated with \mathbf{x}_{it} presents a significant challenge when it comes to constructing a consistent estimator.

In the following sections, we set out three alternative estimators for panel data models based on the within-group (WG), first-difference (FD), and correlated random effects (CRE) estimators used for linear panel models. The three estimators are consistent under the fixed effects assumption but are recommended to be used in specific frameworks ([Cameron and Trivedi, 2005](#)). That is, WG is more efficient when there is no serial correlation, and inconsistent with lagged-dependent variables; FD more efficient with serial correlation and consistent with lagged-dependent variables; CRE is preferred with many time-invariant variables.

3.3 Correlated Random Effects Estimation

Correlated Random Effects (CRE) estimators based on [Mundlak \(1978\)](#) involve modifying (2) to include explicitly the correlation between individual heterogeneity term α_i and predictor variables d_{it} and \mathbf{x}_{it} in the model. Below, we develop a CRE for PLR panel model (2) using this approach.

Were the analyst to know $l_{it} = l_0(\mathbf{x}_{it})$ then a CRE estimator could be based straightforwardly on structural model $y_{it} - l_{it} = d_{it}\theta_0 + \alpha_i + u_{it}$ rather than (2) by fitting the reduced-form model obtained by a) expanding $\alpha_i = \bar{\mathbf{z}}_i\boldsymbol{\pi}_0 + a_i$, where $\bar{\mathbf{z}}_i = T^{-1} \sum_{t=1}^T \mathbf{z}_{it}$ is the vector of individual-specific means of $\mathbf{z}_{it} = (\mathbf{x}_{it}, d_{it})$ and $\boldsymbol{\pi}_0$ is the coefficient of the linear projection of α_i onto the span of (mean-centred) $\bar{\mathbf{z}}_i$, and b) exploiting the orthogonality of $a_i + u_{it}$ and $\bar{\mathbf{z}}_i$ ([Wooldridge, 2010](#), sec. 2.3). This approach would be robust to α_i being non-linear in $\bar{\mathbf{z}}_i$ provided the linear projection exists under the data generating process.

Knowledge of l_{it} is unrealistic but, even were only the parametric form of $l_0(\mathbf{x}_{it})$ known, the approach above can only be used if l_0 is linear such that $l_0(\mathbf{x}_{it}) = \mathbf{x}_{it}\mathbf{l}_0$. This is because the orthogonality of a_i and $\bar{\mathbf{z}}_i$ also implies orthogonality of a_i and \mathbf{z}_{it} , which allows joint estimation of θ_0 , \mathbf{l}_0 and $\boldsymbol{\pi}_0$ based on $y_{it} = \mathbf{x}_{it}\mathbf{l}_0 + d_{it}\theta_0 + \bar{\mathbf{z}}_i\boldsymbol{\pi}_0 + a_i + u_{it}$. However, linear projections cannot guarantee the orthogonality of non-linear $l_0(\mathbf{x}_{it})$ and α_i , so stronger assumptions about α_i are needed. This also affects things when we wish to learn l_0 and m_0 without specifying the functional form of either and, from the discussion in the last section, we also know that the fixed effects assumptions on α_i and c_i confound learning l_0 and m_0 directly from the observed data.

To overcome both of these challenges, we propose the following Mundlak-like *modelling assumption* to induce a reduced-form random effects model:

$$\text{ASM.7 Mundlak model: } \alpha_i = \{\bar{\mathbf{x}}_i - \mathbb{E}(\bar{\mathbf{x}}_i)\}\boldsymbol{\pi}_0^* + a_i \text{ such that } \mathbb{E}(a_i|X_i) = 0.$$

Note that $\boldsymbol{\pi}_0^*$ is a model parameter and not simply the coefficient of the linear projection of α_i onto the span of X_i (and $\bar{\mathbf{x}}_i$ is explicitly mean centred to emphasise that $\mathbb{E}(\alpha_i) = 0$); that the

⁴Under a random effects assumption, we note that [Sela and Simonoff \(2012\)](#) developed an algorithm for using tree-based learners to estimate non-causal partially linear regression models.

residual of the Mundlak-type model is conditional on X_i follows from Chamberlain's generalization of Mundlak for α_i to depend on $\mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}$, or X_i , but with the wave-specific coefficients of each \mathbf{x}_{it} constrained to be equal (Wooldridge, 2010, Section 11.3.2).

Assumption [ASM.7](#) leads to the following constraints on causal model (1):

$$l_0(\mathbf{x}_{it}, \alpha_i) = l_0(\mathbf{x}_{it}, \bar{\mathbf{x}}_i, a_i) \text{ and } m_0(\mathbf{x}_{it}, \alpha_i) = m_0(\mathbf{x}_{it}, \bar{\mathbf{x}}_i, a_i). \quad (3)$$

Then in place of [ASM.6](#) we make the following assumption about (3):

ASM.8 *Additively separable*: $l_0(\mathbf{x}_{it}, \bar{\mathbf{x}}_i, a_i) = l_0(\mathbf{x}_{it}, \bar{\mathbf{x}}_i) + a_i$ and $m_0(\mathbf{x}_{it}, \bar{\mathbf{x}}_i, a_i) = m_0(\mathbf{x}_{it}, \bar{\mathbf{x}}_i) + \lambda a_i$.⁵

Combining the assumptions above leads to

$$\begin{aligned} y_{it} &= v_{it}\theta_0 + l_0(\mathbf{x}_{it}, \bar{\mathbf{x}}_i) + r_{it} \\ v_{it} &= d_{it} - m_0(\mathbf{x}_{it}, \bar{\mathbf{x}}_i) - \lambda a_i, \end{aligned} \quad (4)$$

where $r_{it} = a_i + u_{it}$ satisfies $\mathbb{E}(r_{it} | v_{it}, \mathbf{x}_{it}, \bar{\mathbf{x}}_i) = 0$. The feasible learning tasks here are, therefore, to learn $l_0(\mathbf{x}_{it}, \bar{\mathbf{x}}_i)$ from the data $\{y_{it}, \mathbf{x}_{it}, \bar{\mathbf{x}}_i\}_{t=1}^T$, and also to predict the residuals v_{it} to plug into the structural equation. While learning l_0 from the sample data is straightforward, there are different ways to obtain predictions of v_{it} . The first way is to learn $m_0(\mathbf{x}_{it}, \bar{\mathbf{x}}_i)$ from the data $\{d_{it}, \mathbf{x}_{it}, \bar{\mathbf{x}}_i\}_{t=1}^T$, and save the residuals $\hat{v}_{it} = d_{it} - \hat{m}(\mathbf{x}_{it}, \bar{\mathbf{x}}_i)$. However, this ignores individual heterogeneity $c_i = \lambda a_i$ when predicting v_{it} , so we favour a second way, namely, learning $m_0^*(\mathbf{x}_{it}, \bar{\mathbf{x}}_i, \bar{d}_i)$ from the data $\{d_{it}, \mathbf{x}_{it}, \bar{\mathbf{x}}_i, \bar{d}_i\}_{t=1}^T$ to obtain $\hat{v}_{it} = d_{it} - \hat{m}(\mathbf{x}_{it}, \bar{\mathbf{x}}_i, \bar{d}_i)$. This improves accuracy by predicting c_i as well as m_0 .⁶

3.4 Within-Group Estimation

Alternative approaches in the (partially) linear context involve modelling transformations of the observed data that effectively condition out the problematic fixed effect α_i . The within-group (WG) approach involves modelling the time-demeaned (individual mean-centred) outcomes $\ddot{y}_{it} = y_{it} - T^{-1} \sum_{t=1}^T y_{it} = y_{it} - \bar{y}_i$ and treatments $\ddot{d}_{it} = d_{it} - \bar{d}_i$, which under PLR panel model (2) implies

$$\begin{aligned} \ddot{y}_{it} &= \ddot{d}_{it}\theta_0 + \ddot{l}_0(\mathbf{x}_{it}) + \ddot{u}_{it}, \\ \ddot{d}_{it} &= \ddot{m}_0(\mathbf{x}_{it}) + \ddot{v}_{it}, \end{aligned} \quad (5)$$

where

$$\ddot{l}_0(\mathbf{x}_{it}) = l_0(\mathbf{x}_{it}) - T^{-1} \sum_{t=1}^T l_0(\mathbf{x}_{it}) \text{ and } \ddot{m}_0(\mathbf{x}_{it}) = m_0(\mathbf{x}_{it}) - T^{-1} \sum_{t=1}^T m_0(\mathbf{x}_{it})$$

⁵Specifying $c_i = \lambda a_i$ is a without loss of generality simplification that effectively ensures the two random effects are perfectly correlated but have distinct variances.

⁶The justification of the second way comes from the case where d_{i1}, \dots, d_{iT} are multivariate normal (given X_i and $\bar{\mathbf{x}}_i$) with respective conditional means $m_0(\mathbf{x}_{i1}, \bar{\mathbf{x}}_i), \dots, m_0(\mathbf{x}_{iT}, \bar{\mathbf{x}}_i)$ and common homoskedastic variance $\sigma_u^2 + \sigma_a^2$ and covariance σ_a^2 , where $\sigma_a^2 = \text{var}(a_i | \mathbf{x}_{it})$ and $\sigma_u^2 = \text{var}(u_{it} | \mathbf{x}_{it})$. Elementary calculations for multivariate normal distributions give $\mathbb{E}(d_{it} | \mathbf{x}_{it}, \bar{\mathbf{x}}_i, \bar{d}_i) = m_0(\mathbf{x}_{it}, \bar{\mathbf{x}}_i) + \bar{d}_i - \bar{m}_0(\mathbf{x}_{it}, \bar{\mathbf{x}}_i) \approx m_0(\mathbf{x}_{it}, \bar{\mathbf{x}}_i) + c_i$ where $\bar{m}_0(\mathbf{x}_{it}, \bar{\mathbf{x}}_i) = T^{-1} \sum_{t=1}^T m_0(\mathbf{x}_{it}, \bar{\mathbf{x}}_i)$. So the problem is to learn $m_0^*(\mathbf{x}_{it}, \bar{\mathbf{x}}_i, \bar{d}_i) \equiv m_0(\mathbf{x}_{it}, \bar{\mathbf{x}}_i) + \bar{d}_i - \bar{m}_0(\mathbf{x}_{it}, \bar{\mathbf{x}}_i)$: the role of $\bar{m}_0(\mathbf{x}_{it}, \bar{\mathbf{x}}_i)$ is to constrain $\sum_s m_0(\mathbf{x}_{is}, \bar{\mathbf{x}}_i, \bar{d}_i) = \bar{d}_i$ but this constraint is treated as implicit to be picked up by the learner.

are generally functions of X_i (that is, \mathbf{x}_{it} and also $\mathbf{x}_{i1}, \dots, \mathbf{x}_{it-1}, \mathbf{x}_{it+1}, \dots, \mathbf{x}_{iT}$) but not $\ddot{\mathbf{x}}_{it} = \mathbf{x}_{it} - T^{-1} \sum_{t=1}^T \mathbf{x}_{it}$. In the linear case, where $l_0(\mathbf{x}_{it}) = \mathbf{x}_{it} \mathbf{l}_0$, $m_0(\mathbf{x}_{it}) = \mathbf{x}_{it} \mathbf{m}_0$ and both \mathbf{l}_0 and \mathbf{m}_0 are conformable vectors of regression coefficients, the learning problem is simply to estimate θ_0 by regressing the transformed outcome on the transformed treatment and transformed predictors. When \ddot{l}_0 and \ddot{m}_0 are non-linear, however, standard learning becomes more difficult because both functions depend on X_i .

