DISCUSSION PAPERS IN ECONOMICS

Working Paper No. 04-18

Fifty Years of Goodman's Identity: Its Implications for Regression-Based Inference

Jeffrey S. Zax
Department of Economics, University of Colorado at Boulder
Boulder, Colorado

December 2004

Center for Economic Analysis
Department of Economics

University of Colorado at Boulder
Boulder, Colorado 80309

FIFTY YEARS OF GOODMAN'S IDENTITY: ITS IMPLICATIONS FOR REGRESSION-BASED INFERENCE

Jeffrey S. Zax
Professor
University of Colorado at Boulder
Department of Economics
256 UCB
Boulder, CO
80309-0256

Telephone: 303-492-8268
FAX: 303-492-8960
e-mail: zax@colorado.edu

5 December 2004

ABSTRACT

This paper examines the implications of Goodman's Identity for estimation and inference in linear regression. Its empirical implementation requires the assumption of random coefficients or measurement error. Under the former, regression can be surprisingly potent but is typically misused. With one application of Goodman's Identity, regression can test the neighborhood model, aggregation bias and effects of covariates. Models with more than two groups are completely identified and yield more powerful tests. However, most implementations unwittingly impose the neighborhood model, weight incorrectly and offer meaningless R^2 values as "validation". Moreover, regression is essentially useless for most models requiring two applications of Goodman's Identity, including those of voting with unknown group-specific turnout rates.

Goodman (1953) asserts that the parameters in problems of ecological inference are related by an identity. He proposes that, under appropriate conditions, regression analysis can recover them. This proposal has subsequently been the basis of countless regression-based applications. However, many implications of Goodman's identity for these applications have not previously been explored.

This paper demonstrates that regression techniques are both much more and much less powerful than is generally understood. The literature has concluded that empirical techniques cannot distinguish between the neighborhood model and Goodman's identity as the underlying source of observed data. It has also concluded that the form of aggregation bias, if present, is not identifiable in linear specifications. Lastly, it tends to ignore many plausible covariates of the behavior at issue.

This paper demonstrates that, in a generalized linear specification of Goodman's regression with feasible corrections for heteroskedasticity, valid tests of aggregation bias, the neighborhood hypothesis and the presence of covariates are possible. With more than two groups in the population, linear aggregation bias is identifiable. However, these properties hold only where a single application of Goodman's identity is sufficient to describe the behavior at issue.

At least one important application, the comparison of voting choices across groups, usually requires two applications of Goodman's identity. The first addresses the unobserved turnout rates within groups. The second addresses their unobserved vote choices. In this context, regression-based estimators have expected values that depend on multiple parameters, usually in nonlinear combinations. Identification is possible, if at all, only in models that are considerably more restricted or considerably more complex than those that typically appear in the literature.

Section I presents the general behavioral model that underlies Goodman's regression in the context of a single application of Goodman's identity. Section II discusses the neighborhood model, aggregation bias, heteroskedasticity and weighting in the context of this model, under the assumption that measurement error is absent. Section III extends this model to the R×C case, in which more than two groups are present in the population and more than one characteristic or choice is at issue. Section IV explores the difficulties of regression-based inference when the behavioral model requires two applications of Goodman's identity. Section V concludes.

I. The behavioral model for Goodman's regression

Goodman's identity (Goodman, 1959, 612) relates the proportion of a population with a particular characteristic or making a particular choice to the proportions of the population comprised by its two constituent groups. Let

x_i = the proportion of the population in area *i* that belongs to group 1,

$1-x_i$ = the proportion of the population in area *i* that belongs to group 2, and

y_i = the proportion of the population in area *i* with the characteristic or choice at issue.

The relationship between these three quantities in area *i* is the identity[1]

$$y_i \equiv \beta_{1i} x_i + \beta_{2i}\left[1 - x_i\right], \qquad \textbf{(1)}$$

where

[1] Throughout, square brackets contain quantities that are the objects of explicit algebraic operations. Parentheses contain arguments to functions.

β_{1i} = the proportion of group 1 in area *i* with the characteristic or making the choice, and

β_{2i} = the proportion of group 2 in area *i* with the characteristic or making the choice,

are the two unknown parameters of interest.[2]

Equation 1 can be rewritten as

$$y_i \equiv \beta_{2i} + \left[\beta_{1i} - \beta_{2i}\right]x_i . \quad \textbf{(2)}$$

Equation 2 demonstrates that the proportion of the population with the characteristic or making the choice can be represented as a linear function of the share of group 1 in the population. This suggests an apparent analogy between equation 2 and the conventional representation of the linear regression model.

Accordingly, Goodman (1953, 664 and 1959, 612) suggests that the parameters in this behavioral identity can be estimated by an Ordinary Least Squares (OLS) regression of y_i on x_i, with observations on *n* different areas displaying a variety of values for x_i. In this example, "Goodman's regression" is

$$y_i = b_0 + b_1 x_i + e_i . \quad \textbf{(3)}$$

Goodman asserts that, under appropriate conditions, this regression yields b_0 and b_1 as unbiased

[2] This is the "two-party, no abstention" case of Achen and Shively (1995, 30) and the "basic model" in King (1997, chapter 6). The "ecological inference problem" is often stated as the challenge of recovering parameters governing individual behavior from aggregate data (Robinson (1950, 352), Goodman (1953, 663) and King (1997, 7), as examples). However, the parameters of Goodman's identity describe behavior at the aggregate level, here the "area". Achen and Shively (1995) discuss the problem of deriving macrorelations from microfoundations (pages 23-25) and present behavioral models in which the aggregate parameters in Goodman's identity become explicit functions of individual-level parameters (chapters 2 and 4). King (1997, 119-122) discusses some difficulties with this approach.

estimators of β_{2i} and $\beta_{1i}-\beta_{2i}$ (1953, 664 and 1959, 612).[3]

However, the behavioral model that underlies OLS regression specifies that the dependent variable is only partially determined by the explanatory variable. It also depends upon a random component that is additive and orthogonal to the explanatory variable (Greene (2003, 10-11)). The properties of this random variable allow the conventional empirical OLS model to yield unbiased estimators.

In contrast, Goodman's identity is exact. In the example of equation 2, the value of y_i is completely determined by the value of x_i. Consequently, the analogy between Goodman's regression and the conventional linear regression model is superficial. The true properties of estimators from Goodman's regression must be derived analytically from the implications of the identities upon which they are based, rather than by analogy from those of conventional OLS estimators.

As written, the parameters of Goodman's identity are not identifiable. In the example of equation 2, a different identity holds for each area. Each area requires two unique parameters, but provides only one observation (Achen and Shively (1995, 12), King (1997, 39)).

Under equation 2, the empirical regression of y_i on x_i given in equation 3 would be meaning-less. The expected value of the slope coefficient would be

$$E(b_1) = \frac{\sum_{i=1}^{n}[x_i - \bar{x}]E(y_i)}{\sum_{i=1}^{n}[x_i - \bar{x}]x_i} = \frac{\sum_{i=1}^{n}[x_i - \bar{x}]\beta_{2i}}{\sum_{i=1}^{n}[x_i - \bar{x}]x_i} + \frac{\sum_{i=1}^{n}[x_i - \bar{x}][\beta_{1i} - \beta_{2i}]x_i}{\sum_{i=1}^{n}[x_i - \bar{x}]x_i}. \quad \textbf{(4)}$$

[3] This analogy is common in subsequent literature. Kousser (2001, equation 13) is an example.

Here, the second equality replaces $E(y_i)$ with y_i as a consequence of equation 2, and then replaces y_i with its equivalent in terms of x_i, as given there. Without further assumptions, the two ratios to the right of the second equality in equation 4 cannot be simplified. Consequently, the estimated slope coefficient is not interpretable. It is certainly not an unbiased estimator of the difference $\beta_{1i}-\beta_{2i}$ for any value of *i*.

Clearly, Goodman's identity requires assumptions that reduce the number of underlying parameters in order to be empirically useful. However, this is not sufficient. If, in the example of equation 2, the behavioral parameters β_{1i} and β_{2i} were assumed to be constant across all areas, the identity would be the same for each area: $\beta_{1i}=\beta_1$ and $\beta_{2i}=\beta_2$ for all *i*. It could be rewritten without the *i* subscripts in terms of only two unknown parameters:

$$y_i \equiv \beta_1 x_i + \beta_2\left[1 - x_i\right]. \quad \textbf{(5)}$$

Data on y_i and x_i from only two areas would be sufficient to determine these parameters exactly, because equation 5 is exact. An empirical regression of y_i on x_i would be unnecessary. Were it undertaken, the slope coefficient would be identically $\beta_1-\beta_2$ and the intercept would be identically β_2. The prediction errors for each observation would be identically zero and R^2 would be identically one.[4]

These last two characteristics are entirely absent from the literature. This implies that the assumption of parameter constancy across observations, alone, is not sufficient to endow

[4] Similarly, the area-specific parameters of equations 1 or 2 could be identified for area *i* if the parameters β_{1i} and β_{2i} were constant over time, x_i and y_i were observed twice, and the group share x_i was different for the two observations. In this case, y_i would necessarily also vary across the two, again providing an exact solution. The regression of equation 3 would still yield the incomprehensible results of equation 4. Regressions using only repeated observations for a single area would achieve a perfect fit. Although many empirical examples, such as that of voting behavior, offer repeated observations within area, only Lewis (2004) and Quinn (2004) appear to have explored this identification strategy.

