A Careful Consideration of CLARIFY

Simulation-Induced Bias in Point Estimates of Quantities of Interest

Carlisle Rainey[†]

September 5, 2022

Abstract

Some work in political methodology recommends that applied researchers obtain point estimates of quantities of interest by simulating model coefficients, transforming these simulated coefficients into simulated quantities of interest, and then averaging the simulated quantities of interest (e.g., CLARIFY). But other work advises applied researchers to directly transform coefficient estimates to estimate quantities of interest. I point out that these two approaches are not interchangeable and examine their properties. I show that the simulation approach compounds the transformation-induced bias identified by Rainey (2017), adding bias with direction and magnitude similar to the transformation-induced bias. I refer to this easily-avoided additional bias as "simulation-induced bias." Even if researchers use simulation to estimate standard errors, they should directly transform maximum likelihood estimates of coefficient estimates to obtain point estimates of quantities of interest.

[†]Carlisle Rainey is Associate Professor of Political Science, Florida State University, 540 Bellamy, Tallahassee, FL, 32306. (crainey@fsu.edu).

Introduction

Political scientists employ maximum likelihood (ML) to estimate a variety of statistical models. ML estimators have desirable and widely understood properties. But for many research questions, the model coefficient estimates do not directly interest the researcher. Following King, Tomz, and Wittenberg (2000), researchers often use the coefficient estimates to compute substantively meaningful quantities of interest, such as predicted probabilities, expected counts, marginal effects, and first differences. The literature offers two methods to compute point estimates for these quantities of interest.

Researchers estimate quantities of interest either by simulating quantities of interest and then averaging (e.g., King, Tomz, and Wittenberg 2000) or by directly transforming coefficients into quantities of interest (e.g., Herron 1999).[1] In practice, researchers' choice between these two approaches seems idiosyncratic rather than principled, depending on their preferred software package rather than any statistical criteria. Further, the methodological literature has not distinguished or compared the two approaches to estimating quantities of interest.

How does the simulation approach compare to directly transforming coefficients? Which should we prefer? Or are the two approaches interchangeable, as the literature seems to imply? Rainey (2017) shows that directly transforming coefficients into quantities of interest creates "transformation-induced" bias. I show that when researchers use the average of simulations to estimate quantities of interest, they replicate the logic of transformation-induced bias and add an additional, unnecessary bias to their estimates. I refer to this additional bias as "simulation-induced bias." I show that simulation-induced bias occurs *in addition to*

[1] CLARIFY for Stata (Tomz, Wittenberg, and King 2003) and Zelig for R (Imai, King, and Lau 2008; Choirat et al. 2018) simulate quantities of interest and find the average. The package dynsimpie for Stata also reports the average of simulations (Philips, Rutherford, and Whitten 2016*b*; Jung et al. 2020). The `margins` command in Stata (StataCorp 2017), the margins package in R (Leeper 2018), and the `predict()` function in R for the glm (R Core Team 2018) and polr (Venables and Ripley 2002) classes directly transform coefficients into the quantities of interest.

transformation-induced bias and is approximately the same size and direction. While this bias is usually small relative to the standard error, methodologists should not recommend methods that add unnecessary bias to point estimates. Instead, we should recommend methods that better adhere to the usual justifications.

The Current Practice

The Plug-In Estimator

First, researchers can directly transform the coefficient estimates into quantities of interest. The invariance property of ML estimators allows a researcher to find the ML estimate of a function of a parameter (i.e., a quantity of interest) by first using ML to estimate the model parameter and then applying the function to that estimate (or "plugging in the estimate") (King 1998, pp. 75–76; Casella and Berger 2002, pp. 320–321). I refer to this approach as the "plug-in" estimator. Importantly, the plug-in estimator *remains an ML estimator*. Thus, it retains all of the (desirable) properties of ML estimators (e.g., consistency, asymptotic efficiency).

As a concrete example, suppose a researcher uses ML to estimate a statistical model in which $y_i \sim f(\theta_i)$, where $i \in \{1, \ldots, N\}$ and f represents a probability distribution. The parameter θ_i connects to a design matrix X of k explanatory variables and a column of ones by a link function $g(\cdot)$, so that $g(\theta_i) = X_i\beta$, where $\beta \in \mathbb{R}^{k+1}$ represents a vector of parameter with length $k+1$. The researcher uses ML to compute estimates $\hat{\beta}^{\text{mle}}$ for the parameter vector β. I denote the function that transforms model coefficients into quantities of interest as $\tau(\cdot)$. For example, if the researcher uses a logit model and focuses on the predicted probability for a specific observation X_c, then $\tau(\beta) = \text{logit}^{-1}(X_c\beta) = \dfrac{1}{1 + e^{-X_c\beta}}$. The researcher can use the invariance property to quickly obtain a ML estimate of the predicted probability:

$$\hat{\tau}^{\text{mle}} = \tau\left(\hat{\beta}^{\text{mle}}\right) = \text{logit}^{-1}\left(X_c\hat{\beta}^{\text{mle}}\right) = \frac{1}{1+e^{-X_c\hat{\beta}^{\text{mle}}}}.$$

The Average-of-Simulations Estimator

Second, researchers can use the average of simulated quantities of interest as the point estimator. King, Tomz, and Wittenberg (2000) suggest the following approach:

1. *Fit the model.* Use ML to estimate the model coefficients $\hat{\beta}^{\text{mle}}$ and their covariance $\hat{V}\left(\hat{\beta}^{\text{mle}}\right)$.
2. *Simulate the coefficients.* Simulate a large number M of coefficient vectors $\tilde{\beta}^{(i)}$, for $i \in \{1, 2, \ldots, M\}$, using $\tilde{\beta}^{(i)} \sim \text{MVN}\left[\hat{\beta}^{\text{mle}}, \hat{V}\left(\hat{\beta}^{\text{mle}}\right)\right]$, where MVN represents the multivariate normal distribution.
3. *Convert simulated coefficients into simulated quantity of interest.* Compute $\tilde{\tau}^{(i)} = \tau\left(\tilde{\beta}^{(i)}\right)$ for $i \in \{1, 2, \ldots, M\}$. Most quantities of interest depend on the values of the explanatory variables. In this situation, researchers either focus on a specific observation (typically some kind of "average case") or average across all sample observations (Hanmer and Kalkan 2013). In any case, the transformation $\tau(\cdot)$ includes this choice.[2]
4. *Average the simulations of the quantity of interest.* Estimate the quantity of interest using the average of the simulations of the quantity of interest, so that $\hat{\tau}^{\text{avg}} = \frac{1}{M}\sum_{i=1}^{M}\tilde{\tau}^{(i)}$.[3]

I refer to this as the "average-of-simulations" estimator. But what are the properties of this estimator?

While the estimates it provides are sometimes similar to well-behaved plug-in estimates, King, Tomz, and Wittenberg (2000) develop their method informally. Much of their theoretical

[2] As King, Tomz, and Wittenberg (2000) note, this step might require additional simulation, to first introduce and then average over fundamental uncertainty. I ignore this additional step since (1) it is usually not necessary and (2) including it does not affect my argument.

[3] In the discussion that follows, I assume no Monte Carlo error exists in $\hat{\tau}^{\text{avg}}$. In other words, I assume that M is sufficiently large so that $\hat{\tau}^{\text{avg}} = \text{E}\left[\tau\left(\tilde{\beta}\right)\right]$, where $\tilde{\beta} \sim \text{MVN}\left[\hat{\beta}^{\text{mle}}, \hat{V}\left(\hat{\beta}^{\text{mle}}\right)\right]$.

argument happens quickly when they write "we draw many plausible sets of parameters from their posterior or sampling distribution" (p. 349).[4] One might justify their method from a frequentist perspective by first thinking of their method as "informal" Bayesian posterior simulation (Gelman and Hill 2006). This is helpful because the theory and practice of simulating from a posterior distribution are well-developed and widely understood. The Berstein-von Mises theorem (Van der Vaart 2000, pp. 140-146) guarantees, under relatively weak assumptions, that posterior simulations are asymptotically equivalent to the simulation procedure suggested by King, Tomz, and Wittenberg. And because the point estimator $\hat{\beta}^{\text{mle}}$ (and functions of $\hat{\beta}^{\text{mle}}$) is consistent, then the mean of the simulations (and the mean of functions of the simulations) are consistent as well. Therefore, one can defend $\hat{\tau}^{\text{avg}}$ on the grounds that it is a consistent estimator of τ.

