A Consensus on Second-Stage Analyses in Ecological Inference Models

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

with
Michael C. Herron and Kenneth W. Shotts
Department of Political Science, Northwestern University, 601 University Place, Evanston, IL 60208-1006
e-mail: m-herron@northwestern.edu
e-mail: k-shotts@northwestern.edu

1 Introduction

Since Herron and Shotts (2003a; hereinafter HS), Adolph and King (2003; hereinafter AK), and Herron and Shotts (2003b; hereinafter HS2), the four of us have iterated many more times, learned a great deal, and arrived at a consensus on this issue. This paper describes our joint recommendations for how to run second-stage ecological regressions, and provides detailed analyses to back up our claims.

Our research applies to the problem of estimating how the unobserved parameters β_i^b and β_i^w (e.g., the fraction of blacks and whites, respectively, who vote in precinct i) vary with observed exogenous variables Z_i (e.g., campaign spending in precinct i), using two other observed variables, X_i (e.g., the proportion of people in the precinct who are black) and T_i (e.g., the proportion of people who vote), and using King's (1997) ecological inference model (popularly known as EI).

We offer three main conclusions. First, the best approach is to use the extended EI model for estimation (i.e., to include Z_i in the model from the start and estimate all parameters jointly). Second, in most applications, the best way to use the extended model is not to look at the coefficients on Z_i in EI's systematic component, as some have done, but instead to condition on the precinct-level data in a way we describe later. Third, point estimates from weighted least squares (WLS) second-stage regressions are useful because they have substantively negligible bias in a diverse variety of situations (knowable in advance from the observed data), even if Z_i is omitted from the first-stage ecological inference and even when a second-stage analysis based on least squares (LS) would be biased.

Copyright 2003 by the Society for Political Methodology

Following AK and HS2, this paper does not study the consequences of a vector-valued Z_i or of statistical properties other than unbiasedness. We suspect our results may be more general, but this must be studied.[1]

2 Running and Interpreting the Extended Model

Statistical inference can be based on a *population* or a *superpopulation* approach. The different approaches identify different quantities to estimate and have different implications for robustness to modeling assumptions. Because, in linear regression models, both lead to the same least squares estimator, the distinction is not well known among political scientists. However, for many other models, including the basic and extended EI models, the differences can be very important (see AK, Section 6).

In the population approach to ecological inference, the parameters of interest are those that characterize the districts as they exist for the elections under study, such as β_i^b—the fraction of blacks who voted in the election being analyzed. In judicial litigation concerning redistricting and in many related political science studies, the parameters of interest are those for the elections under study; therefore, the population approach is appropriate. In other situations, the unobserved parameter β_i^b is treated as merely a realization from some higher level "superpopulation" parameters. For more general theories of politics, we might hope to specify and estimate the parameters of a superpopulation model that give rise to the population parameters as unobserved realizations, and which, in turn, give rise to the observed data. The higher level of generalization involved in superpopulation models can thus impose a cost to the analyst in terms of less robustness to model misspecification. Because superpopulation parameters are in a sense farther from the data than population parameters, increased model dependence (i.e., less robustness to model misspecification) is a natural consequence. As a result, of course, "validating [superpopulation] models can be difficult in practice" (Rao 1999, p. 16). Because almost the entire difficulty with making ecological inferences is model specification, we are not optimistic about inferences on superpopulation parameters in ecological inference, at least not with current technology or without additional assumptions. But even if one prefers the superpopulation approach, summarizing the results of the population approach, as we describe later, may still be the best alternative.

In EI, the superpopulation approach would be to use the maximum likelihood (or maximum posterior) estimate of the coefficient on Z_i in the systematic component for the mean in the extended model (i.e., α^b and α^w in King 1997, Eq. 9.2). Our preferred population approach conditions on the data and so estimates a set of conditional densities, $P(\beta_i^b \mid T_i, Z_i)$ for $i = 1, \ldots n$, which is the full summary of information about how β_i^b varies with Z_i. Conditioning on the data is especially helpful in ecological inference because it changes an unbounded estimation problem (as the maximum likelihood or posterior for α is an unconstrained maximization problem) into one that is strictly bounded within known bounds set by the data.

