Pergamon

Electoral Studies 20 (2001) 243–263

Electoral Studies

www.elsevier.com/locate/electstud

Latent groups and cross-level inferences

Wendy K. Tam Cho [*]

Department of Political Science and Department of Statistics, University of Illinois at Urbana-Champaign, 361 Lincoln Hall, 702 South Wright Street, Urbana, IL 61801-3696, USA

Accepted 29 March 2000

Abstract

It is common for the only available data on interesting political phenomena to be aggregated at a level above the micro-unit in question. Analysis of voting behavior in elections for which survey data are unavailable is a case in point: one often must draw inferences about voters by analysing districts or precincts. Using OLS to analyse aggregate data (i.e. ecological regression) implicitly assumes constancy of parameters across aggregate units. This assumption is rarely tenable, since the aggregation process usually generates macro-level observations across which the parameters describing individuals vary. A key step in aggregate data analysis, then, is to identify covariates that separate macro-units into subgroups in which individual behavior is roughly constant. A switching regression context is proposed where the state-defining variable measures homogeneity between the macro-level units. © 2001 Elsevier Science Ltd. All rights reserved.

Successful statistical models have been devised for many of the most fundamental, interesting problems in social science. For instance, advances in survey techniques in the last 30 years now allow us to make reliable inferences from relatively small samples. These successes, however, translate only to situations in which data are available on the group of interest. Models for how to make cross-level inferences remain plagued by widespread disagreement. The ability to make consistently reliable inferences to individuals from aggregate data would be monumental, particularly for those who study election returns. Unfortunately, a “solution” to this problem is not likely to be forthcoming since the ecological inference problem, as an instance of an ill-posed inverse problem, fundamentally has no unique solution. To complicate

[*] Corresponding author. Tel.: +1-217-333-9588.
E-mail address: wendy@cho.pol.uiuc.edu (W.K.T. Cho).

PII: S0261-3794(00)00011-1

matters, choosing among the array of possible solutions amounts to choosing among assumptions in building a model. If the models were robust to violations in the assumptions, this would not be problematic. However, they are not, as the assumptions in fact determine the results (Freedman et al., 1991).

Despite these inauspicious conditions, some real-life applications demand the use of ecological inference models. Situations where the only available data are aggregated at a level other than the level of interest are common. The Voting Rights arena in the United States is just one prominent example. In order to render decisions in these cases, the judge must determine how minority voters cast their ballots. If there were individual-level data, these inferences would be straightforward. However, these data are virtually never available. Instead, the available data are votes at the precinct-level. Hence, we are faced with the non-trivial problem of inferring individual-level behavior from the aggregated precinct-level data. Unfortunately for judges who must render a decision, a general description of the difficulties of the estimation problem is neither sufficient nor helpful. The issue in these cases, whether relief will be granted under the Voting Rights Act, needs to be resolved definitively. Hence, although no unique solution exists, a solution is demanded and must be provided. Moreover, although expert witness testimony in court cases is restricted to evidence that is rooted in a theory that has been developed, reviewed, and validated by the relevant academic community,[1] this situation is obviously not attainable in Voting Rights cases where there is no scholarly consensus on ecological inference models. Instead, it is well-known that making individual-level or cross-level inferences from aggregate data is problematic (Robinson, 1950; Goodman, 1953; Theil, 1971; Freedman et al., 1991; Achen and Shively, 1995; Ansolabehere and Rivers, 1997; Cho, 1998).