Goodman's identity with empirical relevance. Moreover, the literature universally ignores the implication of exact solutions embodied in equations 2 or 5 in favor of statistical formulations such as equation 3. This implies that the collective intuition expects some random element in the behaviors at issue.

In sum, sensible interpretations of Goodman's regression in equation 3 require two types of elaborations in Goodman's identity. First, an assumption must be adopted to reduce the number of parameters in equation 1. Second, an assumption must endow it with random components.

The first requirement can only be satisfied by specifying that the parameters for each area are fixed functions of a limited number of variables:

$$\beta_{1i} = f_1(x_i, z_{1i}) \text{ and } \beta_{2i} = f_2(x_i, z_{2i}). \quad \textbf{(6)}$$

This reduces the number of parameters to that necessary to characterize f_1 and f_2.[5]

In addition, equation 6 is the only formulation that preserves Goodman's identity while expanding it to include covariates of y_i other than x_i. In particular, it is the only formulation that can explicitly incorporate "aggregation bias", the possibility that the proportion of a group with the characteristic at issue depends on that group's share in the area's population. Equation 6 admits this through the explicit presence of x_i in f_1 and f_2.[6] The vectors z_{1i} and z_{2i} contain any other determinants of the proportions of the two groups with the characteristic at issue.

---

[5] This is a general form for the model of "deterministic heterogeneous transition rates" in Achen and Shively (1995, 39-45).

[6] "The assumption that the coefficients are independent of the regressors is the critical problem in ecological inference." (Rivers (1998, 442)). King (1997, 40) states that this assumption is "wrong" and Achen and Shively (1995, 13) characterize it as "always dubious" (page 13). Both assert, correctly, that if this assumption is false, typical specifications of Goodman's regression are biased. The latter add, again correctly, that the bias cannot be corrected through weighting (page 51, footnote 19).

The second requirement cannot be satisfied by simply adding a random component directly to the right side of equation 1. As should be obvious, this tactic fails because it invalidates Goodman's identity. However, random components can be embedded in all of the quantities already present in that identity.

First, the parameters can contain random as well as deterministic components. This "random coefficients" formulation is nearly explicit in Goodman (1959, 612), where he asserts that parameters will vary across areas but share the same expected value.[7] It is absent from most of the subsequent literature, but is central to Achen and Shively (1995) and King (1997). Here, it implies that equation 6 be augmented as

$$\beta_{1i} = f_1(x_i, z_{1i}) + \varepsilon_{1i} \text{ and } \beta_{1i} = f_2(x_i, z_{2i}) + \varepsilon_{2i}, \quad \textbf{(7)}$$

where $E(\epsilon_{1i})=E(\epsilon_{2i})=0$, ϵ_{1i} and ϵ_{2i} are orthogonal to x_i, z_{1i} and z_{2i}.[8]

Second, the population share x_i may be measured with error. If x_i is the true value, the measured value x_i^* would differ from it by an additive random error:

$$x_i^* = x_i + \nu_i, \quad \textbf{(8)}$$

where $E(\nu_i)=0$ and ν_i is orthogonal to x_i.[9] For example, analyses of voting behavior often

---

[7] Goodman (1953) refers to the constants in his identity as both "parameters" and "average probabilities" (pages 664 and 663, respectively). This apparent ambiguity may have been an early anticipation of the random coefficients model.

[8] The "sophisticated Goodman model" of Achen and Shively (1995, 51) sets f_1 and f_2 constant in equation 7. These functions can presumably be more complicated in their "extended sophisticated Goodman model" (page 68).

[9] Chapter 3 of Achen and Shively (1995) is essentially an exposition of the substantial challenges that measurement error presents in the context of Goodman's regression. Lichtman (1974, 422) also identifies measurement error as an important concern in ecological regression. Irwin and Lichtman (1976, 415-416) point out that aggregation may create correlations between x_i and the unobserved component of y_i, as well. Equation 8 reverses the conventional notation, in which the superscript asterisk identifies the

compare votes in an election from one year with population proportions from a census in another. If these proportions can change, the measured proportion may not accurately reflect the relevant electorate.

Together, equations 1, 7 and 8 yield a general restatement of Goodman's identity:

$$y_i \equiv \left[f_1\left(x_i^* - \nu_i, z_{1i}\right) + \varepsilon_{1i} \right]\left[x_i^* - \nu_i\right] + \left[f_2\left(x_i^* - \nu_i, z_{2i}\right) + \varepsilon_{2i} \right]\left[1 - \left[x_i^* - \nu_i\right]\right],$$

or

$$\begin{aligned} y_i \equiv\ & f_2\left(x_i^* - \nu_i, z_{2i}\right) + \left[f_1\left(x_i^* - \nu_i, z_{1i}\right) - f_2\left(x_i^* - \nu_i, z_{2i}\right)\right]x_i^* \\ & + \left[\begin{array}{l} f_1\left(x_i^* - \nu_i, z_{1i}\right) - f_2\left(x_i^* - \nu_i, z_{2i}\right)\nu_i \\ + \left[\varepsilon_{1i} - \varepsilon_{2i}\right]\left[x_i^* - \nu_i\right] + \varepsilon_{2i} \end{array}\right]. \end{aligned} \qquad \textbf{(9)}$$

Equation 9 demonstrates that this generalization is still an identity. Nevertheless, it has the statistical character that is absent in equation 2 but present in equation 3. The two deterministic terms to the right of the identity in equation 2 have their counterparts in the first two terms to the right of the identity in equation 9. However, equation 9 contains a third term to the right of the identity, which consists of all of the random elements introduced through the assumptions of random coefficients in equation 7 and measurement error in equation 8.

Equation 9 demonstrates that appropriate estimation of Goodman's identity, under this complete generalization, presents substantial challenges. First, the measurement error ν_i appears both among the explanatory variables and the unobserved component of y_i. This ensures that the

---

true value (Greene (2003, 84)). This is convenient below, where measurement error is disregarded.

OLS formulas for b_0 and b_1 will yield inconsistent estimators (Greene (2003, 85)). Second, for most choices of interest, z_i could plausibly contain many elements, perhaps in nonlinear combinations. OLS estimators will generally suffer from bias if the specifications of f_1 or f_2 are incorrect.[10]

These challenges also represent generic sources for any unsatisfactory results that may arise from estimations of equation 3. For example, these estimates can yield values for b_0 and $b_1 - b_0$ that are outside the logical bounds of zero and one.[11] Achen and Shively (1995) suggest that this problem could arise if f_1 and f_2 are incorrectly assumed to be constant (page 15), or if measurement error is present (chapter 3). More generally, inconsistency or specification bias are inherent threats to estimates of Goodman's regression. Either or both could be responsible for almost any inadequacy observed in actual examples.

II. Goodman's regression in the absence of measurement error

Empirical implementation of equation 9 requires some response to its challenges. The most severe problem, measurement error, may be remediable through instrumental-variables techniques. However, these techniques do not appear to have been attempted in the ecological regression literature.[12] The rest of this paper therefore defers discussion of this issue, and

[10] Goodman (1959, 612-3) identifies this problem. It reappears in, as examples, Hanushek, Jackson and Kain (1974), Lichtman (1974) and Kousser (2001, 108).

[11] Achen and Shively (1995, 75) conclude that "(l)ogically impossible estimates in ecological regression ... are encountered perhaps half the time, and more often as the statistical fit improves. Ecological regression fails, not occasionally, but chronically." King (1997, 57) states that failures occur "often". In contrast, Kousser (2001, 117-8) asserts that impossible estimates are relatively infrequent.

[12] For example, Achen and Shively (1995, 35, footnote 5) note that, in the study of consecutive elections, attrition and accession to the electorate will ordinarily generate measurement error in the

assumes that x_i is measured without error.

In the absence of measurement error, the assumption of random coefficients is necessary to endow Goodman's identity with any random component, as well as to allow for explicit representation of aggregation bias and other covariates. Its presence in the literature since at least Goodman (1959) indicates that it has intuitive appeal as well. It therefore represents the most pragmatic strategy for interpreting Goodman's regression. Without measurement error, equation 9 becomes

$$y_i = f_2(x_i, z_{2i}) + \left[f_1(x_i, z_{1i}) - f_2(x_i, z_{2i}) \right] x_i + \left[\varepsilon_{1i} x_i + \varepsilon_{2i} \left[1 - x_i \right] \right]. \qquad \textbf{(10)}$$

The expected value of the residual term in equation 10 is zero. This term is also uncorrelated with the deterministic component of y_i. Therefore, OLS estimates of the functions f_1 and f_2 will be unbiased if the empirical equation represents them correctly (Greene (2003, 44)). In particular, aggregation "bias" will not bias OLS estimators if the regression equation correctly specifies the form in which x_i enters f_1 and f_2.

