

**THE STATISTICAL PROPERTIES OF SINGLE REGRESSION  
ESTIMATES OF VOTING BEHAVIOR WHEN TURNOUT IS  
UNKNOWN**

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

Telephone: 303-492-8268  
FAX: 303-492-8960  
e-mail: [zax@colorado.edu](mailto:zax@colorado.edu)

25 October 2007

## **ABSTRACT**

When turnout rates among subgroups within the electorate are unknown, description of voting behavior requires two applications of Goodman's identity. This description is frequently implemented, empirically, with a single variant of "Goodman's regression" which is inconsistent with the identities and does not identify their parameters. The expected values and variances of the dependent variable and all regression statistics are unknown. Simulations demonstrate that the regression mis-estimates true preferences at rates far in excess of acceptable statistical standards. The related techniques of "homogeneous area analysis" and "correlation analysis" are afflicted with similar problems, and similarly unreliable. All three should be discarded.

## **1. Introduction**

The analysis of voting preferences among subgroups in the electorate, when turnout for each subgroup is unobserved, can be modeled by two applications of Goodman's identity (Goodman, 1953). King (1997) provides an explicit statistical implementation of this model. Rosen, et al. (2001) and Wakefield (2004) provide alternative implementations that may also be appropriate.

Despite the availability of these techniques, a variant of "Goodman's regression" is frequently adopted for the purpose of identifying voting preferences. Cho (1998), Kousser (2001), Collet (2005) and Liu (2007) are recent examples. This paper examines the properties of single regression techniques in this context.

## **2. Voting choices, group turnout and Goodman's identities**

Kousser (2001, 110, referencing Achen and Shively, 1995) asserts that the estimation of transition matrices relating partisan voting patterns in two successive elections and the comparison of voting patterns across two different racial or ethnic groups in the same election are the two principle applications of Goodman's (1953) regression. These two applications have radically different statistical characters.

With transition matrices, the composition of the population voting in the second election is known. All voters are characterized by their behavior in the first election, identified either by the alternative for which they voted or by abstention. This context is completely described by a single application of Goodman's identity. Zax (2007) demonstrates that Goodman's regression has attractive properties in this case.

The analysis of voting patterns across racial or ethnic groups is analogous to that of

transition matrices only in the rare circumstance where the racial or ethnic composition of voters is known. Liu (2007) and, apparently, Epstein and O'Halloran (1999) are examples. However, in virtually all elections of interest, the racial or ethnic composition of the electorate is known but the composition of the voters is not.

Therefore, the essential distinction between the analyses of transition matrices and of voting behavior is that, in the latter, the parameters of interest refer to the behavior of a subset of the population whose composition is unknown. The distribution of voters across racial or ethnic groups will differ from that of the electorate if turnout rates differ by group. Therefore, the complete analysis of group behavior requires separate identification of turnout propensities and choices made by actual voters.<sup>1</sup>

Consequently, the typical application of Goodman's approach to the analysis of voting behavior entails two applications of Goodman's identity (King, 1997, 68-71). Zax (2005) presents the underlying structural model, summarized here.

The first application of Goodman's identity relates the observed turnout in any precinct to the observed composition of the electorate and the unobserved turnout rates among the two groups. Formally, the observed characteristics of precinct  $i$  include

$$\begin{aligned} x_i &= \text{the proportion of the electorate in precinct } i \text{ that belongs to group} \\ &\quad \text{D,} \\ 1 - x_i &= \text{the proportion of the electorate in precinct } i \text{ that belongs to group} \end{aligned}$$

---

<sup>1</sup> For example: If all of the group D electorate votes and 70% of its members choose candidate W, only 30% prefer candidate B. The same share prefers candidate B if 50% of the group D electorate abstains and only 20% choose candidate W. However, group D voters prefer candidate W to candidate B in the first case and candidate B to candidate W in the second.

R, and

$T_i$  = the turnout rate in precinct  $i$ , the ratio of the number of votes cast to the number of potential voters.

The unobserved parameters which determine group-specific turnout rates in precinct  $i$  are<sup>2</sup>

$\beta_{Di}$  = the turnout rate among group D voters in precinct  $i$ , the ratio of the number of votes cast by group D voters to the number of potential group D voters, and

$\beta_{Ri}$  = the turnout rate among group R voters in precinct  $i$ , the ratio of the number of votes cast by group R voters to the number of potential group R voters.

Goodman's identity specifies the exact relationship between the observed turnout and observed composition of the electorate in precinct  $i$  as

$$(1) \quad T_i \equiv \beta_{Di}x_i + \beta_{Ri}(1 - x_i) \equiv \beta_{Ri} + (\beta_{Di} - \beta_{Ri})x_i.$$

The relationship in equation 1 is generally of secondary interest. It is, however, a necessary precursor to the relationship of primary interest, between the votes received by a particular candidate and the composition of the electorate. Vote totals are also observed:

$y_{Bi}$  = the ratio of the number of votes received by candidate B to the size of the electorate in precinct  $i$ .

However, the voting preferences of D and R group members are unobserved:

---

<sup>2</sup>  $\beta_{Di}$ ,  $\beta_{Ri}$ ,  $\lambda_{Di}$  and  $\lambda_{Ri}$  correspond to  $\beta_i^b$ ,  $\beta_i^w$ ,  $\lambda_i^b$  and  $\lambda_i^w$  in King (1997).

$\lambda_{Di}$  = the ratio of votes cast by group  $D$  voters for candidate  $B$  to the number of votes cast by group  $D$  voters in precinct  $i$ , and

$\lambda_{Ri}$  = the ratio of votes cast by group  $R$  voters for candidate  $B$  to the number of votes cast by group  $R$  voters in precinct  $i$ .

The ratio of actual  $D$  voters to the size of the electorate is

$$(2) \quad T_{Di} \equiv \beta_{Di} x_i.$$

Similarly, the ratio of  $R$  voters to the size of the electorate is

$$(3) \quad T_{Ri} \equiv \beta_{Ri} x_i.$$

Both of these ratios are unobserved, because they depend upon the parameters  $\beta_{Di}$  and  $\beta_{Ri}$ .

With these definitions, the second application of Goodman's identity relates the observed number of votes accruing to candidate  $B$  to the unobserved numbers of group  $D$  and group  $R$  voters, all as proportions of the electorate in precinct  $i$ ,

$$(4) \quad y_{Bi} \equiv \lambda_{Di} T_{Di} + \lambda_{Ri} T_{Ri}.$$

With the substitution of equations 2 and 3, equation 4 becomes

$$(5) \quad y_{Bi} \equiv \lambda_{Di} \beta_{Di} x_i + \lambda_{Ri} \beta_{Ri} (1 - x_i).$$

Equation 5, Goodman's identity for the relationship between the observed allocation of votes and the observed composition of the electorate, depends on four unobserved population parameters. It can be reformulated as

$$(6) \quad y_{Bi} \equiv \lambda_{Ri} \beta_{Ri} + (\lambda_{Di} \beta_{Di} - \lambda_{Ri} \beta_{Ri}) x_i.$$

As discussed in Zax (2005), this equation does not provide a foundation for statistical estimation

because it holds exactly. This foundation requires that turnout rates  $\beta_{Di}$  and  $\beta_{Ri}$  and cohesion rates  $\lambda_{Di}$  and  $\lambda_{Ri}$  vary randomly across precincts:<sup>3</sup>

$$(7) \quad \beta_{Di} = \beta_D + \varepsilon_{Di}, \lambda_{Di} = \lambda_D + \nu_{Di}, \beta_{Ri} = \beta_R + \varepsilon_{Ri} \text{ and } \lambda_{Ri} = \lambda_R + \nu_{Ri}.$$

With this formulation, the parameters  $\beta_D$  and  $\beta_R$  are, respectively, the underlying turnout propensities among members of group D and group R. The underlying propensities to vote for candidate B among group D and group R voters are  $\lambda_D$  and  $\lambda_R$ , respectively.<sup>4</sup>

The parameters  $\beta_D$ ,  $\beta_R$ ,  $\lambda_D$  and  $\lambda_R$  represent all that is common across precincts in the deterministic component of the behaviors of groups D and R. The random disturbances  $\varepsilon_{Di}$ ,  $\varepsilon_{Ri}$ ,  $\nu_{Di}$  and  $\nu_{Ri}$  distinguish the actual turnout and cohesion rates in each precinct,  $\beta_{Di}$ ,  $\beta_{Ri}$ ,  $\lambda_{Di}$  and  $\lambda_{Ri}$ , from these parameters and from rates in other precincts. With the assumption that each of these disturbances has expected value equal to zero, these parameters are the expected values of the precinct-specific turnout and cohesion rates for groups D and R:

$$E(\beta_{Di}) = \beta_D, E(\lambda_{Di}) = \lambda_D, E(\beta_{Ri}) = \beta_R \text{ and } E(\lambda_{Ri}) = \lambda_R.$$

The disturbances  $\varepsilon_{Di}$ ,  $\varepsilon_{Ri}$ ,  $\nu_{Di}$  and  $\nu_{Ri}$  have, respectively, variances of  $\sigma_{D\varepsilon}^2$ ,  $\sigma_{R\varepsilon}^2$ ,  $\sigma_{D\nu}^2$  and  $\sigma_{R\nu}^2$ . The analysis here assumes that aggregation bias is absent. Formally, this requires that all disturbances are uncorrelated with  $x_i$ .<sup>5</sup>

The relationships between the four disturbance terms are characterized by six

---

<sup>3</sup> King (1997) makes the same assumption, implied by Goodman (1959, 612).