However, the small sample properties of the average-of-simulations remain poorly understood. Methodologists and applied researchers seem to assume that $\hat{\tau}^{\text{avg}}$ is interchangeable with $\hat{\tau}^{\text{mle}}$. But below, I show that the average-of-simulations algorithm compounds the transformation-induced bias described by Rainey (2017) and adds a similar bias to the estimate that I refer to as "simulation-induced bias."

The Popularity of Simulation in Political Science

King, Tomz, and Wittenberg (2000) is among the most important methods papers published in political science. As of 2021, it has received about 4,500 citations according to Google Scholar and 1,800 according to ISI, making it the most cited paper in the *American Journal of Political Science*. But researchers do not use their simulation approach primarily for the point estimates—averaging simulations is an unnecessarily complex method to obtain a

[4]King, Tomz, and Wittenberg's claim that the draws are from the (frequentist) sampling distribution is incorrect. Take the sample mean as an example. If researchers could manipulate a sample to simulate draws from the sampling distribution, then they could compute the population mean with arbitrary precision, because the mean of the sampling distribution is the population mean. Of course, the precision of the sample mean is limited by the sample size–researchers cannot simulate from the sampling distribution.

point estimate. Instead, researchers use simulation to estimate standard errors for arbitrary quantities of interest. The standard deviation of the simulated $\tilde{\tau}$ serves as an easy, generic alternative to the tedious delta method (Herron 1999).[5] But if the researcher uses simulation to estimate the standard error, then it is tempting to just average the simulations to obtain the point estimate. And many authors suggest this shortcut.

King, Tomz, and Wittenberg (2000) explicitly recommend the average-of-simulations to obtain point estimates. They write: "[a]verage the simulated values to obtain a point estimate" (p. 351). King, Tomz, and Wittenberg's advice has echoed throughout the discipline, and methodologists continue to recommend the average-of-simulations. Zhirnov, Moral, and Sedashov (2022*c*), for example, explicitly recommend the average-of-simulations for their point estimates.[6]

Extensions of King, Tomz, and Wittenberg's approach are developed in other highly-cited methods work. Tomz, Tucker, and Wittenberg (2002) extend the approach to multiparty electoral data. Berry, DeMeritt, and Esarey (2010) recommend the approach for testing interactive hypotheses with binary outcomes. Philips, Rutherford, and Whitten (2016*a*) apply the approach to dynamic compositional data. Zhirnov, Moral, and Sedashov (2022*c*) extend the approach distribution-weighted average marginal effects. Brambor, Clark, and Golder

[5] Alongside the point estimate, researchers usually estimate the standard error. I recognize that a discussion of the properties of point estimates raises seemingly-urgent questions about the behavior of standard error estimates. But standard errors fall outside the scope of this paper. The omission is not as unsatisfying as it initially seems. From a frequentist perspective, point and interval estimates require separate consideration because distinct theoretical tools are used to evaluate the two. Rather than evaluate a suite of interval estimators, I point researchers to three options with well-understood frequentist properties for creating confidence intervals (in addition to King, Tomz, and Wittenberg's proposal): non-parametric bootstrap, the parametric bootstrap, and the delta method (see Efron and Tibshirani 1993 and Efron and Hastie 2021 for an overview).

[6] Popular software implements King, Tomz, and Wittenberg's (2000) suggestion to average the simulations. In 2021, the R package Zelig (Choirat et al. 2018) received more than 75,000 downloads from RStudio's CRAN mirror alone. It has received over 675,000 downloads since the mirror began. The authors' popular CLARIFY software (Tomz, Wittenberg, and King 2003) has received almost 2,000 citations. Indeed, "CLARIFY" has become a victim of genericization in political science: Rainey (2016) writes "I compute and plot estimates[...] for predicted probabilities, first-differences, and second-differences using CLARIFY-like simulation" (pp. 634-635). Likewise, Zhirnov, Moral, and Sedashov (2022*c*) provide software which supplies the average-of-simulations (Zhirnov, Moral, and Sedashov 2022*a*,*b*).

(2006), the most cited paper *Political Analysis* and the 7th most cited paper ever published in any political science journal, recommend the approach to compute point estimates of marginal effects for models of limited dependent variables. Each of these papers recommends averaging the simulations to obtain point estimates.

But researchers do not need to obtain point estimates by averaging simulations, even if they use simulation to estimate standard errors. Indeed, Herron (1999) and King and Zeng (2002) recommend obtaining point estimates using the plug-in principle and estimating the standard error using simulation. Foundational texts such as Casella and Berger (2002, pp. 320–321) and Greene (2012, p. 521) recommend the plug-in estimator. King (1998) recommends the plug-in estimator.

While many papers recommend computing quantities of interest and suggest either the plug-in or average-of-simulations estimator, none defend their choice or discuss the alternative. Thus, the political methodology literature has treated the average-of-simulations and the plug-in estimators as interchangeable. As an example, Hanmer and Kalkan (2013) use the plug-in estimator in their Monte Carlo simulations but the average-of-simulations in their empirical illustrations. As another example, King, Tomz, and Wittenberg (2000) recommend the average-of-simulation to political scientists, but King and Zeng (2002) recommend the plug-in estimate to medical researchers. However, I show that the average-of-simulations and plug-in estimators are not interchangeable. Instead, the average-of-simulations introduces a meaningful bias analogous to the transformation-induced bias identified by Rainey (2017). Fortunately, the plug-in estimator avoids this bias and is at least as easy to compute as the average-of-simulations.

A Theory of Simulation-Induced Bias

Before developing the theory of simulation-induced bias, I review the concept of transformation-induced bias from Rainey (2017). As Rainey (2017) shows, transforming unbiased model coefficient estimates can introduce bias into estimates of quantities of interest. Rainey (2017, p. 404) decomposes the bias in the estimate of the quantity of interest, which he refers to as total τ-bias, into two components: transformation-induced τ-bias and coefficient-induced τ-bias. Rainey (2017) defines these as

$$\text{total } \tau\text{-bias} = \underbrace{\mathrm{E}\left[\tau\left(\hat{\beta}^{\text{mle}}\right)\right] - \tau\left[\mathrm{E}\left(\hat{\beta}^{\text{mle}}\right)\right]}_{\text{transformation-induced}} + \overbrace{\tau\left[\mathrm{E}\left(\hat{\beta}^{\text{mle}}\right)\right] - \tau\left(\beta\right)}^{\text{coefficient-induced}}. \tag{1}$$

Transformation-induced τ-bias behaves systematically. The shape of the transformation $\tau(\cdot)$ determines the direction of the bias. In general, any strictly convex (concave) $\tau(\cdot)$ creates upward (downward) transformation-induced τ-bias. The direction and magnitude of the coefficient-induced τ-bias depend on the choice of $\tau(\cdot)$ and the bias in the coefficient estimates, but an unbiased estimator $\hat{\beta}^{\text{mle}}$ implies the absence of coefficient-induced τ-bias.

But Rainey (2017) does not consider the average-of-simulations estimator. This raises the question: Does the average-of-simulations estimator $\hat{\tau}^{\text{avg}}$ suffer the same transformation-induced bias as the plug-in estimator $\hat{\tau}^{\text{mle}}$? I now turn to the average-of-simulations estimator and develop the idea of "simulation-induced bias."

If the transformation of coefficient estimates into an estimate of the quantity of interest is always convex (or always concave), then Jensen's inequality allows the simple statement relating $\hat{\tau}^{\text{avg}}$ and $\hat{\tau}^{\text{mle}}$ given in Lemma 1 .

Lemma 1 *Suppose a nondegenerate ML estimator $\hat{\beta}^{mle}$. Then any strictly convex (concave) $\tau(\cdot)$ guarantees that $\hat{\tau}^{avg}$ is strictly greater (less) than $\hat{\tau}^{mle}$.*

This result is intuitive. Since I simulate using a multivariate normal distribution, $\tilde{\beta}$ has a symmetric distribution. But the distribution of $\tau(\tilde{\beta})$ is *not* symmetric. If $\tilde{\beta}$ happens to fall below the mode $\hat{\beta}^{\text{mle}}$, then $\tau(\cdot)$ pulls $\tau(\tilde{\beta})$ in toward $\hat{\tau}^{\text{mle}}$. If $\tilde{\beta}$ happens to fall above the mode $\hat{\beta}^{\text{mle}}$, then $\tau(\cdot)$ pushes $\tau(\tilde{\beta})$ away from $\hat{\tau}^{\text{mle}}$. This creates a right-skewed distribution for $\tau(\tilde{\beta})$, which pushes the average $\hat{\tau}^{\text{avg}}$ above $\hat{\tau}^{\text{mle}}$. See Appendix A for the proof.

