SIXTY YEARS OF GOODMAN'S IDENTITY:

ITS UNRECOGNIZED IMPLICATIONS FOR

REGRESSION-BASED INFERENCE

Jeffrey S. Zax

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

23 January 2018

SIXTY YEARS OF GOODMAN'S IDENTITY:

ITS UNRECOGNIZED IMPLICATIONS FOR REGRESSION-BASED INFERENCE

ABSTRACT

This paper examines the implications of Goodman's identity for estimation and inference using linear regression. Estimation requires the assumption of either random coefficients or measurement error. Under the former, regression is surprisingly potent: it 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, typical regression estimates of Goodman's identity do not exploit these capabilities. Instead, most implementations unwittingly impose the neighborhood model, weight incorrectly and offer meaningless R^2 values as "validation".

Keywords: Goodman's Identity, ecological regression, neighborhood model, aggregation bias

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 many regression-based applications.[1] However, the implications of Goodman's identity for these applications have not been thoroughly explored.

This paper demonstrates that regression techniques can be much more 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 a generalized linear specification of Goodman's regression with feasible corrections for heteroskedasticity provides valid tests of aggregation bias and the neighborhood hypothesis. These tests are not possible in most alternative estimation techniques, and are certainly more burdensome in those techniques that nominally permit them. In addition, unbiased estimates of the effects of covariates are available even if other components of the model, such as the exact form of aggregation bias, may be unidentified. Lastly, with more than two groups, aggregation bias can be identified without parameter restrictions.

Section I of this paper 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. 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 can be at issue. Section IV 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 making the choice at issue.

The relationship between these three quantities in area *i* is Goodman's identity:[2]

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

where

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.[3]

Equation 1 can be rewritten as

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

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

$$y_i = a + bx_i + e_i . \qquad (3)$$

Goodman asserts that, under appropriate conditions, this regression yields *a* and *b* as unbiased estimators of β_{2i} and $\beta_{1i}-\beta_{2i}$ (1953, 664 and 1959, 612).[4]

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), for

1

example). The properties of this random variable allow the conventional empirical OLS model to yield unbiased estimators.

In contrast, Goodman's identity is exact. In 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. This model cannot reveal the true properties of estimators from Goodman's regression. Instead, those properties must be derived analytically from the implications of the identities upon which they are based.

As written, the parameters of Goodman's identity are not identifiable. In 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)).

If the regression of equation 3 were to be calculated, its results would be meaningless. The slope of the regression would not contain any random components. Consequently, it would be a constant rather than an estimator. Its value would be

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

The second equality of equation 4 replaces y_i with its equivalent in terms of x_i, as given in equation 2.

Without further assumptions, the two ratios to the right of the second equality in equation 4 cannot be simplified. Consequently, the slope that would be estimated by the regression of equation 3 is not interpretable. It is certainly not equal to the difference $\beta_{1i}-\beta_{2i}$ for any *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, for example, the behavioral parameters were assumed to be constant across all areas, $\beta_{1i}=\beta_1$ and $\beta_{2i}=\beta_2$ for all *i*. Equation 2 would then be the same for each area, and depend on only two unknown parameters:

$$y_i \equiv \beta_1 x_i + \beta_2[1 - x_i]. \qquad (5)$$

Data on y_i and x_i from only two areas would determine these parameters exactly, because equation 5 is exact. An empirical regression of y_i on x_i would be unnecessary. It would yield a slope coefficient identically equal to $\beta_1-\beta_2$ and an intercept identically equal to β_2. The prediction

errors for each observation would be identically zero and R^2 would be identically one.[5]

These last two characteristics do not appear in the literature. This implies that the assumption of parameter constancy across observations, alone, cannot endow Goodman's identity with empirical relevance. Moreover, the exact solutions embodied in equations 2 or 5 have been ignored in favor of statistical formulations inspired by equation 3. This implies that the collective intuition anticipates 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 functions of a limited number of variables:

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

This reduces the number of parameters to that necessary to characterize f_1 and f_2.[6]

Equation 6 has two additional virtues. First, it incorporates "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. This is represented by the explicit presence of x_i in f_1 and f_2.[7] Second, equation 6 explicitly includes covariates of y_i other than x_i. The vectors $\mathbf{z_{1i}}$ and $\mathbf{z_{2i}}$ contain any other determinants of the proportions of the two groups with the characteristic or making the choice at issue.

Neither aggregation bias nor covariates can be introduced explicitly into equation 1. The appearance of either would violate the identity. The quantities y_i and x_i cannot be expressed as

functions of z_i because they are observable directly. Consequently, equation 6 is the only formulation that can preserve Goodman's identity, yet expand it to incorporate aggregation bias and covariates.