Equation 10 reveals an important principle of specification. The first term to the right of the equality indicates that f_2 appears in the expanded Goodman's identity without transformation. However, the second term to the right of the equality indicates that the difference $f_1 - f_2$ is interacted with x_i. Therefore, a general empirical specification requires that these interactions appear in the estimated equation.

---

explanatory variable. However, they conclude that "(t)hese fine points are always ignored in practice." Judge, Miller and Cho (2004) offer an attempt to confront them.

A. Goodman's regression and the "neighborhood model"

These interactions provide a test for a very controversial assumption. The "neighborhood model" (Freedman, et al. (1991) and Klein, Sacks and Freedman (1991)) assumes that, within any area, the proportions of each group with the characteristic or making the choice at issue are identical. Variation in y_i across areas arises from variations in the determinants of that characteristic or choice, rather than from variations in characteristics or choice proportions across groups coupled with variations in population composition across areas.

This assumption requires that, at a minimum, the deterministic components of the choices for each group within an area are the same. In other words, the neighborhood model is nested in the model of equation 10, where it imposes the restriction that $f_1=f_2=f$. With this restriction, equation 10 becomes

$$y_i = f(x_i, z_i) + \left[\varepsilon_{1i} x_i + \varepsilon_{2i}\left[1 - x_i\right]\right]. \quad \textbf{(11)}$$

If, as in Freedman, et al. (1991), f is linear in x_i and does not depend on z_i, then equation 11 is linear in x_i. In this case, it has the same form as equation 3. This similarity has suggested that empirical evidence regarding the relationship between y_i and x_i cannot distinguish between the neighborhood model and Goodman's identity ((Freedman, et al. (1991, 682), Klein, Sacks and Freedman (1991), Lichtman (1991, 787), Achen and Shively (1995, 14), King (1997, 41-44) and Kousser (2001, 105-7), as examples).

However, equation 10 demonstrates that, if $f_1 \neq f_2$ but both are still linear functions of x_i alone, the correct empirical specification in this example appends a quadratic term in x_i to equation 3. The neighborhood model implies that the coefficient on this term equals zero. This implication is

testable, and the model itself is therefore falsifiable.[13] With more general specifications for f_1 and f_2, the neighborhood model's implication that $[f_1 - f_2]x_i=0$ should be similarly testable.[14]

This point can be illustrated in a general linear multivariate model of f_1 and f_2. This model includes virtually all empirical examples of Goodman's regression as special cases.[15] It specifies that the arguments of f_1 and f_2 are the same: $z_{1i}=z_{2i}=z_i$. In addition, f_1 and f_2 are linear in all arguments:

$$f_1(x_i, z_i) = \beta_1 + \beta_{10} x_i + \sum_{j=1}^{k} \beta_{1j} z_{ij} \text{ and}$$
$$f_2(x_i, z_i) = \beta_2 + \beta_{20}\left[1 - x_i\right] + \sum_{j=1}^{k} \beta_{2j} z_{ij}, \quad \textbf{(12)}$$

[13] This test is implicit, though unacknowledged, in King (1997, 41-4). The previous literature does not offer a precise general specification of the neighborhood model. Equation 10 demonstrates that the specification in the text is the "weak form" of this model. The "strong form" would also require $\epsilon_{1i}=\epsilon_{2i}$. This imposes the restriction of homoskedasticity on the empirical error terms, which can presumably also be tested. The heteroskedasticity in the unrestricted form of equation 10 is discussed below.

[14] The more restrictive neighborhood model, in which f is a constant, is also testable. It asserts that y_i does not depend on x_i. Equation 3 tests this restriction, which would be rejected if b_1 were statistically significant. Freedman, et al. (1991, 682) suggest the "nonlinear" neighborhood model as an alternative in which $\beta_{1i}=\beta_{2i}=y_i$. Mechanically, this is a tautology rather than a model, because it reduces Goodman's Identity in equation 1 to $y_i \equiv y_i$. It is simply a restricted version of the "model" represented by King's tomography plots (1997, figure 6.3, as an example). These plots demonstrate that, algebraically, an infinite number of pairs of values for β_{1i} and β_{2i} satisfy Goodman's Identity, as reformulated in King's equation 6.27, for each area. For each area, the nonlinear neighborhood model simply chooses the single pair that satisfies the restriction $\beta_{1i}=\beta_{2i}$. It is not evident that this pair has any greater claim to validity than any other on the same tomography line. Any "model" consisting of one pair of values from each of these lines will "fit" the data perfectly, by absorbing all degrees of freedom. At the same time, any procedure of this type will have no predictive value because it is "nihilistic" (Kousser (2001, 105): It implicitly asserts that scientific analysis is not applicable because voting behaviors across areas have nothing in common. If this assertion is unacceptable, than all procedures of this type, including the nonlinear neighborhood model, are irrelevant.

[15] In fact, Achen and Shively (1995, 13 and 73, footnote 14) assert that "(j)ust one technique for handling ecological data has been widely adopted in practice: the linear (unextended) version of Goodman ecological regression". This is the model of equation 10, in which f_1 and f_2 are both constants. In the terms of equation 12, this model assumes that $\beta_{1j}=\beta_{2j}=0$ for all *j*.

where z_i is a vector, *k* represents the number of covariates in z_i and z_{ij} represents the *j*th covariate in z_i.

Under equation 12, equation 10 becomes

$$y_i = \left[\beta_2 + \beta_{20}\right] + \left[\beta_1 - \beta_2 - 2\beta_{20}\right]x_i + \sum_{j=1}^{k} \beta_{2j} z_{ij} + \left[\beta_{10} + \beta_{20}\right]x_i^2 + \sum_{j=1}^{k} \left[\beta_{1j} - \beta_{2j}\right] z_{ij} x_i + \left[\varepsilon_{1i} x_i + \varepsilon_{2i}\left[1 - x_i\right]\right]. \quad \textbf{(13)}$$

Each of the elements of z_i appears linearly and interacted with x_i.[16] The latter variable appears in both linear and quadratic terms.

Consequently, the appropriate estimating equation would be

$$y_i = a + bx_i + \sum_{j=1}^{k} c_j z_{ij} + dx_i^2 + \sum_{j=1}^{k} h_j z_{ij} x_i + e_i, \quad \textbf{(14)}$$

where e_i represents the empirical residual term. The estimated coefficients *a*, *b*, c_j, *d* and h_j would be unbiased estimators of $\beta_2+\beta_{20}$, $\beta_1-\beta_2-2\beta_{20}$, β_{2j}, $\beta_{10}+\beta_{20}$ and $\beta_{1j}-\beta_{2j}$, respectively. The difference h_j-c_j would be an unbiased estimator of β_{1j}. Linear combinations of all identified parameters would be estimated without bias by the same linear combinations of the corresponding estimators.[17] β_1, β_{10}, β_2 and β_{20} would not be individually identified.[18]

---

[16] Achen and Shively (1995, 40, footnote 8) also note that the complete multivariate linear specification of Goodman's identity requires interaction terms.

[17] According to Achen and Shively (1995, 58) and King (1997, 32-3), linear combinations of the area-specific parameters are often of interest. Kousser (2001, 107) suggests them as specification checks.

[18] Equation 12 specifies f_2 as a function of the group 2 proportion $[1-x_i]$ for consistency with the analysis of Goodman's regression when the population contains more than two groups, in section III below. With only two groups, f_2 could be specified as a function of the group 1 proportion x_i instead. This

In terms of equation 13, the neighborhood model imposes the restrictions that $\beta_1=\beta_2$ and $\beta_{1j}=\beta_{2j}$ for all *j*. The former restriction is not testable, but the latter implies that all of the h_j should be statistically indistinguishable from zero. The F-test of this joint null hypothesis is therefore a strong test of the neighborhood model. If the null hypothesis is rejected, the neighborhood hypothesis is clearly false.[19]

This section demonstrates that the linear specification of equation 11, in which the function f is not interacted with x_i, imposes the neighborhood model as a maintained hypothesis. This specification is nearly universal, but the concomitant adoption of the neighborhood model is almost surely inadvertent. Kousser (2001), who is explicitly hostile to this model (pages 105-7, 110) is a particularly ironic example (pages 110-5). Even partisans of the neighborhood model should be obligated to include the interaction terms, in order to test its implications.

B. Goodman's regression, aggregation bias and covariates

The model of equation 13 and its empirical implementation in equation 14 address two other issues that are central to ecological investigations, and previously taken to be intractable. First, aggregation bias is present in equation 13 and its components, β_{10} and β_{20}, are not identified.[20]

---

specification allows the additional identification of β_2. However, this identification does not alter any of the substantive results described here.

[19] The power of this test may be limited. Even if evidence supports the assertion that $\beta_{1j}=\beta_{2j}$ for all *j*, it is still possible that $\beta_1 \neq \beta_2$ and the neighborhood hypothesis is false. The difference $\beta_1 - \beta_2$ is not identified in equation 13, and therefore cannot be tested in equation 14. However, it may be identifiable if additional restrictions apply to equation 12. For example, if f_2 is free of aggregation bias, $\beta_{20}=0$ and β_{10}, β_1 and β_2 are identified.

