Abstract

This paper introduces and proves asymptotic normality for a new semi-parametric estimator of continuous treatment effects in panel data. Specifically, we estimate an average derivative of the regression function. Our estimator uses the panel structure of data to account for unobservable time-invariant heterogeneity and machine learning methods to flexibly estimate functions of high-dimensional inputs. We construct our estimator using tools from double de-biased machine learning (DML) literature. We show the performance of our method in Monte Carlo simulations and also apply our estimator to real-world data and measure the impact of extreme heat in United States (U.S.) agriculture. We use the estimator on a county-level dataset of corn yields and weather variation, measuring the elasticity of yield with respect to a marginal increase in extreme heat exposure. In our preferred specification, the difference between the estimates from OLS and our method is statistically significant and economically significant. We find a significantly higher degree of impact, corresponding to an additional $1.18 billion in annual damages by the year 2050 under median climate scenarios. We find little evidence that this elasticity is changing over time.

Keywords: Average derivative, machine learning, panel data, climate change
JEL Classification: C14, C21, C55, Q51, Q54
1 Introduction

Estimating continuous treatment effects in panel data is essential for many important questions. For example, environmental researchers study how health or economic outcomes change with exposure to high temperatures, air pollution, or other continuous environmental factors. In economics, applied work commonly studies the impact of a continuous variable such as price (e.g. for demand models in industrial organization) or distance (e.g. for evaluating programs in urban economics) in panel settings. Panel data lets researchers control for time-constant factors that are not recorded in the data, by using observations of the same unit over several time periods. In the widely used fixed effects approach, researchers model time-constant factors as indicator variables for each unit in the data (Wooldridge 2005). Most applications using the fixed effects approach for panel analysis rely on linear models and do not attempt to flexibly model high dimensional covariates and their interactions. This can cause significant biases if the linear model is not correctly specified, or if there is heterogeneity in effects.

Machine learning (ML) approaches like Lasso provide tools to estimate flexible functions of high-dimensional data, but standard models can introduce bias and do not incorporate fixed effects terms. Through regularization and overfitting, ML can result in biased estimates and invalid confidence intervals (Chernozhukov, Chetverikov, et al. 2017). There are a growing number of double/debiased machine learning (DML) approaches that address these biases and can construct valid confidence intervals (Chernozhukov, W. Newey, and Singh 2018, Chernozhukov, W. K. Newey, and Singh 2022, Rothenhäusler and Yu 2019). Current DML approaches for continuous treatments do not account for common features of panel data, such as the fixed effects term.

This paper fills this gap by introducing a DML estimator for continuous treatment effects in panel settings with additive fixed effects. We present a new panel data estimator of average derivatives that allows for high dimensionality and heterogeneity. After introducing the estimator, we show that it performs well in simulation trials and prove that it is asymptotically normal. In simulations, we show that our estimator maintains low bias and valid confidence intervals, while standard Lasso does not. We then use the estimator to study an important question in agricultural economics, the impact of extreme heat exposure on crop yields. Our work relates to the panel, DML, continuous treatment effect, and climate economics literature. First, we discuss the applied and theoretical motivation for the estimator. We then discuss our application to measure damages to U.S. agriculture from extreme heat exposure. We show that in both cases, our DML approach allows greater model flexibility while preserving low standard errors.

Flexible continuous effects in panel data are under-explored in the theoretical literature relative to binary treatment effects. While there are many interesting flexible panel methods for binary treatment effects, including factor models and synthetic controls, there are no such methods for continuous effects. The paper studying binary treatment effects in panel flexibly that is most related to ours is Belloni et al. (2016). This paper uses a fixed-effects approach, as we do, and allows for the outcome model to be a general function of covariates estimated with ML, but does not allow for general treatment effect heterogeneity as we do. Chernozhukov,
W. K. Newey, and Singh (2022) and Chernozhukov, Goldman, et al. (2017) allow for general treatment effect heterogeneity and continuous treatment variables. However, Chernozhukov, Goldman, et al. (2017) considers the dynamic panel and imposes the potentially strong assumption of sequential exogeneity, and Chernozhukov, W. K. Newey, and Singh (2022) does not consider panels with few observations per unit. Both these papers use a random correlated effects approach, which imposes a restriction on the relationship between the covariates and the unobserved fixed effects. Our fixed-effects approach does not impose that restriction.

This paper focuses on the average derivative, although there are many other potentially interesting continuous treatment effects. For example, Colangelo and Lee (2020) and Klosin (2021) study the dose-response curve, and Chernozhukov, W. K. Newey, and Singh (2022) considers average policy effects, average treatment effects, and the average equivalent variation bound. The methods in our paper could be adapted to other objects, such as average policy effects or conditional average derivatives. We focus on the average derivative because it can be estimated at parametric root $n$ rates and because there are many applications of the average derivative in practice. Many papers estimate an elasticity, which is equivalent to estimating an average derivative when the outcome variable is the logarithm of a raw value. The average derivative is also studied by, among others, Rothenhäsler and Yu (2019) and Imbens and Newey (2009).

We also contribute to the DML literature more generally by introducing a new optimization procedure to estimate the de-biasing component. We show that the optimization procedure improves performance in Monte Carlo simulations. Furthermore, though our paper is the first to use DML to estimate average derivatives with a fixed effects panel data structure, we contribute more generally to the literature on average derivatives by introducing a new analytical method for estimating average derivatives that is very stable in simulations.

We apply our estimator to study the impact of extreme heat on United States (U.S.) corn yields. We estimate the elasticity of corn yield with respect to a marginal increase in extreme heat after controlling for a set of weather variation compiled by Abatzoglou (2013). Our estimator is well suited to this setting because of the importance of fixed effects modeling to isolate the impact of weather shocks (Deschênes and Greenstone 2007) and the nonlinear relationships between continuous environmental factors such as temperature and precipitation (Schlenker and Roberts 2009).

This application is most closely related to Schlenker and Roberts (2009), which estimates this same elasticity. They introduce a parsimonious model that captures the response of crop yields to temperature and precipitation. Our approach differs because we consider flexible functions and interactions between these terms. Using our approach, we project economic damages from extreme heat by the year 2050 that are $1.18 billion more than projections using the weather variation and linear model from Schlenker and Roberts (2009). We estimate the value of this elasticity over time, and cannot reject that the elasticity remains constant over time. This corroborates findings from Schlenker and Roberts (2009) and Burke and Emerick (2016) who conclude that there has been limited adaptation to extreme heat in U.S. agriculture.

The application is also related to a literature in environmental economics on studying the
economic impacts of climate change. Rode et al. (2021) estimates elasticities of energy use with respect to temperature, to construct a global estimate of the social cost of carbon from energy use. Burke, Hsiang, and Miguel (2015) and Hsiang et al. (2017) estimate economic damages of climate change globally and within the U.S., respectively, by combining estimates of elasticities of multiple economic sectors with respect to temperature. These approaches model the impacts of temperature by separating heat exposure into bins and estimating a coefficient for each bin, a model advocated by Dell, Jones, and Olken (2014) and Hsiang (2016). These approaches can have limitations for high-dimensional weather variation, as the number of variables grows exponentially if the researcher wishes to include interactions between weather factors. Our project is the first, to our knowledge, to estimate these elasticities using a machine learning approach, allowing us to flexibly model higher-dimensional variation.

Our project is also related to a growing field applying machine learning techniques in environmental economics. Many researchers have used machine learning for predictive properties, such as forecasting crop yields (Crane-Droesch 2018), filling missing data to track global poverty (Jean et al. 2016), or more effectively assign treatment (Knittel and Stolper 2019). Recent work has begun using machine learners for measurement tasks. Deryugina et al. (2019) uses a machine learning approach to measure the costs of air pollution. Stetter, Mennig, and Sauer (2022) use a de-biased machine learning approach to measure effectiveness of an agricultural intervention. Our paper is the first application in environmental economics, to our knowledge, to use double machine learning to estimate continuous treatment effects.

The paper is structured in the following way. Section 2 sets up the framework of the paper, introduces the parameter of interest, and presents our estimator. Simulation design and results are given in Section 3. Section 4 covers our application. Section 5 concludes.

2 Estimation

2.1 Notation and Definitions

We work in a panel data setting with \( n \) individuals and \( T \) time periods. As is often the case in economic data, we assume that \( n \) is large but \( T \) is small. We assume we have independent and identically distributed data \((W_1, \ldots, W_n)\) where the \( W_i = \{(X_{i,t}, D_{i,t}, Y_{i,t})\}_{t=1}^T \) are copies of a random variable \( W \) with support \( \{W = \mathcal{X} \times \mathcal{D} \times \mathcal{Y}\}_{t=1}^T \), with a cumulative distribution function (cdf) \( F_{Y|DX}(Y, D, X) \). We use capital letters to denote random variables and lowercase letters to denote their possible values. For each unit in a population \( X_{i,t} \in \mathbb{R}^h \) denotes a vector of covariates, with \( h \) potentially large, and \( D_{i,t} \in \mathbb{R} \) denotes the treatment variable.

For a given variable \( X \), we use the notation \( \Delta X_{i,t} := X_{i,t} - X_{i,t-1} \) for the first difference transformation. For the first difference transformation of a function \( f \) of a variable \( X \), we apply the function \( f \) before taking the difference: \( \Delta f(X_{i,t}) := f(X_{i,t}) - f(X_{i,t-1}) \).

Define \( | \cdot |_1 \) as the \( \ell_1 \) norm; that is, \( |\beta|_1 = \sum_{j=1}^p |\beta_j| \) where \( \beta_j \) is the \( j^{\text{th}} \) component of \( \beta \) and \( p \) is the length of \( \beta \).
2.2 Parameter of Interest

We are estimating a general additive fixed effects panel model.

\[ Y_{i,t} = a_i + \gamma_0(D_{i,t}, X_{i,t}) + \epsilon_{i,t} \quad E[\epsilon_{i,t}|a_i, X_{i,1}, \ldots, X_{i,T}, D_{i,1}, \ldots, D_{i,T}] = 0 \]  

(1)

Here \( a_i \) represents individual fixed effects, and \( \gamma \) is a flexible high dimensional function of treatment, covariates, interactions, and higher order terms. We assume that \( \gamma \) is constant throughout time and that it can be estimated well with Lasso. This assumption implies a form of sparsity on covariates. However, we do not assume that the fixed effects \( a_i \) are sparse.

