Transformation-Induced Bias

Unbiased Coefficients Do Not Imply Unbiased Quantities of Interest*

Carlisle Rainey†

Abstract

Political scientists commonly focus on quantities of interest computed from model coefficients rather than on the coefficients themselves. However, the quantities of interest, such as predicted probabilities, first differences, and marginal effects, do necessarily not inherit the small sample properties of the coefficient estimates. Indeed, unbiased coefficients estimates are neither necessary nor sufficient for unbiased estimates of the quantities of interest. I characterize this transformation-induced bias, calculate an approximation, illustrate its importance with a hypothetical example, and discuss its importance to methodological research.

Political scientists use a wide range of statistical models $y_i \sim f(\theta_i)$, 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 $g(\theta_i) = X_i\beta$. For example, for the binary logit, f represents the Bernoulli probability mass function, g represents the logit function.

The researcher usually estimates β with maximum likelihood (ML), and, depending on the choice of g and f, the estimate $\hat{\beta}$ might have desirable small sample properties. However, ML does not produce unbiased estimates in general. For this reason, methodologists frequently use Monte Carlo simulations to assess the small sample properties of estimators and provide users with rules of thumb about appropriate sample sizes. For example, the ML estimates of β for the binary logit are biased away from zero, leading Long (1997, p. 54) to suggest that "it is risky to use ML with samples smaller than 100, while samples larger than 500 seem adequate."

Although methodologists tend to focus on estimating the model coefficients, substantive

*All computer code necessary for replication is available at github.com/carlislerainey/transformation-induced-bias.

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

researchers tend to focus on a "quantity of interest." A quantity of interest is simply a *transformation* τ of the model coefficients. Examples include marginal effects, first and second differences, predicted probabilities and expected values, and risk ratios (King, Tomz, and Wittenberg 2000).

Fortunately, the invariance principle allows the researcher to calculate estimates of the quantities of interest from the coefficient estimates in a principled manner. The invariance principle states that if $\hat{\beta}$ is the ML estimate of β, then for any function τ, the ML estimate of $\tau(\beta)$ is $\tau(\hat{\beta})$ (King 1998, pp. 75-76, and Casella and Berger 2002, pp. 320-321). That is, researchers can simply transform the ML estimates of the model coefficients to obtain an ML estimate of the quantity of interest. Of course, if $\hat{\beta}$ is a consistent estimator of β, then $\tau(\hat{\beta})$ must be a consistent estimator of $\tau(\beta)$. But the invariance principle raises an important question: Does $\tau(\hat{\beta})$ inherit the small sample properties of $\hat{\beta}$, such as unbiasedness or approximate unbiasedness? The answer is no; the estimates of the quantities of interest do not inherit the small sample properties of the coefficient estimates.

This subtle, yet crucial, point reveals a disconnect between the work done by substantive scholars and that done by methodologists. Methodological work tends to focus on obtaining excellent estimates of the model coefficients, while substantive research tends to focus on estimating quantities of interest.

Much methodological research implicitly suggests that an approximately unbiased coefficient estimate is necessary and/or sufficient for an approximately unbiased estimate of the quantity of interest. Classically, Nagler (1994) uses Monte Carlo simulations to assess the small sample properties of the scobit model coefficients, but he focuses on marginal effects and predicted probabilities in his illustrative application. Recently, Nieman (2015) uses simulations to assess the small sample properties of the coefficients in his strategic probit with partial observability, but he focuses his illustrative application on the predicted probability of civil war. In order to provide more compelling tools for substantive scholars, we must extend our evaluations beyond coefficient estimates to the quantities that substantive researchers typically care about.

