Why the 1980s Measures of Racially Polarized Voting Are Inadequate for the 1990s

ARTHUR LUPIA and KENNETH McCUE

In this paper, we attempt to clarify some of the confusion that surrounds the measurement of racially polarized voting. This clarification is necessary because the determination of whether or not racially polarized voting exists is often a critical component of the evidence presented in Voting Rights Act (Section 2) litigation. We first show that the correlation coefficient should never be used to measure voting polarization by relating the statistic to the individual behavior that it is supposed to be describing. We then compare the estimates of polarized voting that are provided by other commonly used measures with individual behavior in order to show that the Voting Rights disputes of the 1990s will require different and more carefully specified measures than are currently in use.

I. INTRODUCTION

In this paper, we address some of the problems that have characterized previous attempts to measure racially polarized voting. The measurement of racially polarized voting became increasingly important in 1982 when amendments were made to the Voting Rights Act (Section 2).[1] The amendments significantly altered the types of evidence that litigants could use in their attempts to overturn or protect an allegedly vote-diluting electoral system. For the purposes of this discussion, we define "electoral system" as a districting plan that divides an area into districts of roughly equal population, where each district sends their own representative to a legislature. Added to the previous law was a "results test,"[2] which is a list of factors that are presumed to be concurrent with electoral discrimination against the members of a protected minority group. An existing electoral system can be overturned in a courtroom by showing that several factors listed in the "results test" work together in a way that systematically prevents minority group members from obtaining representation.

One important component of the "results test" is the answer to the question: "Does there exist *racially polarized voting* (the tendency of individuals to vote for persons of their own race) in the electorate?" Polarized

LAW & POLICY, Vol. 12, No. 4, October 1990

ISSN 0265-8240 $3.00

voting within certain types of electoral systems can have a great impact on the extent to which different individuals will be represented. For instance, if district lines are drawn to give a geographically compact minority group a minority in all districts and a majority group engages in racially polarized voting, then the majority group can effectively prevent minority group members from being represented in the legislature. Alternatively, district lines can be drawn to give the minority group a majority in some of the districts, which may increase the likelihood that the minority group is represented in the legislature. The relationship between the location of district boundaries and political representation make the demonstration of the presence (or absence) of racially polarized voting an important piece of evidence in a court's decision as to whether an electoral system should persist or be redrawn.[3]

Establishing the existence of racially polarized voting requires expert witnesses, usually social scientists and sometimes statisticians, to make statistical assertions about the voting behavior of certain types of individuals (for instance, blacks and whites). The use of the secret ballot, and the fact that the domain of many Voting Rights Act disputes includes small jurisdictions or past events, means that individual-level survey instruments for these cases are usually not available and cannot be created. Therefore, experts cannot directly examine the behavior of individual voters and are forced to use statistical methods to infer individual level behavior from aggregated voting returns.

As time and our understanding of statistics have progressed, different methods have been used to test hypotheses of racially polarized voting in an electorate. It has been claimed for some time that one particular method, the correlation coefficient, is an inaccurate way to measure racially polarized voting (Loewen, 1982: 182–185). However, such a claim is usually made with an incomplete or incorrect explanation of why it is a bad measure. These incomplete claims have led some in the legal community to believe that the correlation coefficient is still a viable measure of racially polarized voting under certain circumstances.[4]

In section II, we show that the correlation coefficient does not provide a reliable representation of individual level voting behavior and should *never* be used. In particular, we show that use of the correlation coefficient can lead to one of the following two erroneous conclusions:

1. Existence of racially polarized voting established when, in fact, it does not exist.
2. Non-existence of racially polarized voting established when, in fact, it does exist.

These erroneous conclusions can lead to either a decision to overturn a "legal" electoral system or a decision to keep an "illegal" (or discriminatory) electoral system in place. Either mistaken decision is likely to affect which

groups and interests will be represented in a legislature, which in turn will influence the policies that affect everyone.

In Section III, we examine the merits of alternative measures of racially polarized voting. Since problems with the use of the correlation coefficient have been known to many experts for some time, bivariate and multivariate regression methods have sometimes been used to supply polarized voting estimates in more recent Voting Rights Act cases. While these methods are commonly presumed to provide better estimates of racially polarized voting than the correlation coefficient, we believe that these methods have, at times, been misapplied by litigants to Section 2 litigation. This misapplication causes the presentation of statistics that are not necessarily representative of the degree of racially polarized voting in an electorate. We show the consequences of this misapplication by comparing the estimates of individual behavior provided by previously used bivariate and multivariate regression methods to the actual behavior of individuals. This comparison allows us to demonstrate that the complex multi-ethnic voting rights disputes of the 1990s will require different and more carefully specified measures than those that are currently in use. We then state several conditions that the new methods should satisfy in order to be consistent with the individual behavior they are attempting to measure and argue that only when this consistency can be demonstrated should any method of measuring racially polarized voting be acceptable for use in voting rights litigation.

Since the legislative redistricting that will take place after the 1990 census is certain to produce a new round of Section 2 lawsuits, we believe that the time to clear up existing confusion about which statistical methods should be used in these cases is now. It is important to show why, in a concise but general sense, neither the correlation coefficient nor certain bivariate and multivariate regression methods should be used as a method for determining the existence and extent of racially polarized voting.

II. THE INADEQUACY OF THE CORRELATION COEFFICIENT

In most Voting Rights Act cases, individual level voting returns are not available. Therefore, any estimation of the relationship between race and individual level voting behavior must be conducted using observations that provide an aggregated description of the voting behavior of many individuals. In these circumstances, the correlation coefficient is one measure that experts have used to estimate the relationship between an individual's racial identity and his or her voting behavior. In this section, we provide a simple derivation of the relationship between individual behavior and the correlation coefficient and show why this statistic should never be used in voting rights litigation.

Let us first consider what the correlation coefficient is. In Voting Rights Act cases, it is primarily used as a measure of the linear association (or

relationship) between the proportion of the total vote received by a particular candidate, and the proportion of the population who are members of a particular group (we use the word "group" to refer to a partition of the electorate, and assume that each individual is a member of one and only one group).[5] This measure is a function of these proportions over all electoral units (such as precincts, parishes, etc.) which are within the relevant electoral jurisdiction, and hence is a statistic.

A mathematical definition of the correlation coefficient is as follows. Let N be the number of electoral units in the relevant jurisdiction, with i designating a particular electoral unit. Voting precincts, parishes and census tracts are examples of electoral units that are commonly employed in the measurement of polarized voting.[6] The typical domain (or jurisdiction) of this type of case is an area within a city, a whole city, a county or a state. Let X_i be the proportion of electoral unit i who are members of a group called "black" ($0 \leq X_i \leq 1$). Let $1 - X_i$ be the proportion of individuals in electoral unit i who are members of a group called "white." Also, let $V_i (0 \leq V_i \leq 1)$ be the proportion of electoral unit i that voted for a particular candidate. $\bar{X}$ is the mean (or average) proportion of blacks over all of the electoral units within the relevant jurisdiction, and $\bar{V}$ is the mean proportion of votes for a candidate over all of the electoral units within the entire jurisdiction. With these definitions we can display a mathematical formula for the correlation coefficient as follows:

$$\text{Corr}\,(V,X) = \frac{\Sigma_{i=1}^{N}(X_i - \bar{X})(V_i - \bar{V})}{[[\Sigma_{i=1}^{N}(X_i - \bar{X})^2\,\Sigma_{i=1}^{N}(V_i - \bar{V})^2]^{1/2}].} \qquad (1)$$

The correlation coefficient can take on values ranging from -1 to 1. If the value of the correlation coefficient is 1, that is an indication that X_i and V are perfectly correlated. Thus, for example, when the value of X_i moves in a certain direction (*i.e.* the proportion of the electorate that is black increases across precincts), the value of V moves in the same direction (*i.e.* the proportion of the vote received by the candidate increases) with a constant magnitude. In other words, if one electoral unit is 45 % black, a second electoral unit is 48 % black and the correlation coefficient is 1, this should indicate that the candidate will receive 3 % more of the vote in the second electoral unit than she did in the first electoral unit. If the correlation coefficient is -1, then when the value of X_i or V moves in a certain direction, the value of the other variable moves in the opposite direction with the same magnitude. A correlation coefficient of 0 is indicative of no relationship between "percent black" and "percent voting for the candidate." Figure 1 provides examples of how "percent voting for the candidate" and "percent black" could be related for certain values of the correlation coefficient.

The values of X_i and V in Equation 1 come from observations of aggregated individual behavior. In order to relate the correlation coefficient to individual level voting behavior, we assume that the concept "the

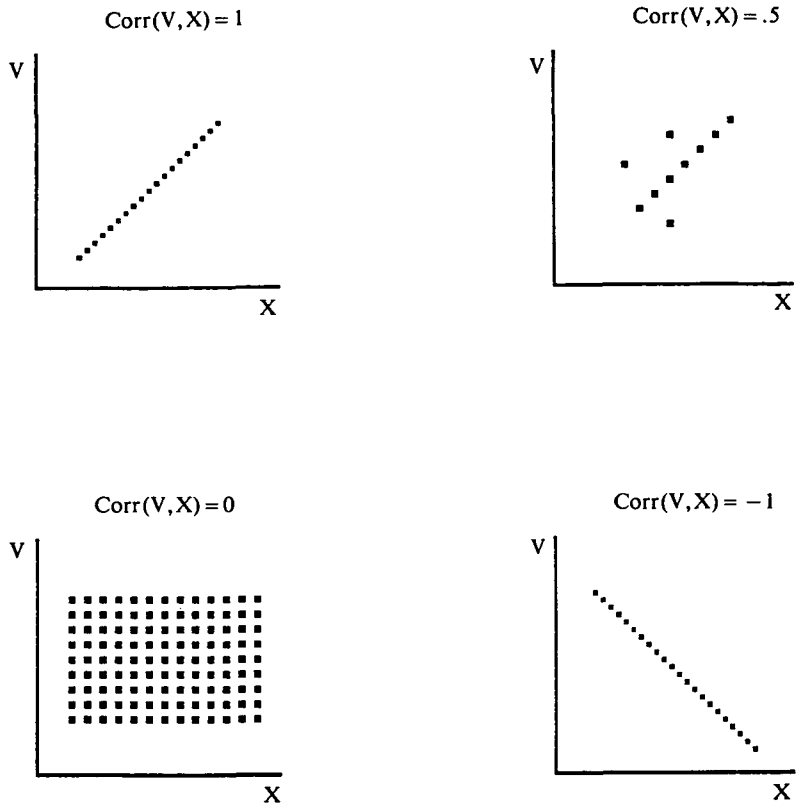


Figure 1 The correlation coefficient.

probability that a certain individual casts a vote for one particular candidate" is valid.[7] It is true that an individual either votes for a particular candidate or does not; however, when we consider an individual as a member of a group, a probabilistic interpretation of individual behavior makes sense. For example, if seventy percent of a group supports a candidate, we would expect that, if we chose a person from that group at random, there would be a seventy percent chance that that particular person supported the candidate.[8] Throughout this paper, we use this concept of individual voting behavior to designate p_1 as the probability that a member of group one will support a particular candidate, and p_2 as the probability that a member of group two will support that candidate.