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 again would invalidate Goodman's identity. However, random components can be embedded in all of the quantities already present to the right of the identity.[8]

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.[9] It is absent from most of the subsequent literature, but is central to Achen and Shively (1995), Ansolabehere and Rivers (1995) and King (1997). Here, it implies that equation 6 be augmented as

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

where $E(\varepsilon_{1i})=E(\varepsilon_{2i})=0$, ε_{1i} and ε_{2i} are orthogonal to x_i, $\mathbf{z_{1i}}$ and $\mathbf{z_{2i}}$.[10]

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:

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

where ν_i has $E(\nu_i)=0$ and is orthogonal to x_i.[11] For example, analyses of voting behavior often compare votes in an election from one year with population proportions from a census in another. If these proportions change, the measured proportions may differ from the relevant proportions.

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

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

or

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

Equation 9 demonstrates that this generalization is still an identity. Nevertheless, it has the statistical character that is absent in equation 2 and 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 that includes 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 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. OLS estimators will generally suffer from bias if the specifications of f_1 or f_2 are incorrect.[12]

These challenges also represent general sources for unsatisfactory results in any estimations of equation 3. For example, estimations can yield values for b_0 and $b_1 - b_0$ that are outside the Duncan-Davis bounds (King (1997, chapter 5)) and even the logical bounds of zero and one.[13] 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 daunting 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.[14] The rest of this paper therefore defers discussion of this issue, and assumes as a maintained hypothesis that x_i is measured without error. In this and the following sections, the discussion focuses on tests for possible misspecifications given this assumption and, for the most part, the additional assumption of linearity in the underlying model.

In the absence of measurement error, random coefficients are necessary to endow Goodman's identity with any random component. Its presence in the literature since, arguably, 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

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

The expected value of the residual term in equation 10 is zero:

$$E\left[\varepsilon_{1i} x_i + \varepsilon_{2i}\left[1 - x_i\right]\right] = x_i E\left[\varepsilon_{1i}\right] - \left[1 - x_i\right] E\left[\varepsilon_{2i}\right] = 0.$$

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, correct empirical specifications require that these interactions appear in the estimated equation.

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

These interactions provide a compelling 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 across areas, rather than from variations in characteristics or choice proportions across groups coupled with variation 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 imposes the restriction

$$f_1 = f_2 = f \quad (11)$$

on equation 10.[15] This implies that the second term to the right of the equality in equation 10, the interaction term in which $f_1 - f_2$ multiplies x_i, disappears, because $f_1 - f_2 = 0$.

Under the neighborhood model, then, equation 10 becomes

$$y_i = f(x_i, \mathbf{z_i}) + \left[\varepsilon_{1i} x_i + \varepsilon_{2i}\left[1 - x_i\right]\right]. \quad (12)$$

The restriction that $f_1 - f_2 = 0$ is testable, regardless of the specifications for f_1 and f_2, by comparing the explanatory power of the estimating equations for equations 10 and 12. Therefore, the neighborhood model can always be distinguished empirically from Goodman's identity. This is true whether or not aggregation bias or covariates are present, and regardless of the functional form that defines their presence.

For example, in the simplest form of Goodman's identity, covariates and aggregation bias are both absent. Consequently, equation 7 is

$$f_1(x_i, \mathbf{z_i}) = \beta_1 \text{ and } f_2(x_i, \mathbf{z_i}) = \beta_2. \quad (13)$$

Goodman's identity, as reformulated in equation 10, becomes

$$y_i = \beta_2 + \left[\beta_1 - \beta_2\right] x_i + \left[\varepsilon_{1i} x_i - \varepsilon_{2i}\left[1 - x_i\right]\right]. \quad (14)$$

In this reformulation, y_i is determined by a constant and a linear term in x_i. Equation 3 is the appropriate empirical representation of this relationship.

Equation 11, applied to the version of Goodman's identity in equation 13, specifies f_1 and f_2 as identical constants, $f_1 = f_2 = \beta_1$. With this restriction, the neighborhood model of equation 12 becomes

$$y_i = \beta_1 + \left[\varepsilon_{1i} x_i + \varepsilon_{2i}\left[1 - x_i\right]\right]. \quad (15)$$

In this model, y_i does not depend on x_i at all. Its empirical counterpart is

$$y_i = a + e_i. \qquad (16)$$

In the absence of covariates and aggregation bias, the estimating equation for Goodman's identity is equation 3 and the estimating equation for the neighborhood model is equation 16. It is obvious that the difference between the two can be tested. If, in estimates of equation 3, *b* is statistically different from zero, then equation 16, and the version of the neighborhood model that it represents, must be rejected.

Previous papers have incorrectly asserted that Goodman's identity and the neighborhood hypothesis cannot be distinguished empirically (Freedman, et al. (1991, 682), Klein, Sacks and Freedman (1991), Lichtman (1991, 787), Achen and Shively (1995, 14), King (1997, 41-44), Kousser (2001, 105-7) and Wakefield (2004, 397), as examples). These assertions are based on an inappropriate comparison between the version of Goodman's identity in equation 13, which does not embody aggregation bias, and the "linear model" version of the neighborhood model proposed by Freedman, et al. (1991), which does.

