Linear Regression in a Nonlinear World

Nadav Kunievsky[*]

Department of Economics
University of Chicago

April 11, 2024

Abstract

The interpretation of coefficients from multivariate linear regression relies on the assumption that the conditional expectation function (CEF) is linear in the variables. However, in many cases the underlying data generating process is nonlinear. This paper examines how to interpret regression coefficients under nonlinearity. We show that if the relationships between the variable of interest and other covariates are linear, then the coefficient on the variable of interest represents a weighted average of the derivatives of the outcome CEF with respect to the variable of interest. Interestingly, if these relationships are nonlinear, the regression coefficient becomes biased relative to this weighted average. We show that this bias is interpretable, analogous to the biases from measurement error and omitted variable bias under the standard linear model.

Keywords: Linear Regression, Multivariate Regression,Conditional Expectation Function

[*] This manuscript has benefited from multiple conversations with Natalie Goldshtein

1 Introduction

Multivariate linear regression is a fundamental tool across most scientific disciplines. Its main usage is to explore the relationship between different variables, assessing the average change in an outcome variable in response to an increase in the variable of interest (e.g Wooldridge [2015], Weisberg [2005], Greene [2003], Cunningham [2021], Montgomery et al. [2012]). For example, in environmental science, researchers may want to know what an increase in air quality implies for public health outcomes. To do this, researchers usually regress various outcomes on air quality measures, controlling for other variables like socioeconomic status and urbanization. They then interpret the regression coefficient on air quality as the change in health outcomes in response to a unit increase in air quality, holding the control variables fixed. Multivariate linear regression is not only used to describe correlation, but also extremely popular in causal analysis in observational studies. When researchers want to estimate a causal effect, they often operationalize the conditional independence assumptions (Pearl [2009], Cunningham [2021], Angrist and Pischke [2009]) necessary for identifying causal relationships by regressing an outcome variable on the variable of interest and a set of control variables. The regression coefficient is then interpreted as the causal effect of a change in the variable of interest on the outcome.

Whether multivariate linear regression is used to learn about causal effects or the correlation between variables, the clarity of interpretation largely hinges on the linearity of the conditional expectation function. If this function is indeed linear, the regression coefficient represents the constant marginal effects of the variable of interest, while controlling for other variables. However, if the function is non-linear, the regression coefficients might not reflect the marginal effect of the variable of interest accurately. This paper extends the findings of Yitzhaki [1996] to show how the coefficient of the variable of interest in a multivariate regression relates to the derivative of the conditional expectation function with respect to that variable. We demonstrate that when the relationship between the variable of interest and other covariates is linear, the regression coefficient for the variable of interest represents a weighted average of the derivatives of the conditional expectations of the outcome variable, with respect to the variable of interest. These weights resemble those found in Yitzhaki [1996], yet they are conditioned on the covariates and averaged across them. We also demonstrate that if the relationship is non-linear, the coefficient becomes biased relative to this weighted average. This bias is interpretable, and akin to the omitted variable bias and classical measurement error that arise when the data generating process is linear.

There are other interpretations of the coefficient of linear regression. Yitzhaki [1996] demonstrates that in a multivariate regression, each coefficient can be interpreted as a weighted sum of coefficients of all variables from a simple regression of the outcome on a univariate control. This interpretation is hard to form intuition for, and does not align with the standard intuition of linear regression in which the coefficient measures the change in outcomes, ”holding other variables fixed.” Most similar to our results, Angrist and Krueger [1999] shows that, in the case of the coefficient on a discrete treatment variable in a multivariate regression with fully saturated control variables, the regression coefficient on the variable of interest can be thought of as a weighted average of the treatment effect on different groups of “compliers” to the treatment, which is similar to what we demonstrate here for the continuous case and correctly specified model. Additionally, recent papers in econometrics have focused on the interpretation of the regression coefficient as a weighted average of treatment effects. Goldsmith-Pinkham et al. [2022] shows that, in the case where we regress an outcome variable on multiple treatment indicator variables, the coefficient on these variables is generally contaminated by the effects of other treatment variables. Similarly, recent literature has discussed the interpretation of the regression coefficient in difference-in-differences and event study outcome models (Callaway et al. [2021], Roth et al. [2022], Sun and Abraham [2021], de Chaisemartin and D'Haultfoeuille [2022]), and shows how these coefficients can be thought of as different weighted sums of the underlying heterogeneous treatment effects. In our paper, we focus on the case of a continuous variable of interest with a set of control variables and how the imposition of a linearity structure in the estimation stage summarizes the underlying conditional expectation function, and how imposing linearity may generate interpretable biases that should be taken into account when discussing the interpretation of regression coefficients. We also focus on the derivative of the conditional expectation with respect to the variable of interest and not the causal effect, although, as discussed below, the results here can easily inform a discussion on the coefficient of interest when the estimated quantity is the causal effect.