Let us now reconsider the definition of the correlation coefficient provided in Equation 1. With some plausible assumptions about individual behavior it is possible to derive an alternate form of this expression in a manner which makes the relationship between the correlation coefficient and individual behavior relatively clear. This expression (which is derived in Appendix I), is

$$\text{Corr}(V,X) = \frac{(p_1 - p_2)}{[(p_1 - p_2)^2 + a_1 p_1(1-p_1) + a_2 p_2(1-p_2)]^{1/2}} \qquad (2)$$

In Equation 2, a_1 and a_2 are expectations of functions of the spatial distribution of members of groups one and two within the electorate (*i.e.* they depend on where people live in relation to each other). An expectation is statistical nomenclature for the averaging of a function over a particular distribution, the average being obtained by integration. The mean, for example, is simply the expectation of the average value of the distribution.

Equation 2 shows that the value of the correlation coefficient depends not only on the likelihood that members of a particular group voted for a particular candidate (p_1 and p_2), but also on a factor that bears no necessary relationship to the issue of voting polarization. This factor is the distribution of the different groups across electoral units (a_1 and a_2).

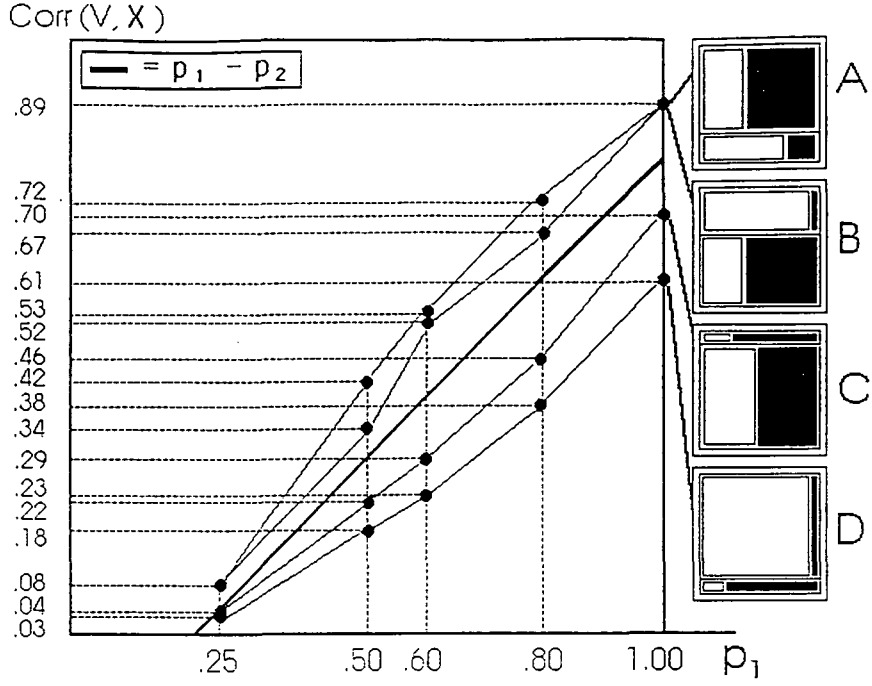
The significance of this finding can easily be seen by comparing two hypothetical electorates, each of which contain members of only two groups.

> Suppose that the electorates are alike in every way, including voting behavior, so that the correlation coefficients of the relationship between group membership and the vote is exactly the same in each electorate. Without changing the voting behavior of any individual, we move one voter in one electorate to another voting precinct in the same electorate. The resulting correlation coefficients are now unequal. The correlation coefficient of an electorate changed although the voting behavior of every individual voter remained constant.

While the distribution of different types of voters within an electorate may be of interest to demographers, differences in the mean and variance of a group's distribution throughout an electorate do not have an obvious place and, in the absence of a (so far non-existent) supportive theoretical argument, should not be included in the determination of the difference in the voting behavior of individuals.

Notice that the correlation coefficient is proportional to the difference $p_1 - p_2$ and will always have the same sign as that difference (so if the first group supports the candidate at a higher rate than the second group, the correlation of the vote for the candidate with that first group will be positive).[9] Using Equation 2, we can also show that as the number of voters within each electoral unit increases, the magnitude of the correlation coefficient becomes greater than the magnitude of the difference $p_1 - p_2$ (in particular, once the number of voters in an electoral unit is over two hundred, the correlation coefficient will almost always be greater than $p_1 - p_2$). As a result, the decision whether to use election precincts or census tracts (which are usually several times larger than election precincts), will produce two different values of the correlation coefficient, even though they are based on the same voting behavior.

We now present some examples that allow us to demonstrate why the correlation coefficient is a bad measure of voting polarization in a wider range of cases than was previously understood. The cases we display in a graphical form are all examples that use Equations 1 and 2. Figure 2 displays values of the correlation coefficient "Corr (V,X)" for four simulated electorates. Each electorate consists of members of only *two* racial or ethnic groups (for instance, black and non-black, or latino and non-latino). X is the proportion of the electorate in group one and $1-X$ is the proportion of electorate in group two.[10]



(probability of support for candidate by member of group 1, with probability of support by member of group 2 held constant at .20.)

KEY

Group Pattern	*Group Name*	*Pct. of Electorate* A	B	C	D
	White	41.5	57.8	43	87.5
	Black	58.5	42.2	57	12.5

Electorate	q	1 - q	c_1	c_2
A	.75	.25	.33	.67
B	.40	.60	.90	.33
C	.10	.90	.25	.45
D	.90	.10	.95	.20

Figure 2 Values of the correlation coefficient for different levels of support and different distributions of ethnic groups.

Each of the four sample electorates contains two types of electoral units. The first type of electoral unit appears with frequency q. c_1 is the percentage of voters in the first type of electoral unit who are members of group one, while $(1-c_1)$ is the percentage of voters in the first type of electoral unit who are members of group two. The second type of electoral unit appears with frequency $1-q$. c_2 is percentage of voters in the second type of electoral unit who are members of group one, while $(1-c_2)$ is the percentage of voters in the second type of electoral unit who are members of group two.

For a better understanding of this notation, consider the following description of Figure 2's sample electorate C (with names of groups added for clarity).

> There are two groups, whites (group one) and blacks (group two), and two types of electoral units or voting precincts (Type I and Type II). A Type I unit is twenty-five percent white ($c_1 = .25$) and seventy-five percent black ($1-c_1 = .75$). A Type II unit is forty-five percent white ($c_2 = .45$) and fifty-five percent black ($1-c_2 = .55$). Of all the electoral units, ten percent are of Type I ($q = .10$) and ninety percent ($1-q = .90$) are of Type II. Electorate C is, thus, forty-three percent white $[(c_1q)+(c_2(1-q)) = .43]$ and fifty-seven percent black $[((1-c_2)q)+((1-c_2)(1-q)) = .57]$.

In the four sample electorates depicted in Figure 2, the racial characteristics of the first type of electoral unit are shown in the top subbox and the racial characteristics of the second type of electoral unit are shown in the bottom subbox. The height of the subboxes are drawn proportional to q and $1-q$. The width of the white and black boxes within the subboxes are drawn proportional to c_1 $(1-c_1)$ within the top subbox and c_2 $(1-c_2)$ within the bottom subbox.

In order to show how the correlation coefficient and individual behavior are related, we have, in Figure 2, varied the value of p_1 while holding p_2 constant at .2. This is but one of many ways that we can show the relationship between the correlation coefficient and individual level voting behavior. A comparison of the electorates, for each value of p_1 displayed, shows that the correlation coefficient can take on many different values for any particular pattern of individual level voting. This comparison shows that the correlation coefficient is not a consistent estimator of voting polarization. How inconsistent is it? Consider the following example that uses Figure 2's electorates.

In electorate D, each member of group two supports a particular candidate with a probability of .8, each member of group one supports the candidate with a probability of .2, $(p_1-p_2 = .6)$, and the correlation coefficient is .38. In electorate A, each member of group one supports the candidate with a probability of .5, each member of group two supports the candidate with a probability of .2, $(p_1-p_2 = .3)$, and the correlation coefficient is .42. Most people would say that there is a much greater difference in the voting behavior of the two groups in Electorate D than in Electorate A, however, the correlation coefficient, as utilized in previous voting rights cases, would indicate that the extent of racially polarized voting is greater in Electorate A!

Depending on how the correlation coefficient is interpreted, it is possible that this measure will lead to a finding of no racial polarization when it actually exists, or vice versa. Let us suppose that a .5 correlation coefficient is the standard against which the existence of racially polarized voting is determined. A .5 standard was used in *Major v. Treen* (1984: 337–338, n. 17) and *Jones v. City of Lubbock* (1984).[11] In the case where the probability that a member of group one supports a particular candidate is .8 and the probability that a member of group two supports the candidate is .2, electorates A and B would be presumed to have racially polarized voting and electorates C and D would not. The use of the .5 standard and the correlation coefficient would mean that for electorates C and D substantively significant racial polarization would not be found to exist when in fact it does exist.

To show the opposite type of mistaken inference, consider the case where the probability that a member of group one supports a particular candidate is .65 and the probability that a member of group two supports the candidate is .2. Electorates A and B would be presumed to have racially polarized voting and electorates C and D would not by the .5 standard, yet there is no difference, across electorates, in the voting behavior of individuals within the groups. In fact, it is the case that, in both of these examples, the only difference in the electorates is in residential patterns. It should be noted that the decision as to which substantive threshold should be implemented to determine racially polarized voting, using any measure, is still being debated.[12] However, the point of this example is that *the correlation coefficient is not the correct way to measure polarization because it can take on different values for electorates voting in exactly the same way.*[13]

Others have suggested that the correlation coefficient should not be used to determine racially polarized voting, however, none has gone far enough to prevent confusion among litigants about whether or not the correlation coefficient can be used. For example, Engstrom and McDonald (1987) have shown that certain bivariate methods, including the correlation coefficient, are inadequate measures of racially polarized voting. They attribute these inadequacies to the presence of more than two racial or ethnic groups in an area. Specifically, they argue that the correlation coefficient can give "systematically biased estimates" when the voting behavior of more than two groups is of interest. For instance, when someone is attempting to estimate the differences between the voting behavior of blacks and whites, while ignoring the voting behavior of a relevant Latino population, proper identification of the actual voting behavior of black and white individuals will not be possible. *Our analysis shows that the problem is not only with the number of groups but also the correlation coefficient itself.*

We now present a final example that uses real data. The data are drawn from the 1983 primary election in the Fourteenth Los Angeles City Council district. These data have been used in two recent Section 2 cases: *Zaldivar v. City of Los Angeles* (1986) and *Garza v. County of Los Angeles* (1990).

