# Analyzing Second-Stage Ecological Regressions: Comment on Herron and Shotts

**Christopher Adolph and Gary King**
*Department of Government, Harvard University,*
*Cambridge, MA 02138*
*e-mail: cadolph@fas.harvard.edu*
*e-mail: king@harvard.edu*

## 1 Introduction

We take this opportunity to comment on Herron and Shotts (2003; hereinafter HS) because of its interesting and productive ideas and because of the potential to affect the way a considerable body of practical research is conducted. This article, and the literature referenced therein, is based on the suggestions in three paragraphs in King (1997, pp. 289–290). Because these paragraphs were not summarized in HS, we thought they might be a useful place to start.

> If a second stage analysis is conducted, least squares regression should probably not be used in most cases, even though it may not be particularly misleading. The best first approach is usually to display a scatterplot of the explanatory variable (or variables) horizontally and (say) an estimate of $\beta_i^b$ or $\beta_i^w$ vertically. In many cases, this plot will be sufficient evidence to complete the second stage analysis.
>
> If it proves useful to have more of a formal statistical approach, and many of the actual values of $\beta_i^b$ fall near zero or one, then some method should be used that takes this into account. The data could be transformed, via a logit or probit transformation, or [the "extended model"] could be applied. . . . Whatever method is chosen, the researcher should be careful to include the fact that some estimates of $\beta_i^b$ are more uncertain than others.
>
> In practice, a weighted least squares linear regression may be sufficient in many applications, with weights based on the standard error of $\beta_i^b$ (or other quantity of interest). Researchers should be careful in applying this simplified method here, and should verify its assumptions with scatterplots. . . . This is not as theoretically elegant a procedure as the more formal set up in [the "extended model"] but it is simple, relatively robust, and probably complete enough to be of use in many applications.

Clearly, the only logically consistent model that has been offered for the issue at hand is the EI extended model that allows the second-stage covariates to be included in the EI estimation procedure. EI software now includes a feature that allows for first differences to be computed if those covariates are included, which should make the extended model somewhat easier to understand and use. (We discuss how to use the extended model in Section 6.)

Any second-stage analysis is by definition a second-best procedure when judged conditional on the model. Still, second-stage analyses—if they give approximately correct answers—have the advantage of being easier to understand, use, and evaluate (because

*Authors' note:* Our thanks to Michael Herron and Kenneth Shotts for helpful comments on an earlier draft.

an estimate of the dependent variable in the second stage is observable); more numerically stable (hence allowing more covariates to be included and studied); computationally faster (allowing more analyses to be examined); more amenable to diagnostics in verifying assumptions (because the actual estimated data points can be observed and the fit can be checked directly); and possibly more robust to certain types of misspecification. The question at hand is if and when such a procedure can be used to produce answers that are approximately correct.

HS's contribution is in pointing out that precinct-level estimates from EI regress to the mean. This "shrinkage" property is indeed a characteristic of EI, and it is also a characteristic of every other Bayesian model. Shrinkage results in optimal estimates, that is, with the smallest possible mean square error. Thus, we agree with the implication of HS's article that the best possible estimate of $\beta_i^b$ (e.g., the fraction of blacks who vote), under HS's assumptions, is that produced by EI. However, HS also make the interesting and correct point that using Bayesian mean posterior estimates with this property, like those given by EI, as dependent variables in least squares (LS) regression can, under some circumstances, produce biased estimates. No one disputes that second-stage regressions are approximations or that they cause problems from some theoretical perspectives; the question at hand is when they cause problems that affect real applied research.

HS study second-stage regression based on LS and conclude that their bias-adjusted slope is preferable. The article does not mention whether the constant term should be adjusted. Their slope adjustment is based on linear approximations within the classic linear econometric theory framework. The problems with HS's approach that we discuss here concern assuming linearity when modeling an inherently nonlinear and bounded relationship (i.e., their algebra by definition miss the information in the bounds), omitting the constant correction from the paper, never computing the correction they propose when conducting simulations, and missing the fact that weighted least squares (WLS) corrects problems they raise. We show how filling in this missing information leads back to the suggestions from the paragraphs quoted previously.

In addition to making second-stage analyses possible by providing the first precinct-level estimates from an ecological inference model, the primary advance of EI was in resolving the half-century long debate between supporters of Goodman's (1959) unbounded linear regression approach and Duncan and Davis's (1953) method of bounds by incorporating all information from both into the same model. HS fall back to Goodman's unbounded approach, and so miss the highly informative deterministic information in the aggregate data.

We replicated HS's simulations without trouble. We then examined how well their adjusted regression procedure fit the observed data based on the estimated $\hat{\beta}_i^b$ and the true (normally unobserved) $\beta_i^b$. We find that correcting the slope but not the constant is considerably worse than an unadjusted regression of $\hat{\beta}_i^b$ on $Z_i$ and the other procedures discussed in HS's paper and the literature we examined. This is true even if we knew the true value of the adjustment. Unless $Z$ has zero mean, the regression line from this method tends to miss the cloud of true $\beta_i^b$ points by a wide margin.

