Pergamon

Political Geography, Vol. 16, No. 8, pp. 675–690, 1997

Printed in Great Britain
0962-6298/97 $17.00 + 0.00

PII: S0962-6298(96)00072-8

Estimating the likelihood of fallacious ecological inference: linear ecological regression in the presence of context effects

GUILLERMO OWEN

Department of Mathematics, Naval Post-Graduate School, Monterey, California, 93943, USA

AND

BERNARD GROFMAN

School of Social Sciences, University of California, Irvine, CA 92697, USA

ABSTRACT. In situations where the only reliable data source is electoral data at the aggregate level for a geographic unit such as voting precincts, social scientists have sought to use ecological regression techniques to recreate the voting behavior of particular groups without committing the ecological fallacy of the sort warned of by Robinson (1950). Until quite recently, the most common use of ecological regression techniques was for the analysis of historical data, e.g., determining the social groups that supported the Nazis. In the US, however, in a development little known to political geographers, within the past decade, social science expert witnesses have testified in hundreds of cases involving minority voting rights—cases where a central issue has been to ascertain the extent of white/Anglo support for black or Hispanic (or more recently, Asian-American) candidates in elections involving candidates of more than one race/ethnic background. One issue in several recent voting rights cases has been whether linear regression methods can validly be used to model racial voting patterns, with experts for defendant jurisdictions claiming that the presence of contextual effects may invalidate the assumptions underlying linear models and give rise to an ecological fallacy. For data on voting patterns in small geographic units such as voting precincts, we show that the presence of a plausible type of race-related contextual effect will, in general, lead to a quadratic relationship between minority share of the electorate and the share of the vote received by the minority candidate. However, when either a substantial proportion of the electorate is located in racially homogeneous or near homogeneous precincts (as often occurs given the patterns of residential concentration in US cities), or when the contextual effects are small, we show that the estimates of the key parameters of interest in voting rights litigation, namely the proportions of minority and non-minority voters who voted for the minority candidate(s), will be very well approximated by a bivariate linear model. In such circumstances, fallacies of ecological inference can be avoided with near certainty, despite the fact that the linear model that has been used to fit the data omits potentially important variables and regression diagnostics for it would show hetero-skedasticity, while a quadratic regression would provide coefficients that are

essentially uninterpretable in terms of the parameter of interest. This result has potentially broader implications for the reliability of ecological inference because it demonstrates that, under plausible assumptions, a simple ecological model can yield accurate results despite theoretical errors in model specification—because those errors prove to be of little or no practical significance.

Introduction

Often political geographers or other social scientists wish to estimate how particular social groups voted, or how voting support in one election translated into voting support in other elections. In situations where reliable survey data is unavailable, social scientists have used electoral data at the aggregate level grouped at the level of some geographic unit such as voting precincts to develop estimates of the parameters of interest through use of ecological regression techniques. In using such techniques, if they seek to move from assertions about what is true of aggregate units to claims about what is true of the individuals inside those units (i.e., from statements that units where people of a given characteristic (set of characteristics) live vote in a particular way to statements about the (average) voting behavior of people with a given characteristic (set of characteristics), then social scientists risk being accused of committing a fallacy of ecological inference of the sort warned of by Robinson (1950).

The combination of availability of survey data and fear of ecological fallacy has greatly diminished the use of ecological techniques for analysis of aggregate data in political science and in sociology. Many recent uses of ecological regression techniques in the published social science literature have been for the analysis of historical data in periods prior to the existence of large-scale social surveys, e.g., determining the social groups that supported the Nazis (Brown, 1982; Falter and Zintl, 1988). Moreover, even when ecological data is used for purpose of analysis, the form of its use is rarely such as to permit detailed inference that is directly interpretable as an estimate of the proportion of members of a given group that exhibit a given behavior (Grofman, 1995). In the US, however, in a development little known among political geographers, within the past decade, following the lead of scholars such as Goodman (1953) and Duncan and Davis (1953), in hundreds of cases involving minority voting rights, social science expert witnesses have testified in court by making use of aggregate ecological techniques to estimate the voting of different racial or ethnic blocs. In these cases a central factual issue has been the extent of white/Anglo support for black or Hispanic (or more recently, Asian-American or Native American) candidates in elections involving candidates of more than one race/ethnic background, as compared to the support for those candidates that comes from minority voters.

Experts in voting rights cases have developed a useful modification of Goodman's methods to deal with turnout differences among racial and ethnic groups in situations where the only data available on the racial composition of ecological units is based on total population or voting age population, the so-called double-equation method (Grofman *et al.*, 1985; Grofman and Migalski, 1988; Loewen and Grofman, 1989; Grofman, 1992).[1] However, despite this improvement, uses of ecological regression and related techniques to estimate levels of racial bloc voting have not been without challenge. In several recent voting rights cases experts for defendant jurisdictions have claimed, *inter alia*, that the presence of contextual effects may invalidate the assumptions underlying linear models and give rise to an ecological fallacy that will usually overstate

the extent to which members of a group vote in support of members of their own group in multi-racial election contests (see esp. Freedman *et al.*, 1991; and rebuttals in Grofman, 1991a; Lichtman, 1991; Loewen *et al.*, 1993).

