ECOLOGICAL REGRESSION *VERSUS* THE SECRET BALLOT

by

S P Klein, J Sacks, and D A Freedman

Technical Report No 276
February 1991

Department of Statistics
University of California
Berkeley, California

February 1991

ECOLOGICAL REGRESSION *VERSUS* THE SECRET BALLOT

by

S P Klein
The RAND Corporation, Santa Monica

J Sacks
Department of Statistics
University of Illinois at Urbana-Champaign

D A Freedman[*]
Department of Statistics, UC Berkeley

Abstract

Ecological regression is a statistical mainstay in litigation brought under the Voting Rights Act of 1965. However, this technique depends on the extremely powerful-- and highly questionable-- *constancy assumption*. Loewen and Grofman have proposed five checks to see whether the constancy assumption is valid. We show that these five checks cannot do the job.

Indeed, the five checks can be passed with flying colors even when the constancy assumption fails and, for example, the data conform to the *neighborhood model*. This alternative model makes a diametrically opposite assumption. In brief, ecological regression assumes that race explains all the systematic differences in voting behavior, and other variables do not matter; the neighborhood model puts the emphasis on these other variables. Not surprisingly, the two models give radically different estimates of how groups vote. The five checks cannot determine which model is reliable, or which set of estimates is right.

*. Support from the Miller Institute is gratefully acknowledged.

ECOLOGICAL REGRESSION *VERSUS* THE SECRET BALLOT

The Voting Rights Act, passed in 1965 and amended in 1982, prohibits any "voting qualification or prerequiste to voting, or standard, practice, or procedure...which results in a denial or abridgement of the right of any citizen of the United States to vote on account of race or color..." Under Section 2 of the Act, members of a minority group must have the same opportunity as "other members of the electorate to participate in the political process and to elect representatives of their choice." Discriminatory intent does not have to be proved; only results matter. (96 Stat. 134, 478 U.S. 36.)

This act would be violated if districts were formed in a way that diluted the minority vote. Such dilution can be illustrated by a hypothetical example. Suppose a community has a five-member city council with each district electing its own member. Suppose too that blacks and whites are highly segregated by residence, 40% of the city's residents are black, but boundary lines are drawn so that blacks comprise less than 50% of the residents in every district. Finally, suppose blacks and whites prefer different candidates. Under these conditions, no black may be elected to the city council, and courts are likely to find a Section 2 violation.

The leading vote dilution case is Thornburg v. Gingles (478 U.S. 30, decided in 1986). The Supreme Court ruled that plaintiffs must prove three things to win such cases:

> First, the minority group must be able to demonstrate that it is sufficiently large and geographically compact to constitute a majority in a single-member district.... Second, the minority group must be able to show that it is *politically cohesive*.... Third, the minority must be able to demonstrate that the white majority votes sufficiently as a bloc to enable it...usually to defeat the minority's *preferred candidate*.... [our italics, pp50-51]

The first test is satisfied if a district could be formed so that more than 50% of its voting age citizens are members of the minority group [1]. The second and third tests are satisfied if minority group voters generally prefer a

[1] There is some controversy regarding whether citizens or voting age citizens should be the base for the first test. In Garza et al. v. the County of Los Angeles, the district court ruled that "the citizen voting age population is the appropriate measure in determining geographical compactness" (90 DAR 6161, but see 6141-2, 90 CDOS 8140). In addition, under the constitutional "one-person-one-vote" principle, all the districts must have about the same number of residents.

different candidate than the one preferred by majority group voters, i.e., voting is "racially polarized"; and the candidates preferred by majority group voters generally win.

If all the whites live in one part of the community and all the blacks in another, it is easy to determine how each group votes, in order to apply the second and third *Gingles* tests. But in many communities, there is a reasonable degree of residential integration. Now the secret ballot makes it difficult if not impossible to determine the voting preferences in each group.

Table 1 illustrates the problem for one hypothetical precinct. In this precinct, the black candidate received 60 votes and the white candidate, 90 votes. There were 50 black voters and 100 white voters. The vote for each candidate is a matter of public record; data on the racial or ethnic composition of the electorate are generally available. However, due to the secret ballot, the breakdown of the vote by group is not known. Did all 50 blacks vote for the black candidate, did they split their votes, or did they favor the white candidate? The question marks in the table represent the unknowns. Obtaining accurate estimates for these unknowns is central to voting rights cases, because of the second and third *Gingles* tests.

Table 1. Number of votes cast for each candidate, and number of voters in each racial group. Data are for a hypothetical precinct.

	Black voters	White voters	All voters
Votes for the black candidate	?	?	90
Votes for the white candidate	?	?	60
Total votes cast	50	100	150

Plaintiffs estimate the racial breakdown of the vote for each candidate, using a statistical technique called "ecological regression." And, in many cases, the court has accepted the results [2]. However, ecological regression is a controversial technique which can be relied on only under very special circumstances (Goodman, 1953, 1959). Specifically, this technique assumes the members of a group vote alike regardless of where they live. For example, according to this "constancy assumption," if the blacks in a totally black precinct give 80% of their vote to a candidate, then (apart from random variation) 80% of the blacks in heavily white precincts must also vote for this candidate, and so must 80% of the blacks in mixed precincts.

In *Gingles*, the District Court "relied principally on statistical evidence presented by...Dr. Bernard Grofman" (478 U.S. 52), based on ecological regression, to establish racially polarized voting. From our perspective, courts decide law cases, not statistical principles. Indeed, Loewen and Grofman (1989) proposed five checks to determine when ecological regression can be trusted in voting rights litigation.

In this article, we demonstrate that the checks cannot do the job. We review ecological regression, and show the five checks can be passed with flying colors even if estimates from ecological regression are grossly in error. To illustrate how this happens, we use the neighborhood model (Freedman et al., 1991). This model turns the constancy assumption on its head by positing that (apart from random variation) minority and non-minority voters in each precinct vote alike. For many jurisdictions, the neighborhood model is more plausible than ecological regression, because people who live in the same precinct generally have similar incomes and education, and share other characteristics such as party affiliation that are related to voting behavior.