2 The Univariate Case

We begin by considering the univariate case. Suppose that the underlying data generating process (DGP) is represented by the function:

$$Y = g(T, \epsilon),$$

where Y is the outcome of interest, T is a continuous variable of interest, ϵ represents unobserved variables that influence the outcome and may be correlated with T, and g is the function that delineates the DGP. A researcher interested in the effect of T might estimate its coefficient in the linear model:

$$Y = \alpha + \beta T + u. \tag{1}$$

In the univariate case, Yitzhaki's theorem (Yitzhaki [1996]) provides a method to associate the population regression coefficient with the underlying DGP. Specifically, under certain regularity conditions, Yitzhaki demonstrates that

$$\beta = \int_{-\infty}^{\infty} \frac{\partial E[Y|t]}{\partial t} w(t) dt, \tag{2}$$

where $w(t) = \frac{E[T - E[T]|T>t]\mathrm{P}(T>t)}{\mathrm{Var}(T)}$ and $\int_{-\infty}^{\infty} w(t) = 1$. These weights are maximized at $E[T]$ and are increasing on the left of the maximum value and decreasing on the right, assigning zero weight to the values at the boundaries of the support.[1] Moreover, if ϵ is independent of T, then the regression coefficient provides us with a positively weighted average of the marginal causal effects of T on Y

$$\frac{\partial E[Y|T=t]}{\partial t} = E\left[\frac{\partial g(t,\epsilon)}{\partial t}\right].$$

3 The Naive Regression weighted Effect and Multivariate Regression

In this section, we extend Yitzhaki's result to the more generalized case, where we allow for additional control variables. Assume the DGP is now represented by the following two equations:

$$Y = g(T, \boldsymbol{X}, \upsilon), \tag{3}$$

$$T = h(\boldsymbol{X}, \varepsilon), \tag{4}$$

where υ and ε are unobserved variables that influence the outcome and the variable of interest's value, respectively, and h and g are the underlying causal functions that govern the DGP.

When researchers want to learn on how a change in variable of interest, T, affects the expected outcome variable Y, they often resort to using linear regression. Specifically, they

[1] The weights can also be thought of a density function, but notice that the density is different than the density of T

may estimate the following linear model:[2]

$$Y = T\beta + \boldsymbol{X}\gamma + \epsilon. \tag{5}$$

If researchers do not want to assume that the DGP is linear, they often interpret the coefficient on the variable of interest, β, as an average of the marginal effect of T, holding $\boldsymbol{X}$ fixed. Naturally, one might wonder how Yitzhaki's theorem applies to this multivariate context and how we should interpret β in this scenario.

To answer this question, we first define the "Naive Regression-Weighted Effect" (NRWE) as:

$$NRWE = \mathrm{E}\bigg[\int_{-\infty}^{\infty} \frac{\partial E[Y|t,\boldsymbol{X}]}{\partial t} w(t,\boldsymbol{X})dt\bigg], \tag{6}$$

where $w(T,\boldsymbol{X}) = \frac{E[T-E[T]|T>t,\boldsymbol{X}]\mathrm{P}(T>t|\boldsymbol{X})}{E[\mathrm{Var}(T|\boldsymbol{X})]}$ and the expectation is taken over $\boldsymbol{X}$. This parameter intuitively extends Yitzhaki's interpretation of β from the univariate case to the multivariate case. To see that, first notice that for each $\boldsymbol{X}$-cell, the numerator of the weights assigns the same weight that Yitzhaki's weights would assign in a regression of the outcome variable on T at the particular value of $\boldsymbol{X}$. The denominator of the weights is simply the average over the conditional variance of T, which ensures that the weights sum to 1, in a manner similar to Yitzhaki's original weights. Secondly, if we further assume that v independent of T given $\boldsymbol{X}$, $T \perp\!\!\!\perp v|\boldsymbol{X}$, then the $NRWC$ provides us with a weighted average of causal effects.

