BIAS IN ECOLOGICAL REGRESSION

STEPHEN ANSOLABEHERE AND DOUGLAS RIVERS

ABSTRACT. Bias formulae are derived for ecological regression estimators. These formulae are useful for determining the direction and magnitude of bias in estimation. It is shown that when group cohesion is higher in areas with higher concentrations of group members and when polarization is higher in more homogeneous areas, ecological regression estimates of polarization will tend to be biased upward. Both minority and majority cohesion rates will be overestimated, but the magnitude of the bias will be greater for the estimate of minority cohesion. Ecological regression estimators are also compared to alternative neighborhood estimation techniques.

1. INTRODUCTION

Ecological regression analysis is often used to estimate the voting behavior of different groups when survey data are unavailable. A typical application of ecological regression involves an election between two candidates of different races where the goal is to estimate the rate at which voters of each race voted for the minority candidate. In the absence of a survey, the only available data are usually precinct or district level voting returns along with corresponding data on the racial composition of voters (alternatively, registrants or eligible voters) in those precincts or districts.

Another frequent application of ecological regression is estimating the rate of party switching between pairs of elections without a panel survey of voters. In both cases, one attempts to infer group voting rates from the pattern of district level election returns. If districts were homogeneous (e.g., if districts were either all white or all black), inferences from district level election returns would be straightforward. Insofar as districts are mixed or heterogeneous, there is some ambiguity in inferring how individuals belonging to different groups behaved from district level voting returns.

Date: June, 1995.

Earlier versions of this paper were presented at the 1993 Annual Meeting of the Social Science History Association and the 1994 Annual Meetings of the Midwest Political Science Association and the American Political Science Association.

Lately, particularly in connection with voting rights litigation, there has been some controversy about the reliability of ecological regression estimates (see, for example, the exchanges between Wildgen (1988, 1990) and Loewen (1990) in *The Urban Lawyer*, between Freedman *et al.* (1991) and Grofman (1991) and Lichtman (1991) in *Evaluation Review*, and between Bullock (1991) and Grofman (1991) in *Social Science Quarterly*; see Achen and Shively (1995), King (forthcoming), and Rivers (forthcoming), for a detailed surveys).

We take no position on the general applicability of ecological regression analysis; it provides satisfactory estimates in some contexts and misleading estimates in others. Rather, it is important to understand what causes errors in ecological regression estimates and to determine the likely direction and magnitude of bias in specific applications. Toward this end, we derive several useful results on the bias of ecological regression estimates in this paper. Although a general aggregation bias result appears frequently in the literature on ecological regression (see, e.g., Duncan, Cuzzort, and Duncan, 1961, p. 66; Alker, 1966, p. 76, n. 6; and Langbein and Lichtman, 1978, p. 63, n. 2), the condition is difficult to interpret and little use has been made of it. Our main result (Proposition 2) follows from this more general bias formula, but specializes it to the case of primary interest in the voting context (two groups with a dichotomous vote choice).

In discussing voting by different racial groups, we adopt some terminology common in the voting rights literature (see *Thornburg* v. *Gingles*, 478 U.S. 30 (1986)). We call the rate at which minority voters vote for the minority candidate "minority cohesion." Similarly, the rate at which majority voters vote for the majority candidate is called "majority cohesion," while the rate at which majority voters vote for the minority candidate is called "majority defection." The difference between the rate at which minority and majority voters vote for the minority candidate is referred to as "polarization."

The results derived here identify three factors as causes of bias in ecological regression estimates. None of these factors can be assessed from aggregated election return data; each requires additional information about the pattern of group voting rates in the particular empirical situation, but the general methodology proposed here should be applicable under a wide variety of factual circumstances. The first factor is the relationship between majority defection rates and the racial composition of districts. Without a survey, there is rarely any direct evidence on how majority defection varies across precincts or districts. Nonetheless, there is good reason to suspect systematic variation in defection rates across districts. For example, it might be argued that predominantly minority areas tend to be poorer and, hence,

more liberal and Democratic than predominantly majority areas. If so, majority defection rates will be higher in districts populated by higher fractions of minority voters than in areas with high concentrations of majority voters. Of course, the opposite pattern could hold as well. Wright (1975), for example, demonstrates that whites living in counties with high proportions of African Americans were more likely to vote for Wallace in 1968 than whites living in predominantly white areas. Nonetheless, it should be possible for courts to determine through expert testimony the plausible direction of the relationship between majority defection and district racial composition.

The second factor is the relationship between minority cohesion and district racial composition. Again, it might be argued that minority voters living in areas with high concentrations of minorities would exhibit higher levels of cohesion. Blacks living in white neighborhoods might be atypical of most blacks and exhibit ususual voting behavior.

The third factor is how the degree of polarization—the difference between minority and majority support for the minority candidate—varies with the heterogeneity of districts. Specifically, we require information about whether mixed or heterogeneous districts (that is, districts with roughly equal numbers of minority and majority voters) have higher or lower levels of polarization than homogeneous minority or majority districts. If prejudiced voters tend to live in more homogeneous areas, then polarization would be lower in the more heterogeneous areas.

With information about these three factors, we show how it is possible to determine the direction and (sometimes) the magnitude of the bias in ecological regression estimates. To summarize our results: *When a group's cohesion tends to be higher in areas with higher concentrations of group members and when polarization tends to be higher in more homogeneous areas, ecological regression estimates of polarization will tend to be biased upward. Under these conditions, both minority and majority cohesion will be overestimated, but the magnitude of the bias will be greater for the estimate of minority cohesion.*

The following section sets forth the notation employed in the remainder of the paper. Section 3 reviews a non-parametric procedure—the method of bounds—for determining the range of possible parameter values. Section 4 describes the ecological regression methodology and presents several empirical examples. Section 5 presents the main theoretical results, including an explicit bias formula for ecological regression estimates. Section 6 adds an additional assumption—linear contextual effects—to obtain somewhat sharper bias results. In the final section, an alternative estimation method—the neighborhood estimator proposed by Klein *et al.* (1991)—is compared to the more conventional ecological regression estimator.

2. NOTATION

For concreteness, consider an election between black and white candidates conducted in N districts with n_i voters in district i ($i = 1, \ldots, N$). Let y_i denote the proportion of the vote in district i received by the black candidate and let x_i denote the proportion of voters in district i who were black. Both x_i and y_i are observable, but we do not know what proportions of blacks and whites in the district voted for each candidate. Let p_i denote the proportion of blacks who voted for the black candidate and let q_i denote the proportion of whites who voted for the black candidate, so that

$$y_i = p_i x_i + q_i(1 - x_i). \tag{1}$$

The fraction of vote received by the black candidate in all N districts combined is a weighted average of the vote for that candidate in the N districts,

$$\bar{y} = \sum_{i=1}^{N} \frac{n_i}{n} y_i,$$

where $n = \sum_{i=1}^{N} n_i$ is the total number of voters in the N districts. Similarly, the proportion of blacks in the combined electorate is

$$\bar{x} = \sum_{i=1}^{N} \frac{n_i}{n} x_i.$$

We shall adopt the convention that $\bar{x} < 1/2$ so that blacks are assumed to be in the minority in the combined electorate.