The population of the city of Los Angeles was twenty-seven percent Latino in the 1980 census. Latinos held none of the fifteen city council seats and there did not exist a district that was considered to be a winnable Latino district. Instead, Latinos in Los Angeles were split among a number of districts. One of those districts, the fourteenth, had a sizable Latino population in its southern section and an affluent, conservative, and primarily white, population in its northern section. All Los Angeles city council elections are non-partisan and a primary election is held before the general election. If no candidate receives more than fifty percent of the primary vote, the top two vote recipients in the primary contest the general election, otherwise, the top vote recipient wins the seat.

In the 1983 primary, the incumbent city councilman, Art Snyder, came within three votes of being forced into a runoff against Latino candidate Steve Rodriguez. Snyder faced a recall campaign in 1984 and eventually resigned from office in 1985. Snyder's seat was taken in a special election by (then State Assemblyman) Richard Alatorre. The election of a Latino to this seat did not dissuade the Justice Department from filing a Section 2 lawsuit against the city. After a period of pre-trial maneuvering, the City Council retained one of the authors as a consultant. A database of voting returns in the Fourteenth City Council district was constructed. The author, at that time using the correlation coefficient, came to the conclusion that racially polarized voting was occurring in the Fourteenth City Council District.[14] He advised the City that this type of conclusion would almost certainly be reached by any reputable expert witness and would be another piece of evidence against the City under the "results test." The City, taking into account this evidence, other evidence, and the advice of an outside law firm, decided to settle. New districts were put in place in time for the primary election of 1987.[15]

Table 2 shows the estimates of individual behavior obtained by using the actual precinct level data from the Fourteenth City Council district and three different measures of racially polarized voting. The value labelled "actual" is the correlation coefficient for "percent Latino voters" and "percent of the vote for Snyder (Rodriguez)" we obtained using a standard statistical package, SPSSX. The value labelled "predicted" is the correlation coefficient obtained through the equations derived in the appendix. An estimate of the actual difference between p_1 and p_2 is given in the column labelled $p_1 - p_2$. The values of $p_1 - p_2$ were estimated using a technique that explicitly allows us to make individual level inferences from aggregate level data and is described in the next section.

Note that $p_1 - p_2$ is roughly half the size of the correlation coefficient. If the correlation coefficient was interpreted as the difference between p_1 and p_2, it would be assumed that there is a much higher degree of racially polarized voting than actually exists. Since the difference is so large, the question of which measure the court accepts becomes of great importance to all potential litigants. In this section, we have shown how and why the

Table 1. Data from the Fourteenth Los Angeles City Council District

P	R	S	PLATV	PNLATV	P	R	S	PLATV	PNLATV
3901	.17	.79	.12	.88	3950	.56	.36	.57	.43
3902	.22	.67	.10	.90	3951	.46	.45	.59	.41
3903	.19	.65	.09	.91	3952	.54	.34	.48	.52
3904	.15	.70	.09	.91	3953	.52	.42	.53	.47
3905	.16	.69	.05	.95	3954	.49	.47	.59	.41
3906	.19	.68	.08	.92	3955	.52	.41	.63	.37
3907	.27	.61	.10	.90	3956	.52	.43	.53	.47
3908	.17	.70	.08	.92	3957	.56	.40	.60	.40
3909	.17	.72	.07	.93	3958	.56	.41	.70	.30
3910	.28	.63	.08	.93	3959	.62	.32	.65	.35
3911	.26	.62	.13	.87	3960	.48	.44	.45	.55
3912	.20	.65	.09	.91	3961	.54	.40	.65	.35
3913	.17	.72	.10	.90	3962	.60	.36	.66	.34
3914	.17	.72	.12	.88	3963	.61	.35	.69	.31
3915	.31	.58	.14	.86	3964	.53	.44	.56	.44
3916	.23	.56	.14	.86	3965	.58	.38	.67	.33
3917	.23	.64	.20	.80	3966	.51	.39	.70	.30
3918	.24	.61	.09	.91	3967	.57	.39	.66	.34
3919	.21	.63	.14	.86	3968	.48	.45	.65	.35
3920	.23	.64	.18	.82	3969	.48	.47	.69	.31
3921	.30	.56	.07	.93	3970	.43	.49	.73	.27
3922	.36	.52	.13	.87	3971	.59	.34	.80	.20
3925	.30	.52	.23	.77	3972	.46	.46	.78	.22
3926	.43	.45	.21	.79	3973	.41	.51	.66	.34
3927	.31	.60	.13	.87	3974	.63	.31	.79	.21
3928	.38	.48	.20	.80	3975	.41	.48	.88	.12
3930	.40	.49	.29	.71	3976	.60	.35	.84	.16
3931	.34	.52	.28	.72	3977	.54	.33	.81	.19
3932	.33	.51	.25	.75	3978	.63	.26	.76	.24
3933	.32	.51	.21	.79	3979	.33	.56	.30	.70
3934	.26	.60	.15	.85	3980	.57	.34	.68	.32
3935	.17	.61	.11	.89	3981	.59	.28	.77	.23
3936	.26	.58	.22	.78	3982	.54	.32	.84	.16
3937	.39	.49	.24	.76	3983	.56	.35	.76	.24
3938	.34	.49	.21	.79	3984	.55	.37	.75	.25
3939	.37	.49	.19	.81	3985	.44	.44	.70	.30
3940	.37	.51	.26	.74	3986	.57	.35	.86	.14
3941	.42	.49	.43	.57	3987	.64	.29	.78	.22
3942	.50	.44	.56	.44	3988	.54	.39	.74	.26
3943	.49	.48	.74	.26	3989	.58	.31	.85	.15
3944	.58	.36	.68	.32	3990	.55	.36	.74	.26
3945	.53	.40	.68	.32	3991	.47	.47	.86	.14
3946	.55	.35	.72	.28	3992	.56	.36	.72	.28
3947	.55	.37	.42	.58	3993	.56	.38	.82	.18
3948	.58	.38	.44	.56	3994	.53	.39	.67	.33
3949	.56	.37	.68	.32	3995	.37	.51	.73	.27
					3996	.56	.34	.83	.17
					3998	.26	.58	.11	.89

KEY

P	Precinct Number
R	Percent of the vote for Rodriguez
S	Percent of the vote for Snyder
PLATV	Percent Latino Voters
PNLATV	Percent Non-Latino Voters

Table 2. Correlation Coefficient of Vote with Percent Latino

	Actual	Predicted	$p_1 - p_2$
Snyder	−.851	−.86	−.394
Rodriguez	.883	.92	.498

correlation coefficient provides unreliable estimates of individual behavior and should not be used. In the next section, we examine other measures of racially polarized voting and lay the foundation of a measure that both provides more reliable measures of individual behavior and is consistent with previous Supreme Court opinions about the measurement of racially polarized voting.

III. THEORETICALLY CORRECT MEASURES

Our recommendation that the correlation coefficient no longer be used as a measure of racially polarized voting may be dismissed as irrelevant for those familiar with the Supreme Court's decision in *Thornburg v. Gingles*, but to do so would be in error. That some experts have dismissed the correlation coefficient without fully understanding the *Thornburg* decision has only led to confusion in the legal community as to which measure of racially polarized voting will be accepted by the courts. We know that this confusion has led some to mistakenly believe that only bivariate methods, like the correlation coefficient or bivariate regression, can be used to measure racially polarized voting.[16] The purpose of this section is to show otherwise.

In *Thornburg*, a 1986 Supreme Court case on vote dilution, the justices upheld the use of a bivariate method as a proper way of measuring voting polarization, while not allowing a multiple linear regression model presented by the appellants (the U.S. and the State of North Carolina) (*Thornburg v. Gingles*, 1990: 2,752). A careful examination of this decision reveals that *it was the court's desire for an accurate estimate of individual level voting behavior and not issues of statistical methodology* that led the court to disallow the multivariate method while, simultaneously, allowing the bivariate method. Consider the principle upon which the multivariate method in *Thornburg* was rejected: "It is the difference between the choices made by blacks and whites, and not the reason for that difference that results in blacks having less opportunity than whites to elect their preferred representatives" (*Thornburg v. Gingles*, 1986: 2,773).

The combination of the *Thornburg* decision, which disallowed a multivariate method, the Engstrom and McDonald article, which argued against bivariate methods in multi-ethnic situations, and our result, which argues

against the correlation coefficient as a measure of polarization in all cases, leaves open the question: "What type of method should be used?" While at the current moment there is no definitive answer, we can shed some light on how to address important issues involved in the estimation of racially polarized voting. This exploration has resulted in our ability to produce a model that addresses the difficulties associated with this type of estimation in a relatively satisfying manner.

This model is called the *homogeneity model* and is originally due to Hawkes (1969). The model uses the "individual probability of voting" concept as did our analysis of the last section. It also uses a concept called *homogeneity*. The term homogeneity refers to the relationship among the choices made by individual members of a group. If all members of a group make similar choices, we say that the group is homogeneous. If all groups are homogeneous and all groups are partitions (*i.e.* every individual is a member of one and only one homogeneous group), then we can directly infer the behavior of individuals from the observed behavior of homogeneous groups (as derived in Appendices I and II).

If all groups were homogeneous, and we could easily identify homogeneous groups, making individual level inferences from aggregate level data would be relatively simple. Unfortunately, homogeneity is not something that can be easily observed. Suppose that an electorate is made up of members of only two groups, "blacks" and "whites." If we were to assume that "blacks" and "whites" were homogeneous groups, we could use some type of regression method (such as bivariate regression) to generate estimates of voting behavior for each group. If our assumption were correct, we could use these estimates to describe the behavior of black and white individuals. If this assumption were not correct, or if we had no way to verify whether or not our "homogeneity assumption" was correct, we would not be certain of the relationship between the behavior of the group and the behavior of individuals. It is clear that in order to have any faith in the reliability of our estimates as measures of individual behavior (racially polarized voting), we would want to have some idea whether or not our "groups" were homogeneous.

