Simulation-Induced Bias in Quantities of Interest[*]

Carlisle Rainey[†]

Holger L. Kern[‡]

August 7, 2018

Abstract

Following **?** researchers commonly obtain point estimates of quantities of interest by simulating model coefficients, transforming these simulated coefficients into simulated quantities of interest, and then taking the average of the simulated quantities of interest. In contract, other researchers directly transform coefficient estimates into quantities of interest, using the invariance property of maximum likelihood estimators. These approaches are not equivalent. We show both analytically and empirically that computing quantities of interest using the average of simulations can introduce substantial bias. Instead, researchers should use the invariance property to calculate the maximum likelihood estimates of their quantities of interest.

Political scientists employ maximum likelihood (ML) estimators to model a wide variety of dependent variables. Examples include logit and probit models for binary outcomes; ordered logit and probit models for ordered categorical outcomes; multinomial logit and probit models for unordered categorical outcomes; Poisson and negative binomial regression for count data; and beta regression for fractions. Political scientists propose several ML estimators and researchers regularly use these estimators in political science research (e.g., **?**, **?**; **?**). Many of these models have coefficients that do not directly interest substantive researchers. Instead, researchers usually need to transform the coefficients to compute substantively-meaningful quantities of interest.

? dramatically improves the reporting of quantitative research by urging researchers to focus on substantively meaningful quantities of interest such as predicted probabilities, expected counts, marginal effects, and first differences. Prior to the publication of **?** and the availability of easy-to-use software for Stata (CLARIFY) and R (Zelig), political science research commonly reports lengthy tables of coefficient estimates and pays little or no attention to the substantive interpretation of the estimates beyond their sign and statistical significance.

---

[*]All computer code necessary for replication is available on GitHub.

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

[‡]Holger L. Kern is Assistant Professor of Political Science, Florida State University, 541 Bellamy, Tallahassee, FL, 32306. (hkern@fsu.edu).

The citations demonstrate the impact of King, Tomz, and Wittenberg (2000) on the literature. According to the Web of Science, by January 2018 **?** has received 1,097 citations, making it the fourth most-cited methodology article in political science, at least among those published in the *American Political Science Review*, the *American Journal of Political Science*, and *Political Analysis*. Moreover, their paper has received the second most-citations overall in the *American Journal of Political Science*, falling just 56 citations short of **?**. Google Scholar lists 3,598 citations for **?**, 1,437 citations for **?** (which provides an overview of the CLARIFY software), and 318 citations for Imai, King, and Lau (2008) (which proposes a common framework for statistical analysis and software development centered on Zelig). In short, the advice that **?** offers noticeably improves research in political science and the social sciences more generally.

? also popularize a simulation-based approach to compute quantities of interest. They suggest that researchers simulate model coefficients, transform these simulated coefficients into simulated quantities of interest, and finally summarizes the distribution of the simulated quantities of interest.

Our paper addresses one aspect of **?**'s advice: to use the average of the simulated quantities of interest as the point estimate. We refer to this estimate as the "simulation average estimate" of the quantity of interest. We show both analytically and empirically that the simulation average exaggerates the transformation-induced bias that **?** describes. Instead, we propose that researchers compute ML estimates of quantities of interest using the invariance property of ML estimators.

The invariance property of ML estimators allows a reseacher to find the ML estimate of a function of a parameter by first using ML to estimate the model parameter and then applying the function to that estimate (**?**, pp. 75–76, and **?**, pp. 320–321). More formally, 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 coefficients with length $k + 1$. The researcher uses ML to compute estimates $\hat{\beta}^{\text{mle}}$ for the parameter vector β. We denote to the function that transforms model coefficients into quantities of interest as $\tau(\cdot)$. For example, if researcher uses a logit model and focuses on a 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: $\hat{\tau}^{\text{mle}} = \tau\left(\hat{\beta}^{\text{mle}}\right) = \text{logit}^{-1}\left(X_c\hat{\beta}^{\text{mle}}\right) = \dfrac{1}{1 + e^{-X_c\hat{\beta}^{\text{mle}}}}$.

Software implementations differ in whether they rely on the invariance property or **?**'s simulation-based approach. Some commonly used software, such as margins in Stata, uses the invariance property. Other software, such as CLARIFY for Stata and Zelig for R, reports the simulation average estimate.

Similarly, the methodological literature presented conflicting advice. **?** uses the invariance property to derive an ML estimator for predicted probabilities in limited dependent variable models (and then uses stochastic simulation to compute measures of uncertainty). Even though **?** cites

2

an earlier version of **?**, it does not mention the fact that its approach differs from **?**. Carsey and Harden (2013) follow **?** and use the simulation average estimate. We do not know of any work that compares the properties of these two point estimators. Indeed, it seems that the literature incorrectly treats both approaches as essentially equivalent. However, as we show, the ML estimate has a distinct advantage over the simulation average estimate.