<sup>4</sup> King (1997, 96 equation 6.5) represents the underlying turnout propensities as  $\mathfrak{B}^b$  and  $\mathfrak{B}^w$ . He does not provide notation for the underlying vote preferences.

<sup>5</sup> Non-zero correlations, as in Cho (1998) and Epstein and O'Halloran (1999), would expand the number of inestimable parameters and exacerbate the intractabilities below.

covariances. Two are most relevant to the analysis below. The covariance between  $\epsilon_{Di}$  and  $v_{Di}$  is  $\sigma_{DeV}$ . It describes the extent to which precinct-specific variations in turnout among group D members,  $\epsilon_{Di}$ , are correlated with precinct-specific variations in voting preferences among those of this group who vote,  $v_{Di}$ . Similarly,  $\sigma_{ReV}$  describes the extent to which precinct-specific variations in turnout among group R members,  $\epsilon_{Ri}$ , are correlated with precinct-specific variations in voting preferences among group R voters,  $v_{Ri}$ .

These covariances are almost certainly non-zero. Precinct-specific variations in group turnout propensities are likely to be related to precinct-specific variations in voting preferences for that same group. For example, candidate B might receive more support from members of both groups in the precinct  $i$  where candidate B resides than in other precincts. If so, in a contest with B as a candidate, members of both groups would be more likely to turnout,  $\epsilon_{Di}>0$  and  $\epsilon_{Ri}>0$ , and more likely to vote for that candidate,  $v_{Di}>0$  and  $v_{Ri}>0$ . Consequently,  $\sigma_{DeV}>0$  and  $\sigma_{ReV}>0$ .<sup>6</sup>

Equations 1 and 7 imply that the relationship between precinct turnout and the composition of the electorate within the precinct is

$$(8) \quad T_i \equiv (\beta_D + \epsilon_{Di})x_i + (\beta_R + \epsilon_{Ri})(1 - x_i) \\ = \beta_D x_i + \beta_R (1 - x_i) + \epsilon_{Ri} (1 - x_i) + \epsilon_{Di} x_i.$$

The expected value of precinct turnout is

---

<sup>6</sup> Moreover, group D turnout would be positively related to group R turnout and to group R preferences for candidate B. Similarly, group D preferences for candidate B would be positively related to group R turnout and preferences. These relationships would be described, respectively, by the four across-group covariances  $\sigma_{eDeR}$ ,  $\sigma_{eDvR}$ ,  $\sigma_{vDeR}$ , and  $\sigma_{vDvR}$ . The Appendix demonstrates that the expected values of statistics from Goodman's regression do not depend on these covariances. However, as indicated in footnote 10, these covariances would probably be important in any analysis of the variances of these statistics.

$$(9) \quad E(T_i) = \beta_D x_i + \beta_R (1 - x_i).$$

Equations 5 and 7 imply that the relationship between the determinants of voting preferences,  $\lambda_D$  and  $\lambda_R$ , and the observed ratio of votes for candidate B to the size of the electorate within a precinct is

$$(10) \quad y_{Bi} = (\lambda_D + v_{Di})(\beta_D + \varepsilon_{Di})x_i + (\lambda_R + v_{Ri})(\beta_R + \varepsilon_{Ri})(1 - x_i)$$

$$= x_i(\lambda_D \beta_D + \lambda_D \varepsilon_{Di} + \beta_D v_{Di} + v_{Di} \varepsilon_{Di})$$

$$+ (1 - x_i)(\lambda_R \beta_R + \lambda_R \varepsilon_{Ri} + \beta_R v_{Ri} + v_{Ri} \varepsilon_{Ri}).$$

The statistical character of  $y_{Bi}$  is complicated. As given in equation 10, it depends on four disturbances, rather than one as in conventional regression analysis.<sup>7</sup>

Equation 10 can be rewritten as

$$(11) \quad y_{Bi} = [x_i \lambda_D \beta_D + (1 - x_i) \lambda_R \beta_R]$$

$$+ [x_i \lambda_D \varepsilon_{Di} + x_i \beta_D v_{Di} + (1 - x_i) \lambda_R \varepsilon_{Ri} + (1 - x_i) \beta_R v_{Ri}]$$

$$+ [x_i v_{Di} \varepsilon_{Di} + (1 - x_i) v_{Ri} \varepsilon_{Ri}].$$

The two terms in the first square brackets to the right of the equality in equation 11 are not random. Therefore, they are their own expected values. The four terms in the second square brackets each include only one disturbance. Their expected values are zero.

However, the two terms in the third square brackets of equation 11 each include the

---

<sup>7</sup> In the case of a single application of Goodman's identity, such as equation 8, the dependent variable depends on two disturbances. Zax (2007) demonstrates that the parameters in this case can be estimated with appropriate single-regression techniques.

products of two disturbances. Their expected values are

$$E(x_i v_{Di} \varepsilon_{Di}) = x_i \sigma_{D\varepsilon} \text{ and } E([1-x_i] v_{Ri} \varepsilon_{Ri}) = [1-x_i] \sigma_{R\varepsilon},$$

Consequently, the expected value of  $y_{Bi}$  depends on the four parameters in the identity of equation 5, as well as the within-group covariances between the turnout and cohesion disturbances:

$$(12) \quad E(y_{Bi}) = x_i (\lambda_D \beta_D + \sigma_{D\varepsilon}) + (1-x_i) (\lambda_R \beta_R + \sigma_{R\varepsilon}).$$

None of these parameters can be estimated by regression analysis. Equation 6 suggests that an estimating equation of the form

$$(13) \quad y_{Bi} = b_0 + b_1 x_i + e_i,$$

might be appropriate, where  $b_0$  is the regression intercept,  $b_1$  is the regression slope and  $e_i$  is the regression error term. Equation 13 is an example of “Goodman’s regression”.

However, the expected value of the intercept in equation 13 is<sup>8</sup>

$$(14) \quad E(b_0) = \beta_R \lambda_R + \sigma_{R\varepsilon}$$

The expected value of the slope is

$$(15) \quad E(b_1) = [\beta_D \lambda_D + \sigma_{D\varepsilon}] - [\beta_R \lambda_R + \sigma_{R\varepsilon}].$$

Obviously, these two estimators are insufficient to identify the six parameters upon which they are based.<sup>9</sup> In other words, the application of Goodman’s regression to voting choices by two

---

<sup>8</sup> The Appendix presents derivations of equations 14 and 15.

<sup>9</sup> In equations 14 and 15, the behaviorally implausible assumption that  $\sigma_{1\varepsilon}$  and  $\sigma_{2\varepsilon}$  are equal to zero identifies only the products  $\beta_1 \lambda_1$  and  $\beta_2 \lambda_2$ . The equally implausible assumption of equal turnout rates,  $\beta_1 = \beta_2$ , also fails to identify  $\lambda_1$  and  $\lambda_2$ . Identification requires four assumptions asserting specific values for  $\sigma_{1\varepsilon}$ ,  $\sigma_{2\varepsilon}$ ,  $\beta_1$  and  $\beta_2$ . Any assertion regarding the values of  $\lambda_1$  and  $\lambda_2$  based on the regression of equation 13 must be based on implicit assumptions about these

different electoral groups fails to reveal anything regarding the underlying behavior of either.<sup>10</sup>

### 3. Properties of single equation estimates of voting behavior

In practice, single equation estimates of voting behavior do not estimate equation 13. They employ  $y_{Bi}/T_i$ , the ratio of the votes received by candidate B to the number of votes cast as the dependent variable (Cho, 1998; Kousser 2001; Collet, 2005 and Liu 2007, as examples), rather than  $y_{Bi}$ , the ratio of the votes received by candidate B to the size of the electorate.