$$\mathrm{E}\bigg[\int_{-\infty}^{\infty} \frac{\partial \mathrm{E}[Y|t,\boldsymbol{X}]}{\partial t} w(t,\boldsymbol{X})dt\bigg] = \mathrm{E}\bigg[\mathrm{E}\bigg[\int_{-\infty}^{\infty} \frac{\partial g(t,\boldsymbol{X},v)}{\partial t} w(t,\boldsymbol{X})dt\bigg|\boldsymbol{X}\bigg]\bigg].$$

The Multivariate version of Yitzhaki's theorem, detailed below, reveals that the population regression coefficient, β, is equivalent to the $NRWE$ only when the relationship between the controls and the variable of interest is linear. In other cases, it often yields a biased estimate in relation to the Naive Regression-Weighted Effect.

Proposition 1 (Multivariate Yitzhaki's Theorem)**.** Denote by π the coefficients of $\boldsymbol{X}$ in the population regression of T on $\boldsymbol{X}$. Denote by $\mu(\boldsymbol{X}) = E[T|\boldsymbol{X}]$ and denote the misspecification error by $\Delta(\boldsymbol{X}) = \mu(\boldsymbol{X}) - \pi\boldsymbol{X}$. Assume the first and second moments and conditional moments exist and that the conditional expectations $E[Y|T,\boldsymbol{X}]$ is differentiable with respect to T, then the regression coefficient on the variable of interest, β, in the

[2] Throughout the analysis, we assume that $\boldsymbol{X}$ contains a constant.

population regression, $Y = T\beta + \boldsymbol{X}\gamma + \epsilon$, is given by:

$$\beta = \underbrace{\frac{\mathrm{Cov}(Y,(T-\mu(\boldsymbol{X}))}{\mathrm{Var}(T-\mu(\boldsymbol{X}))+\mathrm{Var}(\Delta(\boldsymbol{X}))}}_{\text{Weighted Effect of } T} + \underbrace{\frac{\mathrm{Cov}(Y,\Delta(\boldsymbol{X}))}{\mathrm{Var}(T-\mu(\boldsymbol{X}))+\mathrm{Var}(\Delta(\boldsymbol{X}))}}_{\text{Misspecification Bias}}$$

$$= \underbrace{E_{\boldsymbol{X}}\left[\int_{-\infty}^{\infty}\frac{\partial E[Y|T=t,\boldsymbol{X}]}{\partial t}w(t,\boldsymbol{X})dt\right]}_{\text{Weighted Effect of } T} + \underbrace{\frac{\mathrm{Cov}(Y,\Delta(\boldsymbol{X}))}{\mathrm{E}_{\boldsymbol{X}}[\mathrm{Var}(T|\boldsymbol{X})]+\mathrm{Var}(\Delta(\boldsymbol{X}))}}_{\text{Misspecification Bias}},$$

where

$$w(t,x) = \frac{\mathrm{E}[T-\mu(\boldsymbol{X})|T>t,\boldsymbol{X}]\mathrm{P}(T>t|\boldsymbol{X})}{\mathrm{E}_{\boldsymbol{X}}[\mathrm{Var}(T|\boldsymbol{X})]+\mathrm{Var}(\Delta(\boldsymbol{X}))} \geq 0.$$

The proof, detailed in the appendix, applies the Frisch-Waugh-Lovell Theorem and integration by parts.[3] The immediate observation from Proposition 1 is that, generally, β does not equal the $NRWE$. This difference is driven by two factors. First, the weights $w(t,x)$, while resembling those of the $NRWE$, do not integrate to 1. These weights induce a bias akin to the classical measurement error attenuation bias (e.g., Wooldridge [2015]), which attenuates the effect of T compared to the $NRWE$. If the variance of the misspecification error is non-zero, $\mathrm{Var}(\Delta(X)) > 0$, then the effect of T weighted in β would be smaller than in the $NRWE$. Note also that the attenuation is generally larger if the set of covariates can explain more of the variable of interest. Specifically, as $\boldsymbol{X}$ better explains T, then $\mathrm{Var}(T|X)$ decreases, leading to a greater attenuation of the weighted effect of T compared to $NRWE$.