Transformation-Induced τ-Bias

As Rainey (2017) shows, using the invariance property to transform ounbiased model coefficient estimates can introduce bias into the estimates of the quantities of interest.[1] **?**, p. 404 decomposes the bias in the estimate of the quantity of interest–the total τ-bias–into two components: transformation-induced τ-bias and coefficient-induced τ-bias. Rainey defines these as

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

The direction and magnitude of the coefficient-induced τ-bias depends 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. Here, we do not consider coefficient-induced τ-bias any further.

Instead, we focus on transformation-induced τ-bias. Its sign can be predicted based on the shape of the transformation that converts coefficient estimates into an estimate of the quantity of interest. In general, any strictly convex (concave) $\tau(\cdot)$ creates upward (downward) transformation-induced τ-bias.

The Average of Simulations

Rather than rely on the invariance property of ML estimators to compute a point estimate for the quantity of interest, **?** suggests the following simulation-based 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

---

[1]If the coefficient estimates are themselves biased, the transformation-induced bias can, but generally does not, offset the bias in the coefficient estimates.

"average case") or average across all sample observations (**?**).[2] In any case, the transformation $\tau(\cdot)$ includes this choice.[3]

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 $\hat{\tau}^{\mathrm{avg}} = \frac{1}{M} \sum_{i=1}^{M} \tilde{\tau}^{(i)}$.[4]

The Average of Simulations Versus the Maximum Likelihood Estimate

As discussed above, the literature offers two ways to estimate quantities of interest: the estimate $\hat{\tau}^{\mathrm{mle}}$ calculated using the invariance property of ML estimators and the average simulation estimate $\hat{\tau}^{\mathrm{avg}}$ calculated using the algorithm described in **?**. How does $\hat{\tau}^{\mathrm{avg}}$ compare to $\hat{\tau}^{\mathrm{mle}}$?

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

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

Proof By definition,

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

Using Jensen's inequality (**?**, p. 190, Thm. 4.7.7), we know that $\mathrm{E}\left[\tau\left(\tilde{\beta}\right)\right] > \tau\left[\mathrm{E}\left(\tilde{\beta}\right)\right]$, so that

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

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

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

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

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

The proof for concave τ follows similarly. ∎

---

[2]We return to **?** in a later section of our paper.

[3]As **?** note, this step might require additional simulation, to first introduce and then average over fundamental uncertainty. We ignore this additional step since it does not affect our argument.

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

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

For a convex transformation, Lemma 1 shows that $\hat{\tau}^{\mathrm{avg}}$ is always larger than $\hat{\tau}^{\mathrm{mle}}$. But does this imply that $\hat{\tau}^{\mathrm{avg}}$ is *more biased* than $\hat{\tau}^{\mathrm{mle}}$? Theorem 1 shows that this is indeed the case.

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

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

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

$0.$ ∎

Regardless of whether the transformation-induced τ-bias is positive or negative, Theorem 1 shows that the magnitude of the bias is *always* larger for $\hat{\tau}^{\mathrm{avg}}$ than for $\hat{\tau}^{\mathrm{mle}}$ for strictly convex or concave $\tau(\cdot)$. We refer to the additional transformation-induced bias in $\hat{\tau}^{\mathrm{avg}}$ compared to $\hat{\tau}^{\mathrm{mle}}$ as *simulation-induced τ-bias,* so that

$$
\text{simulation-induced } \tau\text{-bias} = \underbrace{\left(\mathrm{E}\left(\hat{\tau}^{\mathrm{avg}}\right) - \tau\left[\mathrm{E}\left(\hat{\beta}^{\mathrm{mle}}\right)\right]\right)}_{\text{t.i. } \tau\text{-bias in } \hat{\tau}^{\mathrm{avg}}} - \underbrace{\left(\mathrm{E}\left(\hat{\tau}^{\mathrm{mle}}\right) - \tau\left[\mathrm{E}\left(\hat{\beta}^{\mathrm{mle}}\right)\right]\right)}_{\text{t.i. } \tau\text{-bias in } \hat{\tau}^{\mathrm{mle}}}
$$
$$
= \mathrm{E}\left(\hat{\tau}^{\mathrm{avg}}\right) - \mathrm{E}\left(\hat{\tau}^{\mathrm{mle}}\right).
$$

