Journal of Interdisciplinary History, IV:2 (Autumn 1973), 237–262.

This is the first of three Notes on
Ecological Regression. The two other
Research Notes will appear in succeeding issues

J. Morgan Kousser

Ecological Regression and the Analysis of Past Politics

Regression estimation of cell entries in contingency tables is among the most useful statistical techniques for political historians. Developed in the 1950s by statisticians who were attempting to circumvent the so-called "ecological fallacy," regression estimation has received a good deal of attention in other social scientific disciplines, but surprisingly little in history.[1] In a recent article in this journal, Jones provided a short introduction to the technique pioneered by Leo A. Goodman.[2] In addition, Jones tested its accuracy by comparing survey results with estimates of voting behavior from the 1960 presidential election, and urged the historical profession to utilize the Goodman procedure. Although helpful as far as it goes, Jones's paper does not treat the theory, mathematical background, and assumptions of the method in sufficient detail to enable historians to employ it creatively. Nor does he offer the researcher advice on how to

J. Morgan Kousser is Assistant Professor of History at California Institute of Technology. His book, *Shaping of Southern Politics*, will be published next year.

Several people were kind enough to offer helpful comments on previous versions of this paper: Jim Green, Daniel J. Kevles, John McCarthy, Stephan Thernstrom, and, especially, David Grether. They saved me from numerous errors. Gudmund Iversen and Howard Rosenthal sent me copies of their excellent unpublished papers.

1 Probably the most approachable full-scale discussions in the literature are Donald E. Stokes, "Cross-Level Inference as a Game Against Nature," in Joseph L. Bernd (ed.), *Mathematical Applications in Political Science*, IV (Charlottesville, 1969), 62–83; W. Phillips Shively, "'Ecological Inference': The Use of Aggregate Data to Study Individuals," *American Political Science Review*, LXIII (1969), 1183–1196. On the ecological fallacy and the development of regression estimation, see W. S. Robinson, "Ecological Correlations and the Behavior of Individuals," *American Sociological Review*, XV (1950), 351–357; L. A. Goodman, "Some Alternatives to Ecological Correlation," *American Journal of Sociology* LXIV (1959), 610–624; O. D. Duncan, R. P. Cuzzort, and B. D. Duncan, *Statistical Geography* (Glencoe, 1961), 62–80; Hubert M. Blalock, *Causal Inferences in Nonexperimental Research* (Chapel Hill, 1964), 97–114; the articles by Hayward R. Alker, Jr., Eric Allardt, and Tapani Volkonen in Mattei Dogan and Stein Rokkan (eds.), *Quantitative Ecological Analysis in the Social Sciences* (Cambridge, 1969). For a comprehensive approach and additional references on ecological regression, see Gudmund R. Iversen, "Recovering Individual Data in the Presence of Group and Individual Effects" (Ann Arbor, 1971).

2 E. Terrence Jones, "Ecological Inference and Electoral Analysis," *Journal of Interdisciplinary History*, II (1972), 249–262.

deal with typical difficulties which arise in an actual analysis of past data—for example, the problems of nonlinearity and logically impossible estimates.

The purposes of the present article are to provide historians with a somewhat broader introduction to ecological regression, to outline some strategies for dealing with typical problems encountered in using the technique and with cases which violate the assumptions of simple linear ecological regression, and to indicate how the technique can be used creatively to test different models of electoral behavior. Historians with minimal mathematical and statistical training should be able to comprehend the paper with only minor difficulties.

Historians investigating political events involving fairly large numbers of people have often used simple statistical techniques. Arthur C. Cole, for example, depicted the antebellum Southern party balance and salient socio-economic traits of counties on multicolored maps, from which he inferred various social characteristics of each party's constituency.[3] Key developed political cartography to its highest point in examining the "friends and neighbors" phenomenon and factional stability in the South.[4] More recently, however, the establishment of massive historical data collections, the spread of modern data processing equipment, and the growing number of historians with some knowledge of statistical techniques have made the use of more sophisticated methods both possible and necessary.

The maps and simple statistical techniques often employed by historians and older political scientists might be called methods of common-sense correlation. There are numerous varieties of such methods. To determine the degree of continuity in the support for Populism and Progressivism, for example, a historian might see whether a "Progressive" candidate carried the counties in which Populism had been strong. To determine whether Southern Negroes supported the Populists in the 1890s, one might focus on the election returns from counties with large proportions of Negroes. One trouble with this technique, of course, is that it takes into account only a part of the available data, ignoring counties where Populism gained only little or middling support or where there were few Negroes. To meet this objection, one might divide a state's counties into groups based on a

3 Arthur C. Cole, *The Whig Party in the South* (Washington, D.C., 1913), appendix. Many of the other political histories emanating from Frederick Jackson Turner's seminars in the early part of this century also exhibit such polychromatic maps.

4 V. O. Key, Jr., *Southern Politics in State and Nation* (New York, 1949).

specific characteristic, for instance, the proportion of white men in each, and then compare the Populist percentage in each group of counties. This technique, however, throws away specific county-level data by placing counties into categories. Within each group, different counties will not have exactly the same proportion of white men or Populists or whatever characteristic determines the category; the analyst overlooks these differences by grouping the counties together into larger areal units.

Some historians have been fortunate enough to discover sub-county election returns. Combining these with knowledge derived from censuses or other records of the socio-economic composition of certain townships, precincts, or beats, one can determine how mine workers or Swedish Protestants, for example, voted in several areas. This method works well enough if all the members of the group lived in segregated enclaves, but it may distort reality if some resided in integrated areas. To generalize confidently from intracounty patterns is to overlook the possibility that Catholics, let us say, may vote differently depending on whether they dwell in predominantly Catholic or predominantly Protestant wards.[5]

Regression analysis is a more potent device for discovering the relationships between various characteristics of a population. Suppose we know the proportion of Negroes in each county in a state, as well as the proportion of all voters voting Republican and Democratic in a certain election. For the sake of simplicity, let us assume for the moment that every eligible voter casts a ballot for one of the two parties. We can represent each county as a point on a two-dimensional graph on which the dimensions are race and politics (Figs. 1 and 2). The careful observer would note that there seems to be a positive relationship between race and Republicanism in Fig. 1; the more Negroes in a county, the more Republicans. But "the more . . . the more" does not specify a great deal about the relationship; the really interesting question is "how much more?" Further, how consistent is the relationship? If points "P" and "Q" in Fig. 1 have the same proportion of Negroes and quite different proportions of Republicans, and if "P" and "R" have the same percentage of Republicans though they differ

5 Paul Kleppner generalizes with great confidence from such sub-county returns in his recent book, *The Cross of Culture* (New York, 1970). Although one must admire Kleppner's exhausting, meticulous data-gathering, his method does not overcome the "ecological fallacy." Rather, in terms to be defined below, he ignores possible effects of "grouping."