Here we wish to present new evidence on the likely accuracy of linear ecological regression techniques that is directly relevant to what is still an ongoing controversy (see e.g., Wildgren, 1993; Grofman, 1993a, 1993b; Firebaugh, 1993) we wish also to provide important reasons to believe that ecological inference may, be able under quite plausible conditions, to yield reasonable inferences about (average) group behavior at the individual level even when the linear ecological regression model is not accurately specified due to the failure to take into account aspects of group behavior that can be modelled in terms of social context.

In applying the linear ecological regression to estimate racial bloc voting, we shows that the accuracy of the linear model will depend on the nature of the distribution of minority voting strength and on the magnitude of the probable contextual effects. In particular, if most whites/Anglos live in areas geographically segregated from those of most minority members and/or if most members of the minority group live in geographically segregated neighbourhoods and/or if the magnitude of the context effect is relatively small and can be well approximated by a quadratic relationship, then linear ecological methods will provide good estimates of average levels of bloc voting of both the minority and the non-minority electorate despite the fact that a linear model omits potentially important variables that would give rise to what appear to be contextual effects.

This result has potentially broader implications for the reliability of ecological inference because it demonstrates that, under plausible assumptions, a simple linear ecological model can yield accurate results despite theoretical errors in model specification and even when regression diagnostics for it would show heteroskedasticity associated with curvilinearity, because the errors prove to be of little or no practical significance. Moreover, the results given below suggest that, under certain circumstances, for purposes of ecological inference, simple models may be more appropriate than more complex models because the coefficients in a simple model can be directly translated into estimates of the individual-level parameters of interest (e.g., the estimated proportion of minority voters who support a given minority candidate), whereas, in the presence of multicollinearity, quadratic (and higher order models or multivariate models) may lead to results that are essentially uninterpretable in terms of estimating the (average) behavior of members of a given group.

Bivariate linear and quadratic models

The basic bivariate model

Consider the voting patterns of two groups of voters. For convenience we shall refer to these as white voters and black voters and assume that the two groups are mutually exclusive and exhaustive. We wish to model the voting behavior of the two groups, using aggregate data (gathered, let us say, at the precinct level). In particular we wish to understand what proportion of the votes of each group go to a candidate (or candidates) identified with each group. For simplicity we assume that there is a single black candidate. Using a notation which has been used in the literature on voting rights

issues (see, e.g., Grofman *et al.*, 1985; Grofman and Migalski, 1988; Grofman *et al.*, 1992), let

x = the proportion of the electorate that is white
P'_B = the proportion of the electorate that votes for the black candidate
P'_W = the proportion of the electorate that votes for the white candidate
P'_{WW} = the proportion of white voters who vote for the white candidate
P'_{BW} = the proportion of black voters who vote for the white candidate
P'_{BB} = the proportion of black voters who vote for the black candidate
P'_{WB} = the proportion of white voters who vote for the black candidate

If the relationship between P'_W and x is linear, with the parameters such as P'_{WB} roughly identical across precincts (subject to random error) we will obtain a linear approximation:

$$P'_W = (P'_{WW} - P'_{BW})x + P'_{BW}. \tag{1}$$

Here we have a linear relationship, $y = sc + r$, where

$$s = P'_{WW} - P'_{BW}$$

and

$$r = P'_{BW}.$$

We may rewrite this to solve for P'_{WW} and P'_{BW} to obtain:

$$P'_{WW} = r + s \tag{2a}$$

$$P'_{BW} = r. \tag{2b}$$

Now we may take the values of P'_{WW} and P'_{BW} estimated from ecological regression to be the *mean* values for white and black support of the white candidate.[2]

With certain key modifications, e.g., to take into account different levels of turnout among white and black voters (see Grofman *et al.*, 1985; Grofman and Migalski, 1988) and to impose various types of cross-checks for validity (see, e.g., Loewen and Grofman, 1989; Grofman *et al.*, 1992, Ch. 4), this approach has become the standard approach to estimating patterns of racial bloc voting in voting rights litigation in the US in situations where reliable survey data broken down by race dealing with the relevant elections (i.e., election contests of the type at issue in the litigation and contests involving both minority and non-minority candidates) is not available. Findings based on its use presented in trial testimony by many different expert witnesses testifying both for plaintiffs and for defendant jurisdictions have been accepted by well over 100 courts, including the US Supreme Court (Loewen, 1982; Engstrom and McDonald, 1988; Grofman, 1992; Grofman *et al.*, 1992).

The approach described above can be seen as a straightforward modification of Goodman's cross-temporal ecological estimation technique (Goodman, 1953) so as to make it applicable to cross-sectional voting behavior data.

Contextual effects affecting racial bloc voting

The simple model above assumes that there are only two (mutually exclusive) groups whose behavior we wish to describe, and that it is appropriate to model each group's behavior with a single parameter, e.g. P'_{WW} or P'_{BW} which can be thought of as the probability that a randomly chosen member of the given group will vote/behave/choose in a given way (e.g., vote for the white/Anglo candidate) in a dichotomous choice situation. Such an assumption makes sense if the group members are essentially homogeneous with respect to attributes (other than group membership) that are also strongly related to the behavior in question, or if differences among group members are essentially unrelated to what proportion of the ecological unit consists of members of that group. In other words, the linear model would seem to make most sense if differences in the behavior of members of the group could be thought of as essentially randomly distributed with respect to the independent variable (group membership proportion), i.e., if errors were non-heteroskedastic.