[20] Rivers (1998, 443) asserts that this model is unidentified. King (1997, section 3.2) and Voss (2004, 72-73) provide simple examples. Achen and Shively (1995, chapters 5 and 6) discuss identifying strategies in otherwise underidentified ecological regression models which could be effective here, if behaviorally appropriate. In this case, for example, the assumption that $\beta_1=0$ is sufficient to identify β_2,

However, equation 14 provides a strong test for the absence of aggregation bias.

This bias is present if either β_{10} or β_{20} is nonzero. Therefore, its absence requires $\beta_{10}=\beta_{20}=0$. This implies the restriction $\beta_{10}+\beta_{20}=0$ as a necessary condition. The coefficient on x_i^2, *d*, identifies this sum. Accordingly, if the corresponding t-statistic rejects the the hypothesis that $d=0$, it also rejects the null hypothesis of no aggregation bias.[21]

Second, even in the presence of unidentified aggregation bias, equation 14 identifies the behavioral determinants of the proportions of the two groups making the choice at issue. The coefficients of z_i for groups 1 and 2 are identified by the estimated coefficients $f_j - c_j$ and c_j, respectively. In other words, appropriate controls for aggregation bias allow unbiased estimates of the behavioral determinants of characteristic or choice proportions, even if they do identify the form of the aggregation bias, itself.

C. Heteroskedasticity and weighting in Goodman's regression

The tests for the neighborhood model and the absence of aggregation bias described in the previous subsections are well-defined only if the estimator of the coefficient variance-covariance matrix has appropriate statistical properties. Similarly, individual parameter estimates identify relevant behavioral determinants only if they can be distinguished statistically from zero. Lastly, linear combinations of parameters are only meaningful when associated with valid confidence

---

β_{10} and β_{20}.

[21] This test does not restrict the treatment of z_i in f_1 and f_2. It requires only that these functions be linear in x_i. More complicated functions of x_i would probably suggest analogous tests. Although rejection would definitively establish the presence of aggregation bias, this test again has limited power. Even if *d* is statistically indistinguishable from zero, it is possible that $\beta_{10}=-\beta_{20}\neq 0$. In this case, aggregation bias would be undetectable, but still present.

intervals.

However, the residual term in equation 10 is heteroskedastic (Goodman (1959, 612)).[22] This must be addressed in order to construct any valid test of statistical significance.[23] Achen and Shively (1995, 47-8) and Lewis (2001, 177) discuss feasible strategies in the context of ecological regression.[24] In addition, White heteroskedasticity-consistent standard errors (Greene (2003, 219-220)) can be employed to provide valid tests of hypotheses regarding parameters, without estimating the structural components of the theoretical residual variances.

Unfortunately, these strategies are rarely employed. Instead, "weighting" is the typical response. Weighting corrects for heteroskedasticity only if the weights for each observation are proportional to the inverse of the residual-specific standard deviation (Greene (2003, 225)). The standard deviations relevant to equation 10 are conveniently invariant to the specification of f_1

---

[22] Heteroskedasticity is inherent in random coefficient models (Greene (2003, 318-9). Achen and Shively (1995, 47-8) and Lewis (2001, 177) note that OLS estimates of equation 14 are unbiased where f_1 and f_2 are constants. In fact, heteroskedasticity does not impose bias on OLS estimators regardless of the forms of f_1 and f_2, if those forms are specified correctly (Greene (2003, 193-5)). King (1997, 65-8) asserts that heteroskedasticity can severely distort inference in ecological regression models. The empirical heteroskedasticity that he discusses may be partially attributable to incomplete specifications of f_1 and f_2, which would incorrectly allocate some of the deterministic component of y_i to the residual. In contrast, Achen and Shively (1995, 47-50 and 128) claim that heteroskedasticity is empirically unimportant. However, they are essentially uninterested in inference (page 58).

[23] As an example, Bourke, DeBats and Phelan (2001, 132)) claim that OLS regressions on aggregate data with f_1 and f_2 specified as constants yield results that appear to be similar to known values based on the underlying microdata. However, their standard errors are wrong because they are not corrected for heteroskedasticity. If the true standard errors were small, the estimates might reject the null hypothesis of the known values. If they were large, the estimates might be statistically "close" to many other values that are substantively quite different from the known values. In other words, the numerical comparison between estimated and true values is uninformative unless scaled by accurate standard errors.

[24] These discussions assume that ϵ_{1i} and ϵ_{2i} are uncorrelated with the disturbances for other areas, ϵ_{1j} and ϵ_{2j}. Autocorrelation (Cho (1998, 145-6)) would introduce additional terms in the residual and require additional corrections to standard errors. Inexplicably, Cho reports OLS standard errors for what is apparently equation 3, without any indication that they have been appropriately corrected.

and $f_{2,}$, because neither appears in its residual. They require only estimates of the variances of ϵ_{1i} and ϵ_{2i} and their covariance, in addition to the known value of x_i.

However, the "conventional weight" in ecological regression practice is the reciprocal of the square root of areal population (Kousser (2001, footnote 23)). This weighting is unrelated to the heteroskedasticity evident in equation 10. It will almost surely compound it.[25]

Instead, the R^2 value from estimates of equation 3 is occasionally offered as evidence of statistical significance (Grofman, Migalski and Noviello (1985, 206)) or, more casually, model performance (Kousser (2001, 111-2)). If this value is from a regression that does not correct for heteroskedasticity, its relationship to the statistical tests of interest – the validity of the neighborhood model, the absence of aggregation bias, and the importance of behavioral determinants – is unknown.

If the R^2 value is from a weighted regression, it is almost surely meaningless, whether or not the weights correct for heteroskedasticity,. Unless the weight is the inverse of a variable that occurs linearly in the original specification, the weighted regression will have no constant term (Greene (2003, 226)). In this case, the R^2 value is not bounded between zero and one and does not represent the proportion of variance in the dependent variable attributable to the explanatory variables (Greene (2003, 36-7)).

---

[25] King (1997, 61-5) and Achen and Shively (1995, 57-61) are critical of weighting by the inverse square root of population. In contrast, Kousser (2001) asserts without proof that it corrects for heteroskedasticity (page 112) and that it yields meaningful changes in the values of ecological regression estimators (page 110). Achen and Shively (1995, 58-9) extend this latter argument. Both are wrong. With the correct deterministic specification, weighted least squares estimators are unbiased and consistent for the behavioral parameters with any weighting scheme that is not correlated with the true residuals, including the equal weights of OLS (Greene (2003, 192-5)). In other words, the incorrect population weights may alter point estimates somewhat, but have no effect on their expected values and distort their standard errors. Kousser (2001) is an example of incorrect standard errors afflicted with both the heteroskedasticity of equation 10 and that imposed by inverse square root of population weights.

D. Summary

This section demonstrates that, in the case of a single application of Goodman's identity, Goodman both over- and under-estimates the efficacy of OLS. He conjectures (1959, 612) that "standard methods of linear regression can be used to estimate" the parameters of this identity. While this method can yield unbiased estimates of these parameters, it cannot, without important modifications, subject them to the necessary significance tests.

With these modifications, however, OLS can provide convincing tests of the neighborhood hypothesis, for the absence of aggregation bias and for the presence of covariates. It can also generate the necessary confidence intervals for parameters and parameter combinations. While there may be formulations of the neighborhood model and aggregation bias that are not contained in the relatively general model analyzed explicitly here, it is likely that appropriate extensions of this analysis will preserve the general conclusions.

With these properties, OLS should be an attractive technique for empirical implementations of a single Goodman's Identity. Ecological inference (King (1997)) is comparable in that it yields consistent estimates and valid hypothesis tests. It is superior in that it is more efficient: It explicitly incorporates the restrictions that $0 \leq \beta_{1i} \leq 1$ and $0 \leq \beta_{2i} \leq 1$. Consequently, its estimates are guaranteed to be feasible and should be more precise. This additional precision is apparent in the simulations of Silva de Mattos and Veiga (2004).[26]

However, OLS restricts the number of covariates in z_i only to the extent that degrees of freedom must be sufficient to allow estimation. In contrast, the treatment of covariates in

[26] These simulations restrict f_1 and f_2 to be constants. Within that context, they also demonstrate both the unbiasedness of OLS and the consistency of ecological inference.

ecological inference is computationally burdensome (King (2003, page 49)).[27] Given the interaction terms that must appear in the unrestricted model of equation 10, specifications of f_1 and f_2 with multiple covariates may not be tractable within this method. If so, tests of the neighborhood hypothesis and aggregation bias may be infeasible.

Other estimation techniques should be unambiguously less attractive than OLS for most purposes. As examples, estimates from the model of King, Rosen and Tanner (1999) may be biased (Silva de Mattos and Veiga (2004)), are difficult to calculate, and appear to be substantially more burdensome in the presence of covariates. Again, tests of the neighborhood hypothesis and aggregation bias may be infeasible. The three estimators proposed in Grofman and Merrill (2004) have no known relationships to the underlying parameters, no significance measures and no known extensions to covariates. The neighborhood hypothesis, aggregation bias and covariates cannot be specified in the contexts of their models, much less tested.