Approximate approach: The first practicable approach, inspired by the linear case above, is based on the following approximation:

$$\ddot{l}_0(\mathbf{x}_{it}) = \tilde{l}_0(\ddot{\mathbf{x}}_{it}) + \epsilon_l^{it} \approx \tilde{l}_0(\ddot{\mathbf{x}}_{it}) \text{ and } \ddot{m}_0(\mathbf{x}_{it}) = \tilde{m}_0(\ddot{\mathbf{x}}_{it}) + \epsilon_m^{it} \approx \tilde{m}_0(\ddot{\mathbf{x}}_{it}), \quad (6)$$

where $\tilde{l}_0(\ddot{\mathbf{x}}_{it})$ and $\tilde{m}_0(\ddot{\mathbf{x}}_{it})$ can be learnt from the time-demeaned sample data $\{\ddot{y}_{it}, \ddot{d}_{it}, \ddot{\mathbf{x}}_{it}\}_{t=1}^T$. This approach relies on the approximation errors $\epsilon_l^{it} = \epsilon_l(\mathbf{x}_{it})$ and $\epsilon_m^{it} = \epsilon_m(\mathbf{x}_{it})$ being small, which will be the case if the true functions are linear or the approximation is accurate over regions of the predictors with the strongest support.⁷

Exact approach: The possibility of the approximation above performing poorly motivates an alternative exact approach. We call this the *hybrid* approach because it incorporates the first-stage CRE estimator from Section 3.3 to learn $l_0(\mathbf{x}_{it})$ and $m_0(\mathbf{x}_{it})$ at each wave and then combines these predictions to learn (5).

To derive the hybrid estimator, we first recall that the additive separability assumption [ASM.6](#) is $l_0(\mathbf{x}_{it}, \alpha_i) = \mathbb{E}(y_{it} | \mathbf{x}_{it}, \alpha_i) = l(\mathbf{x}_{it}) + \alpha_i$. Then, under Mundlak model [ASM.7](#),

$$l(\mathbf{x}_{it}) + \alpha_i = l(\mathbf{x}_{it}) + \bar{\mathbf{x}}_i \boldsymbol{\pi}_0 + a_i \equiv l(\mathbf{x}_{it}, \bar{\mathbf{x}}_i) + a_i,$$

where $l(\mathbf{x}_{it}, \bar{\mathbf{x}}_i) = l(\mathbf{x}_{it}) + \bar{\mathbf{x}}_i \boldsymbol{\pi}_0$ can be learnt because $\mathbb{E}(a_i | \mathbf{x}_{it}, \bar{\mathbf{x}}_i) = 0$ holds. Hence, we first obtain $\hat{l}(\mathbf{x}_{it}, \bar{\mathbf{x}}_i)$ from the data on $(y_{it}, \mathbf{x}_{it}, \bar{\mathbf{x}}_i)$, and then estimate WG model (5) as follows:

$$\widehat{\ddot{l}(\mathbf{x}_{it})} = \hat{l}(\mathbf{x}_{it}, \bar{\mathbf{x}}_i) - \frac{1}{T} \sum_{t=1}^T \hat{l}(\mathbf{x}_{it}, \bar{\mathbf{x}}_i), \quad (7)$$

with similarly defined $\widehat{\ddot{m}(\mathbf{x}_{it})}$ estimated using the \hat{m} obtained using sample data $\{d_{it}, \mathbf{x}_{it}, \bar{\mathbf{x}}_i, \bar{d}_i\}$ as discussed in Section 3.3; then the quasi-oracle estimating equations are given by

$$\begin{aligned} \ddot{y}_{it} &= \ddot{d}_{it} \theta + \widehat{\ddot{l}(\mathbf{x}_{it})} + \ddot{u}_{it} \\ \ddot{d}_{it} &= \widehat{\ddot{m}(\mathbf{x}_{it})} + \ddot{v}_{it}. \end{aligned} \quad (8)$$

⁷The existence of such a function is easily justified: a first-order Taylor series expansion of l_0 around some fixed value \mathbf{x} gives $l_0(\mathbf{x}_{it}) = (\mathbf{x}_{it} - \mathbf{x}) \dot{l}_0(\mathbf{x}) + \mathcal{O}(\|\mathbf{x}_{it} - \mathbf{x}\|^2)$, where $\|\cdot\|$ is the L_1 -norm and column-vector $\dot{l}_0(\mathbf{x})$ is the partial derivative of l_0 with respect to \mathbf{x}_{it} evaluated at \mathbf{x} , so that $l_0(\mathbf{x}_{it}) = \ddot{\mathbf{x}}_{it} \dot{l}_0(\mathbf{x}) + \mathcal{O}(\|\mathbf{b}_x\|^2)$, where $\mathbf{b}_x = \sup \|\mathbf{x}_{it} - \mathbf{x}\|$. Hence, there is some $\ddot{\mathbf{x}}_{it} \dot{l}_0(\mathbf{x}) + \mathcal{O}(\|\mathbf{b}\|^2) \approx \ddot{\mathbf{x}}_{it} \dot{l}_0(\mathbf{x})$ minimizing some loss function, where $\bar{\mathbf{b}} = \inf_{\mathbf{x}} \mathbf{b}_x$ and $\bar{\mathbf{x}} = \operatorname{arginf}_{\mathbf{x}} \mathbf{b}_x$ are respectively the smallest bound over the bounded support of all possible \mathbf{x} -centred confounders, and $\bar{\mathbf{x}}$ any value obtaining this bound.

3.5 First-Difference Estimation

The first-difference (FD) estimator is based on the transformation $\Delta y_{it} = y_{it} - y_{it-1}$ and $\Delta d_{it} = d_{it} - d_{it-1}$, which like the WG transformation removes the individual heterogeneity term such that

$$\begin{aligned}\Delta y_{it} &= \Delta v_{it}\theta + \Delta l_0(\mathbf{x}_{it}) + \Delta u_{it} \\ \Delta d_{it} &= \Delta m_0(\mathbf{x}_{it}) + \Delta v_{it},\end{aligned}\tag{9}$$

for $t = 2, \dots, T$.⁸ The first-differenced nuisance parameters are

$$\Delta l_0(\mathbf{x}_{it}) = l_0(\mathbf{x}_{it}) - l_0(\mathbf{x}_{it-1}) \quad \text{and} \quad \Delta m_0(\mathbf{x}_{it}) = m_0(\mathbf{x}_{it}) - m_0(\mathbf{x}_{it-1}),$$

which are generally functions of \mathbf{x}_{it-1} and \mathbf{x}_{it} . When the nuisance parameters are linear, i.e., $l_0(\mathbf{x}_{it}) = \mathbf{x}_{it}\mathbf{l}_0$ and $m_0(\mathbf{x}_{it}) = \mathbf{x}_{it}\mathbf{m}_0$, then $\Delta l_0(\mathbf{x}_{it}) = \Delta \mathbf{x}_{it}\mathbf{l}_0$ and $\Delta m_0(\mathbf{x}_{it}) = \Delta \mathbf{x}_{it}\mathbf{m}_0$ for $t = 2, \dots, T$. Approximate and hybrid estimators are motivated and obtained in a similar fashion to those set out in Section 3.4. The former requires the existence of an approximation such that

$$\Delta l_0(\mathbf{x}_{it}) = \widetilde{\Delta l_0}(\Delta \mathbf{x}_{it}) + \varepsilon_l^{it} \approx \widetilde{\Delta l_0}(\Delta \mathbf{x}_{it}) \quad \text{and} \quad \Delta m_0(\mathbf{x}_{it}) = \widetilde{\Delta m_0}(\Delta \mathbf{x}_{it}) + \varepsilon_m^{it} \approx \widetilde{\Delta m_0}(\Delta \mathbf{x}_{it}), \tag{10}$$

where $\widetilde{\Delta l_0}(\Delta \mathbf{x}_{it})$ and $\widetilde{\Delta m_0}(\Delta \mathbf{x}_{it})$ are approximations of the nuisance functions to be learnt from the sample data $(\Delta y_{it}, \Delta d_{it}, \Delta \mathbf{x}_{it})$. The approximation errors ε_l^{it} and ε_m^{it} must again be small. The hybrid estimator, equivalent to that for the hybrid WG estimator from Section 3.4, again uses the CRE estimator from Section 3.3 to learn the nuisance parameters, from which a quasi-oracle for (9) is constructed using $\{y_{it}, d_{it}, \widehat{l}(\mathbf{x}_{it}, \bar{\mathbf{x}}_i), \widehat{m}(\mathbf{x}_{it}, \bar{\mathbf{x}}_i, \bar{d}_i)\}$ from which inferences about θ_0 can be made.

4 Estimation and Inference

We now set out the DML procedure for estimating θ_0 . Denote the sample units by $\mathcal{W} = \{1, \dots, N\}$. For sample unit $i \in \mathcal{W}$, we potentially observe $W_i = \{W_{it} : t = 1, \dots, T\}$, where $W_{it} = w\{y_{it}, d_{it}, \mathbf{x}_{it}\}$ and w is a transformation of the data (possibly the identity) chosen by the analyst to implement one of the estimators from Section 3.