Fig. 1 through *4* Hypothetical Data Illustrating Least-Squares Regression

Fig. 1

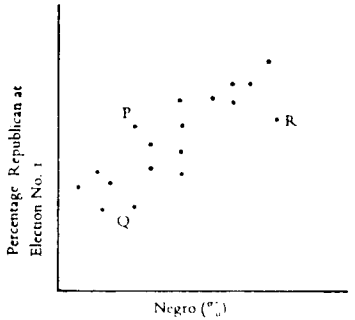


Fig. 2

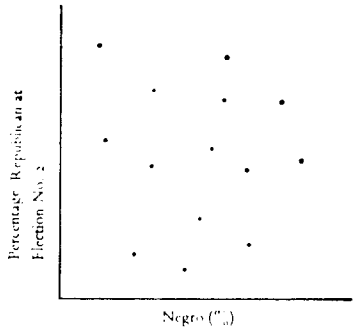


Fig. 3

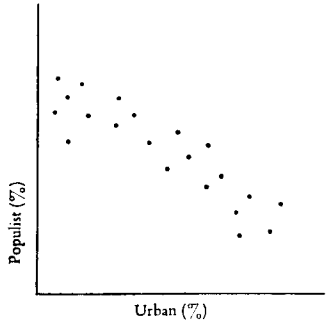
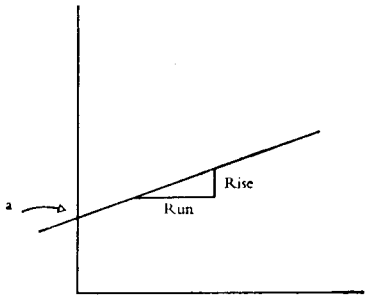


Fig. 4



widely in racial composition, can we be sure there is a relationship between party and ethnic group at all? Moreover, how can we compare relationships between the same variables at different times (Figs. 1 and 2) or between different pairs of variables (Figs. 1 and 2 with Fig. 3)?

Two of the most familiar modern statistical techniques for expressing the relationships between different sets of variables in a common, comparable form are least-squares regression and the correlation coefficient. The least-squares regression line represents a kind of average of all the points on such a graph as Fig. 1 (see Fig. 4). As the mean (average) of a series of numbers is the number which minimizes the difference of the rest of the numbers from itself, the least-squares line is the line which minimizes the sum of the squares of the vertical distances of all the points from itself. The simple regression equation which represents this average has the general form

$$Y = a + bX. \tag{1}$$

In this equation, Y is the "dependent" or predicted variable (conventionally placed on the vertical axis); X, the "independent" or predictor variable; a the point at which the line crosses the Y axis; and b the slope of the line, or "rise" divided by "run" in Fig. 4. The slope of the least-squares line measures the *form* of the relationship between two variables—it answers our "how much" question. If the slope for Fig. 1 was calculated to be 0.5, for instance, we would know that a county with 10 percent more Negroes than another county could be expected, on the average, to have 5 percent more Republicans as well. Furthermore, we could compare the form of this relationship with those depicted in Figs. 2 and 3.[6]

The usefulness of the least-squares line depends on the degree of scatter around it. Just as the mean of the incomes \$10, \$100, and \$100,000 a year tells us little about per capita prosperity, a least-squares line for such data as Fig. 2 would not represent the points very meaningfully. To supplement the slope, we need a measure of dispersion around the least-squares line, that is, of the *strength* of a relationship. Such a measure is the Pearson product-moment correlation coefficient, conventionally denoted "*r*."[7] Despite its familiarity and its enticing property of varying between +1 and −1, the correlation coefficient by itself is not a very good measure of the association between two variables.[8] While the slope of the regression line predicts the amount of increase in one variable if a related variable increases by one unit, the

6 The formulas for "*a*" and "*b*" in the two-variable case are

$$b = \frac{\sum (X-\bar{X})(Y-\bar{Y})}{\sum (X-\bar{X})^2} \quad \text{and} \quad a = \bar{Y} - b\bar{X}$$

$\bar{X}$ and $\bar{Y}$ are the averages of all the X and Y values, respectively. Sigma ($\sum$) shows that the differences are summed over all units. The simplicity of the regression methods presented here should not mislead the reader. Regression analysis can be considerably more complex. See, e.g., N. R. Draper and H. Smith, *Applied Regression Analysis* (New York, 1968).

7 The formula for *r* in the two-variable case has the same numerator as that for *b*, but the denominator for the Pearson coefficient is the product of the standard deviations of both the independent and dependent variables. *n* stands for the number of units in the analysis—e.g., the number of counties in the state.

$$r = \frac{\sum (X-\bar{X})(Y-\bar{Y})}{\sqrt{\frac{\sum (X-\bar{X})^2}{n}}\sqrt{\frac{\sum (Y-\bar{Y})^2}{n}}}$$

8 Hubert M. Blalock, "Causal Inferences, Closed Populations, and Measures of Association," *American Political Science Review*, LXI (1967), 130–137.

correlation coefficient has no correspondingly useful interpretation. Moreover, r is often the occasion for endless quibbles about whether a 0.4 or 0.5 correlation is "low," "moderately high," "fairly high," or what have you. Perhaps the historical profession, only now beginning to publish statistical articles, can learn from its sister disciplines not to substitute a measure of the strength of a relationship (r) for the more important measure of its slope, or form (b).[9] To describe a relationship fully, we need indices of both strength and form.[10] And although b is an adequate index of the form of a linear relationship, r is not the best measure of strength. A better indication of strength is r^2, which measures the percentage of variance in one variable explained by the variance in another. It would be desirable, then, for historians of voting behavior to publish both b's and r^2's for the relationships they measure.

Up to now, we have ignored the fact that election and census returns do not report the race or the place of origin, vote, etc., of individuals, but of groups of individuals. Most available election returns are for counties, and counties are almost never completely homogeneous with respect to any important population characteristic. Therefore, when we correlate county-level election returns or census data, we can only validly generalize about the units on which we have information—counties, not individuals.

Suppose we are trying to decide whether most Negroes supported the Republicans in a particular election, and we know the percentage of Negroes and the percentage of Republicans for each county in a state. Then for each county, we could construct a table like Table 1.

If we calculate a correlation coefficient (r), we use the "marginal" proportions (30, 70, 60, 40) which would be the values of X and Y in equation (1), but we know nothing about the internal cell entries (e, f, g, and h). Indeed, if we knew the internal entries for each county, we probably would not bother to calculate a correlation coefficient, for we would know exactly what percentage of whites and Negroes supported

9 Since a regression equation takes into account the actual numerical data from each county separately, it is almost always superior to the simpler methods of examining the interrelationships between variables discussed previously. Cross-hatched or multicolored maps are more useful only to point out the importance of specifically geographical variables such as "friends and neighbors" effects or regionalism not caused by other social, economic, or political conditions.

10 The comments in this paragraph of the text are confined to linear relationships between two variables. If the data points were distributed on a graph in a nonlinear fashion, one might obtain low r's and b's even if the relationship between two variables was quite strong. It is therefore usually helpful to plot the points on a graph.