[2] Some of the cases in which the court relied on ecological regression results are: United States v. Dallas County Commission, 850 F.2nd 1433 (11th Cir. 1988); McNeil v. City of Springfield, 658 F. Supp. 1015, App. C (C.D. Ill. 1987); and Gomez v. City of Watsonville, 863 F.2nd 1409 (9th Cir. 1988). The court did not believe ecological regression results in Romero v. Pomona; Badillo et al. v. City of Stockton, and ______ v. Fort Lauderdale.

Ecological regression assumes that race (or ethnic group) explains all the systematic differences in voting behavior among precincts, and other variables do not matter. The neighborhood model assumes precinct effects explain everything. Not surprisingly, the two models usually give radically different estimates of how groups vote. Ecological regression can indicate that voting is highly polarized, while the neighborhood model says there is little if any polarization-- and the five checks are passed by both models.

Consequently, the five checks cannot determine which model is more reliable or which set of estimates is right. The implication for voting rights cases is clear: the checks cannot help the courts decide whether ecological regression is trustworthy. In particular, if there is at least some degree of residential integration, polarized bloc voting cannot be inferred from precinct-level voting data. As a result, ecological regression is especially problematic when applied to hispanics in the western part of the U.S. (Freedman et al., 1991).

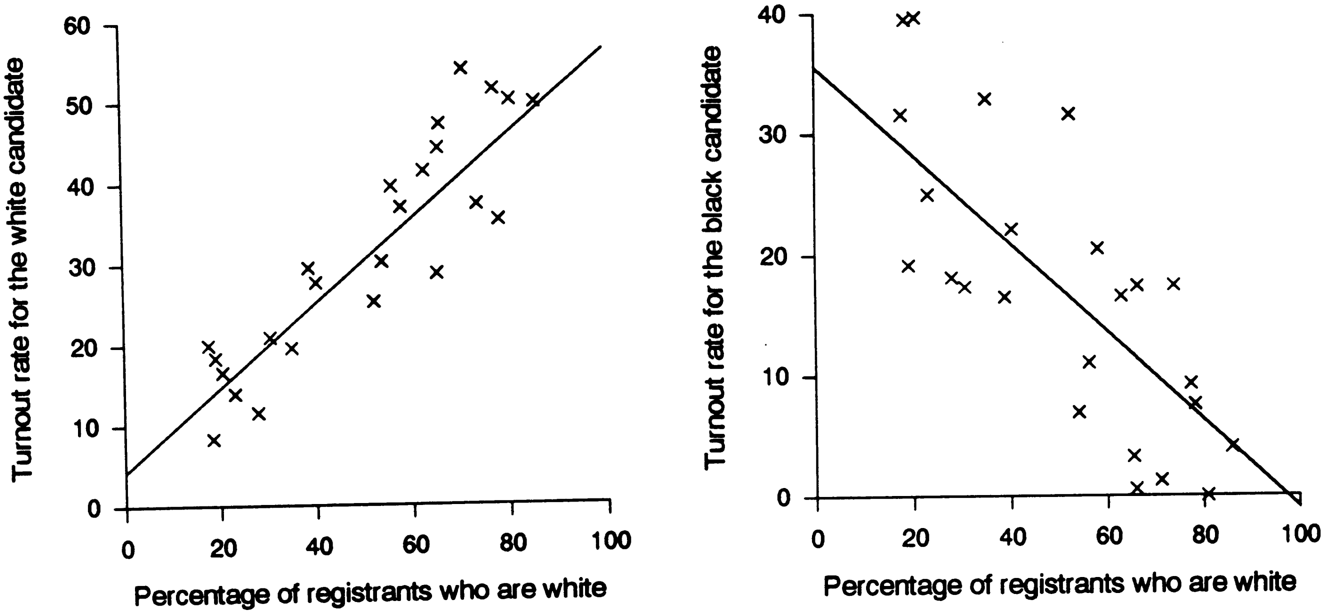
Data are drawn mainly from the example in Loewen and Grofman's article (the 1982 Democratic Primary Runoff for Auditor, Lee County, South Carolina) but also from two recent California cases: Badillo et al. v. the City of Stockton, and Garza et al. v. County of Los Angeles [3]. In the California cases, exit poll data were available. In Los Angeles, data on party affiliation and socio-economic status were available too. So, more rigorous tests of the models were feasible.

[3] In the Stockton and Los Angeles cases, Dr. Grofman consulted for the plaintiffs and we consulted for the defense. The District Court ruled for the defense in Stockton and for the plaintiffs in Los Angeles.

ECOLOGICAL REGRESSION AND THE LINEAR NEIGHBORHOOD MODEL

This section presents the models in the context of the 1982 democratic primary for Auditor in Lee County, South Carolina. There were two candidates; one white, the other black. Basic inputs to ecological regression are the data for each precinct in the county, showing the percentage of registrants in the precinct who are white, and the "turnout rate" for the white candidate: i.e., the percentage of all the registrants in the precinct who came to the polls and voted for the white candidate. These data are shown in the left hand panel of Figure 1.

Figure 1. Left hand panel: Turnout rate for the white candidate plotted against the percentage of whites in the precinct. Right hand panel: Turnout rate for the black candidate plotted against the percentage of whites in the precinct. Regression lines are shown for each plot. (Data on the 1982 Democratic Primary Runoff for Auditor, Lee County, from Loewen and Grofman, 1989.)



A regression equation is fitted to the data. This equation is used to estimate turnout rates by race for the white candidate. In effect, the equation is used to predict the turnout rate for the white candidate in a hypothetical 100% black precinct; that rate is assumed to apply to all black registrants, wherever they live. Similarly, the equation predicts the turnout rate for the white candidate in a hypothetical 100% white precinct; the result is applied to all white registrants. A more detailed explanation of the mechanics will be found in the Appendix. This appendix is divided into Sections A1 through A6, providing technical backup for the arguments presented in the body of the paper; ecological regression is discussed in Section A2.

For example, as shown in Table 2, the estimated turnout rate for the white candidate in a 100% black precinct is 4.18%. Then 4.18% of the black registrants are assumed to turn out for the white candidate in largely black precincts, in largely white precincts, and in mixed precincts. Similarly, the estimated turnout rate for the white candidate in a 100% white precinct is 56.10%. And 56.10% of the white registrants are assumed to turn out for the white candidate in largely black precincts, in largely white precincts, and in mixed precincts. That is the constancy assumption.