The Concepts

As a motivating example, consider the log-linear model

$$\log(\text{income}_i) = \beta_{cons} + \beta_{edu}\text{education}_i + \epsilon_i,$$

where $\epsilon_i \sim N(0, \sigma^2)$, education is measured in years, and income is measured in thousands of dollars. Assuming that the researcher uses the correct model, then least squares, which is also the ML estimator, provides the best unbiased estimator of the coefficients β_{cons} and β_{edu}. However, the researcher is not likely interested in log(income), but in income itself. In particular, she might want to estimate the median income among those with 20 years of education $\text{med}(\text{income} \mid \text{education} = 20) = e^{\beta_{cons}+20\beta_{edu}}$. Because $\text{med}[\log(y)] = \log[\text{med}(y)]$ for a random variable y, one might guess that unbiased estimates of β_{cons} and β_{edu} lead to unbiased estimates of med(income | education = 20), but that is not the case. If we suppose that $N = 10$, $\beta_{cons} = 2.5$, $\beta_{edu} = 0.1$, $\sigma^2 = 1$, and education takes on integers roughly uniformly from 10 to 20, then $\tau(\beta_{cons}, \beta_{edu}) = e^{\beta_{cons}+20\beta_{edu}} \approx \$90k$. To calculate the bias in the estimate of the quantity of interest, though, I simulate 100,000 data sets. For each data set, I estimate the model coefficients and use the coefficient estimates to calculate the quantity of interest. Although $\hat{\beta}_{cons}$ and $\hat{\beta}_{edu}$ are unbiased, the estimate of med(income | education = 20) is strongly biased upward, so that $\text{E}[\tau(\hat{\beta}_{cons}, \hat{\beta}_{edu})] = E(e^{\hat{\beta}_{cons}+20\hat{\beta}_{edu}}) \approx \$106k$.

A similar, but conceptually distinct issue arrises when researchers want to calculate the *mean* from a log-linear model $\log(y) = X\beta + \epsilon$. Many textbooks highlight that $\text{E}[\log(y|X)] \neq \log[\text{E}(y|X)]$, so that $\text{E}(y|X_i) \neq e^{X_i\beta}$ (e.g., Wooldridge 2013). This inequality follows from a transformation of the random component of the model (i.e., ϵ_i). Even if the model coefficients β are *known*, then this inequality holds. But researchers can easily avoid this issue by using the correct transformation $\text{E}(y|X_i) = e^{X_i\beta+\frac{\sigma^2}{2}}$ (assuming for simplicity that σ is known).

However, the bias that interests me flows from a transformation of the *model coefficients*.

Even if the researcher uses the correct transformation $\hat{\tau} = \text{E}(\text{income} \mid \text{education} = 20) = e^{\hat{\beta}_{cons} + 20\hat{\beta}_{edu} + \frac{\sigma^2}{2}}$ (again assuming that σ is known), then $\hat{\tau}$ is biased. That is, even when using the correct transformation, unbiased estimates of the model coefficients do not guarantee unbiased quantities of interest.

But how does a simple transformation of unbiased coefficient estimates induce a large bias in the estimate of the quantity of interest? We usually think about bias as occurring in the model coefficients β, so that

$$\text{coefficient bias} = \text{E}(\hat{\beta}) - \beta.$$

But substantive researchers care mostly about bias in the quantities of interest. For convenience, I refer to the bias in the quantities of interest as τ-bias, so that

$$\tau\text{-bias} = \text{E}[\tau(\hat{\beta})] - \tau(\beta).$$

τ-bias is more complex and subtle than biases in the coefficients. It can be rewritten and decomposed into two components: transformation-induced τ-bias and coefficient-induced τ-bias, so that

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

Any bias in the coefficients passes through to the quantities of interest in the sense that, if the coefficient estimates are biased, then the transformation of the true coefficient is not equal to the transformation of the average coefficient estimate, so that

$$\text{coefficient-induced } \tau\text{-bias} = \tau[\text{E}(\hat{\beta})] - \tau(\beta).$$

But the transformation *itself* introduces bias as well, so that

$$\text{transformation-induced } \tau\text{-bias} = \mathrm{E}[\tau(\hat{\beta})] - \tau[\mathrm{E}(\hat{\beta})].$$

Transformation-induced bias occurs because, in general, $h[\mathrm{E}(y)] \neq \mathrm{E}[h(y)]$ for an arbitrary random variable y and function h.

Little methodology research explicitly recognizes this transformation-induced τ-bias and less fully appreciates its practical importance. Methodologists must become more conscientious of transformation-induced bias; τ-bias can be much larger than coefficient-induced bias and disappear more slowly as the sample size increases.

