Goel, Deepti

Working Paper

Estimator of What? A Note on Teaching Regressions in Introductory Econometrics

GLO Discussion Paper, No. 1646

Provided in Cooperation with:
Global Labor Organization (GLO)

Suggested Citation: Goel, Deepti (2025) : Estimator of What? A Note on Teaching Regressions in Introductory Econometrics, GLO Discussion Paper, No. 1646, Global Labor Organization (GLO), Essen

This Version is available at:
https://hdl.handle.net/10419/323522

Estimator of What? A Note on Teaching Regressions in Introductory Econometrics

Deepti Goel*

August 5, 2025

*Assistant Professor, Pitzer College, 1050 N. Mills Avenue, Claremont, CA 91711; Research Fellow, Global Labor Organization (GLO); email: *deepti_goel@pitzer.edu*

Abstract

The most widely used textbooks on Introductory Econometrics conflate three distinct population parameters: the population regression function (PRF), the conditional expectation function (CEF), and the causal effect. They also incorrectly suggest, and sometimes state, that the Conditional Mean Zero assumption implies causal interpretation of regression coefficients. I highlight these issues and show that by incorporating new notation these limitations can easily be overcome.

JEL codes: A22, C18
Keywords: Regressions, Least Squares, Conditional Mean Zero, Causal Inference

1 Introduction

This note is a result of teaching Introductory Econometrics for several years and finding myself intermittently frustrated when using standard textbooks on regressions. Thanks to a surprisingly less-cited paper by Bryant Chen and Judea Pearl (Chen and Pearl, 2013)[1] titled, Regression and Causation: A Critical Examination of Six Econometrics Textbooks, and the book, Mostly Harmless Econometrics, by Joshua D. Angrist and Jorn-Steffen Pischke (Angrist and Pischke, 2009), I am now able to articulate my discomfiture and offer a way forward in the hope of improving the teaching of regressions in undergraduate classrooms.

Chen and Pearl (2013) examine six widely used textbooks, namely, (Greene, 2012), (Hill et al., 2011), (Kennedy, 2008), (Ruud, 2000), (Stock and Watson, 2011), and (Wooldridge, 2009), and point to a shared shortcoming, namely, that of incomplete notation.[2] Additionally, in my view, there is an unreasonable expectation that readers can intuit from the context what is being left unsaid.[3] Here is what I mean. Consider the Sample (linear) Regression Function (SRF), that most textbooks begin with when teaching regressions:

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X_1 + ... + \hat{\beta}_k X_k + \hat{U} \; where \; \hat{U} \equiv (Y - \hat{\beta}_0 + \hat{\beta}_1 X_1 + ... + \hat{\beta}_k X_k)$$

Most textbooks state (and also prove) that the sample regression coefficient estimator, say $\hat{\beta}_1$, is unbiased. What they do not state explicitly is the population parameter that $\hat{\beta}_1$ is an unbiased estimator of. If there was only one contender for this parameter, then the omission would be innocuous. However, when there are three, it creates confusion. The three plausible population parameters that $\hat{\beta}_1$ could be an unbiased estimator of are:

- The derivative, with respect to X_1, of the Population (linear) Regression Function (PRF) of Y on $X_1, ..., X_k$.
- The derivative, with respect to X_1, of the Conditional Expectation Function (CEF) of Y given $X_1, ..., X_k$, namely $E(Y|X_1, ..., X_k)$.
- The causal effect of X_1 on Y.

I shall denote these three distinct population parameters by β_1^R, β_1^E, and β_1^C, respectively. This new notation allows us to clearly distinguish between the parameters, resulting in greater conceptual clarity.

The essential arguments in this note can be found scattered in different parts of the book, Mostly Harmless Econometrics, by Joshua D. Angrist and Jörn-Steffen Pischke (Angrist and Pischke, 2009). My contribution is to bring them together in one place in order to make a forceful case to change how regressions are introduced and discussed in the most widely used

[1] 64 citations as per Google Scholar as of July 8, 2025.

