Unnecessary Bias

The Consequences of Averaging Simulated Quantities of Interest*

Carlisle Rainey†

Holger L. Kern‡

January 26, 2018

Abstract

Following King, Tomz, and Wittenberg (2000), researchers commonly convert coefficient estimates into an estimate of the quantity of interest using the average of simulations. However, other researchers simply use the invariance property of maximum likelihood estimates to directly convert the model coefficient estimates into the quantity of interest. These approaches are not equivalent, yet researchers rarely justify their choice. I show that the average of simulations can introduce substantial bias compared to the maximum likelihood estimate. In general, when reporting point estimates of the quantity of interest, researchers should report the maximum likelihood estimate, not the average of the simulations.

Political scientists now routinely 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 for ordered categorical outcomes; multinomial logit and probit for unordered categorical outcomes; Poisson and negative binomial regression for count data, and beta regression for fractions (Paolino 2001). We could list numerous additional ML estimators here, many of them proposed by political scientists and regularly used in political science research. However, the model coefficients from these estimators are not always directly informative about the quantities of interest.

Instead of focusing on model coefficients, King, Tomz, and Wittenberg (2000) suggest focusing on substantively meaningful quantities of interest and offer a method for computing these quantities. Their suggestion has improved political science research enourmously. Rather than focusing on model coefficients, almost all researchers rely on more substantively meaningful quantities such as predicted probabilities, first diffences, and marginal effects.

King, Tomz, and Wittenberg (2000) also offer a simulation-based method for computing these quantities of interest. They suggest that the researcher repeatedly (1) draw model coefficients from a multivariate normal distribution, (2) transform the coefficients into the quantity of interest, and

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

†Carlisle Rainey is Assistant Professor of Political Science, Texas A&M University, 2010 Allen Building, College Station, TX, 77843. (crainey@tamu.edu).

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

(3) summarize the distribution of simulated quantity of interest to obtain. The authors suggest that researcher can summarize these simulation to obtain the desired point and interval estimate.

Our analysis examines one piece of King, Tomz, and Wittenberg's (2000) advice: 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. With theory and examples, we show that averaging simulations is suboptimal. In short, it exaggerates the transformation-induced bias described by Rainey (2017). Instead, we propose that political scientists rely on the invariance property to calculate the ML estimate of the quantity of interest.

The invariance property of ML estimators allows the reseacher to find the ML estimate of *a function* of a parameter by first using ML to estimate the parameter and then applying the function to that estimate (King 1998, pp. 75-76, and Casella and Berger 2002, pp. 320-321). We refer to this estimate as the "ML" estimate of the quantity of interest. 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 is connected to a design matrix X of k explanatory variables and a column of ones by a link function g, 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 maximum likelihood to compute estimates $\hat{\beta}^{\text{mle}}$ for the parameter vector β. We refer to the function that transforms model coefficients into quantities of interest as $\tau(\cdot)$. For example, if the researcher using a logit model chooses to focus on a predicted probability for a particular observation X_c, then $\tau(\beta) = \text{logit}^{-1}(X_c\beta) = \dfrac{1}{1+e^{-X_c\beta}}$. The the researcher can use the invariance property $\hat{\tau}^{\text{mle}} = \tau\left(\hat{\beta}^{\text{mle}}\right) = \text{logit}^{-1}\left(X_c\hat{\beta}^{\text{mle}}\right) = \dfrac{1}{1+e^{-X_c\hat{\beta}^{\text{mle}}}}$ to quickly obtain a maximum likelihood estimate of the predicted probability.

Software implementations differ. Some, such as margins in Stata, report the ML estimate of the quantity of interest using the invariance principle. Others, such as Clarify in Stata or Zelig in R, report the simulation average estimate.

The methodology literature offers similarly divided advice. Herron (1999) suggests using the invariance principle to compute the ML estimate of the quantity of interest and using simulation to calculate the measures of uncertainty.[1] Carsey and Harden (2013) follows King, Tomz, and Wittenberg (2000) and recommends that researchers use the simulation average estimate. We do not know of any paper that constrasts the ML and the simulation average estimates. Indeed, it seems that the literature considers both approaches valid and perhaps equivalent. However, we argue that the ML estimate has a distinct advantage over the simulation average estimate.

