UNIVERSITÀ
DI SIENA
1240

QUADERNI DEL DIPARTIMENTO
DI ECONOMIA POLITICA E STATISTICA

Federico Crudu
Giovanni Mellace
Joeri Smits

What does OLS identify under the zero conditional mean assumption?

n. 872 – Febbraio 2022

What does OLS identify under the zero conditional mean assumption?

Federico Crudu[1], Giovanni Mellace[2], and Joeri Smits[3]

[1]University of Siena and CRENoS
[2]University of Southern Denmark
[3]Harvard University

February 22, 2022

Abstract

Many econometrics textbooks imply that under mean independence of the regressors and the error term, the OLS estimand has a causal interpretation. We provide counterexamples of data-generating processes (DGPs) where the standard assumption of zero conditional mean error is satisfied, but where OLS identifies a pseudo-parameter that does not have a causal interpretation. No such counterexamples can be constructed when the assumption needed is stated in the potential outcome framework, highlighting the fact that causal inference requires *causal*, and not just *stochastic*, assumptions.

Keywords: OLS, zero conditional mean error, causal inference.

JEL Classification Numbers: C10, C18, C21, C31.

We thank Christian Dahl, Markus Frölich, Arthur Lewbel, Ye Lu, Pedro Sant'Anna and Rami Tabri for useful conversations.

1 Introduction

In their coverage of ordinary least squares (OLS) regression, many econometrics textbooks invoke the assumption that the regression error is mean independent of the covariates, and has mean zero. It is often claimed or implied that under this assumption (plus additional ones such as linearity and no perfect collinearity), the OLS estimand has a causal interpretation. In this paper, we provide counterexamples to those claims. In the data generating processes (DGPs) of our examples, the zero conditional mean error assumption is satisfied, yet the OLS estimand does not have any causal interpretation: it identifies a pseudo-parameter. We then show that a causal assumption–for instance, one stated in the potential outcome framework–does not allow for such DGPs, permitting identification of a suitably defined causal effect.

A review of econometrics textbooks reveals several examples of causal claims about the OLS estimand under zero conditional mean error. For instance, the graduate textbook by Hayashi (2000) claims that the regression coefficients represent the marginal effects of the regressors. Cameron and Trivedi (2005) and Wooldridge (2010) imply that under zero conditional mean error, the OLS estimand has a structural interpretation. And the introductory textbooks by Wooldridge (2019) and Stock and Watson (2019) both claim that under the zero conditional mean error assumption, the OLS estimand can be interpreted as a *ceteris paribus* effect.

An empirical example in Wooldridge (2019) highlights the problem with these claims. In the example, yield is the regressand and fertilizer the regressor; “if fertilizer amounts are chosen independently of other features of the plots, then [zero conditional mean error] will hold: the average land quality will not depend on the amount of fertilizer. However, if more fertilizer is put on the higher-quality plots of land, then the expected value of u [the error] changes with the level of fertilizer, and [zero conditional mean error] fails.” The first sentence in this quote is true, but the second not necessarily, as we will show. This stems from the fact that a lack of causal relations between variables implies their statistical independence, but the converse is not true (Chalak and White, 2012). This is why causal assumptions are required for a causal interpretation.

What our counterexamples have in common is that the DGPs feature cancellations of parameters that generate stochastic (conditional) independence relations between variables despite the presence of a causal relation between them. The existence of such ‘perverse’ DGPs that generate joint distributions featuring cancellations of parameters has long been acknowledged in other causal inference frameworks. In the Pearl Causal Model (PCM), such joint distributions are termed 'unfaithful' to the DGP, and they are ruled out by the Causal Faithfulness assumption (e.g., Spirtes et al. (2000); Pearl (2009)). In the settable systems extension of the PCM of White and Chalak (2009) and Chalak and White (2012), these instances are referred to as P-stochastic isolation. Such DGPs are special: the set of unfaithful distributions has Lebesgue measure zero. In this paper, we show that these DGPs are ruled out when the assumption required for identification is stated in the potential outcome framework. Indeed, if the assumption needed is stated in the causal language of the potential outcome framework, then OLS *does* identify a causal parameter. This highlights the fact that causal inference requires *causal*, and not just *stochastic*, assumptions.