This transformation of the dependent variable requires a corresponding transformation of equation 6 in order to preserve Goodman's identity. With this transformation, equation 6 becomes

$$(16) \quad \frac{y_{Bi}}{T_i} \equiv \lambda_{Ri} \beta_{Ri} \frac{1}{T_i} + (\lambda_{Di} \beta_{Di} - \lambda_{Ri} \beta_{Ri}) \frac{x_i}{T_i}.$$

The corresponding Goodman's regression would be<sup>11</sup>

$$(17) \quad \frac{y_{Bi}}{T_i} = b_0 \frac{1}{T_i} + b_1 \frac{x_i}{T_i} + \frac{e_i}{T_i}.$$

However, the regressions in Cho (1998), Kousser (2001), Collet (2005) and Liu (2007)

---

parameters which are generally unstated and, for that matter, unexamined.

<sup>10</sup> Zax (2007) demonstrates that the variance of  $y_{Bi}$  consists of 21 terms, incorporating third- and fourth-order moments of the joint distribution of  $\epsilon_{Di}$ ,  $\epsilon_{Ri}$ ,  $v_{Di}$  and  $v_{Ri}$ . Therefore, even if equation 13 could estimate the parameters of interest, it would be effectively impossible to construct valid confidence intervals or tests of statistical significance.

<sup>11</sup> Proper estimation of equation 17 requires recognition that this equation has no intercept and that the heteroskedasticity inherent in Goodman's regression is compounded because the error term for each precinct is divided by turnout for that precinct.

are of the form

$$(18) \quad \frac{y_{Bi}}{T_i} = d_0 + d_1 x_i + e_i,$$

Equation 18 is not an empirical representation of Goodman's identity because it is not consistent with the transformation required to convert equation 6 into equation 16. This, in itself, implies that the relationships between the estimators  $d_0$  and  $d_1$  and the parameters of interest,  $\beta_1$ ,  $\beta_2$ ,  $\lambda_1$  and  $\lambda_2$ , are unknown.<sup>12</sup>

Moreover, the slope from the OLS regression of equation 18 would be

$$d_1 = \frac{\sum_{i=1}^n [x_i - \bar{x}] \frac{y_{Bi}}{T_i}}{\sum_{i=1}^n [x_i - \bar{x}] x_i},$$

where  $n$  is the number of precincts. Its expected value depends on the expected value of  $y_{Bi}/T_i$ :

$$(19) \quad E(d_1) = \frac{\sum_{i=1}^n [x_i - \bar{x}] E\left(\frac{y_{Bi}}{T_i}\right)}{\sum_{i=1}^n [x_i - \bar{x}] x_i}.$$

---

<sup>12</sup> The conceptual confusion exemplified by the use of equation 18 to estimate equation 16 is apparent in the units of  $d_1$ . The units on the dependent variable are

$$\frac{\text{votes for candidate B}}{\text{all votes}}. \text{ Those on the explanatory variable are } \frac{\text{Group D electorate}}{\text{total electorate}}.$$

Therefore, the units for  $d_1$  must be  $\frac{(\text{votes for candidate B})(\text{total electorate})}{(\text{all votes})(\text{Group D electorate})}$ . These

units, and, consequently, the coefficient, are uninterpretable. Loewen and Grofman (1989, 590-591) agree that equation 18 is internally inconsistent.

Combining equations 8 and 10,  $y_{Bi}/T_i$  is

$$(20) \quad \frac{y_{Bi}}{T_i} = \frac{(\lambda_D + \nu_{Di})(\beta_D + \varepsilon_{Di})x_i + (\lambda_R + \nu_{Ri})(\beta_R + \varepsilon_{Ri})(1 - x_i)}{(\beta_D + \varepsilon_{Di})x_i + (\beta_R + \varepsilon_{Ri})(1 - x_i)}.$$

Equation 20 demonstrates that both the numerator and denominator of the expression for  $y_{Bi}/T_i$  contain disturbance terms. Therefore, the expected value of  $y_{Bi}/T_i$  cannot be derived, at least without extensive additional assumptions regarding the joint distribution of these terms.<sup>13</sup> In the absence of these assumptions, the expected value of  $d_1$  is unknowable.

The same is true of  $d_0$ . The OLS constant is  $d_0 = \bar{y}_0 - d_1\bar{x}$ . Its expected value is

$$(21) \quad E(d_0) = E(\bar{y}_B^*) - \bar{x}E(d_1),$$

where  $\bar{y}_B^*$  is the average of  $y_{Bi}/T_i$  over all precincts. Both expectations to the right of the

equality in equation 21 are unknown. Therefore, the same is true of  $E(d_0)$ .

With  $x_i$  set equal to zero,  $\hat{\lambda}_R = d_0$  is taken to be the estimated ratio of votes for candidate B to all votes cast in a precinct where the population is uniformly of group R. With  $x_i$  set equal to one,  $\hat{\lambda}_D = d_0 + d_1$  is taken to be the estimated ratio of votes for this candidate to all votes cast in a precinct where the population is uniformly of group D. These statistics are offered as the single equation estimators of  $\lambda_R$  and  $\lambda_D$ , the structural voting preferences of the two groups.

---

<sup>13</sup> Any such assumptions would have to be nonintuitive. Moreover, the structural model already contains 14 parameters: two describing turnout in equation 1, two describing voting preferences in equation 4, four variances and six covariances for the disturbance terms. The ecological regression literature does not discuss the latter ten parameters, to say nothing of the additional parameters referred to in footnote 10. It is clearly not prepared to engage in an elaboration requiring yet more parameters.

$E(\hat{\lambda}_R)$  and  $E(\hat{\lambda}_D)$  depend on  $E(d_0)$  and  $E(d_1)$ . Therefore, it is evident that they are unknown as well. Moreover, given the number of parameters that must enter into the expected values of their components, it is inevitable that  $E(\hat{\lambda}_R)$  and  $E(\hat{\lambda}_D)$  are not  $\lambda_R$  and  $\lambda_D$ . In other words, the single equation estimators of voting preferences are biased, and the biases are uncharacterizable.

In sum, the regression of equation 18, despite its popularity, has no relationship with Goodman's identity and yields estimators that are formally uninterpretable. In other words, this regression reveals nothing about  $\beta_D$ ,  $\beta_R$ ,  $\lambda_D$  and  $\lambda_R$ .

Two additional single regression statistics appear as attempts to estimate the extent of polarization. The first is the  $R^2$  statistic from the regression of equation 13 (as examples, Grofman, Migalski and Noviello. 1985, 206; Loewen and Grofman, 1989, 596; Lichtman, 1991, 774 and Grofman, Handley and Niemi, 1992, 91).

However, it is well known that, in this equation<sup>14</sup>

$$(22) \quad R^2 = b_1^2 \frac{V(y_{Bi})}{V(x_i)}$$

Equation 15 demonstrates the relationship between  $b_1$  and the parameters of interest does not allow for any definitive statement regarding the values of those parameters. According to footnote 10, even less can be said about those values based on  $V(y_{Bi})$ . Clearly, the combination of these two statistics, as represented by  $R^2$ , reveals nothing about  $\lambda_D$ ,  $\lambda_R$  or their difference.<sup>15</sup>

---

<sup>14</sup> The proof of equation 22 is provided in, as examples, the introductory econometrics textbooks by Goldberger (1998, 18-19), Gujarati (2003, 84-85) and Murray (2006, 185-186).

<sup>15</sup> Additionally, equation 11 demonstrates that the regression of equation 13 is heteroskedastic. If heteroskedasticity is not corrected, the  $R^2$  statistic cannot be transformed into an F-test for statistical significance. If it is corrected, the  $R^2$  statistic can be less than zero or greater than one (Zax, 2007).