III. Goodman's regression with multiple groups and characteristics

Both the positive and negative aspects of Goodman's regression are exaggerated when the number of groups in the population is greater than two.[28] All parameters in the linear multivariate model analogous to that of equation 12 are identified, and the test for aggregation bias has much

[27] King (1997, page 170) suggests that covariates might be addressed by estimating β_{1i} and β_{2i} with ecological inference under the assumption that f_1 and f_2 are constants, and then regressing these estimates on covariates. Redding and James (2001) is an example. This strategy implicitly acknowledges that these covariates should have appeared in the initial specification of f_1 and f_2. The consequences of this misspecification are, predictably, difficult to ascertain (Adolph and King (2003), Adolph, King, Herron and Shotts (2003) and Herron and Shotts (2003a, 2003b)).

[28] Achen and Shively (1995, 34-38 and 129-131) provide a brief discussion of Goodman's identity and regression in the context of transition matrices with more than two electoral choices.

greater power than in the two-group case. However, heteroskedasticity is more complicated.

The case with three groups illustrates these points. Augment the notation of section I as follows:

x_{1i} = the proportion of the population in area *i* that belongs to group 1,

x_{2i} = the proportion of the population in area *i* that belongs to group 2,

$x_{3i}=1-x_{1i}-x_{2i}$ = the proportion of the population in area *i* that belongs to group 3, and

β_{3i} = the proportion of group 3 in area *i* with the characteristic or making the choice at issue.

The analogues to the identities in equations 1 and 2 are then

$$\begin{aligned} y_i &\equiv \beta_{1i}x_{1i} + \beta_{2i}x_{2i} + \beta_{3i}\left[1 - x_{1i} - x_{2i}\right] \\ &\equiv \beta_{3i} + \left[\beta_{1i} - \beta_{3i}\right]x_{1i} + \left[\beta_{2i} - \beta_{3i}\right]x_{2i}. \end{aligned} \tag{15}$$

In this case, the analogue to the linear multivariate model of equations 7 and 12 is

$$\beta_{ri} = f_r\left(x_{ri}, z_{ij}\right) + \varepsilon_{ri} = \beta_r + \beta_{r0}x_{li} + \sum_{j=1}^{k}\beta_{rj}z_{ij} + \varepsilon_{ri}, \tag{16}$$

where *r*=1,...,3 identifies the group. The substitution of equation 16 into equation 15 yields

$$\begin{aligned} y_i = &\left[\beta_3 + \beta_{30}\right] + \left[\beta_1 - \beta_3 - 2\beta_{30}\right]x_{1i} + \left[\beta_2 - \beta_3 - 2\beta_{30}\right]x_{2i} \\ &+ \sum_{j=1}^{k}\beta_{3j}z_{ij} + \left[\beta_{10} + \beta_{30}\right]x_{1i}^2 + \left[\beta_{20} + \beta_{30}\right]x_{2i}^2 + 2\beta_{30}x_{1i}x_{2i} \\ &+ \sum_{j=1}^{k}\left[\beta_{1j} - \beta_{3j}\right]z_{ij}x_{1i} + \sum_{j=1}^{k}\left[\beta_{2j} - \beta_{3j}\right]z_{ij}x_{2i} \\ &+ \left[\varepsilon_{1i}x_{1i} + \varepsilon_{2i}x_{2i} + \varepsilon_{3i}\left[1 - x_{1i} - x_{2i}\right]\right]. \end{aligned} \tag{17}$$

The model of equations 15 and 16 contains 2+*k* parameters in addition to those in equation 13, for a total of 3[2+*k*] parameters. However, the regression of equation 17 estimates 3+*k* coefficients in addition to those in equation 14. The additional coefficient is attributable to the interaction term in $x_{1i}x_{2i}$, which conventional practice would be ordinarily, if incorrectly, omit.

As a consequence of this interaction term, the number of coefficients in the three-group regression of equation 17 equals the number of underlying parameters. All are therefore identified, in contrast to the two-group regression of equation 13.[29] As in equation 13, significance tests on the estimated values for β_{1j}, β_{2j} and β_{3j} indicate whether covariates are important.

Equation 17 also provides complete tests for the neighborhood model and for the presence of aggregation bias. The neighborhood model implies 2*k*+4 restrictions on the regression of equation 17. The requirements that $\beta_1=\beta_2=\beta_3$ and $\beta_{10}=\beta_{20}=\beta_{30}$ imply four restrictions: the absolute values of the coefficients on x_{1i}, x_{2i}, x_{1i}^2, x_{2i}^2 and $x_{1i}x_{2i}$ should be identical. The requirement that $\beta_{1j}=\beta_{2j}=\beta_{3j}$ implies 2*k* restrictions: the coefficients on $z_{ij}x_{1i}$ and $z_{ij}x_{2i}$ should all equal zero. The failure of any of these restrictions would invalidate the neighborhood model.

Aggregation bias is present if $\beta_{10}\neq 0$, $\beta_{20}\neq 0$ or $\beta_{30}\neq 0$. The null hypothesis that it is absent, $\beta_{10}=\beta_{20}=\beta_{30}=0$, implies three restrictions: The coefficients on x_{1i}^2, x_{2i}^2 and $x_{1i}x_{2i}$ should all be equal to zero. The failure of any of these restrictions indicates that aggregation bias is present.

This test is more powerful than that in the case of two groups because the three restrictions can be simultaneously satisfied if and only if aggregation bias is truly absent, $\beta_{10}=\beta_{20}=\beta_{30}=0$. For

---

[29] The coefficients on z_{ij} identify β_{3j}. With these results, the coefficients on $z_{ij}x_{1i}$ and $z_{ij}x_{2i}$ identify β_{1j} and β_{2j}, respectively. The coefficient on $x_{1i}x_{2i}$ identifies β_{30}. With this result, the coefficients on x_{1i}^2 and x_{2i}^2 identify β_{10} and β_{20}, respectively and the constant identifies β_3. With this last result and the identification of β_{30}, the coefficients on x_{1i} and x_{2i} identify β_1 and β_2, respectively.

example, if $\beta_{20}=-\beta_{30}\neq 0$, the second restriction would hold but the third would fail. Therefore, the failure of any one of these restrictions indicates unambiguously that aggregation bias is present.

At the same time, the last line of equation 17 demonstrates that the residual in this regression contains three random components, rather than the two of equation 13. The variance of the random component for each area therefore depends on the population proportions of all three groups in that area, the variances of the three group-specific random components and the three unique covariances among them. Regression estimates must correct for the consequent heteroskedasticity in order to test any of the restrictions implied by the neighborhood hypothesis or the hypothesis of aggregation bias.

As the number of groups increases beyond three, the number of interaction terms between x_{ki} and x_{mj} proliferates more rapidly than the number of underlying parameters. Consequently, models with $R>3$ groups are actually overidentified: They are based on $2+k$ parameters for each group, or $R[2+k]$ parameters in all. However, they estimate $R[2+k]+\frac{1}{2}R[R-3]$ coefficients. Therefore, $\frac{1}{2}R[R-3]$ restrictions are necessary in order to ensure that the estimates are consistent with the underlying model. The effect of these restrictions on the explanatory power of the regression provides a test of the underlying specification of Goodman's Identity.[30]

The number of alternative characteristics or choices has many fewer implications for Goodman-based estimation than does the number of groups in the population. The identities of equations 1 and 15 do not depend on this number, and are therefore valid regardless of its value. Consequently, the estimations of equations 13 and 17 do not depend on the number of alternatives.

[30] Proofs of these claims and those in the remainder of this section are available from the author.

Analogous identities and estimating equations would apply to any additional alternatives. However, they would ordinarily be based on parameters that were specific to these alternatives. Identification in each would be based on the results above. Multiple characteristics or choices would provide additional leverage for identification across equations only if the underlying behavioral theory indicated that equations for different alternatives shared common parameters.

In any case, with random coefficient specifications that are linear in the same variables for all groups and all C alternatives, the number of informative equations is always $C-1$. All equations are constrained by the requirements that the proportions of the population possessing each characteristic or making each choice must sum to one, as must the corresponding proportions within each group. Consequently, the last equation is always implied by the first $C-1$equations.

An election with two candidates provides an example: The fraction of a group in the electorate that chooses to cast its votes for one candidate chooses not to cast its votes for the other. If y_i represents the proportion of votes cast for the first candidate, $1-y_i$ represents the proportion of votes cast for the second. Equation 13, with both sides multiplied by -1 and augmented by one, expresses the relationship between the vote share of the second candidate and the explanatory variables.

As is evident, estimation of this equation would be uninformative. It depends on the same set of parameters as in equation 13. Moreover, neither this transformation of equation 13 nor its combination with the original version are sufficient to identify β_1, β_{10}, β_2 and β_{20}.

Ecological estimation with more than two groups and more than two choices is known generically as the "R×C model". This section demonstrates that once again, OLS, properly specified, should be a relatively attractive estimation technique for this model. Estimates are

unbiased and valid standard errors are available. With more than two groups, identification is complete and may imply testable restrictions. Tests of the neighborhood hypothesis and aggregation bias are straightforward.

