Linear Regression in a Nonlinear World: The Multivariate Case

Nadav Kunievsky

July 3, 2023

In economics and other social sciences, researchers frequently document the causal effect of a variable of interest on an another outcome variable. In non-experimental scenarios, this is usually achieved using a conditional independence assumption[1]. To operationalize this assumption when analyzing the data, researchers often employ multivariate linear regression. They regress an outcome variable on the variable of interest and a set of control variables, thereby ensuring the conditional independence needed to identify the causal effect. It is widely known that if the underlying Data Generating Process (DGP) is linear, the structural parameters that govern the outcome can be identified from the population regression coefficients. A natural question arises then how does the population regression coefficient on the variable of interest is related to the underlying marginal causal effect of the variable of interest, when the DGP is not linear. In this note, we demonstrate, using a generalization of Yitzhaki [1996], how the coefficient of the variable of interest is determined by the underlying DGP. Furthermore, we demonstrate how, when the DGP is non-linear, the imposition of linear structure on the model can be thought of as an omitted variable bias and/or a measurement error.

We begin by considering the univariate case. Suppose that the underlying Data Generating Process (DGP) is represented by

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

where Y is the outcome of interest, T is a continuous variable of interest, ϵ are unobserved variables that influence the outcome, and g is the function that delineates the DGP. A researcher who is interested in the effect of T might choose to estimate the coefficient of T in the following linear model

$$Y = \alpha + \beta_1 T + u \tag{1}$$

In the univariate scenario, 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

[1] It's worth noting that in many instances researchers assume conditional mean independence. However, for the sake of discussion, we focus on conditional independence, which forms the backbone of nonparametric identification arguments Pearl [2009]

that

$$\beta_1 = \int_{-\infty}^{\infty} \frac{\partial E[Y|t]}{\partial t} w(t) dt \tag{2}$$

where $w(t) = \frac{E[T - E[T]|T>t]\mathrm{P}(T>t)}{\mathrm{Var}(T)}$ and $\int_{-\infty}^{\infty} w(t) = 1$. Moreover, if ϵ is independent of T, then the regression coefficient provides us with a positively weighted average of the average marginal effects of the variable of interest

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

In this note, we inquire whether Yitzhaki's result extends to the more generalized case where researchers assume that T is independent of the unobserved variables, given a set of control variables, $\boldsymbol{X}$. Specifically, suppose the DGP is now represented by

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

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

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. We further suppose that $T \perp\!\!\!\perp \upsilon | \boldsymbol{X}$.[2]

When researchers aim to utilize the conditional independence assumption and estimate some functional of the causal effect distribution, they often resort to linear regression. Specifically, they estimate the following linear model[3]:

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

Often, researchers interpret the coefficient on the variable of interest, β, as an average (or some average) of the marginal causal effect of T. 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 Causal Effect" (NRwCE) as

$$NRwCE = E_X\left[\int_{-\infty}^{\infty} \frac{\partial E[Y|t, \boldsymbol{X}]}{\partial t} w(t, \boldsymbol{X}) dt\right] \tag{6}$$

[2]This can also be framed in potential outcomes notation as $T \perp\!\!\!\perp Y_t | X$, $Y = \int_t \mathbb{1}\{T = t\} Y_t dt$

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

where $w(T, \boldsymbol{X}) = \frac{E[T-E[T]|T>t,\boldsymbol{X}]\mathrm{P}(T>t|\boldsymbol{X})}{E[\mathrm{Var}(T|\boldsymbol{X})]}$.

This parameter intuitively extends Yitzhaki's interpretation of β from the univariate case to the multivariate case. Firstly, 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 this 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, just like in the univariate case, if ϵ is independent of T given $\boldsymbol{X}$, the $NRwCE$ provides us with a weighted average of causal effects.

$$\mathrm{E}_X\left[\int_{-\infty}^{\infty}\frac{\partial \mathrm{E}[Y|t,\boldsymbol{X}]}{\partial t}w(t,\boldsymbol{X})dt\right] = \mathrm{E}_X\left[\mathrm{E}\left[\int_{-\infty}^{\infty}\frac{\partial g(t,\boldsymbol{X},\upsilon)}{\partial t}w(t,\boldsymbol{X})dt\middle|\boldsymbol{X}\right]\right]$$

The generalized form of Yitzhaki's theorem, detailed below, reveals that the population regression coefficient, β, is equivalent to the naive regression weighted causal effects 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 Causal Effect.