This "linear model" specifies that y_i does not depend on covariates. With linear aggregation bias, the model is:[16]

$$f(x_i, z_i) = \beta_1 + \beta_{10} x_i. \qquad (17)$$

In this case, the deterministic component of equation 12 is linear in x_i:

$$y_i = \beta_1 + \beta_{10} x_i + \left[\varepsilon_{1i} x_i + \varepsilon_{2i}\left[1 - x_i\right]\right]. \qquad (18)$$

The empirical counterpart to this equation represents y_i as dependent on a constant and a linear term in x_i, which is the specification in equation 3.

This demonstrates that equation 3 is the empirical representation of both the simplest form of Goodman's identity, equation 14, and the "linear model" of Freedman, et al. (1991), equation 18. The fact that these two very different models imply the same estimating equation has been interpreted to indicate that empirical evidence regarding the relationship between y_i and x_i cannot distinguish between Goodman's identity and the neighborhood model

This interpretation is false because the comparison between equations 14 and 18 is inappropriate. The neighborhood model of equation 18 embodies linear aggregation bias but the version of Goodman's identity in equation 14 does not. In other words, the neighborhood model of equation 17 is not a restricted version of Goodman's identity 13, as required by equation 11. The appropriate comparison to the neighborhood model of equation 18 requires linear aggregation bias in Goodman's identity, as well.

With linear aggregation bias, equation 7 demonstrates that Goodman's identity requires[17]

$$f_1(x_i, \mathbf{z_i}) = \beta_1 + \beta_{10} x_i \text{ and } f_2(x_i, \mathbf{z_i}) = \beta_2 + \beta_{20}[1 - x_i]. \qquad (19)$$

Equation 19, with the restriction of equation 11, yields the neighborhood model of equation 17, as it should. Furthermore, it implies that Goodman's identity, as represented in equation 10, is

$$\begin{aligned} y_i &= [\beta_2 + \beta_{20}[1 - x_i]] + [[\beta_1 + \beta_{10} x_i] - [\beta_2 + \beta_{20}[1 - x_i]]] x_i + [\varepsilon_{1i} x_i + \varepsilon_{2i}[1 - x_i]] \\ &= [\beta_2 + \beta_{20}] + [\beta_1 - \beta_2 - 2\beta_{20}] x_i + [\beta_{10} - \beta_{20}] x_i^2 + [\varepsilon_{1i} x_i + \varepsilon_{2i}[1 - x_i]]. \end{aligned} \qquad (20)$$

Equation 20 contains a quadratic term in x_i. Therefore, equation 3 is not the appropriate empirical counterpart. Instead, equation 20 must be estimated by an equation of the form

$$y_i = a + b x_i + d x_i^2 + e_i, \qquad (21)$$

where the notation is consistent with the general model presented in equation 24 below.

The regression of equation 21contains only three estimators, *a*, *b* and *d*. In contrast, the model of equation 20 contains four parameters, β_1, β_{10}, β_2 and β_{20}. None of these parameters are individually identified.

However, tests of the neighborhood hypothesis do not require individual identification of all parameters, because this hypothesis does not specify their individual values. Instead, it specifies only that $\beta_1=\beta_2$ and $\beta_{10}=\beta_{20}$. In other words, the neighborhood model in this context requires only that the differences $\beta_1-\beta_2$ and $\beta_{10}-\beta_{20}$ both equal zero.[18]

Equation 21 identifies the second of these differences with an unbiased estimate, the slope *d*. If the neighborhood hypothesis is false, this slope should be non-zero. If it is correct, this slope should be zero, the quadratic term in equation 19 should be unnecessary, and equation 3 should adequately represent the sample relationship between y_i and x_i.

Consequently, the question of whether the neighborhood model is valid or not, in the presence of linear aggregation bias, can be rephrased as the empirical question of whether equation 3 or equation 21 is a better fit to the data. If *d* differs significantly from zero, then the regression of equation 21 is superior. In this case, the neighborhood model must be false.

The tests of the neighborhood model with neither covariates nor aggregation bias in the comparison of equations 14 and 15, and without covariates but with linear aggregation bias in the comparison of equations 17 and 19, are examples of the more general test available in the comparison of equations 12 and 10. These tests are more powerful if y_i depends on covariates.