Only two other estimation techniques are available for the R×C problem: ecological inference (King (1997, chapter 15)) and the binomial-beta hierarchical model (Rosen, Jiang, King and Tanner (2001)). Both are computationally burdensome when f_r is constant for all r.[31] More complicated specifications of f_r would compound the difficulties. The questions of how the restrictions implied by the neighborhood model or the absence of aggregation bias would be imposed in these techniques are, as of now, not only unanswered, but unasked.

IV. Goodman's regression with two Goodman's identities

Kousser (2001, 110) asserts that the estimation of transition matrices relating partisan voting patterns in two successive elections and the comparison of voting patterns across two different racial or ethnic groups in the same election have been the two principle applications of Goodman's regression. These two applications have dramatically different statistical characters.

The transition matrix problem is described adequately by a single application of Goodman's identity. The electorate in the first election is exhaustively divided into groups of known size voting for each of the available alternatives, and a residual group of known size choosing abstention. The dependent variables measure the known proportions of the electorate in each group in the second election. The results of the previous sections apply to the question of how

[31] King (1997, chapter 15) suggests a simplification relying on iterative applications of the bivariate truncated normal distribution. This strategy may be subject to biases (Ferree (2004)).

electoral choices in the first election are related to choices in the second.

In contrast, a single application of equation 1 is usually insufficient to compare the voting patterns of two groups defined by some characteristic other than their previous electoral behavior. In the unusual circumstance where the allocation of actual voters across racial or ethnic groups is known, equation 1 applies directly, with x_i defined as the proportion of all voters constituted by members of group 1 and y_i defined as the proportion of votes received by a candidate. In this case, β_{1i} and β_{2i} would represent the shares of group 1 and group 2 voters choosing that candidate. These shares would be interpreted as indicating the preferences of voters within each group.

However, in most circumstances, the composition of the electorate is known rather than that of the voters. With x_i defined as the proportion of the electorate constituted by members of group 1, β_{1i} would represent the share of the group 1 electorate that chooses the candidate in question. In this case, it would incorporate the group 1 abstention rate, as well as the rate at which group 1 members who vote choose that candidate. Either of these rates may be parameters of interest, but their combination is not, of itself.[32]

The essential distinction here is that the parameters of interest refer to the behavior of a subset of the population whose composition is unknown. In this case, the application of Goodman's approach to aggregate data requires the separate identification of turnout propensities and choices made by actual voters. This requires two applications of Goodman's identity (King (1997, 68-

[32] This is obvious in a simple example. If all of the group 1 electorate votes and 70% of its members choose candidate 2, only 30% prefer candidate 1. The same share prefers candidate 1 if 50% of the group 1 electorate abstains and only 20% choose candidate 2. In either case, β_{1i} would be .3. However, group 1 voters would prefer candidate 2 in the first case and candidate 1 in the second.

71)).[33]

The turnout parameters appear in the identity

$$T_i \equiv \beta_{1i} x_i + \beta_{2i}\left[1 - x_i\right], \qquad \textbf{(18)}$$

where

T_i = the observed turnout rate in area *i*, the ratio of the number of votes cast to the number of potential voters,

β_{1i} = the unobserved turnout rate among group 1 voters in area *i*, the ratio of the number of votes cast by group 1 voters to the number of potential group 1 voters, and

β_{2i} = the unobserved turnout rate among group 2 voters in area *i*, the ratio of the number of votes cast by group 2 voters to the number of potential group 2 voters.

β_{1i} and β_{2i} can be estimated through Goodman's regression as described in section II.

The choices made by actual voters from the two groups appear in the second identity

$$y_{1i} \equiv \lambda_{1i} w_{1i} + \lambda_{2i} w_{2i}, \qquad \textbf{(19)}$$

where

y_{1i} = the observed ratio of the number of votes received by candidate 1 to the size of the electorate in precinct *i*.

w_{1i} = the unobserved ratio of votes cast by group 1 voters to the size of the electorate in area *i*,

w_{2i} = the unobserved ratio of votes cast by group 2 voters to the size of the electorate in area *i*,

---

[33] Grofman, Migalski and Noviello (1985, 204) and Grofman, Handley and Niemi (1992, 86) are examples of work in which this distinction, and its consequences discussed below, are ignored. The approach here treats voting choice as conditional on turnout choice. This distinction is somewhat artificial. A formulation such as that of Sanders (1998), in which abstention is an intermediate "voting" choice when voters are approximately indifferent between candidates, is more natural. However, Sanders (1998) implements this formulation with microdata, and does not explore its aggregation properties. As of now, Goodman's identity is the only available basis for the interpretation of aggregate voting data.

λ_{1i} = the unobserved ratio of votes cast by group 1 voters for candidate 1 to the number of votes cast by group 1 voters in area *i*, and

λ_{2i} = the unobserved ratio of votes cast by group 2 voters for candidate 1 to the number of votes cast by group 2 voters in area *i*.[34]

The regression analogue to equation 19 is not in the form of Goodman's regression because it requires two explanatory variables, w_{1i} and w_{2i}. Regardless, it cannot estimate λ_{1i} and λ_{2i} directly because both of these variables are unobserved.

However, by the above definitions $w_i=\beta_{1i}x_i$ and $1-w_i=\beta_{2i}x_i$. Therefore, equation 19 becomes

$$y_{1i} \equiv \lambda_{1i}\beta_{1i}x_i + \lambda_{2i}\beta_{2i}\left[1 - x_i\right]. \quad \textbf{(20)}$$

Equation 20 establishes the identity between the observed shares of group 1 and group 2 members in the electorate and the observed ratio of candidate 1 votes to the size of the electorate. It combines the two identities of equations 18 and 19.

As in section I, this identity requires the random coefficients assumption in order to introduce random components and a parameterization of the variations across areas in the determinants of group-specific vote choices. This assumption for λ_{1i} and λ_{2i} is analogous to that for β_{1i} and β_{2i} in equation 7,

$$\lambda_{1i} = g_1(x_i, z_{1i}) + \nu_{1i} \text{ and } \lambda_{2i} = g_2(x_i, z_{2i}) + \nu_{2i}. \quad \textbf{(21)}$$

Again, ν_{1i} and ν_{2i} have expected values equal to zero and are uncorrelated with x_i, z_{1i} and z_{2i}. Incorporating equations 7 and 21, and dropping function arguments for clarity, equation 20 becomes

$$y_{1i} \equiv \left[g_1 + \nu_{ii}\right]\left[f_1 + \varepsilon_{1i}\right]x_i + \left[g_2 + \nu_{2i}\right]\left[f_2 + \varepsilon_{2i}\right]\left[1 - x_i\right], \quad \textbf{(22)}$$

[34] β_{1i}, β_{2i}, λ_{1i} and λ_{2i}, here correspond to β_i^b, β_i^w, λ_i^b and λ_i^w in King (1997).

or

$$y_{1i} \equiv g_2 f_2 + \left[g_1 f_1 - g_2 f_2\right] x_i + \left[x_i\left[g_1 \varepsilon_{1i} + f_1 \nu_{1i}\right] + \left[1 - x_i\right]\left[g_2 \varepsilon_{2i} + f_2 \nu_{2i}\right]\right] + \left[x_i \varepsilon_{1i} \nu_{1i} + \left[1 - x_i\right] \varepsilon_{2i} \nu_{2i}\right]. \quad \textbf{(23)}$$

The comparison between equations 10 and 23 demonstrates two radical differences. First, f_1, g_1, f_2, and g_2, the deterministic components of y_{1i}, enter equation 23 only nonlinearly. Therefore, individual parameters appear only in products.

Second, y_{1i} in equation 23 depends on four disturbances, rather than two as in the case of a single Goodman's identity, or one as in conventional regression analysis. The residual terms in the second line of equation 23 are linear combinations of random components with expected values equal to zero, as in equation 10. However, the third line of equation 23 contains two nonlinear combinations of random components. Their expected values are

$$E\left(x_i \varepsilon_{1i} \nu_{1i}\right) = x_i \sigma_{1\varepsilon\nu} \text{ and } E\left(\left[1 - x_i\right] \varepsilon_{2i} \nu_{2i}\right) = \left[1 - x_i\right] \sigma_{2\varepsilon\nu}, \quad \textbf{(24)}$$

where $\sigma_{1\epsilon\nu}$ and $\sigma_{2\epsilon\nu}$ are the covariances between ϵ_{1i} and ν_{1i} and between ϵ_{1i} and ν_{1i}, respectively. These covariances enter into the expected value of the dependent variable,

$$E\left(y_{1i}\right) \equiv g_2 f_2 + \left[g_1 f_1 - g_2 f_2\right] x_i + x_i \sigma_{1\varepsilon\nu} + \left[1 - x_i\right] \sigma_{2\varepsilon\nu}. \quad \textbf{(25)}$$

Unobserved characteristics of an area that affect turnout for a group within that area are likely to be related to voting preferences for that group, and vice versa. For example, if members of a group have an idiosyncratically strong preference for a particular candidate, this preference may stimulate an idiosyncratically high turnout. Therefore, the covariances in equations 24 and 25

between the random components of turnout and vote share will ordinarily be non-zero.[35]

In consequence, identification is much more difficult in contexts that require two rather than one application of Goodman's identity. It is impossible in all common cases. This includes the specification that is almost universal in the analysis of voting choices by two different racial or ethnic groups in a single election.