An Approximation for the Simulation-Induced Bias in $\hat{\tau}^{\mathrm{avg}}$

Theorem 1 guarantees that $\hat{\tau}^{\mathrm{avg}}$ is more biased than $\hat{\tau}^{\mathrm{mle}}$. But it this bias trivial or substantively important? Monte Carlo experiments allow one to assess this directly, but an analytical approximation provides a helpful rule of thumb. We approximate the simulation-induced τ-bias in $\hat{\tau}^{\mathrm{avg}}$

as

$$\text{simulation-induced } \tau\text{-bias in } \hat{\tau}^{\text{avg}} = \underbrace{\left(\text{E}\left(\hat{\tau}^{\text{avg}} \right) - \tau \left[\text{E}\left(\hat{\beta}^{\text{mle}} \right) \right] \right)}_{\text{t.i. } \tau\text{-bias in } \hat{\tau}^{\text{avg}}} - \underbrace{\left(\text{E}\left(\hat{\tau}^{\text{mle}} \right) - \tau \left[\text{E}\left(\hat{\beta}^{\text{mle}} \right) \right] \right)}_{\text{t.i. } \tau\text{-bias in } \hat{\tau}^{\text{mle}}}$$

$$= \text{E}\left(\hat{\tau}^{\text{avg}} \right) - \text{E}\left(\hat{\tau}^{\text{mle}} \right)$$

$$= \text{E}\left(\hat{\tau}^{\text{avg}} - \hat{\tau}^{\text{mle}} \right)$$

$$= \text{E}\left(\text{E}\left[\tau\left(\tilde{\beta} \right) \right] - \tau\left(\hat{\beta}^{\text{mle}} \right) \right)$$

$$= \text{E}\left(\underbrace{\text{E}\left[\tau\left(\tilde{\beta} \right) \right] - \tau\left[\text{E}\left(\tilde{\beta} \right) \right]}_{\substack{\text{approximated in Eq. 1,} \\ \text{p. 405, of ?}}} \right)$$

$$\approx \text{E}\left[\frac{1}{2} \sum_{r=1}^{k+1} \sum_{s=1}^{k+1} H_{rs}\left(\hat{\beta}^{\text{mle}} \right) \hat{V}_{rs}\left(\hat{\beta}^{\text{mle}} \right) \right], \tag{2}$$

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 τ 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 is similar to the approximation for the transformation-induced τ-bias for $\hat{\beta}^{\text{mle}}$, which, adjusting notation slightly, **?**, p. 405, Eq. 1 computes as

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

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

When we compare Equations 2 and 3, we yet again compare the *average of a function* with the *function of that average*. Therefore, Equations 2 and 3 are not exactly equal. But, as a rule of thumb, we should expect them to be similar. And to the extent that this is the case, the simulation-induced τ-bias in $\hat{\tau}^{\text{avg}}$ is about the same as the transformation-induced τ-bias in $\hat{\tau}^{\text{mle}}$. This implies that the τ-bias in $\hat{\tau}^{\text{avg}}$ will be about *twice* as large as the 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 **?** 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

(**?**). And remember that the estimate $\hat{\beta}^{\text{avg}}$ includes both the transformation-induced bias and the similarly-sized simulation-induced bias.

The Intuition

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 scenario in which $y_i \sim N(0, 1)$, 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 maximum likelihood estimator $\hat{\mu}^{\text{mle}} = n^{-1} \sum_{i=1}^{n} y_i$ of μ, but ultimately cares about the quantity of interest $\tau(\mu) = \mu^2$. The researcher can use the invariance property to compute the ML estimate $\tau(\mu)$ as $\hat{\tau}^{\text{mle}} = \left(\hat{\mu}^{\text{mle}} \right)^2$. Alternatively, the researcher can use the simulation-based approach, estimating $\tau(\mu)$ as $\hat{\tau}^{\text{avg}} = \frac{1}{M} \sum_{i=1}^{M} \tau \left(\tilde{\mu}^{(i)} \right)$, where $\tilde{\mu}^{(i)} \sim N \left(\hat{\mu}^{\text{mle}}, \frac{1}{\sqrt{n}} \right)$ for $i \in \{1, 2, \ldots, M\}$.

The true value of the quantity of interest is $\tau(0) = 0^2 = 0$. However, the maximum likelihood 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 $\hat{\mu}^{\text{mle}}$ is almost surely different from zero, it is clear that $\hat{\tau}^{\text{mle}}$ is biased upward. Moreover, even if $\hat{\mu}^{\text{mle}} = 0$, $\tilde{\mu}^{(i)}$ almost surely does not equal zero. If $\tilde{\mu}^{(i)} \neq 0$, then $\left(\tilde{\mu}^{(i)} \right)^2 > 0$. Thus, $\hat{\mu}^{\text{avg}}$ is larger than the true value $\tau(\mu) = 0$ with probability one.