We therefore begin by extending HS's slope adjustment procedure, within their linear econometric framework, by developing an adjustment for the constant term. We follow the procedure that we believe Herron and Shotts would have used if they had tried.[1] As it turns out, this fully adjusted second-stage regression method dominates

[1] Indeed, after we finished a draft, Herron and Shotts told us that they had the same constant correction in a draft version of their paper, but took it out in the final version to save space. Because HS suggest correcting the slope,

the slope-only procedure. Indeed, the partially adjusted procedure in HS is never called for.

Because the only estimators HS offer for their partial adjustment procedure assume either knowledge of the truth being estimated or that some unknown parameters can be estimated without error, we develop an estimator without these flaws and apply it to create full adjustments. We find that this fully empirical version of the full adjustment procedure is, with two partial exceptions, dominated by (unadjusted) WLS. For one, when the bounds are wide, the true adjustment factor could make a noticable difference, but the adjustment itself cannot be estimated reliably, and so the procedure cannot be applied. However, in this case, of course, researchers should not be running EI in the first place because of extreme model dependence. For the other, nonlinear relationships obviously cannot be well modeled by any linear second-stage regression, and so in that case, a logistic (or other nonlinear) regression procedure is best. Thus, the only situation in which the adjustment would lead to improvements is when the bounds are wide and the true value of the adjustment is known, which describes the Monte Carlo procedures HS used to evaluate their method, but of course not any real application. (Even in this unrealistic situation, the full correction takes an hour to run, compared to a few seconds for WLS.) Because researchers can easily detect which situation applies from the available aggregate data, they can always take appropriate action.

We conclude that researchers with narrow enough bounds to run first-stage EI should first examine a scatterplot of the covariate horizontally by the *bounds* on $\hat{\beta}_i^b$ vertically (such as King, 1997, Fig. 13.2, p. 238) because this scatterplot requires no assumptions at all. We can sometimes learn a lot from such a graph, including especially a check for nonlinearities (which normally occur when the points are near zero or one). If nonlinearities are not apparent, then WLS should be used because it offers approximately unbiased estimates in second-stage regressions. We also suggest a scatterplot of the covariate horizontally by the estimate $\hat{\beta}_i^b$ vertically to examine the data being run.

## 2 A Fully Adjusted Second-Stage Regression Model

We try to stick to HS's notation where possible and present the fully adjusted method with their adjustment as a special case.[2]

First, let the expectation of $\beta_i^b$ conditional on $Z_i$ be approximated by

$$E\left(\beta_i^b\right) = \alpha_R + \gamma_R Z_i, \tag{1}$$

so that estimates of $\alpha_R$ and $\gamma_R$ are the immediate goal of the analysis. Also, let the expectation of $\hat{\beta}_i^b$ conditional on $Z_i$ be

$$E\left(\hat{\beta}_i^b\right) = \alpha_U + \gamma_U Z_i, \tag{2}$$

and do not mention constant term corrections, when we refer to the *HS partially adjusted procedure*, we are referring to what readers would conclude that Herron and Shotts advocate, even though we now know (from our subsequent correspondence) that they also believe the corrections discussed in their paper are incomplete for real applications.

[2] Our notation deviates from HS only to ensure logical consistency. For example, their Eqs. (7) and (8) have different dependent variables, $\beta_i^b$ and $\hat{\beta}_i^b$, set equal to the same entire right side of the equation in both cases ($\alpha + \gamma' Z_i + \nu_i$). Later in their paper, they allow $\gamma$ to differ between the two (calling the first $\gamma_R$ and the second $\gamma_U$), but do not change the constant term (or error terms). We fix these issues and others.

which is, of course, well estimated by LS. HS then assumes that the error in estimating $\hat{\beta}_i^b$ by EI is a linear function of $\beta_i^b$.

$$E\left(\hat{\beta}_i^b - \beta_i^b\right) = \delta_0^b + \delta_1^b \beta_i^b. \tag{3}$$

Then solving Eq. (3) for $E(\hat{\beta}_i^b)$ and substituting the result into Eq. (2) gives a more informative version of Eq. (1):

$$E\left(\beta_i^b\right) = \left(\frac{\alpha_U - \delta_0^b}{1 + \delta_1^b}\right) + \left(\frac{\gamma_U}{1 + \delta_1^b}\right) Z_i. \tag{4}$$

Thus, we know that the quantities of interest, $\alpha_R$ and $\gamma_R$, can be expressed as the intercept and slope of Eq. (4). If we can estimate the components of each, we can derive a consistent estimator, at least when HS's linearity and unboundedness assumptions are not too far off.

## 3 An Improved Estimation Procedure

HS offer an estimation algorithm for correcting the slope term in their Section 7.2. This procedure is intuitive, and we can easily generalize it to provide a correction for the intercept as well. Unfortunately, the procedure itself is flawed for two other reasons. First, it conditions on the point estimate for the parameters of the truncated bivariate normal, $\breve{\psi}$, and of $\delta_0^b$ and $\delta_1^b$, and thus assumes the absence of estimation uncertainty. Ignoring uncertainty would bias standard errors and confidence intervals, of course, which perhaps is why HS do not calculate these. However, because their estimation procedure is nonlinear [because of the ratios in Eq. (4)], ignoring estimation uncertainty also affects their point estimates in finite samples. Second, the procedure calls for drawing only a single simulation of $(\beta_i^b, \beta_i^w)$ for each observation. As a result, the estimate includes substantial Monte Carlo approximation error. The error can be eliminated by running their entire procedure many times and averaging. Although fixing these problems would not have substantially changed the estimates presented in their paper, they matter in some applications. We therefore develop and use a new estimation algorithm that corrects these problems, as well as provides the ability to compute standard errors and confidence intervals, which were not available in HS's version.[3]

