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 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 collection of covariates X_i by alink function g, so that $g(\theta_i) = X_i\beta$. The researcher usually estimates β with maximum likelihood (ML), and, depending on the choice of g and f, these estimates possess might have desirable small sample desirable properties. ML does not produce unbiased estimates in general, though. For this reason, methodologists frequently use Monte Carlo simulations to assess small sample properties of estimators and provide users with rules of thumb about appropriate sample sizes.

Although methodologists tend to focus on estimating the model coefficients, substantive researchers tend to focus on "quantities of interest" (King, Tomz, and Wittenberg 2000)–*transformations* τ of the model coefficients. Examples include marginal effects, first and second differences, predicted probabilities and expected values, and risk ratios. Fortunately, the invariance properties allows one to calculate estimates of the quantities of interest from

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

‡Carlisle Rainey is Assistant Professor of Political Science, University at Buffalo, SUNY, 520 Park Hall, Buffalo, NY 14260 (rcrainey@buffalo.edu).

the coefficient estimates. The invariance property states that if $\hat{\theta}$ is the ML estimate of θ, then for any function τ, the ML estimate of $\tau(\beta)$ is $\tau(\hat{\theta})$ (King 1998, pp. 75-76, and Casella and Berger 2002, pp. 320-321). Of course, if $\hat{\theta}$ is a consistent estimator of θ, then $\tau(\theta)$ is a consistent estimator $\tau(\hat{\theta})$. But the invariance principle raises an important question: Does $\tau(\hat{\theta})$ necessarily inherit desirable small sample properties of $\hat{\theta}$, such as 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, but crucial point reveals a disconnect between the work done by methodologists, which tends to focus on coefficients, and the work done by substantive scholars, which tends to focus on a transformations of the coefficients. Much (though certainly not all) of the methodological research that we do 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 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 focuses his illustrative application on predicted probability of civil war. In order to provide more compelling tools for substantive scholars, we must extend our focus beyond coefficients 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, income is measured in thousands of dollars. Assuming that we use the correct model, then least squares (the ML estimator) provides the best unbiased estimator of the coefficients β_{cons} and β_{edu}. However, we are not nterested in log(income) directly. Instead, we are interested in income itself, in particular, the median income among those with 20 years of education med(income|education $= 20$) $= e^{\beta_{cons}+20\beta_{edu}}$. 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 quantity of interest, I simulate 100,000 data sets and calculate the quantity of interest for each. Although $\hat{\beta}_{cons}$ and $\hat{\beta}_{edu}$ are unbiased, the estimate of med(income|education $= 20$) is strongly biased upward, with an expected value of about $\$106k$. But how does this simple transformation of unbiased estimates of the coefficients induce such 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

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

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

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

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

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

Any coefficient bias passes through to the quantities of interest in the sense that, if the coeffi-

cient estimates are biased, then the transformation of the true coefficient is not equal to the transformation of the average coefficient estimate if the, so that

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

Because $g[\mathrm{E}(X)] \neq \mathrm{E}[g(X)]$ in general, But the transformation *itself* introduces bias as well, so that

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

The bias occurs because, in general, $g[\mathrm{E}(X)] \neq \mathrm{E}[g(X)]$ for an arbitrary random variable X and function g.

Little methodology research explicitly recognizes this transformation-induced τ-bias and less fully appreciates its importance. Below I offer a general characterization using Jensen's inequality, and approximation to its magnitude, an example that illustrates its significance, and a discussion of its importance to methodological research.

A Characterization

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 as well as approximation of the magnitude of the bias.

Theorem 1 *Suppose a generic (non-degenerate) ML 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,

$\int_B \tau(b)S_{\hat{\beta}}(b)db > \tau\left[\int_B bS_{\hat{\beta}}(b)db\right]$, 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 generically, given that $\tau(b)$ might contain a mixture of strictly convex and strictly concave regions or 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. 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.

An Approximation

Next, I approximate the magnitude of the the transformation-induced τ-bias using 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

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

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

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

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

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

If H is constant then the approximation is exact.

Equation 1 does not depend on a strictly convex or concave transformation. The approximation is reasonable if $H(\beta)$ resembles the curvature of τ over most of the sampling distribution. As long as τ is not highly non-linear (e.g., $\left|\frac{\partial^3 \tau}{\partial \beta_r \partial \beta_s \partial \beta_t}\right| >> 0$), 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 and/or sample size. As the sample size grows large, Σ shrinks to zero, which drives the bias to zero as well. This observation matches the fact 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 greater bias.

An Example

Many substantive researchers realize that logistic regression estimates are biased away from zero in small samples. However, fewer realize that a simple penalty applied to the likelihood function can nearly eliminate this bias (Firth 1993). In this example, I illustrate the importance of accounting for transformation-induced bias. I show that obtaining unbiased coefficients can be counterproductive if researchers instead focus on marginal effects.