We apply Lasso after applying a set of basis functions to transform \( \{D_{i,t}, X_{i,t}\} \) into a high-dimensional set of covariates. We define a \( p \times 1 \) dictionary of basis functions \( b \) that transforms our original vector of covariates, so that \( b(D_{i,t}, X_{i,t}) \in \mathbb{R}^p \). Basis functions can include any desired transformations of the covariates, such as polynomial terms or interactions between variables. The assumption that Lasso estimates \( \gamma \) implies that there exists a sparse parameter vector \( \beta_0 \in \mathbb{R}^p \) such that \( \gamma_0(D_{i,t}, X_{i,t}) = b(D_{i,t}, X_{i,t})'\beta_0 \).

We choose these modeling assumptions because they match those commonly used in applied work, while relaxing functional form assumptions on \( \gamma_0 \). We assume the additive error term \( \epsilon_{i,t} \) is mean zero conditional on the history of covariates, an assumption called strict exogeneity that is frequently used in applied work (Wooldridge 2010).

Our estimation target is the average of a continuous treatment effect, the moment function \( m \) given in (2). In our leading application, we consider the average derivative:

\[ \tau_0 = \mathbb{E}[m(W_{i,t}, \gamma_0)] = \mathbb{E} \left[ \frac{\partial \gamma_0(D_{i,t}, X_{i,t})}{\partial D_{i,t}} \right] \]  

(2)

Note here that the expectation is over both \( D_{i,t} \) and \( X_{i,t} \), and that the target derivative is only a function of \( D_{i,t} \) and not its lagged value. The causal interpretation of \( \tau_0 \) is the average causal effect of a marginal increase in treatment. The average derivative has been studied by, among others, Imbens and Newey (2009) and Rothenhäusler and Yu (2019). When \( y_{i,t} \) is in log scale, this parameter captures the elasticity of \( y \) with respect to a marginal change in \( D \).

To account for the fixed effect term \( a_i \), we introduce a first-differenced version of (1). In short panels, it is not possible to consistently estimate \( a_i \). By taking a first difference, we can remove the time-invariant factor \( a_i \) and consistently estimate \( \gamma_0 \) (Wooldridge 2010).

\[ Y_{i,t} - Y_{i,t-1} = \gamma_0(D_{i,t}, X_{i,t}) - \gamma_0(D_{i,t-1}, X_{i,t-1}) + \epsilon_{i,t} - \epsilon_{i,t-1} \]  

(3)

\[ \Delta Y_{i,t} = \Delta \gamma_0(D_{i,t}, X_{i,t}) + \Delta \epsilon_{i,t} \]  

(4)

Because we apply the first difference transformation after taking the function \( \gamma_0 \) of our data, our linear representations of \( \gamma_0 \) and \( \Delta \gamma_0 \) share the same parameter vector \( \beta_0 \). That is, if \( \gamma_0(D_{i,t}, X_{i,t}) := b(D_{i,t}, X_{i,t})'\beta_0 \), then \( \Delta \gamma_0(D_{i,t}, X_{i,t}) := \Delta b(D_{i,t}, X_{i,t})'\beta_0 \).

Note that we can express the estimation target in terms of the average derivative of \( \Delta \gamma_0 \):

\[ \tau_0 = \mathbb{E} \left[ \frac{\partial \gamma_0(D_{i,t}, X_{i,t})}{\partial D_{i,t}} \right] = \mathbb{E} \left[ \frac{\partial \Delta \gamma_0(D_{i,t}, X_{i,t})}{\partial D_{i,t}} \right] \]  

(5)
2.3 Estimator

We construct a de-biased estimator of the average derivative using two Lasso estimated high dimensional functions, \( \Delta \hat{\gamma} \) and \( \hat{\alpha} \). The first, \( \Delta \hat{\gamma} \), is an estimate of the regression function \( \Delta \hat{\gamma} \) in equation 4, how we estimate this function is explained in detail in Section 2.3.1. The second, \( \hat{\alpha} \) is a de-biasing term. We describe \( \hat{\alpha} \) in more detail and explain how we construct it in Section 2.3.2 after introducing the full estimator. In short one can consider \( \Delta \hat{\gamma} \) to be the initial ML estimate of the effect, and the \( \hat{\alpha} \) to be an add on that de-biases this initial estimate.

We use a cross-folds procedure to reduce the risk of overfitting. First the researcher chooses the number of splits \( L \) (\( L = 5 \) is commonly used). Then each unit’s indices are randomly partitioned into the \( L \) equally sized groups. We use \( \ell \) to denote these groups, \( \ell = 1, \ldots, L \). Denote observations in group \( \ell \) by \( W_\ell \). All observations from a single panel unit are placed in the same fold. This is important because the data in different folds should be independent, and there is a dependence within the observations of a single unit. Our functions \( \hat{\gamma}_\ell \) and \( \hat{\alpha}_\ell \) are trained using observations not in group \( \ell \). The full estimator takes these two functions \( \Delta \hat{\gamma}_\ell \) and \( \hat{\alpha}_\ell \) from each fold and uses them to construct our estimate \( \hat{\tau} \):

\[
\hat{\tau} = \frac{1}{n(T-1)} \sum_{\ell=1}^L \sum_{i=\ell}^T \hat{\tau}_{\ell;i,t}
\]

\[
\hat{\tau}_{\ell;i,t} = \frac{\partial \Delta \hat{\gamma}_\ell(D_{i,t}, X_{i,t})}{\partial D_{i,t}} + \hat{\alpha}_\ell(D_{i,t}, X_{i,t}, D_{i,t-1}, X_{i,t-1})(\Delta Y_{i,t} - \Delta \hat{\gamma}_\ell(D_{i,t}, X_{i,t}))
\]

(6)

To compute the asymptotic variance of our estimator, it is necessary to account for correlation of our target within panel units. We assume that errors have a constant correlation within a panel unit but are uncorrelated between panel units. Let \( \hat{\tau}_{\ell;i} = 1/(T-1) \sum_{t=2}^T \hat{\tau}_{\ell;i,t} \). Then the asymptotic variance is:

\[
\hat{V} = \frac{1}{n(T-1)} \sum_{\ell=1}^L \sum_{i=\ell}^T \left\{ \sum_{t=2}^T (\hat{\tau}_{\ell;i,t} - \hat{\tau})^2 + 2 \sum_{t=2}^{T-1} \sum_{t'=t+1}^T (\hat{\tau}_{\ell;i,t} - \hat{\tau}_{\ell;i}) (\hat{\tau}_{\ell;i,t'} - \hat{\tau}_{\ell;i}) \right\}
\]

(7)

Assumptions and the proof for asymptotic normality of our estimator are given in Appendix A. We work with \((T-1)\) time periods rather than \(T\) because we removed one time period by first-differencing the data.

In the following subsections, we describe how we compute the derivative of the regression function and how we estimate the bias correction term \( \alpha \).

2.3.1 Derivative of Regression Function

There are two general steps to find the derivative of the regression function \( \frac{\partial \Delta \hat{\gamma}_\ell(D_{i,t}, X_{i,t})}{\partial D_{i,t}} \), in (6):

1. Estimate \( \Delta \hat{\gamma}_\ell \) for each fold.
(a) Transform the covariates \{D_{i,t}, X_{i,t}\} using a flexible specification. We do so by using polynomial basis functions of terms and interactions, although other approaches like kernel functions or splines could be used as long as the derivatives are bounded. Let \( b(D_{i,t}, X_{i,t}) \) denote the resulting \( p \times 1 \) dictionary of functions. Then let \( \Delta b(D_{i,t}, X_{i,t}) := b(D_{i,t}, X_{i,t}) - b(D_{i,t-1}, X_{i,t-1}) \). We set each function in the dictionary \( b \) to have mean 0 and variance 1; Appendix B.3 discusses this procedure.

(b) Find a vector of coefficients \( \hat{\beta} \) for our dictionary such that \( \Delta \hat{\gamma}(D_{i,t}, X_{i,t}) := \Delta b(D_{i,t}, X_{i,t})' \hat{\beta} \) is a sparse linear approximation of \( \Delta \gamma_0(D_{i,t}, X_{i,t}) \). We do so by solving the following Lasso problem:

\[
\hat{\beta} = \arg\min_\beta \left\{ \frac{1}{n(T-1)} \sum_{i=1}^{n} \sum_{t=1}^{T}(\Delta Y_{i,t} - \Delta b(D_{i,t}, X_{i,t})' \beta)^2 + r_L|\beta|_1 \right\}
\]

This procedure depends on the regularization weight \( r_L \), which we determine by finding values that minimize test-set error in a cross-folds procedure. This procedure is described in Appendix B.2.

2. Calculate the derivative.

We calculate the derivative analytically. Many DML approaches use numeric differentiation; we discuss this alternative in Appendix B.1. Our procedure uses the estimate of \( \hat{\beta}_t \) from the previous step to compute the derivative and derivatives of each function in our dictionary of basis functions.

(a) Construct the dictionary \( b_D \), a \( p \times 1 \) dictionary of derivatives of each basis function in \( b \). For each basis function \( b^j \) for \( j = 1, \ldots, p \) in our dictionary of basis functions, define its derivative as follows:

\[
b^j_D(D, X) = \frac{\partial b^j(D, X)}{\partial D}
\]

(b) Estimate the average derivative as:

\[
\mathbb{E}\left[ \frac{\partial \Delta \hat{\gamma}(D_{i,t}, X_{i,t})}{\partial D_{i,t}} \right] = \mathbb{E}[b_D(D_{i,t}, X_{i,t})' \hat{\beta}]
\]

Example 1. Consider a simple setting where \( X_{i,t} \in \mathbb{R} \), and where \( \gamma_0(D_{i,t}, X_{i,t}) = D_{i,t}^2X_{i,t}. \) Our basis function dictionary is \( b(D_{i,t}, X_{i,t}) = \{D_{i,t}, X_{i,t}, D_{i,t}^2X_{i,t}\} \). In our linear representation, \( \beta_0 = \{0, 0, 1\}. \)

In step 1, we obtain an estimate \( \hat{\beta} \) using Lasso. In step 2, we first define the derivative of the basis functions. Here, \( b_D(D_{i,t}, X_{i,t}) = \{1, 0, 2D_{i,t}X_{i,t}\} \). The estimated average derivative is then: \( \hat{\beta}_1 + \mathbb{E}[2D_{i,t}X_{i,t}]\hat{\beta}_3 \), where \( \hat{\beta}_j \) is the \( j \)-th component of \( \hat{\beta} \).