The accurate procedure that fully represents the uncertainty of HS's corrections is computationally slow (taking about an hour to do one run). Also, the standard errors for the fully adjusted method appear larger than those obtained from LS by about 70% and WLS by about 30%. For the intercepts, which under full adjustment combine the uncertainty of three parameters, the standard errors are, on average, 34 times larger than LS and 25 times larger than WLS. Thus, any reduction in bias that may occur is probably outweighed by

[3] To define our revised estimation algorithm, let a symbol with a tilde ($\sim$) denote a value of that quantity randomly drawn from its posterior density. For a given $X$ and $T$: (1) run EI on $X$ and $T$; (2) regress $\hat{\beta}_i^b$ (which comes from EI) on $Z_i$, yielding the estimated intercept $\hat{\alpha}_U$ and slope $\hat{\gamma}_U$; (3) draw $\tilde{\breve{\psi}}$ from its posterior provided by EI; (4) take $p$ draws of $(\tilde{\beta}_i^b, \tilde{\beta}_i^w)$ from a truncated normal density with parameter vector $\tilde{\breve{\psi}}$; (5) compute a new $T_i$ as $\tilde{T}_i = \tilde{\beta}_i^b X_i + \tilde{\beta}_i^w (1 - X_i)$; (6) run EI on $X_i$ and $\tilde{T}_i$ to yield estimates $(\tilde{\hat{\beta}}_i^b, \tilde{\hat{\beta}}_i^w)$ for all $i$; (7) regress $(\tilde{\hat{\beta}}_i^b - \beta_i^b)$ on $\beta_i^b$ to estimate $\delta_0^b$ and $\delta_1^b$, which we label $\tilde{\hat{\delta}}_0$ and $\tilde{\hat{\delta}}_1$; (8) draw $\tilde{\alpha}_U$ and $\tilde{\gamma}_U$ from the posterior provided by LS; (9) compute the adjusted intercept as $(\tilde{\alpha}_U - \tilde{\hat{\delta}}_0^b)/(1 + \tilde{\hat{\delta}}_0^b)$ and adjusted slope as $\tilde{\gamma}_U/(1 + \tilde{\hat{\delta}}_i^b)$; (10) repeat steps (3)–(9) a sufficient number of times to eliminate Monte Carlo approximation error; and (11) average the simulations in step (10) to get point estimates, take their standard deviation for standard errors, or sort them and use percentile values for confidence intervals.

the substantial increase in variance. This is an intuitive result, given that the bias adjustment takes the form of a ratio, and both the numerator (which is LS) and denominator contain estimation uncertainty. (We confirmed this result with a small number of Monte Carlo experiments.) Is is also consistent with the experience of others trying to create bias adjustments in a variety of models outside the field of ecological inference.

## 4 HS's Monte Carlo Simulation Procedure

We explain in this section that HS's Monte Carlo procedure is inappropriate for evaluating second-stage regressions. The procedure is to draw the true value of the dependent variable, $\beta_i^b$, from the EI model without covariates. Then they create the covariate for the second-stage explanatory variable $Z_i$ endogenously as equal to $\beta_i^b$ plus random noise. This procedure has three flaws.

The first flaw is that the Monte Carlo procedure can be interpreted in three logicially inconsistent ways and, although HS do not discuss which interpretation was intended, none of the three make the procedure valid. The first, and in our view most plausible, interpretation is that by creating $Z_i$ with random error, the procedure induces immense errors-in-variables attenuation bias, quite apart from any attenuation bias that may occur as a result of the Bayesian shrinkage in the EI estimate.

This problem can be seen clearly by studying the parameters of the model HS created from which to draw their Monte Carlo data. In this model, the slope of the coefficient on the covariate in the second-stage regression is 1 (in their notation, $\gamma_R = 1$).[4] However, the estimates of this slope from their simulations of the *true* $\beta_i^b$ (i.e., without any attenuation bias in the dependent variable at all) regressed on $Z$ gives a drastically biased estimate. This slope estimate does not appear in their article, but we were able to replicate their Table 2 exactly, and in our Table 1 present these numbers. As can be seen, whereas the theoretical value of $\gamma_R$ is 1 according to this interpretation, in each case their estimates indicate that $\hat{\gamma}_R$ is never larger than 0.19 and on average about 0.07. Thus, because the unadjusted method is unable to recover the coefficients without shrinkage when using the true value of the dependent variable, this Monte Carlo setup is inappropriate for assessing a dependent variable that is estimated (by EI or otherwise).

A second interpretation of the HS Monte Carlo procedure is that $\gamma_R$ is the causal effect of $Z_i$ on $\beta_i^b$. Because many second-stage regressions are designed to be causal, this is often the most appropriate intepretation. Because $Z_i$ was created endogenously, no matter how one exogenously changes $Z_i$, $\beta_i^b$ will not budge, and hence, by this interpretation, the Monte Carlo procedure is setting the causal effect to zero: $\gamma_R = 0$. Thus, because the numbers in Table 1 are uniformly greater than zero, we know that regressing the true $\beta_i^b$ on $Z_i$ overestimates $\gamma_R$. Therefore, because the unadjusted method is unable to recover the true coefficients when using the true value of the dependent variable (i.e., even though no shrinkage occurs or estimation error in the dependent variable exists), the Monte Carlo setup under this second interpretation is also inappropriate for assessing a dependent variable that is estimated (by EI or otherwise).

