Individual Choice and Ecological Analysis

Kenneth F. McCue*

California Institute of Technology

July 17, 1995

*Research Scientist, Environmental Quality Laboratory, California Institute of Technology.

Abstract

The use of the linear probability model in aggregate voting analysis has now received widespread attention in political science. This article shows that when the linear probability model is assumed to be consistent for the choice of the individual, it is actually a member of a general class of models for estimating individual responses from aggregate data. This class has the useful property that it defines the aggregate analysis problem as a function of the individual choice decisions, and allows the placement of most aggregate voting models into a common probabilistic framework. This framework allows the solution of such problems as inference of individual responses from aggregate data, estimation of the transition model, and the joint estimation and inference from individual and aggregate data. Examples with actual data are provided for these techniques with excellent results.

1 Introduction

The problem of obtaining consistent estimators for parameters of the individual behavior from aggregate data has bedeviled the social sciences ever since the problems were illustrated by Robinson (1950). The method which has had the most enduring popularity is that proposed by Goodman (1953, 1959), which, if there are two groups in the electorate (say white and blacks) voting on a candidate, says to estimation the following equation by linear regression

$$Y = X_1\beta_1 + X_2\beta_2 + U, \tag{1}$$

where Y is the votes cast for a candidate in an electoral unit, X_i are the number in group i, and β_i are the parameters to be estimated. One purpose of this article is to examine what the implications are for (1) when it is assumed that the parameters estimated from (1) are unbiased when estimated with individual data, that is, if the following equation is estimated,

$$y = x_1\beta_1 + x_2\beta_2 + u \tag{2}$$

where y is zero or one depending upon whether the individual chose the candidate, x_1 is one if the voter is in group one, zero otherwise, and vica versa for x_2.

The insight which is necessary to relate this problem to models where a probability law is assumed on this individual is simply this: when the estimation of (2) is unbiased (or consistent) for the parameters β_i then the residuals and the probability distribution of those residuals is completely determined, as is shown in standard textbooks on the subject such as Maddala (1983). Knowledge of this distribution does several things. First, it allows the derivation of an expression for possible "bias" due to grouping in the aggregate estimator when the estimator derived from individual data would be consistent, a problem which has recently attracted scholarly interest (see Palmquist (1993), Ansolabehere and Rivers (1994), and King (1995)). Second, and more importantly, it allows the placement of (1) into a more general class of models which have desirable properties, such as estimation by maximum likelihood and coherence with the modelling of individual choice in political analysis. This general class of models also allows a discussion of the problems of estimating aggregate models in terms of the problems of estimating individual models.

The strategy of this article, is as follows. The first section examines the linear probability model under the light about what is known about the distribution of the residuals. The second section examines the linear probability model into the more general framework of probability theory, and suggests what implications that general framework has for problems in aggregate analysis. The next three sections suggests some specific applications in political science that derive from that more general framework. The first of these sections derives the distribution of the estimators from the two-stage linear probability model for estimating both turnout and vote choice (first introduced by Kousser (1973)) and some potential modifications which have been suggested to that procedure by such authors as Shivley (1975,1991) and King (1995). The next section includes a demonstration of the misspecification of the traditional transition model (also the Goodman model) under consistency of the individual choice decision and the derivation and test of a consistent estimator. The last of these sections examines the estimation of parameters from both aggregate and individual data and

methods of combining and comparing such estimators. These three sections show the power of placing the aggregate analysis problem directly in the context of probability theory, as each solves a previously unsolved problem in aggregate analysis.

2 The Linear Probability Model

For ease of exposition rewrite (2)

$$\begin{aligned} y &= \beta_1 + x_2(\beta_2 - \beta_1) + u, \\ y &= \alpha + x_2\beta + u, \\ y - \bar{y} &= x_2\beta + u, \\ w &= xB + \epsilon, \end{aligned} \tag{3}$$

with the first equality deriving from the fact that $x_1 = 1 - x_2$, and where $\bar{y}$ is the mean of the y's, w is defined as $y - \bar{y}$, and $x = x_2$. Since $\alpha = \bar{y}$ when (2) is correctly specified, (3) also produces unbiased estimates of $B = \beta = \beta_2 - \beta_1$.

If there are n individuals grouped into N electoral units, let R be an N cross n grouping matrix with the i^{th}, j^{th} element of R being one if the j^{th} individual is in the i^{th} electoral unit, zero otherwise.[1] The matrix form of (3) is

$$W = XB + E, \tag{4}$$

where X is the matrix of the x, W is the matrix of the w, and E is the matrix of the ϵ, all of these matrices being n by 1. Then premultiplying (4) by R gives

$$RW = RXB + RE, \tag{5}$$

An estimation of (5) by the method of least squares gives

$$\begin{aligned} \hat{B} &= [X'R'RX]^{-1}X'R'RW, \\ &= B + \frac{X'R'RE}{X'R'RX}. \end{aligned}$$

Now, the j^{th} element of $X'R'$ (which is 1 cross N) is of the form $\sum_{i=1}^{n_j} x_{ji}$, where n_j are the number of individuals in electoral unit j, and x_{ji} is the value of x for the i^{th} individual in electoral unit j. The j^{th} element of RE (N cross 1) can be found similarly to be $\sum_{i=1}^{n_j} \epsilon_{ji}$. This then gives

$$\frac{X'R'RE}{[X'R'RX]} = \frac{\sum_{j=1}^{N}\sum_{i=1}^{n_j} x_{ji}\sum_{i=1}^{n_j}\epsilon_{ji}}{\sum_{j=1}^{N}\sum_{i=1}^{n_j} x_{ji}\sum_{i=1}^{n_j} x_{ji}} \tag{6}$$

[1] More complicated forms of the grouping matrix are available, which create group means for the individuals in the group (see Erbing (1989) or Palmquist (1993)). This form is sufficient for the purposes here.

Taking the expectation of this conditional on the x, if the expectation of each ϵ_{ji} is zero, then the sum of the expectations is zero. Thus the aggregate model given by (5) is conditionally unbiased.

All of the above has been done without any knowledge of the distribution of the residuals. This then raises the question of why there is now a large amount of work finding formulas for "biased" estimators (see Palmquist, Ansolabehere and Rivers, and King), all of them derived under the assumption that (4) is the true model. The thinking here is that the grouping process is correlated with the error term,[2] so that $E[\sum_{i=1}^{n_j} \epsilon_{ji}] \neq 0$. Note that since the expectation of a sum of random variables (the ϵ_{ji}) is the sum of the expectations that implies that the individual expectations are also non-zero. Since the residuals are still the same (if the individual model is to be unbiased) this implies that the probabilities of the residuals occurring within this group have been changed, while still remaining the same for the overall collection of individuals. Can the probability be altered while still maintaining consistency of the individual level estimation?

Here is where knowledge of distribution of the residuals under the assumption of consistency of the estimators from the individual equation is needed. Equation (2) falls in the class of linear probability models, that is, a regression model in which the dependent variable y is a binary variable taking the value of one if the event occurs and zero if it does not. Obviously, the residuals can only take on the values $1-\alpha-\beta x$ and $-\alpha-\beta x$, the first value occurring when $y=1$ (or $w=1-\alpha$), the second when $y=0$ (or $w=-\alpha$). Note

$$E[\epsilon|x] = (1-a)(1-\alpha-\beta x) + a(-\alpha-\beta x) \tag{7}$$

where $1-a$ is the probability of $(1-\alpha-\beta x)$ occurring and a is the probability of $-\alpha-\beta x$ occurring. Then equating this to zero gives $a = 1-\alpha-\beta x$, so the residuals of (3) are $1-\alpha-\beta x$ with probability $\alpha+\beta x$ and $-\alpha-\beta x$ with probability $1-\alpha-\beta x$. Maddala points out that $\alpha+\beta x$ has to be interpreted as the probability of y occurring given the x, which is also the probability of the residual with value $1-\alpha-\beta x$ occurring, so the derived probability is the same as the interpreted probability. This conditional unbiasedness implies consistency, that is, the plim $X'E/n = 0$.[3]

These results can then be used to alter the probabilities of the residuals occurring inside

[2]Usually. There is also a school of thought which assumes that aggregation in and of itself produces bias. See Firebaugh (1978), Erbing and Palmquist. The results of the next section are needed to examine this claim in its full generality. It should be noted, though, that if there is a variable which is correlated with the error term and does not bias the individual level estimate, then one can condition the error term on that variable and obtain an error term with a reduced variance by the Rao-Blackwell theorem. Thus the estimation of the individual-level model will be inefficient. Another implication is that one of the usual methods of determining the adequacy of a regression model, correlating estimated residuals with variables not in the model, will show a non-zero correlation with the correlated variable, leading, by the usual procedure, of putting this variable into the model, and thus estimating a misspecified model.

[3]Unbiasedness implies consistency but one can have consistency without unbiasedness. In the linear probability model, though, since plim $X'E/n$ = plim $\sum_{y=1}(1-\alpha-\beta)/n$ − plim $\sum_{y=0}(\alpha+\beta)/n$ = Pr$(y=1|x=1)(1-\alpha-\beta)$ − Pr$(y=0|x=1)(\alpha+\beta)$, the interpretation of Pr$(y=1) = \alpha+\beta x$ implies that Pr$(y=1|x=1) = \alpha+\beta$ and Pr$(y=0|x=1) = 1-\alpha-\beta$, and these probabilities establish conditional unbiasedness, so consistency implies unbiasedness in this case.

Table 1: Joint PDF of $f(\epsilon_{t-1}, \epsilon_t)$

	$\alpha + \beta x$	$1 - \alpha - \beta x$
$\alpha + \beta x$	$(\alpha + \beta x)^2 + \theta_{11}$	$(\alpha + \beta x)(1 - \alpha - \beta x) + \theta_{12}$
$1 - \alpha - \beta x$	$(\alpha + \beta x)(1 - \alpha - \beta x) + \theta_{21}$	$(1 - \alpha - \beta x)^2 + \theta_{22}$

any grouping while still maintaining the the unbiasedness at the individual model. Presumably the process by which this altering of probabilities takes place is that there is a serial correlation between between observed residuals at time $t-1$ and at time t, and between these two times some entity rearranges matters based on the observations of the residuals at time $t-1$. Keeping the requirement that the linear probability model estimated from (4) be consistent at both time $t-1$ and time t, write the joint pdf of the residuals $\epsilon_{t-1}, \epsilon_t$ as given in Table 1.[4] In Table 1 the correlational terms θ are written as deviations from the product of the individual probabilities. It can easily be confirmed that $\theta_{11} + \theta_{12} = 0$, $\theta_{11} + \theta_{21} = 0$, $\theta_{12} + \theta_{22} = 0$, and $\theta_{21} + \theta_{22} = 0$. All of these equalities taken together imply that $\theta = -\theta_{12} = -\theta_{21} = \theta_{11} = \theta_{22}$. Consistency with non-negativity of probabilities also implies that θ is less than the minimum of $(\alpha + \beta x)^2$ and $(1 - \alpha - \beta x)^2$.[5]

Conditional probabilities will be needed and can also be calculated from Table 1. They are

$$\begin{aligned}
\Pr(\epsilon_t = 1 - \alpha - \beta x | \epsilon_{t-1} = 1 - \alpha - \beta x) &= \alpha + \beta x + \frac{\theta}{\alpha + \beta x}, \\
\Pr(\epsilon_t = -\alpha - \beta x | \epsilon_{t-1} = 1 - \alpha - \beta x) &= 1 - \alpha - \beta x + \frac{\theta}{\alpha + \beta x}, \\
\Pr(\epsilon_t = 1 - \alpha - \beta x | \epsilon_{t-1} = -\alpha - \beta x) &= \alpha + \beta x + \frac{\theta}{1 - \alpha - \beta x}, \\
\Pr(\epsilon_t = -\alpha - \beta x | \epsilon_{t-1} = -\alpha - \beta x) &= 1 - \alpha - \beta x + \frac{\theta}{1 - \alpha - \beta x}.
\end{aligned} \tag{8}$$

Define $n_{j1}(x,t) = \{n_j(x,t) | n_j(x,t) = 1 - \alpha - \beta x\}$, the number of individuals for which the residuals have a value of $1 - \alpha - \beta x$ at time t, and $n_{j2}(x,t) = \{n_j(x,t) | n_j(x,t) = -\alpha - \beta x\}$, the number of observations for which the residuals have a value of $-\alpha - \beta x$ at time t. Let π_{ijk} be the number of observations from $n_{ji}(x, t-1)$ which are transferred from electoral

[4]The joint pdf display in Table I is a member of the class of distributions called the bivariate binomial distribution, about which more will be discussed later (see Stuart and Ord (1987), section 5.50).