As an illustration of how conditioning on the data (and thus the information in the bounds) adds robustness to estimates derived from the extended EI model, we draw from

[1] We have also not studied the accuracy of standard errors, but the usual model-based standard errors, even when they are accurate, would still be a relatively minor portion of the uncertainty in most applications of ecological inference, most of which is the result of specification uncertainty.

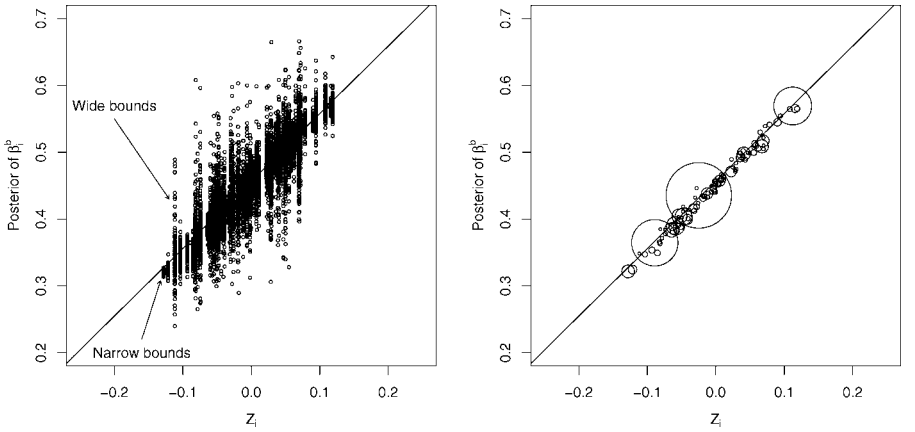

Fig. 1 The population approach: Conditional distributions of $P(\beta_i^b \mid T_i, Z_i)$ plotted vertically by Z_i horizontally. Simulations from the conditional densities are plotted in the left graph. Because the most informative densities (i.e., with the narrowest bounds and smallest variance) are hardest to see there (see the two labeled examples), we also represent in the right graph each density as a circle with area inversely proportional to the variance.

the extended model setting the true $\alpha^b = 1$, with informative bounds.[2] Then we run the extended EI model, but with a highly inaccurate prior on α^b of $N(0, 0.01)$. Because we are interested in α for the purpose of this illustration, this is a particularly extreme form of model misspecification; effectively, we have stacked the deck against ourselves. The result is a superpopulation parameter estimate of $\hat{\alpha}^b = 0.01$ (with a standard error of 0.01) that is far from the true value of $\alpha^b = 1$, and thus dramatically illustrates the nonrobustness of the superpopulation approach, at least to this type of misspecification.

Fortunately, and in spite of the inappropriateness of the model, conditioning on the data under the population approach yields a much better estimate, which we display as a graph of simulations from the posterior estimates of β_i^b plotted against Z_i (Fig. 1). As a summary of this plot, we have drawn a WLS regression line through these posterior distributions, which in this example has a slope of 1.01 (with a standard error of 0.003), which is obviously very close to the true value of 1.00 (weights for WLS were computed from the variance of each conditional posterior). In other cases, in which the plot shows evidence of nonlinearity, some other summary (or simply showing the plot itself) is a better approach. For some purposes, we wish to pay attention to the estimation dependence across simulations in different densities; this can be done by running WLS for each simulation separately, which in this example increases the standard error but does not change the point estimate. (Note that however one summarizes this set of conditional distributions, this procedure is not EI-R because Z_i is in the first-stage run; therefore, none of the issues with that approach, discussed in subsequent sections, apply here.)