Table 2. Ecological regression estimates for turnout rates by racial groups for the white and black candidate; 1982 Democratic Primary Runoff for Auditor, Lee County. (Data from Loewen and Grofman, 1989.)

Estimated turnout rate for	By black voters	By white voters
White candidate	4.18%	56.10%
Black candidate	35.57%	-.95%
All candidates	39.75%	55.15%

A second regression equation is used to estimate group turnout rates for the black candidate. This equation is fitted to precinct-level data, showing the percentage of registrants in the precinct who are white, and the turnout rate for the black candidate. (Right hand panel, Figure 1.)

The next step is to compute group "support rates" for the white candidate: i.e., the percentage of each group's voters who cast their ballots for that candidate. For example, based on the estimates in Table 1, 4.18%+35.57% = 39.75% of the black registrants voted; and 4.18/39.75 = 11% of the black voters cast their ballots for the white candidate. Thus, the black support rate for the white candidate is estimated as 11%. See the first line in Table 3. Similarly, the white support rate for the white candidate is estimated as 102%. More about impossible values-- like -.95% or 102%-- below.

Table 3. Estimated support rates for the white candidate, by blacks and whites.

	By black voters	By white voters
Ecological Regression	11%	102%
Linear Neighborhood	53%	67%

The linear neighborhood model gives alternative estimates for group turnout and support rates. Ironically, these estimates are based on exactly the same equations as ecological regression-- with a twist. Instead of predicting what would happen in hypothetical precincts that are 100% black or 100% white, the linear neighborhood model uses the equations on the real precincts, to estimate each group's turnout for each candidate. Summing across precincts gives overall turnout and support rates (Section A3). The support rates are shown in the second line of Table 3.

According to the neigborhood model, the blacks and whites in a precinct vote the same way. However, racial groups are not spread evenly across precincts. Some precincts have a higher percentage of whites, some have a lower percentage. And voting preferences differ from one precinct to another. As a consequence, when the community is taken as a whole, the neighborhood model may indicate that the groups vote differently.

In Lee County, the linear neighborhood model estimates the black support rate for the white candidate as 53%, and the white support rate for the white candidate as 67%. According to the model, the groups do have different voting preferences. But the difference is much smaller than the 11% *versus* 102% suggested by ecological regression.

The models agree on the total number of votes cast for each candidate, but disagree on the breakdown of votes by racial group (Section A3, Tables 8 and 9). Of course, precinct vote totals for each candidate are observable; indeed, they are a matter of public record. The breakdown by racial group is not observable-- due to the secrecy of the ballot. *The models agree perfectly on the observables, but disagree radically on the unobservables.*

THE FIRST THREE CHECKS

Ecological regression rides on the validity of the constancy assumption. Loewen and Grofman discuss "contextual effects," which are threats to this assumption. For example, the constancy assumption would be violated if whites and blacks who lived in higher-income areas voted for politically conservative candidates while those who lived in poorer areas voted for liberal candidates.

Loewen and Grofman say that "five checks can uncover error owing to contextual effects" (p.598). Checks I, II, and III are discussed in this section; IV and V, in the next section [4].

Check I-- Impossible Estimates. According to Loewen and Grofman "First, and most obviously, any estimate...lower than 0 percent or higher than 100 percent implies a problem" (pp.598-599). Their 102% support rate from ecological regression is "technically impossible (as it is greater than one hundred percent)." They recommend that it "be 'rounded' down to the largest possible number, 100 percent" (p.594). Similarly, they round negative rates up to 0.

Check I is important, but Loewen and Grofman do not pay attention to it. In general, their strategy for dealing with physically impossible estimates only sweeps the problem under the rug. Indeed, they convert evidence against their model to evidence of polarization. Thus, plaintiffs' experts often argue that support estimates over 100% demonstrate strong polarization. "Such 'impossible' estimates generally indicate true levels of racial bloc voting very close to 100%" (Loewen and Grofman, 1989, p.599). However, there are many examples where ecological regression gives support rates over 100%, while exit polls show the true support rates to be well below 50% (Section A6 and Freedman et al., 1991, Table 9).

[4] We can make one additional empirical test of the constancy assumption in the Lee County data. The assumption implies that "blacks turn out, rollon, and vote similarly in all precincts" (p.603). Therefore, the constancy assumption would also be violated if the blacks who lived in largely black precincts were more likely to vote than were the blacks who lived in other precincts.

This breakdown of the constancy assumption did occur in Lee County. An analysis of the data in Table 1 of Loewen and Grofman's article shows there was a statistically significant correlation ($r = .45$, $p <.05$) between the percentage of registrants in the precinct who were black and the percentage of black registrants who turned out to vote.

From a statistical perspective, if a model gives impossible estimates, then the model is wrong and its results are untrustworthy even when they fall into the range of the possible [5]. Ecological regression generates impossible estimates at a high rate. In the Los Angeles County redistricting case, ecological regression yielded impossible estimates, ranging from -5% to 231%, in 5 of the 12 partisan elections analyzed by the plaintiffs (Freedman et al., 1991, Table 3). In a contest for supervisor that took place the day after the Court ruled, ecological regression estimated that 113% of the hispanics supported hispanic candidates while -10% of the hispanics supported black candidates.

To sum up, ecological regression often fails Check I. However, the linear neighborhood model usually passes. For example, it estimated the support rates in the Auditor contest as 53% and 67%, rather than 11% and 102% [6].

[5] There is one escape hatch: Estimates may fall outside the possible range by small amounts which can be explained on the basis of sampling error. In Lee County, the impossible values can probably be explained that way; but in Los Angeles and Stockton, results from ecological regression were too impossible to explain by sampling error.

[6] Occasionally, the linear neighborhood model will produce impossible results for individual precincts that are heavily white or heavily black. Freedman et al. (1991) propose a nonlinear variant of the neighborhood model, which always produces estimates that are physically possible.

Check II-- Compare the Estimate of Total Votes Cast for Each Candidate with the Actual Total. To estimate the total vote for a candidate, Loewen and Grofman multiply the estimated turnout rate by the number of registrants in each group. The products are summed to estimate the total vote, and the estimate is compared to the actual total. For example, according to ecological regression, 4.18% of the 5534 black registrants turned out for the white candidate; and so did 56.10% of the 2624 white registrants (Table 2). The estimated number of votes for the white candidate is

$$.0418 \times 5534 + .5610 \times 2624 = 231 + 2624 = 2855.$$

This estimate matches the actual number, so ecological regression passes Check II. Unfortunately, the check is tautological. The regression methodology insures that the estimated total vote matches the actual total, candidate by candidate [7]. Furthermore, since the linear neighborhood model and the ecological regression model use the same equations, both models pass or fail Check II to exactly the same degree (Section A3).