1. For example, when we estimate \( \Delta \gamma_t \) we set \( b(D_{i,t}, X_{i,t}) \) to be a third order polynomial set of the covariate variables and interactions between \( D \) and each covariate in \( X \).
2.3.2 De-biasing term

Now the next term of equation (6) is the de-biasing term, \( \hat{\alpha}(D_{i,t}, X_{i,t}, D_{i,t-1}, X_{i,t-1}) \). Our \( \alpha \) is based on the methods of Chernozhukov, W. K. Newey, and Singh (2022), which use the Riesz Representation theorem. In short this theorem states that for a given linear functional \( m \), there exists a function \( \alpha \) such that the following holds for any function \( b \):

\[
\mathbb{E}[m(W)] = \mathbb{E}[^{\alpha}(W)b(W)]
\]

(11)

In our case using the definition of the moment and our derivative bases, we have \( m(W, b) = \Delta b(D_{i,t}, X_{i,t}) \). These basis functions are the same standardized basis functions as in Section 2.3.1. Plugging this into the above yields:

\[
\mathbb{E}[\Delta b(D_{i,t}, X_{i,t})] = \mathbb{E}[\alpha(D_{i,t}, X_{i,t}, D_{i,t-1}, X_{i,t-1})\Delta b(D_{i,t}, X_{i,t})]
\]

(12)

Because this equality holds regardless of the function \( b \), it is possible to estimate \( \alpha \) from data independently of estimating the function \( \gamma \). In order to construct an estimate \( \hat{\alpha} \), we first assume that \( \alpha_0 \) has a sparse linear form:

\[
\alpha_0(D_{i,t}, X_{i,t}, D_{i,t-1}, X_{i,t-1}) = \Delta b(D_{i,t}, X_{i,t})\rho_0
\]

Our estimate is then \( \hat{\alpha}(D_{i,t}, X_{i,t}, D_{i,t-1}, X_{i,t-1}) = \Delta b(D_{i,t}, X_{i,t})\hat{\rho} \). We follow Chernozhukov, W. K. Newey, and Singh (2022) and find \( \hat{\rho} \) to minimize the squared loss between \( \alpha_0 \) and \( \hat{\alpha} \):

\[
\hat{\rho} = \arg\min_{\rho} \{ \mathbb{E}[(\alpha_0(D_{i,t}, X_{i,t}, D_{i,t-1}, X_{i,t-1}) - \Delta b(D_{i,t}, X_{i,t})\rho)^2] + r_\alpha|\rho|_1 \}
\]

(13)

Additional details about this solution are given in Appendix B.4. Chernozhukov, W. K. Newey, and Singh (2022) provides an iterative procedure to solve this problem; we also introduce an alternate approach using an optimization package to solve the minimization problem.

Our implementation using an optimization package guarantees that we find an optimal solution to this minimization problem. Iterative approaches may fail to converge to the true value of the parameter, with finite training time. Given the convex nature of the minimization problem, we are able to find an exact solution using optimization software. We use the Python package CvxPy (Diamond and Boyd 2016) to set up the problem and the convex optimization software Mosek (ApS 2021) to solve the problem. In simulation trials, we compare the performance of iterative and optimizer approaches for determining \( \hat{\alpha} \).

With an estimate \( \hat{\alpha} \), we can estimate the average derivative using only the \( Y_{i,t} \) observations:

\[
\hat{\tau} = \mathbb{E}[\hat{\alpha}(D_{i,t}, X_{i,t}, D_{i,t-1}, X_{i,t-1})\Delta Y_{i,t}]
\]

We now provide an example to illustrate the role of the Riesz Representer function in a simple setting.

**Example 2.** Consider the average derivative of a function \( \gamma \) of \( X \) where \( X \sim N(0, 1) \). Via integration by parts after expanding the expectation, \( \mathbb{E}\left[\frac{\partial \gamma(X)}{\partial X}\right] = \mathbb{E}[X\gamma(X)] \).
That is, the Riesz Representer is $\alpha_0(X) = X$ when $X \sim N(0,1)$ and $m$ is the derivative operator. The Riesz Representer depends on the distribution of data and the operator $m$, but does not depend on the choice of function $\gamma$.

If our basis function is the identity (i.e. $b(X_i) = X_i$), then in our linear representation above, $\rho_0 = \{1\}$. We construct an estimate $\hat{\rho}$ using our Lasso procedure, and then estimate the average derivative as: $\hat{\tau} = E[X_i \hat{\rho} Y_i]$.

3 Simulations

We demonstrate the performance of our estimator through a Monte Carlo simulation exercise. We use the data generating process (DGP) defined below to create 1000 different datasets. We report the true value of the derivative, our estimate, the average bias of our estimates, and mean squared error (MSE) between true and estimated values. We compare the performance of our estimator with OLS on the untransformed set of covariates (OLS Linear), OLS on the basis function transformation of covariates (OLS Poly), Lasso without a bias correction term, and DML using an iterative estimation procedure (DML Iterative).

DGP : $N = 1000$ number of individuals, $T = 2$ number of time periods, $h = 20$ original number of $X$ covariates. Our basis function transformation takes third order polynomials of each variable, then adds interactions between each $\{D, X^{(j)}\}$ pair and their polynomials. We have a total of $p = 244$ covariates after applying the basis function transformation.

We generate outcome variables according to the following function:

$$Y_{i,t} = a_i + D_{i,t} + D_{i,t}^2 + D_{i,t}^3 + D_{i,t}X_{i,t}^{(1)} + .1\theta X_{i,t} + \epsilon_{i,t}$$

(14)

To match real-world panel settings, we impose a correlation between $a_i, D_{i,t}$ and $X_{i,t}$. We set $\theta$ so that the $j^{th}$ element is $\theta_j = 1/j^2$. Fixed effects $a_i$, covariates $X_{i,t}$, and random noise $\epsilon_{i,t}$ are drawn from Gaussian distributions: $a_i \sim N(1,1)$, $X_{i,t}^{(j)} \sim N(a_i, 1) j = 1, \ldots, h$, and $\epsilon_{i,t} \sim N(0,1)$, while the treatment is correlated with $X_{i,t}$ but includes simulation draws from a Beta distribution: $D_{i,t} \sim .1\theta X_{i,t} + Beta(1,7)$

We show the results of this simulation trial by summarizing the bias and MSE in Table 1 and plotting the distribution of the error in Figure 1.

Table 1 demonstrate the value of our approach for applied research. If an applied researcher were faced with our simulated data with 20 covariates and they wanted to use classic OLS methods they could run either a classic linear regression with 20 covaraiates (“OLS Linear”) or they could run a more flexible linear regression with polynomial terms of these 20 base covaraiates leading to 244 covariates (“OLS Poly”). Note that in order to recover the average derivative using OLS Poly, the researcher must use either derivatives of the basis functions, as described in Section 2.3.1, or numerical differentiation.

Researchers may prefer OLS Linear to OLS Poly because the large number of regressors in OLS Poly can can lead to large standard errors, even though OLS Linear may be biased due to misspecification. The bias and variance trade off of OLS Linear and OLS Poly is exactly what we see in the table. OLS Linear has smaller standard errors than OLS Poly, which is
| method                      | DML | DML Iterative | Lasso | OLS Linear | OLS Poly |
|----------------------------|-----|---------------|-------|------------|----------|
| True Value                 | 2.96| 2.96          | 2.96  | 2.96       | 2.96     |
| Average Derivative         | 2.958| 2.94          | 2.683 | 3.246      | 2.939    |
| Bias                       | -0.002252| -0.02015      | -0.2771| 0.2861     | -0.0208  |
| Standard Deviation         | 0.2991| 0.341         | 0.3581| 0.3311     | 0.5573   |
| MSE $\tau$                | 0.08013| 0.1069        | 0.1957| 0.1836     | 0.3009   |
| Coverage                   | 0.924| 0.962         | 0.224 | 0.886      | 0.95     |
| MSE $\gamma$ In Sample    | 1.951| 1.951         | 1.951 | 2.338      | 1.515    |
| MSE $\gamma$ Cross Folds  | 2.048| 2.048         | 2.048 | 2.454      | 10.04    |

Table 1: Summary of derivative estimates from 1000 bootstrap trials of our simulation procedure. Bias is the average of the estimated value of the derivative minus the true value of the derivative in each simulation draw. “MSE $\tau$” is the mean squared error between the true average derivative and the estimated average derivative in each simulation draw, while “MSE $\gamma$ in sample” and “MSE $\gamma$ cross folds” refer to the mean squared error of regression from own-sample and out-of-sample estimation.

why applied researchers may prefer it, however OLS Linear has a huge bias relative to OLS Poly.

Our DML method provides an alternative that preserves the low standard errors of OLS Linear and the low bias of flexible modeling such as OLS Poly. Due to the regularization of Lasso, our estimates have considerably lower standard errors as OLS Poly even though they use the same set of basis functions. Figure 1 visualizes this benefit - the error from DML and OLS Poly are both approximately centered around 0, but the distribution of errors using OLS Poly is much wider. In this simulation trial, our method has lower bias than OLS Poly and lower standard errors than OLS Linear.

Reassuringly, our DML procedure results in an approximately de-biased estimate of the average derivative. This is especially clear relative to Lasso without any correction. The bias of base Lasso is 100 times larger than the bias of our proposed method. Our estimates also have substantially lower bias than OLS using the untransformed covariates (“OLS Linear”) and lower bias than OLS using the basis function transformation (“OLS Poly”). We expect these results, as misspecification of OLS Linear can induce a bias in estimating the average derivative, and OLS Poly may be overfitting the data (as shown by the high error in MSE cross folds). Our optimization-based DML approach results in lower bias and standard deviation in the result than DML Iterative, although both dramatically improve over Lasso without a correction. This comparison was done by creating our estimator and then using two different algorithms to solve for the de-baising term.

We use the mean squared error of estimating the true parameter (MSE $\tau$) to compare results, incorporating differences in both bias and variance. The MSE $\tau$ using our estimation procedure is the lowest among all models considered, roughly half the magnitude from Lasso or OLS Linear and one quarter the magnitude from OLS Poly. The MSE $\tau$ is closest to the result using DML Iterative, and shows that our general optimization solution to de-biasing is leading to lower bias.
4 Application

In our application, we measure the elasticity of corn yields with respect to extreme heat exposure and use this to project the damages from expected changes in the distribution of extreme heat following climate change. The treatment variable $D_{i,t}$ is annual aggregate exposure to temperatures above 29 °C; for simplicity we refer to this variable as extreme heat. This variable roughly captures the amount of heat stress a plant experiences. As Schlenker and Roberts (2009) demonstrate, crop yields are generally decreasing in extreme heat, while increasing in heat exposure below 29 °C. The covariates $X_{i,t}$ include other weather features, such as heat exposure below this temperature threshold, precipitation, humidity, and solar radiation. Heat exposure is measured in Growing Degree Days (GDD), the amount of time a crop is exposed to temperatures between an upper and lower bound during the March-August growing season.