[2] Chen and Pearl (2013) assert that these textbooks lack notation to capture the causal effect of X on Y as defined by the 'do-operator' (see section 2.3), resulting in a conflation of causal effects with the moments of joint probability distributions, such as with $E(Y|X)$.

[3] Both Wooldridge 2025 and Stock and Watson 2019 still have these limitations. I have not checked the latest editions of the other textbooks.

textbooks on Introductory Econometrics.

In what follows, I define the three parameters in section 2; section 3 discusses how they may be linked with each other; section 4, states the population parameter(s) that the sample regression coefficient is an unbiased/consistent estimator of, paying attention to the assumptions needed to make these claims; section 5, provides a more complete interpretation of the 'Unbiasedness of OLS' as stated in Wooldridge 2025; and I conclude in section 6.

2 Defining the Population Parameters

Consider the population of random variables $\{X_1, X_2, ..., X_k, Y\}$. For clarity of thought, let $X_1, ..., X_k$ be functionally independent. I begin by defining the three population parameters, β_1^R, β_1^E, and β_1^C, in that order, to emphasize that they are conceptually distinct.

2.1 The Derivative of the PRF

Suppose we are interested in predicting Y given information on $X_1, ..., X_k$, and we are only interested in predictors that are 'linear in parameters,' i.e., of the form $b_0 + b_1X_1 + b_2X_2 + ... + b_kX_k$ where $b_0, b_1, b_2..., b_k$ are real numbers. We define the best predictor using the least squares criterion: the best predictor is that which minimizes the expected value of squared prediction errors, where a prediction error is defined as the difference between the actual value of Y and our predicted value. Our problem can be compactly written as:

$$\begin{aligned} &minimize\ wrt\ \{b_0, b_1, ..., b_k\}:\ E[(Y - (b_0 + b_1X_1 + ... + b_kX_k))^2] \\ &\equiv minimize\ wrt\ \{b_0, b_1, ..., b_k\}:\ \frac{\sum_{i=1}^{N}(Y_i - (b_0 + b_1X_{i1} + ... + b_kX_{ik}))^2}{N} \end{aligned}$$

where N is the population size. If real numbers, $\beta_0, \beta_1, ..., \beta_k$, solve the problem, then, $\beta_0 + \beta_1X_1 + \beta_2X_2 + ... + \beta_kX_k$ is defined as the Population (linear) Regression Function (PRF), and $\{\beta_0, \beta_1, ..., \beta_k\}$ are the PRF coefficients. In the (k+1) dimensional population scatter of $\{X_1, ..., X_k, Y\}$, $\hat{Y} = \beta_0 + \beta_1X_1 + \beta_2X_2 + ... + \beta_kX_k$ is the unique, nonrandom, *linear surface of best fit* for Y in terms of $X_1, X_2, ..., X_k$.

β_1^R is defined as follows:

$$\begin{aligned} \beta_1^R &\equiv \frac{\partial PRF}{\partial X_1} \\ &= \frac{\partial(\beta_0 + \beta_1X_1 + \beta_2X_2 + ... + \beta_kX_k)}{\partial X_1} \\ &= \beta_1\ (\text{given } X_1, ..., X_k \text{ are functionally independent}) \end{aligned}$$

Defining β_1^R entailed imposing a functional form requirement, namely, $\{\beta_0, \beta_1, ..., \beta_k\}$ enter linearly in the predictor function. In this sense, β_1^R is parametric in nature.

2.2 The Derivative of the CEF

In everyday life we frequently, and instinctively, try to guess the outcome of an uncertain phenomenon given information on other related phenomena. For example, a student when deciding the optimal allocation of their time might think about their expected test score were they to devote different amounts of time towards studying. Similarly, a firm trying to maximize sales, might think about expected sales given different money amounts spent on advertising. In statistics, the name for these heuristic guesses is the conditional expectation.