[7] The totals will match exactly if the regression equation weights precincts by the number of registrants (Freedman et al., 1991). With other weights, the match is approximate: if there is considerable variation in the sizes of the precincts and this variation is related to racial makeup and voting preferences, Check II may fail.

Loewen and Grofman say (p.590 n.2) they weighted precincts by turnout, but that gives slightly different coefficients from the reported ones. In fact, their equations result from weighting precincts by number of registrants.

Loewen and Grofman obtain an estimate of 2882 votes for the white candidate, compared to the actual value of 2855. The discrepancy results from a slip (p.599): They mistakenly used black and white turnout rates for the white candidate of 3.9% and 57%, rather than the rates from their regression equation (Figure 2, p.594).

Check III-- Visual Inspection of Scatterplots. Inspection of scatterplots for the turnout data (Figure 1) shows that the points tend to follow the regression line, as the constancy assumption requires. A curvilinear relationship would contradict the constancy assumption.

The implications deserve closer scrutiny. The constancy assumption does require a linear trend. But a linear trend does not rule out contextual effects, because these effects can be linearly related to the percentage of minority registrants in the precinct. For example, Freedman et al. (1991) found linear contextual effects in the Los Angeles data. Thus, Check III looks only at one special way that the constancy assumption can fail.

Loewen and Grofman also say that a high correlation between voting and race at the precinct level indicates that "contextual effects are not a major problem" (p.598). This assertion is wrong, and the reason is the same. Indeed, the linear neighborhood model-- which is based on contextual effects-- makes the same predictions about the observables as the ecological regression model. Both models use the same scatterplots and involve the same correlation coefficients. Check III, even supplemented by examination of correlations, cannot determine which model is right, if either.

Table 4 sums up the discussion of the first three checks. Ecological regression often fails Check I; the neighborhood model usually passes. Both models pass or fail Checks II and III to exactly the same degree.

Table 4. Comparison of models on Checks I, II, and III: Lee County.

		Is the check passed?		
Check	What is investigated?	Ecological regression	Linear neighborhood	Do the models agree?
I	Are estimates physically possible?	No	Yes	No
II	Does estimated total match actual total?	Yes	Yes	Yes
IIIa	Is the relationship linear?	Yes	Yes	Yes
IIIb	Is the correlation high?	Yes	Yes	Yes

CHECKS BASED ON HOMOGENEOUS PRECINCTS

We turn now to Checks IV and V, and introduce a variant called "Check VI", which puts the idea behind Checks IV and V into its clearest form. These checks are based on the comparison of model predictions to actual results in homogeneous precincts. Indeed, all three checks are the same, when applied to precincts that are truly homogeneous. Ordinarily, however, there is some residential integration even in the most extreme precinct. Check IV ignores heterogeneity; Checks V and VI adjust for it, but in different ways. We present the checks, then give a critique.

Check IV. Lee County does not have precincts that are all black or all white. Therefore, Loewen and Grofman focus on four "heavily black precincts" that are 79% or more black in registration; and on four "heavily white precincts" that are 77% or more white in registration.

In the heavily black precincts, the actual support rate for the white candidate (from both races combined) was 32%. Ecological regression estimated the black support rate for the white candidate as 11%. Check IV compares the estimated 11% with the actual 32%.

Similarly, in the heavily white precincts, the support rate for the white candidate (from both races combined) was 86%. The ecological regression estimate of the white support rate is 102%. Check IV compares the estimated 102% with the actual 86%.

There is a noticeable discrepancy between estimates and actuals. A partial explanation is that the heavily black precincts have many white registrants, just as there are many black registrants in the heavily white precincts. The actuals are based on mixed precincts, while the ecological regression estimates are for hypothetical precincts that are 100% black or 100% white.

Check V. This check relies on data from the same heavily black and heavily white precincts that were used in Check IV. But now, these data are used to extrapolate the support rates to hypothetical 100% black or 100% white precincts. The extrapolations are compared to the estimates derived from ecological regression (Section A4).

In a hypothetical 100% black precinct, the extrapolated support rate for the white candidate is 15.5%. Likewise, the extrapolated support rate for the white candidate in a hypothetical 100% white precinct is 103.8%. These extrapolations are compared to the estimated support rates of 11% and 102% from ecological regression. Agreement is rather good. But the linear neighborhood model also estimates a support rate of 11% in purely black precincts, and 102% in purely white precincts (Section A3). So Check V does not help.

Check VI. The ecological regression model can estimate, precinct by precinct, the number of black registrants who voted for the white candidate, as well as the number of whites who voted for the white candidate. So can the linear neighborhood model (Section A3). Vote totals for the candidates can then be found, by combining blacks and whites.

Table 5 shows the results for the four heavily black precincts (in the aggregate), and the four heavily white precincts. In each group of precincts, estimated support rates-- from blacks and whites combined-- can be computed and compared with actuals (Table 6). Both models pass Check VI: indeed, they make identical estimates of total votes for each candidate in each group of precincts.

Table 5. Estimated number of votes cast for each candidate in the heavily black and heavily white precincts, by black voters and by white voters.

	Heavily Black		Heavily White	
	ER	LNM	ER	LNM
Black candidate				
Black voters	398	321	73	14
White voters	-2	75	-7	52
Total	396	396	66	66
White candidate				
Black voters	47	157	8	93
White voters	146	36	443	358
Total	193	193	451	451
Total	589	589	517	517

Notes: ER = ecological regression, LNM = linear neighborhood model. Estimates rounded to integers.

Table 6. Estimated and actual percentage of votes cast for the white candidate in heavily black and heavily white precincts, by blacks and whites combined.

Estimation procedure	Heavily Black	Heavily White
Ecological regression	32.8%	87.2%
Linear neighborhood model	32.8%	87.2%
Actuals	32.2%	85.6%

Note: Actual support rates are known for all voters combined, but are not known for separate racial groups-- due to the secret ballot.