4.1 Data Description

We work with two datasets for the main estimation: corn yield from the USDA’s Survey of Agriculture, and weather records from Abatzoglou (2013). To form projections about the impacts under climate change, we use data about projected future impacts across the U.S. that are generously shared by Burke and Emerick.

For corn yields, we use county-level reports from the USDA Survey of Agriculture, supplemented with additional data generously provided by Burke and Emerick (2016). We use the metric bushels per acre, and focus on U.S. counties east of the 100°West meridian. This re-
region is commonly used in studies of U.S. agriculture because the region west of this meridian relies on heavily subsidized irrigation. In our sample period, over 90% of corn in the U.S. was grown in this region. This dataset also includes the planted area of corn per county in each year.

For weather observations, we use the GridMET weather dataset from Abatzoglou (2013). This dataset contains a rich set of daily weather variables since 1979 at high spatial resolution (4 km) across the United States. In our main results we follow Schlenker and Roberts (2009) and include only precipitation, beneficial heat exposure, and damaging heat exposure. In Appendix D.1, we include additional results with the dataset’s 9 raw variables: specific humidity, precipitation, minimum relative humidity, maximum relative humidity, surface downwelling shortwave flux in air (a measure of solar radiation), minimum air temperature, maximum air temperature, wind speed, and wind direction. We generate a county-level daily dataset by taking the average of daily weather from all grid observations within each U.S. county.

To construct a dataset we can use to study the affects of weather on annual crop yields, we aggregate daily weather observations up to the March-August growing season. We take averages of the variables specific humidity, minimum relative humidity, maximum relative humidity, surface downwelling shortwave flux in air (a measure of solar radiation), wind speed, and wind direction. We take the sum of precipitation. We then construct the variables beneficial and damaging heat exposure, total growing season heat exposure below and above (respectively) 29 °C. This gives us a county-level dataset with rich weather variation for our empirical application.

4.2 Empirical Results

In this section we present our primary empirical results. In our primary results the set of covariates we use comes from the current literature (Schlenker and Roberts 2009; Burke and Emerick 2016) and use total growing season precipitation and heat exposure below and above 29 °C. The specification is our preferred one as it allows for the cleanest comparison between DML and the current literature given that it only uses the variables currently used by applied researchers. In another set of empirical results in Appendix D.1 we use a second set of covariates. This second set of covariates includes go beyond the data used in applied work and include all weather variables at the total growing season level. Though this gives us more variables to work with, but may also make the interpretation of our effect harder to understand. We include results using this second set as they may be of interest and because the difference between estimates becomes smaller. In section we are comparing results using the first common set of weather variables and focus on the changes from DML relative to OLS Linear and not the changes from including additional weather features.

We estimate the model using the standardization, differencing, cross folds, analytical derivative, and de-biasing procedures described in Section 2. We use third order polynomial basis functions. We consider interactions up to depth 2 of all variables. Estimation of Lasso for

2. GridMET also includes several derived variables, such as Energy Release Component, for specific ecological applications.
both the regression problem and the Riesz representer uses the optimization package Mosek (ApS 2021). Table 3 shows the results from the main analysis.

We weight estimates by the area of corn planted when computing coefficients or average derivatives. This choice is common in the applied literature, as the object of interest is the elasticity of yield in an acre of corn with a marginal increase in damaging heat exposure. We follow Burke and Emerick (2016) and weight each observation by the area of corn planted in that county in that year. Appendix D.2 includes results without this weighting choice. Without weighting the difference between OLS and DML estimates become even larger.

The average derivative estimate can be interpreted as the elasticity of corn yield with respect to additional exposure to extreme heat. That is, each estimate is the percentage by which yields change with an additional growing degree day of heat exposure above 29 °C over the growing season. The magnitude is relatively large - this suggests that an increase of a single growing degree day is associated with corn yields declining by 0.51%-0.56%. This finding is line with findings from other analyses of damages to crop yields from extreme heat.

MSE from both in-sample and out-of-sample prediction is slightly higher for OLS Linear than any of the flexible methods. There is not a large difference between the MSE of the OLS Poly and Lasso models. This indicates that, for these sets of weather variables, OLS Poly does not over-fit the data. This is likely because we use a large data-set.

To illustrate the significance of these estimates, Figure 3 shows the projected overall change in crop yield by 2050 due solely to an increase in damaging heat exposure. We apply the estimated elasticity to climate projections from a range of scenarios. To generate these projections, we find the weighted average degree of expected warming from 2015 to 2050 under various climate scenarios, where we weigh the degree of warming by the average area of corn cultivated per county. Climate scenarios are derived from 18 global climate models running the A1B emissions scenario. The per-county degree of expected warming under each climate scenario are generously shared by Burke and Emerick (2016). We compare the OLS Linear and DML estimates.

These results illustrate that modeling assumptions significantly impact overall damage projections with the simple set of weather variables. There are large sources of uncertainty resulting from the varying climate projections, though the difference becomes smaller when using all weather covariates as seen in Appendix D.1. The projected damages are quite significant - in the median emissions scenario, log yields are 0.37-0.41 lower than a world that does not experience climate change. To put this in dollar value, the 2017 Census of Agriculture reported the total value of sales of corn in the United States as $51.2 billion. That range of damage estimates translates to a dollar value of $16.1-$17.3 billion (in 2017 dollars). Our preferred estimate, DML with the widely used simple set of weather covariates, finds an estimated damage of $17.3 billion.

The difference between OLS Linear and DML estimates is statistically significant and economically meaningful in our primary empirical results. OLS Poly and DML find similar results, although the standard error from OLS Poly is significantly higher than that of DML.

3. Burke and Emerick (2016) estimate this same elasticity in a range from -0.0036 to -0.0062, depending on the specification.
Table 2: Temperature and Precipitation

| method          | OLS Linear | OLS Poly | Lasso  | DML    |
|-----------------|------------|----------|--------|--------|
| Average Derivative | -0.005193  | -0.005657 | -0.005659 | -0.005664 |
|                  | (9.859e-05) | (0.0001352) | (1.049e-05) | (7.292e-05) |
| MSE γ In Sample  | 0.08093    | 0.07796   | 0.07805 | 0.07805 |
| MSE γ Cross Folds | 0.08097    | 0.07835   | 0.07835 | 0.07835 |
| Number of Observations | 63662     | 63662    | 63662  | 63662  |
| Number of Covariates | 3         | 36       | 36     | 36     |

Table 3: Estimates of elasticity of corn yields with respect to increase in growing season exposure to extreme heat, using two sets of weather covariates. Standard errors (in parentheses) are clustered at the county level. See text for estimation details.

Figure 2: Temperature and Precipitation

Figure 3: Extrapolating impacts of extreme heat to crop yields by the year 2050, using elasticities from OLS Linear and DML. Each dot represents a central estimate from a model, and the error bar represents the 95% confidence interval. Dotted line represents the median value across climate models. See text for estimation details.
(a) OLS Linear
Trend 3.43e-05, p-val 0.433

(b) DML
Trend 7.56e-05, p-val 0.538

Figure 4: Estimating elasticity of corn yield with respect to extreme heat over time. We use our estimation procedure on 2-period samples from 1980 through 2019. Line shows central estimate, and grey band shows 95% confidence interval.

The difference between OLS Linear and DML is significant with a p value of 0.00012, while the difference between OLS Linear and OLS Poly is significant with a p value of 0.0055. Median damage estimates using DML instead of OLS correspond to an additional $1.18 billion.

These values overstate the degree of damages from climate change for several reasons, but still provide a valuable illustration of the economic significance of using this measurement technique. This is not a complete projection of climate change damages, but only highlights the contribution of the elasticity estimated above using various modeling assumptions. Also note that an elasticity is the marginal impact at the observed distribution of weather characteristics; as the distribution of weather patterns changes with climate change, this elasticity will likely change. Additionally, this estimate does not attempt to account for adaptation to climate change.

We use our estimation procedure to study how the elasticity has changed over time, using 2-period panels from 1980 through 2019. A key advantage of our DML procedure is that it allows de-biased estimation of flexible functions, even in relatively short panels. This allows us to study how the elasticity has changed over time. This research design is similar to Barreca et al. (2016), who study the mortality response from temperature during separate decades across the 20th century to conclude that there has been significant adaptation to extreme heat exposure.

Our estimates suggest that the degree of damages from extreme heat has remained approximately stable from 1980 to 2019. Figure 4 shows these estimates, as well as the value and statistical significance of the trend in elasticity. The trend and standard errors are computed using weighted least squares, weighted by the inverse of the sum of the variance of the elasticity estimate and the residual from the regression of elasticity on the year. Using both methods with both sets of weather variables, we are unable to reject the null hypothesis at the $p = 0.05$ level, suggesting little adaptation to climate change. This finding is in line with the literature on adaptation in U.S. agriculture. Schlenker and Roberts (2009) and Burke and Emerick (2016) also find limited evidence of adaptation to extreme heat.
5 Conclusion

In this paper, we presented the first DML based estimator for continuous treatment effects in the classic fixed effect panel setting. It is also the first DML estimator in this setting to allow for general unrestricted heterogeneity in treatment effects. Our estimator is one of the first methods for flexibly estimating continuous treatment effects in high-dimensional panel data, and our estimation approach improves upon the current state-of-the-art methods in simulations. We gave clear guidance on how to implement our methods so they can be of use to an applied audience.

Our estimator is of broad applied interest - it allows for fixed effects in panel data, unrestricted heterogeneity in treatment effects, and ML tools, and keeps standard errors small even with very flexible models. This allows the researcher to flexibly model high-dimensional data while addressing unobservable fixed effects and machine learning bias.

This is one of the first applications of DML in environmental economics. We applied the estimator to study how extreme heat impacts crop yields. In our application, our DML estimator leads to estimates of treatment effects that are significantly different from the estimates based on OLS. Using a widely used set of weather variation, the difference is significant at the $p = 0.001$ level and economically meaningful. Our method predicts an effect of warmer temperature that is $1.18$ billion larger than current OLS based estimates.
References

Abatzoglou, John T. 2013. “Development of gridded surface meteorological data for ecological applications and modelling.” *International Journal of Climatology* 33 (1): 121–131.

ApS, MOSEK. 2021. *MOSEK Optimizer API for Python 9.3.6*. [https://docs.mosek.com/latest/pythonapi/index.html](https://docs.mosek.com/latest/pythonapi/index.html).

Barreca, Alan, Karen Clay, Olivier Deschênes, Michael Greenstone, and Joseph S Shapiro. 2016. “Adapting to climate change: The remarkable decline in the US temperature-mortality relationship over the twentieth century.” *Journal of Political Economy* 124 (1): 105–159.