The second source of bias in β compared to the NRWE is the misspecification bias, driven by the covariance between the misspecification error and the outcome variable. To better understand this bias, we consider different DGPs. First, let us assume $g(T,\boldsymbol{X},\upsilon) = \beta T + \gamma\boldsymbol{X} + \upsilon$, and allow h to be unrestricted. In this case, the outcome equation is correctly specified. By using the standard argument from the consistency of the OLS, we understand that the population regression coefficient equals the structural β, and, is therefore, trivially equal to the $NRWE$ parameter.[4] Next, let us consider the case in which h, the function governing the intensity of the variable of interest, is linear in $\boldsymbol{X}$. In this case, $\Delta(\boldsymbol{X}) = 0$ for all $\boldsymbol{X}$, and Proposition 1 shows that both the bias term and $\mathrm{Var}(\Delta(\boldsymbol{X}))$ equal zero, which implies that β equals the Naive Regression Weighted Effect.

Finally, let us consider the case where both $g(T,\boldsymbol{X},\upsilon)$ and $h(\boldsymbol{X},\epsilon)$ are non-linear in

[3] Angrist and Krueger [1999] demonstrated for the discrete case, with a fully saturated regression, a similar equivalence between a discrete equivalent of $NRWE$ and the regression coefficient

[4] To see this through the lens of Proposition 1, notice that:

$$\mathrm{Cov}(Y,T-\mu(\boldsymbol{X})) + \mathrm{Cov}(Y,\mu(X)-\pi\boldsymbol{X}) = \mathrm{Cov}(Y,T-\pi\boldsymbol{X}) = \beta\mathrm{Cov}(T,T-\pi X).$$

Divide by the denominator to get β.

$\boldsymbol{X}$. In this scenario, the population regression coefficient doesn't yield a weighted average of treatment effects. Instead, it gives a weighted average of the marginal effect of T and an additional bias term, which is introduced by the correlation between the outcome variable Y and $\Delta(X)$. These $\Delta(X)$ terms represent deviations of the conditional expectations from their best linear approximation.[5] If these deviations are systematically correlated with the outcome variable, our estimate will be biased. If, however, we find that $\mathrm{Cov}(Y, \Delta(\boldsymbol{X})) = 0$ but $\mathrm{Var}(\Delta(\boldsymbol{X})) \neq 0$, then the bias term in Proposition 1 becomes zero, but the measured effect is attenuated due to the variance of the mismeasurment error in the denominator. Even when control variables enter the outcome equation linearly, there can still be a bias if the treatment variable, T, has a nonlinear effect on the outcome. Specifically, consider $g(T, \boldsymbol{X}, v) = f(T) + \boldsymbol{X}\gamma + v$. In this scenario, β would not, in general, equal $NRWE$. To illustrate this, let us again assume that $T \perp\!\!\!\perp v|\boldsymbol{X}$ and $E[v|X] = 0$. Then, we have:

$$\begin{aligned}\mathrm{Cov}(Y, \Delta(\boldsymbol{X})) &= \mathrm{Cov}(f(T) + \boldsymbol{X}\gamma + v, T - \pi X - (T - \mu(\boldsymbol{X})) \\ &= \mathrm{Cov}(f(T), \Delta(\boldsymbol{X}))) \neq 0,\end{aligned}$$

where the second equality stems from the mean zero assumption, and the fact that the residuals are not correlated with $\boldsymbol{X}$. This example demonstrate that even when the conditional expectation function is linear in the control variables, misspecification bias can arise generally, if $f(T)$ and $\boldsymbol{X}$ are correlated. Hence, in the process of selecting control variables for regression analysis, it is imperative for researchers to prioritize examining how the variable of interest interacts with the control variables. This approach is more important than assessing the impact of control variables on the outcome variable.

A different perspective on the resulting bias can be achieved by expressing $\Delta(\boldsymbol{X})$ as the difference between residuals:

$$\Delta(X) = \mu(X) - \pi X = \underbrace{T - \pi X}_{\substack{\text{Unexplained Due to} \\ \text{Linearity Restrictions}}} - \underbrace{T - \mu(X)}_{\substack{\text{Fundamentally} \\ \text{unexplained}}} .$$

This equation illustrates that the bias arises from the variation in T that could potentially be explained using the control variables, but remains unexplained due to the linear constraints of the model.[6] Consequently, if these unexplained components are correlated with the outcome variable, additional bias is introduced because the model does not account for these components. Conversely, in the absence of such correlation, attenuation bias emerges due to the misestimation of the control variables' effect on the variable of interest,

[5] Recall that the coefficients provide the best linear approximation to the conditional expectations. See, for example,(Angrist and Pischke [2009])

[6] Notice that $(T - \mu(X)) - (T - \pi X)$ is the residual from a linear projection of $T - \mu(X)$ on $T - \pi X$.