The root of the problem is that the standard estimation techniques are not reasonable when the parameters are correlated with the regressors, a condition known as “aggregation bias”. Suppose one is investigating rates of support for different candidates amongst racial groups using precinct-level election returns and data on the population’s racial composition. The data exhibit no aggregation bias if group voting rates are constant, that is, the same in every precinct. Freedman et al. (1991) call this condition the “constancy assumption”. In this context, the constancy assumption is that, outside of random variation, all groups tend to vote for the candidates in the same proportions, regardless of precinct of residence. If this assumption holds, then aggregate data analysis is straightforward. Using OLS, one can regress candidate vote shares on racial share variables to estimate the parameters that describe voting habits for each racial group. Parameters which are constant will not be correlated with any set of regressors, and so cross-level inferences are simple. A strong assumption underlying this model, then, is that all minorities vote alike regardless of, for example, which neighborhood they live in, their income, or their partisanship. In real data, this assumption is generally false. Instead, since individuals in the same

[1] See *Frye v. United States*, 293 F. 1013 (1923) and *Daubert v. Dow Pharmaceuticals*, 509 U.S. 579 (1993).

geographic unit tend to resemble each other in unmeasured ways, the more common result will be varying parameters which may well be correlated with the regressors.

Ultimately, because the ecological inference problem is fundamentally unidentified, there is no solution. Attempts to derive a general estimator are thus not likely to be fruitful. Adopting a fatalistic view of the problem, however, is detrimental to practitioners who are constrained to find some solution. The goal of this paper is to provide some insight into the problem by examining the assumptions that determine the results and to propose a test for choosing between hypotheses about the underlying individual-level behavior.

1. Replacing constancy with approximate constancy

A useful insight is that while the constancy assumption may not be reasonable for the entire data set, it may be roughly reasonable for subsets of the data. In other words, while *all* minorities in a data set on the 1992 U.S. Presidential election may not have voted for Clinton at the same rates regardless of precinct of residence, all minorities who lived in wealthy precincts may have voted for Clinton at roughly the same rates. Likewise, all minorities who resided in predominantly Democratic precincts may have behaved similarly, as may those who lived in predominantly Republican precincts. The idea that “behavioral clustering” would occur is not, of course, new to the study of elections or to theories of political behavior. The empirical evidence strongly suggests that geographical units often house definable, politically important characteristics (e.g. Berelson et al., 1954; Putnam, 1966; Miller, 1977; Huckfeldt, 1979).

In the context of aggregate data analysis, if one could specify subsets of the data wherein the constancy assumption is reasonable, the estimation task would be considerably easier. The idea behind the model in this paper is that there are often latent groups in aggregate data wherein the individuals of interest behave homogeneously. These latent groups are defined by some set of variables (e.g. income, race, and/or partisanship). The problem, then, is to determine which variables identify the latent groups best. Choosing these variables properly is fundamental to the ecological inference problem. If we condition on the right variables, the parameters will be mean independent of the regressors, and estimation will be straightforward. If we condition on the wrong variables, the model will be mis-specified and unhelpful. This paper proposes one means of identifying and incorporating appropriate covariates. Ultimately, other methods of covariate selection may be proposed and refined. In that respect, I make no claims about presenting any type of unique solution to the ecological inference problem. Rather, I propose one method of gaining insight into the problem by addressing aggregation bias through an attempt to account for group homogeneity.

It is unlikely that conditioning on any set of covariates will produce subsets of the data where the constancy assumption will hold *exactly*. Instead, we can hope to fulfill only an approximately constant assumption. That is, even if we successfully surmise that there are two groups with distinct behavior in our data set, there is likely to be some variation within the two distinct groups to model. However, as we will see, how one models this within-group variation (e.g. via a random coefficients model or a Markov switching process) is not nearly as consequential as conditioning on the proper covariates in the first place.[2]

King (1997) has recently proposed that the "solution" to the ecological inference problem can be found in a random coefficient model. Unlike OLS models, where the parameters are assumed to be constant across observations, the King model assumes that the parameters vary according to a truncated bivariate normal distribution. If this distribution describes the underlying data well, then the parameters will be mean independent of the regressors. If it does not describe the data well, aggregation bias will still be problematic.[3] When aggregation bias exists, the King model performs poorly, just as OLS performs poorly on data with aggregation bias. Consider the results of a Monte Carlo simulation in Fig. 1. Here, we have the simplest of ecological inference problems, as we are modeling a binary dependent variable (say, vote or not) with a binary independent variable (black or white). The aggregate equation regresses turnout rate on percent black. The parameters β^b and β^w represent black and white turnout rates. The data were generated randomly from a truncated