[5]The θ term can be made random and the following will also go through but since as will be seen in the a later section that the linear probability model is not the best way to model voting behavior this extension is not made. The probabilities can be allowed to vary between time $t-1$ and t but that just creates complications without adding to the examination of "bias" due to grouping. Incidentally, this makes it clear that if there is contemporaneous correlation, then the linear probability model is biased at the individual level, as then $\alpha + \beta x$ is no longer the probability of the event occurring given x. The results of the next section are necessary to handle this type of dependency.

unit k to electoral unit j between time $t-1$ and time t (it is assumed that realizations at time $t-1$ are observed). Define $\delta_{ji} = \sum_{k=1}^{N} \pi_{ijk} - \sum_{k=1}^{N} \pi_{ikj}$, so that δ_{ji} is the overall transfer of individuals whose residuals have a value of $1-\alpha-\beta x$ (i = 1) or $-\alpha-\beta x$ (i = 0). Then the following equality always holds:

$$n_j(x,t) = n_{j1}(x,t-1) + \delta_{j1} + n_{j2}(x,t-1) + \delta_{j2}. \tag{9}$$

The quantity of interest is now the expected value of $n_{j1}(x,t) + n_{j2}(x,t)$, as this is also the expectation of $\sum_{i=1}^{n_j} \epsilon_{ji}$. Using the conditional probabilities calculated above in (9) one has[6]

$$\begin{aligned}
\mathrm{E}[n_{j1}(x,t)] &= \Pr[\epsilon_t = 1-\alpha-\beta x | \epsilon_{t-1} = 1-\alpha-\beta x](n_{j1}(x,t-1)+\delta_{j1}) \\
&\quad + \Pr[\epsilon_t = 1-\alpha-\beta x | \epsilon_{t-1} = -\alpha-\beta x](n_{j2}(x,t-1)+\delta_{j2}) \\
&= (\alpha+\beta x)[n_j(x,t-1)+\delta_{j1}] \\
&\quad + \frac{\theta}{(\alpha+\beta x)(1-\alpha-\beta x)} \Big\{ (1-\alpha-\beta x)[n_{j1}(x,t-1)+\delta_{j1}] \\
&\quad - (\alpha+\beta x)[n_{j2}(x,t-1)+\delta_{j2}] \Big\}
\end{aligned} \tag{10}$$

and

$$\begin{aligned}
\mathrm{E}[n_{j2}(x,t)] &= \Pr[\epsilon_t = -\alpha-\beta x | \epsilon_{t-1} = 1-\alpha-\beta x](n_{j1}(x,t-1)+\delta_{j1}) \\
&\quad + \Pr[\epsilon_t = -\alpha-\beta x | \epsilon_{t-1} = -\alpha-\beta x](n_{j2}(x,t-1)+\delta_{j2}) \\
&= (1-\alpha-\beta x)[n_j(x,t-1)+\delta_{j2}] \\
&\quad + \frac{\theta}{(\alpha+\beta x)(1-\alpha-\beta x)} \Big\{ (\alpha+\beta x)[n_{j2}(x,t-1)+\delta_{j2}] \\
&\quad - (1-\alpha-\beta x)[n_{j1}(x,t-1)+\delta_{j1}] \Big\}.
\end{aligned} \tag{11}$$

Then the sum of the expected numbers for each residual times that residual gives the expected value,

$$\begin{aligned}
E\left[\sum_{i=1}^{n_j} \epsilon_{ji}\right] &= \sum_{x=0}^{1} \{(1-\alpha-\beta x)\mathrm{E}[n_{j1}(x,t)] - (\alpha-\beta x)\mathrm{E}[n_{j2}(x,t)]\} \\
&= \sum_{x=0}^{1} \left\{ \frac{\theta}{(\alpha+\beta x)(1-\alpha-\beta x)} [(1-\alpha-\beta x)[n_{j1}(x,t-1)+\delta_{j1}] \right. \\
&\quad \left. - (\alpha+\beta x)\,[n_{j2}(x,t-1)+\delta_{j2}]] \right\}.
\end{aligned}$$

To obtain the overall expression, note $\sum_{i=1}^{n_j} x_{ji} = n_j(1,t)$, and so (6) becomes

$$E\left[\frac{X'R'RE}{X'R'RX}\right] = \sum_{j=1}^{N} \frac{n_j(1,t)}{\sum_{k=1}^{N} n_k(1,t)^2}$$

[6] All of these expectations are conditional on x, $n_j(x,t)$, and $n_j(x,t-1)$.

$$\sum_{x=0}^{1}\left\{\frac{\theta}{(\alpha+\beta x)(1-\alpha-\beta x)}\left[(1-\alpha-\beta x)[n_{j1}(x,t-1)+\delta_{j1}]\right.\right.$$
$$\left.\left.-(\alpha+\beta x)[n_{j2}(x,t-1)+\delta_{j2}]\right\}\right. \tag{12}$$

This expression then completely typifies the nature of bias assuming that the linear probability model is true and unbiased estimated on the individual level at both time $t-1$ and time t. King (1995) gives five different formulas for bias and shows they are equivalent (these formulas include both the Ansolabehere and Rivers and the Palmquist formulations). Therefore, under the assumption of unbiasedness of the linear probability model, they are all equivalent to the above formula.[7]

If $n_j = 1$ for all j, then the $\delta_{ji} = 0$, and (12) reduces to

$$E\left[\frac{X'R'RE}{X'R'RX}\right] = \sum_{x=0}^{1}\left\{\frac{\theta}{n(\alpha+\beta x)(1-\alpha-\beta x)}\right.$$
$$\left.\left[(1-\alpha-\beta x)\sum_{j=1}^{n} n_{j1}(x,t-1)-(\alpha+\beta x)\sum_{j=1}^{n} n_{j2}(x,t-1)\right]\right\},$$

for which the plim is equal to zero, since $\text{plim}\sum_{j=1}^{n} n_{j1}(x,t-1)/n = \alpha+\beta x$ and $\text{plim}\sum_{j=1}^{n} n_{j2}(x,t-1)/n = (1-\alpha-\beta x)$ by the assumption of individual consistency at time $t-1$. So without aggregation there is no bias. Even with aggregation if the serial correlation between the individuals at times $t-1$ and t is zero ($\theta = 0$), then there is no bias no matter how one might have arranged the individuals into electoral units. Also, if the $n_{ji}(x,t-1)$ are themselves close to the expected values, which one would expect under random grouping, then

$$E[n_{j1}(x,t-1)] = n_j(x,t-1)\Pr(\epsilon_t = 1-\alpha-\beta x) = n_j(x,t-1)(\alpha+\beta x) \tag{13}$$

and

$$E[n_{j2}(x,t-1)] = n_j(x,t-1)\Pr(\epsilon_t = -\alpha-\beta x) = n_j(x,t-1)(1-\alpha-\beta x) \tag{14}$$

then under these two equations holding (12) becomes

$$E\left[\frac{X'R'RE}{X'R'RX}\right] = \sum_{j=1}^{N}\frac{n_j}{\sum_{k=1}^{N} n_k^2}$$
$$\sum_{x=0}^{1}\left\{\frac{\theta}{(\alpha+\beta x)(1-\alpha-\beta x)}[(1-\alpha-\beta x)\delta_{j1}-(\alpha+\beta x)\delta_{j2}]\right\},$$

which shows clearly that, starting with an unbiased arrangement of individuals into groups, bias is created by the movement of individuals between electoral precincts by a number disproportionate to their expected number of residuals in the population.

The process described here requires an observation on the realizations at one time and then selection on the basis of of that observation for rearrangement of voters between electoral units before the next observation.

[7] If the condition of individual unbiasedness is dropped, then all the formulas for "bias" are essentially useless as B can take on any value.

How likely are these manipulations, that is, how probable is it that someone is observing the errors and then performing rearrangements based on those? One consideration is that the lowest level of electoral units in this country is usually the precinct. The selection process outlined above implies that whatever entity is moving individuals around has knowledge of the behavior of the individual. This will rarely be the case. Furthermore, the assignment of voters to election precincts is done primarily by a county registrar of voters who has no reason to assign individuals from one precinct to another on the basis of their observed behavior (which the registrar won't know in any case because the ballots are unmarked). And since the registrar cannot move individuals from one district to another (they would merely be assigned from one precinct to another within the district, and so those individuals would still vote for the same slate of candidates), there would be no particular reason to do so. Therefor it would seem *a priori* that aggregation at the level of the precinct would seem to meet every criteria for random grouping and estimations done at the aggregate level will be consistent *if* (3) holds. What happens when (3) does not hold will also be examined from a probabilistic point of view.[8]

Finally, a question which might be addressed is why researchers have concentrated so much attention on the linear probability model, a model which in general is not used anymore in individual choice problems?[9] The answer has to be that it seems at first glance to be the most natural way of presenting the problem, that is, if $w = xB + \epsilon$, it is straightforward to add up this equation (which is what the grouping matrix does) for members of an electoral unit to get an aggregate equation to estimate. By apparent analogy with the usual least-squares estimation, no distribution on the error term need be assumed, aside from the usual zero mean and conditional unbiasedness, to ensure consistency of the estimator. This makes it a more general method than the method of maximum likelihood, which requires the specification of a probability law for consistent estimation. As has been seen, though, if one does assume unbiasedness of the linear probability model when estimated at the individual level, then the residuals are completely determined, and so the advantage of least squares (not having to specify a probability law on the residuals) is lost. It is to what these residuals imply for the theory of aggregate analysis which is discussed in the next section.

[8]Since the lowest level of electoral behavior which is usually observed is the precinct (surveys almost never have the number of observations necessary to begin to make this type of non-random assignment), this is the unit that one might expect to be manipulated. There might be, for example, a definite partisan advantage for the movement of precincts between districts on the basis of observed realizations, and the agency that might be doing so, say a state legislature during a redistricting, could very well have a such a motivation. The results of the next section will be needed to examine the effects of this type of behavior.

[9]King (1986) states that the use of this model "can yield predicted probabilities greater than one or less than zero, heteroskedacticity inefficient estimates, biased standard errors and useless test statistics." (Maddala (1983) gives a discussion of some of the problems associated with this model-see page 16). The reason it actually works in this case is that estimation of the individual level equation (3) by least squares gives the frequency estimate of those with an x value of one choosing the candidate. Examination of (6) shows that this equation, without "bias", also gives the frequency estimate.

3 Relation to Probability

This section examines the question if there is a probability model induced by the linear probability model, what does it imply about the aggregation process? The short answer is everything, with some qualifications to be discussed in this section. The well-known result (Rao (1973)) is that if one has a sum of independent random variables Y_{ji}, then the sum of these random variables is the convolution of these random variables (and its characteristic function is the product of the characteristic functions of the Y_{ji}), that is,

$$Y_j = Y_{j1} + \ldots + Y_{jn_j} \tag{15}$$

This equation deserves to be called the fundamental equation of aggregate analysis, because it provides a theoretical basis for understanding any theory of aggregate inference which is consistent with a probabilistic choice model defined on the individual.[10] This is the equation used explicitly in Lupia and McCue (1990) (page 384), Brown and Payne (1987) (equation 3.1), Hawkes (1969) (page 70), and MaCrae (1977) (equation 3.1), and there are a number of the implications which are now discussed.

The most obvious and useful inference to draw from (15) is that if one knows the probability law on the individuals, one know the probability law describing the distribution of the sum Y_j. *Ipso facto*, then, there exists a consistent estimator for the distribution for the sum, namely, the estimator which comes from the method of maximum likelihood. Theoretically, then *if one knows the probability laws on the individual choices in an electoral unit, one knows the probability law on the sum of those choices in the electoral unit.* Therefore the entire theory of aggregate inference is resolved, and furthermore resolved in such a manner that proposed estimators (at least any which are based on a probability law for the individual) can be compared.[11] Knowing the theory, of course, and implementation in practice are two separate matters. For one thing, there are limitations, on the method of maximum likelihood, just as there are limitation on other methods of estimation. Aside from the usual regularity conditions (which are assumed to be met), the possible causes of problems in aggregate analysis all have individual-level counterparts. The various forms of misspecification are given in Table 2, and will be discussed in the order given.

The first problem listed in Table 2 is that there are too many parameters to estimate the model. At the individual level, suppose the Y_{ji} are parameterized by p_{ji}. Even if one had access to individual data, any estimation by the method of maximum likelihood would

[10] This equation holds if the Y_{ji} are not independent, if $\Pr(Y_j < w)$ is defined as the integral of the distribution function of the joint distribution of the Y_{ji} over $Y_{j1} + \ldots + Y_{jn_j} < w$ (see Stuart and Ord, Section 7.26). If this is the case, though, the distribution of the sum is no longer defined as the product of the characteristic functions. One procedure to follow is to replace any of the dependent Y_{ji} with the random variable representing their actual distribution. Thus is Y_{j1} and Y_{j2} are assumed to be dependent with one another and with no other individuals, then $Y_{j1,2}$, representing the joint distribution, would be used in the place of $Y_{j1} + Y_{j2}$. The question of dependency is discussed in more detail below.

[11] One can abandon the requirement for a probability law on the individual. Linear programming models basically do that (see Claggett and Van Wingen (1993)). It should be noted, though, that if there is a probability law on the individual, even if it is not modelled, then the estimator from aggregate data is still a function of the individual probability laws.

Table 2: Problems in Aggregate and Individual Analysis

Aggregate Problem	Individual Problem
Too many parameters to estimate model	Same, but less restrictive
Modelling different probability laws as same	Same
Modelling dependent choices as independent	Same
Modelling wrong functional form	Same
Aggregation "bias"	Choice-based sampling

break down, because there is one parameter for every individual. The usual restriction which is placed on this model, then, is to parameterize the p_{ji} by some parametric function q, so that $p_{ji} = q(z_{ji}, \tau)$, where z and τ are conformable vectors. Then if the number of τ don't grow with the sample size, under broad regularity conditions, one can obtain a consistent estimator of τ and hence of p_{ji}. At this point (15) can be estimated, at least in theory. The difficulty for aggregate estimation comes from the fact that convolutions of, say, binary random variables, can produce a large number of terms in the actual convolution (on the order of 10^{300} with 500 binary random variables with different values p_{ji}). This is beyond any current ability to calculate, even with modern computers.[12] Therefore, another restriction must be applied.

One restriction is to assume that a number of the Y_{ji} have common distributions and that the distributions convolve into known distributions. It is now shown that the problem described at the beginning of the paper, that of the linear probability model, follows exactly this restriction, since individual consistency implies that one knows the distribution of the residuals. To show this restriction explicitly, note that first there are two groups in the electorate (corresponding to $x = 0$ and $x = 1$). Each member of either group can make either one of two choices, a vote for the candidate or a vote not for the candidate.[13] The probability of choosing the candidate is the same as $\Pr(1 - xB = \epsilon) = \alpha + \beta x$ (and the probability of not choosing is $1 - \alpha - \beta x$), so there are two types of bernoulli distributions in any electoral precinct (since $x = 0$ or $x = 1$). If there are $n_j(i)$ individuals of type i ($i = 0, 1$) in electoral precinct j, the bernoulli's with probability $\alpha + \beta$ (those corresponding to $x = 1$) of choosing the candidate convolve to a binomial with parameters $(n_j(1), \alpha + \beta)$, while the bernoullis with probability α (those corresponding to $x = 0$) convolve to a binomial with parameters $(n_j(0), \alpha)$. Therefore the probability of observing Y_j individuals choosing the candidate is simply the convolution of these two distributions.