2 Identification failure under the zero conditional mean assumption

For the simple regression model

$$Y_i = \gamma + \lambda D_i + \varepsilon_i,$$

it is often claimed or implied–as we reviewed in Section 1–that if $E(\varepsilon_i|D_i) = E(\varepsilon_i) = 0$, then the OLS estimand of λ has a causal interpretation (e.g., a marginal effect or a ceteris paribus effect).

We will now present a DGP where $E(\varepsilon_i|D_i) = E(\varepsilon_i) = 0$, but where the OLS estimand of λ identifies a pseudo-parameter that does not have a causal interpretation.

We will also show that if we impose an assumption on the potential outcomes instead, then the OLS estimand *does* identify a meaningful parameter, and it is not possible to construct such a counterexample as long as the potential outcomes are well defined.

In particular, we assume that the data are generated as follows

$$\begin{aligned}
U_i &\sim \mathcal{N}(0,1), \\
D_i &= I(U_i > 0), \\
Y_i &= \kappa D_i + \alpha U_i + \nu_i, \\
\kappa &= -2\alpha\frac{\phi(0)}{\Phi(0)}, \\
E(\nu_i|D_i) &= E(\nu_i) = 0.
\end{aligned}$$

Given this DGP, the regression error term ε_i can be written as

$$\varepsilon_i \equiv Y_i - \gamma - \lambda D_i = \kappa D_i + \alpha U_i + \nu_i - \gamma - \lambda D_i.$$

Moreover, since $E(U_i|D_i = 1) = E(U_i|U_i > 0) = \frac{\phi(0)}{\Phi(0)}$ and $E(U_i|D_i = 0) = E(U_i|U_i \leq 0) = -\frac{\phi(0)}{\Phi(0)}$, their difference is

$$E(U_i|D_i = 1) - E(U_i|D_i = 0) = 2\frac{\phi(0)}{\Phi(0)}.$$

Putting it all together, we have that

$$\begin{aligned}
E(\varepsilon_i|D_i = 1) - E(\varepsilon_i|D_i = 0) &= \gamma - \gamma + \alpha[E(U_i|D_i = 1) - E(U_i|D_i = 0)] \\
&+ E(\nu_i|D_i = 1) - E(\nu_i|D_i = 0) + \kappa - \lambda \\
&= 2\alpha\frac{\phi(0)}{\Phi(0)} + \kappa - \lambda \\
&= 2\alpha\frac{\phi(0)}{\Phi(0)} - 2\alpha\frac{\phi(0)}{\Phi(0)} - \lambda \\
&= -\lambda.
\end{aligned}$$

This shows that our standard textbook assumption $E(\varepsilon_i|D_i) = E(\varepsilon_i) = 0$ is satisfied if and only if $\lambda = 0$ even though the effect of D_i on Y_i is not in general equal to zero. The reason cancellation occurs in this example is that the selection bias, caused by omitting U_i from the model, is exactly equal in magnitude and of the opposite sign as the causal effect of D_i on Y_i. This makes the statistical error ε_i and D_i mean independent, but λ does not have a causal interpretation.