The formal definition of the conditional expectation of Y given information on $X_1, ..., X_k$ is as follows:

- If Y is discrete,

$$E(Y|X_1, ..., X_k) = \sum y * f_{Y|X_1,...,X_k}(y)$$

- If Y is continuous,

$$E(Y|X_1, ..., X_k) = \int y * f_{Y|X_1,...,X_k}(y)dy$$

where $f_{Y|X_1,...,X_k}$ is the conditional probability mass/density function of Y given $X_1, ..., X_k$. When viewed as a mapping from the space of $X_1, ..., X_k$ to the space of Y, the conditional expectation is referred to as the Conditional Expectation Function (CEF). As it turns out, the CEF is the solution to an important prediction exercise; perhaps this why we think about it so much when making decisions. Next, I describe the problem that the CEF solves which should motivate why it is the best predictor of Y given $X_1, ..., X_k$.

Suppose, as in section 2.1, we are again interested in predicting Y using $X_1, ..., X_k$, but unlike section 2.1 where we had constrained our predictor to be *linear* in parameters, we now allow it take any form. We continue to define the best predictor using the least squares criterion. This problem can be compactly written as:

$$\begin{aligned} &minimize\ wrt\ m(.):\ E(Y - m(X_1, ..., X_k))^2 \\ &\equiv minimize\ wrt\ m(.):\ \frac{\sum_{i=1}^{N}(Y_i - m(X_{i1}, ..., X_{ik}))^2}{N} \end{aligned}$$

where $m(X_1, ..., X_k)$ is *any* function mapping $X_1, ..., X_k$ to Y. It can be shown that the CEF, namely $E(Y|X_1, ..., X_k)$, is the solution of this unconstrained minimization problem. If $m^*(X_1, ..., X_k)$ is used to denote the optimal $m(.)$, then, $m^*(X_1, ..., X_k) = E(Y|X_1, ..., X_k)$ In the same (k+1) dimensional population scatter of $\{X_1, ..., X_k, Y\}$, referred to in section 2.1, $\hat{Y} = E(Y|X_1, ..., X_k)$ is the unique, nonrandom, *surface of best fit* for Y in terms of $X_1, X_2, ..., X_k$.[4]

[4]Given that the PRF and CEF are solutions to the constrained and unconstrained versions of the same problem, respectively, the CEF will always be at least as good as the PRF at predicting Y, and at times it will be strictly better.

If $E(Y|X_1, ..., X_k)$ is differentiable,[5] then, β_1^E is defined as follows:

$$\beta_1^E \equiv \frac{\partial E(Y|X_1, ..., X_k)}{\partial X_1}$$

Defining β_1^E did not involve making any functional form assumption about $E(Y|X_1, ..., X_k)$. In this sense, β_1^E is non-parametric in nature.

2.3 The Causal Effect

For an individual i, the causal effect of a variable, say X_1, on another variable, say Y, is defined as the change in Y_i when only X_{1i} changes, and everything else that affects Y_i remains as is. In this *ceteris paribus* scenario, the different values of Y_i corresponding to each value that X_{1i} can possibly take, are called *potential outcomes*. Most modern Econometrics textbooks define a *causal effect* using this notion of potential outcomes (Rubin, 2005).

For individual i, the causal effect of changing X_{1i}, from say value a to value b, on outcome Y_i, is defined as the corresponding change in their *potential outcomes*.[6] Let $Y_i^{X_{1i}=a}$ and $Y_i^{X_{1i}=b}$ be the potential outcomes at values a and b, respectively. The causal effect of change in X_1 when it changes from a to b is defined as:

$$\beta_1^{C_i} \equiv \frac{Y_i^{X_{1i}=b} - Y_i^{X_{1i}=a}}{a - b}$$

The division by $a - b$ expresses the effect in terms of a unit change in X_1. The causal effect in the population is the average of the individual causal effects.

$$\beta_1^C \equiv E(\beta_1^{C_i})$$

Like β_1^E, β_1^C is also non-parametric in nature as it does not involve making any functional form assumption on how the potential outcomes change as X_1 changes.