But there are many reasons to expect that errors will be non-random with respect to the group membership proportion. For example, in a partisan context, poor blacks may be somewhat more likely to vote for black Democratic candidates than may wealthier blacks (more of whom might be Republican), and we might expect that the greater the proportion of blacks in a precinct the higher the proportion of those blacks that will be poor. Thus we might expect that inner-city blacks would be somewhat more likely to support black Democratic candidates than would black voters who live in predominantly white suburbs. If so, then errors will not be uncorrelated and linear estimates will be biased.[3]

In order to improve model specification, there are two ways in which we might seek to compensate for differences in group voting behavior that are related to the independent variable. On the one hand we might use a multivariate approach in which we sought to separately estimate, say, the voting behavior of rich blacks and poor blacks and/or rich whites and poor whites (perhaps by introducing a multiplicative dummy variable, perhaps by introducing a control for median income in the precinct). On the other hand, if the expected differences in the voting behavior of a group's members were closely linked to that group's proportion of the population, we might maintain our univariate model but explicitly introduce contextual effects tied to the group's percentage in the ecological unit.

It is the latter approach that we pursue here.[4] In so doing we act as if minority population proportion in the ecological unit is a surrogate for a whole host of other variables that may give rise to differences in the (voting) behavior of members of a group. In practical terms this is not an unreasonable assumption, since, e.g., inner-city blacks will differ on a large number of dimensions from blacks who live in predominantly white areas—albeit what they will share is their race and this may for certain purposes be virtually all that matters.

A univariate quadratic model

Let us modify the bloc voting model of the first section by considering what happens if we posit a linear context effect of the form:[5]

$$P'_{WW} = a1(1 - x) + b_1 \tag{3a}$$

$$P'_{BW} = a2(1 - x) + b_2. \tag{3b}$$

Here we are positing that the extent of white or black support for white candidates is contingent on the racial composition of the precinct. It is apparent from these expressions that the greater the (absolute) value of the parameters a_1 and a_2 relative to b_1 and b_2 the greater the contextual effect.

Using our earlier notation:

$$\begin{aligned} P'_W &= (a1(1-x) + b_1)x + (a_2(1-x) + b_2)(1-x) \\ &= (a1 + b_1)x - a_1x^2 + (a_2 + b_2) - a_2x - (a_2 + b_2)x + a_2x^2 \\ &= (a2 - a_1)x^2 + (a_1 + b_1 - 2a_2 - b_2)x + (a_2 + b_2). \end{aligned} \tag{4}$$

Thus, if there is a (linear) context effect, the vote for the white candidate is a *quadratic* function of the proportion black (minority) in the electorate,[6] which we may represent as $y = Cx^2 + Bx + A$.

The context effect may work in one of two directions: for example, either white support for the black candidate could increase with percent black or it could decrease. The second type of context would occur under the 'black threat' hypothesis posited by Key (1949). James Alt (1994) finds this type of context effect present when he finds anti-black activity by whites in the pre-Voting Rights Act South most likely to occur in the jurisdictions where there are the greatest number of blacks. Similarly, it is well known that white support for the Democratic presidential nominee has, since 1964, been lower, on average, in states with very high black populations (Black and Black, 1992). On the other hand, of course, there are also situations where it would be plausible to expect that white behavior would look most like black behavior in the areas of greatest black concentration. For example, in the US, if we were looking at local elections that were partisan, then whites in heavily black precincts might be more likely to be Democratic supporters than would whites elsewhere and thus more likely to vote for a black Democratic nominee than would whites elsewhere. This would generate a context effect of the first type.

However, as long as either the context effect is small (i.e., a_1 (a_2) is smaller than $b_1(b_2)$), *or* a significant fraction of the electorate is located in racially homogeneous or near homogeneous precincts, *and* we also have a considerable range in both the dependent and independent variables, the errors caused by nonlinearity should be relatively trivial. In the next section we will make this assertion more precise.

The goodness of fit of a linear model when the 'true' model is a quadratic context effect

For the posited quadratic relationship we cannot recover the exact nature of the contextual relationship between support for the white candidate and racial composition (i.e., the parameters specified in *Equations* 3a and 3b) from the coefficients of the fitted regression equation because we have more parameters (four: a_1, b_1, a_2, b_2) to estimate than we have coefficients of our estimating equation (three: A, B, C). Nonetheless, as we will show below, for any particular racial population distribution across precincts, we can use the best-fit coefficients of the quadratic to estimate mean values of *differences* between black and white behavior, i.e., to estimate $P'_{BW} - P'_{WW}$ ($= P'_{WW} - P'_{BW}$).