The first component of DML is the Neyman-orthogonal score function on which to base estimation. The derivation of this score follows Chernozhukov et al. (2018, Section 2.2.2) (see Appendix C for an outline). We first need to define a generic score function for the three panel data estimators as the product of the error terms, i.e.,

$$\boldsymbol{\psi}^\perp(W_i; \theta, \boldsymbol{\eta}) = \mathbf{v}_i \Sigma_0^{-1} \mathbf{u}_i, \tag{11}$$

where \mathbf{u}_i is a column vector of structural residuals and \mathbf{v}_i a row vector of treatment residuals based on one of the models for correlated random effects (4), within-group (5) or first-difference estimator; and Σ_0 is a conformable variance-matrix for \mathbf{u}_i .

The second component of DML is sample splitting, which involves randomly partitioning

⁸The model does not depend on α_i but auto-correlation induced by Δu_{it-1} and Δu_{it} having u_{it-1} in common should be accounted for in variance estimation.

the individual sample units into K equi-sized *folds*. Denote the units in fold $k = 1, \dots, K$ by $\mathcal{W}_k \subset \mathcal{W}$ and let \mathcal{W}_k^c be its complement such that $N_k \equiv |\mathcal{W}_k| = N/K$, $|\mathcal{W}_k^c| = N - N_k$ and, because the folds are mutually exclusive and exhaustive, $\mathcal{W}_k \cap \mathcal{W}_j = \mathcal{W}_k \cap \mathcal{W}_k^c = \emptyset$ and $\mathcal{W}_k \cup \mathcal{W}_k^c = \mathcal{W}_1 \cup \dots \cup \mathcal{W}_K = \mathcal{W}$. Let $\boldsymbol{\eta} = (l, m)$ be the vector of nuisance parameters with population value $\boldsymbol{\eta}_0 = (l_0, m_0)$. For $K > 2$, the larger complementary \mathcal{W}_k^c is used to learn the potentially complex $\boldsymbol{\eta}$, and \mathcal{W}_k for the relatively simple task of estimating θ_0 . ML is used to learn the nuisance parameters from the data from the units in complementary \mathcal{W}_k^c and the learnt prediction rule denoted by $\hat{\boldsymbol{\eta}}_k$. This procedure is repeated for each fold.

The DML estimator $\hat{\theta}$ is then the solution to

$$\frac{1}{N_k} \sum_{k=1}^K \sum_{i \in \mathcal{W}_k} \psi^\perp(W_i; \theta, \hat{\boldsymbol{\eta}}_k) = \mathbf{0}, \quad (12)$$

where $\hat{\boldsymbol{\eta}}_k$ is used to predict the nuisance parameters for the units in fold \mathcal{W}_k . The final estimated causal parameter is the median across the k -folds.

Rather than estimate Σ_0 in (11) using a two-step procedure, we set it to equal the identity matrix and estimate a heteroskedasticity and cluster-robust variance-covariance matrix for $\hat{\theta}$ as follows: for fold k , estimate

$$\hat{\sigma}_k^2 = \hat{J}_k^{-1} \left\{ \frac{1}{N_k} \sum_{i \in \mathcal{W}_k} \psi^\perp(W_i; \theta, \hat{\boldsymbol{\eta}}_k) \psi^\perp(W_i; \theta, \hat{\boldsymbol{\eta}}_k)' \right\} \hat{J}_k^{-1}$$

where $\hat{J}_k = N_k^{-1} \sum_{i \in \mathcal{W}_k} \sum_t v_{it}^2$ and $\psi^\perp(W_i; \theta, \hat{\boldsymbol{\eta}}_k)'$ is the transpose of $\psi^\perp(W_i; \theta, \hat{\boldsymbol{\eta}}_k)$. The final variance of the causal parameter, $\hat{\sigma}^2$, is the median variance across the k -folds plus a finite-sample correction, $(\hat{\theta}_k - \hat{\theta}_{median})^2$, weighted by the number of units in the cluster to account for the variation introduced by sampling splitting (Chernozhukov et al., 2018, p. C30).

5 Monte Carlo Simulation

5.1 Simulation Design

To assess the performance of the ML-driven estimators defined above, we generate data under variations of the following PLR panel model:

$$y_{it} = d_{it}\theta + l_0(\mathbf{x}_{it}) + \alpha_i + u_{it} \quad (13)$$

$$d_{it} = m_0(\mathbf{x}_{it}) + c_i + v_{it} \quad (14)$$

$$\alpha_i = 0.25 \left(\frac{1}{T} \sum_{t=1}^T d_{it} - \bar{d} \right) + 0.25 \frac{1}{T} \sum_{t=1}^T \mathbf{x}_{it,k} + a_i, \text{ for } k = \{1, 3\} \quad (15)$$

$$a_i \sim N(0, 0.95), \mathbf{x}_{it} \sim N(0, 5), c_i \sim N(0, 1). \quad (16)$$

where α_i is the fixed effect modelled as the Mundlak (1978)'s device, a_i and c_i are random effects.

We consider three alternative designs for the nuisance parameters m_0 and l_0 that vary in the level of non-linearity and non-smoothness of the functional forms.

Design 1 (DGP1): Linear in the nuisance parameters

$$m_0 = \frac{1}{4}\mathbf{x}_{it,1} + \mathbf{x}_{it,3}$$

$$l_0 = \frac{1}{4}\mathbf{x}_{it,1} + \mathbf{x}_{it,3}$$

Design 2 (DGP2): Non-linear and smooth in the nuisance parameters

$$m_0 = \cos(\mathbf{x}_{it,1}) + \frac{1}{4} \frac{\exp(\mathbf{x}_{it,3})}{1 + \exp(\mathbf{x}_{it,3})}$$

$$l_0 = \frac{\exp(\mathbf{x}_{it,1})}{1 + \exp(\mathbf{x}_{it,1})} + \frac{1}{4} \cos(\mathbf{x}_{it,3})$$

Design 3 (DGP3): Non-linear and discontinuous in the nuisance parameters

$$m_0 = \frac{1}{4} (\mathbf{x}_{it,1} \cdot \mathbb{1}[\mathbf{x}_{it,1} > 0]) + \frac{1}{2} (\mathbf{x}_{it,1} \cdot \mathbf{x}_{it,3})$$

$$l_0 = \frac{1}{2} (\mathbf{x}_{it,1} \cdot \mathbf{x}_{it,3}) + \frac{1}{4} (\mathbf{x}_{it,3} \cdot \mathbb{1}[\mathbf{x}_{it,3} > 0]),$$

where $\mathbb{1}(z) = 1$ if z is true otherwise $\mathbb{1}(z) = 0$. A visual representation of the functional forms of the nuisance parameters under the three designs is in Figure 1 that plots l_0 and m_0 over the variables x_1 (on the left) and x_3 (on the right) of each graph while setting the other variable equal to zero. The sample of cross-sectional observations is $N=1,000,000$ with $T = 10$ from which we sample while conducting the Monte Carlo simulations.

The nuisance parameters are learnt with LASSO, CART, and RF.⁹ The hyperparameters of the ML learner are tuned in each Monte Carlo simulation via a grid search (Bergstra and Bengio, 2012) over specified parameter values, where five distinct values per hyperparameter are randomly selected by the algorithm within each evaluation (more details on the tuning algorithm are provided in Section D). LASSO uses the penalisation parameter, λ , equivalent to minimum mean cross-validated error. CART and RF choose the optimal hyperparameters with grid search (see Table 1 for a summary). The tuned hyperparameters of CART are the complexity parameter, minimum number of observations in terminal node, maximum depth of any node of the final tree. The tuned hyperparameters of RF are the number of trees, minimum number of observations in terminal node, maximum depth of any node of the final tree; fixed hyperparameters are the number of covariates randomly sampled to split at each node (set as the maximum number), and the importance criterion.

The number of original variables is $p = 30$ but only two of these (x_1 and x_3) are relevant with the rest noise, as shown in the three designs. LASSO with the extended dictionary uses a design matrix augmented with polynomials of order three and interaction terms of all the included regressors. For WG/FD-hybrid and CRE estimators, the total number of variables is $2p$ because the individual-specific means are included as outlined by Mundlak (1978). The number of waves is $T = 10$ throughout but the number of individuals varies $N = \{100, 1000, 4000\}$ to compare finite-sample performance for small, medium and large sample sizes.¹⁰ To reduce the computational

⁹The R packages we used are `cv_glmnet` for LASSO, `rpart` for CART, and `ranger` for RF.

¹⁰We originally run simulations for $N = 10,000$, but the results are on average similar to those for $N = 4,000$. Because tuning RF with $N = 10,000$ requires considerable computational time, we decided to rely on $N = 4,000$ for large

time, each size- N Monte Carlo sample replication is drawn randomly with replacement from a pseudo-population of 1,000,000 individuals. Each simulation run is based on $R = 100$ Monte Carlo replications.¹¹

In all simulations, we use the Neyman orthogonal partialling-out (PO) score following Chernozhukov et al. (2018)’s algorithm 2 and five-fold cross-fitting. Cross-fitting samples individuals so that each cross-sectional unit i along with its full time series is assigned to a unique fold k .

5.2 Simulation Results

Monte Carlo simulation results are displayed for each estimator and learner (LASSO, CART, RF). The figures also allow us to contrast the DML results with those obtained using conventional Ordinary Least Square (OLS) estimation for each estimator (WG, FD, and CRE).

Figures 2-4 report the average bias, root mean squared error (RMSE), and the ratio of the standard errors (SE) and the standard deviation (SD) of the estimated causal parameter $\hat{\theta}_N$.¹² The red dots correspond to linear DGP1, the black diamonds to non-linear smooth DGP2, and the blue triangles to non-linear and discontinuous DGP3. Figures 3-4 display the results for the approximation approach on the top, and for the hybrid approach at the bottom.

Under DGP1 and DGP2, the approximate WG/FD estimators have small bias and accurate SE (the ratio of SE to SD is close to 1) for DML and OLS. However, under non-linear discontinuous DGP3, the performance of the approximations is poor (as is that of the linearity-based learners) with severe upward bias of the causal parameter and substantial under-estimation of SD. Conversely, this is not the case for the CRE and WG/FD-hybrid estimators: DML leads to considerable bias reduction even for DGP3 using both tree-based approaches and LASSO with extended dictionary. The picture is not perfect because DML-RF is biased for $N = 100$, and SE is downward biased for the tree-based learners for all three sample sizes. The best performance comes from the extended-dictionary LASSO, which allows valid inference by having small bias and a ratio of SE to SD very close to one.