Discussion. Checks IV, V, and VI cannot validate the constancy assumption. Indeed, ecological regression and the linear neighborhood model both pass or fail these checks to exactly the same degree (Table 7). Why the coincidence? The linear neighborhood model and the ecological regression model predict the same total support rates for each candidate, precinct by precinct. Furthermore, they predict the same total support rates in hypothetical 100% black precincts, or 100% white precincts. The two models have to make the same predictions, because they use the same regression equations (Section A3).

The linear neighborhood model incorporates strong contextual effects, violating the constancy assumption. The ecological regression model depends on the constancy assumption. Since both models pass Checks IV through VI, these checks cannot demonstrate the validity of the constancy assumption.

Table 7. Summary of checks involving homogeneous precincts.

Check	What does the model estimate?	Benchmark	Do models pass to the same degree?
IV	Vote in 100% precincts	Actual vote in extreme precincts	Yes
V	Vote in 100% precincts	Extrapolation to 100% precincts from extreme precincts	Yes
VI	Vote in extreme precincts	Actual vote in extreme precincts	Yes

Notes: A "100% precinct" is one that is 100% white or 100% black. Following Loewen and Grofman, an "extreme" black precinct is 79% or more black in registration; an "extreme" white precinct is 77% or more white.

We now comment on these checks one at a time. Check IV is flawed because it is does not correct for heterogeneity. Check V is flawed, because the extrapolation is unjustified [8]. Check VI seems to be the most sensible of the three, and the agreement between estimates and actuals is quite striking. Unfortunately, this agreement says little about voting behavior. If the turnout data show a linear trend-- which they do in Lee County-- Check VI must be passed (Section A5). Thus, Check VI covers the same ground as Check III; the agreement between estimates and actuals only confirms the linearity of the data.

---

[8] There are two types of ecological regression, single-equation and double-equation. This article uses the double-equation method, as do Loewen and Grofman. With linear support data, the line used to make the extrapolations in Check V is almost the regression line from the single-equation method. Thus, Check V validates the double-equation method by reference to the single-equation method. We find this quite curious, since Loewen and Grofman have been very critical of the single-equation method.

More technically, the single-equation method agrees with the double-equation method when total turnout rates (blacks and whites combined, all candidates combined) are not systematically related to the racial makeup of the precincts. Thus, Check V tests this feature of turnout rates, and the linearity of the support rate data.

CIRCULAR ARGUMENTS

According to Loewen and Grofman, Check V "amounts to a test of the assumption made in ecological regression...that blacks turn out, rollon, and vote similarly in all precincts" (p.603). However, to justify Check V, Loewen and Grofman "assume that racial groups behaved consistently across precincts. That is...black voters in the white precincts voted like black voters in black precincts, and that white voters in black precincts voted like white voters in white precincts" (p.602). This is simply a paraphrase of the constancy assumption. In other words, the constancy assumption has been used to check itself.

Loewen and Grofman offer the same circular reasoning in defense of Check IV: "persons in integrated neighborhoods usually vote like other members of their racial group rather than members of the 'opposite' group" (p.602). This hypothesis about racial constancy is essential for making inferences about how all blacks vote on the basis of the vote in the heavily black precincts. Again, the constancy assumption has been used to check itself.

SUMMARY AND CONCLUSIONS

Ecological regression is used by plaintiffs in voting rights litigation to determine how groups vote. This procedure is based on the very powerful-- and highly questionable-- constancy assumption. Specifically, ecological regression assumes that (apart from random variation) members of a racial or ethnic group vote the same way, no matter where they live.

The neighborhood model turns the constancy assumption on its head by positing that, apart from random variation, whites and blacks in each precinct vote alike. Not surprisingly, the neighborhood model estimates of group support rates are radically different from the ecological regression estimates.

Loewen and Grofman propose five checks to validate the constancy assumption. These checks cannot do the job. Ecological regression often fails Check I. Check II is a mathematical tautology. Checks III, IV and V are passed to exactly the same degree by both the linear neighborhood model and the ecological regression model, although these two models make very different assumptions about voting behavior. The five checks cannot tell which model is right, if either.

In Los Angeles, data were available on party affiliation and socio-economic status. Exit poll data were available too. In our opinion (Freedman et al., 1991), these data showed the constancy assumption to be quite wrong. The neighborhood model fitted the facts much better than did ecological regression. For Lee County, the constancy assumption may hold, or it may not; the neigborhood model may be better, or worse. The five checks do not help in deciding the issue, in Lee County or anywhere else.

REFERENCES

Freedman, D., S. Klein, J. Sacks, C. Smythe, and C. Everett, 1991. Ecological regression and voting rights. *Evaluation Review* (in press)

Freedman, D., R. Pisani, R. Purves, and A. Adhikari, *Statistics*. 2nd edition. (New York: W.W. Norton & Co., Inc., 1991.)

Goodman, L. 1953. Ecological regression and the behavior of individuals. *American Sociological Review* 18:663-4.

Goodman, L. 1959. Some alternatives to ecological correlation. *American Journal of Sociology* 64:610-25.

Grofman, B., M. Migalski, and N. Noviello, 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* 7:199-233.

Loewen, J. and B. Grofman, 1989. Recent developments in methods used in vote dilution litigation. *The Urban Lawyer* 21:589-604.

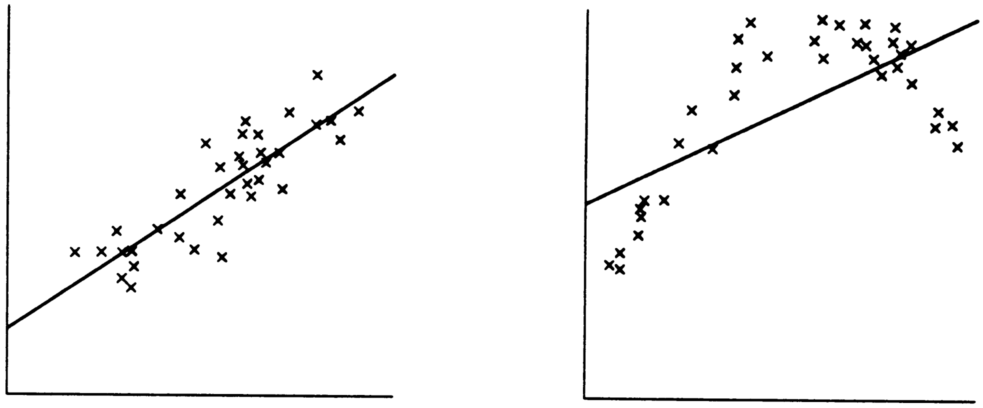
Loewen, J., 1990. Sand in the bearings: Mistaken criticisms of ecological regression. *The Urban Lawyer* 22:503-513.