Uniform distribution

If there were a uniform distribution of voters across the geographic units, then the quadratic function given by

$$y = cx^2 + bx + a \tag{5}$$

would have a linear least squares best fit given by

$$y \approx cx + bx + a - c/6 = (c + b)x + a - c/6 \tag{6}$$

since, for a uniform distribution of x over the [0, 1] interval,

$$x^2 \approx x - 1/6$$

is best in a least squares sense.

Recall, now, that we have posited (Equation 4) that

$$y = (a_2 - a_1)x^2 + (a_1 + b_1 - 2a_2 - b_2)x + (a_2 + b_2).$$

Hence, our linear approximation to this equation can be restated, using *Equation* (6), as

$$y \approx (b_1 - b_2 - a_2)x + (5a_2 + 6b_2 - a_1)/6. \tag{7}$$

For the uniform distribution, by entering the slope and intercept of *Equation* (7) into *Equation* (2), the estimate of the critical mean racial bloc voting parameters derived from linear ecological regression could be approximated as:

$$\begin{aligned} P'_{BW} &= r \\ &= (5a_2 + 6b_2 - a_1)/6 \end{aligned} \tag{8a}$$

$$\begin{aligned} P'_{WW} &= r + s \\ &= (5a_2 + 6b_2 - a_1)/6 + (b_1 - b_2 - a_2) \\ &= (6b_1 - a_2 - a_1)/6. \end{aligned} \tag{8b}$$

Now we are in a position to see how well the estimate of white and black voting behavior obtained from a linear ecological regression approximates the values derived from what we take to be the 'true' quadratic contextual effects model. However, rather than looking at expected errors in estimating the individual parameters P'_{BW} and P'_{WW} we shall focus on the estimated *difference* between black and white voting behaviour (i.e., $P'_{BB} - P'_{WB} = P'_{WW} - P'_{BW}$) since it is this difference that is critical in judging whether or not voting is polarized along racial lines:

$$P'_{WW} - P'_{BW} = a_2 + b_2 - b_1) = s. \tag{9}$$

But, for the previously given contextual model, using *Equation* (3), the actual mean difference between white and black voting behavior (e.g., average $P'_{WW} - P'_{BW}$) would be given by

$$\frac{1}{h}\int (1-x)\,(a_2(1-x)+b_2)\,d\mu - \frac{1}{1-h}\int x(a_1(1-x)+b_1)\,d\mu \tag{10}$$

where μ is the distribution of x, and

$$h = \int x\,d\mu.$$

With a uniform distribution over the [0, 1] interval, we will have $h = 1/2$, and *Equation* (10) becomes

$$2\int_0^1 (1-x)\,(a_2(1-x)+b_2)\,dx - 2\int_0^1 x(a_1(1-x)+b_1)\,dx$$
$$= 2a_2/3 + b_2 - b_1 - a_1/3 \tag{11}$$

Clearly, if a^1 and a_2 are small relative to the other parameters, then the linear approximation to the quadratic is a very good approximation. Now a_1 is a measure, in effect, of the difference in white voting between when whites are in the majority and when whites are in the minority; a_2 is similarly a measure of the difference in black voting between when blacks are in the majority and when blacks are in the minority. The order of magnitude of the effect (the difference between *Equations* 9 and 11) would be given by

$$(a_1 + a_2)/3$$

If there were small context effects of the second type, i.e., if both a_1 and a_2 were positive, so that both white and black voters were more likely to support black candidates in the heavily blacker areas than elsewhere, then linear ecological regression could be expected to slightly overestimate the extent of racial bloc voting, but the values of expressions (9) and (11) will not be greatly different unless it should happen that the effects measured by a_1 and a_2 are of the same order of magnitude as the differences between the average white voter and the average black voter—*which seems very unlikely in the US socio-political context.* Indeed, if there were small context effects of the first type, i.e., if a_1 and 2_2 were of opposite sign, so that there was less support from whites for black candidates in more heavily black areas, then linear ecological regression could be expected to be virtually perfect in approximating the average extent of racial bloc voting differences to the extent that a_1 and a_2 were of comparable magnitude.

Density concentrated at end points

Of course, the variable x is not likely to be uniformly distributed over the unit interval. In most US cities, at least when we are dealing with black and white population groupings, x is concentrated near the end-points 0 and 1. We might therefore imagine a different density:

$$f(x) = \begin{cases} \frac{1}{2}k & 0 \le x \le k \\ 0 & 0 \le x \le 1-k \\ \frac{1}{2}k & 1-k \le x \le 1 \end{cases} \tag{12}$$

where k is a small positive number, of the order (perhaps) of 0.1.

For this assumed distribution, the least squares approximation would be

$$x^2 \approx x - q$$

where q is the average value of $x - x^2$, for the density given by (12). This is readily calculated to equal

$$q = (k/2) - (k^2/3), \tag{13}$$

and the least squares approximation to the quadratic given by *Equation* (5) will be

$$y \approx (b + c)x + a - cq \tag{14}$$

with q as in *Equation* (13).

As before, we have

$$y = (a_2 - a_1)x^2 + (a_1 + b_1 - 2a_2 - b_2)x + (a_2 + b_2).$$

Hence, our linear approximation to this equation can be restated, for our new assumptions about distribution, using *Equation* (14), as

$$y \approx (b_1 - b_2 - a_2)x + a_2 + b_2 - (a_2 - a_1)q. \tag{15}$$

Here q is as defined in *Equation* (13).

