B61

# Throwing Darts at Double Regression—and Missing the Target*

Bernard GROFMAN, *University of California, Irvine*

To improve estimates of the extent of racial bloc voting in U.S. elections for which no reliable survey data on voter choice are available, social scientists have developed a "double regression" approach based on use of aggregate data at the precinct level. This methodology, intended to compensate for differential levels of turnout among minority and nonminority eligible voters, has been used in testimony involving minority voting rights issues in scores of federal trials where evidence about the nature of voting patterns in election contests involving both minority and nonminority candidates is legally required. Wildgen makes four claims about this ecological regression methodology: (1) that this regression-based technique should not be called "regression," (2) that double regression should not be called "bivariate," (3) that it is invariably upwardly biased in its estimates of bloc voting, and (4) that it is incorrect to claim that this methodology "cures the ecological fallacy." The first criticism is absurd on its face, and the attempts to give it a "sensical" meaning lead Wildgen into other errors. The second assertion is a trivial definitional quibble. The third complaint is dead wrong, and the fourth is directed against a straw man largely of Wildgen's own imagination.

**For over a decade Wildgen has testified on behalf of defendant jurisdictions** whose electoral practices have been challenged in court as discriminatory against racial or linguistic minorities. In such voting rights cases the litigants must provide to the court their estimates of what proportion of the votes of black voters and of white voters are received by the various black and white candidates in contests where there are black candidates competing. These estimates are then used, along with other information, to answer the question of whether or not voting in the challenged jurisdiction is racially polarized to a degree that is legally significant.[1] Wildgen (1990),

*Direct correspondence to Bernard Grofman, School of Social Sciences, University of California, Irvine, CA 92717. Research was partially funded by Ford Foundation grant no. 446740-47007. I am indebted to the Word Processing Center, School of Social Sciences, UCI, for manuscript typing and to Dorothy Gormick for bibliographic assistance. The Ford Foundation is not in any way responsible for the opinions expressed in this essay.

[1] Elections in which there is at least one black and at least one white candidate are the elections which courts have held to be most probative of polarization (see Grofman, Handley, and Niemi, 1992; Grofman and Handley, 1992). We use black as a shorthand for "minority."

*SOCIAL SCIENCE QUARTERLY*, Volume 74, Number 3, September 1993

What is really unlikely is that all of these errors could have occurred by chance rather than through conscious arithmomancy. Does this compounding of errors protect us from the ecological fallacy? No more so, I suspect, than did the high correlations Robinson dealt with in documenting the ecological fallacy. Yet there is no easy way around double regression. Advocates of double regression spurn polls, their impossible results prove what they set out to prove, and other aggregate methods are not "standard." In other words, double regression is incapable of disproof. SSQ

## REFERENCES

Bullock, Charles. 1991. "Misinformation and Misperceptions: A Little Knowledge Can Be Dangerous." *Social Science Quarterly* 72:834–39.

Engstrom, Richard L. 1990. "Getting the Numbers Right: A Rely to Wildgen." *Urban Lawyer* 22:495–502.

Engstrom, Richard L., and Michael D. McDonald. 1988. "Definitions, Measurements, and Statistics: Weeding Wildgen's Thicket." *Urban Lawyer* 20:175–91.

Flanigan, William, and Nancy Zingale. 1985. "Alchemist's Gold: Inferring Individual Relationships from Aggregate Data." *Social Science History* 9:71–91.

Goodman, Leo. 1953. "Ecological Regression and the Behavior of Individuals." *American Sociological Review* 19:663–64.

Grofman, Bernard. 1991a. "Multivariate Methods and the Analysis of Racially Polarized Voting: Pitfalls in the Use of Social Science by the Courts." *Social Science Quarterly* 72:826–33.

———. 1991b. "Straw Men and Stray Bullets: A Reply to Bullock." *Social Science Quarterly* 72:840–43.

Grofman, Bernard, Michael Migalski, and Nicholas Noviello. 1985. "The 'Totality of Circumstances Test' in Section 2 of the 1982 Extension of the Voting Rights Act: A Social Science Perspective." *Law & Policy* 7:199–223.

Gujarati, Damodar. 1978. *Basic Econometrics*. New York: McGraw-Hill.

Loewen, James W. 1990. "Sand in the Bearings: Mistaken Criticisms of Ecological Regression." *Urban Lawyer* 22:503–13.

Loewen, James W., and Bernard Grofman. 1989. "Recent Developments in Methods Used in Vote Dilution Litigation." *Urban Lawyer* 21:589–604.