Table 1 Ecological Correlation[a]

	REPUBLICAN (%)	DEMOCRAT (%)	TOTAL
NEGRO (%)	e	f	30%
WHITE (%)	g	h	70%
TOTAL	60%	40%	100%

a e in Table 1 is the proportion of Negro Republicans in the whole voting population, black and white; f, the proportion of Negro Democrats in the voting population, and so on for g and h. It should be evident that e, f, g, and h are *not* equal to the proportion of Negroes who voted Democratic, etc.

the Republicans and Democrats. The disadvantage of not knowing them is that many sets of cell entries could fit the same marginal proportions. For example, we could have a table in which $e=30$, $f=0$, $g=30$, $h=40$, in which case equal percentages of blacks and whites would have supported the Republicans, while the majority of the whites and no Negroes supported the Democrats. On the other hand, our table could read $e=0$, $f=30$, $g=60$, $h=10$, in which case the vast majority of the whites could be found in the Republican camp opposing all of the Negroes and a small minority of white Democrats. Yet in calculating the correlation coefficient both cases would be treated as exactly the same! In other words, we calculate an r on the basis of the variation *between* counties (the marginals for each county), but we do not take into account the variation within the counties (the internal cell entries); therefore, we cannot generalize about individual behavior.

We might conceptualize this situation in the following formula:

$$r_I = K_1 r_W + K_2 r_B \tag{2}$$

where r_I is the correlation between two variables on the individual level (unknown), r_W is the correlation within the counties (unknown), r_B is the correlation between all the counties (known) and K_1 and K_2 are constants.[11] The problem, known as the "ecological fallacy" because the counties are ecological units, is that of treating r_B as though it were r_I—making inferences about individuals when we only have data on aggregations of individuals. As Robinson showed in his 1950

11 If we knew the cell entries for each county, we could observe the relation between them and the marginals for all counties. If the cell entries varied in a manner directly related to the marginals, e.g., if the cell entries were approximately the same for counties with similar marginals, then r_W in equation (2) would be equal to zero, and r_B would then equal r_I. But we do not know the cell entries, so we do not know what the real value of r_W is. For the mathematical derivation of equation (2), see Duncan, Cuzzort, and Duncan, *Statistical Geography*.

article, not only is r_B not necessarily equal to r_I; they may even have different *signs*. The moral for historians is: *Do not use correlations computed from aggregated data as if they were individual correlations.*[12]

But the situation of a scholar who possesses only ecological data is not entirely hopeless. If he can justifiably make certain assumptions about the data, the student can employ ecological regression to skirt some of the statistical traps. Ecological regression also gives more intuitively meaningful statistics than correlation coefficients.[13]

Consider Table 2, which is similar to Table 1 in some respects.

Table 2 Ecological Regression

	REPUBLICAN (%)	DEMOCRAT (%)	TOTAL
NEGRO (%)	P_{11}	P_{12}	X_1
WHITE (%)	P_{21}	P_{22}	X_2
TOTAL	Y_1	Y_2	

One difference between this table and the previous one is that X's and Y's have been substituted for the numbers in the marginals in order to give the table greater generality. The other difference is that P's have been substituted for e, f, g, and h. P_{11} is the proportion of the Negroes who voted Republican, whereas e was the proportion of Negro Republicans in the whole population; in equation terms, this means that

$$P_{11} = \frac{e}{X_1} \quad \text{or} \quad e = P_{11}X_1. \tag{3}$$

Similarly, P_{12} is the percentage of Negroes who voted for the Democrats, and so on for P_{21} and P_{22}.[14] Using the P notation gives the table the desirable property that

$$P_{11} + P_{12} = P_{21} + P_{22} = 100\% = 1.00; \tag{4}$$

12 Several recent historical works use ecological correlations as if they were individual correlations. See, for example, F. Sheldon Hackney, *Populism to Progressivism in Alabama* (Princeton, 1969); Michael Paul Rogin, *The Intellectuals and McCarthy: The Radical Specter* (Cambridge, Mass., 1967); Rogin and John L. Shover, *Political Change in California: Critical Elections and Social Movements, 1890–1966* (Westport, Conn., 1970).

13 One can only estimate cell entries if the variables can be broken down conveniently into mutually exclusive classes. For instance, voters can be divided into Republican, and Democratic groups. If one has the per capita wealth figures by county, however, one cannot obtain estimates of how poor whites, for example, voted. For these sorts of analyses, one must use the standard simple and multiple regression techniques.

14 The subscripts on the P's denote their position in the matrix of rows and columns; thus, P_{11} is in the first row, first column, and P_{12} is in the first row, second column.

this is why we substitute P's for e, f, g, and h. (It should be evident from these definitions that

$$P_{11} + P_{12} \neq X_1 \quad \text{and} \quad P_{11} + P_{21} \neq Y_1.) \tag{5}$$

We also know by definition that

$$X_1 + X_2 = Y_1 + Y_2 = 100\%, \tag{6}$$

since the Negroes and the whites together compose the whole voting population, and likewise with the Republicans and Democrats.

Once we understand the properties of the table, it is relatively simple to explain how to estimate the cell entries by using least-squares regression. Expressed in our earlier terminology, the total proportion of Republicans in the population (Y_1) is the sum of the Negro Republicans (e) and white Republicans (g). To convert this equation to the terminology of Table 2, we remember that,

$$e = P_{11}X_1 \tag{7}$$

and

$$g = P_{21}X_2. \tag{8}$$

Therefore, on the average,

$$Y_1 = P_{11}X_1 + P_{21}X_2. \tag{9}$$

Similarly,

$$Y_2 = P_{12}X_1 + P_{22}X_2. \tag{10}$$

We also know that

$$X_1 + X_2 = 1.0. \tag{11}$$

Consequently,

$$X_2 = 1 - X_1. \tag{12}$$

Substituting this equation into equation (9), we have

$$Y_1 = P_{11}X_1 + P_{21}(1 - X_1). \tag{13}$$

Multiplying out the right side and rearranging terms, we have

$$Y_1 = P_{11}X_1 + P_{21} - P_{21}X_1, \tag{14}$$

$$Y_1 = P_{21} + (P_{11} - P_{21})X_1. \tag{15}$$

artic Now, equation (15) has the same form as equation (1), the simple di‹ast-squares regression equation given earlier:

$$Y = a + (b)X \tag{1}$$

$$Y_1 = P_{21} + (P_{11} - P_{21})X_1. \tag{15}$$

Therefore, we can use the values for X_1 (percent Negro) and Y_1 (percent Republican) in a regression equation, and fill in P_{21} and P_{11} by solving for the regression coefficients a and b according to the usual formulas (stated in footnote 6).

To elaborate, P_{21} is simply a. And since

$$b = P_{11} - P_{21}, \tag{16}$$

it follows algebraically that

$$P_{11} = b + P_{21} \quad \text{or} \quad b + a. \tag{17}$$

And since by definition each row of P's adds up to 100 percent (Negro Republicans and Negro Democrats comprise all the Negroes), we can easily calculate P_{12} and P_{22}. Because

$$P_{11} + P_{12} = 1.00,\ P_{12} = 1.00 - P_{11}. \tag{18}$$

Likewise,

$$P_{22} = 1.00 - P_{21}. \tag{19}$$

Therefore, given the percentages of Republicans, Democrats, Negroes, and whites in each county in a state in one election (i.e., the marginals in Table 2), we can estimate the proportion of Negroes who voted Republican for the state as a whole, as well as the other internal cell entries.