The second auxiliary statistic is the result of “homogeneous precinct analysis”.<sup>16</sup> Collet (2005) is a recent example. This statistic is the average vote share cast for a particular candidate in precincts where a single group constitutes all or almost all of the electorate. This average is

$$(23) \quad \bar{y}_{BH}^* = \frac{\sum_H \left( \frac{y_{Bi}}{T_i} \right)}{n_H},$$

where H represents the subset of precincts identified as “homogeneous” and  $n_H$  represents the number of these precincts.<sup>17</sup>

This statistic has been the subject of substantial criticism.<sup>18</sup> However, its statistical foundations have not been explored. Although previously unacknowledged, homogeneous precinct analysis is simply a restricted version of the single regression in equation 18, in which the slope of that regression is constrained to equal zero and the sample is limited.

If  $d_1=0$ , the sum of squared prediction errors from this regression is

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n \left( \frac{y_{Bi}}{T_i} - d_0 \right)^2.$$

---

<sup>16</sup> This technique is variously referred to as “homogeneous case analysis” (Grofman, Migalski and Noviello, 1985, 203) “homogeneous areas analysis” (Loewen and Grofman, 1989, 600), “extreme case analysis” (Lichtman, 1991, 274) and “homogeneous precinct analysis” (Grofman, Handley and Niemi, 1992, 85).

<sup>17</sup> “Homogeneous precincts” have been defined as those where  $x_i \geq .9$  (Grofman, Migalski and Noviello, 1985, 203),  $x_i \geq .9$  “if possible” (Loewen and Grofman, 1989, 600) and  $x_i \geq .8$  (Lichtman, 1991, 774 and Grofman, Handley and Niemi, 1992, 85).

<sup>18</sup> Langbein and Lichtman (1978), Freedman, Klein, Sacks, Smyth and Everett (1991), Klein, Sacks and Freedman (1991), Firebaugh (1993), Land (1993), Achen and Shively (1995, 94-97 and 107-114), King (1997, 46-55, chapters 9 and 11), Cho (1998), Epstein and O’Halloran (1999), Bourke, DeBats and Phelan (2001, 128) and Kousser (2001, 103) are examples.

The resulting normal equation is

$$\sum_{i=1}^n \left( \frac{y_{Bi}}{T_i} - d_0 \right) = 0.$$

The solution reveals that the intercept estimator is the average value for the dependent variable,

$$d_0 = \frac{\sum_{i=1}^n \left( \frac{y_{Bi}}{T_i} \right)}{n}.$$

If this restricted regression is applied to only the subsample of those precincts designated as “homogeneous”, then

$$d_0 = \bar{y}_{BH}^*.$$

The restriction  $d_1=0$  implies that  $x_i$  does not affect  $y_{Bi}$ . In other words, homogeneous precinct analysis is predicated on the assumption that electoral composition in a precinct does not affect voting behavior in that precinct.<sup>19</sup> This is clearly inconsistent with the motivation for the entire inquiry, which is that voting behaviors differ by group.<sup>20</sup>

---

<sup>19</sup> Paradoxically, advocates of homogeneous precinct analysis (Grofman, 1991 and Lichtman, 1991 as examples) have attacked this assumption in its role as the foundation for the “nonlinear neighborhood model” (Freedman, et al., 1991a, 1991b) without recognizing that it is the basis for their own technique.

<sup>20</sup>  $R^2$  analysis and homogeneous precinct analysis are also mutually inconsistent. In appropriate applications of regression with one explanatory variable,  $R^2$  can be transformed into an F-statistic which tests the statistical significance of the slope  $d_1$  (Goldberger, 1998, 84; Gujarati, 2003, 257; Murray, 2006, 352). In this context, higher values of  $R^2$  correspond to higher probabilities that  $d_1$  is different from zero. In the context of equation 13, this relationship has not been confirmed. Nevertheless, the expectation must be that higher values of  $R^2$  for this equation suggest that the homogeneous precinct restriction of  $d_1=0$  is increasingly unacceptable.

This conceptual confusion restates itself in the statistical properties of  $\bar{y}_{BH}^*$ . As the discussion above demonstrates, the expected value of  $d_0$  is unknown even when  $d_1$  is estimated. The omission of  $x_i$  presumably creates a specification bias which further complicates  $E(d_0)$ .<sup>21</sup> Consequently,  $\bar{y}_{BH}^*$  is even less informative about the parameters of interest than are  $d_0$  and  $d_1$ .<sup>22</sup>

#### 4. Simulations

The previous section demonstrates analytically that estimators from the regression of equation 18 do not identify the parameters of interest or provide proper measures of statistical significance. However, that analysis cannot characterize the inherent biases completely. Three sets of simulations provide additional insight into the empirical performance of these estimators.<sup>23</sup>

The primary focus of these simulations is on the relationship between variations in voting preferences and variations in the accuracy of  $\hat{\lambda}_D$ . In the first set of simulations, group D and group R voters are both indifferent between candidates B and W:  $\lambda_D=.5$  and  $\lambda_R=.5$ . In the second,

---

If so, then homogeneous precinct analysis would be potentially valid only when  $R^2$  is low.

<sup>21</sup> As examples, Goldberger (1998, 108 and 111), Gujarati (2003, 215-217 and 510-513) and Murray (2006, 311-314) discuss specification bias.

<sup>22</sup> The only exceptions are when  $x_i$  equals exactly one for all  $i \in H$  or exactly zero for all  $i \in H$ . In the first case, equation 20 reduces to  $\frac{y_{Bi}}{T_i} = \lambda_D + v_{Di}$ . The average of this value over all precincts in H is an unbiased estimator of  $\lambda_D$ . Similarly, in the second case,  $\frac{y_{Bi}}{T_i} = \lambda_R + v_{Ri}$ . The average of this value over all precincts in H is an unbiased estimator of  $\lambda_R$ .

<sup>23</sup> Zax (2005) summarizes simulations regarding the performance of double regression estimators for the parameters in equation 5. Figures 1, 2 and 3 summarize results regarding the performance of single regression estimators of  $\lambda_D$ , derived from the same simulated data.

their preferences differ strongly:  $\lambda_D=.6$  and  $\lambda_R=.4$ . In the third, their disagreement is almost complete:  $\lambda_D=.9$  and  $\lambda_R=.1$ .

Within each simulation, variations in the combinations of  $\sigma_{Dev}$ ,  $\sigma_{Rev}$ ,  $\beta_D$  and  $\beta_R$  explore the sensitivity of estimates for  $\lambda_D$  to variations in the parameters which determine the values of  $E(b_0)$  and  $E(b_1)$ , as given in equations 14 and 15. The values for the remaining eight parameters referred to in footnote 13 do not vary across simulations.<sup>24</sup>

Variations in the number of simulated “precincts” investigate changes in the performance of these estimates as the amount of underlying information changes. Within each simulation, the composition of the electorate varies uniformly across precincts. Comparisons between simulations which do and do not contain homogeneous precincts assess the sensitivity of estimates to the presence of extreme cases.<sup>25</sup>

With given values for  $\sigma_{Dev}$  and  $\sigma_{Rev}$ , the simulations assign disturbances randomly for each precinct from two truncated bivariate normal distributions, one each for the pair of  $\epsilon_{Di}$  and  $v_{Di}$  and the pair of  $\epsilon_{Ri}$  and  $v_{Ri}$ . The truncation is necessary to ensure that vote shares lie within the feasible range of zero to one, inclusive.<sup>26</sup> With these disturbances, the value of  $x_i$  and the values

---

<sup>24</sup> In these simulations, the four across-group covariances are zero and the ex ante standard deviations of the four disturbances are .3, largely for convenience. There is neither empirical evidence nor intuition regarding their values. None of these parameters appear in equations 14 and 15. This suggests that variations might not have large impacts on point estimates of  $\lambda_D$ . However, as indicated in footnote 10, they would almost certainly appear in the variances of any regression statistics, were those statistics calculated correctly. Consequently, variations in these parameters would probably have important effects on confidence intervals.

<sup>25</sup> All simulations assign parameter values that are independent of  $x_i$ . Consequently, aggregation bias is absent here, as in the analytical discussion of section 2.

<sup>26</sup> The truncated multivariate normal distribution is the natural assumption in this context (King, 1997, chapter 6). The underlying distribution includes one dimension for each of the four disturbance terms. The simulations here require only the truncated bivariate normal distributions

for  $\beta_D$  and  $\beta_R$ , equation 8 yields the value for  $T_i$ . With these values and those for  $\lambda_D$  and  $\lambda_R$ , equation 10 yields the value for  $y_{Bi}$ .