A third way to interpret the true value in HS's simulations is conditional on *each* randomly generated set of $\tau_i$'s ($i = 1, \ldots, n$), and hence conditional on each randomly generated $Z_i$. By this interpretation, each random draw of a set of $n$ observations from the Monte Carlo

[4] Under this first interpretation of the HS Monte Carlo setup, $E(\beta_i^b) = \alpha_R + \gamma_R Z_i$, where $Z_i = \beta_i^b + \tau_i$ and $E(\tau_i) = 0$ and so $E(Z_i \mid \beta_i^b) = \beta_i^b$. Hence $E(\beta_i^b) = \alpha_R + \gamma_R E(\beta_i^b + \tau_i) = \alpha_R + \gamma_R \beta_i^b$, which implies that $\alpha_R = 0$ and $\gamma_R = 1$.

**Table 1** Comparing the true slope on $Z$ in a second-stage regression in the HS Monte Carlo procedure with the estimate based on $\beta_i^b$ (rather than $\hat{\beta}_i^b$) as the dependent variable

| *Model parameters* | *HS estimates* |
|---|---|
| $\breve{\psi}$ | $\hat{\gamma}_R$ |
| (0.5, 0.5, 0.1, 0.1, 0) | 0.04 |
| (0.75, 0.5, 0.1, 0.1, 0) | 0.04 |
| (0.75, 0.75, 0.1, 0.1, 0) | 0.04 |
| (0.9, 0.9, 0.1, 0.1, 0) | 0.02 |
| (0.5, 0.5, 0.32, 0.1, 0) | 0.19 |
| (0.6, 0.6, 0.1, 0.32, 0) | 0.04 |
| (0.9, 0.1, 0.32, 0.32, 0) | 0.15 |
| (0.5, 0.5, 0.1, 0.1, 0.3) | 0.04 |

*Note*. The true value of $\gamma_R$ depends on the interpretation used (described in the text). This table replicates the results of simulations presented in HS's Table 2. All results are averages over 100 simulations. The difference and ratios presented in HS's Table 2 were successfully replicated and are not shown here.

data generating process produces a *different* true value of $\gamma_R$, the value of which is neither set nor observed by the researcher. This value can be estimated by a LS regression of the true $\beta_i^b$ on $Z_i$, and it can also be estimated by EI-R. Even if we assume that the LS estimator is better (because it uses the true $\beta_i^b$), comparing the two estimators does not reveal which is closer to the unknown $\gamma_R$. The Monte Carlo procedure can reveal only how close the estimators are to each other because the target is unknown and changing over iterations and the estimators are correlated. By this interpretation, of course, the procedure would miss the whole point of running Monte Carlo experiments in the first place—creating a world in which we know the true quantity being estimated and then seeing how good an estimator is at recovering the known parameter value.

Although by this last interpretation we cannot know the different true $\gamma_R$ in each iteration, we can compute its expected value, that is, the estimand implied by the LS estimator of $\beta_i^b$ on $Z_i$. However, computing the average $\gamma_R$ requires nonlinear approximations,[5] which in any event were not computed in HS and so the result was not offered as a standard of comparison. Because the Monte Carlo design under this interpretation bears little resemblence to how second-stage regressions are normally thought of, the standard would have little relevance. In any event, the advantage of proper Monte Carlo experiments is that these after-the-fact guesses can be avoided from the start.

Whichever interpretation one has of the HS Monte Carlo procedure, it also suffers from two other serious problems. First, it artificially rules out the possibility of nonlinear relationships between $\beta_i^b$ and $Z_i$ created by the bounds. That is, by constructing the explanatory variable, $Z_i$, from the bounded $\beta_i^b$'s (plus a normal disturbance), the procedure artificially restricts the relationship between the two to be linear and never affected by the [0,1] bounds

[5] That is, under this third interpretation, $\gamma_R = C(\beta_i^b, Z_i)/V(Z_i) = C(\beta_i^b, \beta_i^b + \tau_i)/[V(\beta_i^b) + V(\tau_i)] = [1 + 1/4V(\beta_i^b)]^{-1}$, because $V(\tau_i)$ was set at 1/4. In this expression, $V(\beta_i^b)$, in turn, is a nonlinear function of the parameters of the truncated bivariate normal.

on $\beta_i^b$. Within this framework, testing the robustness of the HS adjustment to the sort of nonlinear relationships that crop up sometimes in applied research is impossible.

Finally, *HS's Monte Carlo simulations do not test the actual adjustment procedure they propose*. All the results they present in their Section 6 rely on the *true* $\beta_i^b$'s to estimate $\delta_0^b$ and $\delta_1^b$, even though this is the one quantity users of second-stage regressions by definition lack. Appropriately, HS recommend a different estimation procedure (in their Section 7.2) for use when $\beta_i^b$ is unknown, but leave this procedure untested. Hence, HS's article offers no direct evidence that adjustment would be less biased or more efficient than unadjusted LS in practice.