From *Equation* (2), we would estimate the critical mean racial bloc voting parameters derived from the slope and intercept of *Equation* (15) as

$$\begin{aligned} PP'_{BW} &= r \\ &= a_2 + b_2 - (a_2 - a_1)q \end{aligned} \tag{16a}$$

$$\begin{aligned} P'_{WW} &= r + s \\ &= a_2 + b_2 - (a_2 - a_1)q + (b_1 - b_2 - a_2) \\ &= b_1 - (a_2 - a_1)q. \end{aligned} \tag{16b}$$

Using the linear model, the *estimated* difference between black and white voting behavior (e.g., $P'_{WW} - P'_{BW}$) is again given by

$$\text{est. } P'_{WW} - P'_{BW} = a_2 + b_2 - b_1. \tag{17}$$

But, for the quadratic contextual model, using *Equation* (3), and our new assumptions as to the nature of the distribution of x, the actual difference between black and white voting behavior (from *Equation* 10) would now be given by

$$\begin{aligned} \text{avg. } P'_{BB} &= \tfrac{1}{k}\left\{ \int_0^k (1 - x)\,(a_2(1 - x) + b_2)\,dx + \int_{1-k}^{1} (1 - x)\,(a_2(1 - x) + b_2)\,dx \right\} \\ &= a_2 + b_2 - ka_2(1 - 2k/3) \end{aligned}$$

while

$$\text{avg. } P'_{WB} = \frac{1}{k}\left\{\int_0^k x\,(a_2(1-x)+b_1)\,dx + \int_{1-k}^{1} x\,(a_1(1-x)+b_2)\,dx\right\}$$
$$= b_1 + ka_1(1-2k/3)$$

Thus we obtain, on the average,

$$P'_{BB} - P'_{WB} = a_2 + b_2 - b_1 - k(a_1 + a_2)\,(1 - 2k/3) \tag{18}$$

and the difference between the estimated value of racial differences from the linear ecological regression, given by (17), and the 'true' average value from the quadratic context effects model, given by (18), is this last term:

$$\text{error} = k(a_1 + a_2)\,(1 - 2k/3). \tag{19}$$

Since we have assumed that k is small, say of the order of 0.1 or smaller, this error will be quite small—unless the sum $a_1 + a_2$ is actually quite large compared to the true value of $P'_{BB} - P'_{WB}$, i.e., unless there is considerably greater difference between the behavior of blacks [whites] when they are in the majority and blacks [whites] when they are in the minority than there is between that of the average black and that of the average white.

Since this seems to be far from the true case in our current socio-political context, we conclude that, as far as obtaining the average difference between white and black voting is concerned, our linear approximation will be quite good. Again, depending on the sign of $a_1 + a_2$, we should get either a small underestimate (if $a_1 + a_2 < 0$) of the magnitude of the difference between black and white voting support for the black candidate or a small overestimate (if $a_1 + a_2 > 0$) of that magnitude. Of course, as noted earlier, if a_1 and a_2 are of opposite signs, then the magnitude of the error will be even closer to zero.

As a simple example, let us assume we have $a_1 = 0.2$, $b_1 = 0.2$, $a_2 = 0.3$, $b_2 = 0.5$, $k = 0.1$. Note this means that whites' support for blacks runs from 0.2 to 0.4, and that of blacks for blacks runs from 0.5 to 0.8 depending on context. In this case the estimated difference would be (by *Equation* 17)

$$P'_{BB} - P'_{WB} = a_2 + b_2 - b_1 = 0.6$$

while the actual average difference, by (18), would be

$$P'_{BB} - P'_{WB} = a_2 + b_2 - b_1 - k(a_1 + a_2)\,(1 - 2k/3) = 0.5533.$$

The error, 0.0467, is not quite negligible but certainly small.

If, instead of this, we have $a_1 = -0.2$, $b_1 = 0.2$, $a_2 = 0.3$, $b_2 = 0.5$, $k = 0.1$, we will have the same estimate for the difference, but the actual value will be 0.5907. In this case (because a_1 and a_2) have different signs) the error is close to negligible.

Explained variance

In terms of variance explained the linear model will also usually be a very good fit under the assumptions stated above. For example, for our second assumption as to the

distribution of x, the one in which x is concentrated near its endpoints (see *Equation* 12), the total variance of x^2 is approximately $(3 - 6k + 5K^2)/12$. The amount of variance given by the linear approximation is $(3 - 6k + 4k^2)/12$. Thus, for small k, the linear approximation accounts for a fraction approximately equal to $1 - k^2/3$ of the total variance. For example, for $k \leq 0.1$, the linear approximation accounts for well over 99% of variance.