This can be illustrated in a relatively general linear multivariate model of f_1 and f_2. This model 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:

$$f_1(x_i, \mathbf{z_i}) = \beta_1 + \beta_{10} x_i + \sum_{j=1}^{k} \beta_{1j} z_{ij} \text{ and}$$
$$f_2(x_i, \mathbf{z_i}) = \beta_2 + \beta_{20}[1 - x_i] + \sum_{j=1}^{k} \beta_{2j} z_{ij}, \quad (22)$$

where $\mathbf{z_i}$ is a vector, k is the number of covariates in $\mathbf{z_i}$ and z_{ij} is the jth covariate in z_i. Equation 22 apparently includes all empirical examples of Goodman's regression, including those of equations 2 and 20, as special cases.[19]

Under equation 22, equation 10 becomes

$$y_i = [\beta_2 + \beta_{20}] + [\beta_1 - \beta_2 - 2\beta_{20}] x_i + \sum_{j=1}^{k} \beta_{2j} z_{ij} + [\beta_{10} + \beta_{20}] x_i^2 + \sum_{j=1}^{k} [\beta_{1j} - \beta_{2j}] z_{ij} x_i + [\varepsilon_{1i} x_i + \varepsilon_{2i}[1 - x_i]]. \quad (23)$$

Each of the elements of $\mathbf{z_i}$ appears linearly and interacted with x_i.[20] The latter appears in both linear and quadratic terms.

Consequently, the appropriate estimating equation would be

$$y_i = a + b x_i + \sum_{j=1}^{k} c_j z_{ij} + d x_i^2 + \sum_{j=1}^{k} h_j z_{ij} x_i + e_i, \quad (24)$$

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.[21] β_1, β_{10}, β_2 and β_{20} would not be individually identified.

In terms of equation 23, the neighborhood model imposes the restrictions that $\beta_1=\beta_2$ and $\beta_{1j}=\beta_{2j}$ for all j from zero through k. For all j from one through k, the restrictions that $\beta_{1j}=\beta_{2j}$ imply

that $\beta_{1j}-\beta_{2j}=0$. This further implies that, in the regression of equation 24, all of the h_j should be statistically indistinguishable from zero. If this hypothesis is rejected by the *t*-test for any of the h_j, the neighborhood model is false.

In addition, under the neighborhood hypothesis the coefficient of x_i in equation 23 is equal to $-2\beta_{20}$, because $\beta_1=\beta_2$. Similarly, the coefficient of x_i^2 in equation 23 is equal to $2\beta_{20}$ because $\beta_{10}=\beta_{20}$. Consequently, this hypothesis asserts that coefficients of x_i and x_i^2 should sum to zero. In the regression of equation 24, this implies that $b+d=0$. Once again, if the relevant *F*-statistic rejects this implication, the neighborhood hypothesis must be false.[22]

In other words, any linear specification of the neighborhood model in the form of equation 23 is empirically distinguishable from the model of Goodman's identity, even though some or even many individual parameters in the model may be unidentified. The neighborhood hypothesis imposes unique, testable constraints on Goodman's identity. The same is almost certainly true of any neighborhood model that might specify the parameters of Goodman's identity as nonlinear combinations of the explanatory variables.

Unfortunately, the most common empirical implementations invoking Goodman's identity omit interactions of f with x_i, yielding the linear form of equation 12. These implementations therefore impose the neighborhood model as a maintained hypothesis, almost surely inadvertently. Kousser (2001), who is explicitly hostile to the neighborhood model (pages 105-7, 110), is an ironic example (pages 110-5).

B. Goodman's regression, aggregation bias and covariates

The model of equation 23 and its empirical implementation in equation 24 address two other issues that are central to ecological investigations, and previously taken to be intractable. First, aggregation bias is present in equation 23 and its affects, β_{10} and β_{20}, are not identified.[23] However, equation 24 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.[24]

Second, even in the presence of unidentified aggregation bias, equation 24 identifies the behavioral determinants of the proportions of the two groups making the choice at issue. The coefficients of $\mathbf{z_i}$ 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 not identify the exact effects of aggregation bias, itself.

C. Heteroskedasticity, weighting and R^2 in Goodman's regression

Tests for the neighborhood model and the absence of aggregation bias 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 intervals.

Inference is complicated in equation 10 because the residual term is heteroskedastic (Goodman (1959, 612); Ansolabehere and Rivers (1995, 9-10)), an inherent property of random coefficient models (Greene (2003, 318-9)). Heteroskedasticity does not impose bias on OLS estimators of f_1 and f_2, if they are specified correctly (Greene (2003, 193-5)).[25] However, it must be addressed in order to construct any valid test of statistical significance.[26]

Achen and Shively (1995, 47-8) and Lewis (2001, 177) discuss feasible strategies in the context of ecological regression.[27] 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 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 values 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 apparent in equation 10. It will almost certainly compound it.[28]

Instead, the R^2 value from 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, then it has no relationship to the statistical tests of interest – the validity of the

neighborhood model, the absence of aggregation bias, and the importance of behavioral determinants.

If the R^2 value is from a weighted regression, it is almost certainly meaningless, whether or not the weights correct appropriately for heteroskedasticity. Unless the weights are the inverses of an explanatory variable that enters 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)).

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 linear model analyzed explicitly here, it is likely that appropriate extensions of this analysis will preserve the general conclusions. For these reasons, OLS should be an attractive technique for empirical implementations of a single Goodman's

identity.