Almost all empirical analyses in political science, no matter what statistical model they use (i.e., not only ecological inference), are based implicitly on a population approach to making inferences. For example, researchers ask whether the 1996 and 2000 elections

[2]We set $\breve{\mathfrak{B}}_i^b = Z_i + 0.44$, $\breve{\mathfrak{B}}_i^w = 0.1Z_i + 0.68$, $\breve{\sigma}_b = 0.005$, $\breve{\sigma}_w = 0.005$, $\breve{\rho} = 0$, $Z \sim N(0, 0.0025)$ and $X \sim \text{Uniform}(0.6, 1)$. As in all data sets generated for this paper, we drew $n = 500$ precincts consisting of 1000 individuals each.

were racially polarized. Indeed, even if they are interested in whether there is racially polarized voting in all elections (or all elections of a certain type), summaries of the set of estimates from the population approach applied to each election in one's data set can be a better way to make superpopulation inferences. If our results given in this section are general, the population approach also has the advantage of being substantially more robust to misspecification, and we recommend it when interpreting the extended EI model.

For other examples of the distinction between population and superpopulation approaches, see Korn and Graubard's (1998) work on variance estimation under complex sampling. For applications in political science, see Gelman and King (1994) and King et al. (2002).

3 Understanding When Second-Stage Regressions Have Substantively Negligible Bias

In this section, we show the precise conditions under which weighted least squares provides estimates with negligible bias in second-stage regressions (we call this procedure *EI-W*, in contrast to EI-R, which was the focus of HS and HS2).[3] The basic intuition is that because Z_i is omitted from the first stage EI run, the only information in the second stage with which to estimate the effect of Z_i comes from the bounds. Because these bounds are known, we have a chance to ascertain when they are sufficiently informative to provide approximately unbiased estimates.

We begin by showing that each of the components of HS2's bias decomposition (the true δ parameters) are almost linear functions of the observable mean bound width and the variance of the bound width across precincts. We do this by simulating a wide variety of data sets, computing the δ parameters in each, and regressing these on the mean and variance of the bound width. Table 1 presents the average R^2 values from these regressions, all of which are very high.[4]

Figure 2 summarizes the results in the rest of the section. We begin by using precinct-level turnout and census data from 385 elections from the last five election cycles in five diverse states (Arizona, Georgia, Maryland, Ohio, and Texas).[5] For each election, we plot the average width of the bounds across precincts (horizontally) by the standard deviation of the bound width (vertically) twice, once for the bounds on β_i^b (as triangles) and once for β_i^w (as circles). The means and standard deviations of bound widths in these data sets display a clear parabolic relationship, resembling the relationship between the mean and standard deviation of a binomial density (and for the same reason: when the mean bound width approaches 0 or 1, the variance must be smaller). We summarize the results via loess-generated 95% confidence intervals.

[3] We use the standard errors of β_i^b to compute weights for the regression of β_i^b on Z_i. See Lewis (2000) for an alternative approach that is also worth considering.

[4] Each row in Table 1 refers to a family of 17 scenarios across which data sets are generated from the same parameters but with varied bounds created by varying the distribution of X. For all scenarios in all 12 families, $\breve{\mathfrak{B}}_i^b = Z_i + 0.44$, $\breve{\mathfrak{B}}_i^w = Z_i + 0.68$, and $\breve{\rho} = 0$. For each row, $\breve{\sigma}_b$, $\breve{\sigma}_w$, and $Z_i \sim (\mu_Z, \sigma_Z^2)$, are set as indicated. For each scenario, we averaged results across 100 simulated data sets. For each scenario in each row, we regressed the error in EI estimates $(\hat{\beta}_i^b - \beta_i^b)$ on a constant term and the true β_i^b and β_i^w, as in HS2, yielding coefficients δ_1 and δ_2. Then, for each family of scenarios, we regressed the different values of δ_1 and δ_2 separately on the mean black and white bound width and the variance of the black and white bound widths. The R^2 for each of these regressions, denoted $R^2_{\delta_1}$ and $R^2_{\delta_2}$, is reported in the table.

[5] These elections include a variety of State House, State Senate, U.S. House, U.S. Senate, and statewide races.