TECHNICAL APPENDIX

This appendix has several sections. Section A1 presents some background on linear regression. Section A2 reviews ecological regression; Section A3, the linear neighborhood model. Section A4 discusses Check V; Section A5, Check VI. Section A6 uses data from Stockton to illustrate how ecological regression can estimate a support rate of over 100% when the actual rate is less than 50%.

A1. Background on regression

Figure 2. Two scatterplots and their regression lines. The left hand plot is linear. The right hand plot is non-linear.



The left hand panel of Figure 2 shows "linear" data, that is, data where the underlying trend is linear. The right hand panel shows "non-linear" data: as x increases, y first increases then decreases. With linear data, the underlying trend is captured by the regression line. In essence, this line picks off the average value of y corresponding to each value of x. By contrast, with non-linear data, the average value of y may be quite far from the regression line (right hand panel of Figure 2).

Figure 3. The regression line picks off the average value of y corresponding to each value of x.

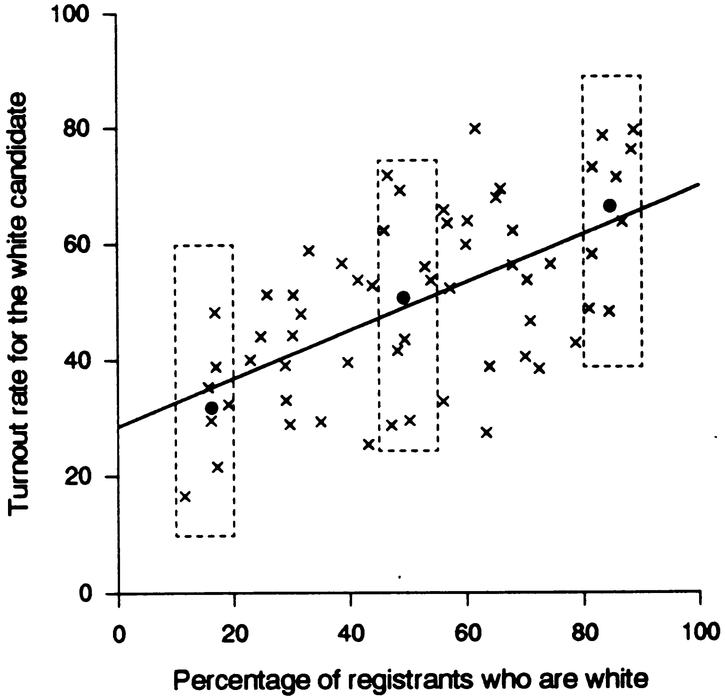


Figure 3 indicates how the regression line picks off the average value of y corresponding to each value of x. This figure shows (hypothetical) election data: the turnout rate for the white candidate in a precinct versus the percentage of registrants in the precinct who are white. By way of illustration, three groups of precincts are marked by rectangles in the figure. The left hand group has 10%-20% white registration; the middle group has 45%-55% white registration; the right hand group has 80%-90% white registration.

A heavy dot marks the center of each of the three groups. (The dot is placed at the average of the data, horizontally and vertically; so it is a little off the center of the rectangle.) With linear data, and a reasonable number of points in each group, the regression line will pass very close to each of the central dots. For more discussion, see (Freedman et al, 1991, sec. 10.2).

A2. How ecological regression works

Basic inputs to ecological regression are plotted in Figure 1. The horizontal axis shows the percentage of registrants in the precinct who are white. The vertical axis in the left hand panel shows the "turnout rate" for the white candidate: i.e., the percentage of all the registrants in the precinct who came to the polls and voted for the white candidate. Each point represents one of the county's 24 precincts. A straight line is fitted to the data by regression. This line is shown in the figure. The equation for the line is

(1) Estimated turnout rate for white candidate = 4.18% + .5192X

In this equation, X stands for the percentage of white registrants in the precinct. In a hypothetical all-black precinct, X = 0%. So the estimated turnout rate for the white candidate is 4.18%:

$$4.18\% + (.5192 \times 0\%) = 4.18\%$$

Similarly, in a hypothetical all-white precinct, X = 100% and the estimated turnout rate for the white candidate is 56.10%.

With ecological regression, the black turnout rate of 4.18% for the white candidate is assumed to be constant across all precincts. Similarly, 56.10% of the white registrants are assumed to turn out for the white candidate in all precincts. The constancy assumption does allow random variation around these levels.

Ecological regression uses a second scatterplot to estimate the turnout rate for the black candidate (right hand panel of Figure 1). As before, the horizontal axis shows the percentage of registrants in the precinct who are white. The vertical axis shows the turnout rate for the black candidate. The straight line fitted to these data has the equation

(2) Estimated turnout rate for black candidate = 35.57% - .3652X

On the basis of this equation, 35.57% of the black registrants and -0.95% of the white registrants are estimated to have turned out for the black candidate. The group turnout rates in Table 2 were estimated from Equations 1 and 2.

The last step is to compute the group "support rate" for each candidate: i.e., the percentage of the group's voters who cast their ballots for that candidate. For instance, according to ecological regression, 39.75% of the black registrants voted: 4.18% + 35.57% = 39.75%. And, of those who voted, 11% supported the white candidate:

$$4.18/39.75 = 11\%.$$

So the black support rate for the white candidate is 11%. The white support rate for the white candidate is computed in a similar fashion:

$$56.10/(56.10 - 0.95) = 102\%.$$

To summarize, ecological regression says that 11% of the blacks and 102% of the whites supported the white candidate. (These group support rates were reported in the first line of Table 3.)

A3. The neighborhood model

Freedman et al. (1991) describe the neighborhood model. According to this model, apart from random variation, the blacks and whites in a precinct vote alike. The linear version of the model uses the same equations as ecological regression. From Equation 1, the estimated turnout rate for the white candidate in a precinct is 4.18% + .5192X, where X is the percentage of registrants in the precinct who are white.