The OLS estimand *can*, however, be given a causal interpretation, if causal assumptions are imposed. Let Y_i^d, $d \in \{0, 1\}$ be the potential outcome unit i would get if we were able to set D_i to d, keeping everything else fixed. Hence, the treatment effect κ_i for unit i equals $Y_i^1 - Y_i^0$. Under the usual Stable Unit Treatment Value Assumption (SUTVA), we can observe only one of the two potential outcomes for each unit according to the observational rule:

$$Y_i = Y_i^1 D_i + Y_i^0 (1 - D_i).$$

From the observational rule we can write

$$\begin{aligned} Y_i &= Y_i^1 D_i + Y_i^0 (1 - D_i) \\ &= Y_i^1 D_i + Y_i^0 - Y_i^0 D_i \\ &= (Y_i^1 - Y_i^0) D_i + Y_i^0 \\ &= \kappa_i D_i + E(Y_i^0) + \eta_i. \end{aligned}$$

The last equation comes from the fact that one can always write the random variable Y_i^0 as $E(Y_i^0) + \eta_i$ with $E(\eta_i) = 0$. Without loss of generality and to simplify exposition we will normalize $E(Y_i^0) = 0$ and $\kappa_i = \kappa \; \forall \; i$ such that

$$Y_i = \kappa D_i + \eta_i.$$

Notice that in the DGP of our counterexample $E(Y_i^0 | D = 1) = \alpha \frac{\phi(0)}{\Phi(0)} \neq E(Y^0 | D = 0) = -\alpha \frac{\phi(0)}{\Phi(0)} \neq 0$. Thus, the weakest possible assumption on the potential outcomes to interpret λ as a causal effect, i.e., $E(Y^0 | D) = E(Y^0)$, is violated. This is because $E(Y^0 | D) = E(Y^0)$ directly assumes zero selection bias and thus no cancellation can occur. This implies that when using a regression model to estimate a causal effect, relying merely on a statistical assumption (zero conditional mean error) does not suffice for identification. To identify a causal effect, one needs to impose causal assumptions.

In Appendix A.1 we show a similar example of a DGP when adding a set of control variables X_i that is not correlated with D_i. We also present an example where D_i depends on X_i. However, in this case, the treatment effect of D_i on Y_i has to be heterogeneous and depend on X_i. The reason is that cancellation occurs when the selection bias and the causal effect are identical in magnitude and of opposite sign. Since the selection bias depends on X_i when D_i is affected by X_i, so must the effect of D_i on Y_i to have cancellation. Finally, in a supplementary online appendix we provide some numerical experiments that confirm our theoretical derivations.

References

A. C. Cameron and P. K. Trivedi. *Microeconometrics: methods and applications*. Cambridge University Press, 2005.

K. Chalak and H. White. Causality, conditional independence, and graphical separation in settable systems. *Neural Computation*, 24(7):1611–1668, 2012.

F. Hayashi. *Econometrics*. Princeton University Press, 2000.

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

P. Spirtes, C. Glymour, and R. Scheines. *Causation, prediction, and search*. The MIT Press, 2000.

J. H. Stock and M. W. Watson. *Introduction to econometrics*. Pearson Education, 2019.

H. White and K. Chalak. Settable Systems: An Extension of Pearl's Causal Model with Optimization, Equilibrium, and Learning. *Journal of Machine Learning Research*, 10(8):1759–1799, 2009.

J. M. Wooldridge. *Econometric analysis of cross section and panel data*. MIT press, 2010.

J. M. Wooldridge. *Introductory econometrics: A modern approach*. Cengage learning, 2019.

Appendix

A Adding control variables

A.1 Control variables that only affect the outcome

Herein, we generalize the above example to include a set of control variables X_i that only affects the outcome Y_i and not our main regressor D_i. We modify our DGP as follows

$$\begin{array}{rcl} \begin{pmatrix} U_i \\ X_i \end{pmatrix} & \sim & \mathcal{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right], \\ D_i & = & I(U_i > 0), \\ \kappa & = & -2\alpha\dfrac{\phi(0)}{\Phi(0)}, \\ Y_i & = & \kappa D_i + \alpha U_i + \delta X_i + \nu_i, \\ E(\nu_i|D_i, X_i) & = & E(\nu_i) = 0. \end{array}$$

