The Role of Simulations in Econometrics Pedagogy

Authors: Anton Bekkerman

This is a postprint of an article that originally appeared in Wiley Interdisciplinary Reviews: Computational Statistics on December 2, 2014. WIREs Computational Statistics

Bekkerman, Anton. “The Role of Simulations in Econometrics Pedagogy.” Wiley Interdisciplinary Reviews: Computational Statistics 7, no. 2 (December 2, 2014): 160–165. doi:10.1002/wics.1342.

Article type: Focus Article

The Role of Simulations in Econometrics Pedagogy

(Article ID)

Anton Bekkerman

Montana State University

Keywords

econometrics, pedagogy, simulation, teaching

Abstract

This article assesses the role of simulation methods in econometrics pedagogy. Technological advances have increased researchers' abilities to use simulation methods and have contributed to a greater presence of simulation-based analysis in econometrics research. Simulations can also have an important role as pedagogical tools in econometrics education by providing a data-driven medium for difficult-to-grasp theoretical ideas to be empirically mimicked and the results to be visualized and interpreted accessibly. Three sample blueprints for implementing simulations to demonstrate foundational econometric principles provide a framework for gauging the effectiveness of simulation analysis as a pedagogical instrument.

Introduction

Axelrod (1997) describes a number of purposes for simulations, four of which are particularly applicable to econometrics: discovery, proof, prediction, and education. Discovery, proof, and prediction arguably provide the greatest scientific value. For example, objectives in simulation-based econometrics research include developing methods to characterize and understand highly complex models, providing empirical support for a particular hypothesis (for which there are few observational data and/or there are high barriers to obtaining additional data), and presenting statistical forecasts of future events.

Applications of simulation methods in modern econometrics can be loosely categorized as being one of four types (Kleiber and Zeileis, 2013). Monte Carlo methods can be used to evaluate small sample power of tests and assess the properties of competing estimators. When tractable analytical functional forms cannot be derived, simulation analyses can help understand and characterize these nonstandard functions and distributions. A third application is resampling or bootstrapping, which is commonly used to characterize properties of estimators (e.g., standard errors) by repeatedly sampling those properties from an approximating distribution that is often generated from observed data. Lastly, there has been substantial growth in the number of econometrics studies that employ simulation-based estimation techniques, such as Bayesian modeling and Markov Chain Monte Carlo (MCMC) methods.

Advances in computing efficiency and affordability have been important catalysts for increasing researchers' ability to access simulation analyses. The rapid improvements in the hardware and software capabilities of personal computers since the early to mid-1990s have had a non-trivial impact on the role of simulation methods in econometrics research (Doornik, 2006). Fontana (2006) notes that beginning at approximately the same time there has been a systematic attempt "at promoting simulation as a methodology in its own right and as a tool to embody and experiment theory." Furthermore, some of the field's most impactful journals call for new econometric methodology to be supported not only by a theoretical justification, but also, at least partially, by empirical evidence, which might be generated through artificial data (Kleiber and Zeileis, 2013). Such improvements in technological capabilities and increased collective support for simulation methods in research has certainly impacted the field of econometrics.

The growth provides some insights about the potential value that simulation methods can offer in understanding new, often complicated concepts in an applied, data-driven framework. Similarly to its application in supporting research, using simulations as an econometrics pedagogical tool can help develop a more effective learning environment (Sterling and Gray, 1991; Kennedy, 1998; DelMas, Garfield, and Chance, 1999; Lane and Tang, 2000). The nearly limitless ability (and flexibility) to empirically mimic and demonstrate foundational concepts provides a more accessible approach to teaching those concepts and can significantly improve students' understanding. This paper outlines three simulation algorithms that can serve as a foundation for developing effective learning opportunities in econometrics education. Applied examples accompany these algorithms and illustrate the types of insights that can be gleaned from using simulations as a pedagogical tool. As simulation-based methods continue to play an important role in the econometrics field, using simulations in learning environments can provide students with a foundation for better grasping fundamental econometric concepts, improving their understanding of existing research, and contributing to future research.

Simulations as a Pedagogical Tool

Despite the increasing role of simulations in econometrics research, these methods have played a less substantial role in econometrics education. In part, this may be because Monte Carlo techniques are usually not taught to students, and there is an underlying perception that setting up a simulation study is too self-evident to commit significant resources for introducing the concepts (Kiviet, 2012). Another reason may be an incomplete knowledge about the potential pedagogical benefits and techniques for teaching econometrics through simulations.