[1] Herron (1999) cites an earlier version of King, Tomz, and Wittenberg (2000), but does not comment on whether researchers should prefer the ML estimate over the simulation average estimate.

Transformation-Induced τ-Bias

While we argue that researchers should use the invariance principle to transform coefficient estimated into the quantity of interest, Rainey (2017) shows that this transformation can introduce bias into the estimates of the quantity of interest. To highly the impact of the transformation, 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 defines these as

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

Note that 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. We do not consider coefficient-induced τ-bias any further.

Instead, we focus on transformation-induced τ-bias. Notice that we can predicted the direction of the transformation-induced bias using the shape of the transformation that converts the model coefficients into the quantities of interest. In general, any strictly convex (concave) $\tau(\cdot)$ creates upward (downward) transformation-induced τ-bias.

The Average of Simulations

Rather than relying on the invariance property to compute the point estimate for the quantity of interest, King, Tomz, and Wittenberg (2000) suggests the following simulation-based approach:

1. *Fit the model.* Use maximum likelihood 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, the researcher might choose to (1) focus either on a particual scenario or (2) compute the average quantity of interest across all observed cases in the data set (Hanmer and Kalkan 2013). We return to Hanmer and Kalkan (2013) in a later section. 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]

[2] As King, Tomz, and Wittenberg (2000) note, this step might require additional simulation, to first introduce and then average over fundamental uncertainty. We ignore this additional step here since it does not affect our argument.

[3] In the discussion that follows, we assume no Monte Carlo error exists in $\hat{\tau}^{\text{avg}}$. In other words, we assume that M is

The Average of Simulations Versus the Maximum Likelihood Estimate

Researchers have two potential estimates of the quantity of interest: the ML estimate $\hat{\tau}^{\text{mle}}$ that they calculate using the invariance principle and the average simulation estimate $\hat{\tau}^{\text{avg}}$ that they calculate using the algorithm described by King, Tomz, and Wittenberg (2000).

Documentation for statistical software does not draw a strong distinction between the two estimates. Indeed, researchers often need to look deep into the details to determine which estimate the software reports. Similarly, many researchers to not report which estimate they report, including the authors. In our previous work, we have used both $\hat{\tau}^{\text{mle}}$ and $\hat{\tau}^{\text{avg}}$ without giving much thought to the choice. Indeed, the differences between $\hat{\tau}^{\text{mle}}$ and $\hat{\tau}^{\text{avg}}$ has not received attention in published research. Typically, we can only determine which estimate reseachers use if (1) they report the software they use and (2) referring to the details in the documentation. But the preceding discussion raises questions. How does $\hat{\tau}^{\text{avg}}$ compare to $\hat{\tau}^{\text{mle}}$? Are they the same? If not, how do they differ? Is one more biased than the other?

If the transformation of estimated model coefficients into estimated quantities of interest is always convex (or always concave), then Jensen's inequality allows the simple statement given in Lemma 1 relating the estimate based on the average of stochastic simulations and the estimate based on the ML invariance property.

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,

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

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

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

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

This result is intuitive. Since we simulate using a multivariate normal distribution, $\tilde{\beta}$ has a symmetric

sufficiently large so that $\hat{\tau}^{\text{avg}} = \text{E}\left[\tau\left(\tilde{\beta}\right)\right]$, where $\tilde{\beta} \sim MVN\left[\hat{\beta}^{\text{mle}}, \hat{V}\left(\hat{\beta}^{\text{mle}}\right)\right]$.

distribution. By definition, $\hat{\tau}^{\text{mle}}$ 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 $\hat{\beta}^{\text{mle}}$, 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 $\hat{\tau}^{\text{mle}}$.

