Unnecessary Bias

Do Not Use the Average of Simulations to Estimate the Quantity of Interest*

Carlisle Rainey†

Holger Kern‡

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.

Suppose the researcher uses maximum likelihood to estimate a statistical model in which $y_i \sim f(\theta_i)$, where $i \in \{1, ..., N\}$ 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 can use maximum likelihood to compute estimates $\hat{\beta}^{\text{mle}}$ for the parameter vector β. Note that we adopt a frequentist perspective, so $\hat{\beta}^{\text{mle}}$ represents a *random variable* that varies across samples.

But researchers usually care about a function τ of the model coefficients β rather than the model coefficients themselves. Following King, Tomz, and Wittenberg (2000), we refer to this function τ as the "quantity of interest." The researcher can use the invariance property $\hat{\tau}^{\text{mle}} = \tau\left(\hat{\beta}^{\text{mle}}\right)$ to quickly obtain a maximum likelihood estimate of the quantity of interest (King 1998, pp. 75-76, and Casella and Berger 2002, pp. 320-321).

Transformation-Induced τ-Bias

The invariance property does not come without cost. The transformation of unbiased model coefficient estimates introduces bias into the estimate of the quantity of interest. If the coefficient estimates are biased, the transformation-induced bias can, but generally does not, offset the bias in the coefficient estimates. To separate the sources of bias, 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.

*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 Kern is Assistant Professor of Political Science, Florida State University, 600 W. College Avenue Tallahassee, FL, 32306.(hkern@fsu.edu).

These are defined as

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

The direction and magnitude of the coefficient-induced τ-bias depends on the choice of τ and the bias in the coefficient estimates, but an unbiased estimator $\hat{\beta}$ implies no coefficient-induced τ-bias. we do not consider coefficient-induced τ-bias any further. Transformation-induced τ-bias, though, can be understood based on the shape of the transformation. In general, any strictly convex (concave) τ creates upward (downward) transformation-induced τ-bias.

The Average of Simulations

Some commonly used software, such margins in Stata (and Leeper's margins port into R), use the invariance property to estimate the quantity of interest. But other software, such as CLARIFY for Stata and Zelig for R, adopts an alternative approach recommended by King, Tomz, and Wittenberg (2000).

King, Tomz, and Wittenberg (2000) suggest the following algorithm:

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 $i \in \{1, 2, ..., M\}$ using $\tilde{\beta}^{(i)} \sim N\left[\hat{\beta}^{\text{mle}}, \hat{V}\left(\hat{\beta}^{\text{mle}}\right)\right]$, where N is the multivariate normal distribution.[1]
3. *Convert simulated coefficients to simulated quantity of interest.* Compute $\tilde{\tau}^{(i)} = \tau\left(\tilde{\beta}^{(i)}\right)$ for $i \in \{1, 2, ..., M\}$. Most quantities of interest depend on the values of the explanatory variables. In this case, the researcher must choose to focus on a particular scenario, or perhaps average across several scenarios (Hanmer and Kalkan 2013). In any case, the transformation τ 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 $\hat{\tau}^{\text{avg.}} = \frac{1}{M}\sum_{i=1}^{M}\tilde{\tau}^{(i)}$.

In the discussion that follows, we assume no Monte Carlo error exists in $\hat{\tau}^{\text{avg.}}$. That is, make M large enough to assume that $\hat{\tau}^{\text{avg.}} = \mathrm{E}\left[\tau\left(\tilde{\beta}\right)\right]$, where $\tilde{\beta} \sim N\left[\hat{\beta}^{\text{mle}}, \hat{V}\left(\hat{\beta}^{\text{mle}}\right)\right]$.

One can hardly overstate the importance of this simulation method to political science research. According to the Web of Science, King, Tomz, and Wittenberg (2000) has been cited 1,097 times, 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*, or *Political Analysis*. It is the second most-cited article *overall* in the *American Journal of Political Science*, falling just 56 citations behind Beck, Katz, and Tucker (1998). Google Scholar suggests that King, Tomz, and Wittenberg (2000) might have as many as 3,598 citations and that Tomz, Wittenberg, and King (2003), which provides and overview of the CLARIFY software, might have as many as 1,437 citations. Make no mistake, the procedures proposed by King, Tomz, and Wittenberg (2000) and software offered by Tomz, Wittenberg, and King (2003) have meaningfully altered (and improved) the practice of political science research.