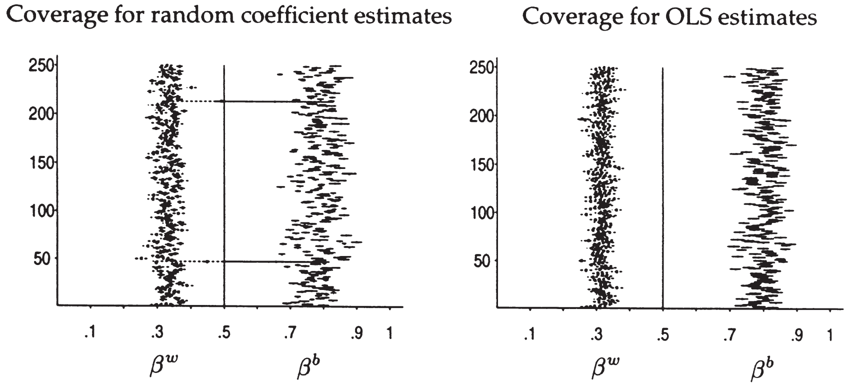

Fig. 1. Error bar plots from a Monte Carlo simulation with data that are inconsistent with the aggregation bias assumption. The true parameter values are marked by the long vertical lines. The error bars to the left of the vertical line are for β^w. The error bars to the right of the vertical line are for β^b. Both β^b and β^w have a true parameter value of 0.5.

[2] While random coefficient models encompass a wide variety of models with origins that far pre-date King's use of them, references to random coefficient models in this paper will focus primarily on their use in King's ecological inference model (King, 1997). In particular, he proposes that a random coefficient model that incorporates a truncated bivariate normal distribution is a "solution" to the ecological inference problem. In this paper, estimates from his random coefficient model will be obtained using his *EzI* software.

[3] A significant problem is that the researcher is unable to surmise whether the distribution is appropriate or not. King advocates some diagnostics, but these diagnostics are severely limited in this regard (Cho, 1998; Freedman et al., 1999). They are often unable to diagnose potential problems with the analysis and are primarily subjective.

bivariate normal distribution with means (on the untruncated scale) $\beta^b = \beta^w = 0.5$, standard deviations $\sigma_b = 0.4$ and $\sigma_w = 0.1$, and correlation $\rho = 0.2$. The true parameter values, $\beta^b = \beta^w = 0.5$, are marked by a vertical line at 0.5 in both plots. The figure shows that even after accounting for the standard errors, if aggregation bias exists in the data set, the estimates from both OLS and King's model are inaccurate and misleading.

King recognizes the problem with aggregation bias, and although he admonishes researchers that covariates may be useful, he supplies no formal method for selecting suitable covariates (see King, 1997, Chapter 16). This is an extremely large and consequential omission. The crux of the problem, understanding the conditions under which data will aggregate consistently and without bias, has not been addressed, but clearly needs to be. Moreover, care needs to be exercised in the selection of covariates, since there is no disagreement that the inclusion of improper covariates produces a biased and inconsistent model (Achen and Shively, 1995; King, 1997; Cho, 1998; Freedman et al. 1998, 1999). Covariates cannot be chosen using only qualitative beliefs if one hopes to obtain reasonable results. Instead, a formal method of selecting suitable covariates is necessary.

I now turn to a discussion of how one might choose variables such that, if the model conditions on these variables, the parameters will be independent of the regressors. A switching regression framework will be adopted. Certainly, a random coefficient framework can be adopted and the results using this framework will be presented as well. However, as we will see, the crucial decision is the inclusion of proper covariates and not whether one incorporates a random coefficient or a fixed parameter model. Once the covariates are chosen, the framework is of distant secondary importance.