For a convex transformation, Lemma 1 shows that $\hat{\tau}^{\text{avg}}$ is always larger than $\hat{\tau}^{\text{mle}}$. I refer to the expectation of this difference between $\hat{\tau}^{\text{avg}}$ and $\hat{\tau}^{\text{mle}}$ as "simulation-induced bias," so that

$$\text{simulation-induced } \tau\text{-bias} = \mathrm{E}\left(\hat{\tau}^{\text{avg}}\right) - \mathrm{E}\left(\hat{\tau}^{\text{mle}}\right).$$

Theorem 1 compares the sum of simulation- and transformation-induced τ-bias in $\hat{\tau}^{\text{avg}}$ to transformation-induced τ-bias in $\hat{\tau}^{\text{avg}}$.

Theorem 1 *Suppose a nondegenerate ML estimator $\hat{\beta}^{mle}$. Then for any strictly convex or concave $\tau(\cdot)$, the sum of the simulation-induced and transformation-induced τ-bias in $\hat{\tau}^{avg}$ is strictly greater in magnitude than the transformation-induced τ-bias in $\hat{\tau}^{mle}$.*

Regardless of the direction of simulation-induced and transformation-induced τ-bias, Theorem 1 shows that the magnitude of the combination in $\hat{\tau}^{\text{avg}}$ is *always* larger than the transformation-induced bias alone in $\hat{\tau}^{\text{mle}}$ for strictly convex or concave $\tau(\cdot)$. The proof follows directly from Jensen's inequality, but see Appendix A for the details.

Theorem 1 shows that $\hat{\tau}^{\text{avg}}$ compounds transformation-induced τ-bias with simulation-induced τ-bias. But is this bias substantively important? An analytical approximation provides a helpful guideline.

I approximate the simulation-induced τ-bias in $\hat{\tau}^{\text{avg}}$ as

$$\begin{aligned}
\text{simulation-induced } \tau\text{-bias in } \hat{\tau}^{\text{avg}} &= \underbrace{\left(\mathrm{E}\left(\hat{\tau}^{\text{avg}}\right) - \tau\left[\mathrm{E}\left(\hat{\beta}^{\text{mle}}\right)\right]\right)}_{\text{t.i. } \tau\text{-bias in } \hat{\tau}^{\text{avg}}} - \underbrace{\left(\mathrm{E}\left(\hat{\tau}^{\text{mle}}\right) - \tau\left[\mathrm{E}\left(\hat{\beta}^{\text{mle}}\right)\right]\right)}_{\text{t.i. } \tau\text{-bias in } \hat{\tau}^{\text{mle}}} \\
&= \mathrm{E}\left(\hat{\tau}^{\text{avg}}\right) - \mathrm{E}\left(\hat{\tau}^{\text{mle}}\right) \\
&= \mathrm{E}\left(\hat{\tau}^{\text{avg}} - \hat{\tau}^{\text{mle}}\right) \\
&= \mathrm{E}\left(\mathrm{E}\left[\tau\left(\tilde{\beta}\right)\right] - \tau\left(\hat{\beta}^{\text{mle}}\right)\right) \\
&= \mathrm{E}\left(\underbrace{\mathrm{E}\left[\tau\left(\tilde{\beta}\right)\right] - \tau\left[E\left(\tilde{\beta}\right)\right]}_{\substack{\text{approximated in Eq. 1,} \\ \text{p. 405, of Rainey (2017)}}}\right) \\
&\approx \mathrm{E}\left[\frac{1}{2}\sum_{r=1}^{k+1}\sum_{s=1}^{k+1} H_{rs}\left(\hat{\beta}^{\text{mle}}\right)\hat{V}_{rs}\left(\hat{\beta}^{\text{mle}}\right)\right], \qquad (2)
\end{aligned}$$

where the remaining expectation occurs with respect to $\hat{\beta}^{\text{mle}}$, $H\left(\hat{\beta}^{\text{mle}}\right)$ represents the Hessian matrix of second derivatives of $\tau(\cdot)$ at the point $\hat{\beta}^{\text{mle}}$ and, conveniently, $\hat{V}\left(\hat{\beta}^{\text{mle}}\right)$ represents the estimated covariance matrix for $\hat{\beta}^{\text{mle}}$.

This approximation appears similar to the approximation for the transformation-induced τ-bias, which (adjusting notation slightly) Rainey (2017, p. 405, Eq. 1) computes as

$$\text{t.i. } \tau\text{-bias } \approx \frac{1}{2}\sum_{r=1}^{k+1}\sum_{s=1}^{k+1} H_{rs}\left[\mathrm{E}\left(\hat{\beta}^{\text{mle}}\right)\right] V_{rs}\left(\hat{\beta}^{\text{mle}}\right), \qquad (3)$$

where $H\left[\mathrm{E}\left(\hat{\beta}^{\text{mle}}\right)\right]$ represents the Hessian matrix of second derivatives of $\tau(\cdot)$ at the point $\mathrm{E}\left(\hat{\beta}^{\text{mle}}\right)$ and $V\left(\hat{\beta}^{\text{mle}}\right)$ represents the covariance matrix of the sampling distribution of $\hat{\beta}^{\text{mle}}$.

When one compares Equations 2 and 3, they yet again compare the *expectation of a function* with the *function of the expectation*. Therefore, Equations 2 and 3 are not exactly equal. But, as a rough guideline, one should expect them to be similar. And to the extent

that the two are similar, the additional simulation-induced τ-bias in $\hat{\tau}^{\text{avg}}$ is about the same as the transformation-induced τ-bias in $\hat{\tau}^{\text{mle}}$.

Because of the similarity between Equations 2 and 3, the simulation-induced τ-bias becomes large under the conditions identified by Rainey (2017) as leading to large transformation-induced τ-bias: when the non-linearity in the transformation $\tau(\cdot)$ is severe and when the standard errors of $\hat{\beta}^{\text{mle}}$ are large. While the transformation-induced τ-bias vanishes as the number of observations grows large, it can be substantively meaningful for the sample sizes commonly encountered in social science research. In standard modeling situations, Rainey (2017) demonstrates that transformation-induced bias can (1) be larger than the bias in the estimates of the model coefficients and (2) shrink to zero more slowly as the sample size increases. By extension, the same claims hold for simulation-induced bias (which, again, appears *in addition to* transformation-induced bias). Appendix C uses a realistic data-generating process to show that transformation-induced and coefficient-induced biases are similar and that both are similar to the well-known small-sample bias in logistic regression coefficients.

The Intuition of Simulation-Induced Bias

To develop the intuition for the theoretical results above, I examine a stylized example with simulations, an alternative analytical approach, and an empirical example.

Using a Drastic, Convex Transformation: $\tau(\mu) = \mu^2$

To develop an intuition for the simulation-induced τ-bias in $\hat{\tau}^{\text{avg}}$, consider the simple (unrealistic, but heuristically useful) scenario in which $y_i \sim \text{N}(0, 1)$, for $i \in \{1, 2, \ldots, n = 100\}$, and the researcher wishes to estimate μ^2. Suppose that the researcher knows that the variance equals one but does not know that the mean μ equals zero. The researcher uses

the unbiased ML estimator $\hat{\mu}^{\text{mle}} = \frac{\sum_{i=1}^{n} y_i}{n}$ of μ, but ultimately cares about the quantity of interest $\tau(\mu) = \mu^2$. The researcher can use the plug-in estimator $\hat{\tau}^{\text{mle}} = \left(\hat{\mu}^{\text{mle}}\right)^2$ of $\tau(\mu)$. Alternatively, the researcher can use the average-of-simulations estimator, estimating $\tau(\mu)$ as $\hat{\tau}^{\text{avg}} = \frac{1}{M} \sum_{i=1}^{M} \tau\left(\tilde{\mu}^{(i)}\right)$, where $\tilde{\mu}^{(i)} \sim \text{N}\left(\hat{\mu}^{\text{mle}}, \frac{1}{\sqrt{n}}\right)$ for $i \in \{1, 2, \ldots, M\}$.

The true value of the quantity of interest is $\tau(0) = 0^2 = 0$. However, the ML estimator $\hat{\tau}^{\text{mle}} = \left(\hat{\mu}^{\text{mle}}\right)^2$ equals zero if and only if $\hat{\mu}^{\text{mle}} = 0$. Otherwise, $\hat{\tau}^{\text{mle}} > 0$. Since $\hat{\mu}^{\text{mle}}$ almost surely differs from zero, $\hat{\tau}^{\text{mle}}$ is biased upward.