Pearl (1995) provides an alternative, but conceptually equivalent, definition of a causal effect in terms of the 'do-operator.' He defines the causal effect of X_1 on Y as the change in $E[Y|do(x_1)]$ as X_1 changes, where $E[Y|do(x_1)]$ is the average value of Y in the population in a scenario where the researcher intervenes to assign the values of X_1 as one would in a randomized controlled experiment.

[5] If $E(Y|X_1, ..., X_k)$ is not differentiable, then, for any two real values a and b that X_1 can take,

$$\beta_1^E \equiv \frac{E(Y|X_1 = a, X_2..., X_k) - E(Y|X_1 = b, X_2..., X_k)}{a - b}$$

where the value of β_1^E may vary depending on the values of a and b.

[6] For any individual, at a point in time, only one of their potential outcomes, namely the one corresponding to the particular value that X_{1i} takes at that time, is *observed*; all other potential outcomes (those associated with other values that X_{1i} could have taken but did not take) remain unobserved. These unobserved outcomes are called called *counterfactuals*.

If $E[Y|do(x_1)]$ is differentiable, then, β_1^C is defined as follows:[7]

$$\beta_1^C \equiv \frac{\partial E[Y|do(x_1)]}{\partial X_1}$$

3 Links between Population Parameters

It should now be clear that the three parameters embody conceptually different features of the population. Next, I discuss under what circumstances would any two of them be the same. For this section, I am assuming that $E(Y|X_1, ..., X_k)$ and $E[Y|do(x_1)]$ are differentiable with respect to X_1.

3.1 Derivative of PRF vis-à-vis Derivative of CEF

If the optimal function, $m^*(.)$, of section 2.2 happens to be linear in parameters, i.e., $m^*(.)$ takes the form $b_0 + b_1X_1 + ... + b_kX_k$, then the CEF and the PRF are one and the same, and $\beta_1^R = \beta_1^E = \beta_1$.

Consider the PRF:

$$\begin{aligned} \hat{Y} &= \beta_0 + \beta_1X_1 + \beta_2X_2 + ... + \beta_kX_k \\ \Leftrightarrow Y &= \beta_0 + \beta_1X_1 + \beta_2X_2 + ... + \beta_kX_k + U \\ where\ U &\equiv Y - (\beta_0 + \beta_1X_1 + \beta_2X_2 + ... + \beta_kX_k) \end{aligned}$$

U is called the error term, it gets its identity from the PRF.

An *assumption* that receives a lot of attention in textbooks is the 'Conditional Mean Zero (CMZ)' assumption which states that $E(U|X_1, ..., X_k) = 0$. Making the CMZ assumption is *equivalent* to assuming that the CEF is linear in parameters. To see this, note that for the PRF stated above,

$$\begin{aligned} &\text{If } E(U|X_1, ..., X_k) = 0 \\ \implies\ &E(Y|X_1, ..., X_k) = \beta_0 + \beta_1X_1 + \beta_2X_2 + ... + \beta_kX_k \end{aligned}$$

On the other hand,

$$\begin{aligned} &\text{If } E(Y|X_1, ..., X_k) = \beta_0 + \beta_1X_1 + \beta_2X_2 + ... + \beta_kX_k \\ \implies\ &E(U|X_1, ..., X_k) = 0 \end{aligned}$$

[7] If $E[Y|do(x_1)]$ is not differentiable, then, for any two real values a and b that X_1 can take,

$$\beta_1^C = \frac{E[Y|do(X_1 = a)] - E[Y|do(X_1 = b)]}{a - b}$$

where the value of β_1^C may vary depending on the values of a and b.

Thus, making the CMZ assumption is equivalent to **assuming** that $\beta_j^R = \beta_j^E \; \forall j \in 1, 2, ..., k$.