The simulations then calculate the ratio of  $y_{Bi}$  to  $T_i$ , the regression of equation 18 and  $\hat{\lambda}_D$ . They repeat this exercise 100 times for each set of assumed parameter values, numbers of precincts and electoral compositions. This is equivalent to 100 replications of a given election in a given district. Each replication is distinguished solely by the samples of disturbance values assigned to the precincts.

#### A. Electoral indifference and Type I error

Figure 1a presents the first of the simulations, in which the two groups are both indifferent between candidates. Here, each simulated “election” takes place in a “district” of 21 “precincts”.

---

for  $\epsilon_{Di}$  and  $v_{Di}$  and for  $\epsilon_{Ri}$  and  $v_{Ri}$ , by virtue of the assumption that the across-group covariances  $\sigma_{vD\epsilon R}$ ,  $\sigma_{\epsilon DvR}$ ,  $\sigma_{\epsilon D\epsilon R}$  and  $\sigma_{vDvR}$  are all zero. “Rejection sampling” (King, 1997, 143-144) simulates these distributions. It draws random pairs from untruncated bivariate normal distributions with specified parameters and discards them if either disturbance term implies that precinct-specific turnout or cohesion rates would lie outside of the feasible range between zero and one: if  $\epsilon_{ji}$  violates the inequality

$$-\beta_j \leq \epsilon_{ji} \leq 1 - \beta_j$$

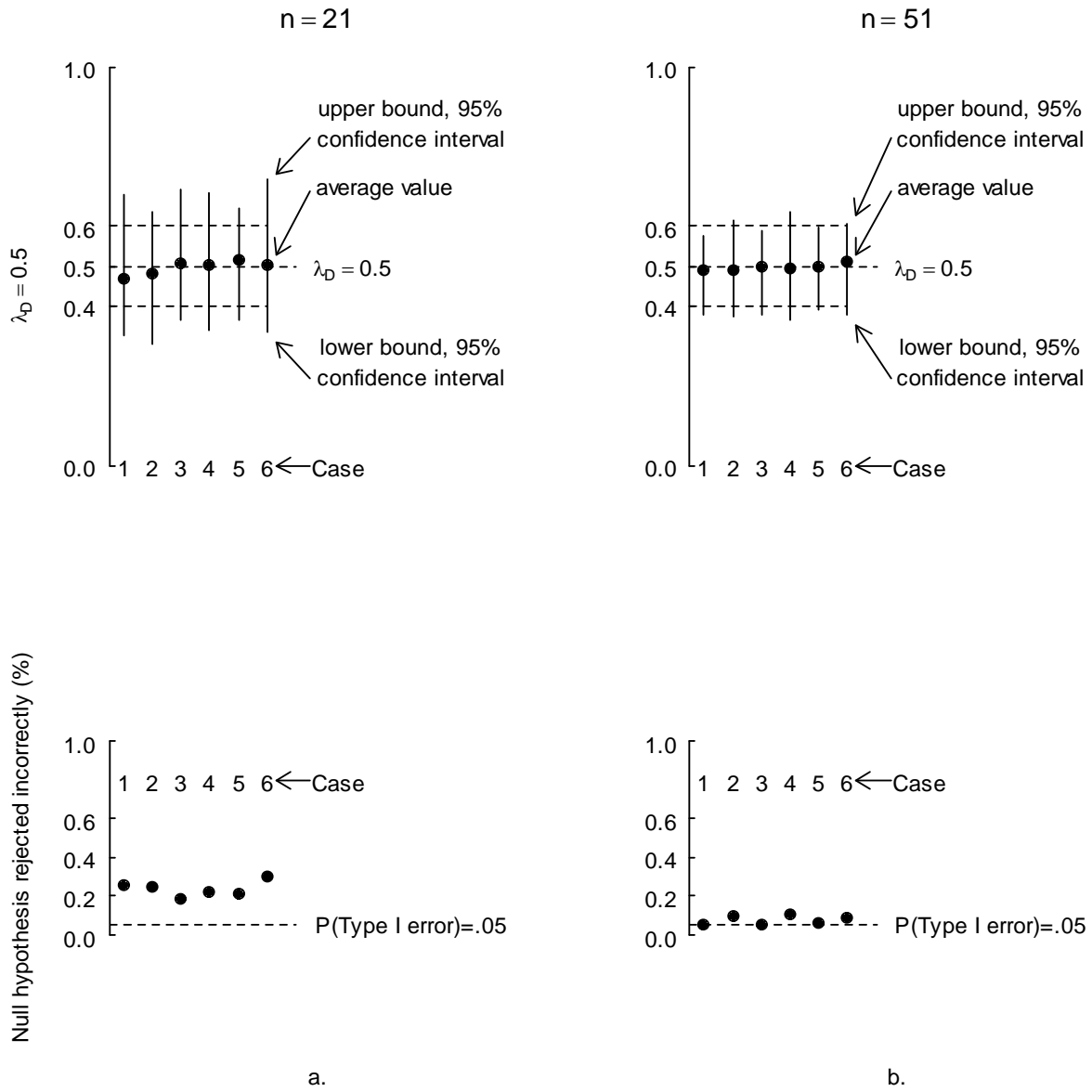
or  $v_{ji}$  violates the inequality

$$-\lambda_j \leq v_{ji} \leq 1 - \lambda_j.$$

Sampling continues for each district until that district has the requisite number of precincts. As a consequence of this procedure, the empirical standard deviations for disturbances and the empirical correlations between disturbances actually included in the sample are smaller in absolute value than the ex ante parameter values.

Figure 1

Single equation estimates of  $\lambda_D = .5$



While these samples are small relative to those in many statistical applications, they are representative of those in Voting Rights litigation regarding local jurisdictions. Precincts vary according to a discrete uniform distribution from homogeneous of group R,  $x_i=0$ , to homogeneous of group D,  $x_i=1$ .

The six cases in figure 1a represent six different combinations of implicit values for  $\sigma_{Dev}$  and  $\sigma_{Rev}$ . These combinations correspond to three explicit ex ante values for each of the within-group correlations between disturbances  $\rho_{Dev}$  and  $\rho_{Rev}$ :  $-.5$ ,  $0$  and  $.5$ . With identical parameter values across groups for all other parameters, only six pairs of correlations  $(\rho_{Dev}, \rho_{Rev})$  are distinct:  $(-.5, -.5)$  for case 1,  $(-.5, 0)$  for case 2,  $(-.5, .5)$  for case 3,  $(0, 0)$  for case 4,  $(0, .5)$  for case 5 and  $(.5, .5)$  for case 6. The pairs  $(0, -.5)$ ,  $(.5, -.5)$  and  $(.5, 0)$  are redundant with cases 2, 3 and 5, respectively, because of symmetry.

In the top panel of figure 1a, the dashed horizontal line at  $\lambda_D=.5$  gives the true value of the parameter of interest. The point associated with each case gives the average value of  $\hat{\lambda}_D$  over the 100 replications in that case. This panel demonstrates that average values of  $\hat{\lambda}_D$  are surprisingly close to the actual value of  $\lambda_D$ , regardless of the values for  $\rho_{Dev}$  and  $\rho_{Rev}$ . In this simulation, the biases in  $\hat{\lambda}_D$  appear to be small and essentially invariant to correlation values.<sup>27</sup>

However, the associated confidence intervals are unacceptably large. The vertical line associated with each case indicates the empirical 95% confidence interval for  $\lambda_D$ . This interval is

---

<sup>27</sup> Zax (2005) demonstrates that, with these parameter values, double regression estimates are biased away from indifference. In this context, single regression appears to be superior.

essentially the range of estimates from the replications after discarding the two or three most extreme estimates at either end.

In the first case of figure 1a, 26 of the 100 simulations yield values of  $\hat{\lambda}_D$  equal to or above .6, or equal to or below .4. If values in these ranges are taken to be inconsistent with the null hypothesis of indifference,  $H_0: \lambda_D=.5$ <sup>28</sup>, these simulations would incorrectly reject that hypothesis in 26% of all cases. This frequency is far in excess of 5%, the conventional limit on acceptable probabilities for type I error.

This is illustrated by the lower panel of figure 1a. The point corresponding to each case represents the proportion of simulations yielding estimates that are inconsistent with the true parameter value  $\lambda_D=.5$ . The dashed line represents the conventional 5% standard for type I error.