Unfortunately, all previous attempts to measure racially polarized voting implicitly assume that the racial group under consideration is a homogeneous group and that "everyone not in that racial group" is also a homogeneous group. In most cases, this implies that all white voters act similarly. In all cases, these assumptions are not questioned or tested. All known methods of estimating individual probabilities can produce substantively inaccurate estimates of racially polarized voting if the groups used in the estimation are not homogeneous.[17] We assert that unless homogeneity assumptions are tested and proven to hold, the relationship between the supposed measures of individual behavior and actual individual behavior will be unclear.[18] Therefore, the reliability of many previous estimates of voting polarization must be called into question.

Thus, we suggest a model that, like all others used before it, depends on the assumption of homogeneous groups in order to produce reliable estimates of individual level voting behavior. One unique advantage of our model is that it allows an explicit test of this statistically and substantively important assumption. The ability to distinguish homogeneous groups from non-homogeneous groups allows us to describe a model that uses aggregate voting data to produce accurate estimates of individual level voting behavior. We now present a simple example in order to show what can go wrong when we attempt to measure individual level voting behavior, using aggregated data, when the homogeneous groups we use in the estimation are, in fact, non-homogeneous.

Suppose we have an electorate that consists of members of only three groups, "blacks," "poor whites," and "non-poor whites." Now suppose that a black candidate contests a particular election. Let the probability of support for the black candidate by a black voter be .9, by a poor white voter be .2, and by a non-poor white voter by .6. One way to estimate the voting behavior of blacks and whites would be to look at individual level data, like an exit poll. If all three groups are homogeneous and all three groups are represented in the poll in numbers proportionate to their number in the population then we can obtain estimates of black and white voting behavior. The calculation of black voting behavior is straightforward: divide the number of blacks in the poll who voted for the black candidate by the total number of blacks. The calculation of white voting behavior depends on the relative number of poor whites and non-poor whites. Suppose that there were twice as many poor whites than non-poor whites. Then we could calculate white voting behavior by the following multi-step process:

1. Divide the number of poor whites in the poll voting for the black candidate by the total number of poor whites;
2. Divide the number of non-poor whites in the poll voting for the black candidate by the total number of non-poor whites; and
3. Multiply the percent support by non-poor whites by two, multiply the percent support by poor whites by one (their respective proportionate weights in the population), add this together, and divide by the sum of the weights (three).

Since the white groups are homogeneous this calculation gives the following result:

$$\frac{(.2\times 1)+(.6\times 2)}{3}=.466$$

46.6 percent of the white voters voted for the black candidate. Notice that the behavior of individuals in each of the homogeneous white groups "poor," (.2), and "non-poor," (.6), is not accurately described by the overall statistic, (.466).[19]

Now, suppose we were to use a bivariate regression technique in order to

estimate individual level voting behavior in the above case. The most frequent application of bivariate regression in Voting Rights Act (Section 2) cases has been to regress "percent vote for the black candidate" on "percent black" plus a constant. The usual estimate of racially polarized voting (the coefficient of "percent black") would depend on the assumption that the group "non-black" is homogeneous. Since "non-blacks" (in this case, "whites") are not a homogeneous group, the coefficients we produce are unlikely to be reliable measures of individual behavior.[20]

We show in Figure 3 exactly what this may mean in terms of the usual bivariate regression. In that figure, we set the support rate of blacks for the black candidate at ninety percent, set that of non-poor whites at sixty percent, as in the previous example. We then vary the support rate of poor whites between zero and one hundred percent. In Figure 3, we display the actual black support, the actual white support and the estimated black and white support as obtained from the method of bivariate regression.[21] Non-poor whites and blacks are assumed to be equally numerous and poor whites half the size of either group, as in the previous example. White support is calculated exactly as above, by using the weighted sum of the two white groups, poor and non-poor.

Certain things are evident from Figure 3. First, if poor whites do not support the candidate at all, we obtain estimates that imply that whites vote for the candidate at a higher rate than blacks even though whites support the candidate at a forty percent level (overall) and blacks support the

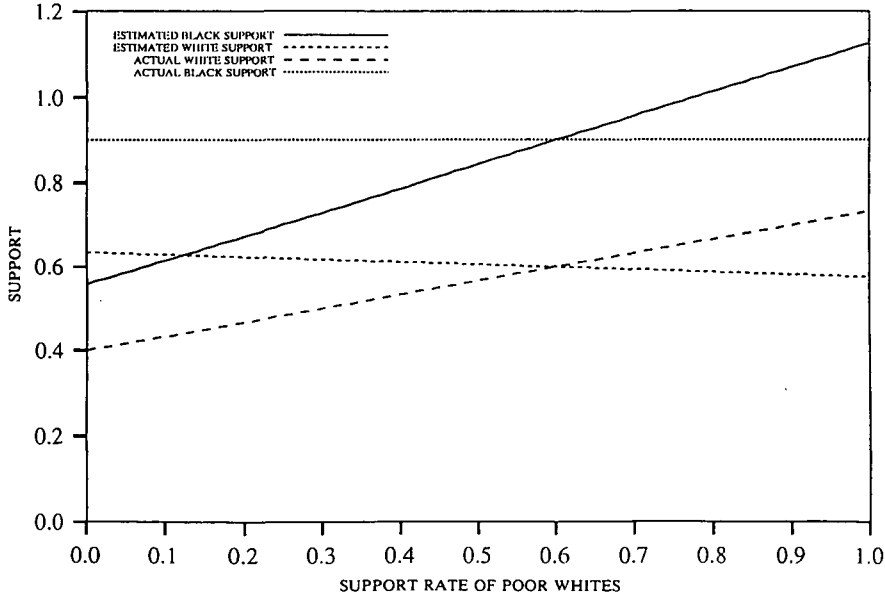


Figure 3 Estimated versus actual support rates

candidate at a ninety percent level! Black support is consistently underestimated until poor white support equals sixty percent, and then is overestimated. Similarly, white support is overestimated until poor white support reaches sixty percent, and then is underestimated. This sixty percent figure is no accident, because it is the support level of the non-poor whites. When both groups of whites have the same support level, they can be treated as one homogeneous group and bivariate regression produces accurate estimates.

This is not just a hypothetical occurrence. Freedman *et al.* (1990) present an example that shows Latino support for Jesse Jackson in the 1988 Presidential primary in San Joaquin county. Bivariate regression predicted that 109 percent of the Latinos had supported Jackson, while exit polls indicated the percent of Latinos that supported Jackson was thirty-five. Due to the residential interleavening of black and Latino voters, and the fact that "Latino" is used as the only regression variable, the erroneous result may very well come from the implicit grouping of whites and blacks together as a homogeneous group (considering the expected disparate response rate of these two groups to a Jackson candidacy).

We now provide a more general consideration of how the homogeneity model allows us to make individual level inferences from aggregate level data. Let us take the example of three groups. We define the number of individuals belonging to homogeneous group one as W_1, the number of individuals belonging to homogeneous group two as W_2, and the number of individuals belonging to homogeneous group three as W_3. Let the respective probabilities of supporting a candidate be p_1, p_2 and p_3 for the three groups. Then we define the expected vote for a particular candidate as:[22]

$$\text{Expected(Votes)} = W_1p_1 + W_2p_2 + W_3p_3.$$

Let us take the example given above of blacks, poor whites, and non-poor whites, with these corresponding to groups W_1, W_2 and W_3. Suppose that in a particular electoral unit there are one hundred blacks, one hundred poor whites, and two hundred non-poor whites. Since, on average, ninety percent of the blacks will vote for the black candidate, twenty percent of the poor whites will vote for a black candidate, and sixty percent of the non-poor whites will vote for the black candidate, we *expect* to observe, in any particular electoral unit, the candidate receiving ninety black votes, twenty poor white votes, and 120 non-poor white votes, for a total of 230 out of four hundred. In fact, we will not observe exactly 230 votes for the candidate every time we encounter this type of electoral unit, but the number of votes we observe should be "close" to 230. The concept of "closeness" in statistics is primarily the specification of an error term. The error term can be thought of as simply the difference between the expected vote and the actual vote observed in the electoral unit. The specification of an error term allows one to estimate a statistical model.[23]

We derive an error term in the appendix of the form of $u(W_1, W_2, W_3, p_1,$

p_2,p_3). This error term is derived from the usual social science assumptions on the voting behavior of individuals (see Appendix II). Thus, the actual vote in the electoral unit (which contains random error from the error term, unlike the expected vote) is

$$\text{Votes} = W_1p_1 + W_2p_2 + W_3p_3 + u(W_1, W_2, W_3, p_1, p_2, p_3).$$

We can use this equation to estimate the probability that an individual in any one of those groups votes for the candidate, from actual aggregate election returns, when the assumptions of the model are met.[24] When we estimate behavior using this model by a method of estimation known as maximum likelihood, a test of our "homogeneity" assumption is possible. That test is called the likelihood ratio test, and is well known to statisticians and statistically sophisticated social scientists (we describe this in Appendix II). We used this test to check whether the estimates of $p_1 - p_2$ given in Table 2 were reasonable, and determined that they were.[25]

The test of the acceptability of the model (*i.e.* the assumption of homogeneous groups) is based primarily on the difference between the expected vote and the actual vote. If the difference fits a theoretically predicted value, then we have evidence that our groups are in fact homogeneous. If the difference fails to fit the predicted value, this would provide evidence that individuals within our groups act differently. We would conclude from a large difference between "expected vote" and "actual vote" that our groups were not homogeneous. In that case, the model should be rejected as the coefficients produced will not be directly related to individual behavior.

How do we proceed when the test rejects the model? Since the model is "rejected" when the difference between the expected vote and the actual vote is "too large," one could always increase the allowable level of the difference by increasing the error term.[26] We do not believe this approach is useful because non-homogeneity in group responses, which biases the individual level estimates of voting, will also increase the variance. Thus, increasing the amount of difference that is acceptable would also allow one to make inaccurate estimates of the individual probabilities and not realize it.

The other course of action one may take is to look for other divisions of the population that might produce homogeneous groups. In the example that we have been using throughout this section, a homogeneity model estimated with the groups "blacks" and "whites" would be rejected because "whites" are not homogeneous. At that point, if one broke the "whites" into "poor whites" and "non-poor whites", one would rerun the model with those groups and find it accepted. The primary problem here is how one goes about finding these groups. McCue (1990) has devised a cluster-analytic methodology, but research on this problem is still in its infancy.