## 5 An Improved Monte Carlo Simulation

Most of our conclusions that follow do not depend on changing the simulation method, but we do so in order to make the results more coherent. To simulate, we follow the logic of the extended EI model. Thus, we first fix $X$, the covariates $Z$ (to values uncorrelated with $X$), the values for the intercept and the slope parameter on $Z$, and the variance and covariance parameters of the truncated bivariate normal. Then, for each simulation, we draw the $\beta_i^b$'s from the extended EI model conditional on $X_i$ and $Z_i$ (without mean centering). The assumption that $X_i$ and $Z_i$ are uncorrelated enables us to run a (first-stage) basic EI model (i.e., with no covariates) without inducing aggregation bias. (The assumption is, of course, less important when the bounds are more informative, but we retain it for simplicity in the simulations that follow.) With this setup, unless the relationship is clearly nonlinear, a regression of the true $\beta_i^b$ on $Z_i$ recovers the intercept and slope coefficients accurately, effectively correcting the problem with the HS procedure.

We use this Monte Carlo setup to illustrate four prototypical situations that in our experience map out the space of applications in which the various second-stage methods work in different ways, at least when we follow the parameter values chosen in HS. (That is, all the simulations we have run look like these plots or, roughly speaking, convex combinations of them.)

First, when ecological data have very wide bounds, EI (and any method of ecological inference) is sensitive to modeling assumptions. In many applications with data such as these, no ecological inference should be conducted unless one has some special auxiliary information about the model assumptions. If one nevertheless proceeds to the second stage, then, because shrinkage probably exists in the $\beta_i^b$'s, the true (unobserved) value of the full adjustment would make for an improvement over LS using the estimated $\hat{\beta}_i^b$'s. Of course, the true adjustment is not known and must be estimated. Unfortunately, it cannot be estimated reliably because EI models $\beta_i^b$ as a random effect constrained to be within the precinct-level bounds. If $Z_i$ is not in the EI first stage (which is true by definition because if it were included, we would not need a second stage), then *the only information in $\hat{\beta}_i^b$ that could be predicted by $Z_i$ comes from the bounds*. The same is true of the adjustment procedure: *The only information with which to estimate $\delta_0^b$ and $\delta_1^b$ comes from the bounds*. If the bounds are relatively uninformative, as we assume in this first prototypical case, then there is little information with which to estimate the adjustment. Of course, this should not be a surprise: An unbounded random effect variable must be unrelated to all measured variables except by chance.

Figure 1 plots the covariate $Z_i$ horizontally by the true $\beta_i^b$ (in the left graph) and the estimated $\hat{\beta}_i^b$ (in the right graph) vertically. Note how the unadjusted least squares line (marked LS) fits the estimated points well (in the right graph), but is attenuated for the true points (in the left). Because the bounds are all very wide, the variances are almost constant

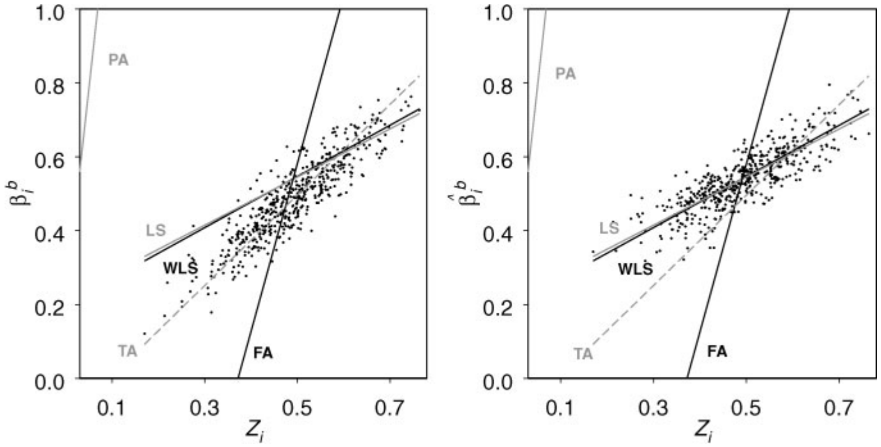

**Fig. 1** Data with wide bounds. Plot of $Z_i$ horizontally by the estimated $\hat{\beta}_i^b$ vertically (in the right graph) and the true $\beta_i^b$ (in the left graph), with fits for the partially adjusted procedure (PA) in HS, least squares (LS) and weighted least squares (WLS) almost on top of one another, the fully adjusted method (FA), and the full adjustment based on the true values of $\delta_0^b$ and $\delta_1^b$ (TA). Clearly, TA fits the true points best, but is unfortunately badly estimated by FA. In this example, insufficient information exists in the bounds with which to make ecological inferences at all. Data were generated from the extended EI model with $X \sim \text{Uniform}\ (0, 0.2)$, $Z \sim \text{Normal}\ (0.5, 0.01)$, $\breve{\mathfrak{B}}_i^b = Z_i - 0.1$, $\breve{\mathfrak{B}}_i^w = Z_i - 0.1$, $\breve{\sigma}_b = 0.05$, $\breve{\sigma}_w = 0.05$, and $\breve{\rho} = 0$.

and so the weighted least squares (marked WLS) line is practically on top of the LS line. The line representing the fully adjusted method using the *true* values of $\delta_0^b$ and $\delta_1^b$ (true adjustment is marked TA) fits the true points well, and would correct for the attenuation. The actual fully adjusted method (marked FA) as estimated from the data also appears, but as a result of the wide bounds, it is not a good estimate, and indeed, is worse than LS and WLS. Figure 1 also plots the partially adjusted method (marked PA) that a researcher might implement based on HS's article, which is more biased than any of the alternatives, a problem that grows in severity as the mean of $Z$ departs from zero. (The PA and FA lines are not exactly parallel because PA is calculated by assuming the estimate $\breve{\breve{\psi}}$ is known exactly, whereas FA uses our estimation procedure.) Because the partial adjustment method is never better than full adjustment, and often dramatically worse, we do not consider it further.