Every statistical procedure is based on certain assumptions that limit its usefulness and undermine the faith we have in its results. To use the example of a technique increasingly popular in political science, "causal modeling" is based on the often dubious assumption that all important variables are represented in the particular causal system. Likewise, however sophisticated his techniques, a statistician's results will be invalid if he cannot assume randomness in the inevitable errors which occur in measuring the variables he manipulates.

In the case of simple ecological regression, the assumption most likely to cause difficulties is that the P's are constant across all ecological units—for example, that the proportion of Negroes voting Republican is the same in every county in the state. To put it another way, we

assume that a sub-group of the population will behave similarly no matter what percentage the group forms of the total population of each county. Actually, it works out mathematically that we need not assume a sub-group behaves in *exactly* the same manner from county to county, but only that its behavior changes *randomly* when its proportion in the population varies.[15]

Empirical evidence indicates that this assumption is often realistic, or, in other words, that simple regression analysis often yields good estimates of the voting behavior of sub-groups of a population. Besides Jones's findings, mentioned earlier, Irwin and Stokes obtained regression estimates that accorded closely with survey data for recent elections in Florida and Germany, respectively.[16] Although there were no political surveys taken in the turn-of-the-century South, my own work allows comparisons between estimates of voting registration by race and the actual totals published by race for Alabama and Louisiana.

To show more specifically how one obtains estimates in a particular case, I will review the steps involved in making the estimates for the first row of Table 3, below. First, I took the number of Negro male adults and the total number of voters registered for every Alabama county and divided each figure by the total estimated adult male population in 1903.[17] The percentage data then formed the input for a least-squares regression program where Y was the percentage of the adult males registered, and X was the percentage of Negro adult males in each county. The computer output gave values for a, b, and r of 0.03, 0.76, and 0.96, respectively. The estimated proportion of blacks registered was simply a or 0.03; the corresponding figure for whites was $a+b$ or 0.79; and the correlation coefficient of 0.96 assured me that the points were tightly packed around the regression line. A glance at the graph of the percentages registered and percentage Negro for each county indicated that the relationship was unmistakably linear.

The estimates in Tables 3 and 4 are quite close to the actual

15 In more formal terms, the expected value of the error terms normally present in such equations as equation (9) is zero. (This is why we excluded the error terms from those equations.) For a proof of this proposition, see Goodman, "Alternatives," 619–620.

16 Galen Arnold Irwin, "Two Methods for Estimating Voter Transition Probabilities" unpub. Ph.D. thesis (Florida State University, 1967); Stokes, "Cross-Level Inference," 73–75. Stokes's regression estimates for British elections did not resemble the survey results so closely. See *ibid.*, 76–83.

17 This last figure was obtained by interpolation from 1900 and 1910 census data.

Table 3 Actual and Estimated Percentages of Voters Registered by Race in Alabama, 1903–1908

	ACTUAL		ESTIMATED	
YEAR	WHITE	NEGRO	WHITE	NEGRO
1903	78	2	79	3
1904	79	2	78	8
1906	75	2	78	5
1908	88	2	91	6

Table 4 Actual and Estimated Percentages of Voters Registered by Race in Louisiana, 1896–1904[a]

	ACTUAL		UNWEIGHTED ESTIMATE		WEIGHTED ESTIMATE	
YEAR	WHITE	NEGRO	WHITE	NEGRO	WHITE	NEGRO
1896	96	93	104	99	96	98
1897	103	96	108	100	102	96
1898	47	10	48	11	49	6
1902	59	3	68	7	48	16
1904	52	1	52	3	52	3

a Louisiana was notoriously slack in purging its registration rolls before 1898.

published figures. I also computed estimates weighted by population for Louisiana because it contained the only really large city in the South during this period. The Alabama estimates never deviate by more than 6 percent of the total adult population, broken down by race. The Louisiana estimates differ from the actual statistics by more than 10 percent of each group in only one instance, the 1902 weighted estimate. It appears that the New Orleans machine, which had retaken the city in 1900, stifled registration for both whites and blacks in 1902, causing the mammoth parish of Orleans to deviate markedly from the statewide pattern, and, consequently, throwing the estimates off.

It is noteworthy, too, that the estimates were fairly accurate in the Louisiana case both when registration totals were inflated and after they were severely trimmed. This fact suggests that the method is reasonably reliable for the South in this period over a broad range of election outcomes.

Furthermore, the estimates in these two states are dependable even when the related correlation coefficients, which measure the extent of

scatter around the regression lines, are quite low. The correlation coefficients corresponding to the Louisiana regression estimates dip as low as 0.02. Consistent behavior across counties is the only necessary assumption for simple regression estimates; a high correlation coefficient is neither a sure sign of such behavior nor a necessary condition for it.[18]

Regression estimation can easily be generalized from two-by-two tables to tables with larger numbers of cells.[19] Up to now, we have excluded nonvoters from our discussion for the sake of simplicity, but

Table 5 Extension of the Regression Estimation Technique to Include Nonvoters

	PERCENTAGE OF TOTAL MALE ADULTS			
	DEMOCRAT	REPUBLICAN	NOT VOTING	TOTAL
BLACK	P_{11}	P_{12}	P_{13}	X_1
WHITE	P_{21}	P_{22}	P_{23}	X_2
TOTAL	Y_1	Y_2	Y_3	

including them in the procedure is straightforward. The equations for calculating the P's in Table 5 are:

$$
\begin{aligned}
Y_1 &= P_{21} + (P_{11} - P_{21})X_1 \\
Y_2 &= P_{22} + (P_{12} - P_{22})X_1 \\
Y_3 &= P_{23} + (P_{13} - P_{23})X_1.
\end{aligned} \tag{20}
$$

If we wanted to study the coherence of a faction or the turnover of electors in two separate elections, we could apply the same procedure. To find out whether the Populists supported or opposed disfranchisement in Alabama, for example, one could estimate the internal cell entries for Table 6. In these actual estimates, I used the following equations:

$$
\begin{aligned}
Y_1 &= P_{31} + (P_{11} - P_{31})X_1 + (P_{21} - P_{31})X_2 \\
Y_2 &= P_{32} + (P_{12} - P_{32})X_1 + (P_{22} - P_{32})X_2 \\
Y_3 &= P_{33} + (P_{13} - P_{33})X_1 + (P_{23} - P_{33})X_2.
\end{aligned} \tag{21}
$$

18 Stokes, "Cross-Level Inferences," 82–83. The important thing to look for on a scatter-plot is whether the points are randomly distributed about the regression line. If the correlation coefficient is high, they will be, but they may be even if r is low.

19 It must be noted that we must still assume behavioral constancy in every cell in the table, an assumption that may be less realistic and less easy to test in tables with a great many cells. See Iversen, "Recovering," 19–20.