This specification requires all deterministic components to be constants: $f_1=\beta_1$, $g_1=\lambda_1$, $f_2=\beta_2$ and $g_2=\lambda_2$ (see footnote 15). The deterministic component of equation 23 is then

$$g_2 f_2 + \left[g_1 f_1 - g_2 f_2\right]x_i = \beta_2\lambda_2 + \left[\beta_1\lambda_1 - \beta_2\lambda_2\right]x_i . \quad \textbf{(26)}$$

The expression to the right of the last equality in equation 25 contains only two terms, a constant and a linear term in x_i. Therefore, equation 3 would be the corresponding estimating equation.[36]

However, the expected values of the estimated coefficients are[37]

$$E\left(b_0\right) = \beta_2\lambda_2 + \sigma_{2\varepsilon v} \quad \textbf{(27)}$$

and

$$E\left(b_1\right) = \left[\beta_1\lambda_1 + \sigma_{1\varepsilon v}\right] - \left[\beta_2\lambda_2 + \sigma_{2\varepsilon v}\right]. \quad \textbf{(28)}$$

Obviously, the two estimated coefficients are insufficient to identify the six parameters upon

---

[35] The discussion here assumes that f_1, g_1, f_2 and g_2 are correctly specified. Any variables incorrectly omitted from these functions will be incorporated in the residuals. Nonzero empirical covariances will also result if variables omitted from different functions are correlated with each other.

[36] Conversely, empirical implementations of equation 3 in contexts where the composition of the population at issue is unknown can only be understood as either imposing the assumptions underlying equation 26, or imposing the neighborhood model with one of g_2 or f_2 constant and the other linear in x_i alone.

[37] Proofs of this and all subsequent statements in this section are available from the author.

which they are based. In other words, the typical application of Goodman's regression to the analysis of voting choices by two different demographic groups is fatally unidentified.[38]

The same is true for other simple specifications. The most parsimonious specification allowing for aggregation bias defines f_1, g_1, f_2 and g_2 as linear in x_i alone:

$$f_1 = \beta_1 + \beta_{10} x_i,\ g_1 = \lambda_1 + \lambda_{10} x_i,$$
$$f_2 = \beta_2 + \beta_{20} x_i \text{ and } g_2 = \lambda_2 + \lambda_{20} x_i.$$

The deterministic component of equation 23 would then be

$$\begin{aligned} g_2 f_2 + \left[g_1 f_1 - g_2 f_2\right] x_i = {} & \beta_2\lambda_2 + \left[\beta_2\lambda_{20} + \beta_{20}\lambda_2 + \beta_1\lambda_1 - \beta_2\lambda_2\right] x_i \\ & + \left[\beta_{20}\lambda_{20} + \beta_1\lambda_{10} + \beta_{10}\lambda_1 - \beta_2\lambda_{20} - \beta_{20}\lambda_2\right] x_i^2 \\ & + \left[\beta_{10}\lambda_{10} - \beta_{20}\lambda_{20}\right] x_i^3. \end{aligned}$$

The corresponding OLS equation is

$$y_{1i} = b_0 + b_1 x_i + b_2 x_i^2 + b_3 x_i^3,$$

with only four coefficients. Their expected values are

$$E(b_0) = \beta_2\lambda_2 + \sigma_{2\varepsilon\nu}, \quad \textbf{(29)}$$

$$E(b_1) = \beta_2\lambda_{20} + \beta_{20}\lambda_2 + \left[\beta_1\lambda_1 + \sigma_{1\varepsilon\nu}\right] - \left[\beta_2\lambda_2 + \sigma_{2\varepsilon\nu}\right], \quad \textbf{(30)}$$

$$E(b_3) = \beta_{20}\lambda_{20} + \beta_1\lambda_{10} + \beta_{10}\lambda_1 - \beta_2\lambda_{20} - \beta_{20}\lambda_2, \quad \textbf{(31)}$$

[38] Zax (2005) demonstrates that "double regression", a common prescription for this problem, fails utterly to resolve it.

and

$$E(b_3) = \beta_{10}\lambda_{10} - \beta_{20}\lambda_{20}. \tag{32}$$

These four terms contain ten parameters. Again, none are identified.

Similarly, the simplest specification allowing for covariates would specify f_1, g_1, f_2 and g_2 as linear in a single covariate z_i:

$$\begin{aligned} f_1 &= \beta_1 + \beta_{11}z_i,\ g_1 = \lambda_1 + \lambda_{11}z_i, \\ f_2 &= \beta_2 + \beta_{21}z_i \text{ and } g_2 = \lambda_2 + \lambda_{21}z_i. \end{aligned} \tag{33}$$

The deterministic component of equation 23 would then be

$$\begin{aligned} g_2 f_2 + [g_1 f_1 - g_2 f_2]x_i = {} & \beta_2\lambda_2 + [\beta_1\lambda_1 - \beta_2\lambda_2]x_i + [\beta_{21}\lambda_2 - \beta_2\lambda_{21}]z_i + \beta_{21}\lambda_{21}z_i^2 \\ & + [\beta_{11}\lambda_1 + \beta_1\lambda_{11} - \beta_{21}\lambda_2 - \beta_2\lambda_{21}]z_i x_i + [\beta_{11}\lambda_{11} - \beta_{21}\lambda_{21}]z_i^2 x_i. \end{aligned}$$

OLS would estimate six terms, a constant and coefficients attached to the variables x_i, z_i, z_i^2, $z_i x_i$ and $z_i^2 x_i$. However, their expected values contain ten parameters: the four intercepts and four slopes in equation 33 and the two covariances in equation 24. Again, none of the parameters are identified.

Identification here is possible, if at all, either in more restricted or more complicated models. As examples of the first option, the three models already discussed could be identified with sufficient *a priori* restrictions on parameter values. Of course, this would be an empty achievement unless these restrictions were consistent with a convincing behavioral model.

As an example of the second option, a general linear specification would define f_1 and f_2 as in equation 12 and g_1 and g_2 analogously:

$$g_1(x_i, z_i) = \lambda_1 + \lambda_{10} x_i + \sum_{j=1}^{k} \lambda_{1j} z_{ij} \text{ and } g_2(x_i, z_i) = \lambda_2 + \lambda_{20} x_i + \sum_{j=1}^{k} \lambda_{2j} z_{ij}.$$

With the two covariance terms of equation 24, this specification contains 4[k+2]+2 parameters. The deterministic component of equation 23 would contain contain [k+2]^2 discrete terms. Accordingly, OLS would estimate one constant and [k+2]^2−1 slopes.

Identification is therefore impossible with two or fewer covariates, $k \leq 2$. If $k \geq 3$, the number of estimates exceeds the number of parameters. This is a necessary condition for identification. It is not sufficient, because, as in the examples of equations 27 through 32, parameters occur in nonlinear combinations in the expected values of empirical estimates. However, in fortuitous circumstances these models may even be over-identified, suggesting the possibility of specification tests.

In principle, it may be possible to identify a wider class of models through careful treatment of the heteroskedasticity evident in equation 23. In contrast to the case of a single application of Goodman's Identity, the residual here depends on all parameters in the deterministic component of y_{1i}. Therefore, empirical heteroskedasticity may be informative about their values.

Unfortunately, the variances of the residual terms also depend on the four disturbance-specific variances and the six unique covariances in the variance-covariance matrix of the disturbances ϵ_{1i}, ν_{1i}, ϵ_{2i} and ν_{2i}. Moreover, they contain terms in the variances and covariances of products of disturbances, such as $x_i^2 V(\epsilon_{1i}\nu_{1i})$ and $x_i[1-x_i]COV(\epsilon_{1i}\nu_{1i}, \epsilon_{2i}\nu_{2i})$. These additional terms depend on higher-order parameters of the four-dimensional disturbance distribution that are not contained in the disturbance variance-covariance matrix.

These considerations suggest that the strategy of identification through heteroskedasticity is

not promising. However, heteroskedasticity must still be addressed in order to perform the inference necessary to validate and interpret any model. For most applications, White heteroskedasticity-consistent standard errors (Greene (2003, 219-220)) will probably prove to be more practical than structural estimation of the components in the theoretical residual variances.

In sum, the only contexts in which OLS regression has any hope of recovering the underlying behavioral parameters from two applications of Goodman's identity are in models that are heavily restricted or relatively rich, specifying at least several covariate determinants of the behavior at issue. As in the case of a single applications of Goodman's identity, these covariates must be interacted with the population proportion x_i in order to avoid the inadvertent imposition of the neighborhood model. Moreover, if covariates affect both the propensity of group *r* members to select into the subpopulation for which the behavior of interest is relevant, g_r, and the propensity of group members to choose that behavior, f_r, they must be appropriately interacted with each other, as well.