Table 1 For a given second-stage covariate and first-stage variance, second-stage error is a nearly deterministic function of the bounds

$\breve{\sigma}_b = \breve{\sigma}_w$	σ_Z	μ_Z	$R^2_{\delta_1}$	$R^2_{\delta_2}$
0.05	0.1	0	.98	.98
0.05	0.1	0.5	.93	.89
0.05	0.5	0	.99	.99
0.05	0.5	0.5	.94	.93
0.1	0.1	0	.99	.99
0.1	0.1	0.5	.99	.99
0.1	0.5	0	.99	.99
0.1	0.5	0.5	.99	.99
0.3	0.1	0	.99	.99
0.3	0.1	0.5	.99	.99
0.3	0.5	0	.99	.96
0.3	0.5	0.5	.99	.95

The circles and triangles in Fig. 2 merely summarize data. To this, we have also added a shaded region in which our analyses (to be discussed below) indicate that EI-W is substantially biased and so should not be used. Outside roughly this shaded region, we find that EI-W has negligible bias, even when EI-R, a second-stage least-squares regression, is strongly biased. Intuitively, small mean bound widths reduce bias by providing more

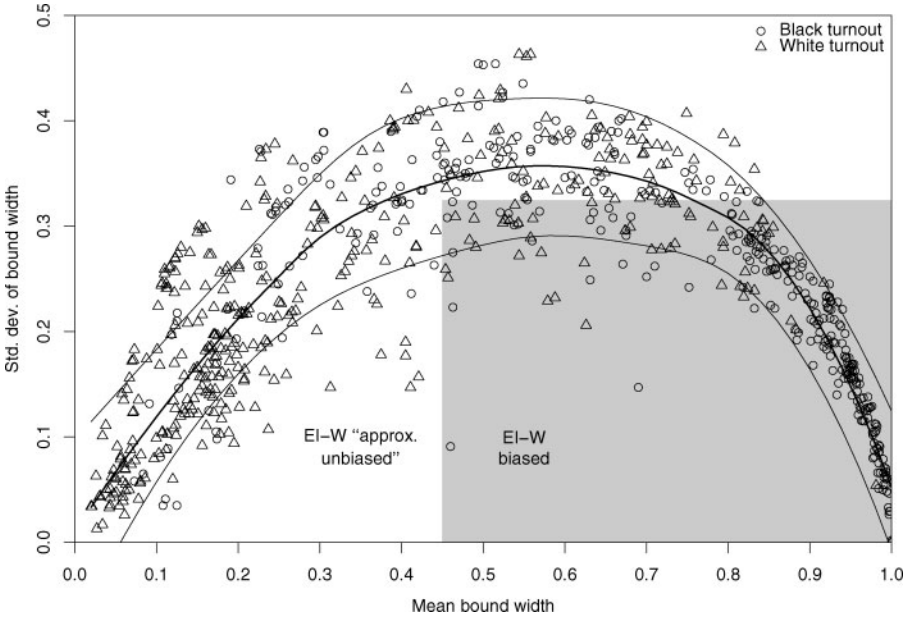

Fig. 2 Mean and standard deviation of the bounds of β_i^b (triangles) and β_i^w (circles) for state elections in Arizona, Georgia, Maryland, Ohio, and Texas. The line was fit with loess (span = 0.5). When data fall within the shaded region, second-stage weighted least squares analyses should not be performed. Dependence between the the two bounds in each election as a result of the tomography line is not shown. The term *approximately unbiased* in this figure is defined with precision by Figs. 3 and 4.

information about the precinct-level parameters, whereas large standard deviations of the bound widths induce a larger correlation between the weights and the β_i's, providing more leverage for the weighted least squares procedure (in EI-W) to reduce the bias (see also AK).