Theorem 1 (Generalized 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 prediction error by $\Delta(\boldsymbol{X}) = \mu(\boldsymbol{X}) - \pi\boldsymbol{X}$. Assume the first and second moments and conditional moments exist, then the treatment coefficient, β, from the population regression, $Y = T\beta + \boldsymbol{X}\gamma + \epsilon$, is given by

$$\begin{aligned}\beta &= \underbrace{\frac{\mathrm{Cov}(Y,(T-\mu(\boldsymbol{X}))}{\mathrm{Var}(T-\mu(\boldsymbol{X}))+\mathrm{Var}(\Delta(\boldsymbol{X}))}}_{\text{Causal Component}} + \underbrace{\frac{\mathrm{Cov}(Y,\Delta(\boldsymbol{X}))}{\mathrm{Var}(T-\mu(\boldsymbol{X}))+\mathrm{Var}(\Delta(\boldsymbol{X}))}}_{Bias} \\ &= \underbrace{E_{\boldsymbol{X}}\left[\int_{-\infty}^{\infty}\frac{\partial E[Y|T=t,\boldsymbol{X}]}{\partial t}w(t,\boldsymbol{X})dt\right]}_{\text{Causal Component}} + \underbrace{\frac{\mathrm{Cov}(Y,\Delta(\boldsymbol{X}))}{\mathrm{Var}(T-\mu(\boldsymbol{X}))+\mathrm{Var}(\Delta(\boldsymbol{X}))}}_{Bias}\end{aligned}$$

where

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

The proof, detailed in the appendix, applies the Frisch-Waugh-Lovell Theorem and integration by parts[4]. The

[4] A similar theorem for the case where T is discrete and the researcher conducts a fully saturated regression can be found at Angrist and Krueger [1999]

initial observation from Theorem 1 is that, in general, β does not equal the $NRwCE$ parameter. The weights, while they resemble the weights of the $NRwCE$, don't sum to 1, and there's an additional bias term driven by changes in the controls and variation in T. To better understand this expression, we evaluate different DGPs.

First, let's 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, therefore, is trivially equal to the $NRwCE$ parameter[5]. Next, let's consider the scenario 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 Theorem 1 shows that both the bias term and $\text{Var}(\Delta(\boldsymbol{X}))$ equal zero, which implies that β equals the naive regression weighted causal effect.

Finally, let's consider the situation where both $g(T, \boldsymbol{X}, \upsilon)$ and $h(\boldsymbol{X}, \epsilon)$ are non-linear in $\boldsymbol{X}$. 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 causal effect of T and a 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[6]. If these deviations are systematically correlated with the outcome variable, our causal estimate will be biased. If, however, we find that $\text{Cov}(Y, \Delta(\boldsymbol{X})) = 0$ but $\text{Var}(\Delta(\boldsymbol{X})) \neq 0$, then the bias term in theorem 1 becomes zero, but we encounter a measurement error issue that attenuates β downwards. 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 $g(T, \boldsymbol{X}, \upsilon) = f(T) + \boldsymbol{X}\gamma + \upsilon$. In this scenario, β would not, in general, equal $NRwCE$. To illustrate this, let's again assume that $T \perp\!\!\!\perp \upsilon | \boldsymbol{X}$ and $E[\upsilon | X] = 0$. Then, we have:

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

where the second equality stems from the mean zero assumption, and the fact that the residuals are not correlated with $\boldsymbol{X}$. This demonstrates that when considering including control variables in the regression, we should concentrate

[5] To see this through the lens of theorem 1, notice that:

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

Divide by the denominator to get β.

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

on their interaction with the variable of interest, and not on their potential impact on the outcome variable. In the accompanying Jupyter Notebook, we demonstrate that the bias and attenuation effects can be significant, such that the regression coefficient may provide limited-to-no information on the effect of the variable of interest.

Therefore, we can see that imposing a linear structure on control variables, while the function $h(\boldsymbol{X}, \epsilon)$ that regulates the intensity of the variable of interest is non-linear, can lead to issues analogous to omitted variable bias and/or measurement error. 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 illustrates that bias is induced by the portion of the variation in T that could potentially be explained using the control variables, but is not explained because we restrict the model to be linear. If these remaining, unexplained components are correlated with the outcome variable, we receive additional bias because we are unable to control for these components. If they are not correlated, we get attenuation bias due to the fact that we mismeasure the control variable function.

To conclude, the Generalized Yitzhaki theorem emphasizes that merely assuming conditional independence and estimating the causal effect using regression is insufficient to ensure the attainment of some functional form of the treatment effect distribution. However, it also provides guidance on how researchers can address these biases when interested in the $NRwCE$ parameter. The simplest approach to obtain an unbiased estimate of $NRwCE$ is to include an estimate of $E[T|\boldsymbol{X}]$ as a control variable in the regression[7]. To compute $E[T|\boldsymbol{X}]$, one can either estimate it nonparametrically Ullah and Pagan [1999], or use prior knowledge on the assignment mechanism (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. Consequently, if researchers wish to use regression, it

[7] This is as the conditional expectation, given the conditional expectation, is a linear function of the conditional expectation $E[T|E[T|X]] = E[T|X]$

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 by plotting $E[T|x_j]$ for different components of X_j. Estimating the conditional expectations for one variable is relatively straightforward and could provide an indication that the linearity assumption is not unreasonable.

Researchers should also bear in mind that the relationship between the treatment and control variables is not invariant to monotonic changes of the treatment. For instance, researchers should be cautious when estimating a linear model where T enters the regression linearly, and a similar model where they use $log(T)$. Both models are unlikely to obtain a weighted average of treatment effects, and at least one of them is likely to suffer from misspecification bias. Hence, researchers should be more conscious of how they model their control variables.

Appendix

The following is a generalization of Yitzhaki [1996] , for the multivariate case. 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 the numerator as

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

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

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

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.

References

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.

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

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

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

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