Moreover, even if $\hat{\mu}^{\text{mle}} = 0$, $\tilde{\mu}^{(i)}$ almost surely differs from zero. If $\tilde{\mu}^{(i)} \neq 0$, then $\left(\tilde{\mu}^{(i)}\right)^2 > 0$. Thus, $\hat{\mu}^{\text{avg}}$ is almost surely larger than the true value $\tau(\mu) = 0$ even when $\hat{\mu} = 0$.

I illustrate this fact clearly by repeatedly simulating y and computing $\hat{\tau}^{\text{mle}}$ and $\hat{\tau}^{\text{avg}}$. Figure 1 shows the first four of 10,000 total simulations. The figure shows how the unbiased estimate $\hat{\mu}^{\text{mle}}$ is translated into $\hat{\tau}^{\text{mle}}$ and $\hat{\tau}^{\text{avg}}$.

First, to find $\hat{\tau}^{\text{avg}}$, I complete three steps: (1) simulate $\tilde{\mu}^{(i)} \sim \text{N}\left(\hat{\mu}^{\text{mle}}, \frac{1}{10}\right)$ for $i \in \{1, 2, \ldots, M = 1,000\}$, (2) calculate $\tilde{\tau}^{(i)} = \tau\left(\tilde{\mu}^{(i)}\right)$, and (3) calculate $\hat{\tau}^{\text{avg}} = \frac{1}{M} \sum_{i=1}^{M} \tilde{\tau}^{(i)}$. The rug plot along the horizontal axis shows the distribution of $\tilde{\mu}$. The hollow points in Figure 1 shows the transformation of each point $\tilde{\mu}^{(i)}$ into $\tilde{\tau}^{(i)}$. The rug plot along the vertical axis shows the distribution of $\tilde{\tau}$. Focus on the top-left panel of Figure 1. Notice that $\hat{\mu}^{\text{mle}}$ estimates the true value $\mu = 0$ quite well. However, after simulating $\tilde{\mu}$ and transforming $\tilde{\mu}$ into $\tilde{\tau}$, the $\tilde{\tau}$s fall far from the true value $\tau(0) = 0$. The dashed orange line shows the average of $\tilde{\tau}$. Notice that although $\hat{\mu}^{\text{mle}}$ is unusually close to the truth $\mu = 0$ in this sample, $\hat{\tau}^{\text{avg}}$ is substantially biased upward.

Second, to find $\hat{\tau}^{\text{mle}}$, I transform $\hat{\mu}^{\text{mle}}$ directly using $\hat{\tau}^{\text{mle}} = \left(\hat{\mu}^{\text{mle}}\right)^2$. The solid green lines show this transformation. The convex transformation $\tau(\cdot)$ has the effect of lengthening the right tail of the distribution of $\tilde{\tau}$, pulling the average well above the mode. This provides the basic intuition for Lemma 1.

The remaining panels of Figure 1 repeat this process with three more random samples.

Each sample presents a similar story — the convex transformation stretches the distribution of $\tilde{\tau}$ to the right, which pulls $\hat{\tau}^{\text{avg}}$ above $\hat{\tau}^{\text{mle}}$.

I repeat this process 10,000 total times to produce 10,000 estimates of $\hat{\mu}^{\text{mle}}$, $\hat{\tau}^{\text{mle}}$, and $\hat{\tau}^{\text{avg}}$. Figure 2 shows the density plots for the 10,000 estimates (i.e., the sampling distributions of $\hat{\mu}^{\text{mle}}$, $\hat{\tau}^{\text{mle}}$, and $\hat{\tau}^{\text{avg}}$). As I show analytically, $\hat{\mu}^{\text{mle}}$ is unbiased with a standard error of $\frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{100}} = \frac{1}{10}$. Both $\hat{\tau}^{\text{mle}}$ and $\hat{\tau}^{\text{avg}}$ are biased upward, but, as Theorem 1 suggests, $\hat{\tau}^{\text{avg}}$ has more bias than $\hat{\tau}^{\text{mle}}$. And as the approximation suggestions, $\hat{\tau}^{\text{avg}}$ has about twice the bias of $\hat{\tau}^{\text{mle}}$. Indeed, the exact bias of each estimator is easy to compute for this simple example. Appendix B shows that the biases are $\frac{1}{n} = \frac{1}{100}$ and $\frac{2}{n} = \frac{2}{100}$ in this example.

Indeed, the exact bias of each estimator is easy to compute for this simple example. Appendix B shows that the biases are 1 = 1 and 2 = 2 in this example.

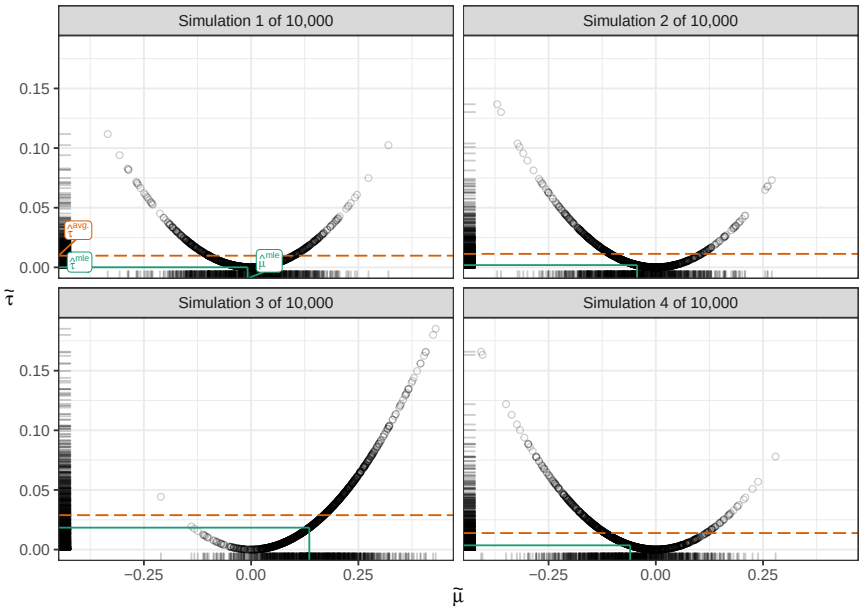

Figure 1: The first four Monte Carlo simulations of $\hat{mu}$. These four panels illustrate the relationship between $\hat{\tau}^{\text{mle}}$ and $\hat{\tau}^{\text{avg}}$ described by Lemma 1 and Theorem 1.

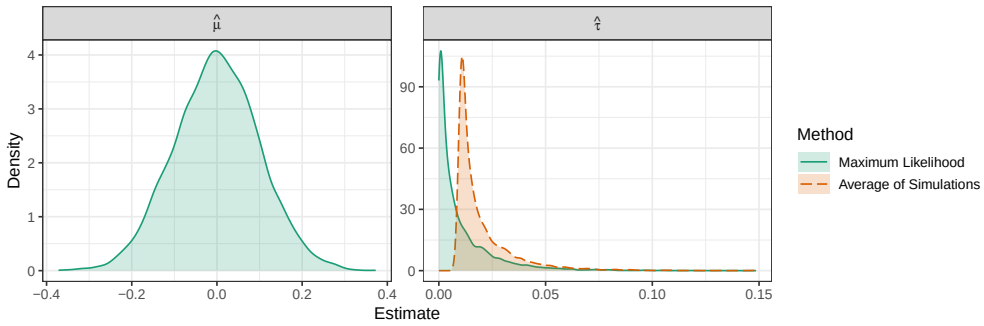

Figure 2: The sampling distributions of $\hat{\beta}^{\text{mle}}$, $\hat{\tau}^{\text{mle}}$, and $\hat{\tau}^{\text{avg}}$.

Using the Law of Iterated Expectations

One can also develop the argument analytically via the law of iterated expectations. It helps to alter the notation slightly, making two implicit dependencies explicit. I explain each change below and use the alternate, more expansive notation only in this section.

The law of iterated expectations states that $\mathrm{E}_Y\left(\mathrm{E}_{X|Y}(X \mid Y)\right) = \mathrm{E}_X(X)$, where X and Y represent random variables. The three expectations occur with respect to three different distributions: E_Y denotes the expectation with respect to the marginal distribution of Y, $\mathrm{E}_{X|Y}$ denotes the expectation with respect to the conditional distribution of $X \mid Y$, and E_X denotes the expectation with respect to the marginal distribution of X.