We now present two analyses and corresponding figures to support the shaded region in Fig. 2. Figure 3 gives a set of difficult cases, with differing means and standard deviations

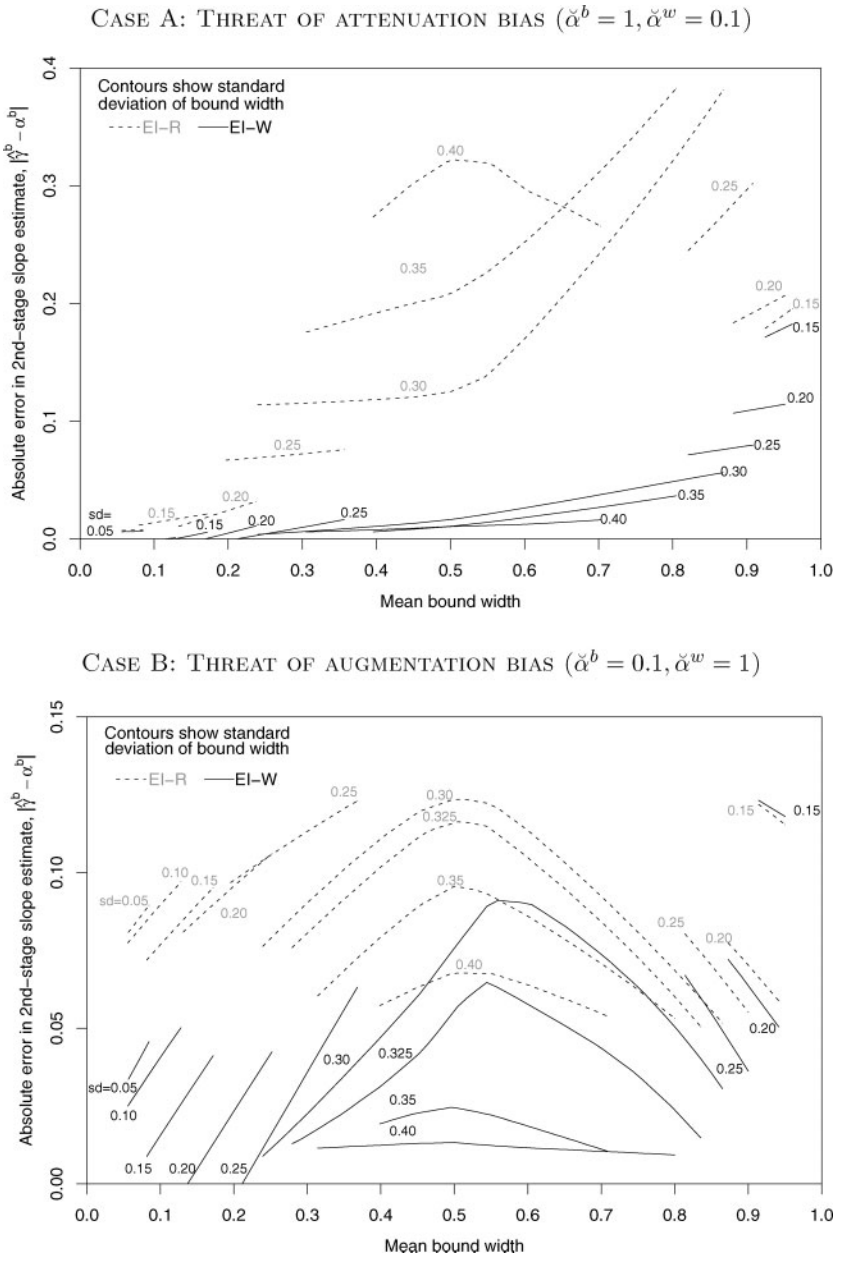

Fig. 3 Bias in EI-R and EI-W second-stage analyses as a function of the mean and standard deviation of the bound width (drawn with loess, span = 0.95). Note how EI-W eliminates much of the bias as compared to EI-R.