Given this DGP, the regression error term ε_i can be written as

$$\varepsilon_i \equiv Y_i - \gamma - \lambda D_i - \beta X_i = \kappa D_i + \alpha U_i + \delta X_i + \nu_i - \gamma - \lambda D_i - \beta X_i.$$

Moreover, since $E(U_i|D_i = 1, X_i) = E(U_i|U_i > 0) = \frac{\phi(0)}{\Phi(0)}$ and $E(U_i|D_i = 0, X_i) = E(U_i|U_i \leq 0) = -\frac{\phi(0)}{\Phi(0)}$, their difference is

$$E(U_i|D = 1, X_i) - E(U_i|D = 0, X_i) = 2\frac{\phi(0)}{\Phi(0)}.$$

Putting it all together, we have that

$$\begin{array}{rcl} E(\varepsilon|D = 1, X_i) - E(\varepsilon|D = 0, X_i) & = & \gamma - \gamma + \alpha[E(U_i|D = 1, X_i) - E(U_i|D = 0, X_i)] \\ & + & E(\nu_i|D = 1, X_i) - E(\nu_i|D = 0, X_i) + \kappa - \lambda \\ & - & \beta X_i + \beta X_i - \delta X_i + \delta X_i \\ & = & 2\alpha\dfrac{\phi(0)}{\Phi(0)} + \kappa - \lambda \\ & = & 2\alpha\dfrac{\phi(0)}{\Phi(0)} - 2\alpha\dfrac{\phi(0)}{\Phi(0)} - \lambda \\ & = & -\lambda. \end{array}$$

This shows that our standard textbook assumption $E(\varepsilon|D, X) = E(\varepsilon)$ is satisfied if and only if $\lambda = 0$, but in this DGP the causal effect of D on Y is $\kappa = -2\alpha\frac{\phi(0)}{\Phi(0)} \neq 0 \; \forall \; \alpha \neq 0$.

A.2 Control variables that affect both the main regressor and the outcome

We now consider a set of control variables X_i that affects both Y_i and D_i. We modify our DGP as follows

$$\begin{aligned} \begin{pmatrix} U_i \\ X_i \end{pmatrix} &\sim \mathcal{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right], \\ D_i &= I(U_i + \pi X_i > 0), \\ \kappa_i &= -2\alpha \frac{\phi(\pi X_i)}{\Phi(\pi X_i)}, \\ Y_i &= \kappa_i D_i + \alpha U_i + \delta X_i + \nu_i, \\ E(\nu|D) &= E(\nu) = 0. \end{aligned}$$

Given this DGP, the regression error term ε_i can be written as

$$\varepsilon_i \equiv Y_i - \gamma - \lambda D_i - \beta X_i = \kappa_i D_i + \alpha U_i + \delta X_i + \nu_i - \gamma - \lambda D_i - \beta X_i.$$

Moreover, since $E(U_i|D_i = 1, X_i) = E(U_i|U_i > \pi X_i, X_i) = \frac{\phi(\pi X_i)}{\Phi(\pi X_i)}$ and $E(U_i|D_i = 0, X_i) = E(U_i|U_i > \pi X_i, X_i) = -\frac{\phi(\pi X_i)}{\Phi(\pi X_i)}$, their difference is

$$E(U_i|D = 1, X_i) - E(U_i|D = 0, X_i) = 2\frac{\phi(\pi X_i)}{\Phi(\pi X_i)}.$$

Putting it all together, gives us

$$\begin{aligned} E(\varepsilon|D = 1, X_i) - E(\varepsilon|D = 0, X_i) &= \gamma - \gamma + \alpha[E(U_i|D = 1, X_i) - E(U_i|D = 0, X_i)] \\ &+ E(\nu_i|D = 1, X_i) - E(\nu_i|D = 0, X_i) + E(\kappa_i|X_i) - \lambda \\ &- \beta X_i + \beta X_i - \delta X_i + \delta X_i, \\ &= 2\alpha\frac{\phi(\pi X_i)}{\Phi(\pi X_i)} + E(\kappa_i|X_i) - \lambda, \\ &= 2\alpha\frac{\phi(\pi X_i)}{\Phi(\pi X_i)} - 2\alpha\frac{\phi(\pi X_i)}{\Phi(\pi X_i)} - \lambda, \\ &= -\lambda. \end{aligned}$$