2. Switching regression with appropriate covariates

Consider a switching regression model in the context of a Voting Rights claim. For the sake of simplicity and mathematical ease, the discussion is limited to the two-group model. The fixed parameter model for two groups is

$$y_i = px_i + q(1-x_i) + e_i \quad i = 1,\ldots, P. \tag{1}$$

In equation 1, let

p = proportion of minority voters who voted for a candidate,
q = proportion of majority voters who voted for a candidate,
x_{i} = proportion of voters in precinct i who are minority,
y_{i} = proportion of the vote in precinct i received by a candidate, and
P = number of precincts.

A switching regression framework allows the parameters to vary.[4] The switching

[4] Indeed, there are many different frameworks that one may adopt to model data that exhibit structural shifts. For instance, the switching between a finite number of regimes can be modeled as stochastic. Suppose there are two regimes. Nature chooses whether an observation falls in regime 1 or 2 with

mechanism is controlled by a latent-group-determining variable, D. In latent group 1, when $D = 0$, the parameters are p_1 and q_1 while in latent group 2, when $D = 1$, the parameters are p_2 and q_2. So, the basic switching regression model can then be formulated as

$$y_i = p_1 x_i + q_1(1-x_i) + e_{1i} \quad i \in I_1 \tag{2}$$

$$y_i = p_2 x_i + q_2(1-x_i) + e_{2i} \quad i \in I_2 \tag{3}$$

where the observations on y are generated by two distinct processes that are indexed by I_1 and I_2.[5]

The introduction of the new state-defining variable transforms the initial constant parameter model from equation (1) to the following varying parameter model by multiplying (2) by $(1-D_\mathrm{i})$, (3) by D_i, and then combining the two equations into the single equation

$$\begin{aligned} y_i &= [D_i p_2 + (1-D_i)p_1]x_i + [D_i q_2 + (1-D_i)q_1](1-x_i) + (1-D_i)e_{1i} + D_i e_{2i} \qquad (4) \\ &= [D_i(p_2-q_2) + (1-D_i)(p_1-q_1)]x_i + [D_i q_2 + (1-D_i)q_1] + [D_i e_{2i} + (1-D_i)e_{1i}]. \end{aligned}$$

with parameters p_1, p_2, q_1, and q_2. These parameters must be estimated, and a procedure for determining which observations fall into which latent groups must be utilized to estimate D.

2.1. The likelihood function

One can estimate the parameters through the method of maximum likelihood. In seeking the likelihood function of expression (4), we assume a fixed nonstochastic

unknown probabilities λ and $1-\lambda$. If the error terms are normally distributed, the likelihood function,

$$L = \prod_{i=1}^{P} \left(\frac{\lambda}{\sqrt{2\pi}\sigma_1} \exp\left\{ -\frac{1}{2\sigma_1^2}(y_i - x_i'\beta_1)^2 \right\} + \frac{1-\lambda}{\sqrt{2\pi}\sigma_2} \exp\left\{ \frac{1}{2\sigma_2^2}(y_i - x_i'\beta_2)^2 \right\} \right),$$

is a mixture of two normal components. Certainly the mixture of distributions is not restricted to the normal distribution and could be modeled as a mixture of truncated normals or logistic distributions to accommodate our restricted area of support. In addition, the probability, λ, can be defined as a function of a set of covariates. Modeling λ as a function of covariates allows a researcher to impose some substantive theory into the structure. In addition, this framework allows one to model the regime-generating process as stochastic, which is essentially what King achieves through the incorporation of the truncated bivariate normal distribution. However, this method may be less than straightforward in a number of instances because the likelihood surface for various parameter combinations may not be concave and/or may have many local maxima. In these instances, the usual estimation procedures may perform very poorly. Hence, while the theory behind the model may be reasonable, we may be hindered by numerical complications. In any event, as we will see, the main issue addressed in this paper, accounting for aggregation bias, is no more easily achieved in this framework than in any other. The same issues remain; and the advantages gained by this framework are not clearly better, especially after one considers the computational complexity.

[5] This model provides a simple approach to the problem that is more reasonable than incorporating the constancy assumption. The "true" model in a given instance may not be as simple as one which incorporates two states. More states may be necessary.