[12] A possibility exists that not all of the permutations need be calculated in order to provide an estimate in these circumstances–that, in fact, a random sample of the permutation might produce a consistent estimate. This line of inquiry is beyond the scope of this paper!

[13] The case of abstentions or indeed not turning out for the election will be discussed later.

Let $p_1 = \alpha + \beta$ and $p_0 = \alpha$. Then one can write the $\Pr(Y_j)$ as

$$\Pr(Y_j) = \sum_{i=0}^{Y_j} \left(\begin{array}{c} n_j(1) \\ i \end{array} \right) p_1^i (1-p_1)^{n_j(1)-i} \left(\begin{array}{c} n_j(0) \\ Y_j - i \end{array} \right) p_0^{Y_j - i} (1-p_0)^{n_j(0)-(Y_j-i)} \qquad (16)$$

where it is understood that $\left(\begin{array}{c} a \\ b \end{array} \right)$ is zero if b is greater than a or b is less than 0. Then estimation of this by maximum likelihood will produce consistent estimators for the α and β.[14] Rather than estimate this likelihood in the above form a normal approximation can be made (see Feller (1950)). This gives

$$Y_j = n_j(1)p_1 + n_j(0)p_0 + u_1 + u_2, \qquad (17)$$

where u_1 is distributed normally with mean zero and variance $n_j(1)p_1(1-p_1)$ and u_2 is distributed normally with mean zero and variance $n_j(0)p_0(1-p_0)$. It is clear that (17) is (1) or (5) but now the distribution of the error terms are known.[15] This is the form of the model given in Lupia and McCue and it is clear that estimating the aggregated linear probability model is the same as estimating the Lupia and McCue model by the method of least squares, assuming no aggregation bias of the type discussed in the last section. Maximum likelihood, though, is of course the preferred method of estimating this equation, because, as Lupia and McCue point out, the equation can be estimated both with and without the assumption that the variances are of the form given above. This can be then be used in a simple chi-squared test.[16]

The next problem in Table 2 is that of modelling different probability laws as the same probability law. This is the equivalent of convolving together in (15) two or more different probability laws into one. Social science descriptors, unlike those attached in the physical sciences, tend to be highly variable. That is, a piece of quartz is still a piece of quartz and implies certain properties about itself, but a "black" or "white" does not necessarily imply anything about the individual. In particular, one can conceive of there being two types of white, poor whites and and non-poor whites, with different probabilities of voting for the candidate. The simplest case of this is when there are three types of individuals in the electorate, all three with different probability laws, and the random variables associated with two of these groups are convolved together under the mistaken belief that they share the same probability law.

This case is treated in detail in Lupia and McCue with numerical and graphical examples and there is no need to repeat it here, but suffice it to say that if this is the case, then the estimate of B in equation (3) is a function of the parameters of the three probability laws and of the the distribution of the types of individuals throughout the electorate, and B can

[14] It is an interesting exercise in probability calculus to show directly that the estimators obtained by the method of maximum likelihood from this equation are consistent.

[15] It should be noted for historicity that Hawkes (1969) was apparently the first to recognize that the nature of the aggregation in ecological regression implied a distribution on the error terms. Hawkes' model will be discussed more fully in the next section.

[16] As is well-known, minus twice the log of the ratio of the restricted to unrestricted likelihood is distributed asymptotically chi-squared.

Table 3: Joint decision probabilities of a dyad

p_{11}	p_{12}	$p_{1.}$
p_{21}	p_{22}	$p_{2.}$
$p_{.1}$	$p_{.2}$	

be a long way from what one believes B is. It is important to realize, though, that if the likelihood ratio test arising out of (17) is not rejected, then one can have some confidence that this is not a problem. Furthermore, the same problem arises if the analysis is done with individual data at the level of the individual, if one estimates two different types of individuals as one type of individual.

The next problem in Table 2 is modelling dependent choices as independent. It might be natural to assume, for example, that voting decision of each member of a dyad[17] are correlated with the voting decision of the other member. If this is the case, then the way to make (15) correct is to model that couple's probability law jointly, having possible outcomes of 0, 1 or 2 votes for the candidate. Then (15) still holds. What if that is not done in individual analysis, that is, what if the probability of choosing a candidate is described by Table 3?[18] If the modeler uses $p_{.1}$ and $p_{.2}$ to model the probability rather than p_{11} and p_{12}, then there is misspecification, and furthermore, it will almost always bias the estimator. Linear regression allows one to not specify the distribution of the error term, but choice-based estimation, which almost always uses some form of the multinomial,[19] is not as forgiving. So it can be seen if there is this endogenous dependence of the voting decision, then individual analysis is also misspecified. Thus while this is a certainly a possible source of error, it is present in both individual and aggregate modelling.

The next error in Table 2 is modelling the wrong functional form. This is obviously a problem in both aggregate and individual level analysis. These problems are compounded when the probabilities are parameterized by other functional forms with their own parameters and covariates. From the standpoint of aggregate modelling, though, an interesting question is how to model processes where the individual choice model does not lend itself easily to the calculation of the convolution. Several different approaches suggest themselves in that case. First, it will usually be the case that some type of central limit theorem applies if there are sufficiently few groups with common parameters, so a normal approximation may

[17]Two individuals residing at the same address with an interpersonal relationship.

[18]Table 3 can be motivated as follows. Assume that there is an underlying continuous variable $y^* = r\rho + v$, and if y^* is above a certain point (call it ψ), the candidate is selected, otherwise, the candidate is not. Now, for the dyad, if the probability of the event $\{v_1 > \psi, v_2 > \psi\}$ is not the product of the probabilities of the events $\{v_1 > \psi\}$, $\{v_2 i > \psi\}$, then the correct probability for the residual for the observed member of the dyad (call that person 1), is $\Pr(v_1|v_2)$. If one looks only at the individual voting decision, and assumes $\Pr(v_1)$ is the correct error term when $\Pr(v_1|v_2)$ is actually the distribution of the error, then there is misspecification if $\Pr(v_1|v_2) \neq \Pr(v_1)$. And in general these will not be equal, if, say, the v_i are jointly distributed normally with non-zero correlation coefficient, a common assumption on joint residuals.

[19]Or even more complicated forms, such as elimination-by-aspects or nested multinomial logit.

be used. Second, method of moments estimates can be used even if distributions cannot be estimated directly since "cumulants cumulate", that is, moments of the sum can be expressed as products of the individual moments (for an application of this approach, see Blischke (1964)). Finally, using (15), it should be possible to calculate the distribution of the sum, if not analytically, then at least numerically. It should be noted that different hypotheses about the nature of the individual decision should lend themselves to different hypothesis observable at the aggregate level, just as at the individual level.

It is now shown that the "bias" calculations discussed in the last section are an additional complication to this estimation problem. Let

$$\frac{(n_{ji}(x,t-1)+\delta_{ji})}{n_j(x,t-1)} = \gamma_{ji},\ i=1,2, \tag{18}$$

Considering the model of serial correlation given in Table 1, one can write down the probability of the individual in electoral unit j as[20]

$$\begin{aligned}\Pr(\epsilon_t = 1-\alpha-\beta x | t \in j) &= \Pr[\epsilon_t = -\alpha-\beta x | \epsilon_{t-1} = 1-\alpha-\beta x]\gamma_{j1} \\ &\quad +\Pr[\epsilon_t = -\alpha-\beta x | \epsilon_{t-1} = -\alpha-\beta x]\gamma_{j2}\end{aligned} \tag{19}$$

then after some algebra,

$$\Pr(\epsilon_t = 1-\alpha-\beta x | t \in j) = \alpha+\beta x+\theta\frac{\gamma_{j1}-\alpha-\beta x}{(\alpha+\beta x)(1-\alpha-\beta x)}. \tag{20}$$

A similar process then leads to

$$\Pr(\epsilon_t = -\alpha-\beta x | t \in j) = 1-\alpha-\beta x+\theta\frac{\alpha+\beta x-\gamma_{j2}}{(\alpha+\beta x)(1-\alpha-\beta x)}. \tag{21}$$

These probabilities can then be place back into (16) or (17) to construct a likelihood.[21]

This likelihood is inestimable in its full generality, since while θ is constant across all parameters, the γ_{ji} as specified here can vary with the electoral units, creating 2 times N parameters to estimate, so the method of maximum likelihood breaks down.[22] The problem, which can be called "bias", comes about because of serial correlation of the residuals and non-random assignment of a previous realization of the individuals with those residuals. An

[20]Implicitly it is being assumed that the conditional probability of the residuals is not changed by a change in the modifying distribution. This seems reasonable here by the the method in which the dependency is derived. For a discussion of this postulate in general choice modules, see Manski and McFadden (1981, page xix).

[21]This is the equivalent of parameterizing the probabilities p_i with parameters θ and γ_{ji}. Parameterization has implications for aggregate estimation which are discussed in Appendix I.

[22]This problem is known in the ecological literature as the "constancy" problem, where, if one considers the two parameters in (1) to vary for every electoral unit, then one always has a perfect fit (and the error term is zero), but it is impossible to estimate the equation. The constancy assumption is then that the two parameters in (1) are constant across all electoral units.

estimable likelihood function can be made with the parameters from equations (20) and (21), if there are additional assumptions about the nature of the "bias". For example, it can be assumed that the γ_{ji} are parameterized by some exogenous variables observed at the level of the electoral unit. Or, it might be assumed that the γ_{ji} are random with some distribution, thus creating a random effects model. By any standard methodology of statistics, however, one should fit equation (17) first as the simplest model, before going on to the model derived from equations (20) and (21). Only if (17) is rejected should the researcher go on to other models.[23]

The analogous problem to "bias" at the individual level is choice-based sampling, which also alters the probabilities of the observations.[24] Choice-based sampling is simply choosing on the basis of observed values of the endogenous variable or on values which are correlated with the observed values (such as previous realizations). Just as with the procedure here, the likelihood of the observations occurring has been altered, and there are ways of adjusting for the changed probability law (see Cosslett (1981), for a demonstration of some of these methods).

It has been argued that when the electoral unit is the precinct there are no *a priori* grounds to expect any bias, that is an outside entity manipulating voters between electoral units for some advantage. Districts, though, are made up of precincts, and, using (15), the distribution of the votes in a district can be represented as

$$D_d = \sum_{j \in d} Y_j = \sum_{j \in d} \sum_{l=1}^{n_j} Y_{jl}. \tag{22}$$

If there is selection on the precincts, this then requires, as was done in the last section with the linear probability model, calculating the probabilities of $\sum_{j \in d} Y_j$, which will now be expressed as conditional on previous realizations. This formula also allows direct estimation of district voting behavior by providing the distribution of the district vote.[25]

Finally, does aggregation in and of itself produce bias, as has been claimed by a number

[23]It should be obvious that it will not, in general, be possible to differentiate "bias" and any of the other types of misspecification describe above, just as in estimating an individual model it is very difficult to determine different types of misspecification.

[24]A distinction needs to be made here between a sample representative of the population and a sample for which one is modelling a regime through what is known in the statistical literature as the ancillarity principle (see Stuart and Ord, section 31.12). Take the famous Literary Digest Poll of 1936, which predicted that Landon would defeat Roosevelt. Since this poll was conducted by telephone, and the country was in the middle of an economic depression, those who had telephones were disproportionately well-off and thus disproportionately inclined to vote Republican. The sample was thus not representative of the population. If, however, one were fitting a model of voting behavior to this sample, one would have $f(x, y) = f(x)f(y|x)$, and one would be estimating y (vote choice) conditional on the x (what are called exogenous characteristics, such as age, income, partisan affiliation, etc), and by the usual standard assumptions, estimates of parameters associated with y would be consistent. What this says is that selection on exogenous characteristics does not invalidate the probability law $f(x, y)$.

[25]Typically, regressions on districts have not been consistent with the assumptions of individual analysis, as they cannot be expressed in terms of (15). Models of presidential selection based on macroeconomic indicators, approval and so forth are also not coherent. Even a model which follows (15) will, as an empirical matter, not work on districts, a problem which is discussed in Appendix I.

of authors? The answer is a qualified no. Equation (15) is *always* true, so any "bias" must be described by this equation. Certainly in the case of independence of the random variables being summed, there is no bias. When the random variables are dependent, this produces the same problems for individual choice-based estimation as it does for an aggregation of those choices, so aggregation *per se* cannot be said to be creating the problem. The case where there needs to be a qualification to the no is that, as in choice-based sampling at the individual level, one can alter the probabilities of the observations occurring away from the distribution one would usually model them. In this case, though, (15) still holds, but the probabilities must be modelled correctly.

This can be shown in complete generality as follows. Suppose one has a regime $y = x\beta+u$, and suppose that there is a random variate z such that u and z are not independent. Then let $\epsilon = u|z$, so that the $\Pr(\epsilon) = \Pr(u|z)$. Then

$$\Pr(u, z) = \Pr(u|z)\Pr(z) = \Pr(\epsilon)\Pr(z) = \Pr(\epsilon, z) \tag{23}$$

Now, suppose one wants to aggregate and this aggregation is a function of the z. The above shows that ϵ and z are independent, so ϵ is independent of any function of the z.[26] In particular, if the aggregation operator R is a function of the z, $R(z)$ is independent of the error term created by conditioning u on z. Thus while aggregation does offer this opportunity for "bias", it can still be handled with a correct modelling of the probabilities, just as can choice-based sampling.

4 Methods of Estimation

There are numerous applications for what has been done in the previous sections of this paper. The first discussed is the distribution of the two-stage procedure under the assumption of the consistency of the linear probability model at the individual level, which has not appeared in the literature before. This two-stage procedure, which estimates both turnout and vote choice, was evidently first created by Kousser (1973). One difficulty with the linear probability model is that it is restricted to modelling events which either occur or don't occur. One way of overcoming this is to estimate two separate equations,

$$Y = X_1\theta_1 + X_2\theta_2 + u \tag{24}$$

and

$$T = X_1\gamma_1 + X_2\gamma_2 + v, \tag{25}$$

where T are the voters who turn out in an electoral unit (as opposed to registered), Y are the votes cast for a candidate in an electoral unit, X_i are the number in group i, and the θ_i and γ_i are the parameters to be estimated. The interest is in the distribution of the functions

$$\hat{\lambda}_i = \frac{\hat{\theta}_i}{\hat{\gamma}_i}, \tag{26}$$

[26] In particular, if u and z are jointly distributed bivariate normal, which would be the usual assumption, the distribution of $u|z$ is independent of u. This is shown by direct calculation for the bivariate normal in Rao, (3d.1.3).