[1] One might use alternative approach to generate $\tilde{\beta}$, such as parametric or non-parametric bootstrap. As long $\tilde{\beta}$ approximately follows a normal distribution, then the intuitions and best practices we offer below still apply.

[2] As King et al. note, this step might require additional simulation, but most times not. Our arguments do not depend on this simplification.

The Average of Simulations Versus the Maximum Likelihood Estimate

Applied researchers seem to use $\hat{\tau}^{\text{avg.}}$ and $\hat{\tau}^{\text{mle}}$ interchangeably. But the preceding discussion raises questions. How does $\hat{\tau}^{\text{avg.}}$ compare to $\hat{\tau}^{\text{mle}}$? Are they the same? How are they different? Is one more biased than the other?

If the transformation is always convex (or always concave), then Jensen's inequality allows the simple statement given in Lemma 1 relating the average of simulations and the maximum likelihood estimate.

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

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), we know that $\mathrm{E}\left[\tau\left(\tilde{\beta}\right)\right] > \tau\left[\mathrm{E}\left(\tilde{\beta}\right)\right]$, so that

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

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

This result is intuitive. By assumption, $\tilde{\beta}$ has a symmetric 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 τ pulls $\tau(\tilde{\beta})$ in toward $\hat{\tau}^{\text{mle}}$. If $\tilde{\beta}$ happens to fall above the mode $\hat{\beta}^{\text{mle}}$, then τ 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 this is the case.

Theorem 1 *Suppose a maximum likelihood estimator $\hat{\beta}^{mle}$. Then for any strictly convex or concave τ, 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 $\hat{\tau}^{\text{mle}}$ for convex or concave τ.

An Approximation for the Additional 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 is it substantial? Monte Carlo experiments allow one to assess this directly, but an analytical approximation provides a helpful rule of thumb.

we approximate the *additional* transformation-induced τ-bias in $\hat{\tau}^{\text{avg.}}$ compared to $\hat{\tau}^{\text{mle}}$ as

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

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 $\hat{\beta}^{\text{mle}}$.

This is 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 rule holds, the *additional* transformation-induced τ-bias in $\hat{\tau}^{\text{avg.}}$ is about the same as the transformation-induced τ-bias in $\hat{\tau}^{\text{mle}}$. Therefore, as a rule of thumb, we suggest that the transformation-induced τ-bias in $\hat{\tau}^{\text{avg.}}$ will be about *double* the transformation-induced τ-bias in $\hat{\tau}^{\text{mle}}$.[3]

Because of the similarity in Equations 2 and 3, the difference between $\hat{\tau}^{\text{avg.}}$ and $\hat{\tau}^{\text{mle}}$ becomes large under the same conditions that Rainey (2017) notes transformation-induced τ-bias becomes large: when the non-linearity

[3]When writing Rainey (2017), we considered using the difference between $\hat{\tau}^{\text{avg.}}$ and $\hat{\tau}^{\text{mle}}$ to estimate (and correct for) the transformation-induced τ-bias in $\hat{\tau}^{\text{mle}}$. This crude form of parametric bootstrap does a reasonable job of reducing the bias, but tends to increase the MSE of the estimated quantity of interest.

in the transformation in severe and when the standard errors of $\hat{\beta}^{\text{mle}}$ are large.