Let p and q denote the proportion of blacks and whites, respectively, in the combined electorate who voted for the black candidate. We will refer to p as "minority cohesion" and to q as the "majority defection." Then it follows, as a matter of accounting, that

$$\bar{y} = p\bar{x} + q(1 - \bar{x}). \tag{2}$$

The aggregate voting rates p and q are weighted averages of the district level voting rates p_i and q_i. For example, $n_i p_i x_i$ of the $n_i x_i$ blacks in district i voted for the black candidate, so

$$p = \frac{\sum_{i=1}^{N} n_i p_i x_i}{\sum_{i=1}^{N} n_i x_i}.$$

Similarly,

$$q = \frac{\sum_{i=1}^{N} n_i q_i (1 - x_i)}{\sum_{i=1}^{N} n_i (1 - x_i)}.$$

Note, however, that the weights used in computing p and q are different: p is obtained by weighting by the proportion of all black voters who voted in

district i, while q is obtained by weighting by the proportion of all white voters who voted in district i. The average proportion of blacks who voted for the black candidate is

$$\bar{p} = \sum_{i=1}^{N} \frac{n_i}{n} p_i,$$

while the average proportion of whites voting for the black candidate is

$$\bar{q} = \sum_{i=1}^{N} \frac{n_i}{n} q_i.$$

In general, $\bar{p}$ will differ from p and $\bar{q}$ will differ from q.

The difference in voting rates between the two groups,

$$\beta_i \equiv p_i - q_i,$$

will be called the *polarization* in district i. The aggregate polarization is $\beta = p - q$ and the average polarization is $\bar{\beta} = \bar{p} - \bar{q}$. Again, $\bar{\beta}$ and β may differ because of the different weights used in computing p and q.

A measure of the heterogeneity of districts is provided by $h_i = x_i(1 - x_i)$. It varies between 0 and $1/4$ and takes its maximum value only when the number of minority group members in a district equals the number of majority group members.

The notation $C(x_i, y_i)$ will be used for the covariance between x_i and y_i. In computing covariances, we shall always weight districts by the number of voters, so

$$C(x_i, y_i) = \sum_{i=1}^{N} \frac{n_i}{n} (x_i - \bar{x})(y_i - \bar{y}).$$

The notation $x_i \perp y_i$ will be used to indicate that $C(x_i, y_i) = 0$.

The variance of x_i is $V(x_i) = C(x_i, x_i)$ and will usually be denoted by σ_x^2. It will be assumed that $\sigma_x^2 > 0$ to avoid degenerate situations. We will also have occasion to use the skewness of x_i,

$$\gamma_x = \sum_{i=1}^{N} \frac{n_i}{n} \left(\frac{x_i - \bar{x}}{\sigma_x} \right)^3.$$

Finally, b_{yx} will denote the slope of a (weighted) least squares regression of y_i on x_i,

$$b_{yx} = C(x_i, y_i)/V(x_i)$$

Similarly, $b_{px} = C(x_i, p_i)/V(x_i)$ will denote the slope of a (weighted) least squares regression of p_i on x_i. The former slope can be computed from the observable data, but the latter generally cannot be computed since p_i is usually unobservable.

3. THE METHOD OF BOUNDS

The marginal proportions x_i and y_i may impose some restrictions upon the voting rates p_i and q_i through the *method of bounds* (Duncan *et al.*, 1952). These bounds are, in fact, a simple consequence of Frechet's (1951) bounds on joint probability distributions. They are easy to derive and can be useful for assessing bias in the model-based procedures described below.

For any probability distribution, Bonferroni's inequality implies that

$$P(X = x, Y = y) \geq \max\{P(X = x) + P(Y = y) - 1, 0\}.$$

Obviously, we also have

$$P(X = x, Y = y) \leq \min\{P(X = x), P(Y = y)\},$$

which gives a version of Frechet's (1951) bound

$$\begin{aligned}\max\{P(X = x) + P(Y = y) - 1, 0\} \\ \leq P(X = x, Y = y) \leq \min\{P(X = x), P(Y = y)\}.\end{aligned}$$

Dividing by $P(X = x)$ gives

$$\begin{aligned}\max\left\{\frac{P(X = x) + P(Y = y) - 1}{P(X = x)}, 0\right\} \\ \leq P(Y = y|X = x) \leq \min\left\{\frac{P(Y = y)}{P(X = x)}, 1\right\},\end{aligned}$$

which implies that

$$p_i^- \leq p_i \leq p_i^+ \qquad \text{and} \qquad q_i^- \leq q_i \leq q_i^+,$$

where

$$\begin{aligned}p_i^- &= \max\{(x_i + y_i - 1)/x_i, 0\} \\ p_i^+ &= \min\{y_i/x_i, 1\} \\ q_i^- &= \max\{(y_i - x_i)/(1 - x_i), 0\} \\ q_i^+ &= \min\{y_i/(1 - x_i), 1\}.\end{aligned}$$

The bounds are tight in the sense that both (p_i^+, q_i^-) and (p_i^-, q_i^+) are feasible (*i.e.*, both satisfy equation (1)). This means that we also have a tight bound for the level of polarization, *viz.*

$$\beta_i^- \leq \beta_i \leq \beta_i^+,$$

where $\beta_i^- = p_i^- - q_i^+$ and $\beta_i^+ = p_i^+ - q_i^-$.

These bounds can be aggregated to obtain a bound for the combined voting rates p and q as well as the aggregate level of polarization β. The aggregated bounds are

$$p^- \leq p \leq p^+ \qquad q^- \leq q \leq q^+ \qquad \beta^- \leq \beta \leq \beta^+,$$

where $\beta^- = p^- - q^+$, $\beta^+ = p^+ - q^-$, and, for example,

$$p^- = \frac{\sum_{i=1}^N n_i x_i p_i^-}{\sum_{i=1}^N n_i x_i}.$$

Unlike the model-based procedures discussed in the remainder of the paper, the method of bounds requires no assumptions about voting rates. Unfortunately, however, the bounds are rarely narrow enough to yield much usable information. Nonetheless, the bounds can be useful in diagnosing the likely magnitude of bias in conjunction with the results obtained in sections 5 and 6.

4. Ecological Regression Analysis

The ecological regression method for estimating p and q relies upon a regression of y_i on x_i. Polarization is estimated by the slope of this regression and majority defection is estimated by the intercept, and minority cohesion is obtained by addition,

$$\hat{\beta} = b_{yx}, \qquad \hat{q} = \bar{y} - b_{yx}\bar{x}, \qquad \hat{p} = \hat{\beta} + \hat{q}.$$

We illustrate the method with four examples.