Table 3: Joint PDF of $f(\epsilon_y, \epsilon_t)$

	$\alpha + \beta x$	$1 - \alpha - \beta x$
$\eta + \phi x$	$\eta + \phi x$	0
$1 - \eta - \phi x$	$-\eta - \phi x + \alpha + \beta x$	$1 - \alpha - \beta x$

which are taken to be the proportion of the group who voted for the candidate, given that they actually turned out. The asymptotic distribution of this estimator is now derived under the assumption of the consistency of the linear probability model when estimated at the individual level.

The problem above is stated in its aggregate form and needs to be related to the linear probability model when estimated at the individual level. Let

$$y = \alpha + x\beta + \epsilon_y \tag{27}$$

and

$$t = \eta + x\phi + \epsilon_t \tag{28}$$

be the individual level linear probability models corresponding to the above aggregate models (here, as in equation (16), $\eta + \phi = \gamma_1$, $\eta = \gamma_2$, $\alpha + \beta = \theta_1$, and $\alpha = \theta_1$). The assumption of consistency on both the turnout and vote choice equations implies that the joint probability of voting in the election and voting for a candidate can be written as in Table 3.[27] Thus under individual consistency, the choices for turnout and vote choice follow a bivariate binomial distribution with the probabilities given in Table 3. Since the probability of an individual voting for a candidate is zero if that individual does not vote, there are only three non-zero entries in this table, and these are completely determined by the marginal entries, and the distribution, as Wishart (1949) points out, becomes a univariate trinomial distribution (that is, the usual trinomial). Since there are only two different values of x, using (15), the individuals in each electoral unit convolve to one of two trinomials. The likelihood for this estimation by a normal approximation to the multinomial is given by Hawkes.[28]

An additional wrinkle occurs when $(\gamma_1, \gamma_2, \lambda_1, \lambda_2)$ are functions of other variables, say $\gamma_k = f(r, \psi_k)$ and $\theta_k = g(r, \rho_k)$, where the r's are variables defined at the level of the aggregate unit (here the optimization is done with respect to ψ_k and ρ_k, which may be vectors). This is known as parameterization and it seems to have been introduced by Miller (1972).[29] This likelihood is also in the literature (see MaCrae or Brown and Payne (in the case where the aggregate compound multinomial is simply an "aggregate" multinomial)).[30]

[27]From (27) and (28) and Table 3 one can interpret (26) as an estimator of $\Pr(y = 1|t = 1)$.

[28]Hawkes model is for the transition model, where the groups are formed by other voting choices, rather than groups formed by exogenous characteristics of the individuals (such as race). The likelihood is still the same, however.

[29]With a linear form. MaCrae used a log-linear form. For a discussion what parameterization accomplishes in aggregate estimation, see Appendix I.

[30]These models are also for the transition model but once again the likelihood is the same. "Aggregate" in the terminology of Brown and Payne means the convolution.

The formula for the proportion voting for a candidate out of those turning out becomes

$$\hat{\lambda}_i = \frac{\sum_{j=1}^{N} \theta(r_j, \hat{\phi}_i) X_j}{\sum_{j=1}^{N} \gamma(r_j, \hat{\psi}_i) X_j}. \tag{29}$$

In either definition of the $\hat{\lambda}_i$ the asymptotic distribution is easily found through standard maximum likelihood theory. Letting $\Lambda = (\lambda_1, \lambda_2)$, defining $\omega_u = (\gamma_1, \gamma_2, \lambda_1, \lambda_2)$ and $\omega_p = (\psi_1, \psi_2, \rho_1, \rho_2)$, then by a corollary of Slutsky's theorem (see Rao, page (385-388)), the asymptotic distribution of $\sqrt{N}\hat{\Lambda}$ in the unparameterized case is simply normal with mean Λ and covariance matrix

$$\mathrm{Var}(\sqrt{N}\hat{\Lambda}) = \left[\frac{\partial \Lambda}{\partial \omega_u}\right] \mathrm{Var}(\sqrt{N}\hat{\theta}, \sqrt{N}\hat{\gamma}) \left[\frac{\partial \Lambda}{\partial \omega_u}\right]^t, \tag{30}$$

and in the parameterized case it is

$$\mathrm{Var}(\sqrt{N}\hat{\Lambda}) = \left[\frac{\partial \Lambda}{\partial \omega_p}\right] \mathrm{Var}(\sqrt{N}\hat{\psi}, \sqrt{N}\hat{\rho}) \left[\frac{\partial \Lambda}{\partial \omega_p}\right]^t. \tag{31}$$

All of the usual inferences on the elements of Λ can be made from the appropriate matrix.

Since maximum likelihood is asymptotically efficient when the correct probability model is specified and the assumption of individual consistency of the linear probability model determines the probability distribution, there is no method of estimation which is more efficient than what has been described above. There are, however, several methods which have been been used or proposed in the literature for which the distribution of Λ can be derived. The first is least squares, and the estimator is then

$$\hat{\lambda}_i = \frac{\theta_i + (X'X)^{-1}X'E_Y}{\gamma_i + (X'X)^{-1}X'E_T}, \tag{32}$$

where E_y is the vector of ϵ_y and E_T is the vector of ϵ_t. The joint distribution of $(X'X)^{-1}X'E_Y$ and $(X'X)^{-1}X'E_T$ is necessary to calculate the asymptotic distribution of this statistic. The three categories of the multinomial are no vote, vote for the candidate, and vote against the candidate, with probabilities $p_1(x) = \eta + \phi x$, $p_2(x) = 1 - \alpha - \beta x$, and $p_3(x) = -\eta - \phi x + \alpha + \beta x$, respectively (set $p(x) = [p_1(x), p_2(x), p_3(x)]^t$). Then if $Z = [Z_{1j}, Z_{2j}, Z_{3j}]^t$ are the respective numbers of voters in the three categories in electoral unit j, this can be approximated by

$$Z = n_j(1)p(1) + n_j(0)p(0) + u(1) + u(0) \tag{33}$$

where the distribution of each $u(x)$ is trivariate normal with mean zero and covariance entries $n_j(x)p_l(x)(1 - p_l(x))$ along the diagonal and $-n_j(x)p_l(x)p_k(x)$ otherwise.[31] Equations (24) and (25) are equivalent to multiplying (33) by a 2 by 3 matrix (call it A) with the (1,2),

[31] See Stuart and Ord, page 487, for a derivation of this approximation.

(2,2) and (2,3) entries having value one and other entries being zero. Then $A[u(1), u(0)]$ is multivariate normal with mean zero and variance

$$\text{Var}(\epsilon_y, \epsilon_z) = \sum_{x=0}^{1} n_j(x) \left[\begin{array}{cc} p_2(x)(1-p_2(x)) & p_1(x)p_2(x) \\ p_1(x)p_2(x) & p_1(x)(1-p_1(x)) \end{array} \right] \tag{34}$$

It is now straight-forward to calculate the asymptotic distribution of $\hat{\lambda}_i$, as the same corollary of Slutksy's theorem applies to the least squares estimators as well.[32] Define Ω_Y as the matrix with $\sum_{x=0}^{1} n_j(x)p_2(x)(1-p_2(x))$ for the jj^{th} entry, zero otherwise, Ω_T as the matrix with $\sum_{x=0}^{1} n_j(x)p_1(x)(1-p_1(x))$ for the jj^{th} entry, zero otherwise, and Ω_{YT} as the matrix with $\sum_{x=0}^{1} n_j(x)p_1(x)p_2(x)$ for the jj^{th} entry, zero otherwise. Assume that the following plims exist: plim $X'X/n = \Sigma_X$, plim $X'\Omega_Y X/n = \Sigma_Y$, plim $X'\Omega_T X/n = \Sigma_T$, and plim $X'\Omega_{YT} X/n = \Sigma_{YT}$. Then

$$\begin{aligned} \text{Var}[\sqrt{N}\hat{\theta}, \sqrt{N}\hat{\gamma}] &= \text{Var}[\sqrt{N}(X'X)^{-1}X'E_Y, \sqrt{N}(X'X)^{-1}X'E_T] \\ &= \left[\begin{array}{cc} \Sigma_X^{-1}\Sigma_Y\Sigma_X^{-1} & \Sigma_X^{-1}\Sigma_{YT}\Sigma_X^{-1} \\ \Sigma_X^{-1}\Sigma_{YT}^t\Sigma_X^{-1} & \Sigma_X^{-1}\Sigma_T\Sigma_X^{-1} \end{array} \right]. \end{aligned}$$

This covariance matrix is then used in (30) and desired inferences can be made from that equation. This distribution has historical interest,[33] as inferences have been made from the Λ derived from least squares estimation without knowledge of the distribution of the estimators.[34]

Several variants of this estimation process are of interest. The first variant is special structures on the error term, usually involving making it a function of the number of individuals or proportion (if using means) in each group, which makes the variance a quadratic function of these numbers (Goodman (1959) seems to have been the first to propose this, and there are models using it all the way up to King (1995)). Under consistency of the linear probability model, however, the variance is a linear function of the number of individuals in each group, so all of these approaches are misspecified. The second variant is using what is known as the method of bounds. This method was introduced by Duncan and Davis (1953) and forms the basis of much of the work of Shively. Recently King (1995) has proposed that the use of information from the methods of bounds in estimation can be used to improve estimation in aggregate voting problems. He makes the very broad claim that "any method of ecological inference that ignores this information is guaranteed to be inefficient," yet under the usual assumptions of the method of maximum likelihood estimators from that procedure are efficient. Can information from the method of bounds improve estimation?

[32]Note that in least squares, $\text{Var}[\sqrt{n}(X'X)^{-1}(X'u)] = \sigma^2(X'X/n)^{-1}$, so that if plim $(X'X/n) = \Sigma_X$ (say), as is usually assumed, then since $\sqrt{n}$ tends to infinity, and the variance tends to a fixed limit, Slutsky's theorem does apply.

[33]It has particularly been used in Voting Rights Cases, such as Gringles and Garza.

[34]It is also possible to obtain an exact distribution for this estimator, as it is the ratio of two normal distributions (see Hinkley (1969)). This distribution is complicated, however, and the usual number of observations available in most ecological situations means that the asymptotic variance will be adequate.

Under consistency of the linear probability model, the answer is no. The method of bounds for a single equation essentially creates a 2 by 2 contingency table with the marginal entries being formed by $n_j(x), x = 1, 0$, and Y_j and $n_j(0) + n_j(1) - Y_j$. Then the cells of the table are those who voted for the candidate versus those who did not by those who are in the group defined by $x = 0$ versus those in the group defined by $x = 1$. Let n_{11j} be the number in the cell defined by those who voted for the candidate and those in the group defined by $x = 1$ (obviously, given the marginals, knowing n_{11j} allows the determination of the other cell entries). Then n_{11j} must be less than the minimum of Y_j and $n_j(1)$. This information is used by King to truncate the distribution of the error terms used in his model.

To use this information, consider the joint distribution of n_{11j}, Y_j, conditional on the $n_j(x)$.[35] Then

$$\Pr(n_{11j}, Y_j | n_j(x)) = \Pr(Y_j | n_j(x)) \Pr(n_{11j} | Y_j, n_j(x)) \tag{35}$$

The likelihood that is usually estimated is derived from $\Pr(Y_j|n_j(x))$ (and given in (16)), and the parameters of interest are p_1, p_2. Additional information would then come from $\Pr(n_{11j}|Y_j, n_j(x))$. This density function is[36]

$$\begin{aligned} \Pr(n_{11j} = k_1 | Y_j, n_j(x)) &= \frac{b(k_1, n_j(1), p_1) b(Y_j - k_1, n_j(1), p_2)}{\sum_{l_1 + l_2 = Y_j} b(l_1, n_j(1), p_1) b(l_2, n_j(1), p_2)}, \\ & \\ & k_1 = 0, ..., \min(n_j(1), Y_j), \end{aligned}$$

where $b(k, n, p)$ is the k^{th} term of the binomial distribution parameterized by n and p. This density is zero whenever k_1 is greater than the minimum of Y_j and $n_j(1)$, so conditioning the error terms when this condition is not met is the same as conditioning on an event of probability zero–in other words, no new information is provided. What is closely related to this method, however, and can improve estimation for the prediction of a quantity for any particular electoral unit, is to condition on the Y_j. That is, use the distribution given above ($\Pr(n_{11j} = k_1 | Y_j, n_j(x))$) to make point estimates for n_{11j} (whence n_{12j}, n_{21j} and n_{22j}). In that sense the King technique does improve estimation.

Shivley, on the other hand, suggests that prior knowledge on the values of the n_{11j} be incorporated into the estimation problem. Certainly, a deterministic bound can be easily be incorporated into (16) by simply restricting the range of maximization for the parameters. The more usual way, though, of incorporating prior beliefs is through specification of a prior distribution representing those beliefs and then integrating the density and that prior to obtain a posterior distribution. The natural (or conjugate) prior for a binomial is the beta distribution, so let the prior for the group defined by $x = 1$ be $B_1(\theta_1, \theta_2)$, and that for group defined by $x = 0$ be $B_1(\phi_1, \phi_2)$, where B_1 is a beta distribution of the first kind (see Maddala (1978), pages 406-411, for definitions). Then (16) becomes

$$\Pr(K_j) = \sum_{i=0}^{K_j} \left(\begin{array}{c} n_j(1) \\ i \end{array} \right) \frac{B(i + \theta_1, n_j(1) + \theta_2 - i)}{B(\theta_1, \theta_2)}$$

[35]Note that n_{11j} is not observed in ecological problems (and so cannot directly be used in estimation), but that the method of bounds is in essence using a function of n_{11j}, so it makes sense to analyze this in terms of the distribution of n_{11j}.