The Intuition

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

To develop an intuition for the additional bias in $\hat{\tau}^{\text{avg.}}$, consider the the simple scenario in which $y_i \sim N(\mu, 1)$, for $i \in \{1, 2, ..., 100\}$. The variance is known to be one. The mean μ is unknown, but equals zero. Suppose the researcher uses the best unbiased estimator $\hat{\mu}^{\text{mle}}$ of μ, but ultimately cares about the quantity of interest $\tau(\mu) = \mu^2$. The researcher can use the invariance property to estimate $\tau(\mu)$ as $\hat{\tau}^{\text{mle}} = \left(\hat{\mu}^{\text{mle}}\right)^2$. Or the researcher might use the simulation approach, estimating $\tau(\mu)$ as $\hat{\tau}^{\text{avg.}} = \frac{1}{M}\sum_{i=1}^{M} \tau\left(\tilde{\mu}^{(i)}\right)^2$, where $\tilde{\mu}^{(i)} \sim N\left(\hat{\mu}^{\text{mle}}, \frac{1}{\sqrt{n}}\right)$ for $i \in \{1, 2, ..., M\}$.

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, $\hat{\mu}^{\text{mle}} > 0$. Since $\hat{\mu}^{\text{mle}}$ rarely equals zero, it is clear that $\hat{\mu}^{\text{mle}}$ is biased upward. But even if $\hat{\mu}^{\text{mle}} = 0$, then $\tilde{\mu}^{(i)}$ rarely equals zero. If $\tilde{\mu}^{(i)} \neq 0$, then $\left(\tilde{\mu}^{(i)}\right)^2 > 0$. Thus, $\hat{\mu}^{\text{avg.}}$ is *always* larger than the true value $\tau(\mu) = 0$.

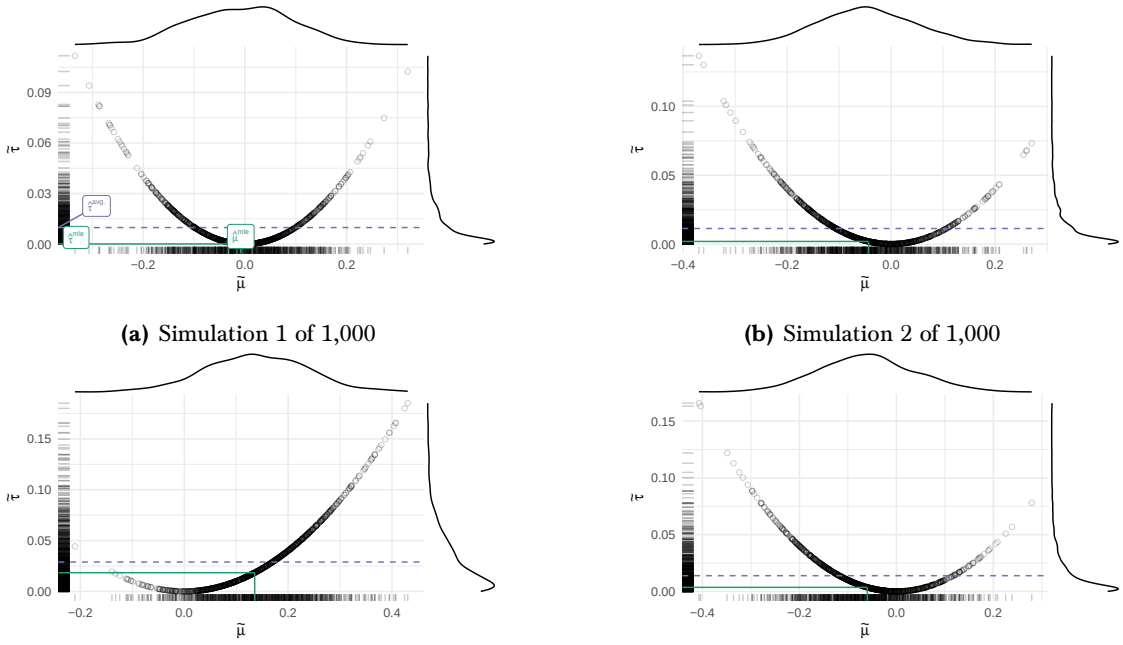
We can see the dynamics even more clearly by repeatedly simulating y and estimating $\hat{\tau}^{\text{mle}}$ and $\hat{\tau}^{\text{avg.}}$. Figures 1a-1d show the first four of 1,000 total simulations. The figures show how the the unbiased estimate $\hat{\mu}^{\text{mle}}$ is translated into $\hat{\tau}^{\text{mle}}$ 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 $i \in \{1, 2, ..., M = 1,000\}$, (2) calculate $\tilde{\tau}^{(i)} = \tau\left(\tilde{\mu}^{(i)}\right)^2$, 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 shows 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 $\hat{\mu}^{\text{mle}}$ estimates $\mu = 0$ unusually well, $\hat{\tau}^{\text{avg.}}$ is biased substantially 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 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 τ has the effect of lengthening the right tail of the distribution of $\tilde{\tau}$, which pulls the average well above the mode. This provides the basic intuition for Lemma 1.