Example 1. Loewen and Grofman (1987) present data from twenty-four precincts in the 1982 election for Auditor of Lee County, South Carolina. Here n_i is the number of persons who voted in the ith precinct, y_i is the proportion of vote received by the black candidate, and x_i is the proportion of voters who were black. A weighted least squares regression of y_i on x_i produced the estimates (with standard errors in parentheses),

$$\begin{array}{rcccccc} \hat{y}_i & = & 0.019 & + & 0.650 & x_i & \qquad R^2 = 0.737 \\ & & (0.062) & & (0.114) & & \end{array}$$

Thus, the estimated minority cohesion rate is $\hat{p} = 0.019 + 0.650 = 0.669$, the estimated majority cohesion rate is $1 - \hat{q} = 0.981$, and the estimated polarization rate is $\hat{\beta} = 0.650$. The bounds are $0.17 \leq p \leq 0.68$ and $0.01 \leq q \leq 0.51$ which provide essentially no information about the degree of polarization ($-0.34 \leq \beta \leq 0.68$).

Example 2. Loewen and Grofman (1987) also provide data on the number of registrants in each precinct of each race as well as the number of voters of each race. This suggests a different ecological regression where n_i is the number of registrants (instead of voters) in the ith precinct, y_i is the turnout rate (the proportion of registrants who voted), and x_i is the proportion of registrants who were black.

$$\begin{array}{rcccccc} \hat{y}_i & = & 0.563 & - & 0.027 & x_i & \qquad R^2 = 0.819 \\ & & (0.049) & & (0.085) & & \end{array}$$

Thus, we obtain the estimates $\hat{p} = 0.536$, $\hat{q} = 0.563$, and $\hat{\beta} = -0.027$. In this example (unlike the preceding one), we can compare these estimates to the actual turnout rates since the identify of voters, but not how they voted, is public information. Here, $p = 0.505$, $q = 0.601$, and $\beta = -0.096$, so the ecological regression method overestimates black turnout and underestimates (by a smaller amount) white turnout. Again, the bounds are uninformative: $0.26 \leq p \leq 0.84$, $0.21 \leq q \leq 0.89$, and $-0.63 \leq \beta \leq 0.63$.

Example 3. A very well known example is provided by Robinson's (1950) paper on the "ecological fallacy." Robinson calculated the correlation between the percentage of each state's population that is foreign born with illiteracy rates using data from the 1930 U.S. Census. Here n_i is the number of persons ten years old and over and x_i and y_i are the proportions of these persons who are foreign-born white and illiterate, respectively.[1] The fitted equation was

$$\begin{array}{rcccccc} \hat{y}_i & = & 0.066 & - & 0.167 & x_i & \qquad R^2 = 0.379 \\ & & (0.007) & & (0.041) & & \end{array}$$

Here the ecological regression produces an impossible estimate for p of -0.101 (in fact, $p = 0.099$), $\hat{q} = 0.066$ underestimates $q = 0.035$, and $\hat{\beta} = -0.167$ has the wrong sign ($\beta = 0.064$). Because illiteracy rates are relatively low, the bounds for p and q are narrow ($0.00 \leq p \leq 0.17$ and $0.02 \leq q \leq 0.05$), but they are not very helpful in determining the level of polarization ($-0.05 \leq \beta \leq 0.15$)

Example 4. Robinson (1950) also discusses the correlation between illiteracy rates (y_i) and the percentage of blacks (x_i) in each state.[2] The corresponding regression equation is

$$\begin{array}{rcccccc} \hat{y}_i & = & 0.020 & + & 0.248 & x_i & \qquad R^2 = 0.842 \\ & & (0.003) & & (0.017) & & \end{array}$$

[1] We were not able to exactly replicate Robinson's state level (as opposed to regional) estimates. It appears that he used illiteracy rates for the population ten years old and over from Table 10 of U.S. Bureau of the Census (1933b, p. 1229) with the percentage of *total* population foreign born from Table 11 of U.S. Bureau of the Census (1933a, p. 35). Hanushek, Jackson, and Kain (1974), whose results we were able to replicate exactly, used these data series for forty-eight states and the District of Columbia. We computed the number of foreign born whites ten years old and over as a percentage all persons ten years old and over for forty-eight states and the District of Columbia.

[2] Robinson apparently used the black percentage of total population from Table 14 of U.S. Bureau of the Census (1933a, p. 38). We computed the percentage of persons ten years old and over who were black from Table 10 of U.S. Bureau of the Census (1933b, p. 1229).

TABLE 1. Cohesion and Defection Rates in Example Datasets

Example	Cohesion p	Defection q	Polarization β
1930 U.S. Census (Robinson, 1950)			
Race-Illiteracy			
Ecological Regression	0.27	0.02	0.25
Neighborhood Model	0.08	0.04	0.04
Actual	0.16	0.03	0.14
Foreign Born-Illiteracy			
Ecological Regression	-0.09	0.06	-0.15
Neighborhood Model	0.03	0.04	-0.01
Actual	0.10	0.03	0.07
1982 Lee County (S.C.) Auditor Election (Loewen and Grofman, 1987)			
Race-Turnout			
Ecological Regression	0.54	0.56	-0.03
Neighborhood Model	0.55	0.55	-0.00
Actual	0.50	0.60	-0.10
Race-Vote			
Ecological Regression	0.67	0.02	0.65
Neighborhood Model	0.40	0.28	0.12
Actual	NA	NA	NA

The bounds are $0.00 \leq p \leq 0.41$, $0.00 \leq q \leq 0.05$, and $-0.05 \leq \beta \leq 0.37$. These estimates are only a little better than those in Example 3, as can be seen from Table 1 which summarizes all four examples.

5. GENERAL BIAS RESULTS

The customary justification for the ecological regression method is to rewrite equation (1) as

$$y_i = q + (p - q)x_i + \epsilon_i \tag{3}$$

where $\epsilon_i = (p_i - p)x_i + (q_i - q)(1 - x_i)$. It is then somewhat loosely argued that least squares can be applied to equation (3) to obtain unbiased estimates of the intercept q and the slope $p - q$.

Because the errors in equation (3) depend on x_i, several problems may arise. First, the errors may be correlated with the regressor x_i, which will bias the estimated cohesion, defection, and polarization rates. Second, the variance of ϵ_i is unlikely to be constant, so the errors are probably heteroskedastic. Third, there is no sampling model for the data so the standard

sampling theory is inapplicable. The first problem is the most serious and is the focus of this paper. The second and third problems bear on the efficiency of the estimates and the precision of the inferences. A brief comment about the error variances is in order before presenting the bias results.