We can see this fact clearly by repeatedly simulating y and estimating $\hat{\tau}^{\text{mle}}$ and $\hat{\tau}^{\text{avg}}$. Figures 1a–1d show the first four of 1,000 total simulations. The figures show how the unbiased estimate $\hat{\mu}^{\text{mle}}$ is translated into $\hat{\tau}^{\text{mle}}$ and $\hat{\tau}^{\text{avg}}$.

First, to find $\hat{\tau}^{\text{avg}}$, we complete three steps: (1) simulate $\tilde{\mu}^{(i)} \sim N \left(\hat{\mu}^{\text{mle}}, \frac{1}{\sqrt{n}} \right)$ 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 and the density plot at the top of each plot show the distribution of $\tilde{\mu}$. The hollow points in Figures 1a–1d show the transformation of each point $\tilde{\mu}^{(i)}$ into $\tilde{\tau}^{(i)}$. The rug plot along the vertical axis and the density plot to the right of each plot show the distribution of $\tilde{\tau}$. Focus on Figure 1a. 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 purple 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 $\hat{\tau}^{\text{mle}}$, we simply 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. Notice that $\hat{\tau}^{\text{mle}}$ corresponds approximately to the mode of the density plot of $\tilde{\tau}$ along the right side of the plot, which falls closer to the true value $\tau(0) = 0$ than $\hat{\tau}^{\text{avg}}$. 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.

Figures 1b–1d repeat this process with three more random samples. In each sample, the story

(a) Simulation 1 of 1,000

(b) Simulation 2 of 1,000

(c) Simulation 3 of 1,000

(d) Simulation 4 of 1,000

Figure 1: These four figures illustrate the relationship between $\hat{\tau}^{\mathrm{mle}}$ and $\hat{\tau}^{\mathrm{avg}}$ described by Lemma 1 and Theorem 1.

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

We repeat this process 1,000 times to produce 1,000 estimates $\hat{\mu}^{\text{mle}}$, $\hat{\tau}^{\text{mle}}$, and $\hat{\tau}^{\text{avg}}$. Figure 2 shows the density plots for the 1,000 estimates (i.e., the sampling distributions of $\hat{\mu}^{\text{mle}}$, $\hat{\tau}^{\text{mle}}$, and $\hat{\tau}^{\text{avg}}$). As we know analytically, $\hat{\mu}^{\text{mle}}$ is unbiased with a standard error of $\frac{1}{\sqrt{n}} = \frac{1}{\sqrt{100}} = \frac{1}{10}$. Both $\hat{\tau}^{\text{mle}}$ and $\hat{\tau}^{\text{avg}}$ are biased upward, but $\hat{\tau}^{\text{avg}}$ is biased more. Theorem 1 shows why this must be the case.
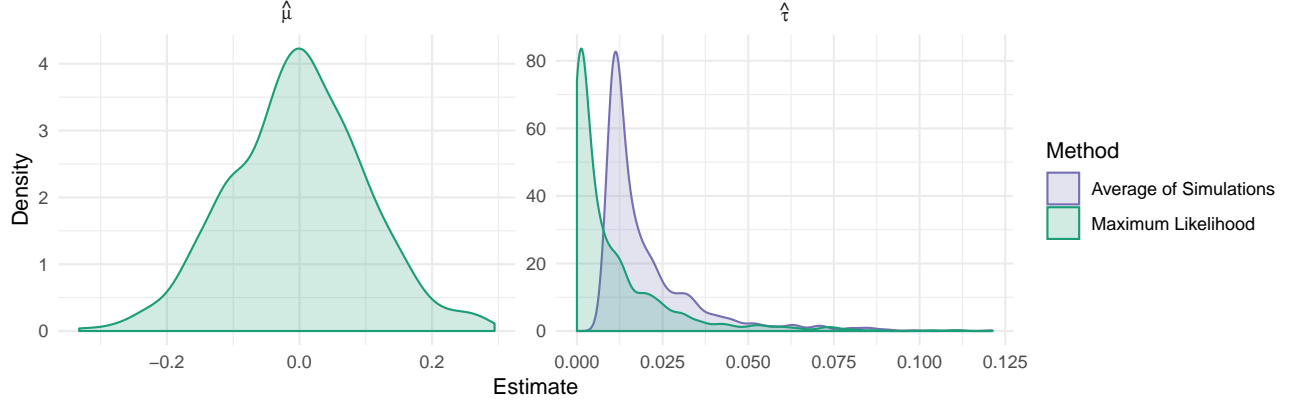
Figure 2: The sampling distributions of $\hat{\beta}^{\text{mle}}$, $\hat{\tau}^{\text{mle}}$, and $\hat{\tau}^{\text{avg}}$.