It is important to note that in a *fully saturated* PRF,[8] the CMZ assumption is true by definition, i.e. it is no longer an assumption but follows from the definition of a fully saturated model. This implies that in a fully saturated PRF, $\beta_j^R \equiv \beta_j^E$. In section 3.4, I give an example of a fully saturated PRF.

3.2 Derivative of CEF vis-à-vis the Causal Effect

Whenever the derivative of the CEF, i.e. $\frac{\partial E(Y|X_1,...,X_k)}{\partial X_1}$, and the derivative of the 'do-operator' function, i.e. $\frac{\partial E[Y|do(x_1)]}{\partial X_1}$ are identical, $\beta_1^E = \beta_1^C$. Ex-ante, there is no reason to believe that $E(Y|X_1, ..., X_k)$ and $E[Y|do(x_1)]$ are the same as they are conceptually different things: $E(Y|X_1, ..., X_k)$ is the first moment of the observed joint distribution of $Y|X_1, ..., X_k$, while $E[Y|do(x_1)]$ is average Y coming from an experiment where the researcher assigns the X_1 values at random in the population of interest. In section 3.4, I present a plausible example of two populations, one where real world conditions are such that $\beta_1^E = \beta_1^C$, and another where because of a different reality, $\beta_1^E \neq \beta_1^C$.

3.3 Derivative of PRF vis-à-vis the Causal Effect

Whenever β_1^R is picking up the ceteris paribus effect of X_1 on Y, $\beta_1^R = \beta_1^C$. For this to happen, the PRF must embody and implement a credible research design (Angrist and Pischke, 2010). The most popular such designs include randomized controlled trials, selection on observables (or matching), difference in differences, instrument variables, and regression discontinuity designs. Crucially, the credibility of the design is rooted in causal thinking aimed at creating the relevant counterfactual.

For example, consider the PRF, say $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + U$ that relies on the 'selection on observables' design to get at the causal effect of X_1 on Y. In that case, whether $\beta_1^R = \beta_1^C$, rests on the 'Conditional Independence Assumption (CIA)' being true. The CIA states that X_1 is as if randomly assigned within each sub-population of the partition defined by the other (besides X_1) explanatory variables, $X_2, ..., X_k$. Whenever, the causal effect is constant across individuals, the CIA holds, and the potential outcome function is linear in $X_2, ..., X_k$, $\beta_1^R = \beta_1^C$. See Angrist and Pischke 2017 for details.

Importantly, The CMZ assumption does not result in any kind of equivalency between β_j^R and β_j^C, or between β_j^E and β_j^C. Angrist and Pischke 2017, Chen and Pearl 2013, and Crudu et al. 2022, all make this point. It is surprising, and unfortunate, that popular textbooks

[8] A fully saturated PRF is one where the explanatory variables are all discrete, and, there are as many regression coefficients as the number of elements in the population partition defined by the explanatory variables. For example, consider the PRF, $\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 * X_2)$ where, X_1 and X_2 are binary. The partition defined by these variables is: $\{(1,1), (1,0), (0,1), (0,0)\}$. Since the PRF has four regression coefficients, namely, $\{\beta_0, \beta_1, \beta_2, \beta_3\}$, this is a fully saturated PRF.

continue to suggest/state that $E(U|X_1, ..., X_k) = 0 \implies \beta_j^R = \beta_j^C$.[9] See section 3.4 for an example where $E(U|X_1, ..., X_k) = 0$, $\beta_j^R = \beta_j^E$, but $\beta_j^R \neq \beta_j^C$. Crudu et al. (2022) also present data generating processes where the CMZ assumption is satisfied but the PRF identifies a pseudo-parameter that does not have a causal interpretation.