Robinson, W. S. 1950. "Ecological Correlation and the Behavior of Individuals." *American Sociological Review* 15:351–57.

Wildavsky, Aaron. 1959. "A Methodological Critique of Duverger's Political Parties." *Journal of Politics* 21:303–18.

Wildgen, John K. 1990. "Vote Dilution Litigation and Cold Fusion Technology." *Urban Lawyer* 22:487–94.

repeating views he has expressed in his recent courtroom testimony—views that have not been persuasive to courts—has taken particular aim at a technique for drawing inferences about racial patterns of voting known as "double regression." In the previous comment, Wildgen reiterates this attack, in essence characterizing the double regression method as a statistical hobgoblin, one which he appears to believe was diabolically conjured up by Grofman in order to mislead federal courts by producing inaccurate overestimates of racial polarization which he views as "incapable of disproof." Because of *very severe* space limitations and because I have elsewhere dealt with the mistaken assertion that valid ecological inference about election outcomes is simply, in principle, impossible,[2] in this brief comment I will focus exclusively on Wildgen's scattershot attack on double regression.

I should first note that, while I would not mind taking credit for the sole invention of the double regression technique, which has become standard in voting rights litigation since 1986 as an improvement on the previous single-equation method, in fact double regression is a technique that was independently arrived at by a number of scholars from several different social science disciplines, including history (Kleppner, 1985; Kousser, 1973, 1974; Lichtman, 1991), as well as political science (myself) and sociology (Loewen, 1982). However, my use of the technique in *Thornburg v. Gingles,*[3] the leading voting rights case of the past decade, and subsequent explication (Grofman, Migalski, and Noviello, 1985; Grofman and Migalski, 1988; Loewen and Grofman, 1989) has become the best known introduction to the method, which has been adapted for use by numerous experts in scores of subsequent voting rights analyses.

Double regression is a straightforward and sensible solution to a problem peculiar to aggregate-level electoral research, namely the need to provide a statistical corrective for divergence between a group's share of the eligible electorate and its share of the actual electorate in situations where only the former is known with surety. In general we wish to know what proportion of the black (and the white) electorate supported the black candidate(s). Absent survey data, the only information we have that directly bears on this question is derived from analysis of the relationship between percent black

[2]See Grofman and Migalski (1988); Grofman, Handley, and Niemi (1992: especially chap. 4); Grofman (1991); and Loewen and Grofman (1989); cf. Grofman (1987, 1992). In my view, the purist position that nothing useful can be said about voting patterns if there are no survey data is absurd. Consider the voting patterns in racially homogeneous areas, for example. Moreover, when used with the proper checks, ecological regression has shown its reliability and has withstood severe attacks, not just from Wildgen, but also from a number of other social scientists and statisticians. In *Garza v. Los Angeles County Board of Supervisors* (D. Cal. 1990), 90 C.D.O.S. 8138 (9th Cir. 199) cert denied, January 1990, for example, in the two elections involving a Hispanic candidate for which exit poll data were available, the citywide ecological regression estimates (based on double regression) and the citywide exit poll data were essentially indistinguishable. Moreover, if barred from reporting actual election returns and inferences based on them, minority plaintiffs would effectively be barred from ever protecting their rights in court.

[3] *Gingles v. Edmisten,* 590 F. Supp. 345 (1984), heard sub nom. *Thornburg v. Gingles,* 478 U.S. 30, 106 S. Ct. 2752 (1986).

among the electorate and the vote share from among that same electorate received by the black candidate. The problem is that, in most voting rights cases, we do not know the black percentage of the actual electorate, only their percentage in the eligible electorate (e.g., blacks of voting age).

To take an extreme situation, let us imagine that all of the eligible whites vote but only 50 percent of the eligible blacks do so. In this situation, the relationship between black share of the electorate and black share of the eligible voters will be nonlinear. Let $x$ be the fraction black among eligible voters in a given precinct. In any given precinct, black share of turnout = $[.5(1 - x)]/[.5(1 - x) + x] = (.5 - .5x)/(.5 + .5x)$. This is a nonlinear function of $x$, a repeating fraction which may be approximated by a polynomial function of order $n$. For example, if black share of eligibles $(1 - x)$ is .5, then black share of turnout is only one-third, but if black share of eligibles is .1, then black share of turnout is only .052, while if black share of eligibles is .9 then black share of turnout is .818. The above formula readily generalizes. If it is the case that $k$ fraction of the eligible whites vote and $j$ fraction of the eligible blacks vote, then black share of turnout = $[j(1 - x)]/[j(1 - x) + kx] = (j - jx)/[j + (k - j)x]$. Again we have a nonlinear function in $x$. Because of this nonlinearity, which becomes a more serious problem the greater the difference between the turnout levels of white and black eligibles, any regression which uses black share of eligibles as a proxy for black share of turnout is potentially suspect.