A Characterization

But how can we characterize the direction of this bias? For strictly convex and strictly concave transformations, Jensen's inequality enables a straightforward characterization of the direction of the transformation-induced τ-bias. This characterization also provides the key intuition for more complicated transformations.

Theorem 1 *Suppose a non-degenerate estimator $\hat{\beta}$. Then any strictly convex (concave) τ creates upward (downward) transformation-induced τ-bias.*

Proof The proof follows directly from Jensen's inequality. Suppose that the non-degenerate sampling distribution of $\hat{\beta}$ is given by $S_\beta(b)$ so that $\hat{\beta} \sim S_\beta(b)$. Then $\mathrm{E}(\hat{\beta}) = \int_B bS_\beta(b)db$ and $\mathrm{E}[\tau(\hat{\beta})] = \int_B \tau(b)S_\beta(b)db$. Suppose first that τ is convex. By Jensen's inequality, $\int_B \tau(b)S_\beta(b)db > \tau\left[\int_B bS_\beta(b)db\right]$, which implies that $\mathrm{E}[\tau(\hat{\beta})] > \tau[\mathrm{E}(\hat{\beta})]$. Because $\mathrm{E}[\tau(\hat{\beta})] - \tau[\mathrm{E}(\hat{\beta})] > 0$, the transformation-induced τ-bias is upward. By similar argument, one can show that for any strictly *concave* τ, $\mathrm{E}[\tau(\hat{\beta})] - \tau[\mathrm{E}(\hat{\beta})] > 0$ and that the transformation-induced τ-bias is downward. ■

In general, researchers do not restrict themselves to a strictly convex or strictly concave τ. For example, typical transformations of logistic regression coefficients, such as predicted probabilities, first and second differences, marginal effects, and risk ratios, all have both convex regions and concave regions. This situation is much more difficult to characterize generally because $\tau(b)$ might contain a mixture of convex and concave regions. Making matters even more difficult, at any particular point b, the multivariate function τ might be convex in one direction and concave in another. In general though, the direction of the bias depends on the *location* of the sampling distribution. But the intuition from Theorem 1 is clear. If most of the sampling distribution is located in a mostly concave region, then the bias will be downward. If most of the sampling distribution is located in a mostly convex region, then the bias will be upward.[1]

An Approximation

While the intuition developed by Theorem 1 helps us understand the *direction* of the bias, how can we assess the *magnitude* of the transformation-induced τ-bias? To approximate the magnitude, I use a a second-order Taylor expansion. First, notice that $\mathrm{E}[\tau(\hat{\beta})] = \mathrm{E}[\tau(\mathrm{E}[\hat{\beta}] + (\hat{\beta} - \mathrm{E}[\hat{\beta}]))]$. Now approximate the term inside the right-hand expectation with a second order Taylor expansion, so that

$$E[\tau(\hat{\beta})] \approx \mathrm{E}\left[\tau[\mathrm{E}(\hat{\beta})] + \sum_{r=1}^{k+1} \frac{\partial \tau[\mathrm{E}(\hat{\beta})]}{\partial \beta_r}[\hat{\beta}_r - \mathrm{E}(\hat{\beta}_r)] + \frac{1}{2}\sum_{r=1}^{k+1}\sum_{s=1}^{k+1} \overbrace{\frac{\partial^2 \tau[\mathrm{E}(\hat{\beta})]}{\partial \beta_r \beta_s}}^{\text{Hessian}=H_{rs}} \overbrace{[\hat{\beta}_r - \mathrm{E}(\hat{\beta}_r)][\hat{\beta}_s - \mathrm{E}(\hat{\beta}_s)]}^{\text{Cov}(\hat{\beta}_r,\hat{\beta}_s)=\Sigma_{rs}}\right]$$

[1] One might wonder about the relevance of these ideas to Bayesian analyses. Indeed, the researcher can usually use MCMC to sample directly from posterior of the model coefficients and, by simple extension, sample the quantity of interest from the posterior distribution. But if the researcher uses the posterior mode as the point estimate, then the identical logic applies. For an alternative point estimate (e.g., posterior mean), the invariance principle no longer holds, so the argument breaks down (i.e., the point estimate of $\tau(\beta)$ is not longer $\tau(\hat{\beta})$. However, regardless of the point estimate the researcher uses, a Bayesian approach does not guarantee an unbiased quantity of interest.