The sampling distributions of $\hat{\theta}_N$ for CRE obtained using tree-based learners are displayed in Figure 5 for $N = 1,000$.¹³ We observe for DGP3 that the sampling distributions are highly non-normal. However, this is less severe under DGP1 and DGP2 where the estimated causal effects are close to normally distributed and SE bias is smaller. This indicates that statistical inference about θ_0 is unreliable using tree-based algorithms because the assumption of asymptotic normality does not hold.

A possible explanation for the non-normality of the sampling distributions is suboptimal hyperparameter tuning of the tree-based algorithms. The importance of optimal hyperparameter tuning for causal modelling has recently been shown for conditional average treatment effect estimators (Machlanski et al., 2023). In particular, we were concerned that we had not tuned over

sample behaviour. From the perspective on an applied researcher, such large data sets are include administrative data or long-running longitudinal studies where most of the respondents are kept in the analysis.

¹¹We run only 100 replications because tuning ML algorithms requires considerable computational time.

¹²SD denotes the standard deviation of the estimator sampling distribution and SE its estimate.

¹³This is a sufficiently large sample to observe the asymptotic behaviour. Similar patterns of over-dispersed distributions are observed for $N = 4,000$ but not for $N = 100$. We do not display those for layout reasons. Their behaviour deteriorates in large samples.

a sufficiently wide range of values in the grid search or taken into account the adaptive nature of optimal hyperparameter choice. For example, [Wager and Walther \(2015, Theorem 1\)](#) show the rate at which the minimum number of observations per leaf for “moderately high-dimensional” cases should increase with N to control the error bounds on the resulting estimates.

To explore this possibility, we used an alternative strategy for hyperparameter tuning for RF. This strategy is based on the hypothesis that the previous strategy led to regularisation-like bias due to under-fitted forests. In short, it involves fixing the maximum depth to 100, building a forest of 1,000 trees, and tuning the minimum node size as in the main simulations (see Table 1). Figure 5 compares the sampling distributions of $\hat{\theta}_N$ using the strategy from Section 5.1 (solid line) with the new ‘partially tuned’ strategy (dashed line). The new strategy forces each random tree to overfit the data and relies on a large forest to average out the overfitting errors. This leads to estimators with larger upward biases (in DGP1 and DGP3 for the WG/FD estimators and all DGPs for CRE) but smaller SDs. This is especially true for the non-linear discontinuous DGP3 where the new strategy leads to a clearly Gaussian normal sampling distribution. However, the new strategy is seen to be unsuccessful, with the results for DGP3 indicating that the analyst must choose between bias and variance when using tree-based methods.

6 Empirical Application

We illustrate the applicability of DML for panel data models with fixed effects by replicating the analysis in [Fazio and Reggiani \(2023\)](#) on voting behaviour after the introduction of the National Minimum Wage (NMW) in the UK in 1999.¹⁴

The study uses the British Household Panel Survey (BHPS), which is a longitudinal survey study for British households running from 1991 until 2009.¹⁵ The survey contains a question asking whether the interviewed individual was ‘*paid the minimum wage in 1999*’. The treated group includes those who have replied affirmatively to this question, which is interacted with an indicator equal to one for waves 9 onward (after the implementation of NMW in 1999) to construct the treatment variable.

Part of the original study estimates the average treatment effect (homogeneous treatment) of NMW on *voting for conservative parties* with OLS. We revisit Specification (2) of Table 5 of [Fazio and Reggiani \(2023\)](#) with DML for partially linear regression models with different base learners (LASSO, CART, RF). The estimating equations for the PO version of the model are

$$Vote_{it} = v_{it}\theta + l(\mathbf{x}_{it}) + \alpha_i + u_{it} \quad (17)$$

$$v_{it} = NMW_{it} - m(\mathbf{x}_{it}) - c_i \quad (18)$$

where $Vote_{it}$ is a dichotomous variable equal to one if the respondent voted for a conservative party in wave t , and zero otherwise; $NMW_{it} = NMW_i \times Post_t$ is the treatment variable, with NMW_i switching to one if the respondent’s hourly pay increased due to the introduction of the NMW,

¹⁴The scope of this replication exercise is not intended to confirm or invalidate their results, but only to show the use of DML with observational panel data.

¹⁵The data can be requested and downloaded from UK Data Service (ukdataservice.ac.uk) upon registration in the platform.

and $Post_t$ taking value of one from wave 9 onward (with the introduction of the NMW) and zero otherwise. Base control variables \mathbf{x}_{it} are the inputs of nuisance functions l and m , and include: age, age squared (not for CART and RF),¹⁶ education, marital status, household size, income of other members, and their individual means. LASSO with the extended dictionary includes non-linear terms of the control variables (i.e., polynomials of order three and interaction terms). Summary statistics of base control variables are shown in Table 2.

Estimation results are displayed in Tables 3 and 4. Table 3 does not include wave and region fixed effects to reflect the DGPs used in our Monte Carlo simulations. In detail, Column (1) displays OLS estimates based on the original specification in Fazio and Reggiani (2023), and the remaining columns show the results obtained with DML using different learners. Table 4 includes wave and region fixed effects as in the original specification. Column (1) corresponds to the original OLS estimates, Column (2) adds the interaction between wave and region fixed effects to Column (1), and the remainder show DML results. We show the causal effects with both the approximation and hybrid WG estimator (respectively, at the top and bottom panels of the tables). The hybrid approach includes the individual means of the control variables (and fixed effects), as required by Mundlak (1978)’s device for CRE. Standard errors (in parenthesis) are clustered at the individual level. Optimal (tuned) hyperparameters used for CART and RF are reported in Table 5.

The estimated causal effects of NMW are similar across (a) estimation approaches (hybrid and approximation) and (b) base learners (LASSO, CART, and RF) when compared with those obtained using OLS in Table 3. This may suggest that the underlying DGP is linear or has similar properties to DGP2 from our simulation study. The causal effect estimates lie between 0.091 and 0.103, with the DML-CART estimates the smallest (equal to 0.091). The extended dictionary LASSO returns the largest effect when used with the approximation approach (equal to 0.103). All DML estimates are statistically significant at 1% level.

With the inclusion of wave and region fixed effects, the estimated causal effects in Table 4 are smaller for every estimator/base-learner combination but differ more widely than before from the OLS baseline in Column (1). This may suggest that the true functional form has properties more similar to DGP3 than before, and that these are being picked up by the learners. More specifically, while LASSO without the extended dictionary, CART and RF produce treatment effects close to the original OLS estimate, this is not observed for LASSO with extended dictionary. That is, LASSO with the extended dictionary estimates smaller effects (0.086 with WG-approximation and 0.079 with WG-hybrid) that are also closer to LS estimates with the interaction between wave and regional fixed effect (included in the extended dictionary) (0.088). This could suggest that the fixed effects are informative in capturing unobserved factors that correlate with both *NMW* and *voting for conservative parties*; it may also signal that other learners are not capturing complex structures of interactions.

In general, it seems that LASSO with dictionary is providing less biased estimates and (from the simulation study) more reliable standard errors, but this requires that the user generates polynomials and interactions of all variables – including the binary variables for the fixed effects. All estimates are statistically significant at least at 5% level.

¹⁶Control variables used in CART and RF do not include polynomial terms because tree-based approaches are non-parametric algorithms that should be able to find interactions between variables and non-linearities in the data.

7 Discussion

DML is already realising the great potential it has for the social sciences, and particularly for leveraging the power of ML for robust estimation of policy-intervention effects. Although the theory underpinning DML is very general, applications of it to panel data have been rare; two notable examples are [Klosin and Vilgalys \(2022\)](#) and [Semenova et al. \(2023\)](#). In this paper, we developed novel DML procedures for estimating intervention effects from panel studies. These procedures are based on a simple extension of the partially linear regression model to panel data. We proposed three estimators – within-group, first-differences, and correlated random effects – that account for the presence of unobserved individual heterogeneity that is potentially correlated with the regressors. For the within-group and first-differences estimators, we proposed two alternative approaches called *approximation* and *hybrid*: the former was found to perform well when the nuisance functions were linear or non-linear and smooth without interactions, while the latter was more robust and performed best in terms of bias when the nuisance functions were non-linear and discontinuous. This is in line with other work showing that the final causal parameter estimate may be adversely affected by the functional forms of the nuisance parameters learnt during the first stage of estimation ([Rudolph et al., 2023](#), Section S1.1 in the Supplementary material).

Our implementation of the DML method is general and widely applicable because it can be used with any ML algorithm (e.g., regression trees and random forest) and not only with statistical learners like LASSO. More negatively, we found that tree-based algorithms require considerable attention from the analyst particularly with regards to hyperparameter tuning to control the bias-variance trade-off. This is in line with recent work emphasising the sensitivity of ML algorithms for causal analysis to hyperparameter choice ([Machlanski et al., 2023](#)). LASSO with an extended dictionary performed the best in terms of bias reduction and statistical inference. Hyperparameter tuning is less challenging for LASSO because it has only one hyperparameter. The disadvantage of using LASSO is that the extended dictionary, specified by the analyst without knowledge of the true functional form of the nuisance parameters, must be rich enough to capture the truth and so can become computationally demanding in terms of memory when $p \gg 30$. Trees, in contrast, do not require the analyst to guess the composition of the dictionary. However, while the focus of our study was to investigate the performance of two widely used families of learner, our findings emphasise the importance of following the widely used practice in ML *not* to rely on one base learner but to use *ensembles* comprising multiple learners (e.g., boosting, stacked learners, and super learners) because these usually outperform single base learners ([Valentini and Masulli, 2002](#)).

The results of our empirical analysis were that DML made no substantive difference to the conclusions of the original study, but the re-analysis can be viewed as a robustness check for non-linearity. However, our simulation study showed that the bias due to incorrectly assuming linearity can be substantial, and that DML can correct for it. Hence, DML has value as an estimation technique in its own right, *and* as a robustness check in analyses where it may be more convenient to report the results obtained using linear models, which are more familiar and easier to understand.