This panel demonstrates that the frequency of type I error varies more across pairs of correlation values than does the average estimate of  $\lambda_D$ . However, it exceeds the standard of 5% by approximately a factor of four or more in every case. Over all 600 simulations, the frequency is 23.8%. Consequently, the regression of equation 18 is not a statistically acceptable strategy for estimating  $\lambda_D$  in the circumstances of this simulation, despite the appearance of modest bias.

In actual circumstances, D typically represents a class of electoral minorities. Frequently, minority turnout rates are much lower than those of majority voters. In these circumstances, the performance of equation 18 is considerably worse than the illustration in figure 1a. If the turnout rate for group D is one-half the turnout rate for group R,  $\beta_D=.25$ , simulations with the same remaining details as figure 1a yield average values of  $\hat{\lambda}_D$  that are somewhat farther from .5 and

---

<sup>28</sup> Zax (2005, 70) discusses the justification for this hypothesis test.

type I errors in 33.0% of all replications.<sup>29</sup>

Figure 1b replicates the simulation of figure 1a with 51 rather than 21 precincts per election. The increase in the number of precincts noticeably increases the accuracy of the single regression estimates. Average values for  $\hat{\lambda}_D$  are closer to the true value of  $\lambda_D=.5$ , and, more importantly, confidence intervals are much smaller. As reported in the lower panel of figure 1b, the frequency of type I error is equal to the conventional threshold of 5% in cases 1 and 3. It is higher in the remaining cases, but only in case 4 does it exceed the minimal standard of 10%.

Unfortunately, this improvement is specific to the specification of this simulation. In figure 1b, the overall frequency of  $\hat{\lambda}_D$  estimates that are inconsistent with the true value of .5 is 7.7%. If, instead, group D turnout is  $\beta_D=.25$ , average values for  $\hat{\lambda}_D$  are again farther from the true value of  $\lambda_D$  and 11% of all simulations incorrectly reject the null hypothesis of indifference.

The simulations represented by figure 1 provide little support for the assertion that vote shares in “homogeneous precincts” are reliable indicators of group D voting preferences. Each of the 600 replications of figure 1a contains three precincts in which  $x_i \geq .9$ . The average value for  $y_{Bi}/T_i$  over these three precincts is outside the interval of .4 through .6, inclusive, in 255. In other words, the hypothesis test above, applied to the homogeneous precinct estimator, would incorrectly reject the null hypothesis of indifference in 42.5% of all cases.

The performance of this test improves slightly in the simulations represented by figure 1b. With 51 precincts in each replication,  $x_i \geq .9$  in six. The average value for  $y_{Bi}/T_i$  over these six

---

<sup>29</sup> The author will provide complete details of these simulations, as well as all others discussed here, upon request.

precincts rejects the null hypothesis of indifference in 172, or 28.7% of all replications.

Additional simulations with the same values of other parameter suggest that the homogeneous precinct estimator requires at least 20 “homogeneous precincts” in order to reduce the rate of type I errors to acceptable levels.

Similarly, the simulations of figure 1 suggest that the  $R^2$  statistic from the regression of equation 13 has no diagnostic value. With either 21 or 51 precincts per election, there is little systematic correlation between this statistic and the accuracy of  $\hat{\lambda}_D$ . If anything, there may be a slight tendency for high values of  $R^2$  to be associated with underestimates of  $\lambda_D$  when both correlations are equal to .5.

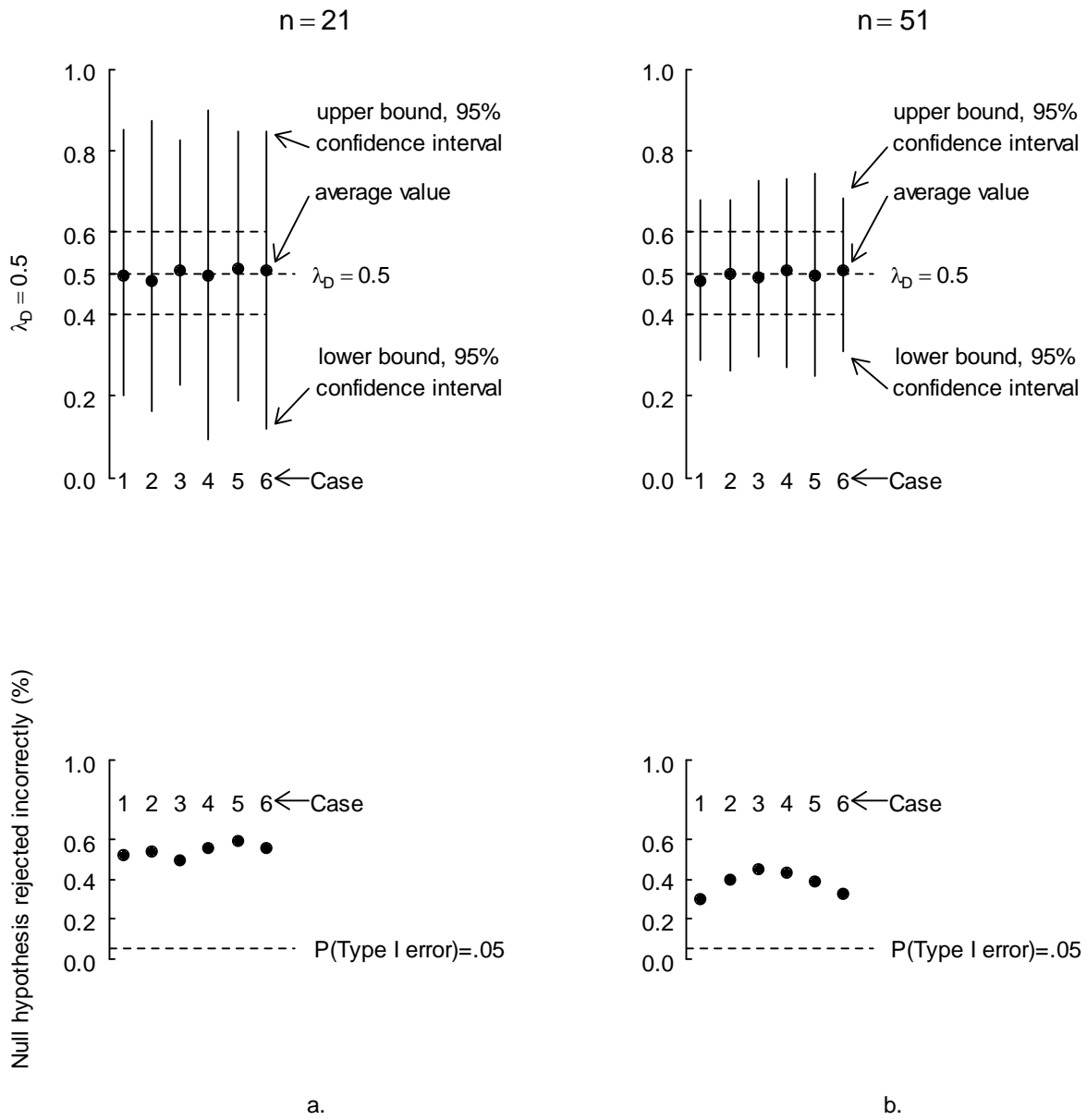
Figure 2 further demonstrates that the performance of  $\hat{\lambda}_D$  depends heavily on the specific simulation context. The simulations reported in figures 2a and 2b replicate those of figures 1a and 1b, with the exception that those in figure 2 contain no homogeneous precincts. Instead,  $x_i$  varies according to a uniform discrete distribution from .3 to .7.

The top panels of figures 2a and 2b indicate that the average values of  $\hat{\lambda}_D$  are not sensitive to the absence of extreme cases. As in figures 1a and 1b, those averages are very close to the true value of .5.

However, the accuracy of any individual estimate of  $\lambda_D$  is much worse. With 21 precincts, empirical confidence intervals typically have lower bounds of .2 or less and upper bounds of .8 or more. Rates of type I error exceed 50% in all six cases. In replications with  $\beta_D=.25$ , rates of type I error reach nearly 60%.

Figure 2

Single equation estimates of  $\lambda_D = .5$ , no homogeneous precincts



The increase in information associated with the sample size of 51 precincts compensates only partially for the loss of information in the extreme cases. All six confidence intervals in the upper panel of figure 2b have lower bounds below .4 and upper bounds above .6. Rates of type I error, as given in the lower panel of that figure, vary from 30% to 45%. With  $\beta_D=.25$ , the rate of type I error across all 600 replications is 42.2%.