From this model, it is a simple matter to obtain a simple summary statistic that describes the voting behavior of any ethnic group which has been broken into more than one homogeneous group. It is accomplished in

exactly the same manner by which we obtained estimates for the voting behavior of "blacks" and "whites" in the hypothetical poll that was discussed near the beginning of this section. Simply multiply the number of individuals in each of the sub-groups (of the ethnic group in question) by the estimated level of support for the candidate in that sub-group, then divide the sum of the support of each of the subgroups by the total number of individuals in the ethnic group.

From the above argument, it is possible to see why the type of regression used by the appellants in the *Thornburg* case could not reproduce individual level estimates. Suppose one has two groups (for instance, blacks and whites) and feels that income is an important factor in the voting decision. One way to model this is to regress vote choice on blacks, whites and income. It is not possible to recover from this procedure, however, the "percent vote for a candidate" by blacks or whites. If one split blacks and whites into poor and non-poor, one could then combine the poor and non-poor blacks together in the manner described above (and the same for the whites), and obtain the overall level of support for the candidate from blacks and whites while explicitly accounting for the possible effect of income. Note that this method, while providing estimates of the effects of income, also addresses the Court's concerns about obtaining "the differences between the choices made by blacks and whites," and indeed, if different ethnic groups vote differently at different levels of income, this breakdown should be made.

Before we conclude, we show the relationship between individual probabilities, the homogeneity model, and other models that have been and are being used to measure racially polarized voting. This relationship is displayed, for the case with two groups, in Table 3. The relationship is shown for two types of aggregated data, *proportions* and *raw counts*. We present this relationship for both types of data to show that some methods of estimation are more dependent on the type of data used than others. Table 4 provides a key to the variables used to relate the coefficients to individual level behavior in Table 3 as well as a brief list of citations showing where the different methods have actually been utilized. We have based this table on the assumption that there are two homogeneous groups in the electorate. If there are not, then all of these methods will produce incorrect estimates, and only the homogeneity model will give evidence that the model is incorrect.[27]

IV. CONCLUSION

We have shown that the correlation coefficient may be biased in the case of two groups (one majority group and one minority group) because this statistic is a function of a factor that is unrelated to voting polarization. When the correlation coefficient is used to demonstrate the relationship between individual voting behavior and racial identity, the correlation

Table 3. The Relationship of Individual Behavior to Estimates Based on Aggregate Data

	Using Proportions	Using Raw Counts
Correlation	$\varrho = \dfrac{(p_1 - p_2)}{[(p_1 - p_2)^2 + a_1 p_1(1-p_1) + a_2 p_2(1-p_2)]^{1/2}}$	$\varrho = \dfrac{\mathrm{Var}(W_1)p_1 + \mathrm{Cov}(W_1, W_2)p_2}{\left[\begin{array}{c}[\mathrm{Var}(W_1)]^{1/2}[p_1^2\mathrm{Var}(W_2) + p_2^2\mathrm{Var}(W_2) + 2p_1p_2\mathrm{Cov}(W_1, W_2) + \\ E(W_1)p_1(1-p_1) + E(W_2)p_2(1-p_2)]\end{array}\right]}$
Bivariate Regression	$V = \alpha + X_1\beta$ $\alpha = p_2$ $\beta = p_1 - p_2$	$V = \alpha + W_1\beta$ $\alpha = \dfrac{S_1^2 E(W_1) - S_{12}E(W_2)}{\mathrm{Var}(W_1)}\, p_2$ $\beta = p_1 + \dfrac{\mathrm{Cov}(W_1, W_2)}{\mathrm{Var}(W_1)}\, p_2$
Maximum Likelihood	$V = X_1\tau_1 + X_2\tau_2$ $\tau_1 = p_1$ $\tau_2 = p_2$	$V = W_1\tau_1 + W_2\tau_2$ (Used in Homogeneity Model) $\tau_1 = p_1$ $\tau_2 = p_2$

Table 4. Key and Citations

ϱ	estimated correlation coefficient of group one and votes for the candidate
p_i	probability that member of group i votes for the candidate
a_1 and a_2	functions of group geographical distributions independent of voting behavior (see Appendix)
Var	Variance
Cov	Covariance
E	Expectation
α	constant term in bivariate regression
X_1	is proportion of electorate in group one (e.g., percent black)
W_1	number of electorate in group one (e.g., blacks)
β	estimated coefficient of group one in bivariate regression
X_2	is proportion of electorate in group two (e.g., percent white)
W_2	number of electorate in group two (e.g. whites)
S_1^2	second moment of $W_1 = \Sigma_i W_{1i}^2$.
S_{12}	$= \Sigma_i W_{1i} W_{2i}$
τ_i	estimated coefficient of group i from maximum likelihood

Correlation	*Kirksey v. Board of Supervisors* (1975) *Major v. Treen* (1983) *Jones v. City of Lubbock* (1984) *Terrazas v. Clements* (1984)
Bivariate Regression	*Butts v. City of New York* (1985) Engstrom and McDonald (1987) Grofman, Migalski and Noviello (1985) *Thornburg v. Gingles* (1986) *Latino Political Action Committee v. City of Boston* (1986) *Garza v. County of Los Angeles* (1990)
Maximum Likelihood	McCue and Lupia (1989)

coefficient is affected not only by the support of the ethnic group but also by the distribution of the groups within an electorate. In a broad range of circumstances, this statistic can give the wrong answer to the question; "Does racially polarized voting exist?" including cases involving only two groups.

Litigants should demand that the statistical methods used to provide evidence in Section 2 cases produce results that can be compared with an exit poll, whether or not an exit poll actually exists. It is clear from the examples given in section two that the correlation coefficient does not provide this type of result. That leaves the bivariate regression model and the homogeneity model. Bivariate (ecological) regression only provides this type of result if the groups it implicitly partitions the electorate into (black versus non-black, latino versus non-latino) are homogeneous. Unfortunately, while all of the ecological regression models depend on the use of homogeneous groups, none allows a test of this assumption.

Thus, in the case of two groups, we prefer the two-variable homogeneity model for the following reason – it is possible to test whether or not the groups we use as independent variables are, in fact, homogeneous. If the groups are homogeneous, our estimates will be reliable estimates of individual behavior. If the groups are not homogeneous, we will not have reliable estimates and we will know to redefine our "groups" in order to obtain reliable estimates of individual behavior.[28] It follows straightforwardly that, for the case of more than two groups, we prefer the N-variable homogeneity model for the same reason. We should emphasize that the search for statistically correct methods of measuring behavior is a field in which there is very active research.[29] Given the substantive implications of the use of regression analysis in the context of voting rights, in particular, and the social sciences, in general, we believe that it merits serious attention in future research.

Combining our result with that of Engstrom and McDonald, we can assert that the correlation coefficient has been shown to be deficient for voting rights litigation, and should no longer be used as a measure of voting polarization. We also assert that closer attention to the relationship between individual behavior and coefficients produced by bivariate and multivariate regression methods is necessary in order to produce a consistent and, presumably, acceptable method of retrieving individual level racially polarized voting measures from aggregated voting data.

ARTHUR LUPIA *is Assistant Professor of Political Science at the University of California, San Diego. He has developed both formal models and empirical instruments for the purpose of exploring the relationship between individual level voting behavior and collective decision making. While an employee of Pactech Data and Research, he studied voting behavior and the effect of alternative districting schemes in Los Angeles county.*

KENNETH MCCUE *and his database services firm, Pactech Data and Research, Inc., have provided reapportionment database services and analyses for the California State Assembly, the Los Angeles City Council, the United States Department of Justice, the city of Boston and the Los Angeles County Board of Supervisors. Dr. McCue holds a Ph.D. in Social Science from the California Institute of Technology and an M.A. in Mathematical Statistics from the University of Kansas. His current academic affiliation is Research Scientist, Environmental Quality Laboratory, California Institute of Technology.*

NOTES

1. Voting Rights Act Amendments (1982) Section 2b and Senate Report 97-417 (1982).
2. The results test was the customary judicial standard in a number of Voting Rights Act cases held before 1982 from *White v. Regester* (1973) to *Mobile v. Bolden* (1980).
3. Including: *Kirksey v. Board of Supervisors of Hinds County* (1976), *McMillan v. Escambia County, Florida* (1981), *Major v. Treen* (1983), *Lee County Branch v. City of Opelika* (1984), *Collins v. City of Norfolk* (1985), *Butts v. City of New York* (1985), *Thornburg v. Gingles* (1986), *McCord v. City of Fort Lauderdale* (1986), *Latino Political Action Committee Inc. v. City of Boston* (1986).
4. To cite a personal example, one of the authors was recently asked by a prominent legal concern to conduct a racially polarized voting analysis. The analysis would provide information as to whether a lawsuit under Section 2 of the Voting Rights Act was advisable. Among the instructions was a request that only estimates obtained through the method of bivariate correlation be submitted. The decision handed down in *Thornburg v. Gingles* (discussed later) was cited as the reason not to use multivariate methods. For reasons that are spelled out in this paper, estimates from another method were submitted as well.
5. The correlation coefficient has a number of rather interesting mathematical interpretations, none of which apply to the measurement of individual level behavior using aggregated data. (See Rodgers and Nicewander, 1988: 59–66).
6. In the estimation of racially polarized voting, the value of N within a given electorate will be determined by the available data. In general, we prefer N to be as high as possible, for estimations of reality by statistical techniques are presumed to be better as the number of observations grow, subject to some technical caveats.
7. It is the usual assumption in social science modelling of choice behavior (see McFadden, 1973). Such an assumption allows us to use probit and logit analysis to interpret individual level choice behavior. A more technical explanation is given in Appendix II.
8. This probabilistic concept of voting is implicit in the Voting Rights Act. That is, the Act seeks to protect groups of individuals who not only share certain racial or ethnic characteristics but also display a certain degree of political cohesiveness. Political cohesiveness is not explicitly defined in the Act, but it clearly does not require absolute (one hundred percent) cohesion.
9. It is also strictly increasing with respect to p_1, that is, if every other term is held constant and p_1 is increased, then $\mathrm{Corr}(X, V)$ increases (this can be shown by differentiating the expression and signing the derivative).
10. In the case where there exists a third group, the correlation coefficient can be an even worse indicator as is indicated in Engstrom and McDonald (1987) and as shown in a broader sense in McCue and Lupia (1989).
11. Coefficients of .5 and higher were deemed "statistically significant." We can only

assume that there was some confusion among the participants in these cases over the difference between "statistical" and "substantive" significance. Statistical significance is dependent on the number of observations in the analysis, a dependency with which their interpretation does not seem to be consistent. In *Major* the number of observations ranged from thirty-one to 428, with the average being about 320. This is not consistent with the true notion of statistical significance, which would require only sixteen observations to give statistical significance (at the .05 level) to a correlation coefficient of .5. If they meant staistical significance, then their analysis is nonsensical. If we assume that they, in fact, meant "substantive significance", our examples clearly show that the use of a .5 standard could lead to incorrect inferences about the existence of racially polarized voting. (From our derivation it is clear that examples for other standards will give similar results.)