Finally, we note that it is relatively simple to extend our method to allow for treatment heterogeneity by estimating *conditional* average treatment effects rather than average treatment

effects *if* the analyst is prepared to specify a parametric model for the heterogeneity in terms of the regressors and across time. However, if the analyst wishes to use ML to learn the heterogeneity function then another method is needed. Further work will extend the *quasi-oracle* approach of Nie and Wager (2021) to static panel models, and further to dynamic models. This would complement the LASSO-based dynamic approach with heterogeneous effects proposed by Semenova et al. (2023) to a more general DML estimator which can be used as part of a conventional ML ensemble.

References

- Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360.
- Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, 47(2):1148 – 1178.
- Baiardi, A. and Naghi, A. A. (2021). The value added of machine learning to causal inference: Evidence from revisited studies. *arXiv preprint arXiv:2101.00878*.
- Belloni, A., Chernozhukov, V., Hansen, C., and Kozbur, D. (2016). Inference in high-dimensional panel models with an application to gun control. *Journal of Business & Economic Statistics*, 34(4):590–605.
- Bergstra, J. and Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(2).
- Cameron, A. C. and Trivedi, P. K. (2005). *Microeconometrics: Methods and Applications*. Cambridge University Press.
- Cengiz, D., Dube, A., Lindner, A., and Zentler-Munro, D. (2022). Seeing beyond the trees: Using machine learning to estimate the impact of minimum wages on labor market outcomes. *Journal of Labor Economics*, 40(S1):S203–S247.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.
- Chernozhukov, V., Newey, W. K., and Singh, R. (2022). Automatic debiased machine learning of causal and structural effects. *Econometrica*, 90(3):967–1027.
- Davis, J. M. and Heller, S. B. (2017). Using causal forests to predict treatment heterogeneity: An application to summer jobs. *American Economic Review*, 107(5):546–550.
- Di Francesco, R. (2022). Aggregation trees. *CEIS Research Paper*, 546.
- Di Francesco, R. (2023). Ordered correlation forest. *arXiv preprint arXiv:2309.08755*.
- Fazio, A. and Reggiani, T. (2023). Minimum wage and tolerance for high incomes. *European Economic Review*, 155:104445.
- Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.
- Heiler, P. and Knaus, M. C. (2021). Effect or treatment heterogeneity? Policy evaluation with aggregated and disaggregated treatments. *arXiv preprint arXiv:2110.01427*.

- Klosin, S. and Vilgalys, M. (2022). Estimating continuous treatment effects in panel data using machine learning with an agricultural application. *arXiv preprint arXiv:2207.08789*.
- Knaus, M. C. (2022). Double machine learning-based programme evaluation under unconfoundedness. *The Econometrics Journal*, 25(3):602–627.
- Knaus, M. C., Lechner, M., and Strittmatter, A. (2022). Heterogeneous employment effects of job search programs: A machine learning approach. *Journal of Human Resources*, 57(2):597–636.
- Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165.
- Lechner, M. (2019). Modified causal forests for estimating heterogeneous causal effects. *arXiv preprint arXiv:1812.09487*.
- Lechner, M. and Mareckova, J. (2022). Modified causal forest. *arXiv preprint arXiv:2209.03744*.
- Lechner, M. and Okasa, G. (2019). Random forest estimation of the ordered choice model. *arXiv preprint arXiv:1907.02436*.
- Machlanski, D., Samothrakis, S., and Clarke, P. (2023). Hyperparameter tuning and model evaluation in causal effect estimation. *arXiv preprint arXiv:2303.01412*.
- Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica*, pages 69–85.
- Nie, X. and Wager, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, pages 299–319.
- Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, pages 931–954.
- Rudolph, K. E., Williams, N. T., Miles, C. H., Antonelli, J., and Diaz, I. (2023). All models are wrong, but which are useful? comparing parametric and nonparametric estimation of causal effects in finite samples. *Journal of Causal Inference*, 11(1):20230022.
- Sela, R. J. and Simonoff, J. S. (2012). Re-em trees: a data mining approach for longitudinal and clustered data. *Machine learning*, 86:169–207.
- Semenova, V., Goldman, M., Chernozhukov, V., and Taddy, M. (2023). Inference on heterogeneous treatment effects in high-dimensional dynamic panels under weak dependence. *Quantitative Economics*, 14(2):471–510.
- Stetter, C., Mennig, P., and Sauer, J. (2022). Using machine learning to identify heterogeneous impacts of agri-environment schemes in the eu: a case study. *European Review of Agricultural Economics*, 49(4):723–759.
- Strittmatter, A. (2023). What is the value added by using causal machine learning methods in a welfare experiment evaluation? *Labour Economics*, 84:102412.
- University of Essex, Institute for Social and Economic Research (2018). British Household Panel Survey: Waves 1-18, 1991-2009. [data collection]. 8th Edition. UK Data Service. SN: 5151, DOI: <http://doi.org/10.5255/UKDA-SN-5151-2>.
- Valentini, G. and Masulli, F. (2002). Ensembles of learning machines. In *Neural Nets: 13th Italian Workshop on Neural Nets, WIRN VIETRI 2002 Vietri sul Mare, Italy, May 30–June 1, 2002 Revised Papers 13*, pages 3–20. Springer.

- Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.
- Wager, S. and Walther, G. (2015). Adaptive concentration of regression trees, with application to random forests. *arXiv preprint arXiv:1503.06388*.
- Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.

A Tables

Table 1. *Hyperparameter tuning*

| Learner | Hyperparameters | Value/interval | Description |
|---------|-----------------|------------------|---|
| Lasso | lambda.min | – | λ equivalent to minimum mean cross-validated error |
| CART | cp | {0.01,0.02} | Prune all nodes with a complexity less than cp from the printout. |
| | minbucket | {5,ceiling(N/2)} | Minimum number of observations in any terminal <leaf> node. |
| | maxdepth | {1,10} | Maximum depth of any node of the final tree. |
| RF | num.trees | {5,100} | Number of trees in the forest. |
| | min.node.size | {5,ceiling(N/2)} | Minimal node size to split at. |
| | max.depth | {1,10} | Maximum depth of any node of the final tree. |
| | mtry | p | The number of covariates, randomly sampled, to split at each node. |
| | importance | impurity | The ‘impurity’ measure is the Gini index for classification, the variance of the responses for regression and the sum of test statistics. |

Note: Hyperparameter tuning for CART and RF is conducted with a random grid search. For RF, nodes with size smaller than min.node.size can occur.

Table 2. *Summary statistics*

| | Mean | SD | Min | Max |
|-------------------|---------|--------|-------|-------|
| Vote Conservative | 0.102 | 0.303 | 0 | 1 |
| NMW | 0.020 | 0.138 | 0 | 1 |
| HH income | 8.082 | 1.312 | -1.32 | 11.46 |
| HH size | 3.145 | 1.213 | 1 | 14 |
| Age | 35.914 | 10.930 | 18 | 65 |
| Age squared | 1,409.3 | 838.9 | 324 | 4,225 |
| Degree | 0.102 | 0.303 | 0 | 1 |
| Married | 0.534 | 0.499 | 0 | 1 |
| Observations | 19,961 | | | |
| No. groups | 4,927 | | | |

Table 3. Replication Results without fixed effects

| | OLS (1) | DML-Lasso (2) | DML-Lasso (3) | DML-CART (4) | DML-RF (5) |
|--|--------------------|-------------------------------|---------------------|---------------------|---------------------|
| <i>Dependent variable: "Vote conservative"</i> | | | | | |
| | | <i>Approximation approach</i> | | | |
| NMW | 0.101** (0.044) | 0.099*** (0.045) | 0.103*** (0.044) | 0.091*** (0.044) | 0.099*** (0.044) |
| | | <i>Hybrid approach</i> | | | |
| NMW | | 0.100*** (0.044) | 0.100*** (0.045) | 0.091*** (0.044) | 0.101*** (0.045) |
| Extended dictionary | No | No | Yes | No | No |
| No. Observations | 19,961 | 19,961 | 19,961 | 19,961 | 19,961 |
| No. Groups | 4,927 | 4,927 | 4,927 | 4,927 | 4,927 |
| <i>Resampling Information</i> | | | | | |
| Estimator | WG | WG | WG | WG | WG |
| No. folds | — | 5 | 5 | 5 | 5 |
| Cross-fitting | — | Yes | Yes | Yes | Yes |
| Score | — | PO | PO | PO | PO |
| DML algorithm | — | 2 | 2 | 2 | 2 |

Note: Column (1) reports the least squares estimates based on Specification (2) in Table 5 in [Fazio and Reggiani \(2023\)](#) without wave and region fixed effects; remaining columns use DML with different learners. Base control variables include: age, age squared (not for CART and RF), education, marital status, household size, income of other members, and their individual means. Column (3) uses an extended dictionary of non-linear terms of the control variables (i.e., polynomials of order three and interactions of the control variables and their individual means). The hybrid approach includes the individual means of the control variables and fixed effects. Standard errors (in parenthesis) are clustered at the individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4. Replication of Specification (2) in Table 5 in [Fazio and Reggiani \(2023\)](#)

| | OLS (1) | OLS (2) | DML-Lasso (3) | DML-Lasso (4) | DML-CART (5) | DML-RF (6) |
|--|--------------------|--------------------|-------------------------------|---------------------|---------------------|---------------------|
| <i>Dependent variable: "Vote conservative"</i> | | | | | | |
| | | | <i>Approximation approach</i> | | | |
| NMW | 0.097** (0.045) | 0.088** (0.045) | 0.095** (0.045) | 0.086*** (0.045) | 0.091*** (0.044) | 0.098*** (0.045) |
| | | | <i>Hybrid approach</i> | | | |
| NMW | | | 0.093*** (0.045) | 0.079** (0.045) | 0.091** (0.042) | 0.095*** (0.048) |
| Wave FE | Yes | Yes | Yes | Yes | Yes | Yes |
| Region FE | Yes | Yes | Yes | Yes | Yes | Yes |
| Wave x Region FE | No | Yes | No | Yes | No | No |
| Extended dictionary | No | No | No | Yes | No | No |
| No. Observations | 19,961 | 19,961 | 19,961 | 19,961 | 19,961 | 19,961 |
| No. Groups | 4,927 | 4,927 | 4,927 | 4,927 | 4,927 | 4,927 |
| <i>Resampling Information</i> | | | | | | |
| Estimator | WG | WG | WG | WG | WG | WG |
| No. folds | — | — | 5 | 5 | 5 | 5 |
| Cross-fitting | — | — | Yes | Yes | Yes | Yes |
| Score | — | — | PO | PO | PO | PO |
| DML algorithm | — | — | 2 | 2 | 2 | 2 |