Usually, the ecological regression is weighted by the number of voters in each district. Weighting by n_i has the advantage of producing a regression line that goes through the point $(\bar{x}, \bar{y})$ which is consistent with the actual proportion of votes received by each candidate. (This property is exploited in Proposition 1 below.) The usual justification offered for this weighting scheme is that it reduces heteroskedasticity (Prais and Aitchison, 1954). It does not, however, eliminate the problem. If p_i and q_i are treated as random variables that are independent of x_i, weighting by the precision of ϵ_i gives efficient estimates. There is no reason, however, to believe that the precision is proportional to n_i. In any event, the method of weighting seems to make little difference to the estimates of the slope and intercept in most ecological regression applications, and we shall adopt the standard practice of weighting the observations by n_i.

The critical issue with ecological regression concerns possible bias in the procedure due to correlation between x_i and ϵ_i. Conditions for these estimators to be unbiased are rarely stated with any precision, nor is it altogether clear what unbiasedness means in the absence of a sampling model for the data. If the data are thought to be sampled randomly from some theoretical population, then it is easy to give conditions for unbiasedness of $\hat{p}$ and $\hat{q}$ as estimators of p and q. For example, if p_i and q_i are independent of x_i and x_i, then $E(\hat{p} - p) = E(\hat{q} - q) = 0$.

For the most part, however, we are interested in making inferences about the actual, unobserved voting rates p and q in a particular election, not about some pseudo-population from which x_i and y_i could be imagined to have been drawn. To avoid such theoretical contortions, we define the bias in estimation to be just the error in estimation,

$$\text{Bias}(\hat{p}) = \hat{p} - p, \qquad \text{Bias}(\hat{q}) = \hat{q} - q.$$

Polarization will be estimated using $\hat{\beta} = \hat{p} - \hat{q}$, so $\text{Bias}(\beta) = \hat{\beta} - \beta = \text{Bias}(p) - \text{Bias}(q)$.[3]

[3]Alternatively, if we adopt the more conventional definition of bias, $\text{Bias}(\hat{\beta}) = E(\hat{\beta}) - \beta$, then the subsequent results hold with slight modification, treating x_i as fixed and p_i and q_i as random variables with $E(p_i) = \bar{p}$ and $E(q_i) = \bar{q}$. For example, Proposition 2 below becomes

$$\text{Bias}(\hat{\beta}) = \frac{\bar{h}}{\bar{x}(1-\bar{x})}((1-\bar{x})E(b_{px}) + \bar{x}E(b_{qx}) - \frac{\sigma_h^2}{\sigma_x^2}E(b_{\beta h}).$$

with the b's denoting weighted least squares regression coefficients.

A simple, but important, observation is that the biases, if any, in ecological regression estimates of minority cohesion and majority defection rates are in opposite directions and the absolute bias is always larger in estimating the minority group's cohesion rate than in estimating the majority group's defection rate. This result follows directly from the numerical properties of (weighted) least squares estimates.[4]

Proposition 1.

$$\text{Bias}(\hat{q}) = -\frac{\bar{x}}{1-\bar{x}}\text{Bias}(\hat{p}),$$

where $0 < \bar{x}/(1-\bar{x}) < 1$.

Proof. Since the (weighted) least squares regression passes through the point $(\bar{x}, \bar{y})$,

$$\bar{y} = \hat{p}\bar{x} + \hat{q}(1-\bar{x}).$$

Also from equation (2) we know that

$$\bar{y} = p\bar{x} + q(1-\bar{x}),$$

so

$$(\hat{p}-p)\bar{x} + (\hat{q}-q)(1-\bar{x}) = 0.$$

Upon rearranging we obtain

$$\text{Bias}(\hat{q}) = -\frac{\bar{x}}{1-\bar{x}}\text{Bias}(\hat{p}).$$

In addition, $0 < \bar{x} < 1/2$ implies $0 < \bar{x}/(1-\bar{x}) < 1$. □

Proposition 1 has important practical implications for the analyis of racially polarized voting. If ecological regression is biased at all, the estimate of minority group voting behavior will be more biased than the estimate of majority group voting behavior. That is, ecological regression may not be suitable for estimates of minority bloc voting. Furthermore, if the bias in estimating minority cohesion is upward (as often must be the case—minority cohesion estimates in excess of 100 percent are not unusual), the cohesion of *both* groups will be overestimated, causing ecological regression to exaggerate the degree of racial polarization as well as the level of racial bloc voting.

[4]If a different weighting scheme is employed, then a different weighted average of the residuals will equal zero, but the results below depend upon districts being weighted by the number of voters. This is, we surmise, the motivation behind the conventional weighting scheme, though its implications appear to be lost on some practitioners, e.g. Loewen and Grofman, 1989, p. 599.

Proposition 1 also implies that the ecological regression estimates of the group voting rates will be unbiased if and only if the ecological regression estimate of polarization ($\hat{\beta}$) is unbiased since

$$\text{Bias}(\hat{p}) = (1-\bar{x})\text{Bias}(\hat{\beta}) \qquad \text{Bias}(\hat{q}) = -\bar{x}\text{Bias}(\hat{\beta}).$$

Consequently, we shall focus on bias in estimating β; in many applications, β is the primary parameter of interest anyway.

Proposition 2 provides necessary and sufficient conditions for bias in ecological regression. Specifically, we show that the bias in $\hat{\beta}$ is a function of three parameters, b_{px}, b_{qx}, and $b_{\beta h}$, which we refer to as "contextual effects."[5] Each of the contextual effects is the slope from an auxiliary (weighted) least squares regression of district-level cohesion (p_i) or defection (q_i) on the minority group's size (x_i) or of polarization on district heterogeneity (h_i).

Proposition 2.

$$\text{Bias}(\hat{\beta}) = \frac{\bar{h}}{\bar{x}(1-\bar{x})}\left((1-\bar{x})b_{px} + \bar{x}b_{qx}\right) - \frac{\sigma_h^2}{\sigma_x^2}b_{\beta h}.$$

Proof. Since

$$\begin{aligned} C(x_i, y_i) &= C(x_i, q_i) + C(x_i, \beta_i x_i) \\ &= C(x_i, q_i) + C(x_i^2, \beta_i) + \bar{\beta}V(x_i) - \bar{x}C(x_i, \beta_i), \end{aligned}$$

it follows that

$$\hat{\beta} - \bar{\beta} = (1-\bar{x})b_{px} + \bar{x}b_{qx} - \frac{\sigma_h^2}{\sigma_x^2}b_{\beta h}.$$

Also,

$$\begin{aligned} \bar{\beta} - \beta &= (\bar{p} - p) - (\bar{q} - q) \\ &= -\frac{C(x_i, p_i)}{\bar{x}} - \frac{C(x_i, q_i)}{1-\bar{x}} \\ &= -\frac{\sigma_x^2}{\bar{x}(1-\bar{x})}\left((1-\bar{x})b_{px} + \bar{x}b_{qx}\right), \end{aligned}$$

so we conclude

$$\begin{aligned} \text{Bias}(\hat{\beta}) &= (\hat{\beta} - \bar{\beta}) + (\bar{\beta} - \beta) \\ &= \left(1 - \frac{\sigma_x^2}{\bar{x}(1-\bar{x})}\right)\left((1-\bar{x})b_{px} + \bar{x}b_{qx}\right) - \frac{\sigma_h^2}{\sigma_x^2}b_{\beta h}, \end{aligned}$$

[5]The ecological regression literature in the 1960's and early 1970's spun off a literature on contextual effects. Our meaning is akin to those contextual effects, but we clarify the particular form of the contextual effect (see Przeworski, 1974).

which proves the result since $\bar{h} = \bar{x}(1-\bar{x}) - \sigma_x^2$. □

Proposition 2 implies a sufficient condition for unbiasedness of the ecological regression estimate of polarization (and of the group cohesion rates): the cohesion and defection rates (p_i and q_i) must be uncorrelated with (x_i) and the polarization rate (β_i) must be uncorrelated with heterogeneity (h_i). Neither one of these assumptions alone is sufficient.