More specifically, for a quadratic relationship given by

$$y = Cx^2 + Bx + A$$

as noted in the text, the least-squares approximation is given by *Equation* (13) as

$$y = (B + C)x + A - Cq.$$

In this instance, the quadratic form has a total variance given by

$$s^2 (A + Bx + Cx^2)$$
$$= (B + C)^2 (1/4 - k/2 - k^2/3) + C^2k^2/12.$$

Here the linear approximation accounts for

$$s^2 ((B + C)x)$$
$$= (B + C)^2 (1/4 - k/2 - k^2/3)$$

of the variance. Thus, the linear approximation accounts for all but the final term, $C^2k^2/12$, of the variance. Since total variance is of the order of magnitude $(B + C)^2/4$, we find that the fraction of the variance not accounted for by the linear approximation is roughly

$$C^2k^2/3(B + C)^2.$$

Assuming, as in the text, that $k \leq 0.1$, we find that the unaccounted for variance can be greater than 5% only when

$$C^2 > 15(B + C)^2, \text{ or } |C| > 3.9\ |B + C|.$$

This in turn can only happen if

(i) B and C are of opposite sign, and
(ii) $0.8 \leq |C/B| \leq 1.3$.

For the contextual model specified by *Equation* (4), we have $C = a_2 - a_1$ and $B = a_1 + b_1 - 2a_2 - b_2$. Hence

$$\frac{C^2k^2}{3(B + C)^2} = \frac{k^2(a_2 - a_1)^2}{3(b_1 - a_2 - b_2)^2}$$

and for this to be greater than 5% (still assuming $k \leq 0.1$), we must have

$$(a_2 - a_1)^2 \geq 15(b_1 - a_2 - b_2)^2$$

or equivalently

$$|a_2 - a_1| \geq 3.9|a_2 + b_2 - b_1|.$$

Note that the right-hand side of this equation is close to $P'_{BB} - P'_{WB}$. Since, in general, the two parameters a_1 and a_2 measure the contextual voting effect in whites and blacks respectively, we find that, even if they are of opposite sign, they would have to be substantially larger than the difference $P'_{BB} - P'_{WB}$ for this last inequality to hold, i.e., *for the unaccounted variance to be greater than 5%, the contextual effect must be considerably greater than the average racial effect.*

To look at numerical results, let us suppose once again that we have $a_1 = -0.2$, $b_1 = 0.2$, $a_2 = 0.3$, $b_2 = 0.5$, and $k = 0.1$, then total variance would be 0.0708. In this case the unaccounted error would be somewhat larger, but still only about 1/3 of the 1 percent of total variance.

Discussion

In estimating the degree to which a linear model approximates an underlying pattern of within group differences that gives rise to contextual effects with a quadratic representation, the issue is not whether or not the linear model fits the data points almost as well as the 'true' quadratic in terms of variance explained, but whether or not using a linear model can be expected to give reasonable parameter estimates for the parameters we care most about, namely P'_{BB} and P'_{WB} (or, equivalently, P'_{WW} and P'_{BW}). Moreover, since, in voting rights litigation, the bottom line is whether or not there is a real difference between black and white (minority and non-minority) levels of support for the black (minority) candidate, even if the linear model is slightly off in its estimates of P'_{WW} and P'_{BW}, taken singly, what matters most is whether or not the linear model accurately captures the *difference* between P'_{WW} and P'_{BW} (or, equivalently, between P'_{BB} and P'_{WB}).

For data on voting patterns in small geographic units such as voting precincts, we have shown that the presence of a plausible type of race-related contextual effect will, in general, lead to a quadratic relationship between minority share of the electorate and the share of the vote received by the minority candidate. But, when either a substantial proportion of the electorate is located in racially homogeneous or near homogeneous precincts (as often occurs given the patterns of residential concentration in US cities) or when the contextual effects are small, we have shown that the estimates of a key parameter of interest in voting rights litigation, namely the difference in proportions of minority and non-minority voters who voted for the minority candidate(s), will be very well approximated by a linear model. In such circumstances, fallacies of ecological inference can be avoided with near certainty, despite the fact that the linear model that has been used to fit the data is underspecified and regression diagnostics would show heteroskedasticity.

Care, however, must be taken not to read more into the generalizability of the above results than is warranted. In particular, all the examples discussed above involve cross-sectional inference within a single election and not cross-temporal analysis of voting patterns across different elections. It is the latter type of analysis that is the most common use of the Goodman (1954) technique. We believe that cross-sectional applications to election data in recent voting rights cases of the variant of Goodman's technique described above do not give rise to quite as many problems as the usual cross-temporal

voting applications of the Goodman technique because: (a) we do not have to worry about the extent to which the pool of potential voters has changed over time; (b) the restriction to a single election makes it more plausible that similar factors are affecting voters of each group; (c) the cross-sectional applications of the Goodman model in the voting rights literature are almost invariably to narrowly delimited geographic areas, again making it more plausible to posit that similar factors are affecting voters of each group; (d) for voting rights cases, data are available on very small units of aggregation in which we normally will have sufficient range in both the dependent and independent variables to develop reliable estimates of the underlying relationships; and (e) for racial groups high levels of political cohesion in cross-racial contests are reasonable to expect given present social realities in the US.[7]