For concreteness, suppose a model explaining the probability of voting as a function of

education

$$\Pr(\text{vote}_i) = \text{logit}^{-1(}\beta_{cons} + \beta_{edu}\text{education}_i),$$

where *vote* indicates whether or not citizen i voted in the election and *education* is measured in years. Let $\beta_{cons} = -2.5$, $\beta_{edu} = 0.2$, and $N = 30$. We are interested in the marginal effect of education. In this case, we might want to calculate the marginal effect of education at a substantively relevant value (e.g., *education* $= 12$ years or *education* $= 16$ years), at all observed values, or perhaps average across the observed data (Hanmer and Kalkan 2013).

To assess the bias in the estimators in the coefficients and the marginal effects, I first create a hypothetical variable *education* that takes on 30 integer values roughly uniformly distributed from 10 to 20. I then simulate 100,000 data sets and compute the coefficients and marginal effects for each data set using ML and penalized ML estimation. As expected, the ML coefficient estimates (intercept and slope) are substantially biased away from zero by about 14%, while the penalized ML estimates are only biased away from zero by about 1%.

At first glance, one might guess that the $\tau(\hat{\beta}^{mle})$ would provide a less biased estimate of $\tau(\beta)$ given the substantial bias in the ML coefficient estimates. However, that does not follow. A *more biased* estimate of β might lead to a less biased estimate of $\tau(\beta)$ if the coefficient- and transformation-induced τ-bias cancel each other out. Indeed, the two biases do cancel out in this hypothetical example. Figure 1 shows that, when calculated from the nearly unbiased $\hat{\beta}^{pmle}$, the estimate of the marginal effect of education on the probability of voting is *always* biased downward. On the other hand, marginal effects calculated using the biased coefficient estimate $\hat{\beta}^{mle}$ are sometimes biased downward, sometimes biased upward, and other times biased little. For six of eleven values of education, the biased ML coefficient estimates lead to less biased marginal effects. The nearly unbiased penalized ML coefficient estimates lead to less biased marginal effects in only four cases. The remaining case is essentially a tie.

If we instead focused only on the marginal effect at *education* $= 12$ or *education* $= 16$, then

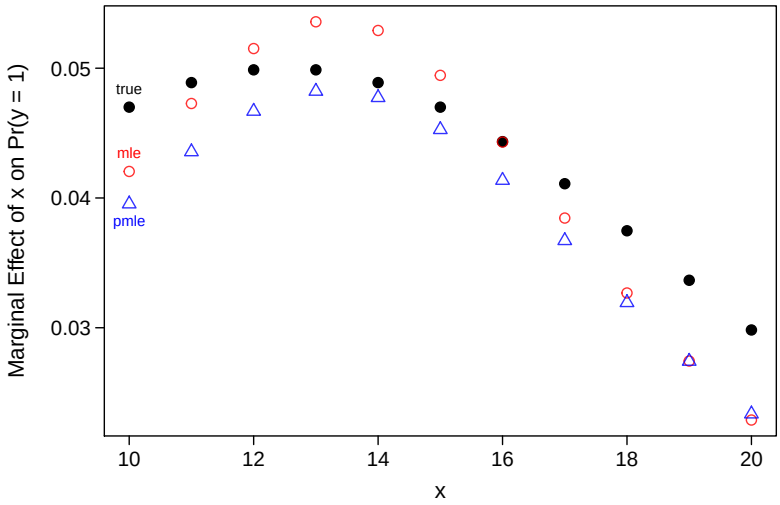

Figure 1: This figure shows the true marginal effects (solid black circles) and the expected value of the marginal effect estimates calculated from ML coefficient estimates (red circles) and penalized ML estimates (blue triangles). Although the ML coefficient estimates are strongly biased away from zero, they provide less biased estimates of the marginal effects for more than half of the observed values of *education*. And while the penalized ML estimates of the coefficients are essentially unbiased, they provide less biased estimates of the marginal effects in only four of the eleven cases and *always* underestimate the marginal effects.

the biased ML coefficient estimates still provides a less biased estimate of the marginal effect. If we follow Hanmer and Kalkan (2013) and average across the marginal effects for the observed data, then the essentially unbiased penalized ML coefficient estimates produce a downward bias of about 10%. The strongly biased ML coefficient estimates, on the other hand, produce a downward bias of only about 2%.

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, it 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 quantities of interest, and the estimate of the quantity of interest does not inherit the small sample properties of the coefficient estimators. Because the quantity of interest often lies in a bounded space, it might approach its asymptotic distribution more slowly. As a result, substantive researchers might need much larger sample sizes than methodologists usually recommend.

Secondly, it has important implications for the bias-variance tradeoff in choosing an estimator. Methodologists usually recognize a tradeoff between bias and variance in estimating parameters. However, the approximation to the transformation-induced bias given in Equation 1 points out an important fact. Greater variance in the coefficient estimates 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.

As methodologists, we cannot ignore transformation induced bias. Nearly unbiased estimators 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, California: Duxbury.

Firth, David. 1993. "Bias Reduction of Maximum Likelihood Estimates." *Biometrika* 80(1):27–38.

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.

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.

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.