Table 6 Populists Opposed Disfranchisement in Alabama

		PERCENTAGE OF ADULT MALES IN REFERENDUM ON RATIFICATION, 1901			
		FOR	AGAINST	NOT VOTING	TOTAL
PERCENTAGE OF ADULT MALES IN GOVERNOR'S RACE, 1894	Democrat	$P_{11} = 56$	$P_{12} = 5$	$P_{13} = 39$	X_1
	Populist	$P_{21} = -1$	$P_{22} = 68$	$P_{23} = 32$	X_2
	Not Voting	$P_{31} = 21$	$P_{32} = 7$	$P_{33} = 72$	X_3
	Total	Y_1	Y_2	Y_3	

The case of extending regression estimation to many-celled tables has substantive as well as statistical importance. *Election analysts should always calculate their totals on the basis of all potential voters.* Excluding nonvoters draws attention away from those who enter and exit from the electorate, voters who may be particularly significant in "critical elections" and in longer periods of change. Moreover, to present election statistics based solely on the percentage of the two-party or multiparty vote is, in effect, to assert that nonvoters do not matter, and that it is unimportant that the political system either excludes many citizens outright or offers them no incentive to vote. In reality, it may be argued that the low turnout rate in United States elections is one of the most significant aspects of our political system.

As the presence of a negative entry in cell P_{21} of Table 6 makes clear, there are times when simple linear regression gives unreliable or logically impossible results. A cell entry below zero or above unity is one sign that the assumption of behavioral consistency across all the counties may have been violated. Another test is to compare, for example, the actual percentage of Republican votes for the state as a whole (obtained directly from the statewide election returns) with the predicted percentage of Republican votes based on the linear regression estimates. To predict the percentages of Republican votes, multiply the estimated percentages of white and blacks voting Republican by the numbers of adult male Caucasians and Negroes, respectively, throughout the state, and divide by the total number of adult males. If the estimated and actual percentages diverge too sharply—more than 5 or 10 percent—recalculate the estimates using a weighted regression formula.[20] Divergent estimates after recalculation are another sign that electoral behavior varied from county to county.

20 One should calculate similar estimates of the percentage voting Democratic and not voting, or whatever the other variables are. This test is not appropriate when the overall proportion of votes for a particular party and the proportions of relevant groups in the population are both close to the means for those figures computed by adding all the county proportions together and dividing by the number of counties. See Goodman, "Alternatives," 614, for a further explanation and for formulas to compute the variances of the estimates.

Furthermore, the historian should test the statistical results against contemporary views on the voting behavior of various groups. Since politicians' jobs depend on judgments of voter feelings, they are often excellent psephologists, and always vocal ones. Fortunately, they tend to preserve their impressions in memoirs, congressional debates, newspapers, etc. In addition to these bits of "impressionistic" evidence, researchers will normally have theories they wish to test. If contemporary judgments, theories, and statistical estimates differ very much, the analyst will, naturally, want to reexamine each.

The final and by far the most important test involves graphing the data and looking for patterns.[21] If the data appear to be arranged in a simple linear fashion, one may accept the estimates as roughly valid. Moreover, finding deviations from behavioral consistency may lead to better explanations of political activity. In my own work, for example, a typical pattern in Tennessee during the 1880s was for the overwhelmingly white counties in East Tennessee and the heavily black areas in West Tennessee to be disproportionately Republican, while the counties between about 10 percent and 30 percent Negro were disproportionately Democratic. In such an instance, one might choose to divide the counties into two or three groups and compute different regression estimates within each group. To classify the ecological units on the basis of their political behavior, peculiarities of their histories, or geographical contiguity is a perfectly legitimate procedure. The estimates of voting patterns within each set of counties may be more valid—i.e., come closer to meeting the assumption of behavioral consistency—than the estimates for the state as a whole. On the other hand, having to make section-by-section estimates greatly increases the time and complexity of computation, data presentation, and interpretation of the findings. Consequently, if a group of counties diverged only slightly from the statewide pattern, one might decide to present the estimates for the state as a whole and note the minor deviant trend parenthetically.[22]

21 All these tests were proposed by Goodman (*ibid.*, 612–614).

22 Separating states into sub-areas also reduces the number of areas used to calculate the estimates and the amount of variation in the value of each variable. For example, the proportion of Negroes in the eighty-eight Tennessee counties in 1890 varied from

The fact that they may fall outside the zero-to-one logical limit does not necessarily mean one should reject the simple linear regression estimates out of hand. My experience shows that one is likely to obtain such estimates whenever a group overwhelmingly supports or opposes a candidate or referendum proposal. For example, contemporaries believed that the Populists overwhelmingly opposed disfranchisement in Alabama. If impressionistic evidence corroborates the estimates, and if the figures are close to the logical limits (say, within 10 percent), then the inadmissible estimates may well be the products of random errors in "sampling." If Alabama had been partitioned into counties according to a different scheme, and Alabamians had voted in nearly the same proportions as they did, we would probably not have obtained the illogical estimate.[23] Such estimates as the one in Table 6 should be interpreted to mean: "Only a very few Populists voted for disfranchisement in Alabama." Whether in these cases one leaves the inadmissible estimates as bald testimony to the imperfections of statistical methodology, or, as Telser suggests, sets the estimates at their limits and recomputes the other estimates accordingly seems chiefly a matter of taste.[24]

Two other methods of circumventing problems, yet continuing to use simple least-squares techniques, should be mentioned. Since estimates may be distorted merely because of the places at which county boundary lines were drawn, one might combine returns from groups of two or three counties and re-compute the estimates. The groups might be formed on the bases of geographical contiguity, or one might simply pair the counties randomly.[25]

One might also switch the rows and columns, regress the percentage Negro on the percentage Republican, and estimate the proportion of Republicans who were black. By multiplying this proportion by the total vote in the state for the GOP, one obtains the estimated number of black Republicans. Divide the latter figure by the black adult male population and one has an indirect estimate of the proportion of blacks

1 percent to nearly 70 percent, whereas the corresponding percentages in the East Tennessee counties were 1 percent and about 30 percent. In smaller, more homogeneous areas, slight variations in county percentages weigh more heavily in the computations than they would in estimates for the state as a whole. Therefore, the estimates for sub-units of the states may be less reliable.

23 Iversen, "Recovering," 8, 20.

24 L. G. Telser, "Least Squares Estimation of Transition Probabilities," in Carl F. Christ (ed.), *Measurement in Economics* (Stanford, 1963), 270–292.

25 See Goodman, "Alternatives," 613.

who voted Republican. (The same procedure can, obviously, be followed to compute indirect estimates of the proportion of whites who voted Republican, the percentage of Negroes who voted Democratic, etc.)