Belloni, Alexandre, Victor Chernozhukov, Christian Hansen, and Damian Kozbur. 2016. “Inference in high-dimensional panel models with an application to gun control.” *Journal of Business & Economic Statistics* 34 (4): 590–605.

Bickel, Peter J, Ya'acov Ritov, Alexandre B Tsybakov, et al. 2009. “Simultaneous analysis of Lasso and Dantzig selector.” *The Annals of statistics* 37 (4): 1705–1732.

Burke, Marshall, and Kyle Emerick. 2016. “Adaptation to climate change: Evidence from US agriculture.” *American Economic Journal: Economic Policy* 8, no. 3 (August): 106–140. issn: 1945774X. [https://doi.org/10.1257/pol.20130025](https://doi.org/10.1257/pol.20130025).

Burke, Marshall, Solomon M Hsiang, and Edward Miguel. 2015. “Global non-linear effect of temperature on economic production” [in en]. *Nature* 527 (7577): 235–239. [https://doi.org/10.1038/nature15725](https://doi.org/10.1038/nature15725).

Chartrand, Rick. 2011. “Numerical differentiation of noisy, nonsmooth data.” *International Scholarly Research Notices* 2011.

———. 2017. “Numerical differentiation of noisy, nonsmooth, multidimensional data.” In *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 244–248. IEEE.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. 2017. “Double/debiased/neyman machine learning of treatment effects.” *American Economic Review* 107 (5): 261–65.

Chernozhukov, Victor, Juan Carlos Escanciano, Hidehiko Ichimura, Whitney K Newey, and James M Robins. 2016. “Locally robust semiparametric estimation.” *arXiv preprint arXiv:1608.00033*.

Chernozhukov, Victor, Matt Goldman, Vira Semenova, and Matt Taddy. 2017. “Orthogonal machine learning for demand estimation: High dimensional causal inference in dynamic panels.” *arXiv*, arXiv–1712.

Chernozhukov, Victor, Whitney Newey, and Rahul Singh. 2018. “De-biased machine learning of global and local parameters using regularized Riesz representers.” *arXiv preprint arXiv:1802.08667*. 
Chernozhukov, Victor, Whitney K Newey, and Rahul Singh. 2021. “A simple and general debiased machine learning theorem with finite sample guarantees.” *arXiv preprint arXiv:2105.15197.*

———. 2022. “Automatic debiased machine learning of causal and structural effects.” *Econometrica* 90 (3): 967–1027.

Colangelo, Kyle, and Ying-Ying Lee. 2020. “Double debiased machine learning nonparametric inference with continuous treatments.” *arXiv preprint arXiv:2004.03036.*

Crane-Droesch, Andrew. 2018. “Machine learning methods for crop yield prediction and climate change impact assessment in agriculture.” *Environ. Res. Lett* 13:114003. [https://doi.org/10.1088/1748-9326/aae159](https://doi.org/10.1088/1748-9326/aae159)

Dell, Melissa, Benjamin F. Jones, and Benjamin A. Olken. 2014. “What Do We Learn from the Weather? The New Climate-Economy Literature.” *Journal of Economic Literature* 52 (3): 740–798. [https://doi.org/10.1257/jel.52.3.740](https://doi.org/10.1257/jel.52.3.740)

Deryugina, Tatyana, Garth Heutel, Nolan H. Miller, David Molitor, and Julian Reif. 2019. “The mortality and medical costs of air pollution: Evidence from changes in wind direction.” *American Economic Review* 109 (12): 4178–4219. ISSN: 19447981. [https://doi.org/10.1257/aer.20180279](https://doi.org/10.1257/aer.20180279)

Deschênes, Olivier, and Michael Greenstone. 2007. “The economic impacts of climate change: Evidence from agricultural output and random fluctuations in weather.” *American Economic Review* 97, no. 1 (March): 354–385. ISSN: 00028282. [https://doi.org/10.1257/aer.97.1.354](https://doi.org/10.1257/aer.97.1.354)

Diamond, Steven, and Stephen Boyd. 2016. “CVXPY: A Python-embedded modeling language for convex optimization.” *Journal of Machine Learning Research* 17 (83): 1–5.

Hsiang, Solomon. 2016. “Climate econometrics.” *Annual Review of Resource Economics* 8:43–75.

Hsiang, Solomon, Robert Kopp, Amir Jina, James Rising, Michael Delgado, Shashank Mohan, D J Rasmussen, et al. 2017. “Estimating economic damage from climate change in the United States.” *Science* 356 (6345): 1362–1369. ISSN: 10959203. [https://doi.org/10.1126/science.aal4369](https://doi.org/10.1126/science.aal4369)

Imbens, Guido W, and Whitney K Newey. 2009. “Identification and estimation of triangular simultaneous equations models without additivity.” *Econometrica* 77 (5): 1481–1512.

Jean, Neal, Marshall Burke, Michael Xie, W. Matthew Davis, David B. Lobell, and Stefano Ermon. 2016. “Combining satellite imagery and machine learning to predict poverty.” *Science* 353 (6301): 790–794. ISSN: 10959203. [https://doi.org/10.1126/science.aaf7894](https://doi.org/10.1126/science.aaf7894)

Klosin, Sylvia. 2021. “Automatic Double Machine Learning for Continuous Treatment Effects.” *arXiv preprint arXiv:2104.10334.*
Knittel, Christopher R, and Samuel Stolper. 2019. *Using machine learning to target treatment: The case of household energy use*. Technical report. National Bureau of Economic Research.

Mboup, Mamadou, Cédric Join, and Michel Fliess. 2009. “Numerical differentiation with annihilators in noisy environment.” *Numerical algorithms* 50 (4): 439–467.

Rode, Ashwin, Tamma Carleton, Michael Delgado, Michael Greenstone, Trevor Houser, Solomon Hsiang, Andrew Hultgren, Amir Jina, Robert E Kopp, Kelly E McCusker, et al. 2021. “Estimating a social cost of carbon for global energy consumption.” *Nature* 598 (7880): 308–314.

Rothenhäusler, Dominik, and Bin Yu. 2019. “Incremental causal effects.” *arXiv preprint arXiv:1907.13258*.

Schlenker, Wolfram, and Michael J Roberts. 2009. “Nonlinear temperature effects indicate severe damages to U.S. crop yields under climate change.” *Proceedings of the National Academy of Sciences of the United States of America* 106 (37): 15594–8. issn: 1091-6490. https://doi.org/10.1073/pnas.0906865106. http://www.ncbi.nlm.nih.gov/pubmed/19717432%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2747166.

Stetter, Christian, Philipp Mennig, and Johannes Sauer. 2022. “Using Machine Learning to Identify Heterogeneous Impacts of Agri-Environment Schemes in the EU: A Case Study.” *European Review of Agricultural Economics*.

Van Breugel, Floris, J Nathan Kutz, and Bingni W Brunton. 2020. “Numerical differentiation of noisy data: A unifying multi-objective optimization framework.” *IEEE Access* 8:196865–196877.

Wooldridge, Jeffrey M. 2005. “Violating ignorability of treatment by controlling for too many factors.” *Econometric Theory* 21 (5): 1026–1028.

———. 2010. *Econometric analysis of cross section and panel data*. MIT press.
A Asymptotic Normality

Proving our estimator is asymptotically normal follows from Theorem 14 of Chernozhukov et al. (2016). Theorem 14 gives us that if Assumptions (1), (2), and (3) below are satisfied, asymptotic normality follows. In this appendix we explain either why our estimator satisfies these assumptions, or give primitive conditions that justify the assumptions.

We start with notation used in the three assumptions before introducing them. Recall from equation (2) from earlier our parameter of interest \( \tau_0 \) is defined to be the solution to the following moment equation.

\[
\tau_0 = \mathbb{E}[m(W, \gamma)] = \mathbb{E}\left[ \frac{\hat{\gamma}_0(D_{i,t}, X_{i,t})}{\partial D_{i,t}} \right]
\]

We define

\[
g(W, \gamma, \tau) = m(W, \gamma) - \tau
\]

We define \( \phi \) to be our first step influence function. In our application \( \phi = \alpha_0(Y - \gamma_0) \).

Adding the first step influence function to our moment function to create our de-biased moment function \( \psi \).

\[
\psi(W, \gamma, \alpha_0, \tau_0) = g(W, \gamma, \tau) + \phi(W, \gamma, \alpha_0, \tau_0)
\]

\[
= g(W, \gamma, \tau) + \alpha_0(Y - \gamma_0)
\]

We also define \( \hat{\delta}_\ell \) below. This is a known as the interaction reminder, it shows up in the decomposition of \( \hat{\psi} - \psi \), and we have to make an assumption about its behavior for asymptotic linearity.

\[
\hat{\delta}_\ell = \hat{\alpha}_\ell(Y - \hat{\gamma}_\ell) - \alpha_0(Y - \gamma_0) - \alpha_0(Y - \gamma_0) - \alpha_0(Y - \gamma_0)
\]

Assumption 1. (mild mean-square consistency) : \( E[\|\psi(W, \tau_0, \gamma_0, \alpha_0)^2]\] < \( \infty \)

\[
i) \int \|g(w, \gamma_\ell, \tau_0) - g(w, \gamma_0, \tau_0)\|^2 F_0(dw) \overset{p}{\to} 0
\]

\[
ii) \int \|\alpha_0(Y - \hat{\gamma}_\ell) - \alpha_0(Y - \gamma_0)\|^2 F_0(dw) \overset{p}{\to} 0
\]

\[
iii) \int \|\hat{\alpha}_\ell(Y - \gamma_0) - \alpha_0(Y - \gamma_0)\|^2 F_0(dw) \overset{p}{\to} 0
\]

4. note in the original paper by Chernozhukov et al. (2016) this object was defined to be capital \( \Delta_\ell \) but given that capital delta is used to describe differences over time we changed notation.

20
These are mild mean-square consistency conditions for $\hat{\gamma}_\ell$ and $(\hat{\alpha}_\ell, \hat{\theta}_\ell)$

Assumption 2. (rate on interaction remainder) For each $\ell = 1, \ldots, L$

$$\sqrt{n} \int \hat{\delta}_\ell(w) F_0(dw) \overset{p}{\rightharpoonup} 0, \int \|\hat{\delta}_\ell\|^2(w) F_0(dw) \overset{p}{\rightharpoonup} 0$$

(19)

Assumption 3. (double robust) For each $\ell = 1, \ldots, L$

i) $\int \phi(w, \gamma_0, \hat{\alpha}_0, \theta_0) F_0(dw) = 0$ with probability approaching one

ii) $\mathbb{E}[\psi(\gamma, \alpha_0, \theta_0)]$ is affine in $\gamma$

As noted in Chernozhukov et al. (2016), in our case because $\mathbb{E}[\psi(\gamma, \alpha_0, \theta_0)]$ is affine in $\gamma$, Assumption 3 imposes no conditions additional to Assumption 1 and 2, but we write it for clarity.