King's maximum-likelihood method of ecological inference (King (1997)) is the principal alternative. His statistical model is comparable to OLS with the White heteroskedasticity-consistent variance estimators in that it yields consistent estimates[29] and valid hypothesis tests. It also allows for explicit treatment of aggregation bias (King (1997, chapter 9). It is superior in that it explicitly incorporates the Duncan-Davis bounds. Consequently, its estimates are guaranteed to be feasible and more precise. This precision is apparent in Silva de Mattos and Veiga (2004).

However, OLS may be preferable to King's method for the purposes of incorporating multiple covariates and testing the neighborhood hypothesis.[30] In the first case, the inclusion of covariates in King's method is computationally burdensome (King (2003, page 49)).[31] Given the many interaction terms that may appear in the unrestricted model of equation 10, specifications of f_1 and f_2 with multiple covariates may not be tractable, at least currently, within this method. If so, tests of the neighborhood hypothesis and aggregation bias may be infeasible.

In the second case, the EzI software package, which provides stand-alone implementation of King's estimation procedure, does not appear to allow the imposition of the constraints implied by the neighborhood hypothesis. It may be possible to implement these restrictions in the EI package of Gauss programs upon which EzI is based, but only with additional programming. In comparison, tests of the neighborhood hypothesis in the OLS context require only the standard t- and F-tests.

Other estimation techniques are unambiguously inferior to OLS for most purposes. The estimation procedure of King, Rosen and Tanner (1999) may yield biased estimates of the underlying model (Silva de Mattos and Veiga (2004)), is difficult to calculate, and is very

burdensome in the presence of covariates. 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 amplified when the number of groups in the population is greater than two.[32] All parameters in the linear multivariate model analogous to that of equation 22 are identified, and the test for aggregation bias has much greater power than in the two-group case. However, heteroskedasticity is more complicated.

The case with three groups illustrates these points. Augmenting the notation of section I:

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

$$\begin{aligned} y_i &\equiv \beta_{1i} x_{1i} + \beta_{2i} x_{2i} + \beta_{3i}\left[1 - x_{1i} - x_{2i}\right] \\ &\equiv \beta_{3i} + \left[\beta_{1i} - \beta_{3i}\right] x_{1i} + \left[\beta_{2i} - \beta_{3i}\right] x_{2i}. \end{aligned} \tag{25}$$

In this case, the analogue to the linear multivariate model of equations 7 and 22 is

$$\beta_{ri} = f_r\left(x_{ri}, z_{ij}\right) + \varepsilon_{ri} = \beta_r + \beta_{r0} x_{li} + \sum_{j=1}^{k} \beta_{rj} z_{ij} + \varepsilon_{ri}, \tag{26}$$

where $r=\{1, 2, 3\}$ identifies the group. The substitution of equation 26 into equation 25 yields

$$\begin{aligned} y_i &= \left[\beta_3 + \beta_{30}\right] + \left[\beta_1 - \beta_3 - 2\beta_{30}\right] x_{1i} + \left[\beta_2 - \beta_3 - 2\beta_{30}\right] x_{2i} + \sum_{j=1}^{k} \beta_{3j} z_{ij} + \left[\beta_{10} + \beta_{30}\right] x_{1i}^2 \\ &+ \left[\beta_{20} + \beta_{30}\right] x_{2i}^2 + 2\beta_{30} x_{1i} x_{2i} + \sum_{j=1}^{k} \left[\beta_{1j} - \beta_{3j}\right] z_{ij} x_{1i} + \sum_{j=1}^{k} \left[\beta_{2j} - \beta_{3j}\right] z_{ij} x_{2i} \\ &+ \left[\varepsilon_{1i} x_{1i} + \varepsilon_{2i} x_{2i} + \varepsilon_{3i}\left[1 - x_{1i} - x_{2i}\right]\right]. \end{aligned} \tag{27}$$

The regression counterpart of equation 27 is

$$y_i = a + b_1 x_{1i} + b_2 x_{2i} + \sum_{j=1}^{k} c_j z_{ij} + d_1 x_{1i}^2 + d_2 x_{2i}^2 + d_{12} x_{1i} x_{2i} + \sum_{j=1}^{k} h_{1j} z_{ij} x_{1i} + \sum_{j=1}^{k} h_{2j} z_{ij} x_{2i} + e_i. \tag{28}$$

The model of equation 27 contains $2+k$ parameters in addition to those in equation 23, for a total of $3[2+k]$ parameters. However, the regression of equation 28 estimates $3+k$ coefficients in addition to those in equation 24. The additional coefficient is attributable to the interaction term in $x_{1i}x_{2i}$. Under conventional practice, this term would be, incorrectly, omitted.

As a consequence of this interaction term, the number of coefficients in the three-group regression of equation 28 equals the number of underlying parameters. Moreover, all are identified, in contrast to the two-group regression of equation 24.[33] As in equation 24, tests of significance on the estimated values for β_{1j}, β_{2j} and β_{3j} indicate whether covariates are important.