Corollary 1. *If* $p_i \perp x_i$, $q_i \perp x_i$, *and* $\beta_i \perp h_i$, *then* Bias$(\hat{\beta}) = 0$.

This condition would, of course, be satisfied if the group voting rates were constant (the so-called "constancy assumption"), but in many applications it is possible to reject constancy via the method of bounds.

The assumptions in Corollary 1 are very strong and usually untenable. Thus, it is useful to consider the sources of bias in ecological regression esimtates. Note that the bias is the difference of two components. The first term in Bias$(\hat{\beta})$ is a weighted average of b_{px} and b_{qx} times a positive coefficient ($\bar{h}/\bar{x}(1-\bar{x})$). The second term enters with a negative sign and is proportional to $b_{\beta h}$. All of the weights in this expression can be calculated from the within and between district variances of x and h. Since district-level group voting rates are unobservable, it is usually not possible to estimate these auxiliary regression coefficients, b_{px}, b_{qx}, and $b_{\beta h}$. Nonetheless, information from other sources may be available about their signs and magnitudes, which, in combination with Proposition 2, may allow us to say something about the bias in ecological regression estimates.

The slopes b_{px} and b_{qx} measure the relationship between district-level minority cohesion and majority defection rates, respectively, and minority group population size. If members of both groups who live in predominantly minority areas vote at higher rates for minority candidates than their counterparts in predominantly majority areas, then b_{px} and b_{qx} will be positive. This is quite plausible in most applications since neighborhood racial composition is usually correlated with other social and economic characteristics that influence voting. For example, whites who live in predominantly black areas tend to have lower incomes than whites who live in predominantly white areas. If lower income voters of both races are more likely to vote for the minority candidate, then the contextual effect will tend to cause an upward bias in the ecological regression estimate of polarization.

The slope $b_{\beta h}$ measures the relationship between polarization and district heterogeneity. If polarization is greater in heterogeneous districts (i.e., districts with x_i close to $1/2$) than in homogeneous districts (i.e., districts with x_i near zero or one), then $b_{\beta h}$ will be positive. A positive correlation between polarization and district heterogeneity will tend to cause a downward bias in the ecological regression estimate of polarization, while the

TABLE 2. Auxiliary Regression Statistics in Example Datasets

Example	Size-Cohesion b_{px}	Size-Defection b_{qx}	Heterogeneity-Polarization $b_{\beta h}$
1930 U.S. Census (Robinson, 1950)			
Race-Illiteracy	0.53	0.06	0.76
Foreign Born-Illiteracy	0.11	-0.26	0.53
1982 Lee County (S.C.) Auditor Election (Loewen and Grofman, 1987)			
Race-Turnout	0.22	0.02	0.76
Race-Vote	NA	NA	NA

reverse would be true of a negative correlation. Unfortunately, less is usually known about the relationship between polarization and heterogeneity than the other contextual effects. In the next section, we consider some restrictions upon the contextual effects that permit bias to be assessed without any independent information about $b_{\beta h}$.

In three of the four examples considered above, it is possible to calculate the contextual effects coefficients, which are shown in Table 2 below. For instance, in Example 2 turnout rates are higher for blacks living in predominantly black areas, which, by itself, would tend to bias the estimate of polarization upward. There is almost no correlation between white turnout and district racial composition, but there is a large positive correlation between district polarization and heterogeneity which will tend to lower the estimate of total polarization.

The coefficients for each contextual effect in the bias formula of Proposition 2 are shown in Table 3. The bias weights, unlike the contextual effects, can always be computed since they depend only upon the x_i's (which are observable). Returning to Example 2, we see that the coefficients of b_{px}, b_{qx}, and $b_{\beta h}$ in the bias formula are 0.41, 0.40, and 0.05, respectively. Thus, the presence of a substantial correlation between district heterogeneity and polarization is not enough to offset the upward bias caused by the correlation between minority cohesion and minority population size.

What the race-turnout example demonstrates is that it is not necessary to have information about all of the contextual effects to be able to characterize the bias of ecological regression estimates. When the coefficient of a contextual effect is small, as in the case of $b_{\beta h}$ in Example 2, even a substantial contextual effect will result in relatively little bias.

The other examples further illustrate this point. Although there are substantial heterogeneity-polarzation contextual effects in every example, only

TABLE 3. Coefficients of Contextual Effects in General Bias Formula for Examples

Example	Coefficient of b_{px}	b_{qx}	$b_{\beta h}$
1930 U.S. Census (Robinson, 1950)			
Race-Illiteracy	0.75	0.08	-0.01
Foreign Born-Illiteracy	0.79	0.12	-0.51
1982 Lee County (S.C.) Auditor Election (Loewen and Grofman, 1987)			
Race-Turnout	0.39	0.46	-0.03
Race-Vote	0.41	0.40	-0.05

in the foreign born-illiteracy example does this contextual effect receive a substantial weight in the bias formula. In the race-illiteracy example, for instance, it is apparent that serious bias could only arise from a correlation between the percentage of blacks in a state's population and black illiteracy rates. It was, of course, well-known that black illiteracy rates in 1930 were much higher in southern states, which also had higher proportions of blacks in their populations. Thus, the substantial upward bias of the estimated polarization rate (0.11) in this example was entirely predictable.

6. LINEAR CONTEXTUAL EFFECTS

Somewhat sharper results about the bias in ecological regression can be obtained if voting rates are assumed to be linear functions of the district racial compostion plus errors uncorrelated with x_i and h_i. We say that *linear contextual effects* hold if p_i and q_i obey

$$p_i = \lambda_0 + \lambda_1 x_i + u_i \tag{4}$$

$$q_i = \eta_0 + \eta_1 x_i + v_i, \tag{5}$$

where $\bar{u} = \bar{v} = 0$, $u_i \perp x_i$, $u_i \perp h_i$, $v_i \perp x_i$, and $v_i \perp h_i$.[6]

The assumption of linear contextual effects amounts to assuming that the errors u_i and v_i in equations (4) and (5), instead of p_i and q_i, are uncorrelated with x_i and h_i. These assumptions would hold, for example, if the micro-level voting model is a multiple linear regression and that the conditional expectation of the excluded macro-level regressors (conditional on x_i) were assumed to be linear in x_i This setup is considerably less restrictive than

[6]Langbein and Lichtman (1978: p. 59) discuss the special case where β_i is a linear function of x_i without error.

that assumed in Corollary 1 and is suggestive of what happens when the relationship between p_i, q_i, and x_i is not too nonlinear.