T, resulting from the linearity restriction.

Under what conditions can we expect that the coefficient on the variable of interest captures the NRWE? First, if our regression is fully saturated and all control variables are discrete, then the linear approximation of the conditional expectation is exact, and the misspecification error is zero, $\mu(x) = 0$ for all x, eliminating both the bias term and the attenuation effect. Another example where the bias is zero occurs when the joint distribution of the explanatory variable (and not necessarily the joint distribution of the outcome variables and explanatory variables) belongs to the Elliptically Contoured distributions[7]; here, the conditional expectation of the variable of interest is linear in the other variables, avoiding the misspecification bias and attenuation bias. Additionally, in some instances, including sufficient interaction terms between variables can approximate the underlying data-generating function, thereby reducing biases (e.g., Hastie et al. [2009]).

In general, the linearity assumption is unlikely to hold, and both misspecification bias and attenuation bias may arise, necessitating a thorough evaluation of the relationships between variables. For instance, if higher values of the control variables X tend to increase both the variable of interest and the outcome variable in a convex manner, then the linear projection will likely underestimate T at high values of X. This situation implies that $\Delta(X)$ is likely to be positive for higher values of X, and consequently, it may be positively correlated with the outcome Y, suggesting $\text{Cov}\,(Y, \Delta(X)) \geq 0$. Such a scenario would induce an upward bias in β compared to the weighted effect component. For example, consider researchers exploring the effect of parent income on a child's years of schooling while controlling for the parent level of education. Previous studies have shown that average income grows exponentially with years of schooling[8] (e.g., Mincer [1974], Heckman et al. [2003]). Hence, the relationship between parent schooling and income is likely to be increasing, and linear projection would underestimate parent income at high values. Since the parent years of schooling are likely to be positively correlated with a child's years of schooling, our estimate for the effect of parent income is likely to be biased. Conversely, if researchers aim to examine the influence of parent educational level on a child's educational attainment while controlling for income, the role of the control variable and the variable of interest is reversed. In this case, the average parent years of schooling is a concave function (log) of parent income. Then, $\Delta(X)$ will likely be negative for higher values of X (linear projection overestimates T at high values X). In this case, if parent income increases a child's years of schooling in a convex manner, then the misspecification bias is likely to be negative.

[7] The Elliptically Contoured distributions famously include the multivariate Gaussian distribution.

[8] Usually the relation is described as log-linear.

3.1 Numerical Illustration

In this section, we demonstrate, using a numerical example, that under different data generating process, the size of two biases, the attenuation and misspecification bias, can be substantial and lead to incorrect conclusions. We assume the following DGP:

$$\begin{aligned} T &= h(X) + \nu \\ Y &= g(T, X) + \epsilon, \end{aligned}$$

where $h(\cdot)$ and $g(\cdot)$ will be defined later and

$$X \sim U(0,5), \quad \nu \sim N(0,1), \quad \epsilon \sim N(0,1).$$

In this DGP, where the conditional distribution of $T|X$ is Normal, we can derive a closed form expression for the weights derived in Proposition 1. Specifically, we have the following:

$$\begin{aligned} &E[T - E[T|X] \mid T > t, X] \cdot P(X > t \mid X) = \\ &\left[h(X) + \sigma \frac{\phi\left(\frac{t-h(X)}{\sigma}\right)}{1 - \Phi\left(\frac{t-h(X)}{\sigma}\right)} - h(X) \right] \cdot \left(1 - \Phi\left(\frac{t - h(X)}{\sigma}\right)\right) = \\ &\sigma\phi\left(\frac{t-h(X)}{\sigma}\right). \end{aligned}$$

Similarly, as the conditional variance of T, is fixed for any x value, we have $E[\mathrm{Var}(T|X)] = \sigma^2$. Therefore, the weights are:

$$w(t, X) = \frac{\sigma\phi\left(\frac{t-\log(X)}{\sigma}\right)}{\sigma^2} = \frac{\phi\left(\frac{t-\log(X)}{\sigma}\right)}{\sigma}$$

which is simply the Normal distribution density function, and we can approximate $NRWE$ numerically using generating samples and estimating the mean derivative in the population

$$\mathrm{E}\left[\frac{\partial h(x)}{\partial x}\right] \approx \frac{1}{n} \sum_{i=1}^{n} \frac{\partial h(x_i)}{\partial x_i}.$$

where n is the number of simulated samples. Similarly, we can approximate the bias term by calculating the sample covariance, $\hat{\mathrm{Cov}}(Y, \hat{h}(X) - \hat{\pi}_T T)$, and the sample variance of the residualized T, where $\hat{\pi}_T T$ is the regression coefficient on T in a linear regression of Y on T.