Econometrics courses provide insight into the intersection of economic principles and the statistical methods for empirically analyzing those principles. For many students, an introductory econometrics course is their first exposure to what can, at first glance, be relatively abstract concepts. Students are expected to grasp and master these concepts

quickly—a task that can make material challenging to learn for many students and similarly challenging to teach for instructors. Simulation methods can play an important role in bridging the learning gap by providing a data-driven medium that allows difficult-to-understand theoretical ideas to be mimicked, altered, and visualized accessibly.

Regression estimation methods are arguably the foundation of empirical econometric analysis. In learning and assessing the quality of potential estimators, students are asked to evaluate the statistical properties of an estimator and address three questions: Is the estimator unbiased? Is the estimator efficient? Is the estimator consistent? Understanding how to answer each question and its relevance is certainly important, but it is perhaps equally as important to also recognize factors that can lead to a biased, inefficient, and/or inconsistent estimator; the magnitude of those impacts; and ways to assess alternative estimators.

Simulations provide an opportunity to fully design the data generation process, introduce aspects that mimic common empirical problems, and assess estimators' quality under numerous alternative conditions. A useful way to demonstrate the application, role, and effectiveness of simulations in this context is with several examples. Each example consists of a blueprint outlining the general algorithm for performing a simulation exercise and an interpretation of results from a sample simulation that uses the algorithm. Each sample simulation assumes a 10,000 observation population with two exogenous variables, $X_1 \sim N(2,2)$ and $X_2 \sim N(0.25X_1, 1)$, an idiosyncratic error term, $\varepsilon \sim N(0,1)$, and the constructed variable, $Y = 1 + 0.5X_1 + 0.75X_2 + \varepsilon$.[1]

Algorithm 1: Estimator Sampling Distribution

One of the more challenging concepts for students to grasp is that estimated sample regression parameters are random variables. Empirically demonstrating this concept can be difficult using a sample of "real" observational data, because it is rare that an entire population is available from which more than a single sample can be drawn. Simulated economic environments, however, provide the ability to produce a population, repeatedly draw samples from this population, and generate simulated sampling distributions.

1. Generate a population matrix $\boldsymbol{X}$ of size $(i \times j)$ and an $(i \times 1)$ vector of exogenous idiosyncratic error terms, ε.
2. Using the data generated in 1, construct an $(i \times 1)$ dependent variable vector, $Y = f(\boldsymbol{X}, \varepsilon; \boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is a known (specified) vector of population parameters.
3. Sample with replacement to obtain n observations for variables Y and $\boldsymbol{X}$, where $n < j$.
4. Estimate $\hat{\boldsymbol{\beta}}$ and store the parameter values.
5. Repeat steps 2–4 m times, where m is reasonably large to ensure that the simulated empirical distribution is a good approximation of the sampling distribution of $\hat{\boldsymbol{\beta}}$ (e.g., $m = 1000$).

After the simulation is completed, distributional properties of the m estimated $\hat{\boldsymbol{\beta}}$ parameters can be assessed using descriptive statistics and empirical density plots. Issues such as sample selection, omitted variables, and measurement error, among others, can be evaluated with respect to their impacts on biasing $E[\hat{\boldsymbol{\beta}}]$ from the known population parameter vector, $\boldsymbol{\beta}$.

Figure 1 shows kernel density functions of simulated sampling distributions for an estimated parameter in a fully specified ordinary least squares (OLS) model and in an OLS model with an omitted variable. That is, each sampling distribution represents values of the estimated parameter $\hat{\beta}_2$ from a fully specified linear model, $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$, and the model, $y = \beta_0 + \beta_2 x_2 + e$. Sampling from the population and estimation was repeated $m = 1000$ times with each sample containing $n = 250$ observations. Figure 1 shows that the kernel density function associated with the simulated sampling distribution of $\hat{\beta}_2$ from the fully specified model is distributed approximately normally around the true population parameter, $\beta_2 = 0.75$. However, the simulated sampling distribution of $\hat{\beta}_2$ from the model that omits the X_1 variable is upward-biased relative to the true parameter value.

[1] Simulation examples were performed using SAS software, and the population data were generated using an initial seed value of 12345.