Note: Column (1) reports the original figures of Specification (2) in Table 5 in [Fazio and Reggiani \(2023\)](#) estimated using least squares; Column (2) adds the interaction between wave and region fixed effects to Column (1); remaining columns use DML with different learners. Base control variables include: age, age squared (not for CART and RF), education, marital status, household size, income of other members. Column (4) uses an extended dictionary of non-linear terms of the control variables and fixed effects (i.e., polynomials of order three and interactions of the control variables). The hybrid approach includes the individual means of the control variables and fixed effects. Standard errors (in parenthesis) are clustered at the individual level. Standard errors (in parenthesis) are clustered at the individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

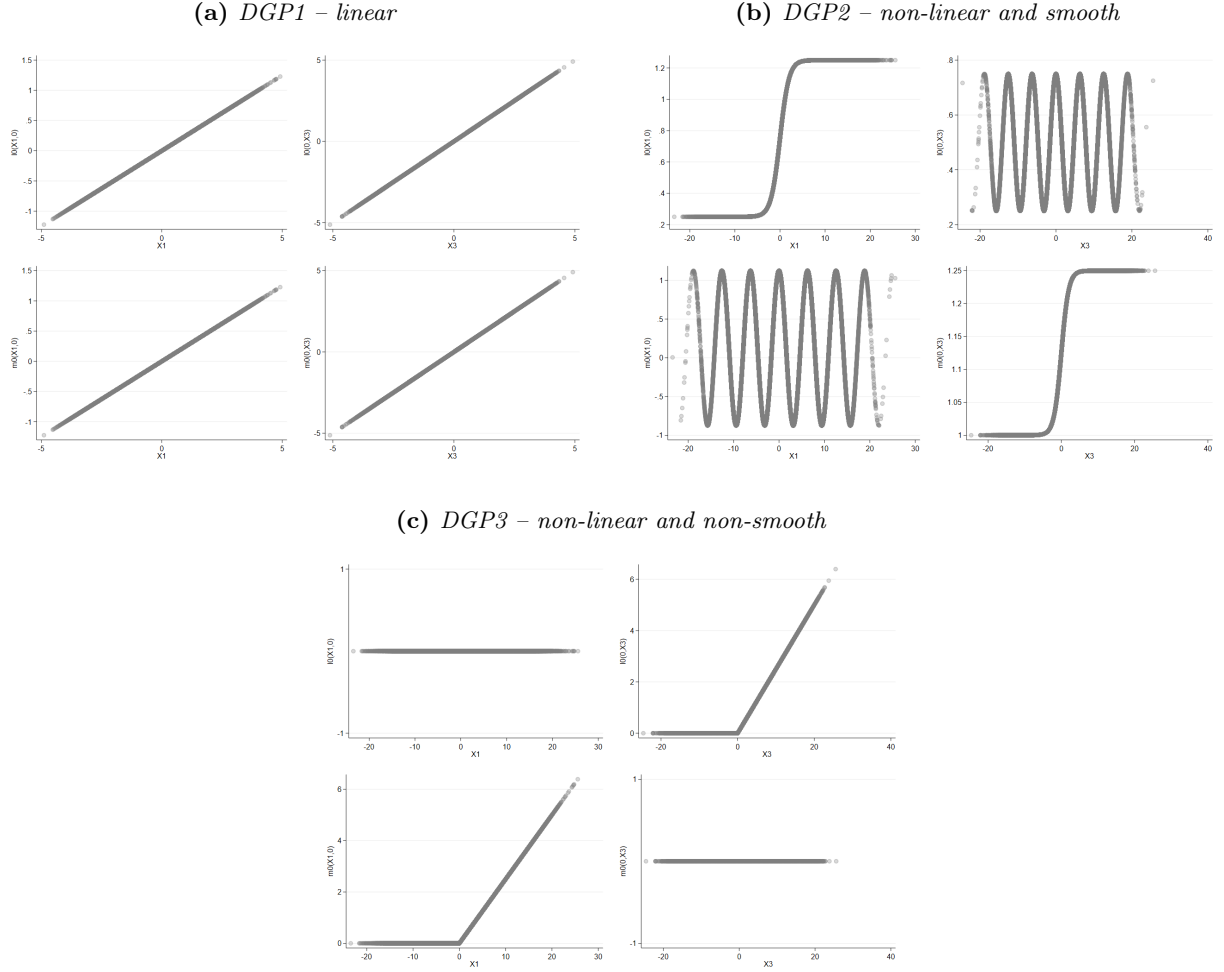
Table 5. Hyperparameter tuning

| Hyperparamters | Table(3) | | Table(4) | |
|--|----------------|----------------|----------------|----------------|
| | Approx | Hybrid | Approx | Hybrid |
| <i>Base learner for nuisance functions: CART</i> | | | | |
| cp* | {0.043, 0.027} | {0.41, 0.41} | {0.044, 0.027} | {0.41, 0.41} |
| minbucket* | {2464, 1688} | {5, 551} | {2464, 1688} | {5, 551} |
| maxdepth* | {4, 3} | {10, 7} | {4, 3} | {10, 7} |
| <i>Base learner for nuisance functions: RF</i> | | | | |
| num.trees | 1000 | 1000 | 1000 | 1000 |
| min.node.size* | {1558, 1170} | {134, 134} | {134, 1429} | {134, 134} |
| max.depth* | {79, 6} | {22, 37} | {37, 85} | {22, 37} |
| mtry | { $p + 1, p$ } | { $p + 1, p$ } | { $p + 1, p$ } | { $p + 1, p$ } |
| importance | impurity | impurity | impurity | impurity |

Note: * indicates the hyperparameters that are tuned with grid search with options (`n_evals` = 10, `resolution` = 20). The optimal value for `cp` is chosen from the interval {0.01, 0.05}; `minbucket*` from {5, $\text{ceiling}(N/2)$ }; `maxdepth*` from {1, 30}; `min.node.size` from the interval {5, $\text{ceiling}(N/2)$ }; `max.depth` from the interval {1, 100}. The hyperparameters in brackets refer to the (optimal) chosen value for m and l , respectively.

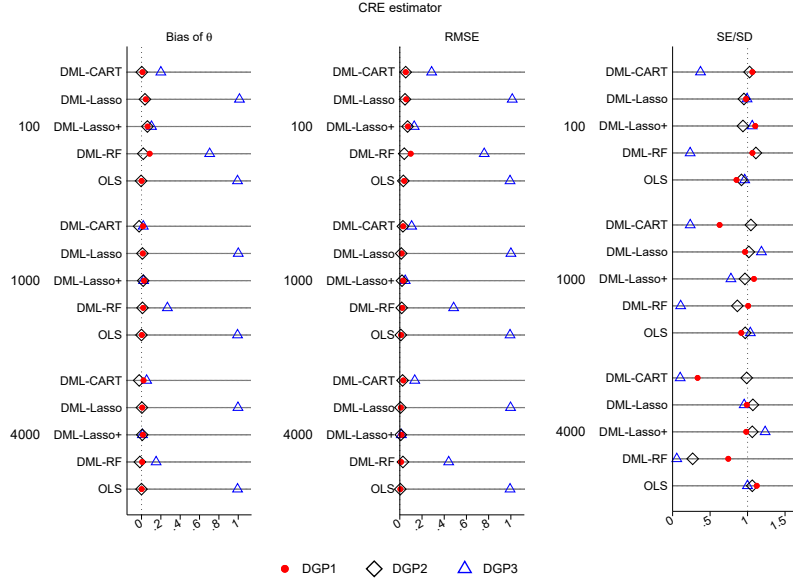
B Figures

Figure 1. Functional form of the nuisance functions



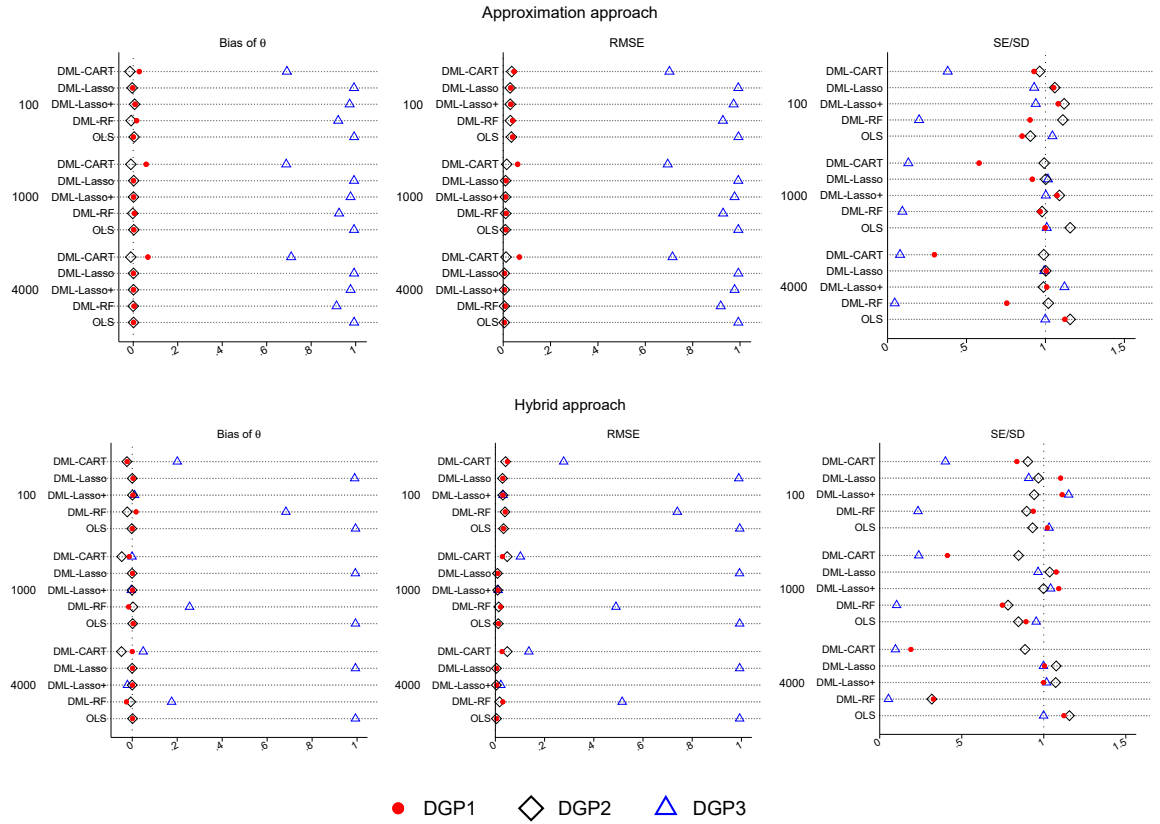
Note: The graphs plot the functional form of the nuisance parameters $l_0(x_1, x_3)$ and $m_0(x_1, x_3)$ – modelled as discussed in Section 5.1 – over x_1 (on the left) and x_3 (on the right) while setting the other variable equal to zero. The sample of cross-sectional observations is $N=1,000,000$ with $T = 10$ from which we sample while conducting the Monte Carlo simulations.