For a convex transformation, Lemma 1 shows that $\hat{\tau}^{\text{avg}}$ is always larger than $\hat{\tau}^{\text{mle}}$. But does this imply that $\hat{\tau}^{\text{avg}}$ is *more biased* than $\hat{\tau}^{\text{mle}}$? 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 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 $\underbrace{\mathrm{E}\left(\hat{\tau}^{\text{avg}}\right) - \tau\left[\mathrm{E}\left(\hat{\beta}^{\text{mle}}\right)\right]}_{\text{t.i. } \tau\text{-bias in } \hat{\tau}^{\text{avg}}} > \underbrace{\mathrm{E}\left(\hat{\tau}^{\text{mle}}\right) - \tau\left[\mathrm{E}\left(\hat{\beta}^{\text{mle}}\right)\right]}_{\text{t.i. } \tau\text{-bias in } \hat{\tau}^{\text{mle}}} > 0.$

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

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 $\hat{\tau}^{\text{avg}}$ than for $\hat{\tau}^{\text{mle}}$ for strictly convex or concave $\tau(\cdot)$.

Because the invariance principle offers an easy method to compute an estimator with less transformation-induced bias, we refer to the additional transformation-induced bias in $\hat{\tau}^{\text{avg}}$ compared to $\hat{\tau}^{\text{mle}}$ as "unnecessary" τ-bias, so that

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

An Approximation for the Unnecessary Bias in $\hat{\tau}^{\text{avg}}$

Theorem 1 guarantees that $\hat{\tau}^{\text{avg}}$ is more biased than $\hat{\tau}^{\text{mle}}$. This raises yet more questions. By how much? Is the bias trivial or substantial? Monte Carlo experiments allow one to assess this directly, but an analytical approximation provides a helpful rule of thumb. We approximate the *unnecessary*

transformation-induced τ-bias in $\hat{\tau}^{\text{avg}}$ as

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

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 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, Rainey (2017, p. 405, Eq. 1) computes as

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

where $H\left[\mathrm{E}\left(\hat{\beta}^{\text{mle}}\right)\right]$ represents the Hessian matrix of second derivatives of τ 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 we compare Equations 2 and 3, we are yet again comparing 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 *unnecessary* transformation-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 transformation-induced τ-bias in $\hat{\tau}^{\text{avg}}$ will be about *double* the transformation-induced τ-bias in $\hat{\tau}^{\text{mle}}$.[4]

Because of the similarity in Equations 2 and 3, the unnecessary bias becomes large under the conditions identified by Rainey (2017) as leading to large transformation-induced τ-bias: when the

[4]This rule of thumb naturally raises the question of whether we can use it to de-bias estimates of quantities of interest by adjusting $\hat{\tau}^{\text{mle}}$ by the difference between $\hat{\tau}^{\text{mle}}$ and $\hat{\tau}^{\text{avg}}$. Some preliminary Monte Carlo evidence suggests that the performance of the bias-adjusted estimator depends strongly on the dgp. Bias-adjustment can lead to smaller MSEs as well as larger MSEs, mirroring the performance of other de-biasing strategies such as bootstrap bias correction (Efron and Tibshirani 1993: ch. 10).

non-linearity in the transformation $\tau(\cdot)$ is severe and when the standard errors of $\hat{\beta}^{\text{mle}}$ are large. While the unnecessary bias vanishes as the number of observations grows large, it can be substantively meaningful for the sample sizes commonly encountered in social science research (Rainey 2017).