Using the Law of Iterated Expectations

We can also develop the intuition behind our argument analytically via the law of iterated expectations. For this it helps to alter the notation slightly, making two implicit dependencies explicit. We 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 w.r.t. the marginal distribution of Y, $\mathrm{E}_{X|Y}$ denotes the expectation w.r.t. the conditional distribution of $X \mid Y$, and E_X denotes the expectation w.r.t. the marginal distribution of X.

Outside of this section, we realize 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 **?** and **?**, we simply 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, we use $\tilde{\beta} \mid \hat{\beta}^{\text{mle}}$ to represent the conditional distribution of $\tilde{\beta}$ and $\tilde{\beta}$ 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 we do steps (1) and (2) just once, but step (3) repeatedly, we have a sample from the

conditional distribution $\tilde{\beta} \mid \hat{\beta}^{\text{mle}}$. If we do steps (1), (2), and (3) repeatedly, then we have a sample from the <u>un</u>conditional distribution $\tilde{\beta}$. The unconditional distribution helps us understand the nature of the simulation-induced τ-bias.

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

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

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

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

If we subtract Equation 5 from Equation 4 we 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 we 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 $\hat{\tau}^{\text{avg}}$ we must subtract Equation 6 from Equation 4. But to move from Equation 4 to Equation 6 we 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 we expect $\hat{\beta}^{\text{mle}}$ and $\tilde{\beta} \mid \hat{\beta}^{\text{mle}}$ to have similar distributions, we should expect the additional swap to roughly double the bias in $\hat{\tau}^{\text{avg}}$ compared to $\hat{\tau}^{\text{avg}}$. It is this additional swap that leads to simulation-induced τ-bias.

Illustrations

We illustrate simulation-induced τ-bias in three ways: (1) by computing the coefficient-, transformation-, and simulation-induced τ-bias for a simple Poisson regression model, (2) by repeating that exercise for a more complex logistic regression model, and (3) by replicating a published article to show that different estimates result from using the invariance property and averaging simulation draws, respectively, to compute quantities of interest.

Marginal Effects in Poisson Regression

As a first illustration, consider the Poisson regression model $y_i \sim \text{Poisson}(\lambda_i)$, where $\lambda_i = e^{(-2+x_i)}$ for $i \in \{1, 2, \ldots, 100\}$. To create x_i, we take 100 i.i.d. draws from a standard normal distribution. Assume that the researcher wants to estimate the marginal effect of x on $\text{E}(y)$, so that $\tau(\beta) = \frac{d\,\text{E}(y)}{dx} = e^{(\beta_{cons}+\beta_x x)}$ for x ranging from -3 to $+3$.

We generate 10,000 data sets and use each data set to estimate $\hat{\tau}^{\text{mle}}$ and $\hat{\tau}^{\text{avg}}$. Note that the transformation is convex, so according to Theorem 1 the τ-bias in both $\hat{\tau}^{\text{mle}}$ and $\hat{\tau}^{\text{avg}}$ should be positive. The rule of thumb suggests about twice as much bias in $\hat{\tau}^{\text{avg}}$ as in $\hat{\tau}^{\text{mle}}$.

Figure 3: The figure shows bias in estimates of the marginal effect in a Poisson regression model. Note that the convex transformation $\tau(\beta) = \frac{d\,\text{E}(y)}{dx} = e^{(\beta_{cons}+\beta_x x)}$ creates a positive bias (see Theorem 1) and that the bias in $\hat{\tau}^{\text{avg}}$ is about twice as large as the bias in $\hat{\tau}^{\text{mle}}$ (compare Equations 2 and 3).

Figure 3 shows the τ-bias in $\hat{\tau}^{\text{avg}}$ and $\hat{\tau}^{\text{mle}}$ (vertical axis) compared to the true value $\tau(\beta)$ (horizontal axis). Notice three features of this plot. First, note that the bias occurs in the expected direction. Because the transformation $\tau(\beta) = \frac{d\,\text{E}(y)}{dx} = e^{(\beta_{cons}+\beta_x x)}$ is convex, the bias is positive. Second, the bias can be substantial, depending on the size of the true marginal effect. For example, when the true marginal effect equals 1 the average τ-bias in $\hat{\tau}^{\text{mle}}$ is about .25 while the τ-bias in $\hat{\tau}^{\text{avg}}$ is about .5. Third, note that over much of the range of the horizontal axis, the bias in $\hat{\tau}^{\text{avg}}$ is about twice as large as the bias in $\hat{\tau}^{\text{mle}}$, as the rule of thumb suggests.