Outside of this section, the reader should understand that the distribution of $\tilde{\beta}$ depends on $\hat{\beta}^{\text{mle}}$ and could be written as $\tilde{\beta} \mid \hat{\beta}^{\text{mle}}$. To remain consistent with previous work, especially King, Tomz, and Wittenberg (2000) and Herron (1999), I use $\tilde{\beta}$ to represent $\tilde{\beta} \mid \hat{\beta}^{\text{mle}}$. The definition of $\tilde{\beta}$ makes this usage clear. In this section only, I use $\tilde{\beta} \mid \hat{\beta}^{\text{mle}}$ to represent the conditional distribution of $\tilde{\beta}$ and use $\tilde{\beta}$ to represent the unconditional distribution of $\tilde{\beta}$. Intuitively, one might imagine (1) generating a data set y, (2) estimating $\hat{\beta}^{\text{mle}}$, and (3) simulating $\tilde{\beta} \mid \hat{\beta}^{\text{mle}}$. If I perform steps (1) and (2) just once, but step (3) repeatedly, I generate a sample from the conditional distribution $\tilde{\beta} \mid \hat{\beta}^{\text{mle}}$. If I perform steps (1), (2), and (3) repeatedly, then I generate a sample from the unconditional distribution $\tilde{\beta}$. The

unconditional distribution helps us understand the nature of the simulation-induced τ-bias.

Applying the law of iterated expectations, I obtain $\mathrm{E}_{\tilde{\beta}}\left(\tilde{\beta}\right) = \mathrm{E}_{\hat{\beta}^{\text{mle}}}\left(\mathrm{E}_{\tilde{\beta}|\hat{\beta}^{\text{mle}}}(\tilde{\beta} \mid \hat{\beta}^{\text{mle}})\right)$. The three identities below connect the three key quantities from Theorem 1 to three versions of $\mathrm{E}_{\hat{\beta}^{\text{mle}}}\left(\mathrm{E}_{\tilde{\beta}|\hat{\beta}^{\text{mle}}}(\tilde{\beta} \mid \hat{\beta}^{\text{mle}})\right)$, with the transformation $\tau(\cdot)$ applied at different points.

$$\tau\left[\underset{\hat{\beta}^{\text{mle}}}{\mathrm{E}}\left(\underset{\tilde{\beta}|\hat{\beta}^{\text{mle}}}{\mathrm{E}}\left(\tilde{\beta} \mid \hat{\beta}^{\text{mle}}\right)\right)\right] = \tau\left[\underset{\tilde{\beta}}{\mathrm{E}}\left(\tilde{\beta}\right)\right] = \tau\left[\mathrm{E}\left(\hat{\beta}^{\text{mle}}\right)\right], \tag{4}$$

$$\underset{\hat{\beta}^{\text{mle}}}{\mathrm{E}}\left(\tau\left[\underset{\tilde{\beta}|\hat{\beta}^{\text{mle}}}{\mathrm{E}}\left(\tilde{\beta} \mid \hat{\beta}^{\text{mle}}\right)\right]\right) = \underset{\hat{\beta}^{\text{mle}}}{\mathrm{E}}\left(\tau\left[\hat{\beta}^{\text{mle}}\right]\right) = \underset{\hat{\beta}^{\text{mle}}}{\mathrm{E}}\left(\hat{\tau}^{\text{mle}}\right), \text{ and} \quad \longleftarrow \text{ Switch } \tau \text{ and an E once.} \tag{5}$$

$$\underset{\hat{\beta}^{\text{mle}}}{\mathrm{E}}\left(\underset{\tilde{\beta}|\hat{\beta}^{\text{mle}}}{\mathrm{E}}\left(\tau\left[\tilde{\beta} \mid \hat{\beta}^{\text{mle}}\right]\right)\right) = \underset{\tilde{\beta}}{\mathrm{E}}\left(\tau\left[\tilde{\beta}\right]\right) = \underset{\tilde{\beta}}{\mathrm{E}}\left(\hat{\tau}^{\text{avg}}\right). \quad \longleftarrow \text{ Switch } \tau \text{ and an E again.} \tag{6}$$

If I subtract Equation 5 from Equation 4, I obtain the transformation-induced τ-bias in $\hat{\tau}^{\text{mle}}$ (see Equation 1 for the definition of transformation-induced τ-bias). To move from Equation 4 to Equation 5 I must swap $\tau(\cdot)$ with an expectation once. This implies that, if $\tau(\cdot)$ is convex, Equation 5 must be greater than Equation 4. This, in turn, implies that the bias is positive.

To obtain the τ-bias in $\hat{\tau}^{\text{avg}}$ I must subtract Equation 6 from Equation 4. But to move from Equation 4 to Equation 6 I must swap $\tau(\cdot)$ with an expectation *twice*. Again, if $\tau(\cdot)$ is convex, then Equation 6 must be greater than Equation 4. However, because one should expect $\hat{\beta}^{\text{mle}}$ and $\tilde{\beta} \mid \hat{\beta}^{\text{mle}}$ to have similar distributions, one should expect the additional swap to roughly double the bias in $\hat{\tau}^{\text{avg}}$ compared to $\hat{\tau}^{\text{mle}}$. This additional swap creates the additional, simulation-induced τ-bias.

Using Holland (2015)

Holland (2015) presents a nuanced theory that describes the conditions under which politicians choose to enforce laws and supports the theoretical argument with a rich variety of evidence. In particular, it elaborates on the *electoral* incentives of politicians to enforce laws. I borrow three Poisson regressions and hypotheses about a single explanatory variable to illustrate how the plug-in estimates can differ from the average-of-simulations estimate.

Holland writes:

> My first hypothesis is that enforcement operations drop off with the fraction of poor residents in an electoral district. So district poverty should be a negative and significant predictor of enforcement, but only in politically decentralized cities [Lima and Santiago]. Poverty should have no relationship with enforcement in politically centralized cities [Bogota] once one controls for the number of vendors.

I use Holland's hypothesis and data to illustrate the behavior of the average-of-simulations and plug-in estimators. I refit Model 1 from Table 2 in Holland (2015) for each city. I then use each model to compute the percent increase in the enforcement operations for each district in the city if the percent of the district in the lower class dropped by half. For example, in the Villa Maria El Triunfo district in Lima, 84% of the district is in the lower class. If this dropped to 42%, then the average-of-simulations estimate suggests that the number of enforcement operations would increase by about 284% (from about 5 to about 20). The plug-in estimate, on the other hand, suggests an increase of 264% (from about 5 to about 17). The plug-in estimate, then, is about 7% smaller than the average-of-simulations estimate—a small, but noticeable shrinkage.

Figure 3 shows how the estimates change (usually shrink) for all districts when I switch from the average-of-simulations estimates to plug-in estimates. Table 1 presents the details for the labeled cases in Figure 3. In Bogota, the estimate shrinks by 11% in Sante Fe and

16% in Usme. In Lima, the estimate shrinks by 5% in Chacalacayo and 7% Villa Maria El Triunfo. The shrinkage is much larger in Santiago, where the standard errors for the coefficient estimates are much larger. The estimate shrinks by about 47% in San Ramon and 53% in La Pintana. The median shrinkage is 7% in Bogota, 2% in Lima, and 36% in Santiago. For many districts in Santiago, the average-of-simulations estimate is about *twice* the plugin estimate. These estimates clearly show that the average of simulations and the ML estimates can differ meaningfully in actual analyses.

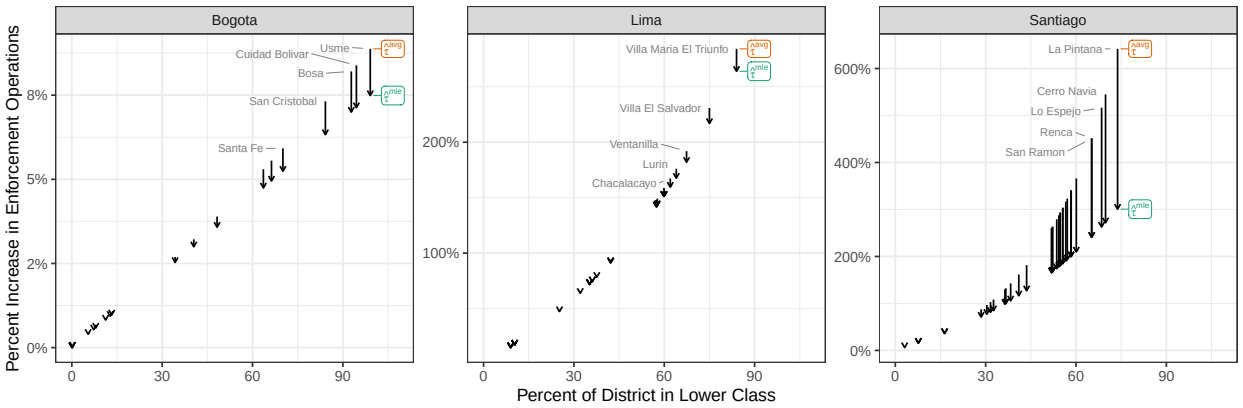

Figure 3: This figure compares the average-of-simulations estimates with the plug-in estimates using three Poisson regression models from Holland (2015). The quantity of interest is the percent increase in the enforcement operations when the percent of a district in the lower class drops by half. The arrows show how the estimates change when I switch from the average-of-simulations to the plug-in estimate.