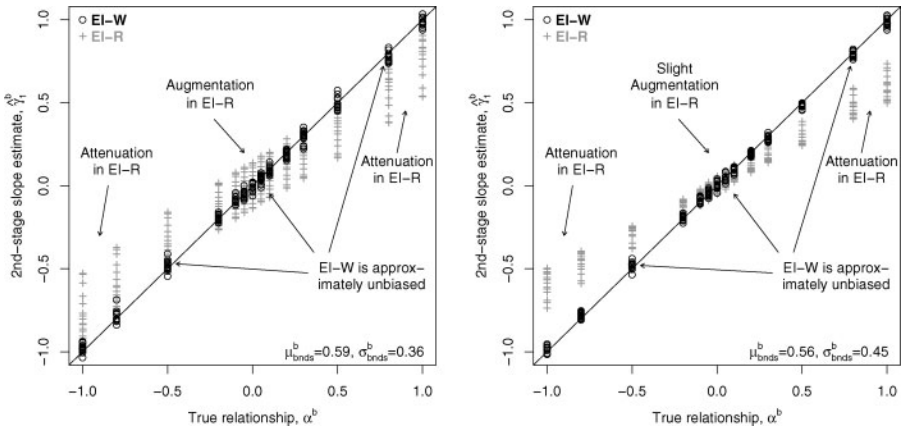

Fig. 4 Bias in EI-R and EI-W second-stage analyses as a function of α^b and α^w. EI-W eliminates most attenuation and augmentation bias in the cases considered.

of the bound width, in which EI-R should fail badly. Case A (the upper panel in the figure) is an example of attenuation bias as indicated in HS and HS2 under EI-R. Data for Case B were generated to fit the ideas in HS2 corresponding to augmentation bias.[6] For each scenario, we consider a wide range of 167 sets of bounds that evenly cover the space of possible bound widths as summarized by their means and standard deviations. For each of these subscenarios, we generate 100 data sets from the extended-EI model, and run EI-R and EI-W. We use the average error and bounds properties for each scenario to produce contour plots of the absolute error in second-stage estimates. The contour lines shown are loess regressions on these simulation results. For clarity and to focus on those cases likely to occur in practical research, we show only those parts of the contour lines for which the mean and standard deviation of the bounds fall inside the 95% confidence region from Fig. 2.

The results in Fig. 3 support HS2's claim regarding bias in EI-R (displayed via dashed contour lines). The figure also supports AK's claim (elaborating King 1997, pp. 289–290) that EI-W (solid contours) is substantially less biased than EI-R for both cases, and has particularly small bias for values of mean and standard deviation of bound widths corresponding to the unshaded region in Fig. 2.

Figure 3 examines simulations for different types of bounds and two fixed values for α^b and α^w. We reverse this setup in Fig. 4 and keep the bounds fixed, and instead we choose a wide range of values for α^b and α^w.[7] In the left panel, the bounds are moderately wide, as in AK's Fig. 2 ($X \sim \text{Uniform}(0.2, 1)$). In the right panel, the bounds are "mixed," as in AK's Fig. 3 ($X \sim \frac{1}{2}\text{Uniform}(0, 0.2) + \frac{1}{2}\text{Uniform}(0.8, 1)$). The results in both plots are given for EI-W (as circles) and EI-R (as plus signs).

[6] In each case, we set $\breve{\mathfrak{B}}_i^b = \breve{\alpha}^b Z_i + 0.44$, $\breve{\mathfrak{B}}_i^w = \breve{\alpha}^w Z_i + 0.68$, $\breve{\sigma}_b = 0.05$, $\breve{\sigma}_w = 0.05$, and $\breve{\rho} = 0$, and draw $Z \sim N(0, 0.01)$, as in AK Figs. 2 and 3. However, unlike AK, we consider cases in which $\breve{\alpha}^b \neq \breve{\alpha}^w$. In Case A, we set $\breve{\alpha}^b = 1$, $\breve{\alpha}^w = 0.1$, which should produce severe attenuation bias in second-stage LS estimates of the relationship between $\hat{\beta}_i^b$ and Z_i. In Case B, $\breve{\alpha}^b = 0.1$, $\breve{\alpha}^w = 1$, which should produce severe augmentation bias.