The Intuition

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

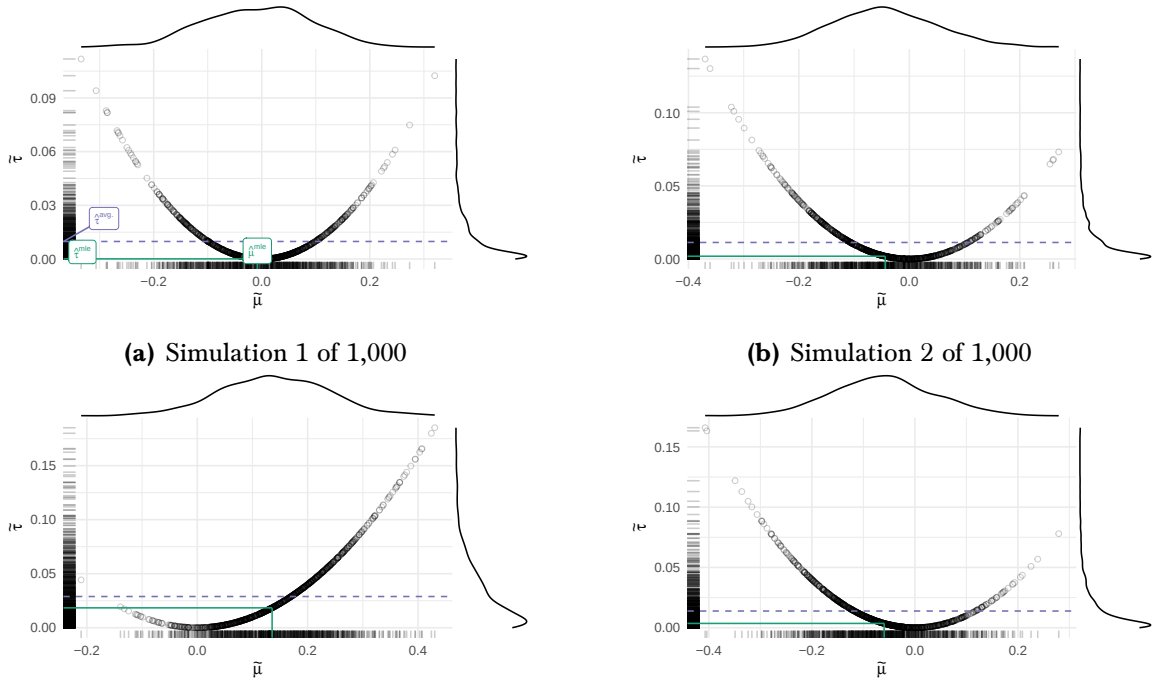
To develop an intuition for the unnecessary bias in $\hat{\tau}^{\text{avg}}$, consider the simple scenario in which $y_i \sim N(\mu, 1)$, for $i \in \{1, 2, \ldots, n = 100\}$ and the researcher wishes to estimate μ^2. For this example, suppose that the researcher know that the variance equals one but does not know that the mean μ equals zero. Suppose the researcher uses the maximum likelihood estimator $\hat{\mu}^{\text{mle}} = n^{-1} \sum_{i=1}^{n} y_i$ of μ, which happened to be the best unbiased estimator 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, $\hat{\mu}^{\text{avg}}$ *always* larger than the true value $\tau(\mu) = \tau(0) = 0^2 = 0$.

We can see the dynamics even more 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 $\hat{\mu}^{\text{mle}}$ and $\hat{\tau}^{\text{mle}}$ are unusually close to the true values $\mu = 0$ and $\tau(\mu) = \mu^2 = 0$ in this sample, $\hat{\tau}^{\text{avg}}$ falls well above the true value $\tau(\mu) = \mu^2 = 0$.

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.



(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: Four figures illustrating the relationship between $\hat{\tau}^{\text{mle}}$ and $\hat{\tau}^{\text{avg}}$ described by Lemma 1 and Theorem 1.

Figures 1b-1d repeat this process three more times to give some sense of how the dynamic changes for different samples. In each case, the story 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 of $\hat{\mu}^{\text{mle}}$, $\hat{\tau}^{\text{mle}}$, and $\hat{\tau}^{\text{avg}}$. Figure 2 shows the density plots for the three empirical sampling distributions. As we would expect, $\hat{\mu}^{\text{mle}}$ 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 more so. Theorem 1 shows why this must be the case.

Using the Law of Iterated Expectations

We can also develop the intuition using a more mathematical approach via the law of iterated expectations. For this it helps if we 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 $\text{E}_Y\left(\text{E}_{X|Y}(X \mid Y)\right) = \text{E}_X(X)$, 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, $\text{E}_{X|Y}$ 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