[36]This probability is also given in Hannan and Harkness (1963), who provide a normal approximation.

$$\left(\begin{array}{c} n_j(0) \\ K_j - i \end{array} \right) \frac{B(K_j - i + \phi_1, n_j(0) + \phi_2 - (K_j - i))}{B(\phi_1, \phi_2)}$$

where B is the beta function. This can then be maximize with respect to θ_1, θ_2, ϕ_1, ϕ_2, and the desired probabilities obtained by the relationship $p_1 = \theta_1/(\theta_1 + \theta_2)$ and $p_2 = \phi_1/(\phi_1 + \phi_2)$.[37] The beta-binomial does not restrict the range of the parameter but other priors can be adopted which do (such as a uniform distribution, appropriately normalized, over $[\underline{p}, \bar{p}]$). No simple closed form prior exists in this case but numerical methods are available.

5 The Transition Model

The next application discussed is that of the transition model. The transition model is once again Goodman's but now one has that there are two races with alternatives $c_1 = 1, \ldots, C_1$ for the first race and $c_2 = 1, \ldots, C_2$ for the second race, and the individual chooses one and only one combination (c_1, c_2). Then there are C_1 equations to be estimated, which are of the form

$$Z_{c_2} = Y_1 \beta_{c_2 1} + \ldots + Y_{C_1} \beta_{c_2 C_1} + \epsilon_{c_2}, \tag{36}$$

where Y_{c_1} are the number of individuals choosing alternative c_1 in race 1 and Z_{c_2} are the number of individuals choosing alternative c_2 in race 2.[38] The interpretation of the $\beta_{c_2 c_1}$ is that it is the conditional probability of making a transition from state c_1 to c_2.[39]

The relationship between (15) and (36) can be calculated as follows. Assume the the choice behavior of each individual follows a multinomial distribution, with the probability of choosing (c_1, c_2) for the i^{th} individual being $p_{c_1 c_2 g}$, where g is the group that individual is in. By the properties of the multinomial, the probability of choosing c_1 for the i^{th} individual (irrespective of the choice of c_2) is $\sum_{c_2=1}^{C_2} p_{c_1 c_2 g}$, and the distribution these choices follows is also multinomial (call this probability law Y_i). Similarly, the probability of choosing c_2 for this same individual is $\sum_{c_1=1}^{C_1} p_{c_1 c_2 g}$ and this probability law is also multinomial (call it Z_i). Under the multinomial, then, the choice behavior in each individual race is derived from the choice behavior in the two races together, and furthermore, each individual race can be estimated consistently.[40].

Now, (36) can be rewritten as

$$Z = Y\beta + \epsilon, \tag{37}$$

[37] In practice the likelihood derived from this probability density function can be approximated with a location parameter and a normal error term.

[38] These choices can include such things as not voting or "rolling-off" from one race to the other.

[39] Conditional since it assumes that the individual was in state c_1 to begin with. The definition is the same as for transition probabilities in Markov chains.

[40] The problem with not using the multinomial is that then the collapse to the individual races is not, in general, consistent with the probability law assumed on the joint choices. One could, though, use (15) to calculate the various marginal distributions. One may want to use a joint distribution different than the multinomial is if one is attempting to infer an asymmetrical effect in vote choice. An example would be coattail voting, where the vote choice for President is assumed to exert an influence on congressional races in particular but not *vica versa* (see Miller (1955)).

where Z is now N by C_2, Y is N by C_1, and β is C_1 by C_2. The least squares estimate[41] of β is

$$\hat{\beta} = [Y'Y]^{-1}Y'Z, \tag{38}$$

Since $Z_j = Z_{j1} + \ldots + Z_{jn_j}$ and $Y_j = Y_{j1} + \ldots + Y_{jn_j}$ from the proceeding paragraph, if there are G groups with common probabilities, Y can be approximated by $X\Theta + U_Y$ and Z can be approximated by $X\Gamma + U_Z$, where X is the N by G matrix of the number of individuals in each group g, Θ is a G by C_1 matrix of weights, Γ is a G by C_2 matrix of weights, U_Y is a N by C_1 matrix and U_Z is a N by C_2 matrix of disturbance terms. So (38) becomes

$$\hat{\beta} = [(X\Theta + U_Y)'(X\Theta + U_Y)]^{-1}(X\Theta + U_Y)(X\Gamma + U_Z), \tag{39}$$

or, taking plims

$$\text{plim } \hat{\beta} = [\Sigma_{YY} + \Theta' S_{XX}\Theta]^{-1}[\Theta' S_{XX}\Gamma + \Sigma_{YZ}] \tag{40}$$

where plim $U_Y'U_Y/n = \Sigma_{YY}$, plim $U_Y'U_Z/n = \Sigma_{YZ}$, and plim $X'X/n = S_{XX}$.[42] This then typifies the nature of the misspecification in the usual transition model.

It turns out, though, that actual transition coefficients can be derived under the assumption of a multinomial distribution on the choices. This derivation is done in Appendix II. To test the model, a dataset is necessary, preferably one of enough specificity that a reasonable determination can be made of whether the model has any utility. The one used here is the Los Angeles County Ballot Image tape from the 1990 General Election, which contains actual ballots (and hence transition probabilities between two races). The transitions are thus in the form of individual ballot images, that is, the exact ballot that an individual voter has cast. The ballots are identified by precinct but not by individual voter within precinct, but they can be used to construct exact transition tables for an individual precinct. The aggregate election results are available by precinct also, and the transition model can be estimated and the estimated parameters checked against the actual transition probabilities.[43]

For a sample test two ballot propositions from the 1990 General election were used, Proposition 131 and Proposition 140. Both of these propositions imposed term limits on the state legislatures and statewide constitutional officers, with 131 being proposed and funded by a Democratic candidate for Governor and 140 being proposed and funded by a conservative Republican County Supervisor from Los Angeles County. 140 was the more draconian of the two, cutting the legislative budget by 40 percent and placing a limit of three two-year terms in the State Assembly (as opposed to six in 131). It is a proposition

[41] The use of least squares for examining this model can be justified for several reasons. First, if the model is correctly specified, it provides a consistent estimator. Second, if the model is either a form of the Hawkes or Brown and Payne model, the term of the likelihood which contains the location parameter dominates the likelihood (both Hawkes and Brown and Payne demonstrate this in their respective articles), so the maximum likelihood estimates will be close to this expression, even if the probability model is misspecified.

[42] The analytical expressions for Σ_{YY} and Σ_{YZ} are given in equations (61), (63), (62), (64) in Appendix II.

[43] There is a wealth of associated census data with these precincts, which can be used as parameterization variables. This data includes such things as ethnicity, age, housing status, and housing cost–census variables from STF3A (this data is associated with the precincts by a complicated method which is interesting in and of its own right and is described in McCue (1994)).

where one would expect considerable differences in voting behavior among voters of different partisan persuasions.

For this analysis two groups are created–those with declared Democratic partisan affiliation (and smaller parties which are leftward leaning) and Republicans and those associated with rightward leaning parties. Only actual voters at the polls were used for this analysis, so there are 3 possible choices for a voter on each proposition–Yea, Nay or Abstain.[44] Various covariates were used, mostly dealing with income and ethnicity. There are 4,618 precincts in this analysis, representing the voting decisions of 1,684,785 individuals. Estimates for the transitions and marginal race choices are given in Table 4 for the left partisans, right partisans, and both groups combined, with the actual transitions and marginal race choices also displayed.

[Table 4 about here]

The most obvious thing to discuss here is that there is very good agreement between the estimated transitions and marginal race choices for the combined groups and the actual transitions and marginal race totals for all groups (and given the information on the ballot tape, these are the only ones that can be checked). The differences in the estimates from the actual totals range from .09 to 5.31. One can see dramatic differences between the voting behavior of the right partisans and the left partisans. In particularly, sixty percent of the right partisans voted no on 131 and yes on 140, whereas less than one percent of the left-leaning partisans made these two choices. Given the different effects of these propositions, and the respective authors, this difference makes substantive sense.

[Figures 1, 2, 3 and 4 about here]

It is also possible to test the formula for misspecification given by (40). In Figure I the actual transition coefficients derived from the ballot image dataset for a number of proposition pairs are plotted against the estimates derived from the least-squares estimator, equation (38). In Figure II these estimates are plotted against the predicted values from the equation which typifies the misspecification in the transition model, equation (40).[45] As can be seen from Figure II, the usual transition model estimators are closely predicted by equation (40) In Figure III, transition coefficients derived under the likelihood given in Appendix II for 756 proposition pairs are displayed.[46] Finally, in Figure IV, actual choice probabilities are plotted against estimated choice probabilities. A comparision of Figure I and Figure III shows that the proposed estimator is far superior for this dataset, while in Figure IV, 95 percent of the estimates fall within plus or minus 6.5% of the actual choice probabilities, meaning that there is sufficient accuracy to make substantive conclusions on from the estimates of the choice probabilities.

[44]Actually, even though the non-voters and the absentee voters have been removed, there are voters who cast invalid ballots–they voted for both yes and no. These voters (who are a very small percent) are coded as Abstain.

[45]Figures I and II are adpoted from McCue (1992).

[46]These pairs are are estimated without parameterization. There are 4,616 precincts in each estimation.

Table 5: Joint Estimation of Models with Individual and Aggregate Data

		Type of model	
		Same	Different
Individual	Known	I	II
Observations	Unknown	III	IV

6 Joint Individual and Aggregate Estimation

The other application of (15) that will be considered here is on the joint use of individual and aggregate data in inference problems, with a particular eye towards testing hypotheses about models of voting behavior set at the level of the individual with aggregate data. It is one of the paradoxes of political science as a discipline that the most commonly observed political behavior, that of election returns, is in general not used in constructing or testing models of voting behavior. It is hoped that the methods presented here will allow such data to be used in such a pursuit.

The conceptualization of the problem is presented in Table 5. There are four cases, formed by a cross of the model and the individual observation. The individual observation is divided into two categories, whether it is known which of the aggregate data groupings the individual observation is in and whether it can be removed from the aggregate grouping (what removed here means will be seen in a moment). The model cross is whether the probability law is the same for both the the aggregate data analysis and the individual data analysis, or whether it is different. All of these cases can be analyzed using (15).

Case I (same model for both aggregate and individual data, individual observations known) is the simplest. Assume that in electoral unit j that the first m_j observations are individual observations. Then assuming independence of all the Y_{j1}, (15) can be written

$$Y_j = \sum_{i=1}^{m_j} Y_{ji} + \sum_{i=m_j+1}^{n_j} Y_{ji}. \tag{41}$$

This is the equation to use to find the probability law for the Y_j. For the linear probability model, if K_{j1} is the number of votes observed from the individual voters, and K_{j2} is the number of votes observed from the remaining aggregate voters, then the likelihood is[47]

$$\begin{aligned} \mathrm{L}(\alpha,\beta) &= \prod_{j=1}^{N}\left\{\Pr[\sum_{i=m_j+1}^{n_j} Y_{ji} = K_{j2}] \prod_{c=0}^{1} \prod_{\{i|i\in j, y=c\}} \Pr(Y_i = c)\right\} \\ &= \prod_{j=1}^{N}\left\{\frac{1}{(X_{j1}(\alpha+\beta)+X_{j2}\alpha)^{\frac{1}{2}}}\exp\left[-\frac{(K_{j2}-X_{j1}(\alpha+\beta)-X_{j2}\alpha)^2}{2(X_{j1}(\alpha+\beta)+X_{j2}\alpha)}\right]\right. \end{aligned} \tag{42}$$

[47] An obvious test for the adequacy of this model is to estimate it with the restriction that the individual and aggregate parameters are the same and unrestricted, then forming the appropriate likelihood ratio test from these two estimations.

$$\prod_{\{i|i\in j,y=1\}} \alpha + x_{2i}\beta \prod_{\{i|i\in j,y=0\}} 1-\alpha - x_{2i}\beta \Big\}.$$

If the model is different between the individual and aggregate data but it is still known which observations of the aggregate sum are the individuals (case II) then equation (42) is still valid but the actual models for the aggregate and the individual estimations will of course be different. The implications of models being different from one another but predicting the same type of outcome (vote choice on a candidate) but both being correctly specified will be considered under case IV, discussed below.

Case III (same model for both aggregate and individual data, individual observations unknown) is sometimes found in Voting Rights Act litigation. While most of the analyses of racially polarized voting center on ecological means of estimation,[48] occasionally an exit poll or voter survey is available which may also provide evidence for or against racially polarized voting. The usual procedure in statistics if one has two estimators and wishes to compare them is to calculate the joint distribution and then test $\hat{e}_1 - \hat{e}_2 = 0$, with the variance being simply $\mathrm{Var}(\hat{e}_1) + \mathrm{Var}(\hat{e}_2) - \mathrm{Cov}(\hat{e}_1, \hat{e}_2)$. Here the $\hat{e}_1$ is the estimator from the aggregate estimation and $\hat{e}_2$ is the estimator from the individual estimation.[49] Then the variance of the aggregate estimator (if done by maximum likelihood) is minus twice the inverse of the matrix of second partials and the covariance between the two estimators can be obtained by the method of Cox (1961). Thus the hypothesis of equal parameters can be tested and failure to reject such a test can be construed as evidence in favor of the claim that racially polarized voting really is being measured accurately (rejection of such a test, on the other hand, implies that there are problems measuring racially polarized voting by either one or both methods). If the test is not rejected then the two estimators can be combined to produce a more exact bound than either estimate by itself.[50]

Cases IV (different models for both aggregate and individual data, individual observations unknown) is what usually confronts the political scientist. Most theories of individual choice behavior in political science deal with presidential vote choice, and those that are quantified (such as funnel of causality, sociotropic, retrospective, and prospective typically have the probabilities parameterized by continuous or at least many-category covariates. There are now two problems. First, as described above, (15) becomes inestimable (though still theoretically correct) if the Y_{ji} do not convolve to a few common distributions. Second, many of the data which go into these models (such as state of the economy, feelings about politicians and parties, etc) are not observed at the aggregate level.[51]

[48]For a discussion of the Voting Rights act and the use of ecological estimation under the "results" test, see Lupia and McCue.