[Insert Figure 1 here]

The omitted variable bias examples demonstrates how simulation methods can be used to illustrate the effects on the central tendency of the sampling distribution. Similar exercises that show sample selection or measurement effects can provide additional insights, because the extent of the bias is conditional on the distribution of the error terms.

Algorithm 2: Asymptotic Properties

Another conceptually difficult topic to master is the large-number properties of estimators. Moreover, it is also unlikely that observational data can be used to demonstrate these properties empirically. Simulations provide an opportunity to examine estimator properties across a broad (effectively limitless) range of data. A classic example of asymptotic estimator properties is consistency—the convergence of the sampling variance around the sampling distribution's central tendency as the number of observations in the sample increases. That is, the variance of the sampling distribution becomes tighter (smaller) around the true population central tendency of a parameter.

1. Generate a population matrix $\boldsymbol{X}$ of size $(i \times j)$ and an $(i \times 1)$ vector of exogenous idiosyncratic error terms, ε.
2. Using the data generated in 1, construct an $(i \times 1)$ dependent variable vector, $Y = f(\boldsymbol{X}, \varepsilon; \boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is a known (specified) vector of population parameters.
3. Sample with replacement to obtain n observations for variables Y and $\boldsymbol{X}$, where $n < j$.
4. Estimate $\hat{\boldsymbol{\beta}}$ and store the parameter values.
5. Repeat steps 2–4 m times, where m is reasonably large to ensure that the simulated empirical distribution is a good approximation of the sampling distribution of $\hat{\boldsymbol{\beta}}$ (e.g., $m = 1000$).
6. Repeat steps 2–5 for different values of n (e.g., $n = 25$, $n = 250$, $n = 2500$).

The empirical density functions for the simulated sampling distributions (under the different assumptions about the size of n) can then be visually assessed. For example, Figure 2 illustrates the simulated sampling distributions of the OLS estimated parameter $\hat{\beta}_1$ in the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$. Sampling from the population and estimation was repeated $m = 1000$ times for each of the three sample size assumptions, $n = 25$, $n = 250$, and $n = 2500$. The figure shows that each simulated sampling distribution is centered around the true population value of the parameter, $\beta_1 = 0.5$, but that the dispersion around the central tendency decreases as the number of sample observations increases. It would be straightforward to alter the simulation algorithm and demonstrate the impacts of estimators that are not consistent.

[Insert Figure 2 here]

Algorithm 3: Estimator Efficiency and Inferences

Numerous issues can cause estimators to be inefficient and/or lead to biased standard error estimates. Within the classic Gauss-Markov linear least squares regression framework, these issues can stem from non-spherical disturbances where data are heteroskedastic or serially autocorrelated. For example, in a food expenditure study, heteroskedasticity may arise as a result of differences in income levels, unobserved factors affecting the dependence of expenditures across time can lead to serial correlation, and/or error terms may be correlated within a particular cluster of observations (e.g., a geographic region) but unrelated to errors outside of that cluster. It can be challenging to understand the impacts of an estimator being inefficient and/or incorrectly estimating standard errors. The following simulation of heteroskedastic data demonstrates an applied approach to assessing these challenges.

1. Generate a population matrix $\boldsymbol{X}$ of size $(i \times j)$ and an $(i \times 1)$ vector of exogenous idiosyncratic error terms, $\nu = f(\boldsymbol{X}, \varepsilon; \alpha)$, where $E[\varepsilon\varepsilon'] = \sigma^2 I_n$ and α is the parameter characterizing the degree of heteroskedasticity. For example, ν may take on the multiplicative form $\nu = \varepsilon \times \exp(\alpha \cdot X)$, where X represents one of the variables in the vector $\boldsymbol{X}$.

2. Using the data generated in 1, construct an $(i \times 1)$ dependent variable vector, $Y = f(\boldsymbol{X}, \nu; \boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is a known (specified) vector of population parameters. Choose a value for the α parameter that corresponds to a desired level of heteroskedasticity.

3. Sample with replacement to obtain n observations for variables Y and $\boldsymbol{X}$, where $n < j$.

4. Estimate $\hat{\boldsymbol{\beta}}$ and store the parameter values. This step can be performed using alternative estimators and standard error correction methods (e.g., OLS, OLS with heteroskedasticity-robust standard errors, weighted least squares) to demonstrate the impacts on estimator efficiency and standard error bias.