This shows that our standard textbook assumption $E(\varepsilon|D, X) = E(\varepsilon)$ is satisfied if and only if $\lambda = 0$, but in this DGP the individual treatment effect $\kappa_i = -2\alpha\frac{\phi(\pi X_i)}{\Phi(\pi X_i)} \neq 0 \;\forall\; \alpha \neq 0$.

What does OLS identify under the zero conditional mean assumption?

Some numerical experiments

Federico Crudu[1], Giovanni Mellace[2], and Joeri Smits[3]

[1]University of Siena and CRENoS
[2]University of Southern Denmark
[3]Harvard University

February 22, 2022

Abstract

In this document, we collect some numerical experiments to support the claims presented in the main text. We use directed acyclic graphs to clarify the role of the variables in the experiments.

Keywords: OLS, zero conditional mean error, causal inference.

JEL Classification Numbers: C10, C18, C21, C31.

1 Introduction

In this document we study the behavior of the OLS estimator under the data generating process (DGP) described in Section 2 of the main text. In addition, we investigate the case where control variables are included (Appendices A.1 and A.2 of the main text). For such a case we distinguish the scenario where the additional controls only affect the outcome and the scenario where they affect both the outcome and the treatment variable. To clarify the role played by the variables in the simulation experiments, we accompany the DGPs with their corresponding directed acyclic graph (DAG) representation.

2 A simple DGP

Let us consider the following model

$$Y_i = \kappa D_i + \alpha U_i + \nu_i \tag{1}$$

where $U_i \sim \mathcal{N}(0,1)$, $\nu_i \sim \mathcal{N}(0,1)$ and $D_i = I(U_i > 0)$.

The parameter of interest is set at $\kappa = -2\alpha\frac{\phi(0)}{\Phi(0)}$, while the auxiliary parameter is $\alpha = 1$. A DAG representation for this DGP is provided in Figure 1, where we see the direct effect $D \to Y$, which is what we want to identify and the fork structure $Y \leftarrow U \to D$, which is a confounding path. From the pair (Y_i, D_i), $i = 1, \ldots, n$ we estimate

$$Y_i = \gamma + \lambda D_i + \varepsilon_i.$$

The sample size for this experiment is set to $n = 1000$ with 10,000 repetitions. The histogram in Figure 2 shows how the estimates accumulate about zero and far away from the true parameter (the vertical red line on the left). As shown in Section 2 of the main text, $\lambda = 0$ immediately implies that the standard assumption $E(\varepsilon_i|D_i) = E(\varepsilon_i) = 0$ is satisfied.

By adding the variable X, we may obtain different graphical structures. In particular, we are interested in the situations where $X \to Y$ and $Y \leftarrow X \to D$ (see Figure 3 and Figure 4). In the first case, X does not further interfere with the identification of the direct effect (besides the confounding

