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

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

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

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

[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

( )

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

[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.

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

[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).

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

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

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 ab>Tu7dmonstrates that t3 266.34 513.66 a6em

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

[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

( ) [( ) ( )] [[ ]]

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

[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

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

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

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

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 re

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*

[22] Heteroskedasticity is inherent in random co

[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

[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) prom131)

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

[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)).

[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

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

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

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 .7(tn.dcf t)Tj36681

[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

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

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

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

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,

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

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

[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

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

and

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

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

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

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.

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

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