12. See Brace, Grofman, Handley and Niemi, 1988: 43–62. It is our opinion, that such a debate, while inherently political on some levels, should also be based on sound and well-tested statistical principles (*i.e.* how does the measure relate to individual behavior, and what does this imply about the behavior of electorates under such a standard), as well as the legal interpretation of what qualities the standard should possess.
13. One additional unpleasant feature of the correlation coefficient is in cases where there are three or more candidates. Let us say candidate one is black, candidate two is white, and candidate three is also white. Suppose that candidate one received forty-nine percent of the black vote and twenty percent of the white vote (we are assuming these are the only two groups). Then the correlation coefficient between the proportion of votes for the black candidate and percent black would be large and positive, even though a majority of the black voters voted for one of the white candidates. The reason is, of course, that the correlation coefficient is proportional to the difference in voting behavior by the two groups, and there is indeed a difference in preference for the black candidate by those two groups. Depending upon how one defines racially polarized voting, this may or may not be it, but it definitely hides the fact that a majority of the blacks voted for a white candidate.
14. Since we have devoted a considerable effort to explaining why the correlation coefficient should not be used, it is somewhat embarrassing to admit to having used it as a consultant. However, it was the accepted legal standard at that time.
15. At that time another Latino, Gloria Molina, was elected to a second Latino district. Molina is now a member of the Los Angeles County Board of Supervisors, having run and won in the Latino seat drawn as a result of the *Garza* settlement, and is perhaps the most visible Latino (and certainly Latina) politician in the country, which shows that Section 2 cases do have important consequences.
16. For example, Freedman *et al.* (1990) state that "A final legal difficulty must be mentioned: in *Gingles*, the Supreme Court ruled against multiple regression." This is clearly an incorrect summary of what the Supreme Court said, as we shall show.
17. This is due to the fact that all known methods either use least squares, or, if maximum likelihood, the "least squares" term dominates (see Brown and Payne, 1986: 452–460). This problem can be considered an example of the "ecological fallacy," but it is important to note that all analyses with aggregate data are not necessarily fallacious. We agree with Hanushek, Jackson, and Kain (1974) that correct specification of an individual level estimation and the manner in which it is transferred into aggregate estimation is the key to whether the fallacy obtains or does not.
18. Our analysis of the correlation coefficient, in the last section, assumed the

presence of homogeneous partitions. Without that assumption, the correlation coefficient provides even less accurate estimates than we generated.

19. In a poll, of course, if we have accurately sampled the population, we can obtain one overall level of support without breaking down into homogeneous groups. One can obtain the same overall level of support by this type of breakdown, though, as we have shown. Notice that if certain groups are oversampled (either by design or inadvertently), this type of adjustment must be made to obtain an accurate indicator of overall support.
20. This is shown in the Appendix and falls under the econometric subject of model misspecification. It has been well-known for decades and is discussed in elementary textbooks like Maddala (1977).
21. Exact details of our sample electorate are given at the end of Appendix II. Basically, blacks are assumed to be highly segregated from non-poor whites and somewhat segregated from poor whites.
22. In actual analyses of racially polarized voting, we obtain V from aggregated voting returns and obtain W_1, W_2 and W_3 from a list of voters. Given a list of voters, we obtain the groups by means of a surname match, or in conjunction with racial, ethnic, demographics or other group information from the most recent census.
23. Most non-statisticians are unaware that statistical models are usually very simple matters if the error terms are left off, typically a linear equation or a relatively simple function. The pages of equations that fill statistical journals are primarily devoted to derivation of distributions of parameters of the models once the error term has been specified and a method of estimation has been selected.
24. Hawkes (1969) shows that this particular model can also be estimated by iterative weighted least squares.
25. Further details of testing on this sort of problem are given in McCue and Lupia (1989).
26. In technical terms, the error term from the homogeneity model (which is binomial) can be replaced with other error terms, such as those derived from the beta-binominal distribution (King, 1989) or an aggregated compound multinomial distribution (Brown and Payne, 1986).
27. We have included raw counts because, to our knowledge, there has been no theoretical justification as to why proportions are used rather than counts.
28. This is highly important in statistical estimation since it is well-known that the ecological regression model often produces impossible estimates. For example, to quote Shivley (1969): "The lack of interest in ecological regression is probably due to the fact that errors in estimation are likely to turn up either as negative percentages or as percentages which are greater than one hundred. This is disheartening to the researcher, and is difficult to present to his colleagues."
29. See Loewen and Grofman (1989: 589–604) for an attempt to establish some conditions for consistency of bivariate regression and Klein, Sacks and Freedman (1990) for a rebuttal.
30. Notice that in the text of the paper X would be X_1 and $1-X$ would be X_2.
31. This was brought to our attention by a reviewer.
32. By the law of iterated expectations, $\mathrm{Cov}(W_1,u)=\mathrm{Cov}(W_2,u)=0$.
33. This level of aggregation is not uncommon in the study of polarized voting.
34. The extension to more than two groups is straightforward and is omitted.
35. The multivariate normal is not essential – any distribution would do, but this approach makes it simple to present the misspecification formula above.
36. This has been noted by Achen (1983).

REFERENCES

ACHEN, CHRISTOPHER (1983) "If Party ID Influences the Vote, Goodman's Ecological Regression is Biased (But Factor Analysis is Consistent)." Paper presented at the American Political Science Association meeting, September 1–4, 1983, at Chicago, Illinois.

BRACE, KIMBALL, GROFMAN, BERNARD N., HANDLEY, LISA R. and RICHARD G. NIEMI (1988) "Minority Voting Equality: the 65 Percent Rule in Theory and Practice," *Law and Policy* 10: 43–62.

BROWN, PHILLIP J. and CLIVE D. PAYNE (1986) "Aggregate Data, Ecological Regression, and Voting Transitions," *Journal of the American Statistical Association* 81: 452–460.

ENGSTROM, RICHARD L. and MICHAEL D. MCDONALD (1987) "Quantitative Evidence in Vote Dilution Litigation, Part II: Minority Coalitions and Multivariate Analysis," *Urban Lawyer* 19: 65–75.

FELLER, WILLIAM (1950) *An Introduction to Probability Theory and its Applications: Volume I.* New York: Wiley and Sons.

FREEDMAN, D.A., KLEIN, S., SACKS, J., SMYTH, C. and C. EVERETT (1990) "Ecological Regression and Voting Rights." Technical Report No. 248, Department of Statistics, University of California, Berkeley, California.

GROFMAN, BERNARD N., MIGALSKI, MICHAEL 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 and Policy* 7: 199–223.

HANUSHEK, ERIC A., JACKSON, JOHN E. and JOHN F. KAIN (1974) "Model Specification, Use of Aggregate Data, and the Ecological Correlation Fallacy," *Political Methodology* 1: 89–107.

HAWKES, A.G. (1969) "An Approach to the Analysis of Electoral Swing," *Journal of the Royal Statistical Society*, Series A, 132: 68–79.

KING, G. (1989) "On Political Methodology." Paper presented at the American Political Science Association meeting, August 30–September 3, 1989, at Atlanta, Georgia.

KLEIN, S.P., SACKS, J. and D.A. FREEDMAN "Ecological Regression Versus the Secret Ballot." Technical Report No. 276, Department of Statistics, University of California, Berkeley, California.

LOEWEN, JAMES W. (1982) *Social Science in the Courtroom: Statistical Techniques and Research Methods for Winning Class Action Suits.* Lexington, Massachusetts: Lexington Books.

LOEWEN, JAMES W. and BERNARD N. GROFMAN (1989) "Recent Developments in Methods Used in Vote Dilution Litigation," *Urban Lawyer* 21: 589–604.

MADDALA, G.S. (1977) *Econometrics.* New York: McGraw-Hill.

MCCUE, KENNETH (1990) "The Inference of Individual Probabilities from Aggregate Data – a Homogeneous Approach." Paper presented at the American Political Science Association meeting, August 30–September 2, 1990, at San Francisco, California.

MCCUE, KENNETH and ARTHUR LUPIA (1989) "An Alternative Statistical Measure for Racially Polarized Voting," Social Science Working Paper 690, California Institute of Technology, Pasadena, California.

MCFADDEN, DANIEL (1973) "Conditional Logit Analysis of Qualitative Choice Behavior," in Paul Zarembka (ed.) *Frontiers in Econometrics.* New York: Academic Press.

RODGERS, JOSEPH LEE and W.A. IAN NICEWANDER (1988) "Thirteen Ways to Interpret the Correlation Coefficient," *The American Statistician* 42: 59–66.

SHIVELY, W. PHILLIPS (1969) "Ecological Inference: The Use of Aggregate Data to Study Individuals," *American Political Science Review* 63: 1183–1196.
UNITED STATES CONGRESS. Senate. Committee on the Judiciary, *Voting Rights Act Extension*, 97th Cong. 2d sess., 1982 S.Rept.417 (Y1. 1/5:97-417).

CASES

Butts v. City of New York (1985) 779 F. 2d 141
City of Mobile, Alabama v. Bolden (1980) 446 U.S. 55
Collins v. City of Norfolk, Virginia (1985) 768 F. 2d 572
Garza v. County of Los Angeles (1990) 918 F. 2d 763
Jones v. City of Lubbock (1984) 730 F. 2d 233
Kirksey v. Board of Supervisors of Hinds County (1976) 528 F. 2d 536
Latino Political Action Committee, Inc. v. City of Boston (1986) 784 F. 2d 409
Lee County Branch v. City of Opelika (1984) 748 F. 2d 1473
Major v. Treen (1983) 574 F. Supp. 325
McCord v. City of Fort Lauderdale, Florida (1986) 787 F. 2d 1528
McMillan v. Escambia County, Florida (1981) 638 F. 2d 1239
Terrazas v. Clements (1984) 581 F. Supp. 1329
Thornburg v. Gingles (1986) 106 S. Ct. 2752
White v. Regester (1973) 412 U.S. 755
Zaldivar v. City of Los Angeles (1986) 780 F. 2d 823

APPENDIX I

DERIVATION OF THE CORRELATION COEFFICIENT

The expected proportion of the vote for a candidate in an electorate with two groups is

$$E(V) = X_1 p_1 + X_2 p_2,$$

where X_i is the proportion of the population that is in group i, and p_i can be thought of as the probability that a member of group i votes for the candidate.[30] Thus $X_1 p_1$ is the expected proportion of the vote that will go to the candidate from group one. (We use proportions where possible for ease of exposition.) We make the simple assumption that the groups are mutually exclusive and collectively exhaustive, that is, each individual in the electorate is a member of one and only one group. We can estimate the value of p_1 and p_2 by an equation such as:

$$V = X_1 p_1 + X_2 p_2 + u$$

Where u is an error term that has an expected value of zero, is conditional on X_1 and X_2, and is assumed to be uncorrelated with X_1 and X_2.