Even if we are using it for cross-sectional analysis of a single election, we do not wish to claim more for bivariate linear ecological regression than is its due. While we have shown that linear ecological regression can be expected to be well-behaved in situations where voting is polarized even when there is a context effect as long as that context effect is not great or as long as we have a reasonable portion of the electorate located in racially homogeneous precincts, if we have neither racially homogeneous or near homogeneous precincts nor a considerable range in both the dependent and independent variables, the presence of substantial contextual effects can give rise to misleading estimates. For example, if we were to look at a US presidential election and use states as our units, the range of variation in the independent variable (percent black) will be limited, and the range in the dependent variable (percent of the two-party vote) will also be limited.[8]

Consider the 1988 presidential election. In 1988, we know from exit poll data that the Dukakis share of the black vote varied little from state to state, but that his share of the white vote varied considerably and was strongly linked to how black was the state (e.g., in a heavily black state like Mississippi, Dukakis received less than 20 percent of the white vote, while in heavily white states his vote averaged above 40 percent).[9] When we use aggregate data to run a linear regression of Democratic share of the two-party vote on percent black in the state, the regression actually shows the Dukakis share of the two-party vote *declining* with black percentage. But, of course, here we are estimating a relationship only over that portion of the curve where the relationship is negative—there simply are no states that are sufficiently black to clearly reveal the actual overwhelming level of black support for Dukakis in the 1988 presidential contest.

While the above example reminds us that caution must be used in interpreting ecological regression results, since voting rights lawsuits are unlikely to be brought except in jurisdictions where the existence of a geographically concentrated minority population makes possible the creation of a district-based remedial plan that concentrates minority population in a small number of districts, when we are looking at data for such lawsuits, involving local-level contests with black candidates and using precincts (not whole states) as our units of analysis, we normally will have sufficient range in both the dependent and independent variables to estimate the underlying relationships. Also, if there are problems with the estimates produced by the linear model, it is very likely that these will be detected by various of the cross-checks on its validity that are standard in the literature.[10]

While we take the modelling above primarily to be general support for the likely validity of estimates about average group behavior derived from the use of linear bivariate ecological regression techniques in the very specific context of racial bloc voting analyses in single elections, the results are also suggestive of types of conditions under which relatively simple ecological regression methods can be reliably used to measure

underlying parameters of voting (and other) behavior and ecological fallacies avoided. What is striking about our results is that they show conditions under which a model known to be misspecified can nonetheless be highly accurate in estimating key parameters of interest.[11]

Acknowledgements

This research was supported by Ford Foundation Grant # 446740-47007 and also draws on previous research that was supported by NSF Grant SES # 88-09392. Neither the National Science Foundation nor the Ford Foundation is in any way responsible for the opinions expressed in this essay. We are indebted to Karen Sadler and the Word Processing Center, School of Social Sciences, UCI for manuscript typing and to Dorothy Green for bibliographic assistance.

Notes

1. In this paper, to avoid notational complexity, we will deal exclusively with single equation ecological regression and we will avoid situations where we need to estimate the behavior of more than two racial groups. However, the results we give can be readily modified to be applicable to the double-equation approach and can, we think, be extended to deal with polychotomous classifications.
2. We omit discussion of how to calculate a confidence range around each of these parameter estimates (see Grofman and Migalski, 1988).
3. Of course, as shown below, the magnitude of that bias may be insignificant.
4. Advocates of the first approach are Lupia and McCue, 1990; Firebaugh, 1993. In the racial bloc voting context, because of the problem of multicollinearity of income with race, and the limited number of data points in most jurisdictions, I am skeptical that this approach will be feasible. Also, since there are numerous control variables (e.g., education, home ownership, recency of in-migration, party registration, as well as income) that might be relevant, choosing among competing multivariate models can become a practical nightmare, a consideration especially relevant in the litigation setting. In future research, one of us, Grofman, intends to work on simulations that should provide evidence that directly bears on the probable utility of various alternative models. Our expectation is that simple models will often prove to be superior to models that, on their face, would appear to be better in that they incorporate more variables known to be linked to the behavior to be understood. (For discussion of a related instance of statistical modelling where more (variables) was not better, see Grofman, 1991b.)
5. See Miller 1977, and discussion of this model in Grofman (1987).
6. The so-called 'neighborhood model' of Freedman *et al.* (1991) can be thought of as a special case of this model, one where $a_2 = a_1$. Not only is this a restrictive condition in general, it can never happen when a_2 and a_1 are of opposite sign. These two parameters will be of opposite sign if there is a context effect in which the willingness of blacks to vote for black candidates increases among blacks who live in blacker precincts but white solidarity behind white candidates also increases the blacker is the precinct population (see below). For further discussion of the Freedman *et al.* (1991) model see Grofman (1991a), Lichtman (1991); and Loewen *et al.* (1993).
7. For further discussion of this point see Grofman (1995).
8. There are also potential complictions caused by differential turnout or roll-on (voter for office) levels among white and black eligible voters that we do not address here. As noted earlier, these can be dealt with using a double equation rather than a single equation approach (see Grofman and Migalski, 1988; Grofman *et al.*, 1992, Ch. 4).
9. See Glazer *et al.*, 1993.
10. See Loewen and Grofman, 1989; Grofman *et al.*, 1992, Ch. 4.

11. Moreover, even when we can posit that the 'true' relationship is quadratic and not linear, fitting a quadratic rather than a linear model to data on voting patterns by racial composition of precincts does not improve our ability to estimate mean differences in white and black voting patterns because the set of equations to derive the (four) needed parameters from the (three) fitted coefficients of the quadratic regression is underdetermined.