5. Repeat steps 2–4 m times, where m is reasonably large to ensure that the simulated empirical distribution is a good approximation of the sampling distribution of $\hat{\boldsymbol{\beta}}$ (e.g., $m = 1000$).

6. Perform stems 2–5 across a range of values for the heteroskedasticity parameter α.

This simulation procedure provides an opportunity to assess the manner in which efficiency and standard error estimates are affected across varying degrees of heteroskedasticity and across various estimators and variance correction techniques. For example, consider that the population model for Y described above is now a function of exogenous variables X_1 and X_2 as well as the error term $\nu = \varepsilon \times \exp(\alpha \cdot X_1)$, where $\alpha = \{0.0, 0.5, 1.0, 1.5, 2.0\}$.[2] For each level of α, samples of size $n = 250$ are used to estimate three models $m = 1000$ times: an OLS estimation with no correction for heteroskedasticity, a weighted least squares (WLS) estimation, and an OLS estimation with White's heteroskedasticity-robust standard errors.

Table 1 presents a comparison of standard error estimates across the different heteroskedasticity levels and estimated models. First, the data show that for each level of heteroskedasticity, the average OLS standard errors across the 1,000 simulations underestimate the more representative scale parameter characterized by the standard deviation of the OLS empirical parameter distribution. This implies that in cases when the data are heteroskedastic, estimated standard errors from OLS regressions are systematically biased. However, a comparison between average White's heteroskedasticity-robust standard errors and the standard deviation of the OLS empirical parameter distribution suggests that White's variance correction provides a reasonable characterization of the true sampling distribution dispersion. Lastly, the standard deviation of the WLS empirical parameter distribution indicates that this estimator is more efficient.

[Insert Table 1 here]

Figure 3 provides additional insights about the efficiency of alternative estimators in the presence of heteroskedasticity. The figure shows the simulated sampling distribution for the estimated parameter $\hat{\beta}_1$ from the OLS and WLS models. The kernel density functions indicate that the sampling distribution of the parameter associated with the OLS model has a larger dispersion than the distribution of the WLS model parameter, even though both estimators remain unbiased, as shown by both empirical distributions being centered around the population parameter value, $\beta_1 = 0.5$. This provides evidence that the WLS estimator is more efficient than the basic ordinary least squares estimator when heteroskedasticity is present.

[Insert Figure 3 here]

Conclusions

Over the past two decades, simulation methods have had an increasing role in the econometrics field, and these trends are likely to persist as hardware and software technologies continue to improve. The increased presence of simulation methods in top econometrics journals provides some evidence that these empirical tools do not simply occupy a small

[2] Note that to maintain a similar scale for ease of interpretation across the different values of α, the population observations of X_1 were generated from $X_1 \sim N(0, 0.2)$. All other distributional assumptions did not change.

niche in econometrics. The same reasons that make simulation methods useful in research (i.e., understanding complex ideas and models through an applied, data-driven medium) also apply in explaining how simulations can be valuable as pedagogical tools. That is, simulation methods can be effective in describing challenging foundational concepts, providing an effectively limitless, flexible learning environment in which theoretical topics can be represented and studied using data. This will improve students abilities to develop a deeper knowledge of essential econometric concepts, efficiently work with large datasets, and acquire and hone the skills to effectively use simulation methods and design new ones.

References

Axelrod, R. 1997. “Advancing the art of simulation in the social sciences.” In *Simulating social phenomena*. Springer, pp. 21–40.

DelMas, R.C., J. Garfield, and B. Chance. 1999. “A model of classroom research in action: Developing simulation activities to improve students statistical reasoning.” *Journal of Statistics Education* 7.

Doornik, J.A. 2006. “The role of simulation in econometrics.” *Palgrave Handbook of Econometrics* 1:787–811.

Engemann, K.M., and H.J. Wall. 2009. “A journal ranking for the ambitious economist.” *Federal Reserve Bank of St. Louis Review* 91.

Fontana, M. 2006. “Simulation in economics: Evidence on diffusion and communication.” *Journal of Artificial Societies and Social Simulation* 9.

Kalaitzidakis, P., T.P. Mamuneas, and T. Stengos. 2011. “An updated ranking of academic journals in economics.” *Canadian Journal of Economics/Revue canadienne d’économique* 44:1525–1538.