[49]The individual estimation can, of course, be the linear probability model, but it can also be such procedures as probit logit, or simply a frequency estimator. The aggregate estimation is assumed to be estimated in a manner consistent with equation (15).

[50]The simplest way to do this is to minimize the variance of a linear combination of $\hat{e}_1$ and $\hat{e}_2$, that is, $\mathrm{Var}(\lambda\hat{e}_1 + (1-\lambda)\hat{e}_2)$. In any case, the combination of information is well know–see Rao, page 389-391.

[51]This is expressed Hanushek, Jackson, Kain (1974) as follows: "[M]any social science theories relate to behavior at the microlevel, and these theories do not always lead to simple aggregations. For example, model of mass voting behavior often require measures of individual attitudes on different issues or toward

It is still possible, however, to make use of information from both the individual and aggregate analysis together through comparison of the probabilities generated from each estimation. If there are G groups in the aggregated estimation, then there are G probabilities coming out of the estimation, $\hat{p}_1, ..., \hat{p}_G$, or, if there is the parameterization $p_g = p_g(r_i, \psi)$,[52]

$$\hat{p}_g = \frac{\sum_{i=1}^{n} p_g(r_i, \hat{\psi}) X_{gi}}{\sum_{i=1}^{n} X_{gi}} \tag{43}$$

From the individual estimation,[53] calculate

$$\hat{q}_g = \frac{\sum_{i \in g} q(s_i, \hat{\eta})}{|\{i | i \in g\}|}, \tag{44}$$

where q is the probability law assumed on the individual, s_i is the data vector associated with the individual, and $\hat{\eta}$ is the estimated vector of (conformable) weights for the data. Calculation of variances for both of these estimators are straightforward and the covariance may be calculated by the method of Cox. The G pairs of probabilities may then be compared simultaneously and, if equality of the probabilities is rejected, other tests are available to determine which probabilities lead to the rejection of the null hypothesis of equal probabilities (see discussion in Stuart and Ord, Section 29.53, and references therein).

7 Discussion

This paper has covered a good deal of material. It has demonstrated that the most commonly used model for aggregate voting analysis is a member of a more general class of models, and that this general class of models has very useful properties. The implications of this analysis allows a straight-forward calculation of potential "bias" through grouping and development of models, which, with suitable restrictions, allow the estimation of such bias. It has derived the distribution of an estimator which has been used in court cases for a number of years. It has provided the solution to a problem which has been "unsolved" for over forty year, that is, the transition model. And it has developed a straight-forward way of combining aggregate and individual data. Furthermore, it has done it within the context of probability theory, from which the entire field of statistics is derived.

This last point deserves emphasis. Probability theory rests on three axioms, and upon accepting those axioms, the methods of mathematics can then be used to ascertain the truth or falsity of an assertion. It one asserts that a probability model is true at the individual level then by (15) one *knows* the probability model at the aggregate level. The key conceptual

the competing candidates. It is difficult to see how a model incorporating such influences could be specified and an appropriate variable measured at the aggregate level."

[52]See Appendix I for a discussion of parameterization.

[53]This assumes that there is sufficient information to place the individual in the same group as that individual would have been in the aggregate group. For most data sets, this would seem to be the case, particularly the ANES ones.

misunderstanding here on the part of the literature in the social sciences, is, of course, that if one has n identical objects producing either one of two values, then the central limit theorem applies, that is, one has a normal approximation to the sum of bernoulli's. When one has whites and blacks in a precinct, and each type of person is assumed to be identical to every other person of that type, one has the sum of two normals. This concept of probabilistic aggregation is missing from nearly all of the papers referenced in this article from political science, history and sociology and is present in nearly all of the articles referenced from statistics.[54] The converse problem exists in the statistical literature, which is practically all concerned with voter transitions. In that body of work, there is no concept of modelling individual decisions and thus no inkling that modelling one voting decision as a linear function of other voting decisions is making that first voting decision a function of *realizations* of random variables, and that those other voting decisions are parameterized by other, exogenous, variables.

An important question (particularly for a substantive field which political science, history and sociology all claim to be) is whether what has been presented here matters, that is, can it make any substantive effects on the answers researchers have been getting from their previous efforts at ecological inference? The answer is a yes, with one qualification. First of all, it is clear that the reformulation of the transition model, which is new to this paper, really is a breakthrough. The methods proposed for combining aggregate and individual data, while straight-forward, offer promise of actually combining what heretofore have been two distinct bodies of evidence. And, of course, the derivation of the formula for "bias" and the distribution of the traditional two-equation turnout and vote equations, under the assumption of consistency of the linear probability model at the individual level, are new. So all of these contribute to the yes side of the ledger.

On the no side of the substantive differences, when one is estimating the response probabilities of two (or more) groups with one election, the the common factor of nearly all the likelihoods or methods in the literature is that they produce the least-squares estimator. That is, (1) is pretty much always the equation and $\hat{\beta} = (X'X)^{-1}X'Y$ is pretty much always the estimator,[55] so none of these methods of estimation will make much difference in the location parameters and hence not make much difference in the *substantiative* interpretation of the response probabilities. This follows obviously from the fact that the least squares estimator is consistent (if not efficient) if the error term in (1) is uncorrelated with the X's. The fact that this type of estimator is inefficient usually does not matter in an ecological problem. Consider a congressional district with $n = 200$ precincts and one homogenous

[54]For example, the knowledge of the distribution of the error terms on the linear probability model have been known since at least Goldberger (1964, page 249-250), but might be considered "obscure"–at least this author was unaware of it, which is certainly one definition of obscure. But being very much aware of the aggregation of probabilities, a search was made for a reconciliation of the claim of any residuals in the individual linear probability model with the implications of the central limit theorem–if the same assumptions are made, then the two approaches *had* to coincide–this is the power of mathematics.

[55]With parameterization, this would be the solution to $\sum(Y_j - X_j\beta(z_j, \tau))^2$.

group with probability p. Then

$$\mathrm{Var}(\hat{\beta}) = \frac{X'\mathrm{E}[uu']X}{(\sum X_j^2)^2} = p(1-p)\frac{\sum X_j^3}{(\sum X_j^2)^2} \tilde{=} p(1-p)\frac{1}{n\mu_x} \tag{45}$$

where μ_x is the average value of x. If there are 500 voters in each precinct then the standard deviation of β is under .01. Thus least squares produces estimates for which the substantive interpretation will not change if some other method of estimation is used.[56]

While the inferences drawn from the estimators under different techniques may not matter substantively, what does matter is whether one draws the substantive inferences to begin with. That is, the central problem of aggregate inference (starting with Robinson) has always been that one is very wrong and one gets both excellent goodness of fit statistics and highly significant coefficients. Before parameterization the problem was persistent physically impossible estimates,[57] after parameterization the problem became estimates which were unlikely.[58] Problems like these suggest problems with the assumptions of the model. Another way to determine whether the assumptions are violated is to look at the goodness of fit of the model,[59] and if the goodness of fit is poor, not to use the coefficients to make substantive inferences.[60]. The important point is that for one to assess the goodness of fit correctly one must be modelling correctly, and thus the results of this paper are important substantively in the following sense: they will tell the researcher if the interpretation of the estimates has validity.

From a larger perspective, then, the synthesis here has been taking the usual social science model of individual choice behavior and the usual statistical model of aggregation of random variables. It is iterated once again that once one accepts the consistency of the individual choice, the ecological "fallacy" is resolved.[61] The techniques presented in this paper, though specifically relating to the aggregation of individual choices, can be used for any type of aggregation for which there is a model at one level of inference and it is being applied to another, aggregated level of inference.

---

[56]Discussions of both the Hawkes and Brown and Payne likelihoods and what terms dominate are given in each of their respective papers.

[57]Shivley (1969): "The lack of interest in ecological regression is probably due to the fact that errors in estimation are likely to turn up either as negative percentages or as a percentages which are greater than one hundred. This is disheartening to the researcher, and is difficult to present to his colleagues." It should be pointed out, though, that one *can* produce results with plausible substantive interpretations given a sufficiently well-constructed database (see Tam (1995)).

[58]Brown and Payne: "A persistent feature of the fitted transition matrices obtained is the zero fitted values in some cells."

[59]One statistic that can be used is $\sum(Y_j - X_j\hat{p})'\hat{\Sigma}(Y_j - X_j\hat{p})$, which is distributed asymptotically chi-squared with an appropriate number of degrees of freedom. This is suggested in Brown and Payne.

[60]Actually determining what "poor" is a subject for another paper

[61]Incidentally, the expected value of the correlation coefficient when calculated from aggregate data (assuming individual level consistency), is given in Lupia and McCue in Table 3. The difference between this value and the one obtained from individual level data, as noted by Robinson is, of course, what brought the problem of ecological inference to everyone's attention to begin with.

Articles Referenced.

- Blischke, W. R. (1964). "Estimating the Parameters of Mixtures of Binomial Distributions", *Journal of the American Statistical Association*,59, 510-528.
- Brown, Philip and Clive Payne (1986). "Aggregate Data, Ecological Regression, and Voting Transitions." *Journal of the American Statistical Association*, 81, 452-460.
- Clagett, Willian and John Van Wingen (1993). "An Application of Linear Programming to Ecological Inference: An Extension of an Old Procedure", *American Journal of Political Science*, 37, 2 633-661.
- Cosslett, Stephen. 1981. "Efficient Estimation of Discrete-Choice Models", in *Structural Analysis of Discrete Data with Econometric Applications*, The MIT Press, Cambridge, Massachusetts, London, England.
- Cox, D. R.. 1961. "Tests of Separate Families of Hypothesis", in *Fourth Berkeley Symposium on Mathematical Statistics and Probability* Vol 1 Berkeley. University of California Press, 105-123.
- Feller, William. (1950). *An Introduction to Probability Theory and its Applications: Volume I* , New York.
- Firebaugh, Glenn. (1978). "A Rule for Inferring Individual-Level Relationships from Aggregate Data", *American Sociological Review*, 43 (August), 557-572.
- Goldberger, Arthur. (1964). *Econometric Theory* , John Wiley & Sons, New York.
- Goodman, Leo. (1953). "Ecological Regression and the Behavior of Individuals", *American Sociology Review*, 18, 663-664.
- Goodman, Leo. (1959) "Some Alternatives to Ecological Correlation", *American Journal of Sociology*, 64, 610-625.
- Hannan, J. and W. Harkness. (1963). "Normal Approximation to the Distribution of Two Independent Binomials, Conditional on Fixed Sum", *Annals of Mathematical Statistics*, 34, 1593-1595.
- Hanushek, Eric, John Jackson, and John Kain. (1974). "Model Specification, Use of Aggregate Data, and the Ecological Correlation Fallacy." *Political Methodology*, 1, 89-107.
- Hawkes, A.G. (1969). "An Approach to the Analysis of Electoral Swing", *Journal of the Royal Statistical Society*, Ser A. 132. 68-79.
- Hinkley, D.V. (1969). "On the Ratio of Two Correlated Normal Random Variables", *Biometrika*, 56, 635-639.
- King, Gary. (1986). "How Not to Lie with Statistics: Avoiding Common Mistakes in Quantitative Political Science." *American Journal of Political Science*, 30, 3, 666-687.

- King, Gary. (1995). "A Solution to the Ecological Inference Problem". Unpublished manuscript.
- Kousser, Morgan. (1973). "Ecological Regression and Analysis of Past Politics", *Journal of Interdisciplinary History*, IV, 2, 237-262.
- Lupia, Arthur and Kenneth McCue (1990) "Why the 1980's Measures of Racially Polarized Voting are Inadequate for the 1990s," *Law & Policy*, 12, 355-388.
- Madalla, G.S. (1977) *Econometrics*, McGraw-Hill Book Company, New York.
- Madalla, G.S. (1983) *Limited-dependent and Qualitative Variables in Econometrics*, University of Cambridge, Cambridge.
- MacRae, E. C. (1977). "Estimation of Time Varying Markov Processes With Aggregate Data", *Econometrica*, 45, 183-198.
- Manski, Charles F. and Daniel McFadden. (1981) *Structural Analysis of Discrete Data with Econometric Applications*, The MIT Press, Cambridge, Massachusetts, London, England.
- McCue, Kenneth F. (1992). "The Inference of Individual Probabilities from Aggregate Data", Presented at the Political Methodology Group Meetings, Harvard University, July 15-18th, 1992.
- McCue, Kenneth F. (1994) *California Statewide Election/Registration Merged to Census Data, 1990-1992.* Berkeley, CA; Institute of Governmental Studies, University of California, Berkeley, 1994.
- Miller, W. L. (1972). "Measures of Electoral Change Using Aggregate Data", *Journal of the Royal Statistical Society*, Series A, 135, 122-142.
- Miller, Warren (1955). "Presidential Coattails", *Public Opinion Quarterly*, 19, 353-368.
- Palmquist, Bradley. (1993) "Ecological Inference, Aggregate Data Analysis of U.S. Elections, and the Socialist Party of American", Ph.D. dissertation, University of California, Berkeley.
- Stuart, Alan and J. Keith Ord. (1987) *Kendall's Advanced Theory of Statistics*, Volume I and II, Oxford University Press, New York.
- Rao, C. Radhakrishna. (1973) *Linear Statistical Inference and its Applications.*, John Wiley & Sons, New York.
- Shively, W. Phillips. (1969) "Ecological Inference: The Use of Aggregate Data to Study Individuals", *American Political Science Review*, 63, 1183-1196.
- Shively, W. Phillips. (1974) "Utilizing External Evidence in Cross-Level Inference", *Political Methodology*, 1, 61-73.

- Shively, W. Phillips. (1991) "A General Extension of the Methods of Bounds, with Special Application to Studies of Electoral Transition", *Historical Methods*, 24, 81-94.
- Tam, Wendy. (1995) "Asians: A Monolithic Voting Block?", *Political Behavior*, 17, 2, 223-249.
- Wishart, John. (1949) "Cumulants of Multivariate Multinomial Distributions", *Biometrika*, 36, 47-58.