In sum, figures 1 and 2 demonstrate that the regression of equation 18 is generally unsuitable for the purpose of estimating the electoral choices of group D when both groups are truly indifferent between candidates. Averages of many election simulations yield estimates of  $\lambda_D$  that are surprisingly close to the true value. However, the estimate provided by any individual election has an unacceptably high probability of being wrong. Rates of type I error in every simulation except those of figure 1b are astronomical. Even in the case of figure 1b, those rates approach acceptable levels only if at least some individual precincts are highly homogeneous.<sup>30</sup>

#### B. Electoral preferences and Type II error

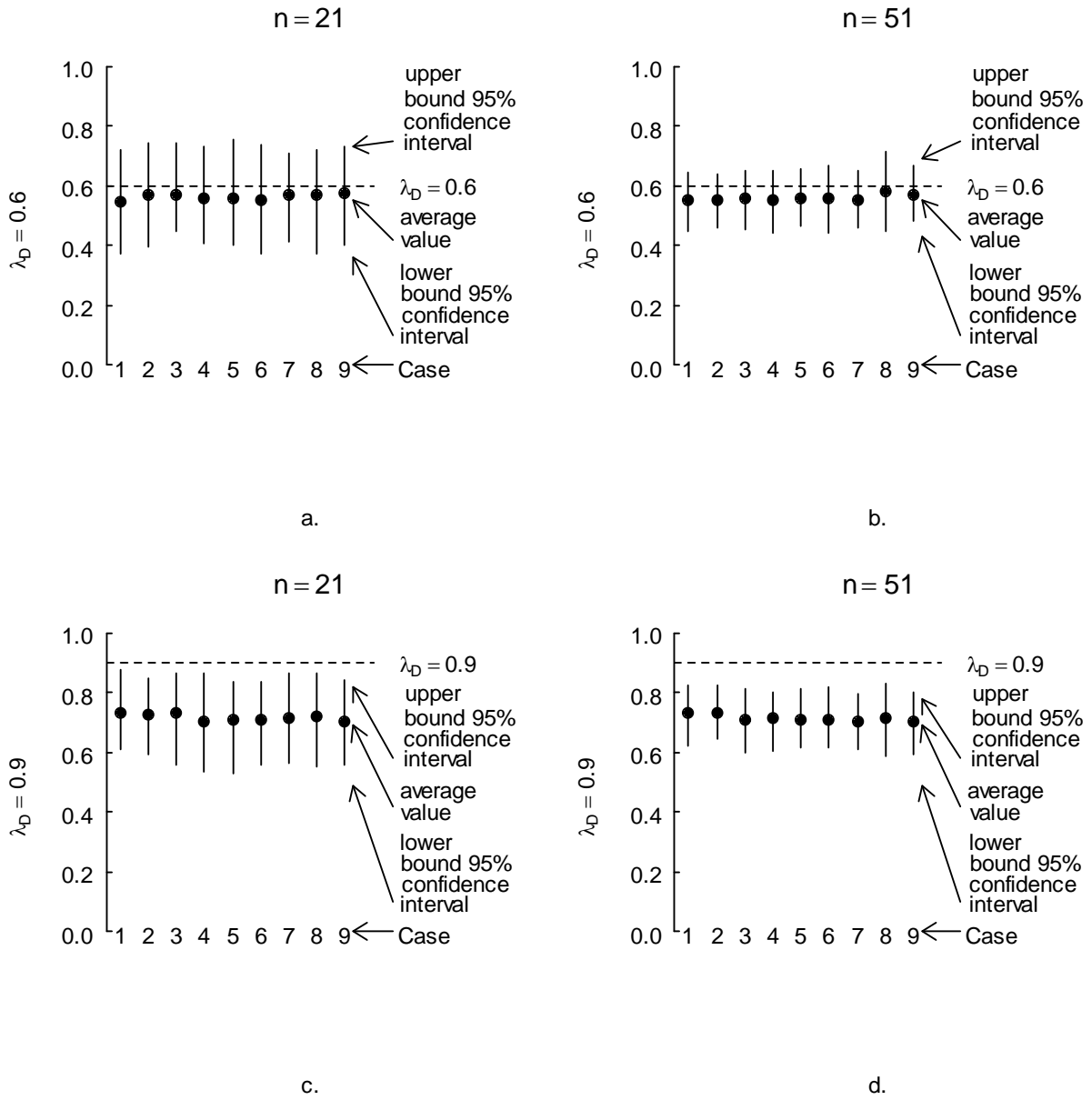
Figure 3 presents simulations in which group D and group R voters have different voting preferences. These differences imply that the cases where  $(\rho_{Dev}, \rho_{Rev})$  take the values  $(0, -.5)$ ,  $(.5, -.5)$  and  $(.5, 0)$  are distinct from those in which their values are  $(-.5, 0)$ ,  $(-.5, .5)$ , and  $(0, .5)$ . Consequently, each of these simulations has nine cases. The values for  $(\rho_{Dev}, \rho_{Rev})$  are  $(-.5, -.5)$  for case 1,  $(-.5, 0)$  for case 2,  $(-.5, .5)$  for case 3,  $(0, -.5)$  for case 4,  $(0, 0)$  for case 5,  $(0, .5)$  for

---

<sup>30</sup> Changes in the value of  $\beta_R$  should have effects on the results that are similar in magnitude to changes in the value of  $\beta_W$  because the simulations in figures 1 and 2 treat groups D and R symmetrically, . Partial simulations suggest that results are not sensitive to the true value of  $\lambda_R$ .

Figure 3

Single equation estimates of  $\lambda_D=0.6$  and  $\lambda_D=0.9$



case 6, (.5, -.5) for case 7, (.5, 0) for case 8 and (.5, .5) for case 9. These simulations yield results that differ, in some circumstances dramatically, from those in which preferences are identical.

Figure 3a presents the results of the nine simulations in which  $\lambda_D=.6$ ,  $\lambda_W=.4$  and the number of precincts is 21. The average point estimate in each is less than the true value. The biases in figure 3a are typically small: average values for  $\hat{\lambda}_D$  are between .548 and .576.

However, these biases are sensitive to other parameter values. If  $\beta_D=.25$  rather than .5, average values for  $\hat{\lambda}_D$  vary from .502 to .603. They are especially low when  $\rho_{Dev}=-.5$ .

Empirical confidence intervals for  $\lambda_D$  in all nine cases of figure 3a include the true value of .6. However, if  $\hat{\lambda}_D$  were an unbiased estimator of  $\lambda_D$  and its values were distributed symmetrically around .6, half of the 900 estimates generated by the simulations of figure 3a should lie below this value. Of the actual estimates, 584 are less than .6. This indicates that  $\hat{\lambda}_D$  is biased in the direction of understating the degree of electoral preference when groups are not indifferent.

Figure 3b replicates this simulation with 51 precincts in each election. Increased information for each election yields improved estimates. All average values for  $\hat{\lambda}_D$  are again below but close to the true value of  $\lambda_D=.6$ . All empirical confidence intervals include the true value and are much smaller than those of figure 3a. However, they are so narrow that the true value lies close to the upper bounds of many. Moreover, 691 of the 900 values for  $\hat{\lambda}_D$  are less than .6.

These distortions are exacerbated with lower turnout among group D. If  $\beta_D=.25$ , average

values for  $\hat{\lambda}_D$  can again be as low as .5, especially if  $\rho_{D\epsilon v} = -.5$ . Confidence intervals shift downwards as well, to the point where the true value is only just within some upper bounds.

These distortions are also exacerbated if voting preferences of the two groups are more disparate. As in figure 3a, each election in figure 3c has 21 precincts. However,  $\lambda_D = .9$  and  $\lambda_R = .1$ . The consequence is that, in all cases, the average value of  $\hat{\lambda}_D$  is less than .74. These averages underestimate the true value of .9 by nearly 20%. They are even smaller with  $\beta_D = .25$ .

None of the confidence intervals in figure 3c contain the true value. At the same time, 65, or 7.2% of all elections yield values of  $\hat{\lambda}_D$  that are less than .6. If estimates in this range are taken as consistent with the null hypothesis that group D voters are effectively indifferent, than these simulations are more likely to commit this type II error than to identify actual group D preferences. This imbalance is even greater with  $\beta_D = .25$ .