First Differences in a Logistic Regression Model

As a more realistic example we compute the τ-bias for a more complex logistic regression model. To find a plausible data-generating process and realistic set of explanatory variables, we base our Monte Carlo study on the classic example of the decision to vote (**????**). In particular, Berry, DeMeritt, and Esarey (2010) uses a large data set with about 100,000 observations to estimate their models. We take borrow the coefficients and explanatory variables from Berry, DeMeritt, and Esarey (2010) to use as the data-generating process and explanatory variables, respective, in our Monte Carlo

study.

For the data-generating process, we use the model specification and coefficients that **?** reports in Table 1, Column 2. To generate the explanatory variables, we randomly draw samples of 100, 200, 400, and 800 observations from the data that **?** uses to fit their model. (Across the Monte Carlo simulations, we generate a new, random outcome variable using these explanatory variables and Berry, DeMeritt, and Esarey's coefficients.)

For the quantity of interest, we focus on 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 bias) depends on the original closing date as well as the values of the other explanatory variables. To capture and present this heterogeneity, we estimate the coefficient-, transformation-, and simulation-induced bias for 250 randomly-chosen observations from the full data.

Figure 4 shows the resulting coefficient-, transformation-, and simulation-induced τ-biases (columns) for the four different sample sizes (rows). Each point shows the bias for one of the 250 observations. (Remember that the true effect, as well as the size of each type of bias, varies across observations). First, note that the simulation-induced τ-bias (right-hand column) has a magnitude similar to the well-known small sample bias in logistic regression coefficients (left-hand column). The magnitude of the simulation-induced τ-bias is similar to the magnitude of the transformation-induced τ-bias (middle column), in line with the approximation we derive in a previous section. 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. We see the same pattern here. Especially for certain observations, the transformation-induced bias remains large even when the coefficient-induced bias has nearly disappeared. Consistent with the previous result that simulation-induced bias approximates the transformation-induced bias, we see that the simulation-induced bias disappears more slowly as well.

Replication

As a final illustration of simulation-induced bias, we replicate an existing study (**?**) and compare the ML estimates of quantities of interest (comuted using the invariance property) to simulation average estimate. We cannot demonstrate bias using this approach (since we do not know the true effects), but we can show that the estimates differ meaningfully.

To unify explanations of U.S. Supreme Court decisions, **?** fits a single probit model that combines the legal and extralegal models of Court decision-making to a data set of 64 decisions. The article models the probability of a conservative decision as a function of whether the Solicitor General filed an Amicus brief (SG = 1) or not (SG = 0) and 10 other explanatory variables. See **?**
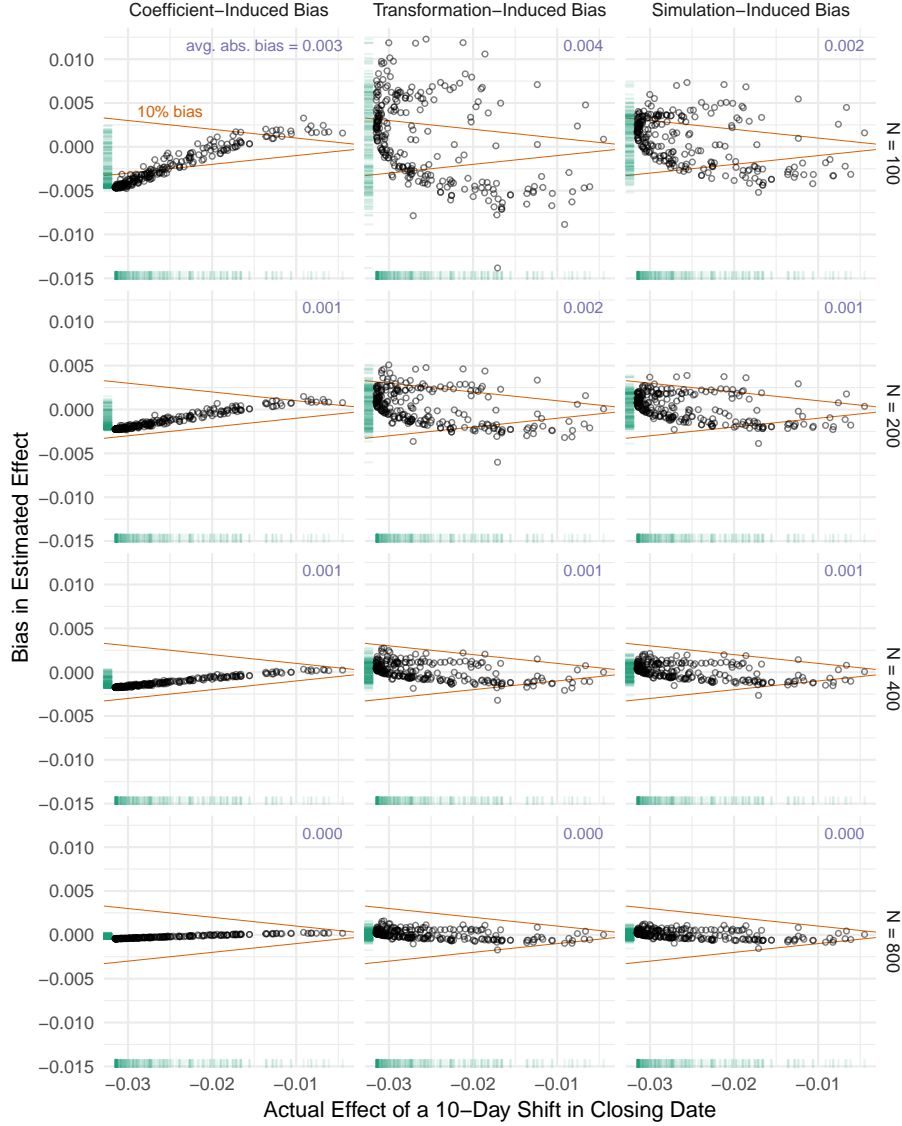
Figure 4: The figure shows the coefficient-induced, transformation-induced, and simulation-induced τ-biases for a logistic regression model based on **?**. The points the fall outside the orange lines have bias greater than 10%. The numbers at the top right corner of each plot report the average absolute bias across all 250 observations.