For our simulation, we generate $1,000,000$ draws and perform a Monte Carlo simulation exercise of 300 iterations. Table 1 below shows the results of these simulations for different data generating processes. The first row demonstrates the argument we've made in section 3. The relationship between the control variable, X, and the variable of interest, T, is convex, and higher values of the control variables are associated with higher outcome values. As the linear model underestimates the true value for high values of X, we get that β is much larger than the NRWE, driven by the misspecification bias. Notice that also the attenuation bias, computed as the difference between the weighted effect of T and $NRWE$, is substantial, where the effect weight is driven to almost zero. The second row demonstrates that even if the data generating process is linear in the variable of interest, the regression coefficient can still be biased. In our example, the NRWE is 1, but the regression coefficient is twice that size, 1.997. In this example, all the bias stems from the misspecification, where, again, the role of the weighted effect has been attenuated to almost zero. The third row demonstrates how the regression coefficient can lead us to wrong coclusions on the link between Y and T. The third row demonstrates that even if T does not affect directly the outcome variable, we may get a significant coefficient in our regression analysis. In our data generating process we get that the coefficient size is driven solely by misspecification bias.

	NRWE	β	**Misspecification Bias**	**Weighted Effect of T**	**Attenuation Bias**
$\mathrm{E}[T\|X] = exp(X)$ $E[Y\|T,X] = sin(T) + X$	-0.0414	0.0049	0.0051	-0.0001	-0.0413
	(0.0007)	(0.0001)	(0.0001)	(0.0000)	(0.002)
$\mathrm{E}[T\|X] = exp(X)$ $E[Y\|T,X] = T + exp(X)$	1.000	1.997	1.994	0.003	0.995
	NA	(0.0001)	(0.0002)	(0.0002)	(0.0733)
$\mathrm{E}[T\|X] = sin(X)$ $E[Y\|T,X] = sin(X) + X^2$	0.000	-0.405	-0.405	0.000	0.000
	NA	(0.0017)	(0.0058)	(0.0061)	(0.001)

Table 1: Simulation Results

Notes: This table presents the results from a Monte Carlo exercise that calculates the decomposition of the regression coefficient β, according to Proposition 1, from the regression model $Y = \beta T + \alpha X + u$, where the data generating process is specified in the first column. The coefficient β is decomposed into the misspecification bias and the weighted effect of T. The last column shows the attenuation bias, calculated as the difference between the Naïve Regression Weighted Effect and the weighted effect of T. Standard deviations of the estimated parameters are in parentheses.

4 Conclusion

Proposition 1 emphasizes the difficulties in interpreting regression coefficients when the underlying data-generating process is not linear. However, it also provides guidance on how researchers can address these biases when interested in the $NRWE$ parameter. The simplest approach to obtain an unbiased estimate of $NRWE$ is to include an estimate of $E[T|\boldsymbol{X}]$ as a control variable in the regression[9]. To compute $E[T|\boldsymbol{X}]$, one can either estimate it nonparametrically (e.g., Ullah and Pagan [1999]), or use prior knowledge of it (Borusyak and Hull [2021]). However, this raises a question: if one can estimate $E[T|\boldsymbol{X}]$ nonparametrically, why choose to estimate the causal effect using regression instead of estimating the entire model nonparametrically? In many cases, researchers opt for linear regression due to its efficiency and stability, two properties that do not always characterize nonparametric estimators, especially when the dimensions of $\boldsymbol{X}$ are large. Therefore, if researchers wish to use regression, it would be insightful to include in their analysis a discussion on the relationship between the control variables and the variable of interest. This can be done, for example, by plotting $E[T|x_j]$ for different components of $\boldsymbol{X}$, or provide theoretical justification for the use of linear controls.

Researchers should also bear in mind that the relationship between the variables of interest and control variables are not generally invariant to monotonic changes. For instance, researchers should be cautious when estimating a linear model where T enters the regression linearly, and a similar model where they use $\log(T)$ instead. Without altering the control variable as well, as Proposition 1 shows, both models are unlikely to obtain a weighted average of changes in the conditional expectation, and at least one of them is likely to suffer from a misspecification bias. Hence, researchers should be more conscious of how they model their control variables.