Conclusion

Many social scientists turn to King, Tomz, and Wittenberg (2000) for advice on interpreting, summarizing, and presenting empirical results. The authors improved empirical research by highlighting the importance of substantively meaningful quantities of interest. I agree with King and Zeng's (2006) summary of the literature: "[w]hether such effects are calculated via

Table 1: This table presents the details for the districts labelled in Figure 3.

		Average of Simulations			ML Estimate			
City	**District**	**% Change[a]**	**From[b]**	**To[c]**	**% Change**	**From**	**To**	**Shrinkage[d]**
Bogota	**Usme**	9%	5.5	5.8	7%	5.4	5.8	16%
	Cuidad Bolivar	8%	6.1	6.4	7%	5.9	6.3	15%
	Bosa	8%	7.1	7.4	7%	6.9	7.4	15%
	San Cristobal	7%	12.6	13.4	6%	12.5	13.3	14%
	Santa Fe	6%	27.3	29.0	5%	26.6	28.0	11%
Lima	**Villa Maria El Triunfo**	284%	5.3	19.5	264%	4.7	17.1	7%
	Villa El Salvador	231%	7.3	23.5	217%	6.8	21.4	6%
	Ventanilla	192%	8.4	23.4	182%	8.2	23.0	5%
	Lurin	176%	6.9	17.4	168%	6.4	17.1	5%
	Chacalacayo	167%	6.7	16.4	159%	6.2	16.1	5%
Santiago	**La Pintana**	642%	1.4	4.0	301%	0.8	3.4	53%
	Cerro Navia	545%	1.5	4.2	272%	1.0	3.6	50%
	Lo Espejo	517%	1.4	4.3	263%	1.0	3.5	49%
	Renca	451%	1.3	4.1	241%	1.0	3.4	47%
	San Ramon	451%	1.2	4.0	241%	1.0	3.3	47%

[a] Quantity of interest; percent change in enforcement operations when the percent in the lower class drops by half.
[b] Enforcement operations when the percent in the lower class equals its observed value.
[c] Enforcement operations when the percent in the lower class equals half its observed value.
[d] Shrinkage in the quantity of interest due to switching from the average of simulations to the ML estimator.

analytical derivation or what is now the more common approach of statistical simulation, political scientists have made much progress in learning how to make sophisticated methods speak directly to their substantive research questions" (p. 132).

Researchers estimate quantities of interest either by averaging simulated quantities of interest (e.g., CLARIFY in Stata, Zelig in R) or using the invariance property of maximum likelihood estimators (e.g., margins in Stata and R). In practice, researchers' choice between these two estimators seems idiosyncratic rather than principled, depending on their preferred software package rather than any statistical criteria. The methodological literature recommends both, but has not distinguished or compared the two approaches to estimating quantities of interest.

When researchers use the average of simulations (King, Tomz, and Wittenberg 2000) to estimate quantities of interest, they replicate the logic of transformation-induced bias (Rainey 2017) and add simulation-induced bias to the estimates. This additional bias is roughly the same magnitude and direction as transformation-induced bias and occurs *in*

addition to transformation-induced bias. While the additional bias is usually small relative to the standard error, methodologists should not recommend methods that add unnecessary bias to point estimates. Instead, we should recommend methods that better adhere to the usual evaluative standards. Even if we recommend using simulation to estimate standard errors, we should not recommend averaging simulations to obtain the point estimate. Instead, researchers should directly transform maximum likelihood estimates of coefficients to obtain maximum likelihood estimates of the quantities of interest. The resulting point estimates inherit the desirable properties of maximum likelihood estimators and avoid unnecessary simulation-induced bias.

References

Altman, Micah, and Michael P McDonald. 2003. "Replication with Attention to Numerical Accuracy." *Political Analysis* 11(3):302–307.

Berry, William D., Jacqueline H. R. DeMeritt, and Justin Esarey. 2010. "Testing for Interaction in Binary Logit and Probit Models: Is a Product Term Essential." *American Journal of Political Science* 54(1):105–119.

Brambor, Thomas, William Roberts Clark, and Matt Golder. 2006. "Understanding interaction models: Improving empirical analyses." *Political analysis* 14(1):63–82.

Casella, George, and Roger L. Berger. 2002. *Statistical Inference*. 2nd ed. Pacific Grove, CA: Duxbury.

Choirat, Christine, James Honaker, Kosuke Imai, Gary King, and Olivia Lau. 2018. *Zelig: Everyone's Statistical Software*.

Efron, Bradley, and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*. New York: Chapman and Hall.

Efron, Bradley, and Trevor Hastie. 2021. *Computer Age Statistical Inference, Student Edition: Algorithms, Evidence, and Data Science*. Vol. 6 Cambridge University Press.

Gelman, Andrew, and Jennifer Hill. 2006. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Greene, William H. 2012. *Econometric Analysis*. 7th ed. Upper Saddle River, New Jersey: Prentice Hall.

Hanmer, Michael J., and Kerem Ozan Kalkan. 2013. "Behind the Curve: Clarifying the Best Approach to Calculating Predicted Probabilities and Marginal Effects from Limited Dependent Variable Models." *American Journal of Political Science* 57(1):263–277.

Herron, Michael C. 1999. "Postestimation Uncertainty in Limited Dependent Variable Models." *Political Analysis* 8(1):83–98.

Holland, Alisha C. 2015. "The Distributive Politics of Enforcement." *American Journal of Political Science* 59(2):357–371.

Imai, Kosuke, Gary King, and Olivia Lau. 2008. "Toward a common framework for statistical analysis and development." *Journal of Computational and Graphical Statistics* 17(4):892–913.

Jung, Yoo Sun, Flávio DS Souza, Andrew Q Philips, Amanda Rutherford, and Guy D Whitten. 2020. "A command to estimate and interpret models of dynamic compositional dependent variables: New features for dynsimpie." *The Stata Journal* 20(3):584–603.

King, Gary. 1998. *Unifying Political Methodology: The Likelihood Theory of Statistical Inference.* Ann Arbor: Michigan University Press.

King, Gary, and Langche Zeng. 2002. "Estimating risk and rate levels, ratios and differences in case-control studies." *Statistics in medicine* 21(10):1409–1427.

King, Gary, and Langche Zeng. 2006. "The dangers of extreme counterfactuals." *Political analysis* 14(2):131–159.

King, Gary, Michael Tomz, and Jason Wittenberg. 2000. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation." *American Journal of Political Science* 44(2):341–355.

Leeper, Thomas J. 2018. *margins: Marginal Effects for Model Objects.*

Nagler, Jonathan. 1991. "The Effect of Registration Laws and Education on US Voter Turnout." *The American Political Science Review* 85(4):1393–1405.

Nagler, Jonathan. 1994. "Scobit: An Alternative Estimator to Logit and Probit." *American Journal of Political Science* 38(1):230–255.

Philips, Andrew Q, Amanda Rutherford, and Guy D Whitten. 2016*a*. "Dynamic pie: A strategy for modeling trade-offs in compositional variables over time." *American Journal of Political Science* 60(1):268–283.

Philips, Andrew Q, Amanda Rutherford, and Guy D Whitten. 2016*b*. "dynsimpie: A command to examine dynamic compositional dependent variables." *The Stata Journal* 16(3):662–677.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing.

Rainey, Carlisle. 2016. "Compression and Conditional Effects." *Political Science Research and Methods* .

Rainey, Carlisle. 2017. "Transformation-Induced Bias: Unbiased Coefficients Do Not Imply Unbiased Quantities of Interest." *Political Analysis* 25:402–409.