for details.[5]

We focus on two quantities of interest: the probability of a conservative decision and the effect of the Solicitor General filing a brief, i.e., the first difference in the probability of a conservative decision. We compute these two quantities of interest for each observation in the data set. Figure 5 compares the estimates. First, consider the estimates of the probability of a conservative decision in Figure 5a. The pattern is clear: when the fitted probability of a conservative decision is less than 50%, the simulation average estimate is larger than the ML estimate. In this region, the transformation (the normal cdf) is convex, leading to positive bias. When the fitted probability of a conservative decision is greater than 50%, the simulation average estimate is smaller than the ML estimate. In this region, the transformation is concave, causing the bias to be negative. When the fitted probability of a conservative decision is close to 50%, the differences between the two estimates are smaller since the transformation is close to linear in this area. The same is true for fitted probabilities close to 0% and 100%.

Second, note that some of the differences between the estimates are quite large. For example, when the ML estimates are around 5%, the simulation average estimate comes in at about 10%. This difference may seem small at first (i.e., only 5 percentage points), but for such observations $\hat{\tau}^{\text{avg}}$ is about twice as large as $\hat{\tau}^{\text{mle}}$.

Figure 5b displays estimates of the effect of the Solicitor General filing an Amicus brief. The largest differences between the two estimators appear in the upper-right corner of the plot. For this group of observations, the simulation average estimate suggests that a brief from the Solicitor General increases the probability of a conservative decision by about 40 percentage points. The ML estimate on the other hand suggests an increase of about 60 percentage points. This difference is certainly meaningful — the maximum likelihood estimate is 50% larger than the estimate based on stochastic simulation.
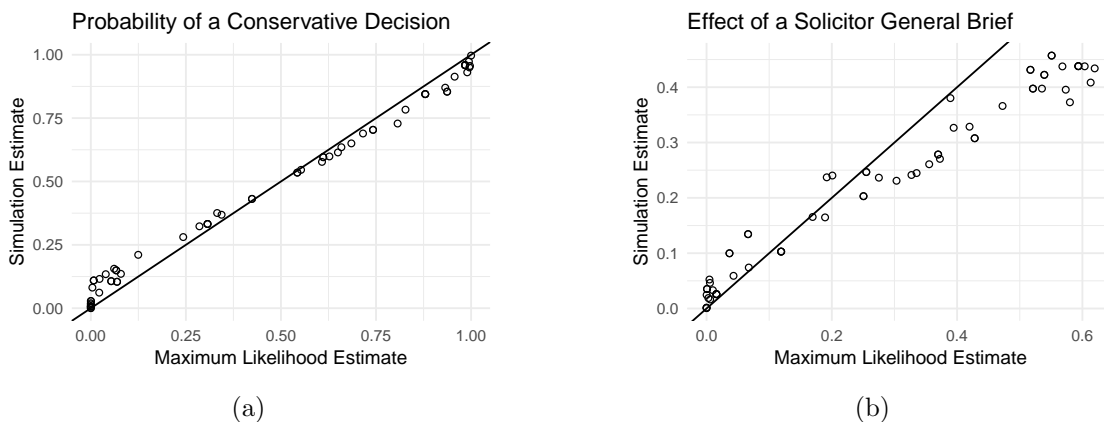

Figure 5: The figure shows the relationship between the simulation average and the maximum likelihood estimate for two quantities of interest. The left panel (a) shows the probability of a conservative decision. The right panel (b) shows the effect of a brief by the Solicitor General on the probability of a conservative decision.