References

ALT, J. (1994) The impact of the Voting Rights Act on black and white registration in the South. In *Quiet Revolution: The Impact of the Voting Rights Act 1965–1990*, ed. C. Davidson and B. Grofman. Princeton University Press, Princeton, NJ.

BLACK, E. AND BLACK, M. (1992) *The Vital South*. Harvard University Press, Cambridge.

BROWN, C. (1982) The Nazi vote: a national ecological study. *American Political Science Review* 76(2), 285–302.

DUNCAN, D. AND DAVIS, B. (1953) An alternative to ecological correlation. *American Sociological Review* **18**, 665–666.

ENGSTROM, R. AND MCDONALD, M. (1988) Definitions, measurements, and statistics: weeding Wildgen's thicket. *Urban Lawyer* 20(1) (Winter), 175–191.

FALTER, J. AND ZINTL, R. (1988) The economic crisis of the 1930's and the Nazi vote. *Journal of Interdisciplinary History* 19(1) (Summer), 55–85.

FIREBAUGH, G. (1993) Are bad estimates good enough for the courts? *Social Science Quarterly* 74(3) (September), 488–496.

FREEDMAN, D. A., KLEIN, S. P., SACKS, J., SMYTH, C. T. AND EVERETT, C. G. (1991) Ecological regression and voting rights. *Evaluation Review* 15(6) (December), 673–713.

GLAZER, A., GROFMAN, B. AND OWEN, G. (1993) A formal model of group oriented voting. Paper presented at the *Annual Meeting of the Public Choice Society*, March 19–21, 1993, New Orleans, LA.

GOODMAN, L. (1953) Ecological regression and the behavior of individuals. *American Sociological Review* **18** (December), 663–664.

GOODMAN, L. (1959) Some alternatives to ecological correlation. *American Journal of Sociology* **64**, 61–625.

GROFMAN, B. (1987) Models of voting. In *Micropolitics Annual*, ed. S. Long, JAI Press, Greenwich CT, pp. 31–61.

GROFMAN, B. (1991a) Statistics without substance: a critique of Freedman *et al.* and Clark and Morrison. *Evaluation Review* 15(6) (December), 746–769.

GROFMAN, B. (1991b) Rejoinder: straw men and stray bullets, a reply to bullock. *Social Science Quarterly* 72(4) (December), 840–843.

GROFMAN, B. (1992) Expert witness testimony and the evolution of voting rights case law. In *Controversies in Minority Voting: The Voting Rights Act in Perspective*, ed. B. Grofman and C. Davidson, pp. 197–229. The Brookings Institution, Washington, DC.

GROFMAN, B. (1993a) Throwing darts at double regression: a rejoinder to Wiuldgren. *Social Science Quarterly*

GROFMAN, B. (1993b) The use of ecological regression to estimate racial bloc voting. *University of San Francisco Law Review*, 27(3), 593–625.

GROFMAN, B. (1995) New methods for valid ecological inference. In *Spatial and Contextual Models in Political Research*, ed. Munrow Eagles. Taylor and Francis, London.

GROFMAN, B., HANDLEY, L. AND NIEMI, R. G. (1992) *Minority Representation and the Quest for Voting Equality*. Cambridge University Press, New York.

GROFMAN, B. AND MIGALSKI, M. (1988) Estimating the extent of racially polarized voting in multicandidate elections. *Sociological Methods and Research* 16(4), 427–454.

GROFMAN, B., MIGALSKI, M., NOVIELLO, N. (1985) The Totality of Circumstances' Test in Section 2 of the 1982 Extensions of the Voting Rights Act: a social science perspective. *Law and Policy* 7(2) (April), 209–223.

KEY, V. O. (1949) *Southern Politics*. Vintage, New York.

LICHTMAN, A. J. (1991) Passing the test: ecological regression analysis in the Los Angeles county case and beyond. *Evaluation Review* 15(6) (December), 770–799.

LOEWEN, J. (1982) *Social Science in the Courtroom*. Lexington Books, Lexington, MA.

LOEWEN, J., BURTON, O. V., BRISCHETTO, R. R. AND FINNEGAN, T. (1993) It ain't broke, so don't fix it: the legal and factual importance of recent attacks on methods used in vote dilution litigation. *University of San Francisco Law Review* 27 (Summer), 737–780.

LOEWEN, J. AND GROFMAN, B. (1989) Comment: recent developments in methods used in voting rights litigation. *Urban Lawyer* **21**(3), 589–604.

LUPIA, A. AND MCCUE, K. (1990) Why the 1980s measures of racially polarized voting are inadequate for the 1990s. *Law and Policy* **12**(4) (October, 355.

MILLER, W. L. (1977) *Electoral Dynamics in Britain since 1918*. MacMillan Press Ltd, London.

ROBINSON, W. S. (1950) Ecological correlations and the behavior of individuals. *American Political Science Review* 5, 351–357.

WILDGREN, J. K. (1993) Social alchemy in the courtroom: the 'double regression hoax'. *Social Science Quarterly*.