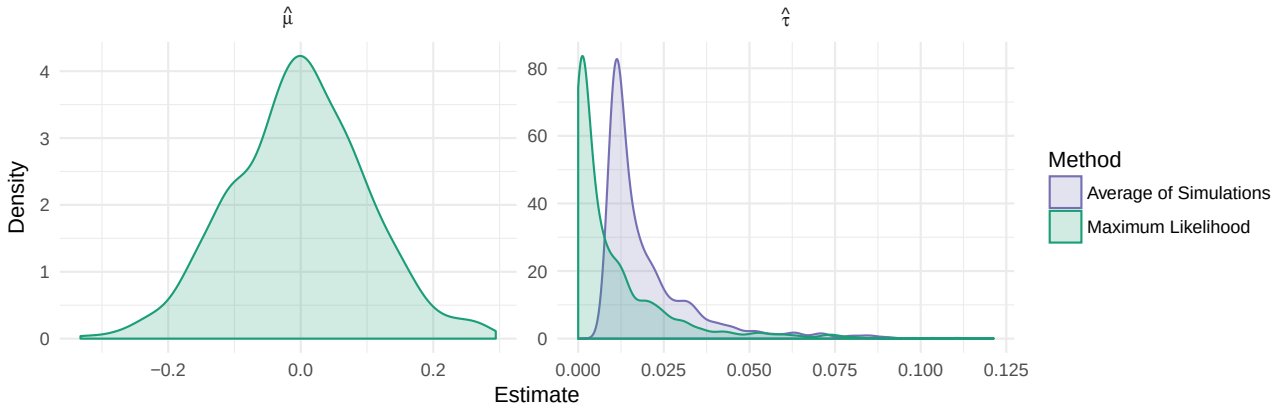


Figure 2: The sampling distributions of $\hat{\beta}^{\text{mle}}$, $\hat{\tau}^{\text{mle}}$, and $\hat{\tau}^{\text{avg}}$.

as $\tilde{\beta} \mid \hat{\beta}^{\text{mle}}$. To remain consistent with previous work, especially King, Tomz, and Wittenberg (2000) and Herron (1999), we simply use $\tilde{\beta}$ to represent $\tilde{\beta} \mid \hat{\beta}^{\text{mle}}$. The definition of $\tilde{\beta}$ makes this 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, then 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 unconditional distribution $\tilde{\beta}$. The unconditional distribution helps us understand the nature of the unnecessary transformation-induced τ-bias.

Applying the law of iterated expectations, we obtain $\mathrm{E}_{\tilde{\beta}}\left(\tilde{\beta}\right) = \mathrm{E}_{\hat{\beta}^{\text{mle}}}\left(\mathrm{E}_{\tilde{\beta}|\hat{\beta}^{\text{mle}}}(\tilde{\beta} \mid \hat{\beta}^{\text{mle}})\right)$. 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 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 transformation-induced 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 $\hat{\tau}^{\text{avg}}$.

Illustrative Simulation

As an 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 instantaneous 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 $+3$.

Following the procedures discussed above, 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 transformation-induced τ-bias in both $\hat{\tau}^{\text{mle}}$ and $\hat{\tau}^{\text{avg}}$ will 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 shows the transformation-induced τ-bias in $\hat{\tau}^{\text{avg}}$ and $\hat{\tau}^{\text{mle}}$ compared to the true value $\tau(\beta)$. Notice three features of this plot. First, the bias is substantial. The relative size of the bias varies, but when the true marginal effect is greater than 0.5, the average transformation-induced τ-bias in $\hat{\tau}^{\text{mle}}$ is about $\frac{1}{3}$ the size of the true effect. For $\hat{\tau}^{\text{avg}}$, the bias is about $\frac{3}{4}$ the size of the true effect. Second, notice 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. Third, notice that 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.

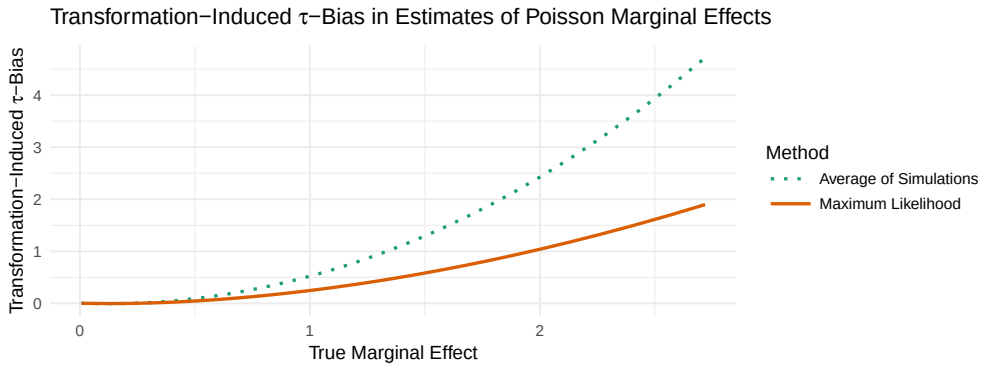


Figure 3: This figure shows the bias in the estimates of the marginal effects in a Poisson regression model. Notice that the convex transformation $\tau(\beta) = \frac{d\,\text{E}(y)}{dx} = e^{(\beta_{cons}+\beta_x x)}$ creates a positive bias (see Theorem 1) and that the bias in $\hat{\tau}^{\text{avg}}$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