8 Appendix I

Why does parameterization work? One possibility is that the individual probability really is parameterized by what is, after all, a contextual variable. Individual models in political science, however, rarely have contextual variables in them, and there is another way of viewing these variables which gives them an interpretation at the level of the individual, using (15). Suppose the number of individuals in the g^{th} group who vote for a candidate have been collapsed from g_H other groups, that is, $K_{gi} = \sum_{j=1}^{G_g} K_{gij}$, where i is the electoral unit. Then

$$\mathrm{E}[K_{gi}] = \mathrm{E}[\sum_{j=1}^{g_H} K_{gij}] = \sum_{j=1}^{g_H} n_{gij} p_{gj}, \tag{46}$$

where n_{gij} are the number of individuals in the g_j^{th} group in the i^{th} electoral unit and $n_{gi} = \sum_{j=1}^{g_H} n_{gij}$.

It is obvious from the above the the expected value of K_{gi} varies as i varies, so the location parameter itself is misspecified and thus the whole model. There is a scenario, though, where the misspecification, while still present, is not as great and it relates directly to reparameterization. In this case, it is assumed that n_{gij}/n_{gi} is constant,[62] then $\mathrm{E}[K_{gi}] = n_{gi}\tilde{p}_g$, where now $\tilde{p}_g = \sum_{j=1}^{g_H} n_{gij} p_{gj}/n_{gi}$, and $\tilde{p}_g$ is now constant. Then the variance under the assumptions of one homogenous group is

$$n_{gi}\tilde{p}_g(1-\tilde{p}_g) = n_{gi}\frac{\sum_{j=1}^{g_H} n_{gij} p_{gj}}{n_{gi}}(1-\tilde{p}_g) = \sum_{j=1}^{g_H} n_{gij} p_{gj}(1-\tilde{p}_g). \tag{47}$$

The true variance is, of course,

$$\mathrm{Var}[K_{gi}] = \sum_{j=1}^{g_H} \mathrm{Var}[K_{gij}] = \sum_{j=1}^{g_H} n_{gij} p_{gj}(1-p_{gj}), \tag{48}$$

and the difference of these two expressions is

$$n_{gi}\tilde{p}_g(1-\tilde{p}_g) - \mathrm{Var}[K_{gi}] = \sum_{j=1}^{g_H} n_{gij} p_{gj}(p_{gj}-\tilde{p}_g). \tag{49}$$

[62] Cases where such constancy might arise are such things as the ratio between men and women. Of course, any such constancy is unlikely to be perfectly constant, thus producing another source of error.

It will not be possible *a priori* to determine whether this variation is greater or less than zero, as it depends on the covariation between the number of individuals of each subgroup of g in the electoral unit and the difference of that subgroup probability from $\tilde{p}_g$. The important thing to note is that, as Brown and Payne point out, the estimation term involving the location parameter is by far the most important in this estimation, and the location parameter under this scenario is constant. Thus, while there is misspecification under this scenario, it will not bias the estimates asymptotically.

The relation to the parameterized estimation is as follows. If one has a covariate r such that $n_{gij} = n_{gj}(n_{gi}, r_j, \psi)$, then

$$\sum_{j=1}^{g_H} \frac{n_{gij} p_{gj}}{n_{gj}} = \sum_{j=1}^{g_H} \frac{n_{gj}(r_j, \psi)}{n_{gj}} p_{gj} = \tilde{p}_g(r_j, \psi), \tag{50}$$

becomes the new expression for $\tilde{p}_g$. From the above, this is a constant location parameter (conditional on the r_j), but the variance is wrong, with the difference for any electoral unit being

$$\sum_{j=1}^{g_H} n_{gj}(r_j, \psi) p_{gj}(p_{gj} - \tilde{p}_g(r_j, \psi)). \tag{51}$$

Thus parameterization (if there is one which acts as described here), will make the location parameter unbiased but it will change the variation. Given that the term where the location parameter is present dominates, however, this will produce a consistent estimate, but it is likely that various measures of goodness of fit (such as likelihood ratios, walds, and so forth) will be higher than they would be ordinarily, even though the parameter estimates would be very close to the actual value. This suggests empirical examination with actual or simulated data will be necessary to determine the exact nature of this effect under parameterization.

An example might help make clear some of the ideas to this appendix. Suppose there are actually two groups with common probabilities but instead only one group is modelled. Then one has

$$Y_j = X_{1j} p_1 + X_{2j} p_2 + u_1 + u_2 = X_j p(r_j, \psi) + u_z, \tag{52}$$

where $X_j = X_{1j} + X_{2j}$. Then,

$$p(r_j, \psi) = \frac{X_{1j} p_1 + X_{2j} p_2}{X_j} = r_{1j} p_1 + r_{2j} p_2, \tag{53}$$

where $r_{ij} = X_{ij}/X_j$. If one actually models $p(r, \psi)$ as the above, the mean vector is the same and the difference of the variances using (51) is

$$X_j r_{1j} p_1 [p_1 - (r_{1j} p_1 + r_{2j} p_2)] + X_j r_{2j} p_2 [p_2 - (r_{1j} p_1 + r_{2j} p_2)] \tag{54}$$

In this case, even while modelling as one group, one could write the likelihood as a function of the p_1 and p_2 with the correct variance and perform an optimization (though of course it would be simpler to simply estimate as two groups).

Usually, though, things won't be that simple. Suppose the two homogenous groups are defined by an income variable, with members of one group being defined as individuals below

a certain level of income, and members of the other group being define above a certain level of income. If the parameterizing variable z is simply the level of income in the electoral unit, as is typical when census data is merged to the electoral unit, it will not provide a perfect tracking for the two groups (as with the above), but instead will be highly correlated with it. Then there is misspecification once again, and no easy analytical way of displaying the misspecification. Once again empirical examination will be necessary to determine what the effects of this misspecification are.

This is probably the appropriate place to comment on district analysis, that is, ecological regression when the electoral unit is the district rather than the precinct. Suppose that parameterization takes place at precinct level. If $Y_{dj} = \sum_i^{n_j} Y_{dji}$ is parameterized at the precinct level, then $Y_d = \sum_{j \in d}^{D_d} \sum_i^{n_j} Y_{dji}$ will be the convolution of D_d random variables with no common distribution. This is the same problem as discussed in section 2, where the likelihood had too many parameters to estimate. Thus any estimation at the district level when parameterization takes place at the precinct level will be misspecified. Only if one can make precinct estimates without parameterization can one then make district estimates, with or without parameterization.

9 Appendix II

The extension made here to the model presented in Lupia and McCue now allows the estimation of transition coefficients between the candidate choice in two separate races. That is, if an individual faces two races on the ballot (such as the race for President and the race for House of Representatives), the assumption is made that the the individual chooses one and only one choice on each race.[63] Thus, if there are C_1 candidates for the first race and C_2 for the second, the individual falls into one of $C_1 C_2$ cells.

It is assumed that there are G groups in the electorate, each group being chosen so that they are homogenous (that is, the individual follows a multinomial probability law which is common to all members of the group). Parameterize the probabilities of the multinomial by the two indices $c_1 = 1, \ldots, C_1$, $c_2 = 1, \ldots, C_2$, with the probability of an individual in group g choosing candidate c_1 in the first race and c_2 in the second race being $p_{c_1 c_2 g}$.

Arranging these probabilities in a table, one has

$$(p_{11g}, \cdots, p_{C_1 C_2 g}) \quad = \quad \begin{array}{ccc|c} p_{11g} & \cdots & p_{1C_2 g} & \beta_{11g} \\ \vdots & \ddots & \vdots & \vdots \\ p_{C_1 1g} & \cdots & p_{C_1 C_2 g} & \beta_{1C_1 g} \\ \hline \beta_{21g} & \cdots & \beta_{2C_2 g} & \end{array} \tag{55}$$

where,

[63] These choices can include such things as not voting or "rolling-off" from one race to the other.

$$\beta_{rcg} = \begin{cases} \sum_{c2=2}^{C_2} p_{cc_2g}, & r = 1 \\ \sum_{c1=1}^{C_1} p_{c_1cg}, & r = 2 \end{cases} \tag{56}$$

Here, then, β_{rcg} being the probability of the individual in group g choosing the c^{th} candidate in the r^{th} race. These β_{rcg}'s are the usual parameter of interest in ecological regression (as opposed to the transition model) and it can be seen that estimation of the transition parameters provides an estimate of the ecological regression coefficients.[64] Thus the solution to one problem provides the solution to another.

Using the probabilities above, the (random) race results can be written using the multivariate normal approximation to the multinomial distribution as

$$\text{Prob}\,(Y_1, \cdots, Z_{C_2}) \;=\; \begin{array}{ccc|c} \sum_g X_g p_{11g} + u_{11g} & \cdots & \sum_g X_g p_{1C_2g} + u_{1C_2g} & Y_1 \\ \vdots & \ddots & \vdots & \vdots \\ \sum_g X_g p_{C_1 1g} + u_{C_1 1g} & \cdots & \sum_g X_g p_{C_1 C_2 g} + u_{C_1 C_2 g} & Y_{C_1} \\ \hline Z_1 & \cdots & Z_{C_2} & \end{array}$$

where X_g is the number of voters in group g and the u are distributed multivariate normal with covariance terms as follows:

$$\text{E}\,[u_{c_1d_2g}u_{d_1c_2h}] = \begin{cases} -X_g p_{c_1d_2g} p_{d_1c_2g}, & g = h,\ c_1 \neq d_1 \text{ or } c_2 \neq d_2 \\ -X_g p_{c_1c_2g}\,(1 - p_{c_1c_2g})\,, & g = h,\ c_1 = d_1 \text{ and } c_2 = d_2 \\ 0 & g \neq h \end{cases} \tag{57}$$

To maximize the likelihood function the covariance matrix must be calculated. The means and covariances are calculated from the following:

$$Y_{c_1} = \sum_{c_2=1}^{C_2} \sum_g (X_g p_{c_1c_2g} + u_{c_1c_2g}) = \sum_g (X_g \beta_{1c_1g} + v_{1c_1g})\,, \tag{58}$$

$$Z_{c_2} = \sum_{c_1=1}^{C_1} \sum_g (X_g p_{c_1c_2g} + u_{c_1c_2g}) = \sum_g (X_g \beta_{2c_2g} + v_{2c_2g})\,, \tag{59}$$

with

$$v_{rcg} = \begin{cases} \sum_{c2=2}^{C_2} u_{cc_2g}, & r = 1 \\ \sum_{c1=1}^{C_1} u_{c_1cg}, & r = 2 \end{cases} \tag{60}$$

[64] Stricly speaking, these are not transition probabilities are being calculating, since the transition probabilities are the conditional probability of moving to a choice in the second race once the first race has been chosen–thus, $t_{c_1c_2g} = p_{c_1c_2g} / \sum_{c_2=1}^{C_2} p_{c_1c_2g}$ is the usual transition probability. The notion of simultaneous choice here makes the parametrization by the cell probabilities more natural and the transition probabilities are simple transforms of these probabilities.

The means are obvious from the above. For the variances, one has

$$\mathrm{Var}\left(Y_{c_1}|X,p\right)=\mathrm{Var}\left(\sum_g v_{1c_1g}\right)=\mathrm{Var}\left(\sum_g\sum_{c_2=1}^{C_2}u_{c_1c_2g}\right)=\sum_g X_g\beta_{1c_1g}\left(1-\beta_{1c_1g}\right), \tag{61}$$

and similarly,

$$\mathrm{Var}\left(Z_{c_2}|X,p\right)=\sum_g X_g\beta_{2c_2g}\left(1-\beta_{2c_2g}\right), \tag{62}$$

$$\mathrm{Cov}\left(Y_{c_1}Y_{d_1}|X,p\right)=-\sum_g X_g\beta_{1c_1g}\beta_{1d_1g}, \tag{63}$$

and

$$\mathrm{Cov}\left(Z_{c_2}Z_{d_2}|X,p\right)=-\sum_g X_g\beta_{2c_2g}\beta_{2d_2g}. \tag{64}$$

Finally,

$$\begin{aligned}
\mathrm{Cov}\left(Y_{c_1}Z_{c_2}|X,p\right) &= \mathrm{Cov}\left(\sum_g v_{1c_1g},\sum_g v_{2c_2g})\right)\\
&= \mathrm{E}\left[\sum_g\sum_{d_2=1}^{C_2}u_{c_1d_2g}\sum_{d_1=1}^{C_1}u_{d_1c_2g}\right]\\
&= \sum_g X_g\left[p_{c_1c_2g}-\sum_{d_2=1}^{C_2}\sum_{d_1=1}^{C_1}p_{c_1d_2g}p_{d_1c_2g}\right]\\
&= \sum_g X_g\left[p_{c_1c_2g}-\sum_{d_2=1}^{C_2}p_{c_1d_2g}\sum_{d_1=1}^{C_1}p_{d_1c_2g}\right]\\
&= \sum_g X_g\left[p_{c_1c_2g}-\beta_{1c_1g}\beta_{2c_2g}\right].
\end{aligned}$$

It is this last covariance term which allows the identification of the transition probabilities. The intuitive interpretation is the same as that of independence in a n by m table–that is, if all of the covariance terms of the Y and the Z are zero, then the choices for the two races are independent (the cell parameter is simply the product of the marginal sums, which in this case are the $\beta's$).