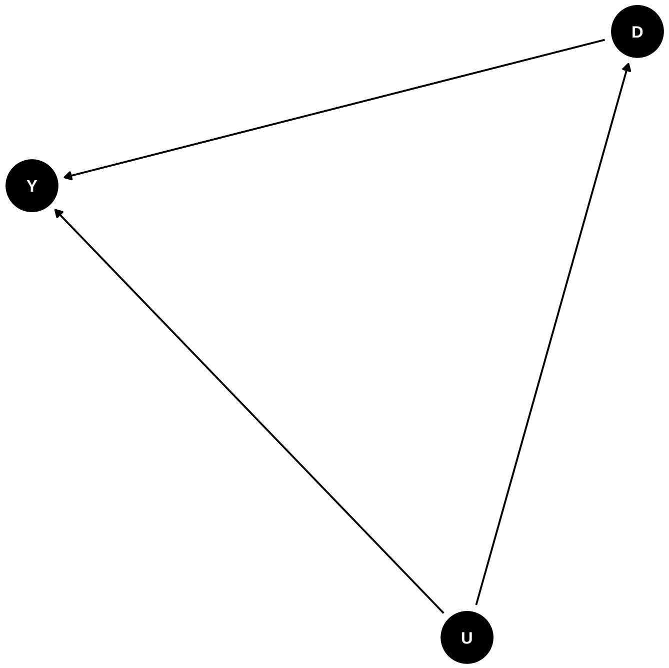

Figure 1: Confounding path and direct causal path with no additional regressors.

effect of U), while in the second case, further confounding effects are avoided by controlling for X.

Let us investigate those cases numerically. For simplicity, we consider X_i to be a scalar. Let us define our model as

$$Y_i = \kappa D_i + \delta X_i + \alpha U_i + \nu_i. \tag{2}$$

The DGP is the same as that introduced in Equation 1 with the addition that $X_i \sim \mathcal{N}(0, 1)$ and $\delta = 1$. The estimated model is

$$Y_i = \gamma + \lambda D_i + \beta X_i + \varepsilon_i.$$

We notice that Figure 5 is similar to Figure 2 which leads us to conclude that the inclusion of covariates cannot help us estimate the causal effect (this can also be evinced from the DAG in Figure 3).

Finally, in the last DGP, the variable X_i has an effect both on Y_i and D_i. As pointed out in Appendix A.2 of the main text, this specification has some implications on the definition of the

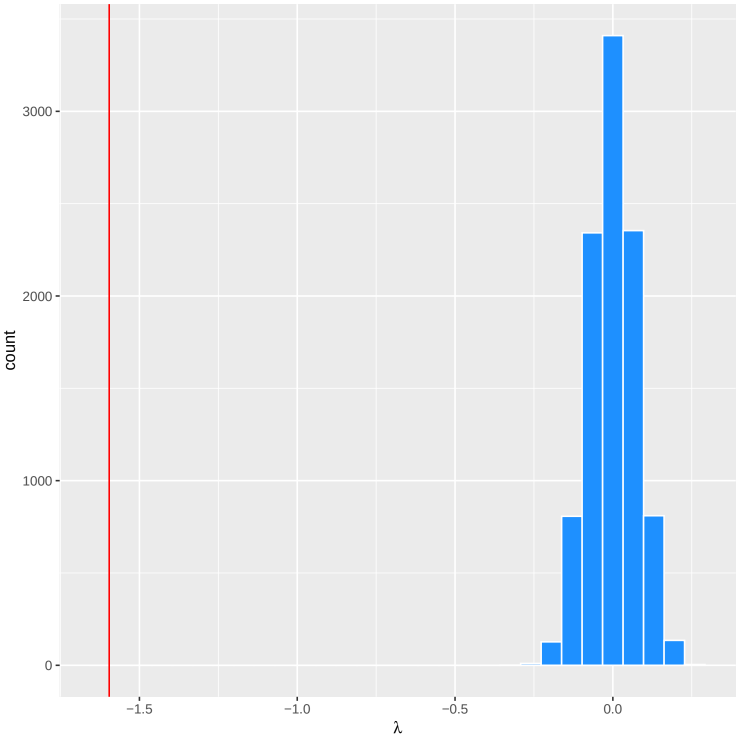

Figure 2: OLS estimates of λ. The red vertical line corresponds to the true value $\kappa = -2\frac{\phi(0)}{\Phi(0)}$.