Consider what happens in an extreme hypothetical where 90 percent of blacks vote for the black candidate and 90 percent of whites vote for the white candidate, and black turnout is, as before, 50 percent of black eligibles while white turnout is 100 percent of white eligibles. Then, if we regress the black candidate's share of the vote on the black share of the eligibles, we obtain $y = .03 + .77x$ as our linear fit if we posit a uniform distribution. This means that we would estimate black vote for the black candidate as roughly 80 percent (.77 + .03, too low) and white vote for the black candidate as 3 percent (also too low). Looking at the graphical representation (omitted), we see that there is considerable heteroscedasticity—as we might expect when fitting a nonlinear function with a linear model. Note that, in this hypothetical, where turnout differences between black and white eligibles are severe, and black turnout is lower than white turnout, the single-equation method causes us to underestimate black vote for the black candidate by 10 percentage points.

When we use the double-equation method, we estimate two equations. The statistical trick used by double regression to cope with the nonlinearity resulting from differential turnout by race rests on the fact we can run linear regressions in which the black candidate's share of eligible (and not actual) voters is used as the dependent variable and regressed on an independent variable which has the same denominator as the dependent variable does (see details in Grofman, Migalski, and Noviello [1985]). Under the above assumptions we obtain $y_1 = .35x + .10$ and $y_2 = -.85x + .90$. This

gives rise to an estimate of the proportion of whites who vote for the black candidate as 10 percent (.10/(.10 + .90)), and an estimate of the proportion of blacks who vote for the black candidate as 90 percent ((.35 + .10)/(.35 + .10 − .85 + .90)). Thus, in this hypothetical, (a) the double-equation method will, unlike the single-equation method, *get the proportions of each race voting for the black candidate exactly right,* and (b) since white turnout is higher than black turnout, the double-equation method will produce estimates of black support for the black candidate that are *higher* than those yielded by the single-equation method. *But here higher estimates are more accurate estimates.* In contrast, if white turnout is lower than black turnout, then the single-equation method will usually yield estimates of black support for the black candidate that are too high, while again the double-equation method will make the appropriate corrections for differential turnout levels.

In like manner, if whites turn out at 50 percent of eligibles and blacks turn out at 100 percent of eligibles (the mirror image of our previous example), then (again assuming a uniform distribution on the proportion of the eligible population that is black) and again assuming that voting is, in fact, near perfectly predicted by racial lines (90 percent of each race voting for the same-race candidate), then the single-equation method would give us black vote for the black candidate as 97 percent and white vote for the black candidate as 20 percent because the regression equation is $y = .20 + .77x$ (graphical representation omitted). These figures are both too high. (Recall that in the previous example the single-equation estimates were too low.) Once again the double-equation method will yield the correct estimate (90 percent) of black support for the black candidate, but now, with black turnout lower than white turnout, the single equation overestimates black support for the black candidate and so the double-equation estimate of black bloc voting will be *lower* than the single-equation estimate. Thus, *sometimes* the double-equation method yields estimates higher than the single-equation method (when the single-equation answer is too low) and *sometimes* the double-equation method yields results higher than the single-equation method (when the single equation answer is too high), but, in general, *the double-equation method yields results closer to the truth than does the single-equation method!*

Wildgen (1993) asserts that the virtual absence of double regression outside of the voting rights context shows that it lacks validity. Rather, its use is largely confined to the courtroom because (a) it can only be used to solve a particular and very special type of problem (see below) and (b) the elections that are of most relevance in voting rights cases are usually local elections for which survey data are not available, while the elections of most interest to political scientists are national (or statewide) elections for which survey data are available. Wildgen (1993) also makes a large fuss about the fact that sometimes inferences based on ecological regression lead to parameter estimates outside the (0, 1) range. First of all, this does not happen often.

Second, point estimates are just that, estimates. Third, to take advantage of the fact that we know vote shares must range between 0 and 1 to constrain estimates to lie within that range is not "cheating," it is just common sense. Fourth, it is easy to show that slight randomness in the underlying relationships can give rise to data sets where (even with double regression) the fitted regressions give us estimates slightly outside the feasible range. Nonetheless, if the data show, say, 108 percent of the blacks voting for the black candidate, almost without exception we can be confident that this tells us that black support for the black candidate is very very high indeed. (For a discussion of the few peculiar counterexamples, involving situations with more than one minority group and no homogeneous precincts, see Grofman, Handley, and Niemi [1992: chap. 4].)