[7] We begin with the scenarios in AK's Figs. 2 and 3. That is, we set $\breve{\mathfrak{B}}_i^b = \breve{\alpha}^b Z_i + 0.44$, $\breve{\mathfrak{B}}_i^w = \breve{\alpha}^w Z_i + 0.68$, $\breve{\sigma}_b = 0.05$, $\breve{\sigma}_w = 0.05$, and $\breve{\rho} = 0$ and draw $Z \sim N(0, 0.01)$. However, unlike AK, we allow the values of $\breve{\alpha}^b$ and $\breve{\alpha}^w$ to range widely, considering every combination of α's which can be made from the following list: -1, -0.8, -0.5, -0.3, -0.2, -0.1, -0.05, 0, 0.05, 0.1, 0.2, 0.3, 0.5, 0.8, 1, for a total of 225 different scenarios. For each scenario, we generated 100 data sets from the extended-EI model, ran EI-R and EI-W, and averaged the results for each.

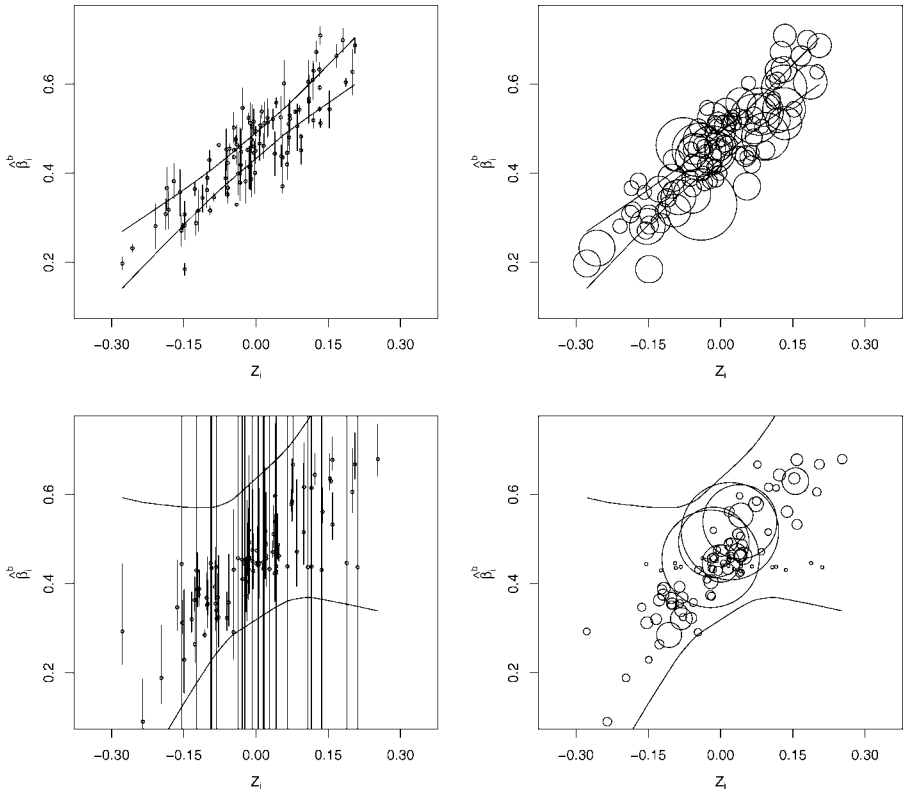

Fig. 5 Diagnostics for extended EI or second-stage ecological regressions.

The results for EI-R match the theory derived in HS2 well, displaying attenuation and augmentation biases just where they are expected. What Fig. 4 adds are EI-W results for the same data sets. Unlike the EI-R results, the EI-W results all appear on or near the 45° line, indicating trivially small levels of bias.

4 Graphically Summarizing Bound Information

Finally, we offer a diagnostic for use in either second-stage analyses or EI extended model runs. We consider two situations and portray them in Fig. 5. For the top row of plots, we generated data so the bounds would be very informative.[8] The top left graph in Fig. 5 plots Z_i horizontally and a vertical line representing the 100% bounds on β_i^b vertically.