Figures 1b-1d repeat this process three more times to give some sense of how the dynamic changes for different data sets y. In each case, the story is similar–the convex transformation τ stretches the distribution of $\tilde{\tau}$ to the right, which pulls $\hat{\tau}^{\text{avg.}}$ above $\hat{\tau}^{\text{mle}}$.

We repeat this process until we have 1,000 different estimates of $\hat{\mu}^{\text{mle}}$, $\hat{\tau}^{\text{mle}}$, and $\hat{\tau}^{\text{avg.}}$. Figure 2 shows the density plot for each of these 1,000 estimates (i.e., the sampling distributions). Notice $\hat{\mu}^{\text{mle}}$ is unbiased with a standard error of $\frac{1}{\sqrt{n}} = \frac{1}{\sqrt{100}} = \frac{1}{10}$. This confirms a basic result from statistical theory. Both $\hat{\tau}^{\text{mle}}$ and $\hat{\tau}^{\text{avg.}}$ are biased upward, but $\hat{\tau}^{\text{avg.}}$ more so. Theorem 1 ensures this must be the case.



(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.

Using the Law of Iterated Expectations

We can also develop the intuition using a more mathematical approach via the law of iterated expectations. However, 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 $\text{E}_Y\left(\text{E}_{X|Y}(X|Y)\right) = E_X(X)$, 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 $X|Y$, 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 $\tilde{\beta}|\hat{\beta}^{\text{mle}}$. To remain consistent with other papers, especially King, Tomz, and Wittenberg (2000) and Herron (1999), we simply use $\tilde{\beta}$ to represent $\tilde{\beta}|\hat{\beta}^{\text{mle}}$. The definition of $\tilde{\beta}$ makes this clear. In this section only, we use $\tilde{\beta}|\hat{\beta}^{\text{mle}}$ to represent the conditional distribution of $\tilde{\beta}$. we 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 $\tilde{\beta}|\hat{\beta}^{\text{mle}}$. If we do steps (1) and (2) just once, but step (3) repeatedly, then we have a sample from the conditional distribution $\tilde{\beta}|\hat{\beta}^{\text{mle}}$. If we do steps (1), (2), and (3) repeatedly, then we have a sample from the unconditional distribution $\tilde{\beta}$. The unconditional distribution helps one understand the additional bias.[4]

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

[4]More technically, we could also define the relevant distributions hierarchically. First, we have $\hat{\beta}^{\text{mle}} \sim s(\beta)$, where $s(\beta)$ represents the sampling distribution of $\hat{\beta}$. Then we have $\tilde{\beta} \sim N\left[\hat{\beta}^{\text{mle}}, \hat{V}\left(\hat{\beta}^{\text{mle}}\right)\right]$.