Linear contextual effects suggest a generalization of the ecological regression equation (3). Imposing linear contextual effects on equation (1) yields

$$y_i = \eta_0 + (\lambda_0 + \lambda_1 - \eta_1)x_i + (\lambda_1 - \eta_1)h_i + \text{error}. \tag{6}$$

Some writers (see, for example, Alt, 1994) have proposed an equation like (6), but it is evident that the four linear contextual effects parameters are not identified from the three regression coefficients in (6). The multiple regression may still be useful as a specification test. For example, if the assumptions of Corrollary 1 hold, then $\lambda_1 - \eta_1 = 0$, so the coefficient $b_{yh\cdot x}$ of h_i in a multiple regression of y_i on x_i and h_i might be used as an estimator of $\lambda_1 - \eta_1$. In two of the four examples (including the race-vote example by Loewen and Grofman (1987) as a model of ecological regression analysis, where $b_{yh} = -0.32$, as well as Robinson's foreign born-illiteracy data, where $b_{yh} = -2.18$), the coefficient of h is not close to zero.

The limitations of this test are obvious. First, there can be substantial bias in the ecological regression estimate if $\lambda_1 = \eta_1$ that the test will have no power against. Second, even under the assumptions of linear contextual effects, $b_{yh\cdot x}$ will be a biased estimate of $\lambda_1 - \eta_1$ if $u_i - v_i$ is correlated with x_i^3. Third, no sampling theory has been proposed for the test statistic.

Assuming linear contextual effects allows us to obtain a simpler expression for the bias in the ecological regression estimate of polarization that does not depend on the heterogeneity-polarization contextual effect. We will show that the bias derived under the assumption of linear contextual effects in Proposition 3 below provides a satisfactory approximation for the bias in the examples considered above.

Proposition 3. *Under the assumption of linear contextual effects,*

$$\text{Bias}(\hat{\beta}) = Ab_{px} + Bb_{qx},$$

where $A = \bar{x} - \sigma_x^2/\bar{x} + \sigma_x\gamma_x$ *and* $B = (1 - \bar{x}) - \sigma_x^2/(1 - \bar{x}) - \sigma_x\gamma_x$.

Proof. Since

$$C(x_i, h_i) = V(x_i) - C(x_i, x_i^2) = \sigma_x^2 \left(1 - 2\bar{x} - \sigma_x\gamma_x\right),$$

$\lambda_1 = b_{px}$, and $\eta_1 = b_{qx}$,

$$\frac{\sigma_h^2}{\sigma_x^2} b_{\beta h} = -(b_{px} - b_{qx}) \left(2(\bar{x} - \frac{1}{2}) + \sigma_x\gamma_x\right).$$

The result then follows from Proposition 2. □

TABLE 4. Exact and Approximate Bias Estimates for Ecological Regression Examples

Example	Exact Bias	Linear Approximation
1930 U.S. Census (Robinson, 1950)		
Race-Illiteracy	0.11	0.12
Foreign Born-Illiteracy	-0.22	-0.20
1982 Lee County (S.C.) Auditor Election (Loewen and Grofman, 1987)		
Race-Turnout	0.07	0.11
Race-Vote	NA	NA

The advantage of Proposition 3 over Proposition 2 is that the bias in the ecological regression estimate of polarization no longer depends on the heterogeneity-polarization contextual effect $b_{\beta h}$. The weights A and B for b_{px} and b_{qx} can be readily computed from the observable data.

Returning to the examples of section 3, we can compare the exact bias in the ecological regression estimate of polarization (using the formula of Proposition 2) with the linear approximation (of Proposition 3). The results are displayed in Table 4. In both of the examples from the 1930 U.S. Census, the linear approximation is very accurate. The linear approximation performs somewhat less well in the race-turnout example from the 1982 Lee County Auditor's election, though the error is not so large as to render it useless.

In the last example of Section 4, the contextual effects (and, hence, the bias) cannot be computed, but the weights on the contextual effects may help in assessing the bias in the ecological regression estimates. Table 5 presents these weights for all four examples.

Consider, for example, the foreign born-illiteracy ecological regression based on the 1930 U.S. Census. Since most immigrants settled in northern urban states, where there were better public school systems and lower illiteracy rates among non-immigrants, b_{qx} is likely to be negative (and was, as Table 2 confirms). It is less clear *a priori* whether foreign born illiteracy rates are positively or negatively correlated with the percentage of state population that is foreign born (in fact, there was a slight positive correlation), but the weight on b_{px} is small enough (0.08) that this contextual effect is unlikely to be a substantial source of bias. We would conclude, therefore, that the ecological regression estimate of polarization (in this case, the difference between foreign and native born illiteracy rates) is likely to be biased downward.

TABLE 5. Coefficients of Contextual Effects in Linear Approximation for Bias

Example	Coefficient of b_{px}	Coefficient of b_{qx}
1930 U.S. Census (Robinson, 1950)		
Race-Illiteracy	0.15	0.68
Foreign Born-Illiteracy	0.08	0.83
1982 Lee County (S.C.) Auditor Election (Loewen and Grofman, 1987)		
Race-Turnout	0.44	0.41
Race-Vote	0.33	0.48

In the race-vote example, both contextual effects, b_{px} and b_{qx}, have positive coefficients in the linear version of the bias formula, though the weight on b_{qx} is somewhat larger than the weight on b_{px}. Since whites living in integrated areas tend to be poorer and more Democratic, they are more likely to vote for a black candidate than whites living in homogeneous white areas; if true, this would make b_{qx} positive. It is also likely that blacks living in predominantly black areas are more likely to vote for a black candidate, which would mean that b_{px} is also positive. Thus, as seems plausible, if b_{px} and b_{qx} are positive, the ecological regression estimate is likely to be biased upwards. Consequently, it follows from Propositions 1 and 3, that more than 1.9 percent of white voters voted for the black candidate and that fewer than 66.9 percent of the black voters voted for the black candidate.