A.1 Primitive Conditions for Assumptions

Primitive Conditions for Assumption (1) Now we go into detail about the primitive conditions for the assumptions starting with Assumption (1) part i). Plugging in the definition of $g$ into the equation, we get that this assumption is the following. Given that our moment is one dimensional we drop the norm notation for clarity.

$$\mathbb{E} \left[ \left( \frac{\partial \hat{\gamma}_\ell}{\partial D_{i,t}} - \theta_0 \right) \frac{\partial \gamma_0}{\partial D_{i,t}} \right]^2 F_0(dw) \overset{p}{\rightharpoonup} 0$$

(20)

$$\mathbb{E} \left[ \left( \frac{\partial \hat{\gamma}_\ell}{\partial D_{i,t}} - \frac{\partial \gamma_0}{\partial D_{i,t}} \right)^2 F_0(dw) \right] \overset{p}{\rightharpoonup} 0$$

(21)

For conciseness this we write this with expectation notation.

$$\mathbb{E} \left[ \left( \frac{\partial \hat{\gamma}_\ell}{\partial D_{i,t}} - \frac{\partial \gamma_0}{\partial D_{i,t}} \right)^2 \right] \overset{p}{\rightharpoonup} 0$$

(22)

Which is a mean square consistency condition for the average derivative. Mean square consistency for the average derivative follows from two primitive assumptions 1) mean square consistency first step, stated in Assumption (1) and 2) mean square continuity for average derivative, stated in Assumption (5). We start by writing down these two assumptions and show how together they imply mean square consistency for the average derivative.

The first primitive assumption is mean square consistency of the first step estimator $\hat{\gamma}$.

Assumption 4. (mean square consistency first step)

$$\|\hat{\gamma}_\ell - \gamma_0\|_{F,2} \overset{p}{\rightharpoonup} 0$$

(23)
The second primitive assumption is mean square continuity for average derivative.

**Assumption 5. (mean square continuity for average derivative)**

First we define a function class $\Gamma_Q$ where $Q < \infty$ and such that for all $\gamma \in \Gamma_Q$

$$\mathbb{E}\left(\frac{\partial \gamma}{\partial D_{i,t}}^2\right) \leq Q[\mathbb{E}(\gamma^2)]^{\frac{1}{2}}$$

(24)

We assume that $\gamma_0 \in \Gamma_Q$ and $\hat{\gamma} \in \Gamma_Q$ with high probability.

Now to show that Assumption (4) and Assumption (5) give us the mean square consistency for the average derivative.

**Proof.**

$$\mathbb{E}\left[\left(\frac{\partial \hat{\gamma}_0}{\partial D_{i,t}} - \frac{\partial \gamma_0}{\partial D_{i,t}}\right)^2\right] = \mathbb{E}\left[\left(\frac{\partial (\hat{\gamma}_0 - \gamma_0)}{\partial D_{i,t}}\right)^2\right]$$

(25)

$$\leq \hat{Q}[\mathbb{E}(\hat{\gamma}_0 - \gamma_0)^2]^{\frac{1}{2}}$$

(26)

$$= \hat{Q}\|\hat{\gamma}_0 - \gamma_0\|_{F,2} \overset{p}{\rightarrow} 0$$

(27)

We can bound the left hand side of equation (22) using the two assumptions. First the linearity of the average derivative give us the first equality below. Then the inequality follows by mean square continuity since it holds for every realization of $\hat{\gamma}$. Then the last line follows by Assumption (5). Hence we have mean square consistency condition for the average derivative.

Please note that further primitive conditions Assumption (5) are given in Chernozhukov, W. K. Newey, and Singh (2021) in Lemma B.4. Though we leave the details to Chernozhukov, W. K. Newey, and Singh (2021), in short the conditions are given in Assumption (6).

**Assumption 6. Primitive conditions for Assumption (5)**

i) $f(d|x)$ vanishes for each $d$ in the boundary of the support of $D$ given $X = x$ almost everywhere

ii) $-\partial_d \log f(d|x)$ is bounded above

iii) $\Gamma$ consists of functions $\gamma$ that are twice continuously differentiable in the first argument and satisfy $\mathbb{E}[\left(\partial_d \gamma(D, X)\right)^2] < \infty$ and $\mathbb{E}[\left(\partial_d^2 \gamma(D, X)\right)^2] < \infty$.

Assumption (6) is satisfied if $\Gamma$ satisfies Sobolev conditions.
**Primitive Conditions for Assumption** [2] The rate condition on the interaction term of Assumption [2] requires that $\hat{\alpha}_\ell$ and $\hat{\gamma}_\ell$ can be estimated fast enough.

$$
\|\hat{\alpha}_\ell - \alpha_0\|_{F,2} \|\hat{\gamma}_\ell - \gamma_0\|_{F,2} = o_p(n^{1/2})
$$

(28)

Rates and conditions for Lasso estimation of $\hat{\gamma}_\ell$ are given in Bickel, Ritov, Tsybakov, et al. (2009). As for $\hat{\alpha}_\ell$, either an approximate sparsity assumption or sparse eigenvalue assumption can justify rates needed for (28) as explained in detail in Chernozhukov, W. K. Newey, and Singh (2022).

**B Details of Estimation Procedure**

**B.1 Analytical vs. Numerical Derivative**

We propose calculating the derivative analytically, as opposed to the numerical methods. In current DML and Auto DML papers, derivatives are computed using numerical differentiation. To explain what we mean by analytical vs numerical let’s consider a function $\gamma(D, X)$, and let’s say we have an estimate of this function $\hat{\gamma}(D, X)$ and we want to estimate the derivative at point $D = D_0$ and $X = X_0$. To estimate derivative numerically we could use Newton’s difference quotient (also known as a first-order divided difference) and pick some small $h$.

$$
\frac{\hat{\gamma}(D_0 + h, X_0) - \hat{\gamma}(D_0, X_0)}{h} = b(D_0 + h, X_0)\hat{\beta} - b(D_0, X_0)\hat{\beta}
$$

(29)

It is well known that this procedure can introduce biases, either through formula error or rounding error. Formula error is introduced because, for most cases, the difference between the true derivative and the numerical approximation is decreasing in $h$. Formula error is most relevant when $h$ is large. Rounding error is introduced during computation with small $h$, as computers must round floating point numbers in order to carry out computation.

There are approaches for reducing the error of numerical differentiation, although these approaches are still under development especially for noisy data. Simple modifications include taking a symmetric difference instead of a one-sided difference. The problem of numerical differentiation with noisy data is more challenging and the subject of ongoing research (Mboup, Join, and Fliess 2009; Chartrand 2011, 2017; Van Breugel, Kutz, and Brunton 2020). By taking analytical derivatives of our basis function, we avoid these numerical challenges.

**B.2 Tuning**

In this section we explain tuning $\hat{\alpha}$ and $\hat{\gamma}$.

We use a data-driven process to select the hyperparameters for Lasso and the Reisz representer function. We select hyperparameters by minimizing loss on the test set during each fold of our cross validation procedure. Let $\hat{\gamma}_\ell$ and $\hat{\alpha}_\ell$ denote the estimates of $\gamma$ and $\alpha$ trained using indices not in set $\ell$ using the above procedures, for the given hyperparameter value.
Let $\hat{\beta}_\ell$ and $\hat{\rho}_\ell$ be the corresponding estimates of parameter vectors in our sparse linear models. Recall that $W_\ell$ denotes all observations in the fold $\ell$. Let $L_\gamma(\gamma, W_\ell; r_L)$ be the sum of squared error of the function $\gamma$ with the hyperparameter $r_L$ on the data in $I_\ell$:

$$L_\gamma(\gamma, W_\ell; r_L) = \sum_i \sum_t (\Delta \hat{\gamma}_\ell(D_{i,t}, X_{i,t}) - \Delta Y_{\ell,t})^2$$

Let $L_\alpha(\alpha, I_\ell; r_\alpha)$ be the loss function of the function $\alpha$ with the hyperparameter $r_\alpha$ on the data in $I_\ell$. That loss function is the sum of the distance between $\alpha_0$ and $\hat{\alpha}_\ell$ using our dictionary of basis functions. See Appendix B.4 for an explanation of this loss function. Minimizing this distance is equivalent to minimizing:

$$L_\alpha(\hat{\alpha}_\ell, W_\ell; r_\alpha) = \sum_i \sum_t -2b_D(D_{i,t}, X_{i,t})^T \hat{\rho}_\ell + \hat{\rho}_\ell \Delta b(D_{i,t}, X_{i,t}) \Delta b(D_{i,t}, X_{i,t})^T \hat{\rho}_\ell$$

Then, select hyperparameter $r_L$ that minimize test-set mean squared error of the regression:

$$r_L = \arg\min_r \frac{1}{n(T-1)} \sum_{\ell=1}^k L_\gamma(\hat{\gamma}_\ell, W_\ell; r)$$

And select hyperparameter $r_\alpha$ that minimize test-set loss of the Riesz representer:

$$r_\alpha = \arg\min_r \frac{1}{n(T-1)} \sum_{\ell=1}^k L_\alpha(\hat{\alpha}_\ell, W_\ell; r)$$

### B.3 Normalization

Many machine learning models perform better when the independent variables are standardized, that is when the data has mean zero and variance 1. In this section, we include some details about how to conduct this standardization so that the researcher is able to recover derivatives after that step.

For each basis function $b^j$ for $j = 1, \ldots, p$ in the dictionary $b$, define the mean and standard deviation of the transformed data: $\mu^j := \mathbb{E}[b^j(W_i)]$ and $\sigma^j := \sqrt{\mathbb{E}[(b^j(W_i) - \mu^j)^2]}$. Their sample equivalents are: $\hat{\mu}^j := \frac{1}{n} \sum_i b^j(W_i)$ and $\hat{\sigma}^j := \sqrt{\frac{1}{N-1} \sum_i (b^j(W_i) - \hat{\mu}^j)^2}$.

To generate a standardized basis function, we apply the following transformation: $\hat{b}^j := (b^j(W_i) - \hat{\mu}^j) / \hat{\sigma}^j$. We assumed that there was some $\beta_0$ such that $\gamma_0(W_i) = \beta_0 b(W_i)$. Then there exists $\tilde{\beta}_0$ such that $\gamma_0(W_i) = \tilde{\beta}_0 \hat{b}(W_i) + C$, where $\tilde{\beta}_0 \sigma^j = \beta_0^j$ for all $j$ components, $\hat{b}$ is the dictionary of all standardized basis functions $\hat{b}^j$, and $C$ is some generic constant. This is easy to confirm via algebraic manipulation. When we take differences to remove the unobserved individual fixed effect, this $C$ term is also removed.