Taking the expectation of the right-hand side eliminates the middle term and allows expressing the final term as a function of the variance of the sampling distribution, so that

$$\mathrm{E}[\tau(\hat{\beta})] \approx \tau[\mathrm{E}(\hat{\beta})] + \frac{1}{2}\sum_{r=1}^{k+1}\sum_{s=1}^{k+1} H_{rs}\Sigma_{rs},$$

where H represents the Hessian matrix of second derivatives of τ at the point $\mathrm{E}(\hat{\beta})$ and, conveniently, Σ represents the covariance matrix of the sampling distribution. Rearranging gives an approximation to the magnitude of the transformation-induced τ-bias, so that

$$\text{transformation-induced } \tau\text{-bias} = \mathrm{E}[\tau(\hat{\beta})] - \tau[\mathrm{E}(\hat{\beta})] \approx \frac{1}{2}\sum_{r=1}^{k+1}\sum_{s=1}^{k+1} H_{rs}\Sigma_{rs}.$$

If H is constant then the approximation is exact. If $\hat{\beta}$ is unbiased, then $\tau[\mathrm{E}(\hat{\beta})]$ can be replaced with $\tau(\beta)$, so that Equation 1 represents both transformation-induced and the total τ-bias.

Equation 1 does not depend on a strictly convex or concave transformation. As long as τ is not highly non-linear (e.g., $\left|\frac{\partial^3\tau}{\partial\beta_r\partial\beta_s\partial\beta_t}\right| \approx 0$), then Equation 1 provides a reasonable estimate of the direction and magnitude of the bias.

Equation 1 quantifies two intuitions. First, the amount of bias depends on the standard error or sample size. As the sample size grows large, Σ_{rs} shrinks to zero, which drives the bias to zero as well. This matches the previous observation that $\tau(\hat{\beta})$ is a consistent estimator of $\tau(\beta)$. Secondly, the amount of bias depends on the curvature in τ. If τ is nearly linear so that $H \approx 0$, then the transformation introduces minimal bias. On the other hand, more curvature, so that $H >> 0$, leads to a large bias.

Two Monte Carlo Simulations

The following two Monte Carlo studies illustrate the importance of accounting for transformation-induced bias in Monte Carlo studies of estimators. Approximately unbiased coefficients are not

enough—one must assess the bias in the quantities of interest as well.

A Hypothetical Model

Many substantive researchers realize that logistic regression estimates are biased away from zero in small samples and use "rules of thumb" to judge whether asymptotic properties, such as asymptotic unbiasedness, approximately apply to a finite sample. When non-events outnumber events, one such rule of thumb requires ten events per explanatory variable (Peduzzi et al. 1996). I show that this rule works quite well choosing a sample size that yields approximately unbiased coefficients, but severely underestimates the sample size needed for approximately unbiased estimates of the marginal effects.

For simplicity, I focus on the model $\Pr(y) = \text{logit}^{-1}(\beta_{cons} + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6)$, where y indicates whether or not each observation experiences an event and the x_j represent fixed explanatory variables that I create by simulating from independent, standard normal distributions. For this simulation, I set $\beta_{cons} = -1$ and $\beta_j = 0.15$ for $j \in \{1, ..., 6\}$. I assume that "approximately unbiased" means a bias of less than three percent, where

$$\text{percent bias} = 100 \times \frac{E[\tau(\hat{\beta})] - \tau(\beta)}{\tau(\beta)}. \tag{1}$$

I vary number of observations N from 100 to 2,500, and, for each sample size, I simulate 100,000 data sets, use each data set to estimate the coefficients, and use the estimated coefficients to calculate the marginal effects. I use these 100,000 estimates to calculate the percent bias given by Equation 1.

Figure 1 shows the bias in the coefficients as the sample size increases. The left panel shows the bias in $\hat{\beta}_{cons}$ and the right panel shows the bias $\hat{\beta}_1$. For $N = 100$, $\hat{\beta}_{cons}$ and $\hat{\beta}_1$ are biased away from zero by about ten percent. However, this bias drops to about three percent for $N = 250$ and nearly disappears for $N = 2,500$. The rule of thumb works well; the bias is

negligible for about $N = 219$.