Equation 28 also provides powerful tests for the neighborhood model and for the presence

of aggregation bias. The neighborhood model implies $2k+4$ restrictions on equation 27. 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, $|b_1| = |b_2| = |d_1| = |d_2| = |d_{12}|$. The requirement that $\beta_{1j}=\beta_{2j}=\beta_{3j}$ implies $2k$ restrictions, the coefficients on $z_{ij}x_{1i}$ and $z_{ij}x_{2i}$ should all equal zero, or $h_{1j}=h_{2j}=0$ for all j. The failure of any one of these restrictions would invalidate the neighborhood model.

Aggregation bias is present if $\beta_{10}\neq0$, $\beta_{20}\neq0$ or $\beta_{30}\neq0$. 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, or $d_1=d_2=d_{12}=0$. 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 example, if $\beta_{20}=-\beta_{30}\neq0$, the second restriction would hold but the third would fail: $d_2=0$ but $d_{12}\neq0$. Therefore, the failure of any one of these restrictions indicates unambiguously that aggregation bias is present.

At the same time, the last three terms of equation 27 demonstrate that the residual in this regression contains three random components, rather than the two of equation 23. 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. Tests of the restrictions implied by the neighborhood hypothesis or the hypothesis of aggregation bias require corrections for the consequent heteroskedasticity.

As the number of groups increases beyond three, the interaction terms between x_{ki} and x_{mj}

proliferate more rapidly than do the underlying parameters. With R groups, Goodman's identity is

$$y_i \equiv \sum_{r=1}^{R-1} \beta_{ri} x_{ri} + \beta_{Ri}\left[1 - \sum_{r=1}^{R-1} x_{ri}\right] \equiv \beta_{Ri} + \sum_{r=1}^{R-1}\left[\beta_{ri} - \beta_{Ri}\right] x_{ri}.$$

The generalized linear specification of the propensities, according to equation 16, is

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

This specification contains k+2 parameters for each of the R groups, or R[2+k] in total.

With this specification, the deterministic part of y_i is

$$\left[\beta_R + \beta_{R0}\left[1 - \sum_{r=1}^{R-1} x_{ri}\right] + \sum_{j=1}^{k} \beta_{Rj} z_{ij}\right] + \sum_{r=1}^{R-1}\left[\left[\beta_r + \beta_{r0} x_{ir} + \sum_{j=1}^{k} \beta_{rj} z_{ij}\right] - \left[\beta_R + \beta_{R0}\left[1 - \sum_{r=1}^{R-1} x_{ri}\right] + \sum_{j=1}^{k} \beta_{Rj} z_{ij}\right]\right] x_{ri}.$$

Rearranging, this becomes

$$\begin{aligned}
&\left[\beta_R + \beta_{R0}\right] + \sum_{r=1}^{R-1}\left[\beta_r - \beta_R - \beta_{R0}\right] x_{ri} + \sum_{j=1}^{k} \beta_{Rj} z_{ij} \\
&+ \sum_{r=1}^{R-1}\left[\beta_{r0} + \beta_{R0}\right] x_{ri}^2 + 2\beta_{R0} \sum_{r=1}^{R-1} \sum_{m=1}^{j-1} x_{ri} x_{mi} \\
&+ \sum_{r=1}^{R-1} \sum_{j=1}^{k}\left[\beta_{rj} - \beta_{Rj}\right] z_{ij} x_{ri}.
\end{aligned}$$

The number of coefficients estimated in the successive terms of this expression is one, R−1, k, R−1, [R−1][R−2]/2 and k[R−1]. Therefore, the total number of estimated coefficients is R[2+k]+½R[R-3]. This exceeds the number of parameters by ½R[R-3], which is equal to zero when R=3 and positive when R>3.

Consequently, models with *R*>3 groups are actually overidentified. In order to ensure that estimates are consistent with the underlying model, ½R[R−3] restrictions are necessary. The

effect of these restrictions on the explanatory power of the regression provides a general test of the underlying specification of Goodman's identity.

In contrast to the number of groups in the population, the number of alternative characteristics or choices has few, if any, implications for Goodman-based estimation. The identities of equations 1 and 25 do not depend on this number, and are therefore valid regardless of its value. Consequently, the estimations of equations 23 and 27 do not depend on the number of alternatives.

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.

This section demonstrates that OLS, properly specified, should be a relatively attractive estimation technique for the "R×C model", that with more than two groups, more than two choices or both. 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.