causal effect. Let us define the DGP as

$$Y_i = \kappa_i D_i + \delta X_i + \alpha U_i + \nu_i \tag{3}$$

with $\kappa_i = -2\alpha\frac{\phi(\pi X_i)}{\Phi(\pi X_i)}$ and $D_i = I(U_i + \pi X_i > 0)$ where ν_i, X_i and U_i are identically and independently sampled from a standard normal distribution. Furthermore, $\pi = \delta = \alpha = 1$. The results of the simulation are displayed in Figure 6. Given that κ_i is not constant, we compute $\bar{\kappa} = \frac{1}{n}\sum_{i=1}^{n}\kappa_i$ for every Monte Carlo replication, and we plot its corresponding histogram alongside the histogram for $\hat{\lambda}$. Also in this case, we find that the estimates for λ are well away from the values of κ_i.

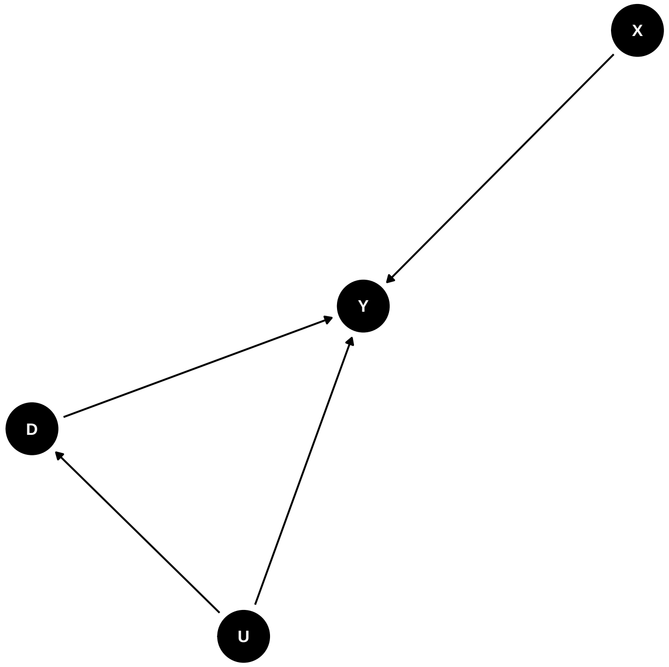

Figure 3: Confounding path and direct causal path where X only affects Y.

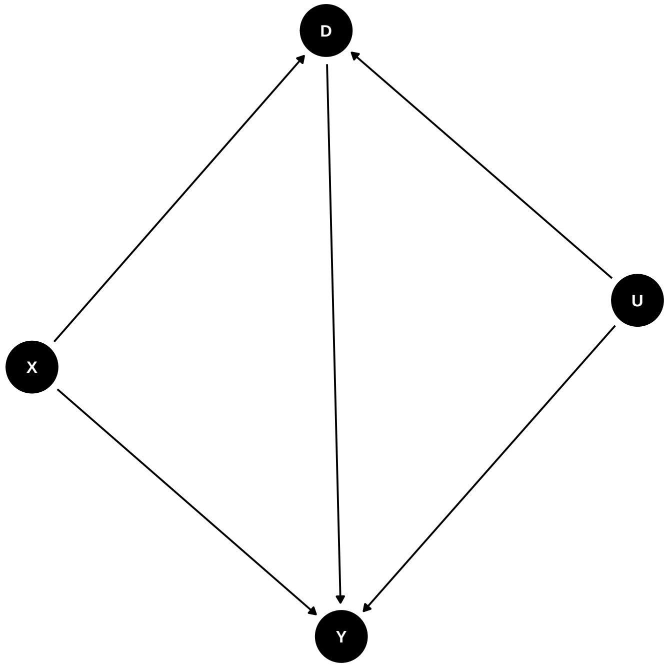

Figure 4: Confounding path and direct causal path where X affects both D and Y.