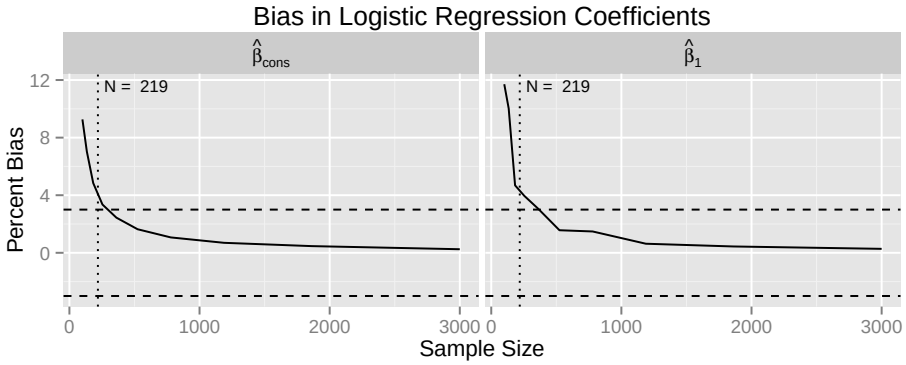

Figure 1: This figure shows the percent bias for the intercept and coefficient for x_1. The rule of thumb requiring ten events per explanatory variable suggests a minimum sample size of about 219. For samples larger than about 250, the bias falls below three percent and it nearly disappears as the sample size approach 2,500.

Figure 2 shows the bias in the estimates of the marginal effects as the sample size increases. The left panel shows the total bias, the middle panel shows the coefficient-induced bias, and the right panel shows the transformation-induced bias. Since the marginal effect of x_1 varies with x_1 itself, I plot the estimates for a range of values of x_1.

Two features stand out. First, small sample bias is much larger for the marginal effects than for the coefficients. For $N = 100$, the estimate of the marginal effect is biased by about -75% for $x_1 = -3$, -50% for $x_1 = -2$, and -25% for $x_1 = -1$. Second, the small sample bias in the estimates of the marginal effects descends to zero more slowly than the coefficient estimates. While the coefficient estimates are approximately unbiased for about $N = 250$, the estimates of the marginal effects retain substantial bias. Indeed, the bias in the estimates of the marginal effects drops below the 3% threshold at about $N = 2,500$—more than ten times the rule of thumb that works well for the coefficients.

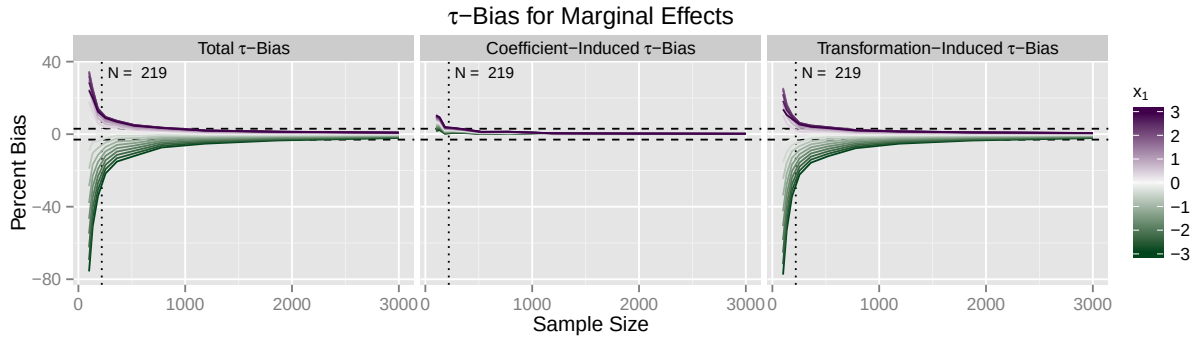

Figure 2: This figure shows the total, coefficient-induced, and transformation-induced τ-bias for the marginal effects. The rule of thumb requiring ten events per explanatory variable suggests a minimum sample size of about 219. However, the bias falls well outside the three percent threshold for this suggested sample size. The estimates fall within the three percent threshold only for sample sizes nearing 2,500—more than ten times the rule of thumb that works well for the coefficients. Also notice that while the coefficient-induced bias receives the most attention from methodologists, the transformation-induced bias is *much* larger.