Rainey, Carlisle, and Kelly McCaskey. 2021. "Estimating logit models with small samples." *Political Science Research and Methods* 9(3):549–564.

StataCorp. 2017. *Stata 15 Base Reference Manual*. College Station, TX: Stata Press.

Tomz, Michael, Jason Wittenberg, and Gary King. 2003. "Clarify: Software for Interpreting and Presenting Statistical Results." *Journal of Statistical Software* 8(1).

Tomz, Michael, Joshua A Tucker, and Jason Wittenberg. 2002. "An easy and accurate regression model for multiparty electoral data." *Political Analysis* 10(1):66–83.

Van der Vaart, Aad W. 2000. *Asymptotic Statistics*. Vol. 3 Cambridge university press.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with S*. Fourth ed. New York: Springer.

Wolfinger, Raymond E., and Steven J. Rosenstone. 1980. *Who Votes?* Yale University Press.

Zhirnov, Andrei, Mert Moral, and Evgeny Sedashov. 2022*a*. "DAME: data-driven interpretation of interactive nonlinear models." https://github.com/andreizhirnov/DAME.

Zhirnov, Andrei, Mert Moral, and Evgeny Sedashov. 2022*b*. "How to use Stata and R code to calculate distribution-weighted average marginal effects (DAME)." https://github.com/andreizhirnov/data-conscious-marginal-effects.

Zhirnov, Andrei, Mert Moral, and Evgeny Sedashov. 2022*c*. "Taking Distributions Seriously: On the Interpretation of the Estimates of Interactive Nonlinear Models." *Political Analysis* pp. 1–22.

Appendix for: "A Careful Consideration of CLARIFY"

A Proofs

A.1 Proof of Lemma 1

Proof By definition,

$$\hat{\tau}^{\text{avg}} = \mathrm{E}\left[\tau\left(\tilde{\beta}\right)\right].$$

Using Jensen's inequality (Casella and Berger 2002, p. 190, Thm. 4.7.7), $\mathrm{E}\left[\tau\left(\tilde{\beta}\right)\right] > \tau\left[\mathrm{E}\left(\tilde{\beta}\right)\right]$, so that

$$\hat{\tau}^{\text{avg}} > \tau\left[\mathrm{E}\left(\tilde{\beta}\right)\right].$$

However, because $\tilde{\beta} \sim \text{MVN}\left[\hat{\beta}^{\text{mle}}, \hat{V}\left(\hat{\beta}^{\text{mle}}\right)\right]$, $\mathrm{E}\left(\tilde{\beta}\right) = \hat{\beta}^{\text{mle}}$, so that

$$\hat{\tau}^{\text{avg}} > \tau\left(\hat{\beta}^{\text{mle}}\right).$$

Of course, $\hat{\tau}^{\text{mle}} = \tau\left(\hat{\beta}^{\text{mle}}\right)$ by definition, so that

$$\hat{\tau}^{\text{avg}} > \hat{\tau}^{\text{mle}}.$$

The proof for concave τ follows similarly. ■

A.2 Proof of Theorem 1

Proof According to Theorem 1 of Rainey (2017, p. 405), $\mathrm{E}\left(\hat{\tau}^{\text{mle}}\right) - \tau\left[\mathrm{E}\left(\hat{\beta}^{\text{mle}}\right)\right] > 0$. Lemma 1 shows that for any convex τ, $\hat{\tau}^{\text{avg}} > \hat{\tau}^{\text{mle}}$. It follows that $\underbrace{\mathrm{E}\left(\hat{\tau}^{\text{avg}}\right) - \tau\left[\mathrm{E}\left(\hat{\beta}^{\text{mle}}\right)\right]}_{\text{s.i. and t.i. } \tau\text{-bias in } \hat{\tau}^{\text{avg}}} > \underbrace{\mathrm{E}\left(\hat{\tau}^{\text{mle}}\right) - \tau\left[\mathrm{E}\left(\hat{\beta}^{\text{mle}}\right)\right]}_{\text{t.i. } \tau\text{-bias in } \hat{\tau}^{\text{mle}}} >$ 0.

For the concave case, it follows similarly that $\underbrace{\mathrm{E}\left(\hat{\tau}^{\text{avg}}\right) - \tau\left[\mathrm{E}\left(\hat{\beta}^{\text{mle}}\right)\right]}_{\text{s.i. and t.i. } \tau\text{-bias in } \hat{\tau}^{\text{avg}}} < \underbrace{\mathrm{E}\left(\hat{\tau}^{\text{mle}}\right) - \tau\left[\mathrm{E}\left(\hat{\beta}^{\text{mle}}\right)\right]}_{\text{t.i. } \tau\text{-bias in } \hat{\tau}^{\text{mle}}} <$ 0. ■

B Additional Analysis of the Drastic, Convex Transformation

In the main text, I develop an intuition for the simulation-induced τ-bias in $\hat{\tau}^{\text{avg}}$ using the simple (unrealistic, but heuristically useful) scenario in which $y_i \sim \text{N}(0,1)$, for $i \in \{1,2,\ldots,n=100\}$, and the researcher wishes to estimate μ^2. Suppose that the researcher knows that the variance equals one but does not know that the mean μ equals zero. The researcher uses the unbiased ML estimator $\hat{\mu}^{\text{mle}} = \frac{\sum_{i=1}^{n} y_i}{n}$ of μ, but ultimately cares about the quantity of interest $\tau(\mu) = \mu^2$. The researcher can use the plug-in estimator $\hat{\tau}^{\text{mle}} = \left(\hat{\mu}^{\text{mle}}\right)^2$ of $\tau(\mu)$. Alternatively, the researcher can use the average-of-simulations estimator, estimating $\tau(\mu)$ as $\hat{\tau}^{\text{avg}} = \frac{1}{M}\sum_{i=1}^{M} \tau\left(\tilde{\mu}^{(i)}\right)$, where $\tilde{\mu}^{(i)} \sim \text{N}\left(\hat{\mu}^{\text{mle}}, \frac{1}{\sqrt{n}}\right)$ for $i \in \{1,2,\ldots,M\}$.

Below, I calculate the bias of each estimator.

B.1 The Bias in the ML Estimator

To simplify the notation below, I use $\hat{\mu}$ in place of $\hat{\mu}^{\text{mle}}$.

First, note that $\hat{\mu} = \frac{\sum_{i=1}^{n} y_i}{n}$ is an *unbiased* estimator so that $\text{E}(\hat{\mu}) = \mu = 0$. We then have the common identity for mean-squared error: $\text{E}\left((\hat{\mu}-\mu)^2\right) = \text{Var}\left(\hat{\mu}\right) - \text{E}(\hat{\mu}-\mu)^2$. Substituting $\mu = 0$, we have $\text{E}\left(\hat{\mu}^2\right) = \text{Var}\left(\hat{\mu}\right) - \text{E}(\hat{\mu})^2$. Substituting $\text{E}(\hat{\mu}) = \mu = 0$, we have $\text{E}\left(\hat{\mu}^2\right) = \text{Var}\left(\hat{\mu}\right)$. Then $\text{E}\left(\hat{\mu}^2\right) = \text{Var}\left(\frac{\sum_{i=1}^{n} y_i}{n}\right) = \frac{1}{n^2}\text{Var}\left(\sum_{i=1}^{n} y_i\right)$. Then, using the identify that the variance of the sum of independent random variables is the sum of their variances, we have $\text{E}\left(\hat{\mu}^2\right) = \frac{1}{n^2}(n \times 1) = \frac{1}{n}$.

Since $\tau = \mu^2 = 0$, the bias in $\hat{\tau} = \left[\hat{\mu}^{\text{mle}}\right]^2$ is $\frac{1}{n} - 0 = \frac{1}{n}$. Because there is no coefficient-induced bias, this is also the transformation-induced bias.

B.2 The Bias in the Average-of-Simulations Estimator

To simplify the notation below, I use $\bar{\tau}$ in place of $\hat{\tau}^{\text{avg}}$.