Kennedy, P.E. 1998. “Using Monte Carlo studies for teaching econometrics.” In W. Becker and M. Watts, eds. *Teaching Economics to Undergraduates: Alternatives to Chalk and Talk*. E. Elgar.

Kiviet, J.F. 2012. *Monte Carlo simulation for econometricians*. Now.

Kleiber, C., and A. Zeileis. 2013. “Reproducible econometric simulations.” *Journal of Econometric Methods* 2:89–99.

Lane, D.M., and Z. Tang. 2000. “Effectiveness of simulation training on transfer of statistical concepts.” *Journal of Educational Computing Research* 22:383–396.

Sterling, J., and M.W. Gray. 1991. “The effect of simulation software on students’ attitudes and understanding in introductory statistics.” *Journal of Computers in Mathematics and Science Teaching* 10:51–56.

Thomson Reuters. 2011. “Journal Citation Reports Social Sciences Edition.”

[**Further Reading**] Barreto, H., and F. Howland. 2005. *Introductory econometrics: using Monte Carlo simulation with microsoft excel*. Cambridge University Press.

[**Further Reading**] Robert, C.P., G. Casella, et al. 2010. *Introducing Monte Carlo methods with R*, vol. 18. Springer.

[**Further Reading**] Rubinstein, R.Y., and D.P. Kroese. 2011. *Simulation and the Monte Carlo method*, vol. 707. John Wiley & Sons.

[**Further Reading**] Wicklin, R. 2013. *Simulating data with SAS*. SAS Institute.

Cross-References

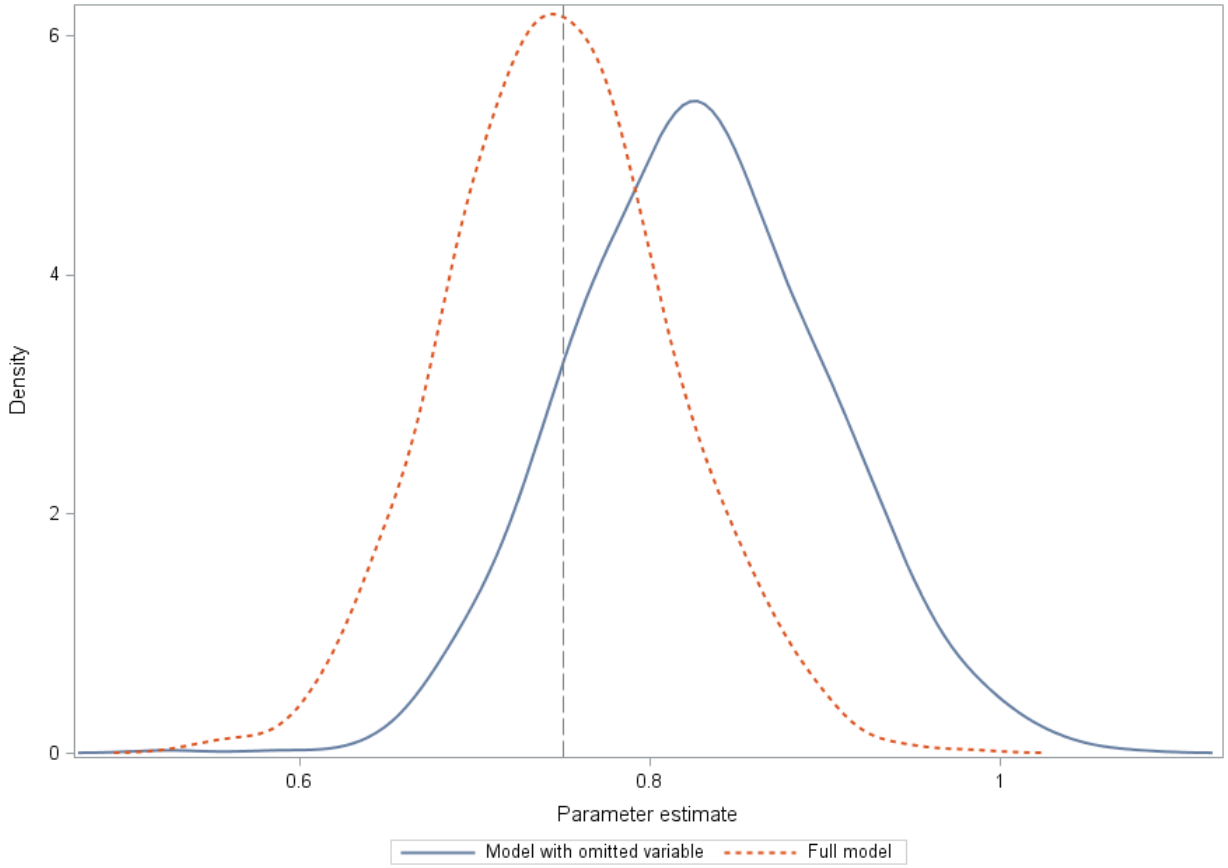
Monte Carlo methods, Resampling, Modeling and simulation