At least four other estimation techniques have been suggested for the R×C problem. King's method for ecological inference (King (1997, chapter 15)) and the binomial-beta hierarchical model (Rosen, Jiang, King and Tanner (2001)) are both computationally burdensome when f_r is constant for all r.[34] More complicated specifications of f_r would compound the difficulties. There appears to be little applied experience with the procedures of Judge, Miller and

Cho (2004) or Greiner and Quinn (2009). In any event, 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. Conclusion

This paper demonstrates that regression-based applications of Goodman's identity can be much more effective than previously understood. OLS estimates of the generalized Goodman's regression in equation 23 are unbiased. They are also heteroskedastic, but corrections are feasible.

With these corrections, OLS estimators provide valid statistical tests for the neighborhood model and aggregation bias. In addition, they provide unbiased estimates and standard deviations for the effects of covariates. Moreover, identification in these models improves as the number of groups in the population increases. With these properties and its well-known flexibility, OLS should be a valuable tool in the analysis of models that, correctly-specified, require a single application of Goodman's regression. In cases with many covariates, it may be superior, despite the risk of estimators outside of the Duncan-Davis bounds and the ingenuity embodied in recent alternatives (King (1997), King, Rosen and Tanner (1999) and Lewis (2004) as examples).

Current applications of Goodman's regression typically fail to realize any of its potential. Most empirical exercises impose the neighborhood assumption, estimate incorrect standard errors and offer neither hypothesis tests nor confidence intervals. Instead, they are distorted by arbitrary weights that exacerbate heteroskedasticity, and justified with meaningless R^2 values. Sixty 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.

Ansolabehere, Stephen and Douglas Rivers (1995) "Bias in ecological regression", working paper, Stanford University, Palo Alto, CA.

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.

Collet, Christian (2005) "Bloc voting, polarization, and the Panethnic Hypothesis: The case of Little Saigon", Journal of Politics, Vol. 67, No. 3, August, 907-933;

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 individuals", 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.

Greiner, D. James and Kevin M. Quinn (2009) "R×C ecological inference: bounds, correlations, flexibility and transparency of assumptions", Journal of the Royal Statistical Society, Series A, Vol. 172, Issue 1, 67-81.

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.

Herron, Michael C. and Kenneth W. Shotts (2004) "Logical inconsistency in EI-based second-stage regressions", American Journal of Political Science, Vol. 48, No. 1, January, 172-183.

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.

Liu, Baodong (2007) “EI extended model and the fear of ecological fallacy”, Sociological Methods & Research, Vol. 36, No. 1, August, 3-25.

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.

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.

Wakefield, Jon (2004) “Ecological inference for 2x2 tables”, Journal of the Royal Statistical Society: Series A (Statistics in Society), Vol. 167, Part 3, July, 385-445.

Endnotes

[1] According to Google Scholar, Goodman (1953) has been cited 563 times. Its successor, Goodman (1959), has been cited an additional 497 times.

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

[3] This is the "two-party, no abstention" case of Achen and Shively (1995, 30) and the "basic model" in King (1997, chapter 6). The parameters of Goodman's identity describe behavior at the aggregate level, here the "area". However, 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)). Achen and Shively (1995) 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 difficulties with models of this sort. Typically, they require data at the individual level or assumptions that effectively impose homogeneity on the deterministic component of individual behavior within an aggregate unit. Therefore, the discussion here follows King (1997, 119) and focuses on the problem of obtaining valid parameter estimates and constructing valid tests of behavior at the aggregate level.

[4] This analogy is common in subsequent literature. As examples, it is explicit in Kousser (2001, equation 13) and implicit in Collet (2005, 914) and Liu (2007, 6).

[5] 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, it appears that only Lewis (2004) and Quinn (2004) have explored this identification strategy.

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

[7] "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).

[8] Measurement error in y_i is also possible. However, the dependent variable in many applications of Goodman's identity measures voting behavior. Official counts of voters and votes are exact, at least for the purpose of determining electoral outcomes. Therefore, it will usually be appropriate to treat these counts as measured without error.

[9] 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.

[10] 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).

[11] Chapter 3 of Achen and Shively (1995) is 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 true value (Greene (2003, 84)). This is convenient below, where measurement error is disregarded. Measurement error is also possible in z_{1i} and z_{2i}, but the consequences would be similar in form and tractability to those associated with measurement error in x_i.

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

[13] 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.

[14] 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.

[15] Equation 10 demonstrates that this is the "weak form" of the neighborhood model. The "strong form" would also require that the random components of group behavior be identical within area, $\varepsilon_{1i} = \varepsilon_{2i}$. This would impose the additional restriction of homoskedasticity on the empirical error terms, which could also be tested. The heteroskedasticity in the unrestricted form of equation 10 is discussed below.

[16] Freedman, et al. (1991, 682) suggest the "nonlinear" neighborhood model as an alternative in which $\beta_{1i}=\beta_{2i}=y_i$. 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 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 arbitrarily chooses the single pair that satisfies the restriction $\beta_{1i}=\beta_{2i}$. Statistically, this "model" is no better than that consisting of any other pair of values from each of these lines. Each of these "models" will "fit" the data perfectly, by absorbing all degrees of freedom. 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 implausible, than all procedures of this type, including the nonlinear neighborhood model, are irrelevant.