3.4 Unifying Example

I have chosen this rather simplistic example as it allows for clarity of thought. Consider the following PRF:

$$\widehat{BirthWt} = \beta_0 + \beta_1 Male$$

Or equivalently,

$$BirthWt = \beta_0 + \beta_1 Male + U \; where \; U \equiv BirthWt - (\beta_0 + \beta_1 Male)$$

$BirthWt$ denotes the birth weight of a new born. $Male$ is an indicator of whether the newborn is male or not: $Male$ takes the value 1 if new born i is male, and 0, otherwise. In this example, by definition,

$$\beta_1 \equiv \beta_1^R$$

Moreover, this is a fully saturated model,[10] and so trivially, the CMZ assumption holds, i.e. $E(U|Male) = 0$. Therefore, in this example

$$\beta_1 \equiv \beta_1^R = \beta_1^E$$

In other words, the regression coefficient on $Male$ also gives the difference between the average birth weights of male and non-male newborns in the population.

Finally, is $\beta_1 = \beta_1^C$ in this example? The answer depends on whether or not one believes that the CIA holds. Because in this example there are no other explnatory variables, the CIA translates to, is being $Male$ randomly distributed in the population of newborns. This in turn depends on our understanding of the real world. In places like India where there is male preference, and female maltreatment and female foeticide are prevalent, being male is manipulable, and is therefore *not as if randomly assigned* in the population. Thus, in the

[9] Both Stock and Watson (2019) and Wooldridge (2025) make this claim. Here is a quote from Chapter 4 of Stock and Watson 2019:"The first least squares assumption for causal inference is that the error term in the linear regression model has a conditional mean of 0 given the regressor X. This assumption holds if X is randomly assigned in an experiment or is as-if randomly assigned in observational data. Under this assumption, the OLS estimator is an unbiased estimator of the causal effect β_1." And, here is a quote from Chapter 2 of Wooldridge 2025: "Section 2-5 will show that we are only able to get reliable estimators of β_0 and β_1 from a random sample of data when we make an assumption restricting how the unobservable u is related to the explanatory variable x. Without such a restriction, we will not be able to estimate the ceteris paribus effect, β_1." In this quote, the assumption restricting how u is related to x is the CMZ Assumption.

[10] There is one binary explanatory variable. The partition it defines is: $\{(1), (0)\}$. Since the PRF has two regression coefficients, $\{\beta_0, \beta_1\}$, it thus meets the definition of a fully saturated model.

Indian population, $\beta_1 \equiv \beta_1^R = \beta_1^E \neq \beta_1^C$.[11] However, in places such as (pre-immigration) Sweden, being male is truly the luck of the draw (as if randomly assigned), and one can claim that CIA holds. In such places, $\beta_1 \equiv \beta_1^R = \beta_1^E = \beta_1^C$.

4 Sample Regression Coefficient as an Estimator

I assume we have access to a random sample of $\{X_1, X_2, ..., X_k, Y\}$ from the population, and there is no perfect multicollinearity among $\{1, X_1, X_2, ..., X_k\}$, both in the population and in the sample. The Sample Regression Function (SRF), is obtained by mimicking in the sample the least squares exercise that is used to obtain the PRF. Suppose the SRF is given by:

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X_1 + ... + \hat{\beta}_k X_k + \hat{U}$$
$$where\ \hat{U} \equiv Y - (\hat{\beta}_0 + \hat{\beta}_1 X_1 + ... + \hat{\beta}_k X_k)$$

I explicitly state the conditions under which the SRF coefficient, $\hat{\beta}_1$, is a valid estimator of each population parameter discussed above.

4.1 $\hat{\beta}_1$ as estimator of β_1^R

An SRF *always* gives a consistent estimator of the corresponding PRF without any extra assumptions (other than access to a random sample and no perfect multi-collinearity). This is because the SRF is the Method of Moments (MOM) estimator of the PRF. Thus, $\hat{\beta}_j$ is a consistent estimator of β_j^R for all $j \in \{0, 1, 2, ..., k\}$. However, $\hat{\beta}_j$ is not always an unbiased estimator of β_j^R. For this we need the CMZ assumption. In other words, $\hat{\beta}_j$ is an unbiased estimator of β_j^R if $E(U|X_1, X_2, ..., X_k) = 0$.