Figure 2. *Simulation results, CRE estimator*



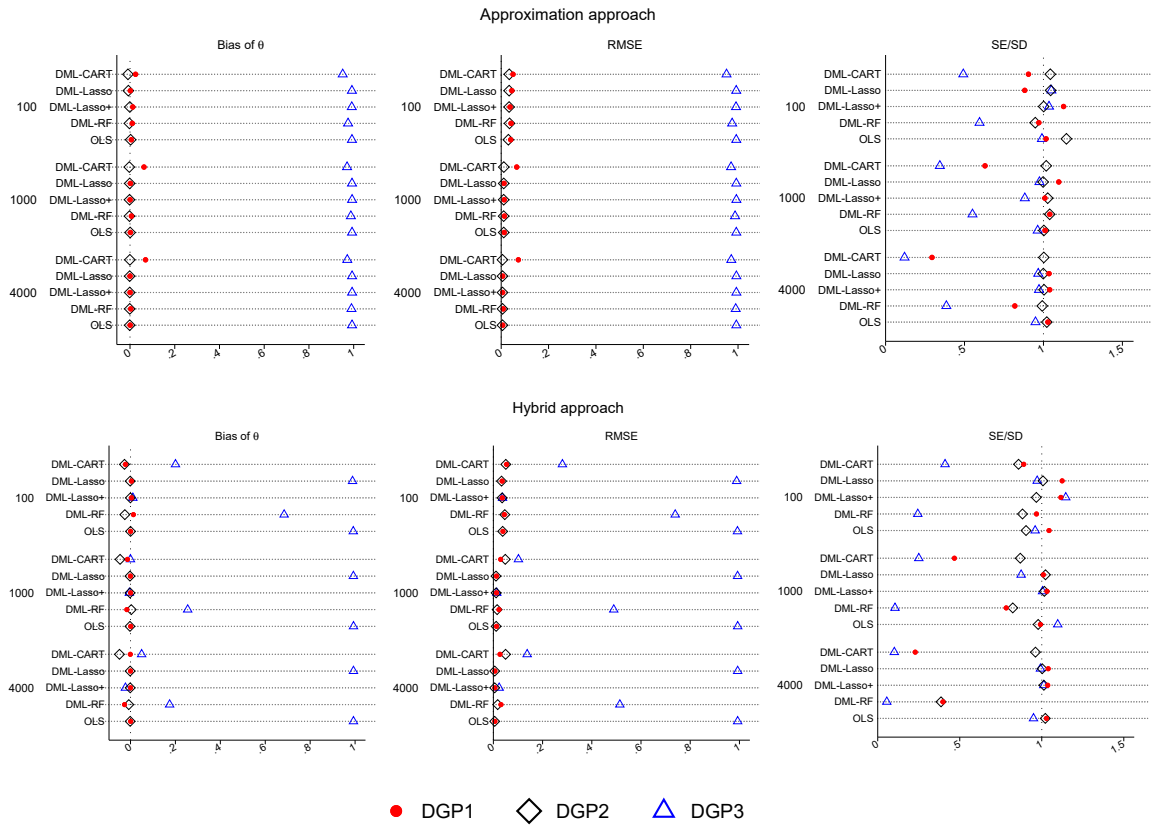
Note: Averages over 100 Monte Carlo replications. DGP1 is linear in the nuisance functions; DGP2 smooth non-linear; DGP3 non-smooth non-linear. Time is fixed to $t = 10$ time periods.

Figure 3. *Simulation results, WG estimator*



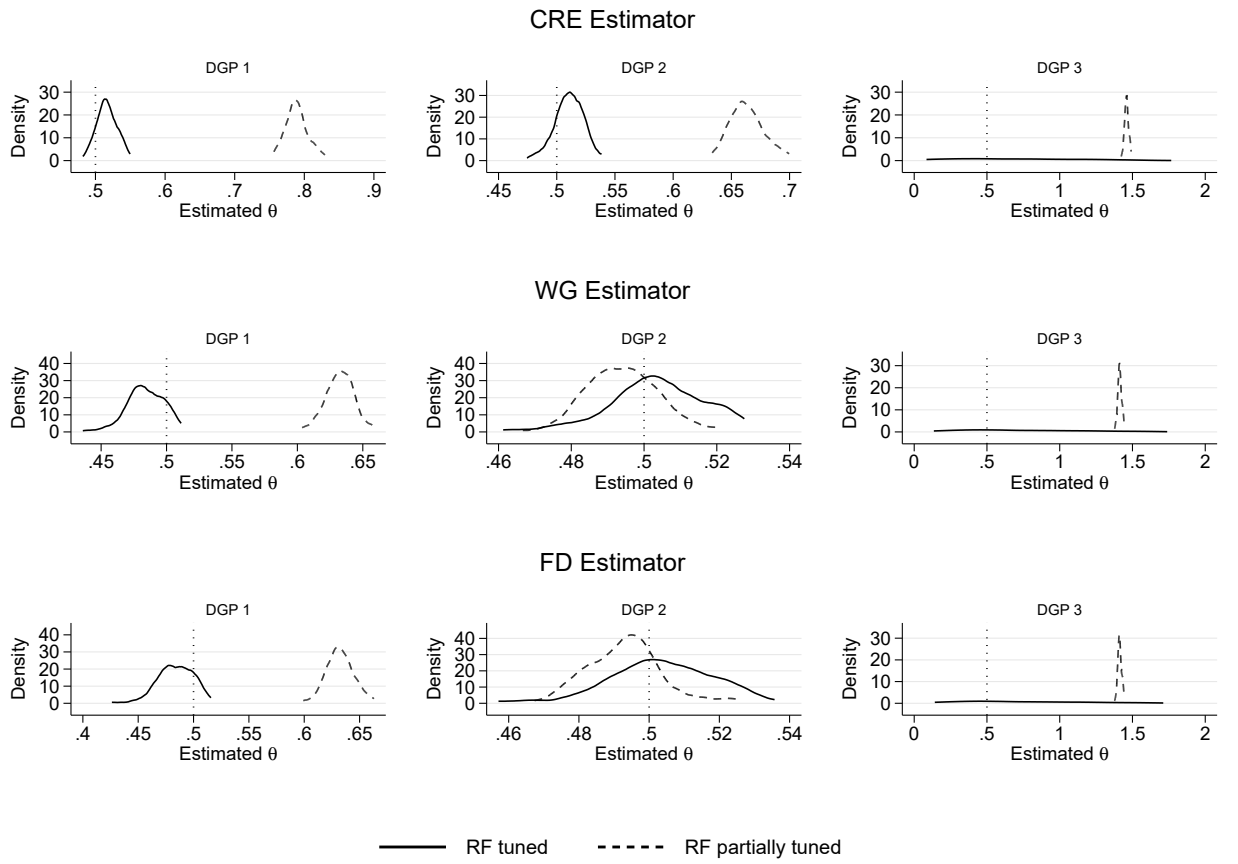
Note: Averages over 100 Monte Carlo replications. DGP1 is linear in the nuisance functions; DGP2 smooth non-linear; DGP3 non-smooth non-linear. Time is fixed to $t = 10$ time periods.

Figure 4. *Simulation results, FD estimator*



Note: Averages over 100 Monte Carlo replications. DGP1 is linear in the nuisance functions; DGP2 smooth non-linear; DGP3 non-smooth non-linear. Time is fixed to $t = 10$ time periods.

Figure 5. Distribution of $\hat{\theta}$, exact approach estimators for $N = 1,000$



Note: CART and RF 'tuned' use tuned hyperparameters from values reported in Table 1. RF 'partially tuned' uses 100 trees, maximum depth is 100, and the minimum node size is tuned. Hyperparameters are tuned via grid search.

C A Neyman Orthogonal Score Function

We adapt the development outlined for the cross-sectional case by [Chernozhukov et al. \(2018, sec. 2.2.4\)](#) for the partially linear panel model

$$y_{it} = \theta_0 d_{it} + g_0(\mathbf{x}_{it}) + \alpha_i + u_{it},$$

a more general version of which is

$$y_{it} = f_{\theta_0}(d_{it}; \mathbf{x}_{it}, t) + g_0(\mathbf{x}_{it}) + \alpha_i + u_{it},$$

where $f_{\theta_0}(d_{it}; \mathbf{x}_{it}, t) = E\{y_{it} - y_{it}(0) | d_{it}, \mathbf{x}_{it}, \alpha_i\}$ is a user-specified model for the causal effect that captures heterogeneity in the effect of d_{it} over \mathbf{x}_{it} or time t or both (but not α_i). This model leads to the IV-style estimator for θ_0 , but we will ultimately show that the Neyman orthogonal score for the PO-style estimator has the same form.

The model above can be written vector-wise as

$$\mathbf{r}_i = \mathbf{y}_i - \mathbf{f}_{\theta_0}(\mathbf{d}_i, X_i) - \mathbf{g}_0(X_i),$$

where $\mathbf{f}_{\theta_0}(\mathbf{d}_i, X_i) = (f_{\theta_0}(d_{i1}; \mathbf{x}_{i1}, 1), \dots, f_{\theta_0}(d_{iT}; \mathbf{x}_{iT}, T))'$, $\mathbf{g}_0(X_i) = (g_0(\mathbf{x}_{i1}), \dots, g_0(\mathbf{x}_{iT}))'$, $\mathbf{y}_i = (y_1, \dots, y_T)'$ and $\mathbf{r}_i = (r_{i1}, \dots, r_{iT})'$, with $r_{it} = \alpha_i + u_{it}$, are all column vectors of length T . By construction, the conditional moment restriction $E(u_{it} | d_{it}, \mathbf{x}_{it}, \alpha_i) = 0$ holds, but further assumptions are generally needed to identify θ_0 . We begin by deriving the score under the assumption that $E(\alpha_i | d_{it}, \mathbf{x}_{it}) = 0$ from which

$$E(\mathbf{r}_i | \mathbf{d}_i, X_i) = \mathbf{0} \tag{19}$$

follows.