Proposition 3 allows us to identify graphically conditions under which ecological regression either over or underestimates the true degree of polarization. Assuming linear contextual effects, $\hat{\beta}$ is unbiased if and only if the contextual effects for p and q are exact linear functions of one another: $b_{qx} = -(A/B)b_{px}$. Although little precise information about b_{px} and b_{qx} may be available, *a priori* information may suggest ranges $[b_{px}^{-}, b_{px}^{+}]$ and $[b_{qx}^{-}, b_{qx}^{+}]$ of plausible values for b_{px} and b_{qx}, respectively; these values correspond to points in the shaded rectangle in the figure. The parallel diagonal lines in the figure show the minimum and maximum level of bias associated with these values of b_{px} and b_{qx}. If A and B are positive, then

$$Ab_{px}^{-} + Bb_{qx}^{-} = \text{minimum bias}$$

and

$$Ab_{px}^{+} + Bb_{qx}^{+} = \text{maximum bias}.$$

Ecological regression estimates will be unbiased if and only if the point (b_{px}, b_{qx}) lies on a diagonal line through the origin with the same slope as

the lines shown. Points (b_{px}, b_{qx}) above this line represent situations where ecological regression will exaggerate the degree of racial polarization, and points below this line are cases where ecological regression understates the degree of polarization. Therefore, if A and B are positive and if we believe that b_{px} and b_{qx} are positive as well, then ecological regression will always overstate the degree of racial polarization.

The quantities A and B need not both be positive. Three other cases deserve mention. If both A and B are negative, then the line depicting unbiasedness of ecological regression remains the same, but points below the line are situations where ecological regression exaggerates the degree of polarization and points above the line are cases where ecological regression understates the polarization. If $A < 0$ and $B > 0$ then the line describing situations where ecological regression is unbiased has a positive slope. Like the case where A and B are both positive, contextual effects above the line will overstate the degree of polarization, and effects below the line will understate the polarization. If $A > 0$ and $B < 0$, the line remains upward-sloping, but points below the line will represent situations where ecological regression overstates polarization and points above the line are cases where ecological regression understates the polarization.

In practice, then, the signs of the the quantities A and B will be quite important in determining the direction of the bias in ecological regression estimates of polarization. Note that $A > 0$ if and only if

$$\gamma_x > \frac{\sigma_x}{\bar{x}} - \frac{\bar{x}}{\sigma_x},$$

and $B > 0$ if and only if

$$\gamma_x < \frac{1-\bar{x}}{\sigma_x} - \frac{\sigma_x}{1-\bar{x}}.$$

In other words, A is positive if the skewness of x exceeds the coefficient of variation of x minus the reciprocal of that coefficient. Similarly, B is positive if the skewness of x exceeds the reciprocal of the coefficient of variation of $1 - x$ minus that coefficient.

The likely magnitude of the bias in the ecological regression estimate of the degree of racial polarization in Example 1 depends on the magnitudes of the auxiliary regression coefficients b_{px} and b_{qx}. Different beliefs about which values of these coefficients are plausible will generate different conclusions about the direction and magnitude of bias. Our purpose here is to show how essentially qualitative information, combined with the precinct-level bounds, can yield relatively precise quantitative estimates of bias.

To illustrate how auxiliary information might be utilized, suppose we believe that the rate at which blacks vote for black candidates is mostly uncorrelated with neighborhood characteristics, but that whites living in predominantly black neighborhoods are more likely to vote for black candidates than whites living in predominantly white areas.

The precinct level bounds on p_i and q_i shown in Figures 2 and 3 show that the data are not inconsistent with these claims. In these figures, the dots represent the percentage the percentage of vote received by the black candidate in each precinct, while the bars represent the range of possible voting rates for each group consistent with the bounds. Ecological regression fits a line through the scatterplot of dots (which are the same in each figure).

Figure 2 shows that the data are uninformative about the voting rates of blacks living in mostly white areas, while there must be considerable variability in the voting rates of blacks living in mostly black areas. The correlation between p_i and x_i could be positive, negative, or zero. Values of b_{px} in the range of $[0.1, 0.2]$ would be consistent with both our prior beliefs and the current dataset.

Figure 3 shows that whites living in homogeneous white areas voted at very low rates for the black candidate, but the data are uninformative about the behavior of whites living in predominantly black precincts. If we assume that whites never vote for the black candidate at higher rates blacks (so the upper bound on q_i is y_i, represented by the dots in the figure), then the maximum value of b_{qx} is about one and, more plausibly, somewhat smaller, say 0.8. Neither data nor theory are very helpful in establishing a lower bound for b_{qx}, but a lower bound of 0.2 is reasonable.

The *a priori* bounds $0.0 \leq b_{px} \leq 0.2$ and $0.2 \leq b_{qx} \leq 0.8$ are, in fact, the values shown by the shaded box in Figure 1. With an assumption of linear contextual effects, we conclude (using Proposition 3 and the data in Table 5) that the bias in $\hat{\beta}$ lies in the interval,

$$0.33 \times 0.1 + 0.48 \times 0.2 \leq \text{Bias}(\hat{\beta}) \leq 0.33 \times 0.2 + 0.48.8,$$

i.e., $0.13 \leq \text{Bias}(\hat{\beta}) \leq 0.45$, approximately. The implied estimate of polarization, after adjusting for bias, would be between 0.20 and 0.52, compared to the original ecological regression estimate of 0.65. Values at the upper end of this range are still consistent with a high level of racially polarized voting, while values at the lower end would suggest a much more modest level of racially polarized voting.

7. NEIGHBORHOOD ESTIMATES

We briefly consider the properties of an alternative estimator suggested by Klein, Sacks, and Freedman (1991). In contrast to ecological regression methods which depend upon a lack of correlation between group cohesion and size, they propose a "neighborhood model" which assumes that all groups within a single neighborhood or district behave similarly. The basic idea behind the neighborhood model is to estimate group voting rates p_i and q_i by the district voting rate y_i. The corresponding estimates of p and q are

$$\tilde{p} = \frac{\sum_{i=1}^{N} n_i x_i y_i}{n\bar{x}} \qquad \tilde{q} = \frac{\sum_{i=1}^{N} n_i (1-x_i) y_i}{n(1-\bar{x})}.$$

This procedure will yield exact estimates of p and q if there is no polarization within each district. The lack of polarization within each district does not imply that total polarization is zero, but it always yields smaller estimates of polarization (in absolute value) than ecological regression. Let $\tilde{\beta} = \tilde{p} - \tilde{q}$ be the estimate of polarization obtained using the neighborhood model.

Proposition 4. *$\tilde{\beta} = \theta\hat{\beta}_b$, where $0 < \theta \leq 1$.*

Proof. From equation (6),

$$\begin{aligned}
\tilde{\beta} &= \frac{\sum_{i=1}^{N} n_i x_i y_i}{\sum_{i=1}^{N} n_i x_i} - \frac{\sum_{i=1}^{N} n_i (1-x_i) y_i}{\sum_{i=1}^{N} n_i (1-x_i)} \\
&= \frac{(1-\bar{x})C(x_i, y_i) + \bar{x}(1-\bar{x})\bar{y}}{\bar{x}(1-\bar{x})} - \frac{\bar{x}\bar{y} - C(x_i, y_i) - \bar{x}(1-\bar{x})\bar{y}}{\bar{x}(1-\bar{x})} \\
&= \frac{C(x_i, y_i)}{\bar{x}(1-\bar{x})} = \frac{\sigma_x^2}{\bar{x}(1-\bar{x})}\hat{\beta}.
\end{aligned}$$

□

The assumption of within district constancy is apparently rather strong (though perhaps no worse than between district constancy), so it is useful to have a general bias result for the neighborhood model.