An Actual Model

To further highlight the practical implications of transformation-induced τ-bias, I use the coefficients reported for Model 1 in Table 1 of Fearon and Laitin (2003, p. 84) and their explanatory variables to conduct a second simulation. Fearon and Laitin use a logit model to assess several hypotheses about the causes of civil war. Using their reported coefficients as the true model parameters and their explanatory variables the predictors (6,327 observations; 106 events; 11 explanatory variables), I repeatedly (1) simulate a new outcome variable, (2) estimate the logit model coefficients with ML, and (3) calculate the quantities of interest. For each observed case, I calculate the predicted probability of a civil war, the marginal effect of a change in per capita income, and the change in probability if per capita income increased by $4,500 (about one standard deviation). Using these simulations, I calculate the total, coefficient-induced, and transformation-induced τ-bias for the three quantities of interest. Figure 3 shows the total and decomposed τ-bais.

This example highlights that the relative importance of coefficient-induced τ-bias and

transformation-induced τ-bias depends on the substantive problem. The top row of Figure 3 shows that the coefficient-induced τ-bias pushes the estimated predicted probabilities downard, while the transformation-induced τ-bias pushes the estimate upward. For most cases, the two biases roughly cancel, though for larger true probabilities, the transformation-induced bias tends to be larger, producing an upward bias in the predicted probabilities. The middle row of Figure 3 shows that, for the marginal effects, the transformation-induced τ-bias tends to overwhelm the coefficient-induced τ-bias. The final row of Figure 3 shows that the opposite holds for the first difference, where almost all of the total τ-bias is due to bias in the estimates of the coefficients.