As Shively has shown, this indirect estimating procedure may be useful in certain cases to reduce bias in the estimates.[26] Nevertheless, computing proportions by both the indirect and direct procedures is likely to be misleading, for the procedures are based on models that will usually be incompatible. In the case of the direct estimate, one may be assuming, for example, that the proportion of Negroes who vote Republican is constant except for random variations across all counties. But the analogous indirect estimate will be grounded on the very different hypothesis that the percentage of Republicans who are black is constant from county to county. In other words, the direct procedure allows us to estimate the total number of Negro and white Republicans in the state, from which we can compute both the proportion of Negroes who voted Republican and the proportion of Republicans who were Negro. But if we stick to the assumption behind the direct procedure, we cannot say we really know whether the proportion of Republicans who were Negro was roughly the same from county to county. All that we can really estimate in this case is the overall statewide proportion of Republican votes cast by Negroes, which may have varied in all sorts of ways from county to county. This overall statewide percentage may have some inherent interest, but it is not the same as an estimate of individual behavior, which is, after all, what we have been seeking. One cannot freely substitute the direct and indirect estimates for each other, for they are based on different assumptions about individual behavior.

If none of the foregoing methods yield reasonable estimates that meet the three tests Goodman proposed, the constant proportions or simple least-squares model is probably seriously misleading, and one should try more complex models. In fact, the constant proportions model is only a special case of a more general model of group and individual effects.[27]

26 For the mathematical proofs and more general discussions, see Shively, "'Ecological Inference,'" 1194–1195; Gudmund R. Iversen, "Estimation of Cell Entries in Contingency Tables When Only Margins Are Observed" unpub. Ph.D. thesis., (Harvard University, 1969), 66–76, 246–247.

27 Iversen, "Recovering," *passim.*

Table 7, which shows both group and individual effects, resembles Table 2, except that the P's are not simply equal to constants, but instead to the sums of constants and other constants multiplied by the Negro percentages for each county. In this specific example, the upper-left-hand cell indicates that a higher proportion of blacks voted Republican in the more heavily Negro counties than in those counties where whites predominated. One might account for this effect of grouping on individual behavior by theorizing that black political solidarity increased as the number of whites in the counties diminished. It is easy to see that the constant proportions model is a special case of the model used in Table 7. The simpler model simply states that $d=f=0$; i.e., that whites as well as blacks voted for each party in roughly the same percentages no matter what the racial composition of each county.

Table 7 Individual and Group Effects

	REPUBLICAN (%)	DEMOCRAT (%)	TOTAL
NEGRO (%)	$P_{11}=c+dX_1$	$P_{12}=1-P_{11}$	X_1
WHITE (%)	$P_{21}=e+fX_1$	$P_{22}=1-P_{12}$	X_2
TOTAL	Y_1	Y_2	

What is more significant is the fact that there is a whole family of equations similar to the one in Table 7 and that exploring them may well lead to new and more sophisticated explanations of past behavior. If the assumption of constant proportions does not seem valid for a particular case, then the historian can go on to try more complex models. For example, contemporaries might have agreed that although the overwhelming proportion of Negroes voted for the GOP in every county, the number of white Republicans tailed off markedly as one approached the "black belt." Concentrating on the left-hand cells in Table 7, one might express this model as

$$P_{11}=c \qquad (22)$$
$$P_{21}=e-fX_1.$$

The P's could be estimated for this model in the following manner:

$$Y_1=P_{11}X_1+P_{21}X_2 \qquad (23)$$

by definition. Substituting into equation (23) the values for the P's given in (22), we have

$$Y_1=cX_1+(e-fX_1)X_2. \qquad (24)$$

And we know

$$X_2=1-X_1. \qquad (25)$$

Therefore,

$$Y_1=cX_1+(e-fX_1)(1-X_1). \qquad (26)$$

Multiplying out and rearranging terms, we have finally

$$Y_1=e+(c-e-f)X_1+fX_1^2, \qquad (27)$$

which is of the general form

$$Y=a+b_1X+b_2X^2. \qquad (28)$$

This is a regression equation which we can solve and then obtain the values of the coefficients we want by the equations

$$e=a$$
$$c=b_1-b_2-a \qquad (29)$$
$$f=b_2.$$

We can then compare the squares of the correlation coefficients, which measure the "percentage of variance explained," for this and other models, and choose the model with the highest R^2.[28]

It should be noted that the cell entries for whites in the model based on equation (22) will not be simple constants; we will not know, in other words, what proportion of whites voted Republican and what percentage voted Democratic in the state as a whole. But we can obtain these percentages, for we will be able to compute the percentage of whites who voted Republican and Democratic for each county. We can then add up all these figures, weighing each by white population if there are large discrepancies in population from county to county, and finally divide by the number of counties. This process gives us the proportion of whites who voted Democratic and Republican across the whole state, and these percentages can be contrasted with white voting behavior in particular counties or groups of counties.

28 Alternatively, one could choose the model which minimizes the sum of squared residuals, or that in which the estimated internal cell entries have the smallest variances. See Howard Rosenthal, "Aggregate Data" (Carnegie-Mellon University, n.d.), mimeo, 20–25. Strictly speaking, these procedures for choosing between models are only appropriate when one can assume that the error terms for each county have the same variance. (The general model assumes uncorrelated error terms, all with expected means of zero.) Unfortunately, methods of compensating for unequal error terms across counties are much too complex to be described here.

In addition to the model just discussed, the individual historian who has special knowledge of a particular set of circumstances could propose others. Using the same example, he might surmise that a better explanation of political behavior would be

$$P_{11} = c \tag{30}$$
$$P_{21} = c - fZ,$$

where Z is some measure of wealth. This model expresses the hypothesis that the blacks voted Republican in constant proportions, while the poorer whites voted Republican in larger percentages than the richer ones. Or the historian might have reason to believe that

$$P_{11}/P_{21} = c. \tag{31}$$

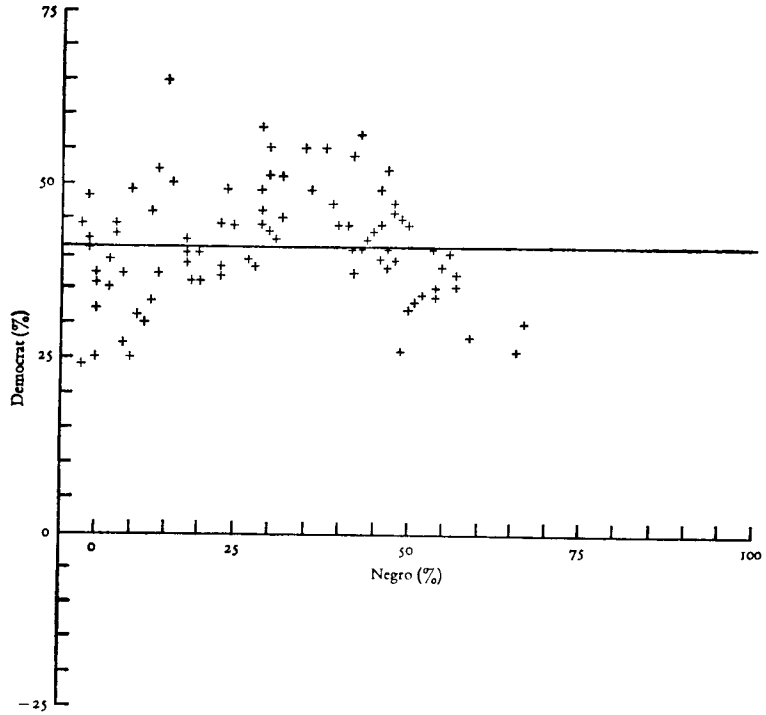
$P_{21} = c - fZ$, a model which differs from the preceding one in assuming that the proportion of blacks voting Republican was not constant, but varied with white voting behavior. One could go on to propose other models—ones with logarithmic terms, ones with different coefficients equal to zero, equal to each other, or equal to ratios of other coefficients, ones with various combinations of variables included.[29]

The point is that each investigator can set up and choose between models which have substantive importance for his data. He may employ regression estimation to test contemporary explanations of behavior; to support or attack previous historians' theories; to ascertain whether religious or social or economic groups divided in a particular election; or to determine the effect of particular events on voting behavior.