Proposition 5.

$$\text{Bias}(\tilde{\beta}) = -\frac{\sigma_h^2 b_{\beta h} + \bar{\beta}\bar{h}}{\bar{x}(1-\bar{x})}.$$

Proof. Proposition 4 implies that

$$\begin{aligned}\tilde{\beta} &= \frac{C(x_i, q_i) + C(x_i^2, \beta_i) + \bar{\beta}\sigma_x^2 - \bar{x}C(x_i, \beta_i)}{\bar{x}(1-\bar{x})} \\ &= -\frac{C(h_i, \beta_i) - \bar{\beta}\sigma_x^2}{\bar{x}(1-\bar{x})} - \frac{C(p_i, x_i)}{\bar{x}} - \frac{C(q_i, x_i)}{1-\bar{x}} \\ &= \beta - \frac{\sigma_x^2 b_{\beta h} + \bar{\beta}\bar{h}}{\bar{x}(1-\bar{x})},\end{aligned}$$

using an argument similar to that in the proof of Proposition 2. □

One implication of Proposition 5 is that if there is some polarization and it is in the same direction in each district (*e.g.*, $p_j \geq q_j$), then the neighborhood model always underestimates the total amount of polarization. In general, however, little can be said about the direction and magnitude of the bias.

The neighborhood model also provides an alternative interpretation for the ecological regression coefficients. It is straightforward to show that using the predicted values from the ecological regression,

$$\tilde{p}_i = \tilde{q}_i = \hat{q} + \hat{\beta}\bar{x}_i,$$

in place of y_i in (eq. no) yields the same estimates $\tilde{p}$ and $\tilde{q}$. Klein *et al.* (1991) call this the "linear neighborhood model," and argue that there is no way to determine whether ecological or neighborhood estimates are closer to the truth using only the aggregate data.

Table 1 in section 4 presents the neighborhood model estimates for the four examples considered earlier. In the three cases where the actual result is known, the neighborhood estimates understate the true degree of polarization. Comparing the neighborhood estimates and ecological regression estimates of p, q, and β in Table 1 suggests that there is no reason to prefer one set of estimates to the other. In the South Carolina auditor's election ecological regression performed better; in the 1930 Census data on Foreign Born and Illiteracy the neighborhood estimates were closer to the actual results; and in the 1930 Race and Illiteracy the actual rates of cohesion, defection, and polarization split the difference between the two models. Which estimates researchers should use depends on contextual information beyond that contained in aggregate demographic and electoral data.

The analysis of bias at the end of Section 6 suggested that the true level of racial polarization is probably between 0.20 and 0.52. Both the neighborhood ecological regression estimates of polarization lie outside this interval—but on opposite sides, with the neighborhood estimate at 0.12 and the ecological regression estimate at 0.65. For many other problems, the pattern

is likely to be similar. In such cases, careful application of the bias results derived here may help to narrow the range of plausible estimates.

8. REFERENCES

C. H. Achen and W. P. Shively, *Cross-Level Inference* (Chicago: University of Chicago Press, 1995).

J. E. Alt, "The Impact of the Voting Rights Act on Black and White Voter Registration in the South," in *Quiet Revolution in the South* (Princeton: Princeton University Press, 1994), pp. 351–77.

H. R. Alker, Jr. "A typology of ecological fallacies," in *Quantitative Ecological Analysis in the Social Sciences* (Cambridge: MIT Press, 1969), pp. 69–86.

C. S. Bullock, "Misinformation and Misperceptions: A Little Knowledge Can Be Dangerous," *Social Science Quarterly*, vol. 72 (1991), pp. 834–39.

O. D. Duncan, R. P. Cuzzort, and B. Duncan, *Statistical Geography: Problems in Analyzing Areal Data* (Glencoe: Free Press, 1961).

M. Frechet, "Sur las tableaux de correlation dont les marges sont donnees," *Annals de Universite de Lyon A*, vol. 20 (1951), pp. 13–31.

D. A. Freedman, S. P. Klein, J. Sacks, C. A. Smyth, and C. G. Everet, "Ecological Regression and Voting Rights," *Evaluation Review*, vol. 15 (1991), pp. 673–711.

B. Grofman, "Straw Men and Stray Bullets: A Reply to Bullock," *Social Science Quarterly*, vo. 72 (1991), pp. 840–43.

B. Grofman, "Statistics Without Substance: A Critique of Freedman *et al.* and Clark and Morrison," *Evaluation Review*, vol. 15 (1991), pp. 746–69.

G. King, *A Solution to the Ecological Inference Problem*, forthcoming.

S. P. Klein, J. Sacks, and D. A. Freedman, "Ecological regression *versus* the secret ballot," *Jurimetrics*, vol. 31 (1991), pp. 393–413.

L. I. Langbein and A. J. Lichtman, *Ecological Inference* (Beverly Hills: Sage, 1978).

A. J. Lichtman, "Passing the Test: Ecological Regression Analysis in the Los Angeles Case and Beyond," *Evaluation Review*, vol. 15 (1991), pp. 770-99.

J. W. Loewen and B. Grofman, "Recent developments in methods used in vote dilution litigation," *The Urban Lawyer*, vol. 21 (1989), pp. 589–604.

S. J. Prais and J. Aitchison, "The Grouping of Observations in Regression Analysis," *Review of the International Statistical Institute*, vol. 22 (1954), pp. 1–27.

A. Przeworski, "Contextual Models of Political Behavior," *Political Methodology*, vol. 1 (1974), pp. 27–61.

D. Rivers, *Statistics of Voting*, (forthcoming).

U.S. Department of Commerce, Bureau of the Census, *Fifteenth Census of the United States: 1930*, vol. II (Washington: Government Printing Office, 1933).

J. K. Wildgen, "Adding *Thornburg* to the Thicket: The Ecological Fallacy and Parameter Control in Vote Dilution Cases," *The Urban Lawyer*, vol. 20 (1988), pp. 155–73.

J. K. Wildgen, "Vote Dilution and Cold Fusion Technology," *The Urban Lawyer*, vol. 22 (1990), pp. 489–502.

DEPARTMENT OF POLITICAL SCIENCE, MASSACHUSETTS INSTITUTE OF TECHNOLOGY, CAMBRIDGE, MA 02139
E-mail address: sda@mit.edu

DEPARTMENT OF POLITICAL SCIENCE, STANFORD UNIVERSITY, STANFORD, CA 94305
E-mail address: rivers@stanford.edu