Second, when the bounds are at least somewhat informative (that is, when few of the bounds are extremely wide), we are in the situation in which we would be more likely to trust ecological inferences using EI (or another method that takes into account the information in the precinct-level bounds). When in addition the relationship is approximately (or locally) linear, we find that LS and WLS usually do as well as, and often better than, the fully adjusted procedure. Figure 2 gives one example in which LS, WLS, and the fully adjusted method all give approximately the same estimates.

Third, when some observations have wide bounds and others have narrow bounds, and $\hat{\beta}_i^b$ is an approximate (or locally) linear function of $Z_i$, (unadjusted) WLS regression is often substantially less biased than LS, and approximately equivalent to or better than the fully adjusted procedure.[6] This is contrary to HS's claims that WLS would not make a difference;

[6] Like all weighted regressions, this procedure would have higher variance than LS. HS studied consistency, and implicitly bias, but did not address other properties, such as efficiency.

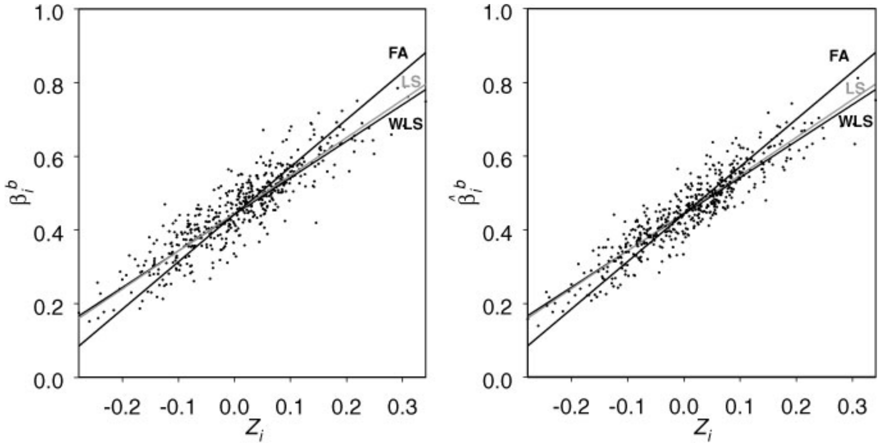

**Fig. 2** Data with informative bounds. Plot of $Z_i$ horizontally by the estimated $\hat{\beta}_i^b$ vertically (in the right graph) and the true $\beta_i^b$ (in the left graph) with fits for least squares (LS) and weighted least squares (WLS) appearing almost on top of one another, and the fully adjusted method (FA). Note how all three methods give almost the same answer. Data were generated from the extended EI model, with $X \sim \text{Uniform }(0.2, 1)$, $Z \sim \text{Normal }(0, 0.01)$, $\breve{\mathfrak{B}}_i^b = Z_i + 0.44$, $\breve{\mathfrak{B}}_i^w = Z_i + 0.68$, $\breve{\sigma}_b = 0.05$, $\breve{\sigma}_w = 0.05$, and $\breve{\rho} = 0$.

what they missed by applying a linear regression framework to this problem with bounds and nonlinearity is that the degree of attenuation is greater when the bounds are wider—as can be seen by the differences between Figs. 1 and 2—and so the weights are correlated with the attenuation bias and can at least partially correct for it.

Figure 3 provides an example of this phenomenon. We generated the data for this figure from the same model as Fig. 2, changing only the parameter values so that the points were