[17] The notation here extends that of equation 5. As there, β_1 and β_2 represent the fixed components of the propensity of each group to make the choice in question. The new parameters, β_{10} and β_{20}, represent the effects of population composition on these propensities. This model and the more general version in equation 22 specify f_2 as functions 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. The fundamental results are identical with either specification.

[18] King (1997, 41-44) also discusses the case of linear aggregation bias without covariates. As here, King concludes that Goodman's model implies equation 20, and that its individual parameters are not identified. However, he does not address the separate question of whether the difference between Goodman's identity and the neighborhood hypothesis is identified.

[19] Equation 22 extends the notation used previously, as discussed in footnote 16. It reduces to equation 19 if $\beta_{1j}=\beta_{2j}=0$ for all *j* from one to k. It reduces to equation 13 if, in addition, $\beta_{10}=\beta_{20}=0$. Achen and Shively (1995, 13 and 73, footnote 14) refer to this latter model when they assert that "(j)ust one technique for handling ecological data has been widely adopted in practice: the linear (unextended) version of Goodman ecological regression".

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

[21] 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.

[22] Collectively, these tests must have very high power. If all fail to reject their respective null hypotheses, the neighborhood model could still be false only in the unlikely event that $\beta_1\neq\beta_2$ and $\beta_{10}\neq\beta_{20}$, yet $\beta_1+\beta_{10}-\beta_2-\beta_{20}=0$.

[23] Rivers (1998, 443) asserts that this model is unidentified. King (1997, section 3.2), Voss (2004, 72-73) and Wakefield (2004, 398) provide additional examples. Achen and Shively (1995,

chapters 5 and 6) discuss identifying strategies 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, β_{10} and β_{20}.

[24] This test does not restrict the treatment of $\mathbf{z_i}$ in f_1 and f_2. It requires only that these functions be linear in x_i. More complicated functions of x_i would presumably suggest analogous tests. This test should have relatively good power. If *d* is statistically indistinguishable from zero, aggregation bias could only be present in the unlikely event that $\beta_{10}=-\beta_{20}\neq 0$.

[25] Achen and Shively (1995, 47-8) and Lewis (2001, 177) note that OLS estimates of equation 24 are unbiased where f_1 and f_2 are constants.

[26] King (1997, 65-8) asserts that heteroskedasticity can severely distort inference in ecological regression models. As an example, the results in Bourke, DeBats and Phelan (2001, 132)) are subject to this distortion because their standard errors are not corrected for heteroskedasticity. Achen and Shively (1995, 47-50 and 128) claim that heteroskedasticity is empirically unimportant. However, they are essentially uninterested in inference (page 58).

[27] These discussions assume that ε_{1i} and ε_{2i} are uncorrelated with ε_{1j} and ε_{2j} for all *i* and *j*. Autocorrelation (Cho (1998, 145-6)) would require additional corrections to standard errors. Inexplicably, Cho reports OLS standard errors without any indication that they have been appropriately corrected.

[28] King (1997, 61-5) and Achen and Shively (1995, 57-61) are critical of weighting by the inverse square root of population. Ansolabehere and Rivers (1995, 10) agree that this weighting almost certainly fails to correct for heteroskedasticity. In contrast, Kousser (2001) asserts without proof that it does correct for heteroskedasticity (page 512) 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.

[29] In general, maximum likelihood estimation is guaranteed to yield consistent estimators, assuming, as throughout this paper, that the underlying behavior is appropriately represented by the model (Greene (2003, 467-8)). King's method is explicitly maximum likelihood. The White heteroskedasticity-consistent variance estimates are equivalent to maximum likelihood estimates (Greene (2003, 520)).

[30] It may also be convenient to take advantage of the simplicity of OLS by testing for aggregation bias in Goodman's regression as described above. If it is absent, King's method can then be invoked without the burden of exploring aggregation bias in that context.

[31] 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)), but

probably unfortunate (Herron and Shotts (2004)).

[32] 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.

[33] The coefficients on the z_{ij} , c_j, identify β_{3j}. With these results, the coefficients on $z_{ij}x_{1i}$ and $z_{ij}x_{2i}$, h_{1j} and h_{2j}, identify β_{1j} and β_{2j}, respectively. The coefficient on $x_{1i}x_{2i}$, d_{12}, identifies β_{30}. With this result, the coefficients on x_{1i}^2 and x_{2i}^2, d_1 and d_2, identify β_{10} and β_{20}, respectively, and the constant a identifies β_3. With this last result and β_{30}, the coefficients on x_{1i} and x_{2i}, b_1 and b_2, identify β_1 and β_2, respectively.

[34] King (1997, chapter 15) suggests a simplification relying on iterative applications of the bivariate truncated normal distribution. This strategy may yield biased estimates of the underlying behavioral model (Ferree (2004)).