First, compute $\text{E}(\bar{\tau} \mid \hat{\mu}) = \text{E}\left[\frac{1}{M}\sum_{i=1}^{M}\left(\tilde{\mu}^{(i)}\right)^2\right] = \frac{1}{M}\sum_{i=1}^{M}\text{E}\left[\left(\tilde{\mu}^{(i)}\right)^2\right]$. Then we have $\text{E}(\bar{\tau} \mid \hat{\mu}) = \frac{1}{M}\sum_{i=1}^{M}\left[\text{Var}(\tilde{\mu}^{(i)}) + \text{E}\left(\tilde{\mu}^{(i)}\right)^2\right]$. Substituting known values, we have $\text{E}(\bar{\tau} \mid \hat{\mu}) = \frac{1}{M}\sum_{i=1}^{M}\left[\frac{1}{n} + \hat{\mu}^2\right]$. Simplifying, we have $\text{E}(\bar{\tau} \mid \hat{\mu}) = \frac{1}{M}\left[\frac{M}{n} + M\hat{\mu}^2\right] = \frac{1}{n} + \hat{\mu}^2$.

Next, apply the law of iterated expectations to find $\text{E}(\bar{\tau}) = \text{E}((\bar{\tau} \mid \hat{\mu}))$. Substituting, we have $\text{E}(\bar{\tau}) = \text{E}(\frac{1}{n} + \hat{\mu}^2)$. Then, simplifying, we have $\text{E}(\bar{\tau}) = \frac{1}{n} + \text{E}(\hat{\mu}^2) = \frac{1}{n} + \frac{1}{n} = \frac{2}{n}$.

The bias in $\hat{\tau}^{\text{avg}}$ is therefore $\frac{2}{n}$. Because simulation-induced bias is defined as $\text{E}\left(\hat{\tau}^{\text{avg}}\right) - \text{E}\left(\hat{\tau}^{\text{mle}}\right)$, the simulation-induced bias in this example is $\frac{2}{n} - \frac{1}{n} = \frac{1}{n}$. Thus, the simulation-induced and transformation-induced bias in this example are exactly equal and the average-of-simulations estimator exactly doubles the bias in the ML estimator.

C Real-World Relevance of Simulation-Induced Bias

To get a sense of the"real world" magnitude and importance of simulation-induced bias, I conduct a Monte Carlo study on estimates and data from Berry, DeMeritt, and Esarey (2010). This paper serves as an ideal example because it is highly cited, uses a large dataset, and builds on a familiar series of substantive and methodological papers, including Wolfinger and Rosenstone (1980), Nagler (1991, 1994), and Altman and McDonald (2003). For the data-generating process, I use the model specification and coefficients that Berry, DeMeritt, and Esarey (2010) report in their Table 1, Column 2. To generate suitably-small samples of explanatory variables, I randomly draw a sample of 100 observations and subsequently add in 100 more, then 200 more, and then 400 more observations to create samples of 100, 200, 400, and 800 observations from the original data set from Berry, DeMeritt, and Esarey (2010).

In the Monte Carlo simulations, I keep the explanatory variables fixed and use the coefficient estimates from Berry, DeMeritt, and Esarey (2010) to simulate 10,000 outcome variables. For each simulated outcome variable, I compute $\hat{\beta}$, $\hat{\tau}^{\text{mle}}$, and $\hat{\tau}^{\text{avg}}$. Finally, I use these estimates to decompose the τ-bias into its coefficient-induced, transformation-induced, and simulation-induced components.

As the quantity of interest, I calculate the change in the probability of turning out to vote if the election registration deadline occurs 10 days sooner (a one standard deviation change). However, the effect of this 10-day shift (and the coefficient-, transformation-, and simulation-induced τ-bias) depends on the original closing date as well as the values of the other explanatory variables. To capture and present this heterogeneity, I estimate the coefficient-, transformation-, and simulation-induced biases for 100 randomly-chosen observations from the full data set.

Figure 4 shows the resulting coefficient-, transformation-, and simulation-induced τ-bias (columns) for the four different sample sizes (rows). Each point shows the bias for one of the 100 observations. Points that fall outside the shaded region have a bias greater than 10%, which I consider substantively meaningful.

The simulation-induced τ-bias (right-hand column) has a magnitude comparable to the well-known small sample bias in logistic regression coefficients (left-hand column, Rainey and McCaskey 2021). The magnitude of the simulation-induced bias is similar to—or perhaps slightly larger than—the magnitude of the transformation-induced bias (middle column), in line with the rough approximation developed in the main text. Second, at least in some scenarios, the simulation-induced bias is sufficiently large to meaningfully affect results. For at least some observations, the simulation-induced bias remains larger than 10% until the sample size exceeds 400. Third, while each type of bias disappears asymptotically, the biases disappear at different rates. Rainey (2017) shows that transformation-induced bias can disappear more slowly than coefficient-induced bias. The same pattern appears for simulation-induced bias. Especially for certain observations, the simulation-induced bias remains large even when the coefficient-induced bias has nearly disappeared.

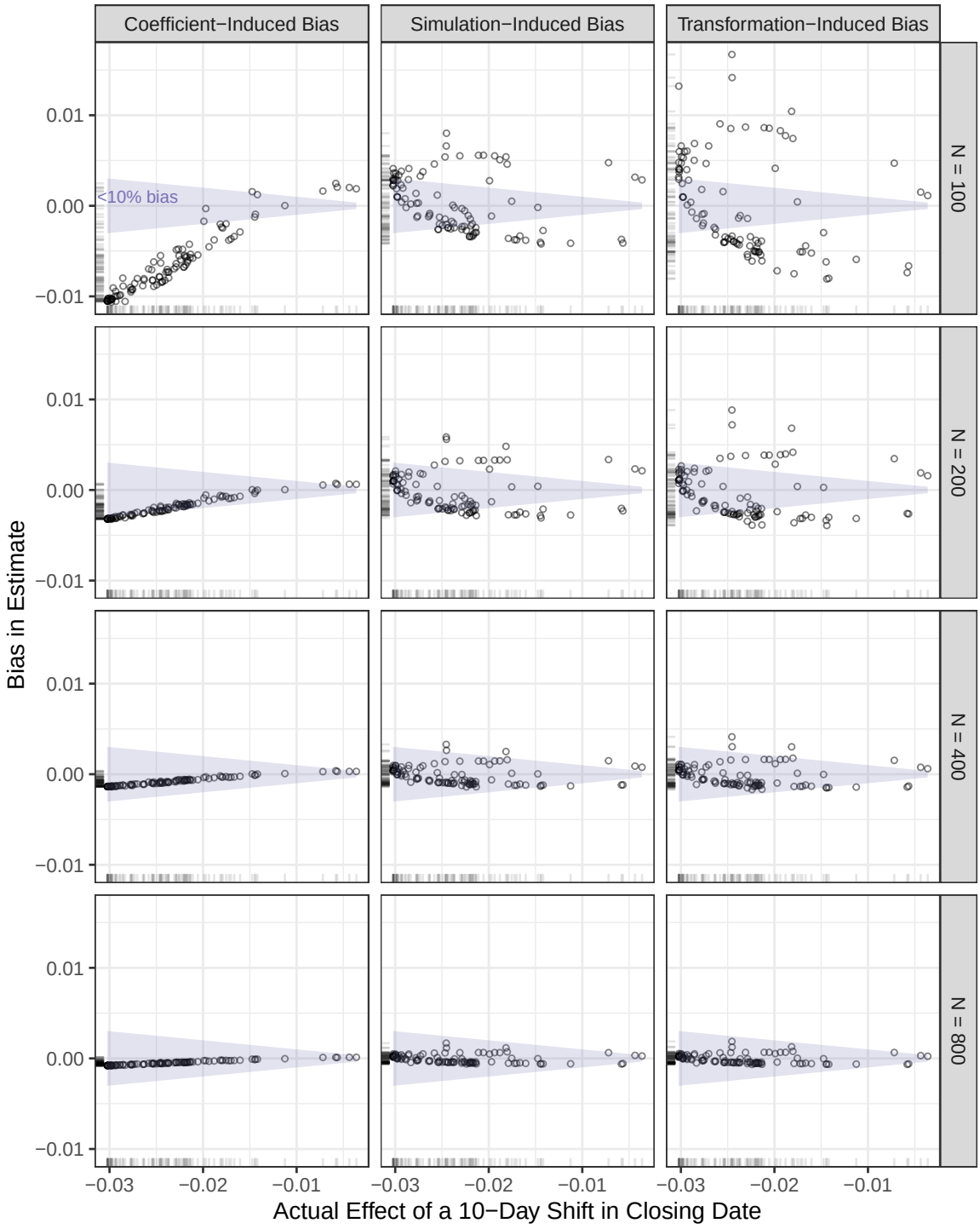

Figure 4: The figure shows the coefficient-induced, transformation-induced, and simulation-induced τ-biases for a logistic regression model based on Berry, DeMeritt, and Esarey (2010). The points that fall outside the shaded region have bias greater than 10%.