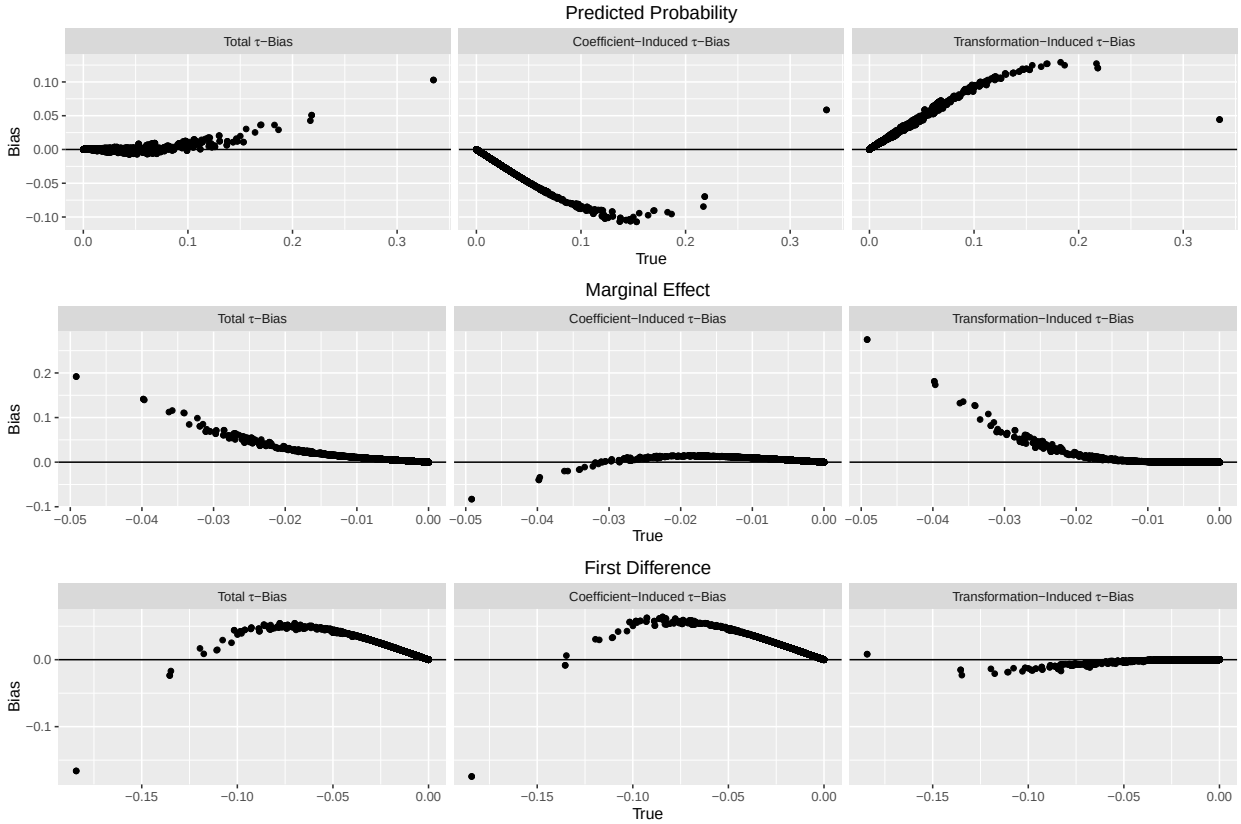

Figure 3: This figure shows the total, coefficient-induced, and transformation-induced τ-bias for three quantities of interest. Each point represents a single observation from Fearon and Laitin's (2003) data set. For each observed case, I calculate the predicted probability of a civil war, the marginal effect of a change in per capita income, and the change in probability if per capita income increased by \$4,500 (about one standard deviation).

The Implications

Quantities of interest do not inherit the small sample properties of the coefficient estimates. This fact has important implications for how we study the small sample properties of estimators.

First, τ-bias has important implications for the sample sizes that methodologists recommend to substantive researchers. Methodologists usually parameterize models so that the coefficients lie in an unbounded space. This allows the coefficient estimates to rapidly approach their asymptotic distribution, which ensures the estimates have acceptable small sample properties. Substantive researchers, though, usually transform these coefficient estimates into a quantity of interest, which, because it often lies in a bounded space, might approach its asymptotic distribution more slowly. As a result, substantive researchers might need much larger sample sizes than methodologists usually recommend. Methodologists must remain conscientious of the quantities of interest to substantive researchers and assess the performance of their estimators in terms of these quantities.

Secondly, τ-bias has important implications for the bias-variance tradeoff in choosing an estimator. Methodologists usually recognize a tradeoff between bias and variance in estimating parameters. Actions intended to remove bias might increase variance and vice versa. However, the approximation to the transformation-induced bias given in Equation 1 points out an important result. Greater variance in the coefficient estimates might lead to increased bias in the quantities of interest. This implies that if an estimator is essentially unbiased, then greater efficiency translates to reduced bias in the quantities of interest. Similarly, small reductions in bias at the expense of large loss in efficiency might lead to greater bias in the quantities of interest. For example, refinements of the usual logit model intended to *reduce* bias in the coefficients, such as heteroskedastic probit or scobit, might actually *increase* bias in the quantities of interest. Methodologists must be aware of this tradeoff when recommending more sophisticated complex estimators to substantive researchers and comparing alternative estimators.

As methodologists, we cannot ignore transformation induced bias. Nearly unbiased es-

timates of coefficients are not enough. We must be aware of the quantities of interest to substantive researchers and calibrate our tools for these quantities.

References

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

Fearon, James D., and David D. Laitin. 2003. "Ethnicity, Insurgency, and Civil War." *American Political Science Review* 97(1):75–90.

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.

Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Advanced Quantitative Techniques in the Social Sciences Thousand Oaks, CA: Sage.

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

Nieman, Mark David. 2015. "Statistical Analysis of Strategic Interaction with Unobserved Player Actions: Introducing a Strategic Probit with Partial Observability." Forthcoming in *Political Analysis*. Manuscript at http://www.marknieman.net/storage/PA_strat_2015.pdf.

Peduzzi, Peter, John Concato, Elizabeth Kemper, Theodore R Holford, and Alvan R Feinstein. 1996. "A Simulation Study of the Number of Events per Variable in Logistic Regression Analysis." *Journal of Clinical Epidemiology* 49(12):1373–1379.

Wooldridge, Jeffrey M. 2013. *Introductory Econometrics: A Modern Approach*. 5th ed. Mason, Ohio: South-Western Cengage Learning.