A Note on ?

? suggests that researchers avoid the *average-case* approach to developing quantities of interest and instead use the and the *observed-value* approach. With either approach, researchers estimate the quantity of interest as a key explanatory variable changes its value. However, in non-linear models researchers must also deal with the other explanatory variables in the model, because these variables alter the quantity of interest. The average-case approach sets the other explanatory variables to central values such as the mean, median, or mode. ?, in contrast, suggests estimating the quantity of interest for all sample observations, leaving the explanatory variables except for the key variable of interest at their observed values, and then averaging the estimates across the sample. Here, our notation captures this choice as part of the transformation $\tau(\cdot)$, so Hanmer and Kalkan's (compelling) argument does not undermine or enhance our own.[6]

Because researchers have not drawn a sharp conceptual distinction between using the simulation average estimate and the ML estimate (computed using the invariance property of ML estimators), ? does not discuss this choice. Since it explicitly builds on ?, we interpret ? as relying on the simulation average estimate when computing quantities of interest. The replication archive for the article confirms that this is indeed the case.

The important point is this: ? draws a distinction between the average-case and observed-value approaches to computingquantities of interest. Our paper draws a distinction between computing estimates of quantities of interest (whether average-case or observed-value based) using invariance properity of ML estimators or using King, Tomz, and Wittenberg's (2000) simulation-based approach. Regardless of whether researchers use the average-case approach or the observed-value approach, the simulation-based approach leads to estimates that include simulation-induced bias that researchers can easily avoid by relying on the invariance property of ML estimators instead.

Conclusion

Many social scientists turn to ?'s seminal article when seeking advice on how to interpret, summarize, and present empirical results. By highlighting the importance of reporting substantively meaningful quantities of interest, it has significantly improved empirical research in political science and neighboring disciplines. However, political scientists following ?'s advice estimate quantities of interest either with the average of simulated quantities of interest (e.g., Clarify in Stata, Zelig in `R`) or using the invariance property of ML estimators (e.g., margins in Stata and `R`). In practice, researchers' choice between the two approaches seems idiosyncratic rather than principled, depending on their preferred software package rather than statistical principles. Even the methodological

---

[6]We generally agree with the arguments in favor of the observed-value approach but recommend that researchers plot the distribution of quantities of interest in addition to providing a summary measure such as their average. See ? for examples.

literature fails to pay attention to differences between the two approaches to estimating quantities of interest.

? stresses the importance of transformation-induced bias, which originates in the non-linear transformation of model coefficient estimates into estimated quantities of interest. As **?** shows, such transformation-induced biases are large when standard errors are large or when the transformation of the model coefficients into quantities of interest is highly non-linear. We show that when researchers use the simulation average to estimate quantities of interest, they roughly double the transformation-induced bias that Rainey (2017) describes. We refer to this unnecessary bias as "simulation-induced bias." The good news is that the fix is easy: we do not have to use the simulation-based approach to estimate a quantity of interest. Instead, we can simply plug model coefficients into the transformation to obtain a ML estimate of the quantity of interest. We recommend that statistical software does this by default.[7]

Finally, if researchers use the invariance property to compute ML estimates of quantities of interest, how should they conduct statistical inference? Commonly employed approaches include the delta method, stochastic simulation, and the bootstrap (e.g., **?**. **?** presents some limited Monte Carlo evidence that these approaches lead to similar inferences but a detailed examination of this question is, to the best of our knowledge, still missing from the literature.

---

[7] Based on communications with Christopher Gandrud, a member of the Zelig Core Team, it appears that Zelig sometimes uses the median of the simulation draws as point estimator (as opposed to the mean). We cannot find any information in the Zelig documentation (Choirat et al. 2017) for when that might happen. Note that the simulation median corresponds to the ML estimate (in expectation) for monotonic transformations. (Even then, simulation introduces Monte Carlo error that the researcher could easily avoid by using the invariance property of ML estimators in the first place.) To see this revisit the previous example where $\tau(\mu) = \mu^2$. Imagine the best-case scenario in which $\hat{\mu}^{\mathrm{mle}} = 0$, so that the estimate based on the invariance property is unbiased. Even then simulated draws of the quantity of interest are almost surely greater than zero. Taking either the mean or the median of the simulated draws will result in an estimator with more bias than the ML estimator.