Example: Supreme Court Decisions

To unify explanation of U.S. Supreme Court decisions, George and Epstein (1992) fit a single probit model that combines the legal and extralegal models of Court decision-making to a data set of 64 decisions. The authors model 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 George and Epstein (1992) for the more details of the model.

We use this model illustrate the potential impact of using the simulation average rather than the maximum likelihood estimate of the quantity of interest. We focus on two potential quantities of interest: the probability of a conservative decision and the effect of the Solicitor General filing a brief. Table 1 summarizes these quantities of interest.

Table 1: This table provides the details of the quantities of interest from George and Epstein's (1992) model of U.S. Supreme Court decisions.

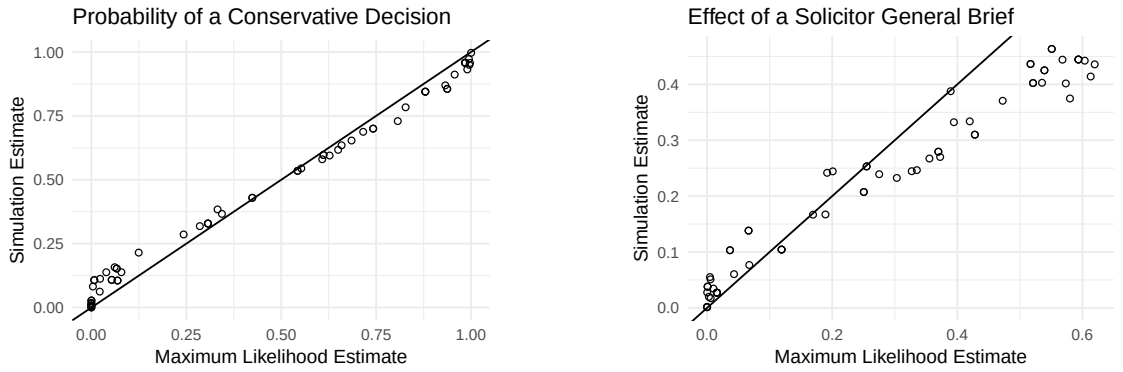
Description	Notation	Change in Key Explanatory Variable	Values for Other Explanatory Variables
probability of a conservative decision	$\tau(\beta) = \Phi(X_c\,\beta)$	none	every observed combination
effect of a Solicitor General brief on the probability of a conservative decision	$\tau(\beta) = \Phi(X_{\text{high}}\beta) - \Phi(X_{\text{low}}\beta)$	for X_{high}, SG = 1, and for X_{low}, SG = 0	every observed combination

For each quantity of interest, we compute an estimate using the average of simulation and maximum likelihood. First, we use both the average of simulations and maximum likelihood to estimate the probability of a conservative decision for each combination of explanatory variables included in the data set. Second, we use both approaches to estimate the effect of a Solicitor General brief on the probability of a conservative decision. We define this effect as the *difference* in the probability of a conservative decision for each observation in the data set, if that observation changed from one in which the Solicitor General *did not* file a brief (SG = 0) to one in which the Solicitor General *did* file a brief (SG = 1).

Figure 4 compares the estimates. First, consider the estimates of the probability of a conservative decision in Figure 4a. The pattern is clear: when the chance of a conservative decision is less than 50%, the average of the simulations is too large. In this region, the transformation (the normal cdf) is convex. When the chance of a conservative decision is greater than 50%, the average of the simulations is too small. In this region, the transformation is concave. When the chance of a conservative decision is closer to 50%, the differences between the average of the simulations and the maximum likelihood estimate are smaller, because the transformation is more linear in this area. The same is true for chances close to 0% and 100%.