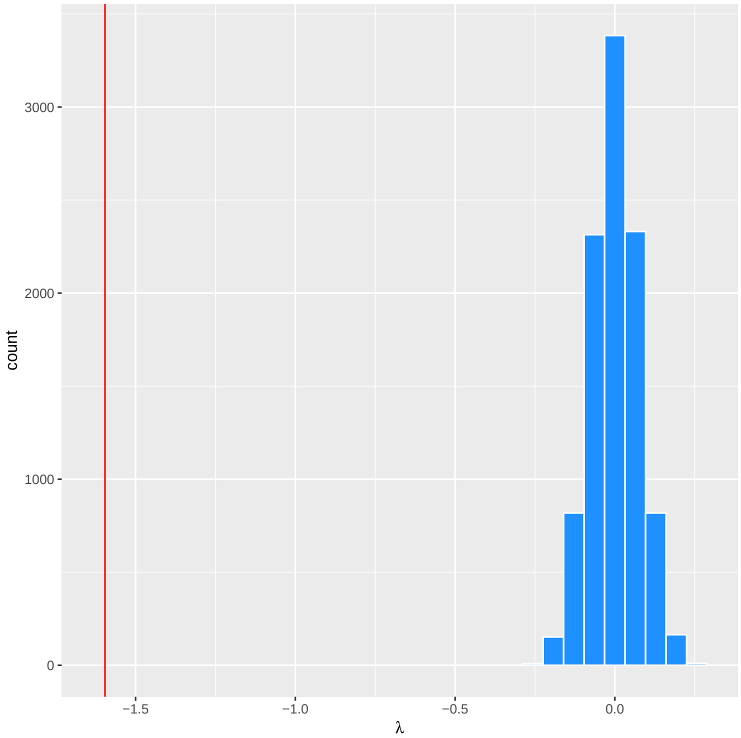

Figure 5: OLS estimates of λ. The red vertical line corresponds to the true value $\kappa = -2\frac{\phi(0)}{\Phi(0)}$.

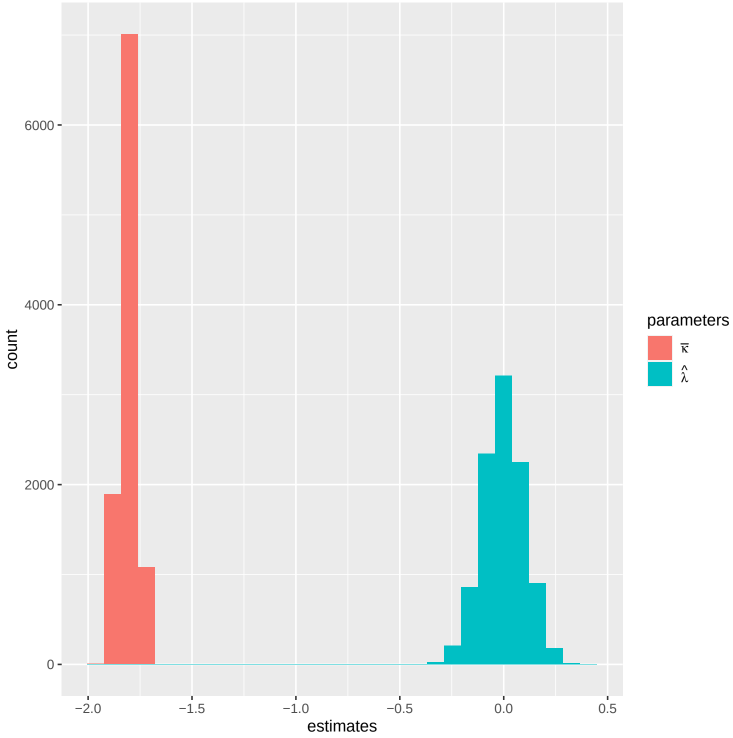

Figure 6: OLS estimates of λ and average value $\bar{\kappa}$.