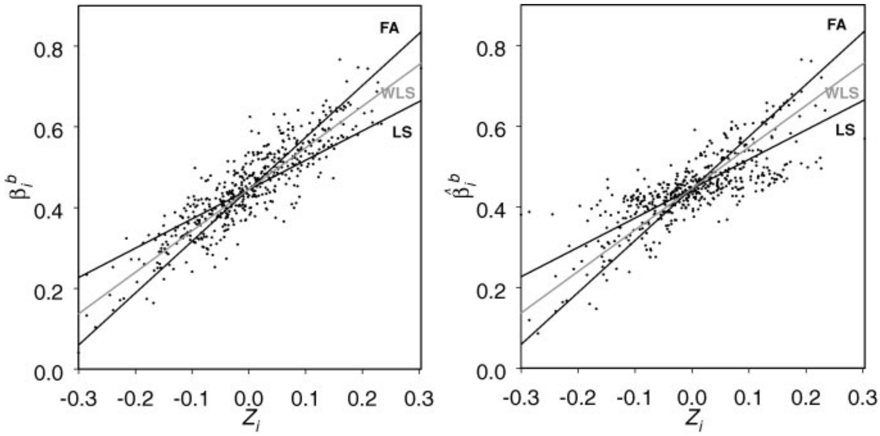

**Fig. 3** Data with narrow and wide bounds. Plot of $Z_i$ horizontally by the estimated $\hat{\beta}_i^b$ vertically (in the right graph) and the true $\beta_i^b$ (in the left graph) with fits for least squares (LS), weighted least squares (WLS), and the fully adjusted method (FA). Note how (unadjusted) WLS corrects for most of the attenuation bias. Data were generated from the extended EI model, with $X \sim \frac{1}{2}\text{Uniform }(0, 0.2) + \frac{1}{2}\text{Uniform }(0.8, 1)$, $Z \sim \text{Normal }(0, 0.01)$, $\breve{\mathfrak{B}}_i^b = Z_i + 0.44$, $\breve{\mathfrak{B}}_i^w = Z_i + 0.68$, $\breve{\sigma}_b = 0.05$, $\breve{\sigma}_w = 0.05$, and $\breve{\rho} = 0$.

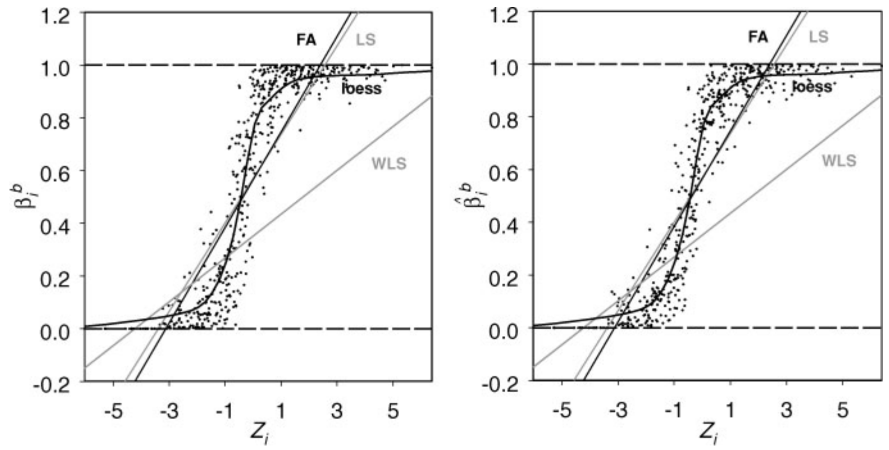

**Fig. 4** Data with a nonlinear relationship. Plot of $Z_i$ horizontally by the estimated $\hat{\beta}_i^b$ vertically (in the right graph) and the true $\beta_i^b$ (in the left graph) with fits for least squares (LS) and the fully adjusted method (FA) almost on top of one another, weighted least squares (WLS), and the better fitting loess regression on the logistic scale (loess). Note how all the linear methods give out of bounds predictions. Data were generated from the extended EI model, with $X \sim$ Uniform (0, 1), $Z \sim$ Normal (0, 4), $\breve{\mathfrak{B}}_i^b = Z_i + 1.16$, $\breve{\mathfrak{B}}_i^w = Z_i + 1.16$, $\breve{\sigma}_b = 0.3$, $\breve{\sigma}_w = 0.3$, and $\breve{\rho} = 0$.

affected by the bounds. The effect of the wide bounds on some observations can be seen by the attenuation in the set of points forming a flatter slope in the right graph (as compared to the left graph which has no such feature). As a result, the LS line is a good deal flatter than it should be (as judged by the fit to the points in the left graph) but the (unadjusted) WLS line corrects for most of the attenuation. The FA line, in contrast, overcorrects for attenuation.

Finally, when the relationship is nonlinear, as is often observably the case because of the bounds, then any (adjusted or unadjusted) linear second-stage procedure can produce impossible results. In this situation, a scatterplot or an appropriate nonlinear procedure is better. The fully adjusted procedure in this situation often produces more out-of-bounds predictions than the unadjusted procedure. (In this case, WLS is also inappropriate, both because the assumption of linearity does not hold and because $\hat{\beta}_i^b$'s at the extremes have standard errors of zero or nearly so. Giving extra weight to these observations tends to bias the estimate of the slope downward.) Figure 4 illustrates these issues. In this example, we also include a nonlinear model by using a loess regression of a logit transformation of $\hat{\beta}_i^b$, $\ln(\hat{\beta}_i^b/(1 - \hat{\beta}_i^b))$, on $Z_i$ and using simulation to compute the regression line. This line (marked loess) clearly gives a far better fit than any of the other methods. It is also the only method that does not extend above 1 or below 0 for $\beta_i^b$ (i.e., into the impossible region) for some values of $Z_i$. (A linear regression on the logit scale would also stay out of the impossible region, but the fit would not be as much of an improvement.)

## 6 The Extended Model

As the only self-consistent second-stage approach, the extended model should probably see more use. We therefore pause briefly here to discuss an important issue about how to use it.

One apparently obvious, but flawed, way to use the extended model is to study the effects of the explantory variables only by looking at the truncated normal parameter estimates ($\alpha^b$ and $\alpha^w$ in King, 1997, p. 170) and their standard errors. The problem with this approach is that it does not include the robustness of the bounds that comes from conditioning on $T_i$.