Table 1: Comparison of Heteroskedasticity Correction Methods across Increasing Levels of Heteroskedasticity (α)

	$\alpha = 0.0$	$\alpha = 0.5$	$\alpha = 1.0$	$\alpha = 1.5$	$\alpha = 2.0$
Avg. OLS Standard Error	0.3105	0.3155	0.3245	0.3425	0.3654
StdDev of OLS Empirical Distribution	0.3101	0.3284	0.3390	0.3851	0.4516
Avg. OLS White's Standard Error	0.3121	0.3201	0.3427	0.3880	0.4441
StdDev of WLS Empirical Distribution	0.3101	0.3180	0.3089	0.3067	0.3082

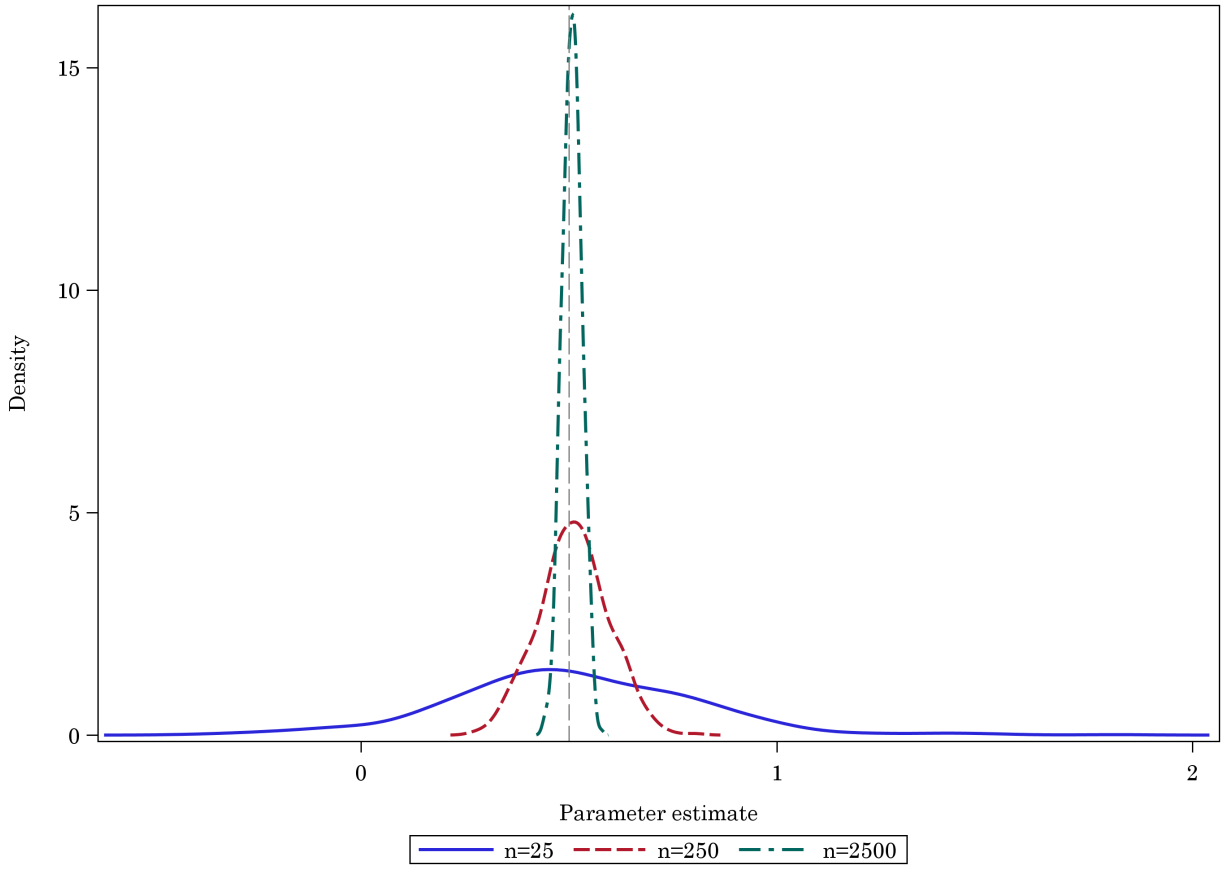
Notes: Average standard errors and standard deviations of empirical sampling distributions are presented for the parameter β_1, which is estimated using the linear model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + v$. In the population model, the error term is specified as $v = \varepsilon \times \exp(\alpha \cdot X_1)$, where $\varepsilon \sim N(0,1)$. A total of 1,000 simulations were performed using a sample size of 250 observations.