Recall that the average derivative is $\mathbb{E}[\beta_0 b_D(W_i)]$. We could also write this in terms of $\tilde{\beta}_0$, with the fact that $\beta_0^j = \tilde{\beta}_0 \sigma^j$. Let $\sigma = \{\sigma^1, \ldots, \sigma^p\}$, and $\sigma^{-1} = \{1/\sigma^1, \ldots, 1/\sigma^p\}$ Then,
we write this relationship more compactly as \( \mathbb{E}[(\hat{\beta}_0 \circ \sigma^{-1})b_D(W_i)] \), where \( \circ \) is elementwise multiplication or the Hadamard product (i.e. \( \hat{\beta}_0 \circ \sigma^{-1} = \{\hat{\beta}_0^1/\sigma^1, \ldots, \hat{\beta}_0^p/\sigma^p\} \)).

In our estimation procedure, we found it more convenient to producing scaled derivatives of each basis function and multiplying by the beta estimate from using scaled data. Define the scaled derivative of each basis function, \( \hat{b}_D^j := b_D^j/\hat{\sigma}^j \). Then, the average derivative is: \( \mathbb{E}[\hat{\beta}_0 \hat{b}_D^j(W_{i,t})] \). These procedures are equivalent, as can be confirmed through algebraic manipulation.

This suggests our procedure to standardize data and recover the derivative:

1. For each basis function \( b^j \) for \( j = 1, \ldots, p \), find \( \hat{\mu}^j \) and \( \hat{\sigma}^j \). Store these estimates.
2. Create the sample standardized basis function and its derivative, \( \hat{b}^j := (b^j(W_i) - \hat{\mu}^j)/\hat{\sigma}^j \) and \( \hat{b}_D^j := b_D^j/\hat{\sigma}^j \).
3. Find an estimate \( \hat{\beta} \) that satisfies the regression \( \mathbb{E}[\Delta y_{i,t}] = \hat{\beta} \Delta \hat{b}(W_{i,t}) \).
   Where \( \Delta \hat{b}(W_{i,t}) := \hat{b}(W_{i,t}) - \hat{b}(W_{i,t-1}) \). This could be via OLS or a cross-folds Lasso procedure.
4. Estimate the average derivative as \( \mathbb{E}[^{\hat{\beta}b_D(W_{i,t})}] \).

Standardization of our basis functions is also relevant for estimating the Riesz representer. After the standardization, our Riesz representer now takes the form \( \hat{\alpha}(W_{i,t}) = \hat{\beta}b(W_{i,t})\hat{\rho} \). Before standardizing the data, we had an estimator of the form:

\[
\hat{\rho}_{\text{original}} = \arg\min_{\rho} -2\hat{M}\rho + \rho'\hat{Q}\rho + \lambda|\rho|_1
\]

where \( \hat{M} = \mathbb{E}[b_D(W_{i,t})] \) and \( \hat{Q} = \mathbb{E}[\Delta b(W_{i,t})'\Delta b(W_{i,t})] \). After applying the standardization, we are taking the derivative of the standardized basis functions, so this estimator now takes the form:

\[
\hat{\rho} = \arg\min_{\rho} -2\hat{M}\rho + \rho'\hat{Q}\rho + \lambda|\rho|_1
\]

where \( \hat{M} = \mathbb{E}[\hat{b}_D(W_{i,t})] \) and \( \hat{Q} = \mathbb{E}[\Delta \hat{b}(W_{i,t})'\Delta \hat{b}(W_{i,t})] \). We use \( \hat{b}_D^j(W_{i,t}) = b_D^j(W_{i,t})/\hat{\sigma}^j \), although this neglects an additional correction because standardization involves estimating the mean and variance of the dataset. Below, we include a derivation of a full bias correction term to account for this estimation, and note this bias correction term is equivalent to the above expression for large \( n \). In our estimation, we therefore use the values \( \hat{b}_D^j(W_{i,t}) = b_D^j(W_{i,t})/\hat{\sigma}^j \) to construct \( \hat{M} \) and estimate \( \hat{\alpha} \) using [30].

In the remainder of this section, we proceed with the full derivation of the term \( \hat{\sigma}b^j(W_{i,t})/\hat{D}_{i,t} \).

Recall that we use the notation \( \xi_D \) to denote the partial derivative of \( \xi \) with respect to \( D \).

5. This form can also be motivated by the chain rule, taking the derivative of the standardized data. In this case, an additional bias correction would be necessary because we are estimating \( \hat{\mu}^j \) and \( \hat{\sigma}^j \). See Appendix [B.3] for more details.
We use this estimator to find \( \hat{\rho} \). Equality comes from the definition \( \hat{\rho} \).

Where the 3rd equality comes from applying the Riesz Representation theorem, and the 4th

\[
\hat{\sigma}_D^j = \frac{1}{2} \left( \frac{\sum_i (b^j(W_i) - \hat{\mu}^j)^2}{N - 1} \right)^{-1/2} \frac{1}{N - 1} \sum_k 2(b^j(W_i) - \hat{\mu}^j)(b^j_D(W_i) - \hat{\mu}_D^j)
\]

\[
\sum_k (b^j(W_i) - \hat{\mu}_D^j)(b^j_D(W_i) - \hat{\mu}_D^j) = (b^j(W_i) - \hat{\mu}_D^j)b^j_D(W_i) - \sum_k (b^j(W_i) - \hat{\mu}_D^j)\hat{\mu}_D^j
\]

\[
\sum_k (b^j(W_i) - \hat{\mu}_D^j)\hat{\mu}_D^j = 0
\]

\[
\sum_k (b^j(W_i) - \hat{\mu}_D^j)b^j_D(W_i) = (b^j(W_i) - \hat{\mu}_D^j)b^j_D(W_i)
\]

\[
\frac{\partial}{\partial \hat{\sigma}_D^j} \frac{b^j(W_i) - \hat{\mu}_D^j}{\hat{\sigma}_D^j} = \frac{b^j_D(W_i)}{\hat{\sigma}_D^j} \left( \frac{N - 1}{N} - \frac{(b^j(W_i) - \hat{\mu}_D^j)^2}{(N - 1)(\hat{\sigma}_D^j)^2} \right)
\]

Note that \( \frac{b^j_D(W_i)}{\hat{\sigma}_D^j} - \frac{b^j_D(W_i)}{\hat{\sigma}_D^j} \left( \frac{N - 1}{N} - \frac{(b^j(W_i) - \hat{\mu}_D^j)^2}{(N - 1)(\hat{\sigma}_D^j)^2} \right) = O(1/N) \). This term is therefore negligible, as it converges to 0 faster than the \( \sqrt{N} \) convergence rate.

### B.4 Riesz representer Details

Our goal is to find the estimator \( \hat{\alpha}(W_{i,t}) \) that minimizes the mean squared error (MSE) between \( \hat{\alpha} \) and \( \alpha_0 \):

\[
\hat{\alpha} = \arg\min_{\alpha} \mathbb{E}[(\alpha_0(W_{i,t}) - \alpha(W_{i,t}))^2]
\]

Plugging in our guess at the functional form, \( \hat{\alpha} = \Delta b(W_{i,t})\hat{\rho} \), we form the following regularized problem:

\[
\hat{\rho} = \arg\min_{\rho} \mathbb{E} \left[ (\alpha_0(W_{i,t}) - \Delta b(W_{i,t})\rho)^2 \right] + \lambda|\rho|_1
\]

\[
= \arg\min_{\rho} \mathbb{E}[\alpha_0(W_{i,t})^2 - 2\Delta b(W_{i,t})\alpha_0(W_{i,t})\rho + \rho^T \Delta b(W_{i,t})' \Delta b(W_{i,t})\rho] + \lambda|\rho|_1
\]

\[
= \arg\min_{\rho} -2\mathbb{E}[b_d(W_{i,t})]\rho + \rho^T \mathbb{E}[\Delta b(W_{i,t})' \Delta b(W_{i,t})]\rho + \lambda|\rho|_1
\]

\[
= \arg\min_{\rho} -2\hat{M}\rho + \rho^T \hat{Q}\rho + \lambda|\rho|_1
\]

Where the 3rd equality comes from applying the Riesz Representation theorem, and the 4th equality comes from the definition \( \hat{M} = \mathbb{E}[b_d(W_{i,t})] \) and \( \hat{Q} = \mathbb{E}[\Delta b(W_{i,t})' \Delta b(W_{i,t})] \).

We use this estimator to find \( \hat{\rho} \); we use an optimization package to find the optimal value of \( \hat{\rho} \). Chernozhukov, W. K. Newey, and Singh (2022) provides an iterative approach.
More generally, for any linear functional $m$ and any functional form for $\hat{\alpha}$:

$$\hat{\alpha} = \arg\min_\alpha \mathbb{E}[(\alpha_0(W_{i,t}) - \alpha(W_{i,t}))^2]$$

$$= \arg\min_\alpha \mathbb{E}[\alpha_0(W_{i,t})^2] - 2\mathbb{E}[\alpha_0(W_{i,t})\alpha(W_{i,t})] + \mathbb{E}[\alpha(W_{i,t})^2]$$

$$= \arg\min_\alpha -2\mathbb{E}[m(W_{i,t}, \alpha)] + \mathbb{E}[\alpha(W_{i,t})^2]$$

Plugging in our estimator form, we get $\mathbb{E}[m(W_{i,t}, \alpha)] = \mathbb{E}[b_D(W_{i,t})\rho]$ and $\mathbb{E}[\alpha(W_{i,t})^2] = \rho'\mathbb{E}[\Delta b(W_{i,t})'\Delta b(W_{i,t})]\rho$, confirming the above result.

### C Additional Simulation Results

Here, we include simulation results from different parameters of the data generating process described in Section 3. Each table summarizes 1000 bootstrap trials, for the specified data generating process. We vary the number of covariates and the number of time periods in comparison to our main simulation results in the paper which have $T = 2$ and 20 covariates.