To illustrate how one might employ and choose between a few of these models, let us look at some actual data. 84.8 percent of the adult males in North Carolina turned out to vote in the hotly contested 1884 gubernatorial race. Since Negroes made up over a third of the adult male population in North Carolina in this period, it is clear that at least half the blacks and a very high proportion of the whites must have turned out. But how did members of each race vote? Were the Democrats, who garnered nearly 54 percent of the votes, simply the "white man's party," or did they also attract a large percentage of black voters through liberal rhetoric, vote-buying, fraud, or intimidation? Did white voters, seething with prejudice and apprehensive of

29 For some of these models, there would probably be no way to estimate all the coefficients.

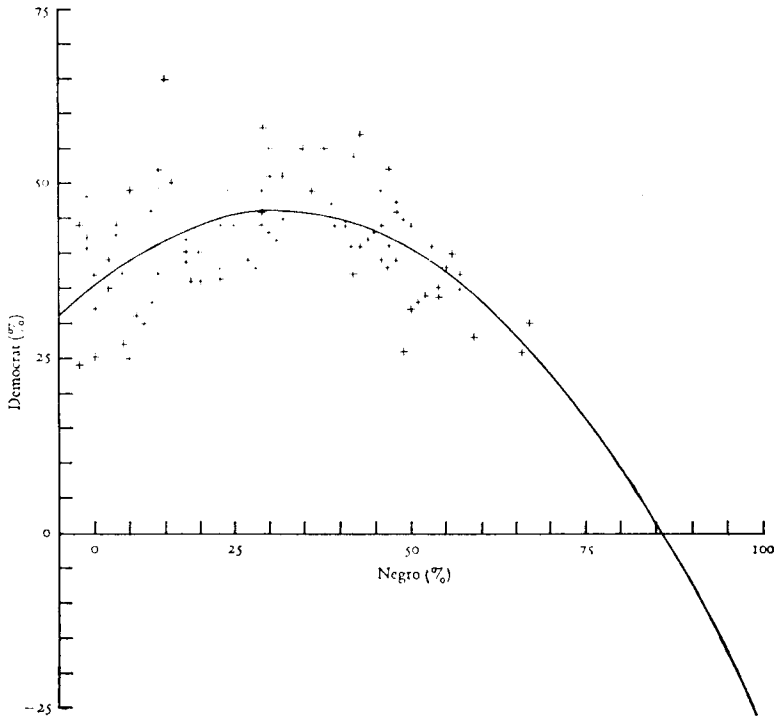
Fig. 5 The Simple Least-Squares Line for the Relation Between the Percent of Adult Males Voting Democratic and the Percentage of Negroes in Each County in the North Carolina Gubernatorial Election of 1884



"Negro domination," leave the "Radical" party to the blacks? If racial and party lines were not precisely coterminous, which whites and which Negroes crossed over?

As Fig. 5 shows, simple least-squares regression does not provide very adequate answers to these questions. The least-squares line representing the relation between the percentage of adult males voting Democratic and the percentage of Negroes in each county is nearly flat and the correlation coefficient is quite low (+0.02). An observer who was mechanically computing correlation coefficients might merely shake his head and continue the quest for high r's elsewhere. If he did, however, he would be deserting too soon, for the graph shows that the points do not fall randomly around the least-squares line. The Democratic percentages appear to rise up to about 30 percent Negro and then decline quickly.

Fig. 6 Curvilinear Regression Line for 1884 North Carolina Election



The evident trend suggests that one might try an equation with an X^2 term—say, $Y = a + b_1X + b_2X^2$, where Y is the Democratic percentage, X the Negro percentage, and a, b_1 and b_2, regression coefficients. Let us hypothesize that although roughly the same proportion of Negroes voted Democratic in each county, the percentage of whites who cast Democratic ballots rose as the black percentage climbed, perhaps because white fears of Negroes surged when blacks approached a majority. In terms of our earlier tables and equations, the model states that the proportion of blacks who voted Democratic was c_1, while the corresponding white proportion equalled $c_2 + c_3X$. (The reader should be able to fit this model into the equation $Y = a + b_1X + b_2X^2$ by following a process similar to that in equations 22 through 29).

When we solve this equation for the given data, it turns out that

$$c_1 = -0.2195$$
$$\text{and} \quad c_2 + c_3X = 0.3618 + 1.537X. \tag{32}$$

If we use the second equation to estimate the number of white Democrats in each county, sum over all counties, and divide by the total number of white adult males, we get a statewide estimate of 84.1 percent for the proportion of whites who voted Democratic. The correlation coefficient 0.529, indicates that the new equation explains about 28 percent of the variance in the Democratic vote (0.529^2); whereas, the simple least-squares line explained only 0.04 percent.

But as the negative value for c_1 and the graph of the curvilinear equation given in Fig. 6 show, there are difficulties with this model. The estimated proportion of Negroes who voted for the Democrats is the same, in this model, as the point at which the regression line intersects with the line corresponding to a theoretically all-Negro county. Yet, even though there were no actual counties above 70 percent Negro, the regression line continues to bend downward, crossing the 100 percent Negro line at the unacceptable value of −0.2195. Now, we could decide that this value indicates that no Negroes voted Democratic (i.e., $c_1 = 0$), and recalculate the white estimate accordingly.[30] But in this case there is a better solution. Since the curve in Fig. 6 seems to rise up to about 30 percent Negro and decline thereafter, why not split the state at 30 percent Negro and run separate least-squares regressions for each group of counties?