To estimate the effect of the variable X_1 on V, vote dilution litigants have commonly used the correlation coefficient. The correlation coefficient for this circumstance is defined as:

$$\text{Corr}(X_1, V) = \frac{\text{Cov}(X_1, V)}{[\text{Var}(X_1)\text{Var}(V)]^{1/2}} \tag{3}$$

(Note that this argument can easily be generalized to N groups. We concentrate on only two groups to simplify the exposition.)

We begin our derivation by taking a closer look at the components of Equation 1.

Since X_1 and X_2 are proportions and make up the entire electorate, we know that $X_1 + X_2 = 1$.

From the definition of covariance, we have:

$$\begin{aligned}\text{Cov}(X_1, X_2) &= \text{Cov}(X_1, 1 - X_1) \\ &= -\text{Var}(X_1),\end{aligned}$$

and also

$$\begin{aligned}\text{Cov}(X_1, X_2) &= \text{Cov}(1 - X_2, X_2) \\ &= -\text{Var}(X_2),\end{aligned}$$

which are relations that we use when deriving $\text{Var}(V)$. Note that the variances of X_1 and X_2 are equal. Thus,

$$\text{Var(V)} = \text{Var}(X_1 p_1 + X_2 p_2 + u)$$

$$= p_1^2 \text{ Var}(X_1) + p_2^2 \text{ Var}(X_2) + \text{Cov}(X_1, X_2) p_1 p_2 + \text{Var(u)}$$
$$= p_1^2 \text{ Var}(X_1) + p_2^2 \text{ Var}(X_1) + (-2\text{Var}(X_1) p_1 p_2) + \text{Var(u)}$$
$$= \text{Var}(X_1)[p_1 - p_2]^2 + \text{Var(u)}.$$

Additionally,

$$\begin{aligned}\text{Cov}(X_1, V) &= \text{Cov}(X_1, X_1 p_1 + X_2 p_2 + u) \\ &= \text{Cov}(X_1, X_1) p_1 + \text{Cov}(X_1, X_2) p_2 \\ &= \text{Var}(X_1) p_1 + \text{Cov}(X_1, X_2) p_2 \\ &= \text{Var}(X_1) p_1 - \text{Var}(X_1) p_2 \\ &= \text{Var}(X_1)(p_1 - p_2)\end{aligned}$$

Substituting the three component parts derived above, we can obtain the value of the correlation coefficient:

$$\text{Corr}(X_1, V) = \frac{\text{Var}(X_1)[p_1 - p_2]}{\text{Var}(X_1)^{1/2}[\text{Var}(X_1)[p_1 - p_2]^2 + \text{Var(u)}]^{1/2}}$$

$$= \frac{\text{Var}(X_1)[p_1 - p_2]}{\text{Var}(X_1)[(p_1 - p_2)^2 + \dfrac{\text{Var(u)}}{\text{Var}(X_1)}]^{1/2}}$$

$$\text{Corr}(X_1, V) = \frac{[p_1 - p_2]}{[(p_1 - p_2)^2 + \dfrac{\text{Var(u)}}{\text{Var}(X_1)}]^{1/2}} \tag{4}$$

Since both $\text{Var}(X_1)$ and $\text{Var}(u)$ are positive, and p_1 and p_2 are always between zero and one, this expression will always be between -1 and 1, as we know to be the case with the Pearson correlation coefficient. This formula is always valid for any variables V and X_1[31] and this is the expression we use to obtain the predicted correlation coefficients in Table 1.

To obtain a more precise understanding of this expression for $\text{Corr}(X_1, V)$, we need to have some idea of the magnitude of $\text{Var}(u)$. In a linear regression analysis, it is usually assumed that u is normally distributed with mean zero and some unknown variance. To obtain an estimate of that unknown variance, we make two simple assumptions about the behavior of individual voters. In addition, we must utilize the raw counts (as opposed to proportions) as this allows us to make plausible normality assumptions on the error term. A more detailed explanation of these assumptions is provided in McCue and Lupia (1989).

We have previously assumed that the two groups made up the entire electorate and that every individual in the electorate is a member of one and only one group. To this we add the assumptions:

1. Within a group, each individual has an equal probability of voting for a candidate (and)

2. Voter decisions are independent. (Such an assumption is made implicitly in any correlation and regression analysis relevant to this case.)

When these assumptions hold, we can view the voting behavior of each group as a number of Bernoulli trials, so that the aggregate distribution will be distributed binomially. As there are two groups, express

$$T = K_1 + K_2$$

where T is the number of votes (raw count) for the candidate, K_1 is the number of votes (raw count) from group one and K_2 is the number of votes from group two. By assuming independence of voter decisions and by letting W_1 be the number of individuals in group E_1 (so that $X_1 = \frac{W_1}{W_1 + W_2}$), we know that K_1 is distributed binomially, with mean $W_1 p_1$ and variance $W_1^{1/2} p_1 (1 - p_1)$. It can be shown as an elementary exercise in probability theory (see Feller, 1950) that we can approximate K_1 with a constant $W_1 p_1$ plus a constant $W_1^{1/2}$ times a normal variate ϵ_1, which has variance $p_1(1 - p_1)$. The same type of approximation is made for K_2. This gives

$$K_1 = W_1 p_1 + W_1^{1/2} \epsilon_1$$
$$K_2 = W_2 p_2 + W_2^{1/2} \epsilon_2$$

Note that by this construction, the normal variates are independent of one another and of the W's.

Given the satisfaction of the two assumptions, the equation representing the vote for a candidate is (substituting in our approximations for K_1 and K_2):

$$T = W_1 p_1 + W_2 p_2 + W_1^{1/2} \epsilon_1 + W_2^{1/2} \epsilon_2,$$

$E(\epsilon_1 | W) = E(\epsilon_2 | W) = 0$ (where $W = (W_1, W_2)$) and ϵ_1 and ϵ_2 are independent of each other, W_1 and W_2.[32]

From this and the relation $V = \frac{T}{W_1 + W_2}$, we have

$$V = X_1 p_1 + X_2 p_2 + \frac{W_1^{1/2}}{W_1 + W_2} \epsilon_1 + \frac{W_2^{1/2}}{W_1 + W_2} \epsilon_2.$$

Now, we can solve for an expression of Var(u):

$$\begin{aligned} \text{Var}(u) &= \text{Var}\left(\frac{W_1^{1/2}}{W_1 + W_2} \epsilon_1 + \frac{W_2^{1/2}}{W_1 + W_2} \epsilon_2 \right) \\ &= \text{Var}\left(\frac{W_1^{1/2}}{W_1 + W_2} \epsilon_1 \right) + \text{Var}\left(\frac{W_2^{1/2}}{W_1 + W_2} \epsilon_2 \right) \end{aligned}$$

Using this expression for the covariance of $W_1^{1/2} \epsilon_1$ and $W_2^{1/2} \epsilon_2$:

$$\begin{aligned} \text{Cov}(W_1^{1/2} \epsilon_1, W_2^{1/2} \epsilon_2) &= E(W_1^{1/2} W_2^{1/2} \epsilon_1 \epsilon_2) - E(W_1^{1/2} \epsilon_1) E(W_2^{1/2} \epsilon_2) \\ &= E\left(W_1^{1/2} W_2^{1/2} E(\epsilon_1 \epsilon_2 | W) \right) - E\left(W_1^{1/2} E(\epsilon_1) \right) \\ &\quad E\left(W_2^{1/2} E(\epsilon_2) \right) \end{aligned}$$

$$\mathrm{Cov}(W_1^{1/2}\epsilon_1, W_2^{1/2}\epsilon_2) = 0.$$

By the two assumptions,

$$p_1(1-p_1) = \mathrm{Var}(\epsilon_1 | W) = \mathrm{E}(\epsilon_1^2 | W)$$
$$p_2(1-p_2) = \mathrm{Var}(\epsilon_2 | W) = \mathrm{E}(\epsilon_2^2 | W)$$

From the definition of variance, we have,

$$\mathrm{Var}\left(\frac{W_1^{1/2}}{W_1+W_2}\epsilon_1\right) = E\left(\frac{W_1}{(W_1+W_2)^2}\epsilon_1^2\right) - \left[E\left(\frac{W_1^{1/2}}{W_1+W_2}\epsilon_1\right)\right]^2$$
$$= E\left[\left(\frac{W_1}{(W_1+W_2)^2}\right) E(\epsilon_1^2 | W)\right]$$
$$- \left[E\left[\left(\frac{W_1^{1/2}}{W_1+W_2}\right) E(\epsilon_1 | W)\right]\right]^2.$$

This gives us

$$\mathrm{Var}\left(\frac{W_1^{1/2}}{W_1+W_2}\epsilon_1\right) = E\left(\frac{W_1}{(W_1+W_2)^2}\right) p_1(1-p_1) \text{ and}$$
$$\mathrm{Var}\left(\frac{W_2^{1/2}}{W_1+W_2}\epsilon_2\right) = E\left(\frac{W_2}{(W_1+W_2)^2}\right) p_2(1-p_2).$$

Then,

$$\mathrm{Var}(u) = E\left(\frac{W_1}{(W_1+W_2)^2}\right) p_1(1-p_1) + E\left(\frac{W_2}{(W_1+W_2)^2}\right) p_2(1-p_2)$$

Recalling our expression of the correlation coefficient from Equation 2, we have

$$\mathrm{Corr}(X_1, V) = \frac{[p_1-p_2]}{[(p_1-p_2)^2 + \dfrac{\mathrm{Var}(u)}{\mathrm{Var}(X_1)}]^{1/2}}.$$

Since,

$$\frac{\mathrm{Var}(u)}{\mathrm{Var}(X_1)} = a_1 p_1(1-p_1) + a_2 p_2(1-p_2),$$

where

$$\frac{E\left(\dfrac{W_1}{(W_1+W_2)^2}\right)}{\mathrm{Var}(X_1)} = a_1 \text{ and } \frac{E\left(\dfrac{W_2}{(W_1+W_2)^2}\right)}{\mathrm{Var}(X_2)} = a_2,$$

we have,

$$\mathrm{Corr}(X_1, V) = \frac{(p_1-p_2)}{[(p_1-p_2)^2 + a_1 p_1(1-p_1) + a_2 p_2(1-p_2)]^{1/2}}. \qquad (5)$$

The two assumptions serve to express the correlation coefficient as an exact function of the underlying probabilities and the expectations of functions of the distributions of the groups. When no such assumption is explicitly made, the divergence of this (or any) measure from the true underlying values can become even greater. Such assumptions have been made implicitly in all estimation carried out with the correlation coefficient or multiple regression models as so far used in voting rights litigation.