In the case of narrow bounds, just plotting the bounds of β_i^b vertically against Z_i highlights all information in the data, as in the top-left figure. Going a step further, we compute 100% confidence intervals around the second-stage regression line, using only the first-stage bounds and the standard assumptions of linear regression, but none of the EI model's assumptions. No logically possible set of β_i^b's could produce a regression line that lies

[8] That is, $X \sim$ Uniform(0.9, 1), $\breve{\mathfrak{B}}_i^b = Z_i + 0.44$, $\breve{\mathfrak{B}}_i^w = 0.68$, $\breve{\sigma}_b = 0.05$, $\breve{\sigma}_w = 0.05$, $\breve{\rho} = 0$, $Z \sim N(0, 0.01)$, and notably, the relationship between β_i^b and Z_i is locally linear.

outside these 100% confidence intervals.[9] The top-right figure shows another way to present the same information: plot each $\hat{\beta}_i^b$ scaled in inverse proportion to the width of the bounds, so that circles representing precincts with narrow bounds have larger area than wide bounds cases. This fixes the same graphical illusion in the same way as in Fig. 1.

The bottom row of plots displays data generated with some narrow and some wide bounds (e.g., $X \sim \frac{1}{4}\text{Uniform}(0, 0.05) + \frac{3}{4}\text{Uniform}(0.8, 1)$). This plot is particularly difficult to read because the uninformative wide bounds obscure the informative narrow bounds and the second-stage 100% confidence intervals are likely to be too wide to serve any purpose for these observations (without more model-dependent inferences). However, the second style of scatterplot, with points inversely proportional to bound width, does serve as a useful diagnostic in such cases (e.g., to assess the possibility of nonlinearity) before proceeding to EI-W or extended EI. For example, a different scatterplot might show the tightly bound cases to be nonlinear in Z, which would require modifying extended EI or EI-W.

5 Concluding Remarks

We recommend using the extended EI model when possible. When running the extended EI model is infeasible because of insufficient time and computational power, or is otherwise inconvenient, we also provide a method of ascertaining when point estimates from WLS second-stage regressions have negligible bias or none at all.

References

Adolph, Christopher, and Gary King. 2003. "Analyzing Second-Stage Ecological Regressions: Comment on Herron and Shotts." *Political Analysis* 11:65–76.

Gelman, Andrew, and Gary King. 1994. "A Unified Method of Evaluating Electoral Systems and Redistricting Plans." *American Journal of Political Science* 38:514–554.

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

Herron, Michael C., and Kenneth W. Shotts. 2003b. "Cross-Contamination in EI-R: Reply." *Political Analysis* 11:77–85.

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

King, Gary, Christopher Murray, Joshua Salomon, and Ajay Tandon. 2002. "Enhancing the Validity and Cross-Cultural Comparability of Survey Research." Manuscript. (Available from http://gking.harvard.edu.)

Korn, E. L., and B. I. Graubard. 1998. "Variance Estimation for Superpopulation Parameters." *Statistica Sinica* 8:1131–1151.

Lewis, Jeffrey B. 2000. "Two-Stage Approaches to Regression Models in which the Dependent Variable Is Based on Estimates." Working paper, University of California–Los Angeles.

Rao, J. N. K. 1999. "Some Current Trends in Sample Survey Theory and Methods." *Sankhya: The Indian Journal of Statistics* 61, Series B, Pt. 1:1–57.

[9] Obtaining the bounds on the second-stage regression line implied by the precinct bounds is straightforward for the bivariate case. Although an infinite number of hypothetical data sets—and corresponding feasible regression lines—could lie inside the p precinct bounds, these lines must lie in a finite (vertical) interval. Thus for any Z_i there must exist a maximal (minimal) regression line above (below) which no feasible regression line exists. To find all maximal and minimal regression lines, we create a data set for each $Z_c \in (Z_1, \ldots Z_i, \ldots Z_p)$ with β_i^b equal to the lower bound if $Z_i \geq Z_c$ and equal to the upper bound otherwise, and vice versa, yielding $2p$ data sets and associated regression lines. The 100% confidence intervals plotted in Fig. 5 are the minimum and maximum of these $2p$ lines at each Z_i. The next version of EI and EzI will include a feature to calculate these second-stage bounds, along with the other developments in HS, AK, and HS2. The software will be available at http://gking.harvard.edu.