With  $\lambda_D = .9$ , increased information does not necessarily improve performance. The elections simulated in figure 3d have 51 precincts. They yield point estimates of  $\lambda_D$  that are similar to those in figure 3c. However, the confidence intervals are substantially narrower. Consequently, the true value  $\lambda_D = .9$  lies even further above the upper bounds of all nine. The only benefit is that the frequency of type II errors,  $.4 < \hat{\lambda}_D < .6$ , is reduced to 1.7%.

With  $\lambda_D = .9$  and 51 precincts, statistical performance deteriorates in all dimensions if  $\beta_D = .25$ . Average estimates of  $\lambda_D$  are two to five percentage points lower across the nine simulations than in figure 3d. Confidence intervals shift down by similar amounts. The frequency of type II errors increases commensurately, to 7.7%.

Homogeneous precinct analysis also fails to provide accurate estimates of  $\lambda_D$  when group D is not indifferent between candidates. In figures 3a and 3b, where  $\lambda_D=.6$ , the average values of  $\bar{y}_{BH}^*$  across all 100 simulations are similar to average values of  $\hat{\lambda}_D$ . However,  $\bar{y}_{BH}^*$  is .4 or less for 10.7% of all simulations in figure 3a and 5.4% of all simulations in figure 3b. In these elections, homogeneous precinct analysis would incorrectly predict that group D prefers candidate B's opponent.

In figures 3c and 3d, where  $\lambda_D=.9$ , average values for  $\bar{y}_{BH}^*$  are approximately two percentage points below the already low averages for  $\hat{\lambda}_D$ . Moreover, with 21 precincts, 17.5% of values for  $\bar{y}_{BH}^*$  are less than .6. With 51 precincts, this occurs in 11% of all elections. In these elections, homogeneous precinct analysis is, incorrectly, consistent with the null hypothesis that group D is indifferent to candidate B.

Correlation analysis is also unreliable. Across all simulations in figure 3, values of the  $R^2$  statistic for the regression of equation 13 are positively correlated with values for  $\hat{\lambda}_D$  from the regression of equation 18. In simulations where  $\lambda_D=.6$ , the consequence is that estimates of  $\lambda_D$  from regressions associated with the highest  $R^2$  substantially overestimate the true value. In simulations where  $\lambda_D=.9$ , estimates of  $\lambda_D$  from regressions associated with the highest  $R^2$  approach but still understate the true value.

Lastly, extreme cases generally limit the biases in the regression of equation 18. The simulations of figure 3, replicated with electoral composition varying from  $.3 \leq x_i \leq .7$  rather than

from  $0 \leq x_i \leq 1$ , yield average values of  $\hat{\lambda}_D$  that are slightly below those of the corresponding simulation in figure 3. Consequently, the biases are slightly larger. In addition, confidence intervals are much larger, to the point where the upper bounds exceed the feasible maximum of one in five of the nine cases with the parameters of figure 3c.

## **5. Conclusion**

This paper demonstrates that conventional single regression estimators of voting preferences for groups within the electorate are unreliable when group-specific turnout rates are unknown. These estimators are almost certainly biased, have unknown variances and no valid measures of statistical significance. Simulations with reasonable parameter values demonstrate that average values for these estimators may be relatively close to true values when voters are indifferent to a candidate, but typically understate the intensity of preferences when they are not. Type I errors, in the first case, and type II errors in the second occur with frequencies well above those that are ordinarily acceptable.

Clearly, these estimators do not have the properties ordinarily required by the empirical social sciences. Their continued use cannot be justified, especially with the ready availability of superior alternatives (King, 1997; Rosen, et al., 2001 and Wakefield. 2004) .

## Appendix

### Proof of equation 14:

The OLS formula for  $b_1$  is

$$b_1 = \frac{\sum_{i=1}^n [x_i - \bar{x}] y_{Bi}}{\sum_{i=1}^n [x_i - \bar{x}] x_i}.$$

Consequently,

$$E(b_1) = \frac{\sum_{i=1}^n [x_i - \bar{x}] E(y_{Bi})}{\sum_{i=1}^n [x_i - \bar{x}] x_i}.$$

Replacing  $E(y_{Bi})$  as indicated by equation 12 yields

$$\begin{aligned} E(b_1) &= \frac{\sum_{i=1}^n [x_i - \bar{x}] [x_i [\lambda_D \beta_D + \sigma_{Dev}] + [1 - x_i] [\lambda_R \beta_R + \sigma_{Rev}]]}{\sum_{i=1}^n [x_i - \bar{x}] x_i} \\ &= [\lambda_D \beta_D + \sigma_{Dev}] - [\lambda_R \beta_R + \sigma_{Rev}]. \end{aligned}$$

### Proof of equation 15:

The OLS formula for  $b_0$  is

$$b_0 = \bar{y}_D - b_1 \bar{x}.$$

Therefore,

$$E(b_0) = E(\bar{y}_D) - \bar{x} E(b_1) = E(\bar{y}_D) - \bar{x} [[\lambda_D \beta_D + \sigma_{Dev}] - [\lambda_R \beta_R + \sigma_{Rev}]].$$

Equation 12 for  $E(y_{1i})$  yields

$$E(\bar{y}_D) \equiv \bar{x} [\lambda_D \beta_D + \sigma_{Dev}] + [1 - \bar{x}] [\lambda_R \beta_R + \sigma_{Rev}].$$

Consequently,

$$E(b_0) = \bar{x} [\lambda_D \beta_D + \sigma_{Dev}] + [1 - \bar{x}] [\lambda_R \beta_R + \sigma_{Rev}] - \bar{x} [[\lambda_D \beta_D + \sigma_{Dev}] - [\lambda_R \beta_R + \sigma_{Rev}]] = [\lambda_R \beta_R + \sigma_{Rev}].$$

## References

- Achen, Christopher H. and W. Phillips Shively (1995) Cross-Level Inference, The University of Chicago Press, Chicago.
- 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) "If the assumption fits: A comment on the King ecological inference solution", Political Analysis, Vol. 7, No. 1, Winter, 143-163.
- 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.
- Epstein, David and Sharyn O'Halloran (1999) "Measuring the electoral and policy impact of majority-minority voting districts", American Journal of Political Science, Vol. 43, No. 2, April, 367-395.
- Firebaugh, Glenn (1993) "Are bad estimates good enough for the courts?", Social Science Quarterly, Vol. 74, No. 3, September, 488-495.
- 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.
- Goldberger, Arthur S. (1998) Introductory Econometrics, Harvard University Press, Cambridge, MA.
- Goodman, Leo A. (1953) "Ecological regressions and behavior of individual", American Sociological Review, Vol. 18, No. 6, December, 663-664.
- Goodman, Leo A. (1959) "Some alternatives to ecological correlation", American Journal of Sociology, Vol. 64, No. 6, May, 610-625.
- Grofman, Bernard (1991) "Statistics without substance: A critique of Freedman et al. and Clark and Morrison", Evaluation Review, Vol. 15, No. 6, December, 746-769.
- Grofman, Bernard, Lisa Handley and Richard G. Niemi (1992) Minority Representation and the Quest for Voting Equality, Cambridge University Press, New York.
- 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.

Gujarati, Damodar N. (2003) Basic Econometrics, Fourth Edition, McGraw-Hill Irwin, Boston, MA.

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

Klein, Stephen P., Jerome Sacks and David A. Freedman (1991) “Ecological regression *versus* the secret ballot”, Jurimetrics Journal, Vol. 31, No. 4, Summer, 393-413.

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

Land, Kenneth C. (1993) “Discriminatory electoral practices, contextual effects, and a new double regression method for the courts” Social Science Quarterly, Vol. 74, No. 3, September, 469-470.

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.

Langbein, Laura and Allan J. Lichtman (1978) Ecological Inference, Sage Publications, Beverly Hills.

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

Loewen, James W. and Bernard Grofman (1989) “Comment: Recent developments in methods used in vote dilution litigation”, The Urban Lawyer, Vol. 21, No. 3, Summer, 589-604.

Murray, Michael P. (2006) Econometrics: A Modern Introduction, Pearson Addison Wesley, Boston, MA.

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

Wakefield, Jon (2004) “Ecological inference for 2×2 tables”, Journal of the Royal Statistical Society: Series A, Vol. 167, No. 3, July, 385-574.

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

Zax, Jeffrey S. (2007) “Fifty years of Goodman’s identity: Its implications for regression-based inference”, working paper, University of Colorado at Boulder, Boulder, CO.