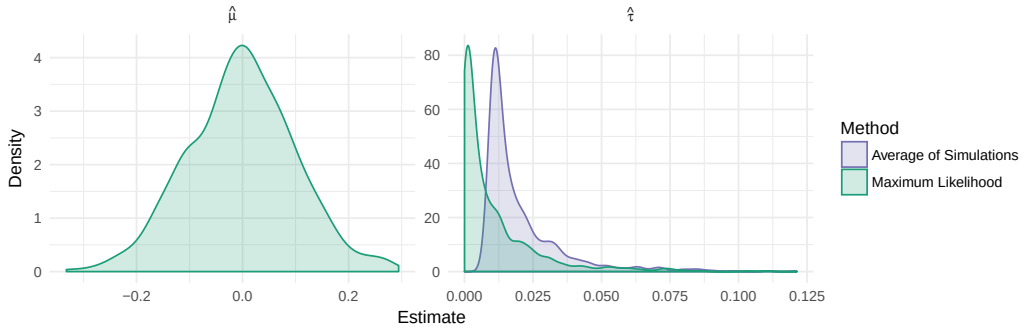


Figure 2: The sampling distributions of $\hat{\beta}^{\text{mle}}$, $\hat{\tau}^{\text{mle}}$, and $\hat{\tau}^{\text{avg.}}$.

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

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

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

If we subtract Equation 5 from Equation 4, then 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 from Equation 5, we must swap τ with an expectation once, so, if τ is convex, then 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 from Equation 6, we must swap τ with an expectation *twice*. Again, if τ is convex, then Equation 6 must be greater than Equation 4. However, because we expect $\hat{\beta}^{\text{mle}}$ and $\tilde{\beta}|\hat{\beta}^{\text{mle}}$ 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 Simulations

Marginal Effects in Poisson Regression

As an illustration, consider the Poisson regression model $y_i \sim \text{Poisson}(\lambda_i)$, where $\lambda_i = e^{(-2+x_i)}$ for $i \in \{1, 2, ..., 100\}$. To create the x_is, we simply took 100 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 Theorem 1 guarantees that 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 $\tau(\beta)$. Especially

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 $\hat{\tau}^{\text{avg}\cdot}$, 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 $\tau(\beta) = \frac{d\,\mathrm{E}(y)}{dx} = e^{(\beta_{cons}+\beta_x x)}$ is convex, the bias is positive. Third, notice that the bias in $\hat{\tau}^{\text{avg}\cdot}$ is about twice as large as the bias in $\hat{\tau}^{\text{mle}}$, as the rule of thumb suggests.

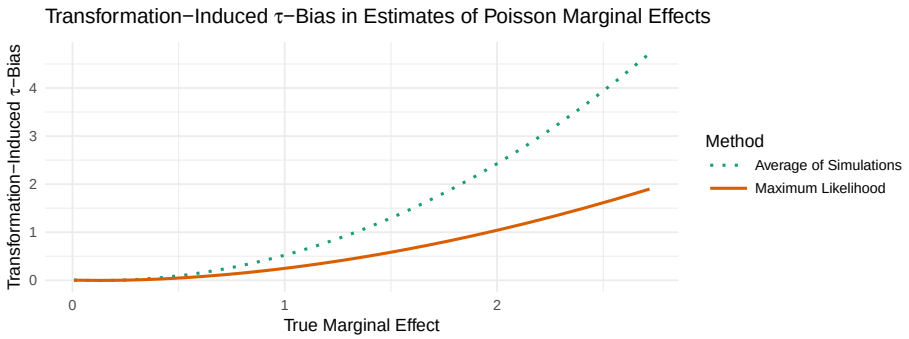


Figure 3: This figure shows the bias in the estimates of the marginal effects in a Poisson regression model. Notice that the convex transformation $\tau(\beta) = \frac{d\,\mathrm{E}(y)}{dx} = e^{(\beta_{cons}+\beta_x x)}$ creates a positive bias (see Theorem 1) and that the bias in $\hat{\tau}^{\text{avg}\cdot}$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