To estimate group turnout rates for the white candidate, the linear neighborhood model uses the actual value of X in the precinct. For instance, 444 of the 567 registrants in Precinct 1 were white; so this precinct was 78.31% white. Inserting 78.31% into Equation 1 yields an estimate of 44.84% for the white candidate's turnout rate in the precinct:

$$4.18\% + (.5192 \times 78.31\%) = 44.84\%.$$

To estimate how many registrants in each group turned out and voted for the white candidate in Precinct 1, the linear neighborhood model applies the 44.84% rate to both the 123 black and 444 white registrants. The estimates are 55 and 199, because 44.84% of 123 = 55 and 44.84% of 444 = 199.

Estimates for the black candidate are done the same way, starting from Equation 2. The linear neighborhood model says that 6.97% of the registrants in Precinct 1 turned out and voted for the black candidate, because

35.57% - (.3652 x 78.31%) = 6.97%.

To calculate how many registrants in each group turned out for the white candidate in Precinct 1, the 6.97% rate is applied to the 123 black and 444 white registrants. The results are 9 and 31, because 6.97% of 123 = 9 and 6.97% of 444 = 31.

By contrast with the linear neighborhood model, the ecological regression model takes Equation 1 and sets X to 0% or 100% in every precinct. The rate of 4.18%, corresponding to X=0%, is applied to the blacks; the rate of 56.10%, corresponding to X=100%, is applied to the whites. The results in Precinct 1: 4.18% of 123 = 5 blacks voted for the white candidate, and 56.10% of 444 = 249 whites voted for the white candidate. Both models estimate a total of 254 votes for the white candidate in Precinct 1; but the breakdown by race is different. Similar conclusions apply to the black candidate (Table 8).

Table 8. Estimates of votes cast in Precinct 1, by ecological regression and the linear neighborhood model.

	Ecological regression	Linear neighborhood
Black candidate		
Black voters	44	9
White voters	-4	31
Total	40	40
White candidate		
Black voters	5	55
White voters	249	199
Total	254	254
Total	294	294

These procedures are repeated for each precinct. In sum, the linear neighborhood model says that 2529 blacks voted in the Auditor contest: 1187 of them turned out for the black candidate and 1342 for the white candidate. Furthermore, 2250 whites voted for Auditor: 737 voted for the black candidate and 1513 for the white candidate (Table 9).

The linear neighborhood model estimates the black support rate for the white candidate as 53%. The arithmetic: 1342/(1342+1187)=53%. Similarly, the white support rate for the white candidate is 67%. (These figures were reported in the second line of Table 3.)

As Tables 8 and 9 show, the models agree on the observables (total votes for each candidate in each precinct). The reason is that the models use the same equations. However, the models disagree on the unobservables-- the vote breakdown by racial group. That is because the models make different assumptions about group voting behavior.

Table 9. Comparison of estimates to actuals: all precincts. (The question marks in the "actuals" column denote unobservable entries.)

	Estimates		
	Ecological regression	Linear neighborhood	Actuals
Black candidate			
Black voters	1968	1187	?
White voters	-44	737	?
Total	1924	1924	1924
White candidate			
Black voters	231	1342	?
White voters	2624	1513	?
Total	2855	2855	2855
Total votes	4779	4779	4779

A4. Check V

The details of Check V, as presented by Loewen and Grofman, are quite complicated. A graphical description of the method (Figure 4) is equivalent, and simpler to explain. The x-axis in Figure 4 is the same as in Figure 1(the percentage of registrants in the precinct who are white). The y-axis shows the white candidate's "support rate", or share of the total votes cast. One of the two solid dots in the figure corresponds to the set of four heavily black precincts; the other, to the set of four heavily white precincts. A straight line is drawn to join the two dots. The equation for the line is

(3) Estimated support rate for white candidate based on only the heavily black and heavily white precincts = 15.5% + .883X.

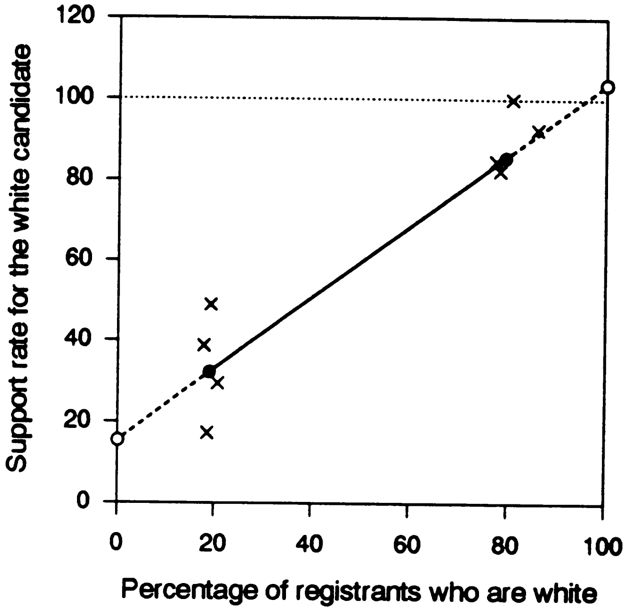
The slope and intercept of the line are derived as follows. In the four heavily black precincts, 18.9% of the registrants were white and the support rate for the white candidate among all voters was 32.2%. The dot corresponding to these four precincts has x=18.9% and y=32.2%. The corresponding values for the four heavily white precincts were 79.4% and 85.6%; the coordinates for the dot are (79.4%,85.6%). The slope of the line joining the dots is .883:

$$(85.6 - 32.2)/(79.4 - 18.9) = .883.$$

The intercept of the line is 15.5%:

$$32.2\% - (.883 \times 18.9\%) = 15.5\%.$$

Figure 4. Support rates for the white candidate plotted against the percentage of whites in the precinct, and extrapolated linearly to hypothetical precincts which are 100% black or 100% white. The solid dot at the left represents the aggregate of the four heavily black precincts. The solid dot at the right represents the aggregate of the four heavily white precincts. The open dots represent the extrapolations.



In a hypothetical 100% black precinct, Equation 3 extrapolates the support rate for the white candidate as 15.5%, because 15.5% + (.883 x 0%) = 15.5%. Likewise, the extrapolated support rate for the white candidate in a hypothetical 100% white precinct is 103.8%. These extrapolations are compared to the estimated support rates of 11% and 102% from ecological regression.