Further, some of the differences are quite large. For example, when maximum likelihood suggests a chance of about 5%, the average of the simulation suggests a chance of about 10%. This difference may seem small at first (i.e., only 5 percentage points), but the average of simulations is about *double*

the maximum likelihood estimate.



(a) **(b)**

Figure 4: The figure shows the relationship between the simulation average and the maximum likelihood estimate two the quantities of interest. The left panel (a) shows the probability of a conservative decision. Notice that the simulation average tends falls about the maximum likelihood estimate when the probability is low–where the transformation is convex–and below the maximum likelihood estimate when the probability is high–where the transformation is concave. The right panel (b) shows the effect of a brief by the Solicitor General on the probability of a conservative decision.

Now consider the estimates of the effect of the Solicitor General filing an Amicus brief in Figure 4b. The largest differences appear in the upper-right corner of the plot. For this group of observations, the average of simulations suggests than a brief from the Solicitor General increases the chance of a conservative decision by about 40 percentage points. On the other hand, the maximum likelihood estimate suggests an increase of about 60 percentage points. This difference is certainly meaningful—the maximum likelihood estimate is 50% larger than the average of the simulations.

A Note on Hanmer and Kalkan (2013)

Hanmer and Kalkan (2013) discusses two approaches to computing quantities of interest: the more commonly used *average-case* approach and their recommended *observed-value* approach. With either approach, researchers estimate the quantity of interest as the value of the key explanatory variable changes. However, in many 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. Hanmer and Kalkan (2013), in contrast, suggests estimating the quantity of interest for all sample observations, leaving their explanatory variables (except for the key variable of interest) at their observed values, and then averaging the estimates across the sample. Here, this choice is implicitly part of the transformation $\tau(\cdot)$, so their (compelling) argument does not undermine or enhance our own.[5]

[5]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 Ai and Norton (2003) for examples.

Because researchers have not drawn a sharp conceptual distinction between using the average of simulation draws and using the ML invariance property, Hanmer and Kalkan (2013) does not discuss this choice. Since it explicitly builds on King, Tomz, and Wittenberg (2000), we interpreted Hanmer and Kalkan (2013) as relying on the average of simulation draws when computing quantities of interest. The replication archive for the article confirms that this is indeed the case.

The important point is this: Hanmer and Kalkan (2013) draws a distinction between the average-case and observed-value approaches to computing quantities of interest. Our paper draws a distinction between estimating quantities of interest (whether average-case or observed-value based) using the average of simulated draws and using the invariance property. Regardless of whether researchers use the average-case approach or the observed-value approach, the simulation average leads to estimates that generally suffer from transformation-induced bias that researchers can easily avoid by relying on the invariance property instead.

Conclusion

Many political scientists turn to King, Tomz, and Wittenberg's (2000) seminal paper when seeking advice on how to interpret, summarize, and present empirical results. By highlighting the importance of reporting substantively meaningful quantities of interest along with the uncertainty, King, Tomz, and Wittenberg (2000) has significantly improved empirical research in political science and neighboring disciplines. However, depending on the statistical software used, political scientists following King, Tomz, and Wittenberg's advice will 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 to compute the ML estimate (e.g., margins in Stata and `R`). In practice, researchers' choice between the two approaches seems idiosyncratic rather than principled. As far as we can tell, researchers' choice depends largely on their prefer software package rather than statistical principles. Even the methodological literature has failed to pay attention to differences between the two approaches to estimating quantities of interest.

Rainey (2017) stresses the importance of transformation-induced bias, which originates in the non-linear transformation of model coefficient estimates into estimated quantities of interest. As Rainey (2017) shows, such transformation-induced biases is large when standard errors are large or when the transformation of the model coefficients into quantities of interest is highly non-linear. We shows that when researchers use the simulation average to estimate quantities of interest, they roughly double the transformation-induced bias that Rainey (2017) describes. The good news is that the fix is easy: do not use the average of simulation draws to estimate quantities of interest. Instead, simply plug model coefficients into the transformation to obtain an estimate of the quantities of interest. We recommend that statistical software does this by default.

References

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

George, Tracey E., and Lee Epstein. 1992. "On the Nature of Supreme Court Decision Making." *American Political Science Review* 86(2):323–337.

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.

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

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.

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