The typical context in which we are interested in knowing the PRF is that of *predicting* Y. No elaborate thinking is needed to justify the inclusion (or exclusion) of particular Xs, and operationally, the final choice of Xs gets determined by 'out of sample prediction fit' of various model specifications. While in the exposition above, we focused on one element of the PRF, namely, β_1^R, typically in prediction applications one is not focused on a specific element, and interest lies in the overall fit of the model.

4.2 $\hat{\beta}_1$ as estimator of β_1^E

Recall that $\beta_1^E \equiv \frac{\partial E(Y|X_1,...,X_k)}{\partial X_1}$. If $\frac{\partial E(Y|X_1,...,X_k)}{\partial X_1}$ is not a constant (and there is no compelling reason why it should be), then $\hat{\beta}_1$ is neither an unbiased, nor a consistent, estimator of β_1^E. If however, the CMZ assumption holds, i.e. $E(Y|X_1, ..., X_k) = 0$, then $\beta_1^E = \beta_1^R \equiv \beta_1$; and $\hat{\beta}_1$ is both an unbiased and a consistent estimator of β_1^E.

As pointed by Angrist and Pischke (2009), even when the CEF is non-linear in parameters,

[11] Perhaps, in India, $\beta_1^R > \beta_1^C$ as the coefficient is not only picking up the causal effect of being male, but also the effects of better nutrition and prenatal care for mothers pregnant with a male child.

the PRF is the best *linear* approximation of the CEF. In other words, if $\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$ is the PRF, then $\{\beta_0, \beta_1, ..., \beta_k\}$ is also the solution to the following problem (Angrist and Pischke, 2009):

$$minimize\ wrt\ \{b_0, b_1, ..., b_k\}:\ E[(E(Y|X_1, ..., X_k) - (b_0 + b_1 X_1 + ... + b_k X_k))^2]$$

Whenever researchers are interested in predicting Y using variables $X_1, ..., X_k$, they will get the most accurate predictions (in terms of minimizing least squared errors), using the CEF. Given that the PRF gives the best linear approximation of the CEF, the PRF must be valued for prediction even when CEF is non-linear in parameters.

4.3 $\hat{\beta}_1$ as estimator of β_1^C

Whether $\hat{\beta}_1$ is a credible estimator of β_1^C, depends on the credibility of the underlying research design that the PRF implements. In a selection on observables research design, credibility fundamentally depends on whether the CIA holds or not.

β_1^C is the causal effect of a specific variable, X_1, on outcome Y. The context in which we are interested in knowing a causal effect is very different from that of wanting to *predict* the outcome, Y. When we are interested in the causal effect of X_1 on Y, we are only interested in whether the PRF coefficient on X_1 gives us the causal effect of interest and we do not care about how the PRF does in terms of predicting Y. A lot of thought must go into what Xs to include and exclude from the model as the credibility of the research design may be compromised when 'bad controls' are included as regressors (Angrist and Pischke, 2009).

5 Sample Regression Coefficient is an Unbiased Estimator of What?

In this section, I closely follow the presentation in section 3.3, titled, 'The Expected Value of the OLS Estimators' in Chapter 3 of Wooldridge, 2025. I restate, verbatim, Theorem 3.1 titled, 'Unbiasedness of OLS' along with the assumptions involved:

"Under assumptions MLR.1 through MLR.4,

$$E(\hat{\beta}_j) = \beta_j,\ j = 0, 1, ..., k$$

for any values of the population parameter β_j. In other words, the OLS estimators are unbiased estimators of the population parameters."

Assumptions MLR.1 through MLR.4 are stated below (also stated verbatim):

- Assumption MLR.1 Linear in Parameters: The model in the population can be written as $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u$ where where $\beta_0, \beta_1, ..., \beta_k$ are the unknown parameters (constants) of interest and u is an unobserved random error or disturbance term.