As presented by Loewen and Grofman, Check V involves an iterative process. We now demonstrate that our graphical method is equivalent. Let f be the overall fraction of whites in the set of four heavily black precincts, and s the overall support rate for the white candidate in those precincts. Let g and t be the corresponding fractions for the set of four heavily white precincts: 0<f<g<1 and 0<s<t<1.

Let s_n be the nth approximation to the support rate for the white candidate in a hypothetical black precinct. Let t_n be the nth approximation to the support rate for the white candidate in a hypothetical white precinct. Loewen and Grofman's iterative process is the following. Set $s_0=s$ and $t_0=t$; and for $n=0,1,2...$,

$$s_{n+1} = \frac{s-ft_n}{1-f}$$

$$t_{n+1} = \frac{t-(1-g)s_n}{g}$$

Loewen and Grofman stop at $n=1$. However, the process can be completed analytically. Let $y=L(x)$ be the equation of the straight line joining (f,s) and (g,t). Equation 3 gives an explicit formula for L, in the Lee County example.

It can be proved mathematically that s_n decreases to $L(0)$ while t_n increases to $L(1)$. The proof is inductive; the main idea can be seen by verifying that $t_1=L(g_1)$, where

$$1-g_1= \frac{f}{g}(1-g).$$

Now $0<f<g$, so $g<g_1<1$ and $t<t_1<L(1)$. The quantities f_n and g_n can be defined analogously; $f_0=f$ and f_n decreases with n; $g_0=g$ and g_n increases with n; $s_n=L(f_n)$ and $t_n=L(g_n)$. Finally,

$$1-g_n<(\frac{f}{g})^n(1-g).$$

So g_n increases to 1. Therefore, the sequence of of points (g_n, t_n) moves upward to the right along the line $y=L(x)$, starting from (g,t) and converging to $(1,L(1))$. In particular, t_n converges to $L(1)$. Likewise, the sequence of points (f_n, s_n) moves downward to the left along the line, starting from (f,s) and converging to $(0,L(0))$. In particular, s_n converges to $L(0)$. **QED**

Our graphical procedure is therefore equivalent to Loewen and Grofman's iteration. But the numerical results differ slightly from the ones they present. The reason seems to be that we completed the iteration; they stopped after two moves.

A5. Check VI

If the turnout data are linear, Check VI will be satisfied. Indeed, the heavily black precincts correspond to the left hand group in Figure 3. The regression line-- which is used by both models-- passes close to the group average. Hence, for this group of precincts, the estimated turnout for the white candidate must be quite close to the actual turnout. Likewise, the estimated total turnout must be close to right. Therefore, the estimated support rate must be about right too, because

$$\text{support rate} = \frac{\text{turnout for the candidate}}{\text{total turnout}}$$

The same reasoning applies to heavily white precincts-- or to a middle group of precincts.

To sum up, Check VI will be satisfied if the turnout data are linear. Since Check VI only uses data at the left and right ends of the scatterplot, it is only a partial test of linearity.

A6. Ecological regression can estimate a support rate to be over 100% even when it is really under 50%

Two ecological regression analyses were conducted on the June 1988 Democratic Presidential Primary in Stockton. The first analysis used all of the city's 130 precincts and the actual election returns in each precinct. The resulting estimate: 109% of Stockton's hispanic voters cast their ballots for Jesse Jackson. However, an exit poll showed that the hispanic support rate for Jackson was only 35%. See Table 10.

When exit polls and ecological regressions differ, plaintiffs' experts generally seem to prefer the regression estimates. For example, according to Loewen and Grofman "surveys only tell how people *say* they voted. Ecological regression based on election returns can yield information as to how categories of people *actually* voted" (p. 589; see also Loewen, 1990, p. 504).

Despite Loewen and Grofman, we consider surveys to be more reliable than ecological regression, but that does not need to be argued here. Instead, we make our point by doing an ecological regression analysis in a situation where voter preferences are known.

In the Stockton exit poll, conducted by the Field Research Corporation, respondents filled out a brief questionnaire as they left the polls, stating their ethnicity and voting preferences. (To insure anonymity, respondents dropped the completed forms into a locked "ballot" box.)

There were a total of 1849 respondents in 39 precincts. These 1849 questionnaires define the "Exit Poll" data set. Each questionnaire reports the respondent's ethnicity (hispanic or non-hispanic), voting precinct, and voting preference-- e.g. for Dukakis or Jackson. In the Exit Poll data set, this expressed preference is the analog of a vote-- a vote which is known.

Table 10. Estimated and actual hispanic support for Jackson in the 1988 Democratic Presidential Primary, Stockton.

	Ecological regression	Neighborhood model	Actual
Citywide	109%	43%	?
Exit Poll	116%	44%	35%

Note: The Citywide data set is based on the 18,803 votes cast in Stockton's 130 precincts. The Exit Poll data set is based on the responses of 1,849 voters in 39 sample precincts.

The Exit Poll data are tabulated so as to mimic the election data that are usually available for a community. This tabulation breaks the link between individuals, their ethnicity, and their preferences. For example, if 10 of the 50 respondents in a precinct said they were hispanic, this precinct is classified as 20% hispanic. Similarly, if 25 of these 50 respondents said they voted for Jackson, his support rate in that precinct was taken as 50%. This procedure was repeated for all of the 39 precincts in the poll.

The additional information available from the poll-- the number of hispanics for Jackson-- is deliberately suppressed for the moment. In effect, ecological regression and the linear neighborhood model both attempt to reconstruct the suppressed data. Ecological regression estimates that 116% of the 298 hispanics in the poll supported Jackson; the linear neighborhood model estimates 44%. In fact, his support rate among the hispanic respondents was 35% (Table 10).

For present purposes, respondents "support" Jackson by checking his name on the exit poll ballot; what happens in the voting both is not material. The conclusion: ecological regression can estimate a support rate over 100% in a situation where the exact rate is known and is well below 50%.

As it happened, the exit poll was very accurate. It predicted that Jackson would receive 39% of all the votes cast in the primary, and he got 35%. (There is an intriguing numerical coincidence: the split among hispanic voters in the poll reflected the actual division of votes in the election.) Moreover, the neighborhood model did a reasonable job at estimating group voting preferences in Stockton; not perfect, but much better than ecological regression.