Letting 1_n be an $nx1$ vector of ones and 1_{nxm} be a nxm matrix of ones, the final covariance matrix becomes

$$\begin{aligned}
\mathrm{Var}\left(Y,Z|X,p\right) &= \sum_g X_g\left[\mathrm{diag}(\beta_g)-\beta_g\beta_g^t+\begin{pmatrix}0 & P_g\\ P_g^t & 0\end{pmatrix}\right]\\
&= \sum_g X_g\begin{bmatrix}\mathrm{diag}(P_g1_{C_2})-P_g1_{C_2xC_2}P_g^t & P_g-P_g1_{C_2xC_1}P_g\\ P_g^t-P_g^t1_{C_1xC_2}P_g^t & \mathrm{diag}(P_g^t1_{C_1})-P_g^t1_{C_1xC_1}P_g\end{bmatrix},
\end{aligned}$$

where the last identity comes from the fact that

$$\beta_g=\begin{pmatrix}\beta_{1g}\\ \beta_{2g}\end{pmatrix}=\begin{pmatrix}P1_{C_2}\\ P^t1_{C_1}\end{pmatrix} \tag{65}$$

and so

$$\beta_g \beta_g^t = \begin{bmatrix} P_g 1_{C_2 x C_2} P_g^t & P_g 1_{C_2 x C_1} P_g \\ P_g^t 1_{C_1 x C_2} P_g^t & P_g^t 1_{C_1 x C_1} P_g \end{bmatrix}. \tag{66}$$

As is usual in these case, p can be parameterized by function of covariates. The form used here is

$$p_{c_1 c_2 g} = \frac{e^{z\theta_{c_1 c_2 g}}}{\sum_{d_1=1}^{C_1} \sum_{d_2=1}^{C_2} e^{z\theta_{d_1 d_2 g}}}, \tag{67}$$

where the z is a vector of covariates (including one) and $\theta_{C_1 C_2 g}$ is set equal to zero for all g for identification.

The relationship between the likelihood presented here and the likelihoods presented in the statistical literature on voting is as follows. For only one race, the likelihood for this model without a reparameterization of the probabilities as a function of exogenous variables and weights is given in Hawkes, the likelihood with such a reparameterization is given in Brown and Payne (for the case when the compound multinomial parameter is infinite). For two races, this likelihood is not given in the literature. Incidentally, this model can be extended to three or more races if some of the multinomial probabilities are restricted (usually to zero).

The arrangement of the data in (55) is an example of a probability distribution called the bivariate multinormal (cumulants for this distribution are given in Wishart). The Hawkes model (and the Brown and Payne model without the compounding by the Dirichlet distribution) are examples of convolutions of univariate multinomials. The convolution for bivariate (or more generally, multivariate) multinomials is a "new" distribution,[65] and it is therefore titled the aggregate multivariate multinomial distribution.

[65]Brown and Payne note that the compound multinomial has been used before but that the "aggregate compound multinomial", that is, the convolution of the compound multinomials, is new. This distribution is thus new in this sense.

FIGURE 1
GOODMAN VS. ACTUAL TRANSITION VALUES

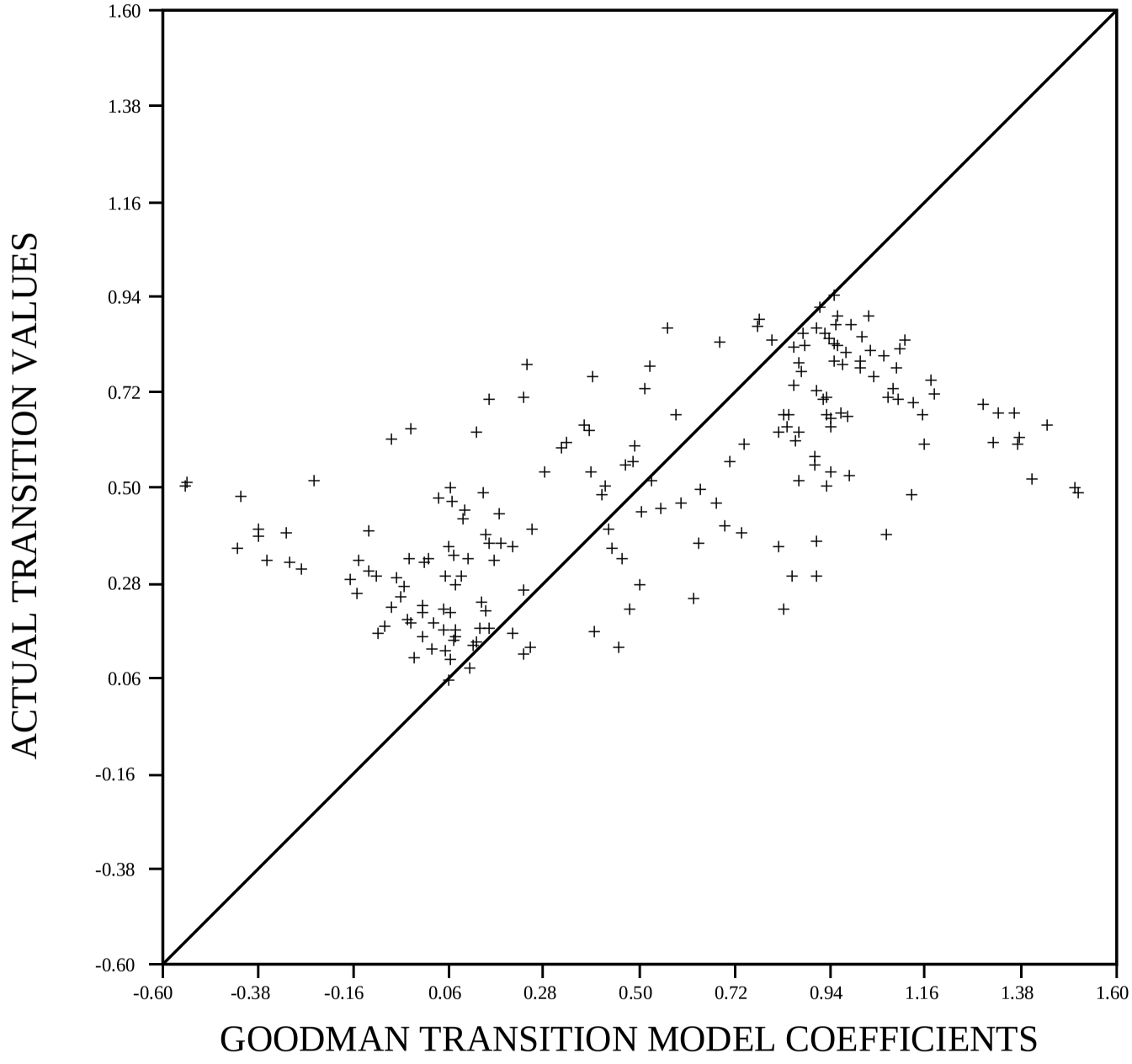

FIGURE 2
GOODMAN VS. PREDICTED VALUES

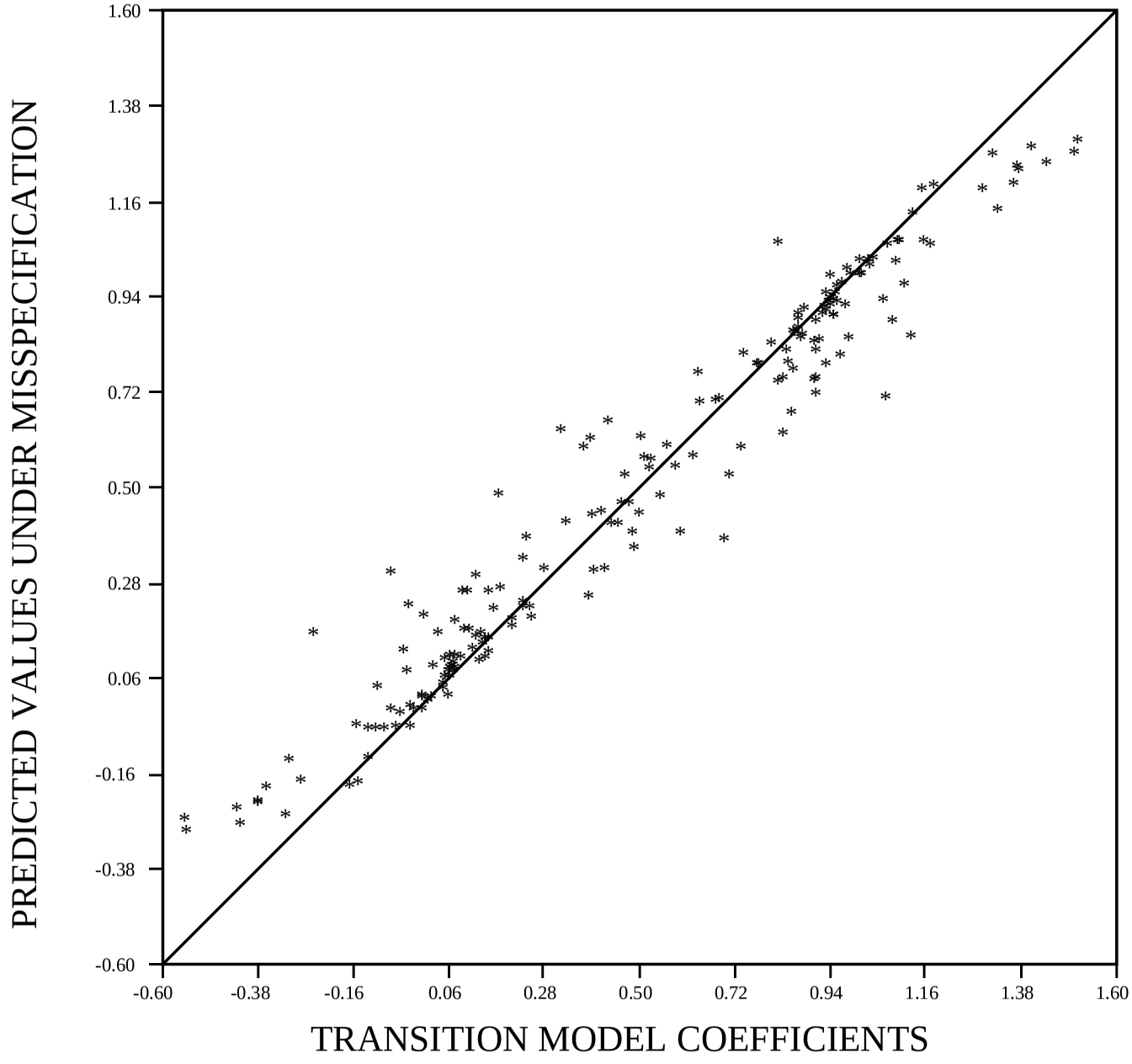

FIGURE 3
INDIVIDUAL MODEL VS. ACTUAL TRANSITION VALUES

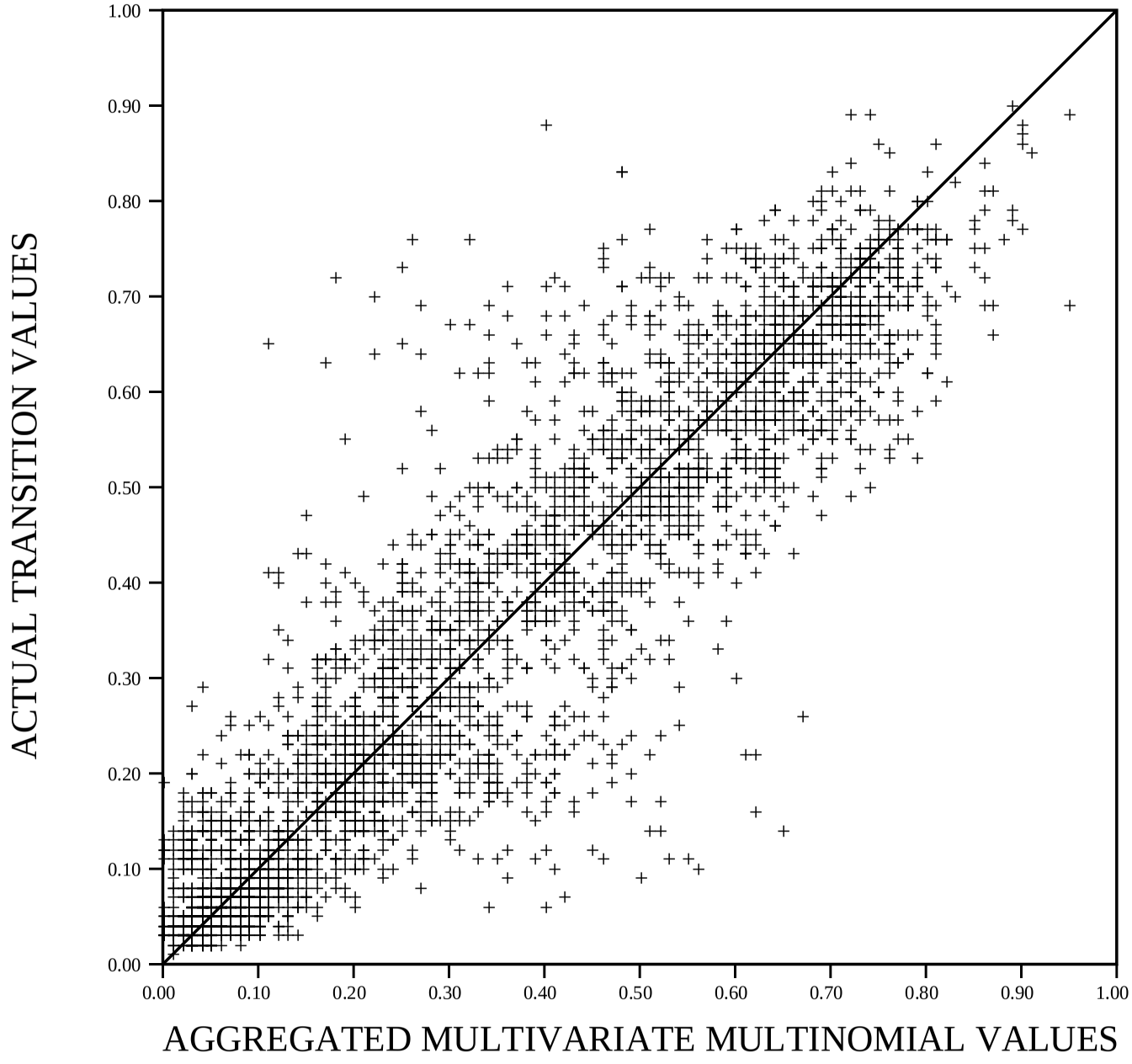

FIGURE 4
INDIVIDUAL MODEL VS. CHOICE PROBABILITIES