A Appendix

A.1 Proof of Proposition 1

Denote by π the coefficients from linear projection of T on $\boldsymbol{X}$. Denote by $f(T)$ and $f(T|\boldsymbol{X})$ the density and conditional density of T. Using Frisch-Waugh-Lovell theorem (Frisch and Waugh [1933]), we have that

$$\beta = \frac{\text{Cov}(Y(T - \boldsymbol{X}\pi)}{\text{Var}(T - \pi\boldsymbol{X})}.$$

We start by focusing the numerator. Denote the conditional expectation of T, condition on $\boldsymbol{X}$ by $\mu(\boldsymbol{X})$. Then we can express the numerator as

$$\begin{aligned}\text{Cov}(Y, T - \boldsymbol{X}\pi) &= \text{Cov}(Y, (T - \mu(\boldsymbol{X}) + \mu(\boldsymbol{X}) - \boldsymbol{X}\pi) \\ &= \text{Cov}(Y, (T - \mu(\boldsymbol{X})) + \text{Cov}(Y, \mu(\boldsymbol{X}) - \boldsymbol{X}\pi).\end{aligned}$$

[9] This is because the conditional expectation, given the conditional expectation, is a trivial linear function of the conditional expectation: $E[T|E[T|X]] = E[T|X]$

Using the law of iterated expectations and integration by parts we can re-express the first term as

$$
\begin{aligned}
\mathrm{Cov}(Y,(T-\mu(\boldsymbol{X})) &= E[Y(T-\mu(\boldsymbol{X})] \\
&= E_{\boldsymbol{X}}[E[Y(T-\boldsymbol{X})|\boldsymbol{X}]] \\
&= E_{\boldsymbol{X}}[E[E[Y|T,\boldsymbol{X}](T-\mu(\boldsymbol{X})|\boldsymbol{X}]] \\
&= E_{\boldsymbol{X}}\left[\int_{-\infty}^{\infty} E[Y|T,\boldsymbol{X}](u-\mu(\boldsymbol{X})f(u|\boldsymbol{X})du\right] \\
&= E_{\boldsymbol{X}}\Bigg[\left[E[Y|u,\boldsymbol{X}]\int_{-\infty}^{t}(u-\mu(\boldsymbol{X})f(u|\boldsymbol{X})dt\right]_{t=-\infty}^{\infty} \\
&- \int_{-\infty}^{\infty}\frac{\partial E[Y|t,\boldsymbol{X}]}{\partial t}\int_{-\infty}^{t}(u-\mu(\boldsymbol{X})f(u|\boldsymbol{X})dudt\Bigg] \\
&= E_{\boldsymbol{X}}\left[\int_{-\infty}^{\infty}\frac{\partial E[Y|t,\boldsymbol{X}]}{\partial t} - \int_{-\infty}^{t}(u-\mu(\boldsymbol{X})f(u|\boldsymbol{X})dudt\right] \\
&= E_{\boldsymbol{X}}\left[\int_{-\infty}^{\infty}\frac{\partial E[Y|t,\boldsymbol{X}]}{\partial t}E[T-\mu(\boldsymbol{X})|T>t,\boldsymbol{X}]p(T>t|\boldsymbol{X})dt\right]
\end{aligned}
$$

Where the last equality follows from the fact that

$$
E[T-\mu(\boldsymbol{X})|T>t,\boldsymbol{X}]\mathrm{P}(T>t|\boldsymbol{X}) + E[T-\mu(\boldsymbol{X})|T\leq \boldsymbol{X}]\mathrm{P}(T\leq t|\boldsymbol{X}) = 0.
$$

Therefore the numerator is given by

$$
\mathrm{Cov}(Y,T-\boldsymbol{X}\pi) = E_{\boldsymbol{X}}\left[\int_{-\infty}^{\infty}\frac{\partial E[Y|t,\boldsymbol{X}]}{\partial t}E[T-\mu(\boldsymbol{X})|T>t,\boldsymbol{X}]p(T>t|\boldsymbol{X})dt\right] + \mathrm{Cov}(Y,\mu(\boldsymbol{X}-\pi\boldsymbol{X}).
$$