Converting the notation used by [Chernozhukov et al. \(2018, Sec. 2.2.4\)](#) to that used in this paper, we have $W \equiv \{\mathbf{y}_i, \mathbf{d}_i, X_i\}$, $R \equiv \{\mathbf{d}_i, X_i\}$ and $Z \equiv \{X_i\}$, with $h(Z) \equiv \mathbf{g}_0(X_i)$ and $m(W; \theta, h(Z)) \equiv \mathbf{r}_i$. Using that $\partial_\theta \equiv \partial / \partial \theta = \nabla_\theta$, we can now define the various quantities needed and then use Lemma 2.6 to give us Neyman orthogonal score as follows:

$$A(R) \equiv -\partial_{\theta'} E[m\{W; \theta, h_0(Z)\} | R] |_{\theta=\theta_0} = -\partial_\theta \mathbf{f}'_{\theta_0},$$

which equals $-\mathbf{d}_i$ if $f_{\theta_0}(d_{it}; \mathbf{x}_{it}, t) = d_{it}\theta_0$; then

$$\Gamma(R) \equiv -\partial_{\nu'} E\{m(W; \theta_0, \nu) | R\} |_{\nu=h_0(Z)} = -I_T,$$

that is, the $T \times T$ identity matrix;

$$\Omega(R) \equiv E[m\{W; \theta_0, h_0(Z)\} m'\{W; \theta_0, h_0(Z)\} | R] = E(\mathbf{r}_i \mathbf{r}_i' | \mathbf{d}_i, X_i) = \Sigma_0(\mathbf{d}_i, X_i),$$

that is, the $T \times T$ within-individual auto-covariance matrix; and

$$\begin{aligned} G(Z) &\equiv \mathbb{E}\{A'(R)\Omega^{-1}(R)\Gamma(R)|Z\}\{\Gamma'(R)\Omega^{-1}(R)\Gamma(R)|Z\}^{-1} \\ &= \mathbb{E}\{\partial_{\theta}\mathbf{f}'_{\theta_0}\Sigma_0^{-1}(\mathbf{d}_i, X_i)I_T|X_i\}\mathbb{E}\{\mathbf{I}_T\Sigma_0^{-1}(\mathbf{d}_i, X_i)\mathbf{I}_T|X_i\}^{-1} \\ &= \mathbb{E}\{\partial_{\theta}\mathbf{f}'_{\theta_0}\Sigma_0^{-1}(\mathbf{d}_i, X_i)|X_i\}\mathbb{E}\{\Sigma_0^{-1}(\mathbf{d}_i, X_i)|X_i\}^{-1}. \end{aligned}$$

Applying their Lemma 2.6 leads to

$$\begin{aligned} \mu(R) &\equiv A'(R)\Omega^{-1}(R) - G(Z)\Gamma'(R)\Omega^{-1}(R) \\ &= \partial_{\theta}\mathbf{f}'_{\theta_0}\Sigma_0^{-1}(\mathbf{d}_i, X_i) - \mathbb{E}\{\partial_{\theta}\mathbf{f}'_{\theta_0}\Sigma_0^{-1}(\mathbf{d}_i, X_i)|X_i\}\mathbb{E}\{\Sigma_0^{-1}(\mathbf{d}_i, X_i)|X_i\}^{-1}\Sigma_0^{-1}(\mathbf{d}_i, X_i) \\ &= [\partial_{\theta}\mathbf{f}'_{\theta_0} - \mathbb{E}\{\partial_{\theta}\mathbf{f}'_{\theta_0}\Sigma_0^{-1}(\mathbf{d}_i, X_i)|\mathbf{X}_i\}\mathbb{E}\{\Sigma_0^{-1}(\mathbf{d}_i, X_i)|X_i\}^{-1}]\Sigma_0^{-1}(\mathbf{d}_i, X_i), \end{aligned}$$

so that the Neyman orthogonal score $\psi^{\perp} \equiv \mu(R)m\{W; \theta, h(Z)\}$ is

$$\psi^{\perp}(W; \theta_0, h_0(Z)) = \left[\partial_{\theta}\mathbf{f}'_{\theta_0} - \mathbb{E}\{\partial_{\theta}\mathbf{f}'_{\theta_0}\Sigma_0^{-1}(\mathbf{d}_i, X_i)|X_i\}\mathbb{E}\{\Sigma_0^{-1}(\mathbf{d}_i, X_i)|X_i\}^{-1} \right] \Sigma_0^{-1}(\mathbf{d}_i, X_i)\mathbf{e}_i. \quad (20)$$

We can further simplify this expression if selection is strongly ignorable such that

$$\mathbf{r}_i \perp\!\!\!\perp \mathbf{d}_i | X_i$$

so that (20) simplifies as

$$\psi^{\perp}(W; \theta_0, h_0(Z)) = \{\partial_{\theta}\mathbf{f}'_{\theta_0} - \mathbb{E}(\partial_{\theta}\mathbf{f}'_{\theta_0}|X_i)\}\Sigma_0^{-1}(X_i)\mathbf{e}_i \quad (21)$$

because $\Sigma_0^{-1}(\mathbf{d}_i, X_i) = \Sigma_0^{-1}(X_i)$. For the simple case we consider in this paper, $f_{\theta_0}(d_{it}; \mathbf{x}_{it}, t) = d_{it}\theta_0$, this is

$$\psi^{\perp}(W; \theta_0, h_0(Z)) = \{\mathbf{d}_i - \mathbb{E}(\mathbf{d}_i|X_i)\}\Sigma_0^{-1}(X_i)\mathbf{e}_i.$$

The equivalent result is obtained for PO model (2) if the model residual can be written $r_{it} = y_{it} - l_0(\mathbf{x}_{it}) - f_{\theta_0}(d_{it}; \mathbf{x}_{it}, t) + f_{\theta_0}\{m_0(\mathbf{x}_{it}); \mathbf{x}_{it}, t\}$, that is, where

$$f_{\theta_0}(d_{it}; \mathbf{x}_{it}, t) - E\{f_{\theta_0}(d_{it}; \mathbf{x}_{it}, t)|\mathbf{x}_{it}, \alpha_i\} = f_{\theta_0}\{d_{it} - m_0(\mathbf{x}_{it}); \mathbf{x}_{it}, t\} + f_{\theta_0}(c_i; \mathbf{x}_{it}, t),$$

and c_i satisfies the random effects assumption. Under this model, $h(Z) \equiv (l_0, m_0)$, $A(R) \equiv f_{\theta_0}\{d_{it} - m_0(\mathbf{x}_{it}); \mathbf{x}_{it}, t\}$, $\Omega(R) = \Omega(Z)$ and $\Gamma(R) = (-I_T, \theta_0 I_T) = \Gamma$ so that

$$G(Z) = E\{A'(R)|Z\}^{-1}\Omega(Z)\Gamma\{\Gamma'\Omega^{-1}(Z)\Gamma\}^{-1} = \mathbf{0}$$

and $\mu(R) = A'(R)\Omega^{-1}(Z)$.

If the random effects assumptions fails and we instead appeal to [ASM.7](#) and [ASM.8](#) then the same results above apply but with $W \equiv \{\mathbf{y}_i, \mathbf{d}_i, X_i, \bar{\mathbf{x}}_i\}$, $R \equiv \{\mathbf{d}_i, \mathbf{X}_i, \bar{\mathbf{x}}_i\}$ and $Z \equiv \{X_i, \bar{\mathbf{x}}_i\}$, with $h(Z) \equiv g_0$ or $h(Z) \equiv (l_0, m_0)$ and $m(W; \theta_0, h_0(Z)) \equiv \mathbf{r}_i$.

D Hyperparameter Tuning

Finding the optimal configuration of hyperparameters (or hyperparameter tuning) of a ML learner is essential to reach state-of-the-art performance in effect estimation, independently of the choice of estimators and learners (Machlanski et al., 2023). Hyperparameter optimization proceeds with trials of different configurations of values of the hyperparameters to tune. Resampling methods – such as, cross-validation (CV) – are used to evaluate the performance of the algorithm in terms of RMSE (when the hyperparameters are numeric). This procedure is repeated for several configurations until a stopping rule is applied (e.g., maximum number of evaluations). Finally, the configuration with the best performance (with, e.g., lowest RMSE) is selected and passed to the learner to train and test the model.

In the DML algorithm, hyperparameter tuning works as follows.

1. When the tuning is on folds, units in the training sample for fold k (W_k^c) are used for tuning. These are subsequently divided, e.g., in five-fold CV to create training and testing inner samples. When tuning is not on folds (default), all data is passed to the tuning procedure, but the composition of the units assigned to the k -th CV fold differs from the corresponding fold in the DML procedure. Then, five-fold CV is instantiated such that the k -th CV fold is the test sample and the rest the training sample.
2. The model is tuned by trying the performance of the learners with different configurations of the hyperparameters. The most common search algorithms are grid search and random search (Bergstra and Bengio, 2012). We use grid search as hyperparameter optimiser, which exhaustively evaluates any possible combination of given hyperparameter values in the grid, conditional to a given resolution (i.e., the number of different values to try per hyperparameter). This method are non-adaptive such that the proposed configuration ignores the performance of previous ones.
3. Each evaluation within the tuning routine selects the best configuration of hyperparameters among all k CV folds, based on the lowest RMSE. Once the tuning algorithm stops (e.g., at the j -th evaluation), the best configuration of hyperparameters among the j results is chosen (based on lower RMSE) and passed to the DML algorithm.
4. The best configuration is set as parameters of the learners of the nuisance parameters. The model is then trained on the complementary set for fold k , W_k^c , and tested on W_k . Predictions for m and l are stored.

The default tuning procedure for DML (not on folds) follows the same sample splitting principle behind DML. There is no separate test set for validation because predictions are done at the DML stage, and the test sample in the learning stage of DML uses different combinations of units in each fold (tuning not on folds).