We are interested in obtaining some understanding of how the value of the a_i's affect the correlation coefficient – in particular, how it relates to $p_1 - p_2$. From the above equation, we can see that $\text{Corr}(X_1, V) < p_1 - p_2$ if and only if

$$1 < (p_1 - p_2)^2 + a_1 p_1 (1 - p_1) + a_2 p_2 (1 - p_2) \tag{6}$$

or

$$1 - (p_1 - p_2)^2 < a_1 p_1 (1 - p_1) + a_2 p_2 (1 - p_2) \tag{7}$$

Now, if we examine the value of a_i,

$$\frac{E\left(\frac{W_i}{(W_1 + W_2)^2}\right)}{\text{Var}(X_i)} = a_i \tag{8}$$

we can see that since $\text{Var}(X_i)$ stays constant as the electoral unit size grows (X_i is a proportion), the numerator goes to zero as the electoral unit size grows. The actual value depends upon the distribution, but for practically any conceivable electorate, two hundred voters in an electoral unit would be enough to state the magnitude of the correlation coefficient is considerably larger than the magnitude of the difference $p_1 - p_2$.

APPENDIX II

INDIVIDUAL LEVEL ESTIMATION WITH AGGREGATED DATA

In this section, we show how a model that estimates probabilities from aggregate data can be derived from a model of individual decision making that is commonly used by statistically sophisticated social scientists. Let us first consider an estimation of individual level behavior when individual level data is available (*i.e.* when we can observe each voter's vote).

Let y^* be a scalar that represents the degree to which a voter prefers one candidate over another. Let X be a vector of individual characteristics and β be a conformable vector of weights (X is known. β is unknown and to be estimated). The particular weighting scheme that characterizes β allows us to describe individual level behavior as it shows which factors influence voter decisions. To estimate the voting behavior of individuals, we must develop a model that allows us to use observations of y^* and X to estimate

β. One model of individual choice we could propose (known to economists as a "random utility model" following McFadden (1974), is as follows:

$$y^* = X\beta + u,$$

where u is an error term that represents the uncertainty of β measured on y^*. In such a model, we assume that u has zero mean and finite variance.

While this model is commonly used to describe individual choice in some situations, it cannot be directly applied to a study of voting behavior since we do not observe y^*. That is, we can observe which candidate the voter voted for but we cannot observe the extent to which the voter prefers one candidate to another. Given this restriction, we must use a discrete choice model to estimate β.

$$\text{Let } y = \begin{cases} 0 \text{ if } y^* < \alpha: & Pr(y=0) = Pr(y^* < \alpha) = 1 - F(X\beta) \\ 1 \text{ if } y^* \geq \alpha: & Pr(y=1) = Pr(y^* \geq \alpha) = F(X\beta) \end{cases}$$

Using this type of model, we can proceed with the estimation of β under several different assumptions. Which method we choose depends on which assumptions we wish to make about the distribution of μ. If we assume that μ is normally distributed, then we use the method of maximum likelihood called *probit*. If we assume that μ is logistically distributed, then we use the method of maximum likelihood called *logit*. Either process involves maximizing the following type of likelihood function (of y and X) with respect to β.

$$\Pi_{y_i=1} F(X_i\beta) \Pi_{y_i=0} 1 - F(X_i\beta)$$

Now, we can begin to consider individual level estimation using aggregated data. We first consider a relatively simple aggregation where each observation gives voting data for two individuals. With this type of data, we can observe three possible ballot permutations. We will either observe candidate i receiving no votes, one vote, or two votes. The probability of each of these events is as follows:

$$\begin{aligned} Pr(V_i=0) &= [(1-F(X_{i1}\beta))(1-F(X_{i2}\beta))] \\ Pr(V_i=1) &= [(F(X_{i1}\beta))(1-F(X_{i2}\beta))] + [(1-F(X_{i1}\beta))(F(X_{i2}\beta))] \\ Pr(V_i=2) &= [(F)X_{i1}\beta))(F(X_{i2}\beta))]. \end{aligned}$$

Then, to find β we maximize the following equation:

$$\Pi_{V_1=0}[(1-F(X_{i1}\beta))(1-F(X_{i2}\beta))] \times \Pi_{V_1=2}[(F(X_{i1}\beta))(F(X_{i2}\beta))] \times \Pi_{V_1=1}[[(F(X_{i2}\beta))(1-F(X_{i2}\beta))] + [(1-F(X_{i1}\beta))(F(X_{i2}\beta))]].$$

Notice that this equation is much more complex than its counterpart for the individual level example. The likelihood function grows even more complex as the level of aggregation grows. Voting data is usually only available for much higher aggregations than we have spoken about. Consider an aggregation of five hundred individuals per observation.[33] As

we expand the number of individuals in an electoral unit the number of permutations explodes. Consider the case where we observe that the candidate receives 250 votes in an electoral unit of five hundred. In this case, alone, there are approximately 10^{300} permutations. When we consider that there are five hundred other possible observations (of the number of votes for candidate i), then it becomes obvious that the likelihood function will be very hard to estimate.

In order to find β in this type of situation, we must make some type of restriction. In general, we restrict the number of different types of voters that there can be, and for purposes of this exposition, we will assume there are only two types of voters (one may think of them as black and white, for example), each with a different probability of voting for a candidate. We also assume that all voters are independent of one another, which allows us to typify each voter as a Bernoulli trial. Thus if there are R_1 individuals in group one, the number of votes cast for the candidate is distributed binomially with mean $R_1 p_1$ and variance $R_1 p_1(1-p_1)$. This binomial variable, in turn, can be excellently approximated by a normal variable with the same mean and variance (Feller, 1950), or

$$R_1 p_1 + R_1 \epsilon_1,$$

where ϵ_1 is distributed normally with mean zero and variance $p_1(1-p_1)$.

Since there are only two types of voters, the number of votes obtained by the candidate is the sum of the two types of voters, which is

$$V = R_1 p_1 + R_2 p_2 + R_1 \epsilon_1 + R_2 \epsilon_2,$$

which is the equation derived in Appendix I. This is the equation used in the homogeneity model. The parameters of interest, p_1 and p_2, can be efficiently and consistently estimated by maximum likelihood (Hawkes shows that iterative least squares can be used), and consistently estimated by least squares, a fact we will use below for misspecification analysis.[34]

A simple test of the adequacy of this model is to maximize it with the constraint that $\mathrm{Var}(\epsilon_i) = p_i(1-p_i)$ and without that constraint. Now, minus twice the log of the ratio of the constrained to the unconstrained will be asymptotically distributed chi-squared with the number of degrees of freedom equal to the number of groups, and this is our likelihood ratio test (other asymptotically equivalent tests, such as the Wald or Lagrange multiplier, may also be used).

Now suppose there are three groups but that instead of estimating

$$V = R_1 p_1 + R_2 p_2 + R_3 p_3 + R_1 \epsilon_1 + R_2 \epsilon_2 + R_3 \epsilon_3,$$

we estimate

$$V = S_1 q_1 + S_2 q_2 + S_1 \theta_1 + S_2 \theta_2,$$

Where $S_1 = R_1$ and $S_2 = R_2 + R_3$ (in other words, we have collapsed groups two and three together and estimated them as one group). In order to

determine the effects of misspecification, we must derive the relation of the estimated probabilities $\hat{q}_i$ to p_i. To do this, we use the method of least squares, which is consistent for this problem under the homogeneity obtaining.

It is shown in elementary textbooks (*e.g.* Maddala, 1977) that

$$E(\hat{q}) = [S'S]^{-1}[S'R]p,$$

where $\hat{q}$ and p are the vectors consisting of $(\hat{q}_1, \hat{q}_2)$ and (p_1, p_2, p_3), respectively. This then gives the relationship of $\hat{q}$ as a function of p. To construct the example in Section III of this paper, we need to construct an electorate. Let t be distributed as a trivariate normal with mean μ and covariance matrix I. Then, we let r be generated by At, where A is an arbitrary three by three matrix. Assume R (and hence S) are realizations of r. Thus r is multivariate normal with mean $A\mu$ and covariance matrix AA'.[35]

Let C be the two by three matrix which collapses groups two and three together, that is

$$\begin{bmatrix} 1.0 & 0.0 & 0.0 \\ 0.0 & 1.0 & 1.0 \end{bmatrix}$$

Then $s = Cr$, and by standard matrix algebra, we have $\text{Cov}(s,r) = C\text{Var}(r,r) = CAA'$, $\text{Var}(s) = \text{Var}(Cr) = CAA'C'$, and $\text{E}(s) = CA\mu$. Then

$$\begin{aligned} \text{plim}E(\hat{q}) &= \text{plim}[\frac{S'S}{n}]^{-1}\frac{[S'\text{R}]}{n}p \\ &= E(s's)E(s'r)p \\ &= [\text{Var}(s) + \text{E}(s)(\text{E}(s))']^{-1}[\text{Cov}(s,r) + \text{E}(s)(\text{E}(r))']p \\ &= [CAA'C + CA\mu\mu'A'C']^{-1}[CAA' + CA\mu\mu'A]p \end{aligned}$$

Thus, if we specify A and μ, we can obtain an expression for $\hat{q}$ as a function of p. For the example in Figure 3, we use $\mu = (200,100,200)$, and A is

$$\begin{bmatrix} 1.0 & 0.1 & -0.9 \\ 0.3 & 1.0 & -0.8 \\ 0.3 & 0.3 & 1.0 \end{bmatrix}$$

so the correlation matrix for r is

$$\begin{bmatrix} 1.00 & 0.63 & -0.38 \\ 0.63 & 1.00 & -0.29 \\ -0.38 & -0.29 & 1.00 \end{bmatrix}$$

This matrix shows the degree of residential interleavening among the three groups.

One point that should be mentioned in passing is that the estimated standard errors on the estimated coefficients of the misspecified model will be very small and the goodness-of-fit (called R^2) will be very high, even though the coefficients can be quite different from what they purport to measure.[36] Thus the accuracy of the model cannot be determined from such standard techniques as R^2, and instead a more sophisticated test such as we propose above must be used.