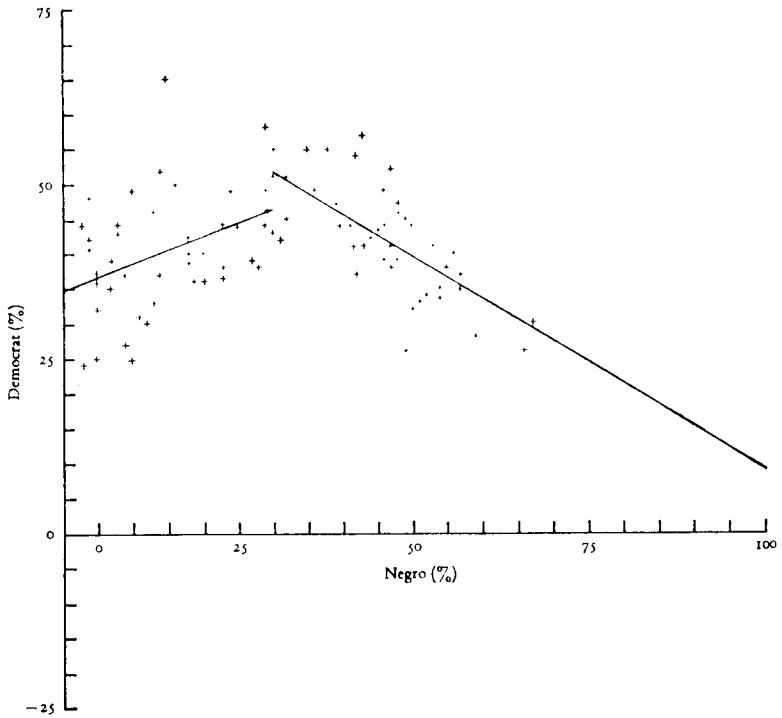
The values for the estimates obtained by this procedure are given in Table 8, the graph in Fig. 7. Among whites, nearly twice as high a proportion voted Democratic in counties over 30 percent Negro as in the "whiter" counties. Negroes reversed that pattern, nearly eight times as high a percentage voting for the Democrats in the under–30

Table 8 Estimated Percentages of Whites and Negroes Voting Democratic in Two Groups of Counties in the 1884 Gubernatorial Race in North Carolina

	PERCENTAGE OF ADULT MALES DEMOCRATIC	
RACE	COUNTIES UNDER 30% NEGRO	COUNTIES OVER 30% NEGRO
WHITE	39.7	77.0
NEGRO	78.5	12.3

30 If we did, the new values would be $c_2 + c_3X = 0.3618 + 1.3175X$, which gives a statewide estimate of 77.3 percent of the whites voting Democratic. One might also be able to find other soluble models which could fit the equation.

Fig. 7 Least-Squares Lines for Counties Above and Below 30 Percent Negro, 1884 North Carolina Election



percent Negro as in the over–30 percent Negro counties.[31] Since this model explains about the same proportion of the variance (about 28 percent) as the second model, and much more than the first, and since all the estimates from it fall in the admissible range, the third model is superior to the other two. One could perhaps, go on testing and refining other models, but the chief points about the procedure are now established.

Table 9 demonstrates the substantive significance of adopting different models. In the first model, whites vote Democratic by less than two-to-one, while nearly as many Negroes vote Democratic as

31 Two interpretations might explain the unreasonably high percentage of Negroes estimated to have voted Democratic in the under–30 percent counties. Either whites became much more Democratic in counties over about 20 per cent Negro than in whiter counties, inflating Democratic totals in the 20–30 percent Negro counties, or the whites split in such counties, and the Negroes, holding the balance of power, traded votes for favorable policies, patronage, or money.

Republican, and the estimate of Negro non-voting is inadmissible. In the second, politics entirely follows racial lines, and three of the four estimates are impossible. In the third, the whites vote Democratic by three-to-one, the Negroes Republican by about the same margin—estimates which are considerably closer to contemporary impressions.[32] Moreover, the third model encourages us to try to explain why white and Negro voting behavior seems to have varied with the racial composition of the counties. Was there a threshold percentage of Negroes in a community below which blacks could not organize to vote in a bloc? Conversely, was there a percentage of blacks below which whites, in relatively peaceful times at least, did not worry very much about the presence of Negroes? To answer these and other similar questions, a historian would have to examine large numbers of "impressionistic" sources, and perhaps the writings of psychologists, sociologists, and other social scientists, as well as additional election data.

Table 9 Three Models of Voting by Race in North Carolina, 1884

		PERCENTAGE OF ADULT MALES		
MODEL	RACE	DEMOCRATIC	REPUBLICAN	NOT VOTING
Statewide least-squares	W	46	28	26
	N	47	59	−6
	R^2	0.04[a]	30.25	42.16
Curvilinear	W	84	−4	n.c.[b]
	N	−22	117	n.c.
	R^2	27.98	43.96	n.c.
Split at 30% Negro	W	59	20	n.c.
	N	24	74	n.c.
	R^2	27.80	42.19	n.c.

a Percentage of variance explained.
b Not computed.

In sum, in addition to providing a good deal of promise of overcoming the "ecological fallacy," the method outlined in this paper

32 Statewide estimates for the third model were computed by multiplying, e.g., the number of white adult males in each group of counties by the estimated percentages of whites voting Democratic, adding the resulting estimated number of white Democrats in both sections, and dividing by the number of white adult males in the state. Similar procedures yielded the statewide percentages of white Republicans, Negro Democrats, and Negro Republicans.

also allows much more sophisticated hypothesis-testing and model-building than the simple correlational methods often used in the profession at the present time. Far from simply mechanical, the procedure gives full range to the historian's creative impulses, while at the same time demanding increased analytical rigor. Like other good statistical methods, regression estimation does not allow the data analyst to rely on the computer to spew out proper interpretations magically. Regression estimation is a way of testing theories put forth by contemporaries and other historians, not a method for manufacturing analyses.

Journal of Interdisciplinary History, IV:2 (Autumn 1973), 263–272.

William O. Aydelotte

Lee Benson's Scientific History: For and Against

Toward the Scientific Study of History: Selected Essays. By Lee Benson (Philadelphia, J. B. Lippincott Company, 1972) 352 pp. $8.75 (Paperback, $3.95)

What Benson means by "scientific" history is, in a definition which he borrows from Ernest Nagel, an attempt "to provide systematic and responsibly supported explanations . . ." (110). He is not trying to make another contribution to the silly argument over whether history is like physics. He simply wants methods of historical scholarship that are more thoughtful and more careful. He does, however, hold that the primary goal of scientific history or history as social science "is to help develop general laws of human behavior" (199). Although earlier attempts by John W. Burgess, Henry Adams, Frederick Jackson Turner, and Charles A. Beard to make history scientific were, Benson believes, unsuccessful, he hopes that some progress in this direction may be possible if historians are properly trained and follow suitable approaches. In this book he has included eight essays, of which all but the last and longest have been previously published. The essays are unequal both in length and in value: some are rather slight; others are of major consequence.

Taken together, the essays constitute a statement of principles, a manifesto, by a student who has devoted a good deal of attention to exploring new approaches to research in American history. Benson explains why scholars have not developed a genuinely scientific historiography, in the sense in which he uses the term, and what, specifically, they can do to make amends. Among his suggestions are: formal methods, including mathematical tools; rigorous testing of hypotheses by a systematic and effective marshalling of evidence, rather than reliance on impressions; a careful formulation of the issues and objectives of research, including the development of theoretical models in which assumptions are identified and examined as closely as the limitations of the human mind will permit; and coordination of the

William O. Aydelotte is Professor of History at The University of Iowa. He is the author of *Bismarck and British Colonial Policy* (New York; rev. ed., 1970) and *Quantification in History* (Reading, Mass., 1971); and, with Allan G. Bogue and Robert W. Fogel, is the editor of *The Dimensions of Quantitative Research in History* (Princeton, 1972).