Next, we turn to the denominator. We can re-express it as

$$
\begin{aligned}
\mathrm{Var}(T-\pi\boldsymbol{X}) &= \mathrm{Var}(T-\mu(\boldsymbol{X})+\mu(\boldsymbol{X})-\pi\boldsymbol{X}) \\
&= \mathrm{Var}(T-\mu(\boldsymbol{X})) + \mathrm{Var}(\mu(\boldsymbol{X})-\pi\boldsymbol{X})) + 2\mathrm{Cov}(T-\mu(\boldsymbol{X}),\mathrm{Var}(\mu(\boldsymbol{X})-\pi\boldsymbol{X})) \\
&= \mathrm{Var}(E[T-\mu(\boldsymbol{X})|\boldsymbol{X}]) + E[\mathrm{Var}(T|\boldsymbol{X})] + \mathrm{Var}(\mu(\boldsymbol{X})-\pi\boldsymbol{X})) \\
&= \mathrm{E}[\mathrm{Var}(T|\boldsymbol{X})] + \mathrm{Var}(\mu(\boldsymbol{X})-\pi\boldsymbol{X})),
\end{aligned}
$$

where we used the law of total variance and the fact that $\mathrm{Cov}(T-\mu(\boldsymbol{X}),\mu(\boldsymbol{X})-\pi\boldsymbol{X}) = E[(T-\mu(\boldsymbol{X})(\mu(\boldsymbol{X})-\pi\boldsymbol{X}))] = 0$, due to the law of iterated expectations, which concludes the proof.

Bibliography

Joshua D. Angrist and Alan B. Krueger. Empirical strategies in labor economics. In *Handbook of labor economics*, volume 3, pages 1277–1366. Elsevier, 1999.

Joshua D. Angrist and Jörn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 2009.

Kirill Borusyak and Peter Hull. Non-random exposure to exogenous shocks. *Working Paper*, 2021.

Brantly Callaway, Andrew Goodman-Bacon, and Pedro H. C. Sant'Anna. Difference-in-differences with a continuous treatment, July 2021. URL https://arxiv.org/abs/2107.02637. First draft on arXiv: July 6, 2021. This draft: January 26, 2024.

Scott Cunningham. *Causal Inference: The Mixtape*. Yale University Press, New Haven, CT, 2021. ISBN 978-0-300-25356-0.

Clément de Chaisemartin and Xavier D'Haultfoeuille. Difference-in-differences estimators of intertemporal treatment effects. Working Paper 29873, National Bureau of Economic Research, March 2022. URL https://econpapers.repec.org/RePEc:nbr:nberwo:29873. Revision Date: July 2023.

Ragnar Frisch and Frederick V. Waugh. Partial time regressions as compared with individual trends. *Econometrica: Journal of the Econometric Society*, pages 387–401, 1933.

Paul Goldsmith-Pinkham, Peter Hull, and Michal Kolesár. Contamination bias in linear regressions. NBER Working Paper 30108, National Bureau of Economic Research, June 2022. URL https://www.nber.org/papers/w30108.

William H. Greene. *Econometric Analysis*. Prentice Hall, 5 edition, 2003.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer New York, 2nd edition, 2009. ISBN 978-0-387-84857-0.

James J Heckman, Lance J Lochner, and Petra E Todd. Fifty years of mincer earnings regressions. *National Bureau of Economic Research Working Paper Series*, (9732), 2003.

Jacob Mincer. *Schooling, Experience, and Earnings*. National Bureau of Economic Research, 1974.

Douglas C. Montgomery, Elizabeth A. Peck, and G. Geoffrey Vining. *Introduction to Linear Regression Analysis*. John Wiley Sons, 2012.

Judea Pearl. *Causality*. Cambridge University Press, 2009.

Jonathan Roth, Pedro H. C. Sant'Anna, Alyssa Bilinski, and John Poe. What's trending in difference-in-differences? a synthesis of the recent econometrics literature. *Journal of Econometrics*, 235:2218–2244, January 2022. doi: 10.1016/j.jeconom.2022.07.001. URL https://econpapers.repec.org/RePEc:arx:papers:2201.01194.

Liyang Sun and Sarah Abraham. Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, 225(2):175–199, April 2021. URL https://econpapers.repec.org/RePEc:arx:papers:1804.05785.

Aman Ullah and Adrian Pagan. *Nonparametric Econometrics*. Cambridge University Press, Cambridge, 1999.

Sanford Weisberg. *Applied Linear Regression*. John Wiley Sons, 3 edition, 2005.

Jeffrey M. Wooldridge. *Introductory Econometrics: A Modern Approach*. Cengage Learning, 6 edition, 2015.

Shlomo Yitzhaki. On using linear regressions in welfare economics. *Journal of Business Economic Statistics*, 14(4):478–486, 1996.