- Assumption MLR.2 Random Sampling: We have a random sample of n observations, $\{(x_{i1}, x_{i2}, ..., x_{ik}, y_i) : i = 1, 2, ..., n\}$, following the population model in Assumption MLR.1.
- Assumption MLR.3 No Perfect Collinearity: In the sample (and therefore in the population), none of the independent variables is constant, and there are no exact linear relationships among the independent variables.
- Assumption MLR.4 Zero Conditional Mean: The error u has an expected value of zero given any values of the independent variables. In other words, $E(u|x_1, x_2, ..., x_k) = 0$.

This is how I interpret the Theorem. Given, MLR.4, $\beta_j = \beta_j^R = \beta_j^E,\ j = 0, 1, ..., k$. There is nothing in the theorem that should lead us to conclude that $\beta_j = \beta_j^C$. For that we would need to think about the validity of the CIA for each x_j.

6 Concluding Remarks

Angrist and Pischke (2017) note the recent shift in the empirical econometrics literature towards using a regression to estimate a causal parameter of interest, denoted as β_1^C in this note. At the same time, a regression is also a workhorse tool for predicting variables, with cutting-edge applications in supervised machine learning. In applications aimed at prediction, interest lies in estimating the surface of best fit, characterized by the coefficient set $\{\beta_j^R\}$ and $\{\beta_j^E\}$ in this note. Since economists are interested in both causal and predictive problems, Introductory Econometrics textbooks must cover all three population parameters, but at the same time they need to be explicit about which parameter is being referenced in a given context. To do so, it is most important to introduce new notation to distinguish between the three parameters, as I have done in this note. I hope the authors of textbooks will revisit their exposition on regressions and incorporate the notation presented here. The edits necessary are relatively minor when compared to the benefits in terms of making Econometrics more accessible to a wider audience with varied research goals.

References

Angrist, J. D. and J. Pischke (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, USA: Princeton University Press.

Angrist, J. D. and J. Pischke (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives 24*(2), 3–30.

Angrist, J. D. and J. Pischke (2017). Undergraduate econometrics instruction: Through our classes, darkly. *Journal of Economic Perspectives 31*(2), 125–144.

Chen, B. and J. Pearl (2013). Regression and Causation: A Critical Examination of Six Econometrics Textbooks. *Real-World Economics Review 65*, 2–20.

Crudu, F., G. Mellace, and J. Smits (2022). What does ols identify under the zero conditional mean assumption? *Department of Economics University of Siena* (872).

Greene, W. H. (2012). *Econometric Analysis* (7 th ed.). New Jersey, USA: Pearson Education.

Hill, C. R., W. E. Griffiths, and G. C. Lim (2011). *Principles of Econometrics* (4 th ed.). New York, USA: John Wiley & Sons Inc.

Kennedy, P. (2008). *A Guide to Econometrics* (6 th ed.). Cambridge, USA: MIT Press.

Pearl, J. (1995). Causal Diagrams for Empirical Research. *Biometrika 82*(4), 669–710.

Rubin, D. B. (2005). Causal inference using potential outcomes. *Journal of the American Statistical Association 100*(469), 322–331.

Ruud, P. A. (2000). *An Introduction to Classical Econometric Theory*. Oxford, UK: Oxford University Press.

Stock, J. H. and M. W. Watson (2011). *Introduction to Econometrics* (3 rd ed.). New York, USA: Addison-Wesley.

Stock, J. H. and M. W. Watson (2019). *Introduction to Econometrics* (4 th ed.). New York, USA: Pearson.

Wooldridge, J. M. (2009). *Introductory Econometrics: A Modern Approach* (4 th ed.). South-Western Cengage Learning.

Wooldridge, J. M. (2025). *Introductory Econometrics: A Modern Approach* (8 th ed.). Mason, OH, USA: Cengage.