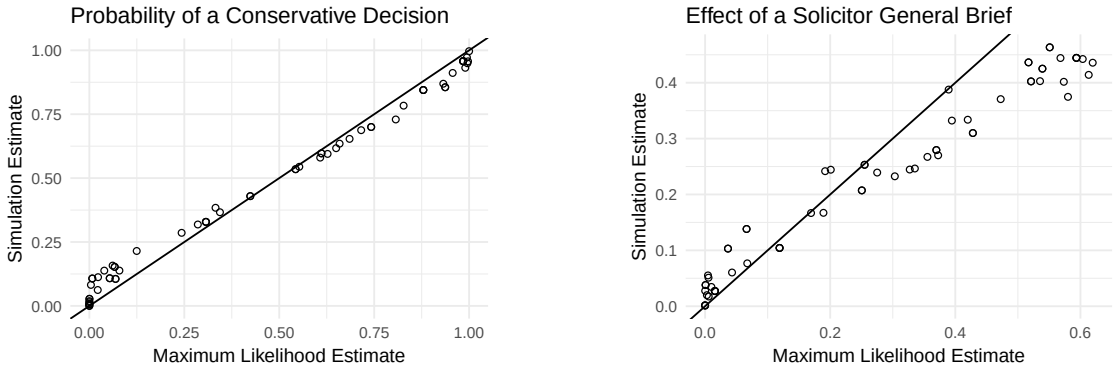


Figure 4: The figure shows the relationship between the simulation average and the maximum likelihood estimate two the quantities of interest. The left panel (a) shows the probability of a conservative decision. Notice that the simulation average tends falls about the maximum likelihood estimate when 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) discuss two approaches to computing quantities of interest: the typical "average-case" approach and their recommended "observed-value" approach. With either approach, the researcher estimates the quantity of interest–the change in the expected value of the outcome variable (e.g., the probability of a

conservative decision) as a key explanatory variable changes from a low value to a high value (e.g., SG changes from 0 to 1). But the researcher 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 at a typical value, such as the median. Hanmer and Kalkan (2013) suggest estimating the quantity of interest for all the observed combinations of the other explanatory variables (like we do above) and then averaging the quantity of interest across the combinations. In our argument, this choice is built into the transformation τ, so their (compelling) argument does not undermine or enhance our own.[5] Because researchers have not drawn a sharp conceptual distinction between the simulation average and maximum likelihood estimates of the quantity of interest, Hanmer and Kalkan (2013) do not adopt a clear position. We read their paper to suggest averaging the maximum likelihood estimates ($\hat{\tau}^{\text{mle}}$) for each observed case (though they do not explicit write this). However, in their Stata code, they average the simulation averages ($\hat{\tau}^{\text{avg}\cdot}$) for each observed case.

When we average the maximum likelihood estimates of the effect of a Solicitor General brief in the left panel (b) of Figure 4 (as Hanmer and Kalkan (2013) seem to suggest in their paper), we obtain an estimated effect of 0.28. When we average the simulation averages (as Hanmer and Kalkan do in their Stata code), we obtain an estimated effect of 0.23. The maximum likelihood estimate is about 22% larger than the simulation average–the choice matters.

The important point is this: Hanmer and Kalkan (2013) draw a distinction between the average-case and observed-value approaches to choosing the quantity of interest. This paper draws a distinction between estimating quantities of interest (whether average-case or observed-value) using the simulation average and the maximum likelihood estimate. Regardless of whether the researcher uses the average-case approach or the observed-value approach, the simulation average is more biased than the maximum likelihood estimate.

Conclusion

Substantive researchers in political science tend to estimate their quantity of interest using the average of simulation (e.g., CLARIFY in Stata, Zelig in R) or using the invariance property of maximum likelihood estimates (e.g., margins in Stata). In practice, researchers' choice between the two seems idiosyncratic rather than principled. But the choice is clear. Rainey (2017) introduces the idea of transformation-induced bias. This paper shows that the average of the simulations roughly *doubles* this bias. In many cases, the researcher has small standard errors. In this case, the additional bias is small. But in other cases, such as when the researcher has large standard errors or focuses on a highly non-linear transformation of the model coefficients, the bias is substantial. And the fix is easy: simply plug the coefficient estimates into the transformation to obtain the maximum likelihood estimates of the quantity of interest. Software should do this by default.

References

Beck, Nathaniel, Jonathan N. Katz, and Richard Tucker. 1998. "Taking Time Seriously: Time-Series-Cross-Section Analysis with a Binary Dependent Variables." *American Journal of Political Science* 42(4):1260–1288.

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

[5] We generally agree with their arguments in favor of the observed-value approach, but we recommend researchers plot the distribution of effects rather than summarizing them into a single average.

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.

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