Figure 1: Impacts of Omitted Variables on the Empirical Estimator Distribution



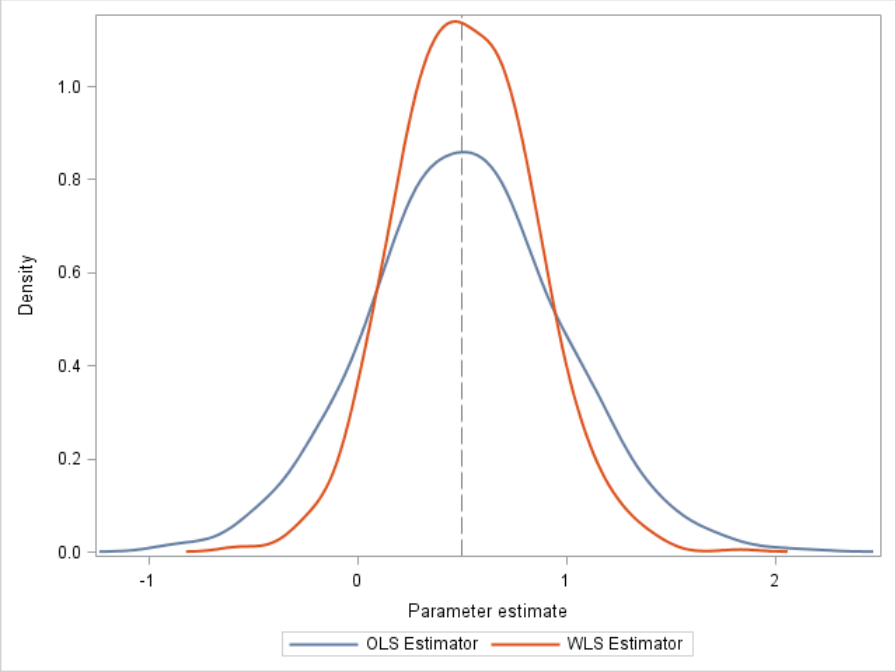
Notes: Kernel density functions of simulated sampling distributions are for the parameter β_2, which is estimated using ordinary least squares for a fully specified linear model, $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$, and a model with an omitted variable, $y = \beta_0 + \beta_2 x_2 + e$. Sampling from the population and estimation was repeated $m = 1000$ times with each sample containing $n = 250$ observations. The population value is $\beta_2 = 0.75$ and is indicated by the dashed vertical line.

Figure 2: Empirical Estimator Distributions under Alternative Sample-Size Assumptions



Notes: Kernel density functions of simulated sampling distributions are for the parameter β_1, which is estimated using the linear model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$. Sampling from the population and estimation was repeated $m = 1000$ times under three sample-size assumptions, $n = 25$, $n = 250$, and $n = 2500$. The population value is $\beta_1 = 0.50$ and is indicated by the dashed vertical line.

Figure 3: Empirical Distributions of OLS and WLS Estimators



Notes: Kernel density functions of simulated sampling distributions are presented for the parameter β_1, which is estimated using the linear model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + v$. In the population model, the error term is specified as $v = \varepsilon \times (X_1^{\alpha})$, where $\varepsilon \sim N(0,1)$. The population value is $\beta_1 = 0.50$ and is indicated by the dashed vertical line.