Models that fulfill these requirements appeared to be absent from the literature of the social sciences. The unfortunate corollary is that most, if not all extant regression statistics for contexts involving two applications of Goodman's Identity are worthless: They have no known relationships to the parameters they purport to estimate. The optimistic response is that, with appropriate construction, these contexts may invite the estimation of empirical models that are much more ambitious and intriguing than previously attempted.

For the moment, only ecological inference (King (1997)) correctly specifies equation 22. Consequently, it is the only technique available that identifies the parameters in g_1, f_1, g_2 and f_2. However, computational limitations currently restrict the specifications of these functions. While

these restrictions are likely to be relaxed as computational power and techniques improve, dramatic improvements will be necessary in order to accommodate truly flexible specifications. Parallel efforts to construct plausible models that are correctly-specified and identifiable in regression may therefore be worthwhile.

V. Conclusion

This paper demonstrates that regression-based applications of Goodman's identity can be much more effective than previously understood. In contexts where a single application of Goodman's identity is sufficient to characterize the behavior at issue, OLS estimates of the generalized Goodman's regression in equation 13 are unbiased. They are also heteroskedastic, but corrections are feasible.

With these corrections, OLS estimators provide valid statistical tests for the neighborhood model, aggregation bias, and the significance of covariates. Moreover, identification in these models improves as the number of groups in the population increases. These results, coupled with the flexibility and tractability of OLS, suggest that correctly specified models should be valuable tools in the analysis of single applications of Goodman's regression, notwithstanding the risk of estimates outside the bounds of zero and one, and the ingenuity embodied in recent attempts to provide improved estimators (King (1997), King, Rosen and Tanner (1999) and Lewis (2004) as examples).

Contexts where the proportions of groups that engage in the behavior at issue are unknown require two applications of Goodman's identity. In these contexts, individual estimates from Goodman's regression do not identify individual behavioral parameters. Identification may be

possible if models contain sufficiently numerous restrictions or explanatory variables, but these options have not been explored. With, again, appropriate corrections for heteroskedasticity, valid tests may be available for the neighborhood model, aggregation bias and the significance of covariates.

This paper also demonstrates that current practice in the application of Goodman's regression typically fails to achieve any of these results. Most empirical exercises specify the implied empirical models incorrectly, ignore heteroskedasticity and offer neither hypothesis tests nor confidence intervals, valid or otherwise. Instead, they are contaminated with arbitrary weights that exacerbate heteroskedasticity, and justified with R^2 values that are meaningless. Clearly, more than 50 years after it was first promulgated, Goodman's Identity has yet to be fully appreciated.

References

Achen, Christopher H. and W. Phillips Shively (1995) Cross-Level Inference, The University of Chicago Press, Chicago.

Adolph, Christopher and Gary King (2003) "Analyzing second-stage ecological regressions: Comment on Herron and Shotts", Political Analysis, Vol. 11, No. 1, Winter, 65-76.

Adolph, Christopher, Gary King, Michael C. Herron and Kenneth W. Shotts (2003) "A consensus on second-stage analyses in ecological inference models", Political Analysis, Vol. 11, No. 1, Winter, 86-94.

Bourke, Paul, Donald DeBats and Thomas Phelan (2001) "Comparing individual-level voting returns with aggregates: A historical appraisal of the King solution", Historical Methods, Vol. 34, No. 3, Summer, 127-134.

Cho, Wendy K. Tam (1998) "Iff the assumption fits ...: A comment on the King ecological inference solution", Political Analysis, Vol. 7., 143-164.

Ferree, Karen E. (2004) "Iterative approaches to R × C ecological inference problems: Where they can go wrong and one quick fix", Political Analysis, Vol. 12, No. 2, Spring, 143-159.

Freedman, David A., Stephen P. Klein, Jerome Sacks, Charles A. Smyth and Charles G. Everett (1991) "Ecological regression and voting rights", Evaluation Review, Vol. 15, No. 6, December, 673-711.

Goodman, Leo A. (1953) "Ecological regressions and behavior of individual", American Sociological Review, Vol. 18, No. 6, December, 663-664.

Goodman, Leo A. (1959) "Some alternatives to ecological correlation", American Journal of Sociology, Vol. 64, No. 6, May, 610-625.

Greene, William H. (2003) Econometric Analysis, Fifth Edition, Prentice Hall, Upper Saddle River.

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

Grofman, Bernard and Samuel Merrill (2004) "Ecological regression and ecological inference", chapter 5 in King, Gary, Ori Rosen and Martin Tanner, eds., Ecological Inference: New Methodological Strategies, Cambridge University Press, New York, 123-143.

Grofman, Bernard, Michael Migalski and Nicholas Noviello (1985) "The 'Totality of the Circumstances Test' in Section 2 of the 1982 Extension of the Voting Rights Act: A social science perspective", Law & Policy, Vol. 7, No. 2, April, 199-223.

Hanushek, Eric A., John E. Jackson and John F. Kain (1974) “Model specification, use of aggregate data, and the ecological fallacy”, Political Methodology, Winter, 89-107.

Herron, Michael C. and Kenneth W. Shotts (2003a) “Cross-contamination in EI-R: Reply”, Political Analysis, Vol. 11, No. 1, Winter, 77-85.

Herron, Michael C. and Kenneth W. Shotts (2003b) “Using ecological inference point estimates as dependent variables in second-stage linear regressions”, Political Analysis, Vol. 11, No. 1, Winter, 44-64.

Irwin, Laura and Allan J. Lichtman (1976) "Across the great divide: Inferring individual level behavior from aggregate data", Political Methodology, Vol. 3, No. 4, 411-439.

Judge, George G., Douglas J. Miller and Wendy K. Tam Cho (2004) “An information theoretic approach to ecological estimation and inference”, chapter 7 in King, Gary, Ori Rosen and Martin Tanner, eds., Ecological Inference: New Methodological Strategies, Cambridge University Press, New York, 162-187.

King, Gary (1997) A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior From Aggregate Data, Princeton University Press, Princeton.

King, Gary (2003) EI: A Program for Ecological Inference, http://gking.harvard.edu/files/ei.pdf.

King, Gary, Ori Rosen and Martin A. Tanner (1999) “Binomial-beta hierarchical models for ecological inference”, Sociological Methods & Research, Vol. 28, No. 1, August, 61-90.

Klein, Stephen P., Jerome Sacks and David A. Freedman (1991) “Ecological regression *versus* the secret ballot”, Jurimetrics, Vol. 31, 393-413.

Kousser, J. Morgan (2001) “Ecological inference from Goodman to King”, Historical Methods, Vol. 34, No. 3, Summer, 101-126.

Lewis, Jeffrey B. (2001) "Understanding King's ecological inference model: A method-of--moments approach", Historical Methods, Fall, Vol. 34, No. 4, 170-188.

Lewis, Jeffrey B. (2004) “Extending King’s ecological inference model to multiple elections using Markov Chain Monte Carlo”, chapter 4 in King, Gary, Ori Rosen and Martin Tanner, eds., Ecological Inference: New Methodological Strategies, Cambridge University Press, New York, 97-122.

Lichtman, Allan J. (1974) "Correlation, regression, and the ecological fallacy: A critique", Journal of Interdisciplinary History, Vol. 4, No. 3, Winter, 417-433.

Lichtman, Allan J. (1991) “Passing the test: Ecological regression analysis in the Los Angeles County case and beyond” Evaluation Review, Vol. 15, No. 6, December, 770-799.

Quinn, Kevin M. (2004) “Ecological inference in the presence of temporal dependence”, chapter 9 in King, Gary, Ori Rosen and Martin Tanner, eds., Ecological Inference: New Methodological Strategies, Cambridge University Press, New York, 207-232.

Redding, Kent and David R. James (2001) "Estimating levels and modeling determinants of black and white voter turnout in the South, 1880-1912", Historical Methods, Fall, Vol. 34, No. 4, 141-158.

Rivers, Douglas (1998) “Review of ‘A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data”, The American Political Science Review, Vol. 92, No. 2, June, 442-443.

Robinson, W.S. (1950) “Ecological correlations and the behavior of individuals”, American Sociological Review, Vol. 15, No. 3, June, 351-357.

Rosen, Ori, Wenxin Jiang, Gary King and Martin A. Tanner (2001) “Bayesian and frequentist inference for ecological inference: the R × C case”, Statistica Neerlandica, Vol. 55, No. 2, July, 134-156.

Sanders, Mitchell S. (1998) “Unified models of turnout and vote choice for two-candidate and three-candidate elections”, Political Analysis, Vol. 7, 89-116.

Silva de Matos, Rogerio and Alvaro Veiga, “A structured comparison of the Goodman regression, the truncated normal, and the binomial-beta hierarchical methods for ecological inference”, chapter 15 in King, Gary, Ori Rosen and Martin Tanner, eds., Ecological Inference: New Methodological Strategies, Cambridge University Press, New York, 351-382.

Voss, D. Stephen (2004) “Using ecological inference for contextual research”, chapter 3 in King, Gary, Ori Rosen and Martin Tanner, eds., Ecological Inference: New Methodological Strategies, Cambridge University Press, New York, 69-96.

Zax, Jeffrey S. (2005) “The statistical properties and empirical performance of double regression”, Political Analysis, Vol. 13, No. 1, January, 57-76.