Wildgen's (1993) most serious charge is that double regression "profits its users through ratcheting up slopes" (p. 477). *As can be seen from the above examples, this assertion is flatly wrong.* Double regression will usually produce higher estimates of bloc voting by blacks than will single-equation regression *if* black turnout is lower than white and lower estimates *if* black turnout is higher than white—but only because, in both instances, the double-equation results more closely mirror reality. Moreover, when Wildgen says that double regression "profits its users through ratcheting up slopes," he implies, wrongly, that (a) the only users of ecological regression are experts in court who are testifying for *plaintiffs* and (b) experts testifying for plaintiffs (including myself, Richard Engstrom, and others) have so little scholarly integrity that they would use methods which they knew produced results favorable to their side (i.e., higher estimates of bloc voting) and then seek to hide that fact from the federal courts while testifying under oath. I have testified for defendant jurisdictions in voting rights cases, including the city of Boston, and the states of Indiana and Rhode Island as well as for plaintiffs. I use exactly the same methodology no matter the side for which I testify. The methods I advocate have been used by other experts for defendants as well, e.g., Kimball Brace, Lisa Handley, Harold Stanley, and Ronald Weber.

The discussion of double regression above, although necessarily an abbreviated one, also allows us to show that another one of Wildgen's central points rest on a trivial definitional quibble. Wildgen says that double regression is not bivariate. But each of the two equations identified above *is* bivariate. Of course, as shown in Grofman, Migalski, and Noviello (1985), the parameters (slope and intercept) in each of the two equations are combined so as to estimate two additional sets of values. (See also the calculations for double regression shown above.)[4]

[4]Wildgen has simply not understood the distinction made in Justice Brennan's opinion in *Thornburg* when Brennan distinguished between approaches that seek to describe the (estimated) actual average voting behavior by race, e.g., what proportion of black voters vote for the black candidate versus what proportion of the white voters vote for the black candidate,

Wildgen also makes the assertion that "double regression is not regression." If by this is meant the claim that the use of two regressions somehow isn't actually regression, that's just bizarre. As far as I can judge, Wildgen is using this cutesy phrasing to express a claim that double regression, *by definition,* violates the standard assumptions of regression analysis. Here there are numerous errors. (a) Wildgen miscalculates various of the values he cites as evidence for double regression failures (see rebuttal in Engstrom [1990]). (b) Wildgen (1993:476) incorrectly asserts that there is no statistical model for the sampling errors in double regression. In fact, that is one of the central points developed in Grofman and Migalski (1988), an article in a mainstream methods journal of which Wildgen is apparently unaware. (c) Wildgen looks at the wrong error terms. He does not realize that the relevant errors to look at are those involved in comparing the *predicted* and the *actual* candidate vote shares at the precinct level—a comparison which can readily be generated from the information provided by the double-equation regressions. When he makes a fuss about the fact that errors in each of the double equations are correlated with the independent variable, he fails to understand that this, rather than being a fault, is actually a justification for the desirability of going to the trouble of making use of the double-equation method. Such correlated errors occur in part because voter *turnout* is correlated with race, exactly as anticipated by the model.

Wildgen's other criticisms of double regression attack a straw man largely of his own creation. He misstates (or overstates) the claims made for the method and then knocks down these claims—in the process displaying how limited is his own understanding of statistics. For example, Wildgen (1993) asserts that the double regression method purports to provide a set of "statistical controls for the effects of *differential participation, socioeconomic conditions,* and *precinct size*" (p. 476; emphasis added). But that claim is just flat wrong. Double regression is intended to control only for the *first* of these, i.e., "racial differences in levels of voter participation" (Engstrom, 1990; quoted in Wildgen, 1993:475–76). How does Wildgen make such a basic error? Answer: by misreading the works he cites and citing out of context.