These results show that, as expected, the bias generally decreases as the number of samples per covariate increases. This is particularly true of OLS Poly. OLS Poly has the lowest bias in these additional trials, but has higher mean squared error in estimating the true derivative (MSE $\tau$) than DML, DML Iterative, or Lasso in all trials. DML has lower bias than DML Iterative in Table 4 and Table 5, and lower MSE $\tau$ in all trials.

| method | DML | DML Iterative | Lasso | OLS Linear | OLS Poly |
|--------|-----|---------------|-------|------------|----------|
| True Value | 2.937 | 2.937 | 2.937 | 2.937 | 2.937 |
| Average Derivative | 2.929 | 2.909 | 2.819 | 3.245 | 2.932 |
| Bias | -0.007124 | -0.02712 | -0.1173 | 0.3082 | -0.004251 |
| Standard Deviation | 0.3187 | 0.3498 | 0.311 | 0.3553 | 0.5017 |
| MSE $\tau$ | 0.09302 | 0.1141 | 0.1021 | 0.2152 | 0.2436 |
| Coverage | 0.906 | 0.942 | 0.287 | 0.845 | 0.942 |
| MSE $\gamma$ In Sample | 1.945 | 1.945 | 1.945 | 2.349 | 1.752 |
| MSE $\gamma$ Cross Folds | 2.033 | 2.033 | 2.033 | 2.411 | 3.59 |

Table 4: Summary of derivative estimates from 1000 bootstrap trials of our simulation procedure. Estimates use $N = 1000$, $T = 2$, and 10 covariates. Flexible basis functions include 3rd order polynomial functions of all terms and all interactions of $D$ and $X$ terms. After applying the basis function transformation, $p = 124$. Bias is the average of the estimated value of the derivative minus the true value of the derivative in each simulation draw. “MSE $\tau$” is the mean squared error between the true average derivative and the estimated average derivative in each simulation draw, while “MSE $\gamma$ in sample” and “MSE $\gamma$ cross folds” refer to the mean squared error of regression from own-sample and out-of-sample estimation.
### Table 5: Summary of derivative estimates from 1000 bootstrap trials of our simulation procedure. Estimates use $N = 1000$, $T = 5$, and 10 covariates. Flexible basis functions include 3rd order polynomial functions of all terms and all interactions of $D$ and $X$ terms. After applying the basis function transformation, $p = 124$. Bias is the average of the estimated value of the derivative minus the true value of the derivative in each simulation draw. “MSE $\tau$” is the mean squared error between the true average derivative and the estimated average derivative in each simulation draw, while “MSE $\gamma$ in sample” and “MSE $\gamma$ cross folds” refer to the mean squared error of regression from own-sample and out-of-sample estimation.

| method                | DML | DML Iterative | Lasso | OLS Linear | OLS Poly |
|-----------------------|-----|---------------|-------|------------|----------|
| True Value            | 2.936 | 2.936        | 2.936 | 2.936      | 2.936    |
| Average Derivative    | 2.935 | 2.922        | 2.864 | 3.239      | 2.936    |
| Bias                  | -0.0006326 | -0.01354    | -0.07144 | 0.3038 | 0.0003044 |
| Standard Deviation    | 0.2258 | 0.2425      | 0.2187 | 0.2272 | 0.2925 |
| MSE $\tau$            | 0.03628 | 0.04374     | 0.03838 | 0.1293 | 0.06965 |
| Coverage              | 0.9   | 0.923        | 0.289 | 0.678     | 0.937   |
| MSE $\gamma$ In Sample | 1.988 | 1.988        | 1.988 | 2.366      | 1.918    |
| MSE $\gamma$ Cross Folds | 2.01 | 2.01         | 2.01  | 2.388     | 2.207   |

### Table 6: Summary of derivative estimates from 1000 bootstrap trials of our simulation procedure. Estimates use $N = 1000$, $T = 5$, and 20 covariates. Flexible basis functions include 3rd order polynomial functions of all terms and all interactions of $D$ and $X$ terms. After applying the basis function transformation, $p = 244$. Bias is the average of the estimated value of the derivative minus the true value of the derivative in each simulation draw. “MSE $\tau$” is the mean squared error between the true average derivative and the estimated average derivative in each simulation draw, while “MSE $\gamma$ in sample” and “MSE $\gamma$ cross folds” refer to the mean squared error of regression from own-sample and out-of-sample estimation.

| method                | DML | DML Iterative | Lasso | OLS Linear | OLS Poly |
|-----------------------|-----|---------------|-------|------------|----------|
| True Value            | 2.964 | 2.964        | 2.964 | 2.964      | 2.964    |
| Average Derivative    | 2.977 | 2.96         | 2.902 | 3.258      | 2.973    |
| Bias                  | 0.01301 | -0.003874     | -0.06199 | 0.2943 | 0.008642 |
| Standard Deviation    | 0.2292 | 0.25         | 0.2222 | 0.2351     | 0.2878  |
| MSE $\tau$            | 0.03879 | 0.04942      | 0.0391 | 0.1245     | 0.07208 |
| Coverage              | 0.8747 | 0.9114       | 0.27  | 0.6976     | 0.9287  |
| MSE $\gamma$ In Sample | 1.98 | 1.98         | 1.98  | 2.369      | 1.838   |
| MSE $\gamma$ Cross Folds | 2.013 | 2.013       | 2.013 | 2.409      | 2.577   |

### D Additional Applied Results

#### D.1 Including New Weather Variables

The results are relatively similar between the two sets of weather variation. The results from the paper are presented again in Table 7a next to the results including more
covariates in Table 7b. OLS Linear finds a lower magnitude of the average derivative than the flexible models (OLS Poly, Lasso, and DML), all of which find relatively consistent estimates of the average derivative. The estimates from OLS Poly have the highest variance among any model, with roughly twice the standard deviation as OLS Linear. DML has a slightly lower variance than OLS Linear, and is significantly higher than Lasso. Based on the simulation results, we expect that the standard errors with Lasso do not accurately reflect the uncertainty around the estimate.

| (a) Temperature and Precipitation | method  | OLS Linear | OLS Poly | Lasso | DML |
|----------------------------------|---------|------------|----------|-------|-----|
| Average Derivative               | -0.005193 | -0.005657 | -0.005659 | -0.005664 |
|                                  | (9.859e-05) | (0.0001352) | (1.049e-05) | (7.292e-05) |
| MSE γ In Sample                  | 0.08093 | 0.07796 | 0.07805 | 0.07805 |
| MSE γ Cross Folds                | 0.08097 | 0.07835 | 0.07835 | 0.07835 |
| Number of Observations           | 63662 | 63662 | 63662 | 63662 |
| Number of Covariates             | 3 | 36 | 36 | 36 |

| (b) All Weather Covariates       | method  | OLS Linear | OLS Poly | Lasso | DML |
|----------------------------------|---------|------------|----------|-------|-----|
| Average Derivative               | -0.005402 | -0.005247 | -0.005298 | -0.005283 |
|                                  | (0.0001026) | (0.0002392) | (1.917e-05) | (7.862e-05) |
| MSE γ In Sample                  | 0.07988 | 0.06552 | 0.06538 | 0.06538 |
| MSE γ Cross Folds                | 0.07994 | 0.06795 | 0.06784 | 0.06784 |
| Number of Observations           | 63662 | 63662 | 63662 | 63662 |
| Number of Covariates             | 9 | 351 | 351 | 351 |

D.2 Agriculture Results Without Weighting

In our main estimates, we weight the results by the area of corn planted per county; the average derivative we estimate in that panel is the effect averaged over all acres of corn. This estimate is commonly used in the literature, including by Schlenker and Roberts (2009) and Burke and Emerick (2016). It may also be of interest to consider the effect without weighting, which gives the effect averaged over all U.S. counties. This appendix includes those results. Table 8 summarizes the estimation results, using each estimation method and the two sets of covariates. Figure 6 shows the extrapolated damages from extreme heat by the year 2050, using the same weighted procedure as in Burke and Emerick (2016) to generate these extrapolations. That range of damage estimates translates to a dollar value of $17.1-$21.7 billion (in 2017 dollars). Our preferred estimate, DML with all weather covariates, finds an estimated damage of $21.7 billion. The difference between OLS Linear and DML estimates is statistically significant and economically meaningful. Median damage estimates using DML instead of OLS correspond to an additional $4.23 and $3.84 billion using the simple set of covariates and the set of all weather variables, respectively. This difference in parameter estimates is significant at the $p = 0.001$ level.

Figure 7 shows the estimated elasticity over time, without weighting to adjust for the crop
level. Using both methods for the simple set of covariates, the mean trend in the elasticity is positive and we reject the null hypothesis at the 0.05 level. Using all covariates, we fail to reject the null hypothesis at the 0.05 level for either estimation method.

(a) Temperature and Precipitation

| method                      | OLS Linear | OLS Poly | Lasso   | DML    |
|-----------------------------|------------|----------|---------|--------|
| Average Derivative          | -0.005738  | -0.007558| -0.007578| -0.007577|
|                             | (8.48e-05) | (0.0001199)| (1.182e-05)| (6.431e-05) |
| MSE γ In Sample             | 0.07901    | 0.07125  | 0.07124  | 0.07124 |
| MSE γ Cross Folds           | 0.07911    | 0.07166  | 0.07161  | 0.07161 |
| Number of Observations      | 63662      | 63662    | 63662    | 63662  |
| Number of Covariates        | 3          | 36       | 36       | 36     |

(b) All Covariates

| method                      | OLS Linear | OLS Poly | Lasso   | DML    |
|-----------------------------|------------|----------|---------|--------|
| Average Derivative          | -0.005586  | -0.007237| -0.007229| -0.007226|
|                             | (9.5e-05)  | (0.0001926)| (1.708e-05)| (6.417e-05) |
| MSE γ In Sample             | 0.07785    | 0.05822  | 0.05802  | 0.05802 |
| MSE γ Cross Folds           | 0.07797    | 0.06054  | 0.06054  | 0.06054 |
| Number of Observations      | 63662      | 63662    | 63662    | 63662  |
| Number of Covariates        | 9          | 351      | 351      | 351    |

Table 8: Using the weighted estimates. Estimates of elasticity of corn yields with respect to increase in growing season exposure to extreme heat, using two sets of weather covariates. Standard errors (in parentheses) are clustered at the county level. See text for estimation details.
Figure 5: Extrapolating impacts of extreme heat to crop yields by the year 2050, using elasticities from OLS Linear and DML. Each dot represents a central estimate from a model, and the error bar represents the 95% confidence interval. Dotted line represents the median value across climate models. See text for estimation details.

Figure 6: Same as figure above using the non-weighted estimates.
Figure 7: Estimating elasticity of corn yield over time with OLS Linear and DML. This figure shows the weighted estimates. We use our estimation procedure on 2-period samples from 1980 through 2019. Line shows central estimate, and grey band shows 95% confidence interval. The trend and standard errors are computed using weighted least squares, weighted by the inverse of the sum of the variance of the elasticity estimate and the residual from the regression of elasticity on the year.

Figure 8: Same figure as above, without weights.