To explain, consider a simpler case: estimating the district-wide fraction of blacks who vote, $B^b$, without covariates. If the model holds (and the number of people per precinct is constant over precincts), a consistent and efficient estimator of this quantity is as follows. (1) Run EI to estimate the five parameters of the truncated bivariate normal. Because they are on the untruncated scale, (2) compute from them (analytically or by simulation) the truncated parameters, which of course includes the mean of the precinct parameters, and hence our estimate.

Suppose, however, that the model is not exactly right. Then we would also want to condition on $T$ so that the precinct-level bounding information can be included in this estimate. To do this, use an alternative estimator: (1) Run EI to estimate the five parameters of the truncated bivariate normal. Then, (2) condition on T and compute estimates of the fraction of blacks who vote in each precinct, $\beta_i^b$, by drawing (as in King, 1997) from the posterior density, $P(\beta_i^b \mid T_i)$, all the mass of which falls within the known bounds. Finally, (3) average the precinct estimates to produce an estimate of $B^b$, as desired originally.

Because it includes the precinct-level bounds, the second estimator is clearly less sensitive to misspecification than the first. Also, because dealing with misspecification is the key issue in making ecological inferences in real research, we see little reason to use the first estimator. An equivalent problem applies in estimating and interpreting $\alpha^b$ and $\alpha^w$ directly: the estimates do not include information from the precinct-level bounds. Although conditional on the model, they are consistent and efficient, they are more sensitive to misspecification. Thus, we suggest the same approach to estimating effects from the extended model: (1) Estimate the parameters of the truncated bivariate normal and $\alpha^b$ and $\alpha^w$. (2) Condition on $T_i$ and compute the conditional densities, $P(\beta_i^b \mid T_i)$. (3) Either display all the densities as a function of $Z$, such as via a scatterplot of simulations from these densities vertically by $Z_i$ horizontally, or summarize them in some way. The result is much less sensitive to misspecification.

## 7 Concluding Suggestions

As suggested in King (1997) and quoted earlier, "The best first approach is usually to display a scatterplot of the explanatory variable (or variables) horizontally and (say) an estimate of $\beta_i^b$ or $\beta_i^w$ vertically. In many cases, this plot will be sufficient evidence to complete the second stage analysis" (p. 289). This approach remains accurate. Indeed, *show us the data* is a good general motto for any statistical analysis, especially those with complex nonlinear and bounded variables such as those resulting from ecological inference. To this scatterplot, we would suggest adding information on the bounds. This can be done by adding a thin vertical line representing the bounds on $\beta_i^b$ for each $(\hat{\beta}_i^b, Z_i)$ point plotted (King, 1997, Fig. 13.2). From this figure, we can then see all the information in the data, precisely how informative the bounds are, and whether the bounds are of constant or variable width.

For researchers who wish a simple approximation to estimating a second-stage relationship instead of the extended model, the information provided in this article provides a guide. If enough information to run a first-stage EI model exists and a scatterplot does not indicate nonlinearity, then weighted least squares is the best approach. Researchers can easily tell which is appropriate by examining the situations discussed in Section 5.

If a more formal statistical approach seems desirable, then a good method must go beyond classical linear econometric theory. It must take into account (a) the nonlinear nature of the problem, (b) the bounded nature of the second-stage dependent variable with the width of the bounds varying over observations, (c) the heteroskedasticity and the correlation between bounds and the shrinkage, and (d) the effect of any possible logical inconsistency of the first and second stages of the analysis (because, of course, two-step statistical methods need not be logically consistent to work well; see, e.g., Meng, 1994). At present, the only model that has been proposed with all these properties is the extended EI model that allows covariates to be included as part of the EI estimation procedure (King, 1997, Ch. 9). HS are correct that this extended model is sometimes only weakly identified, but that is only when the bounds are not narrow enough and $X$ is included among the covariates or highly related to $Z$. In other cases, with narrow bounds or even wide ones when $Z$ is unrelated to $X$, the extended model can be strongly identified and so can be used in many cases without problem. Imai and King (2002) demonstrate how to compute first differences and other quantities of interest from the extended EI model, and they report on extensions of the EI software to make this possible.

Econometric theory and the classic linear regression framework works well for what it was designed. However, in models with nonlinear relationships or sample spaces, or parameter spaces that are highly and differentially bounded, such as in ecological inference problems, political methodologists must look elsewhere or develop their own methods.

HS have made an important contribution by highlighting what turns out to be the shrinkage property of Bayesian point estimates such as those provided by EI. We are in their debt for pointing this out and stimulating the ideas and discussions offered herein.

## References

Duncan, Otis Dudley, and Beverly Davis. 1953. “An Alternative to Ecological Correlation.” *American Sociological Review* 18:665–666.

Goodman, Leo. 1959. “Some Alternatives to Ecological Correlation.” *American Journal of Sociology* 64:610–624.

Herron, Michael C., and Kenneth W. Shotts. 2003. “Using Ecological Inference Point Estimates as Dependent Variables in Second-Stage Linear Regressions.” *Political Analysis* 11:44–64.

Imai, Kosuke, and Gary King. 2002. “Did Illegally Counted Overseas Absentee Ballots Decide the 2000 U.S. Presidential Election?” (Available from http://gking.harvard.edu/preprints.shtml#ballots.)

King, Gary. 1997. *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*. Princeton, NJ: Princeton University Press.

Meng, X. L. 1994. “Multiple-Imputation Inferences with Uncongenial Sources of Input.” *Statistical Science* 9:538–573.