His claim that double regression is intended to deal with the problems of precinct size is based on his confusion of Loewen's defense of *weighting cases* (i.e., voting precincts) by precinct population, with Loewen's defense of the *double* regression methodology. Yet one could choose to weight cases

---

and approaches that seek to "explain" why such proportions are as they are by recourse to factors such as party identification, socioeconomic differences, endorsements by local and community newspapers, differences in campaign issues of the candidates, and so on. The legal standard for bloc voting is that all that matters is the differences, not the reason(s) for them. In terms of this distinction, the double regression is a descriptive method. In contrast, most multivariate methods aim at explanation (e.g., Bullock [1984]; but not all—see discussion in Grofman [1985, 1992], Grofman, Handley, and Niemi [1992: chap. 4]).

even if one were doing a *single* regression of, say, percent black in the precinct on percent vote for the black candidate. Weighting cases is weighting cases, i.e., simply a device to reflect the underlying population distribution given that precincts may vary considerably in numbers of eligible voters. My regression analyses and those of Alan Lichtman, another expert for the U.S. Department of Justice, were replicated by one of the experts for the defendant County of Los Angeles in *Garza,* University of Illinois statistician Jerome Saks. Lichtman had weighted cases; I had not. Saks replicated our analyses using both weighted and unweighted cases. The differences between weighted and unweighted regressions were so trivial that Saks ceased to worry about which was which.

When he provides a brief quote from Engstrom (1990:497) to support his claim that double regression is intended to control for differences in socioeconomic conditions, the error comes from omitting material from Engstrom (1990), including material interior to that quoted, that would make it clear that the quoted phrase "in addition, the socio-economic conditions of the district's black population, conditions that typically relate to political participation, are shockingly low" simply provides a reason why one would expect that the double regression correction for differential levels of black and white turnout (relative to eligibles) would need to be employed.

In sum, Wildgen's wild-eyed attacks on double regression not only never hit the bullseye; they mostly miss the dartboard entirely. On the rare occasions where Wildgen's criticisms are legitimate, as where he criticizes judicial misunderstandings of the correlation coefficient, what he has to say has been said by others, myself included (see, e.g., Grofman, 1985, 1992). SSQ

## REFERENCES

Bullock, Charles S., III. 1984. "Symbolics or Substance: A Critique of the At-Large Election Controversy." *State and Local Government Review* 21:91–99.

Engstrom, Richard L. 1990. "Getting the Numbers Right: A Response to Wildgen." *Urban Lawyer* 22:495–502.

Grofman, Bernard. 1985. "Criteria for Districting: A Social Science Perspective." *UCLA Law Review* 33:77–184.

———. 1987. "Models of Voting." Pp. 31–61 in Samuel Long, ed., *Micropolitics Annual.* Greenwich, Conn.: JAI Press.

———. 1991. "Multivariate Methods and the Analysis of the Racially Polarized Voting: Pitfalls in the Use of Social Science by the Courts." *Social Science Quarterly* 72:826–33.

———. 1992. "Expert Witness Testimony and the Evolution of Voting Rights Case Law." Pp. 197–229 in Bernard Grofman and Chandler Davidson, eds., *Controversies in Minority Voting.* Washington, D.C.: Brookings Institution.

Grofman, Bernard, and Lisa Handley. 1992. "Identifying and Remedying Racial Gerrymandering." *Journal of Law and Politics* 8:746–69.

Grofman, Bernard, Lisa Handley, and Richard G. Niemi. 1992. *Minority Representation and the Quest for Voting Equality.* New York: Cambridge University Press.

Grofman, Bernard, and Michael Migalski. 1988. "Estimating the Extent of Racially Polarized Voting in Multicandidate Elections." *Sociological Methods and Research* 16:427–54.

Grofman, Bernard, Michael Migalski, and Nicholas Noviello. 1985. "The 'Totality of Circumstances' Test in Section 2 of the 1982 Extension of the Voting Rights Act: A Social Science Perspective." *Law & Policy* 7:209–23.

Kleppner, Paul. 1985. *Chicago Divided: The Making of a Black Mayor.* De Kalb: Northern Illinois University Press.

Kousser, J. Morgan. 1973. "Ecological Regression and the Analysis of Past Politics." *Journal of Interdisciplinary History* 4 (Autumn):237–62.

———. 1974. *The Shaping of Southern Politics.* New Haven: Yale University Press.

Lichtman, Alan. 1991. "Passing the Test: Ecological Regression Analysis in the Los Angeles County Case and Beyond." *Evaluation Review* 15:770–99.

Loewen, James. 1982. *Social Science in the Courtroom.* Lexington, Mass.: Lexington Books.

Loewen, James, and Bernard Grofman. 1989. "Recent Developments in Methods Used in Vote Dilution Litigation." *Urban Lawyer* 21:589–604.

Robinson, W. S. 1950. "Ecological Correlations and the Behavior of Individuals." *American Sociological Review* 5:351–57.

Wildgen, John K. 1990. "Vote Dilution Litigation and Cold Fusion Technology." *Urban Lawyer* 22:487–94.

———. 1993. "Social Alchemy in the Courtroom: The 'Double Regression' Hoax." *Social Science Quarterly* 74:471–79.