X matrix. Let $\beta_1 = (q_1, p_1-q_1)$ and $\beta_2 = (q_2, p_2-q_2)$. The likelihood, $L(y \mid \beta_1, \beta_2, \sigma_1^2,\sigma_2^2)$, depends on β_1, β_2, σ_1^2, and σ_2^2. In particular,

$$\log L=-\frac{P}{2}\log 2\pi-\frac{1}{2}\sum_{i=1}^{P}\log[\sigma_1^2(1-D_i)^2+\sigma_2^2D_i^2] \quad (5)$$

$$-\frac{1}{2}\sum_{i=1}^{P}\frac{(y_i-x_i'[\boldsymbol{\beta}_1(1-D_i)+\boldsymbol{\beta}_2D_i])^2}{\sigma_1^2(1-D_i)^2+\sigma_2^2D_i^2}.$$

Since the latent groups are unknown, the D's must be estimated along with the β's and σ's. We assume that D_{i} is a function of s explanatory variables with observations $z_{\mathrm{i}1},\ldots, z_{\mathrm{is}}$ for each precinct i, where i=1,..., P. The state-defining variable D can now be approximated through one of two methods depending on whether we believe that D should provide perfect or imperfect discrimination between latent groups.

2.1.1. Pure groups

If we believe that D_{i} should provide perfect discrimination and be exactly 0 or 1 in every instance, i.e. with certainty, every observation belongs in a certain latent group, then we need to approximate D by D^*, a continuous unobserved variable where D_i^* is distributed

$$D_i^*\sim f(d_i^* \mid z_i\gamma),$$

and γ is a vector of unknown coefficients. We then assume the following relationship with an observed dummy variable, d_{i},

$$d_i=\begin{cases}0 & \text{if } d_i^*\leq 0\\ 1 & \text{if } d_i^*>0.\end{cases} \quad (6)$$

The observed values, then, are just realizations of a binomial process that vary from trial to trial dependent on the covariates. If we approximate $\mathrm{D_i}^*$ by a standardized logistic distribution,

$$D_i^*\sim f(d_i^* \mid z_i\gamma)=\frac{\exp\{d_i^*-z_i\gamma\}}{(1+\exp\{d_i^*-z_i\gamma\})^2},$$

we will have

$$\Pr(D_i=1)=\Pr(D_i^*\leq 0)=\int_{-\infty}^{0} f(d_i^* \mid z_i\gamma)dd_i^*=[1+\exp(-z_i\gamma)]^{-1} \quad (7)$$

This threshold estimator provides a framework with pure groups where shifts between groups are discrete and definite.

When the “pure group formulation” is employed, one places strong constraints on the nature and type of groups that are assumed to exist in the data. In the voting example, for instance, minority voters residing in wealthy precincts may behave distinctly from minority voters living in poor precincts. If pure groups are assumed, each precinct would be classified as either wealthy or poor. None of the precincts would be regarded as falling inbetween this discrete grouping. If this grouping approximates the underlying structure of the data well and the inclusion of these variables results in uncorrelated parameters and regressors, the model will produce good estimates. In some cases, this formulation may be too constraining.

2.1.2. *Hybrid groups*

One way to alleviate the stringent nature of this pure group assumption is to incorporate the notion of a “hybrid group” in place of the discrete groups. In this case, we would use the state-defining variable, D, to distinguish transitions between pure groups in probability. Here, D would be a probability rather than a dichotomous or polychotomous variable as above. So, observation i is in group 1 with probability D_i and in group 2 with probability $(1-D_\mathrm{i})$. The shift between groups is no longer treated as discrete. Instead, observations are allowed to exist in hybrid groups. So, for instance, an observation would not need to be classified as either strictly wealthy or strictly poor but could fall inbetween these two classifications.

To estimate this model, we can approximate D_i by

$$D_i=[1+\exp(-z_i\gamma)]^{-1}. \tag{8}$$

We then replace D_i in equation (5) with the value in equation (8). The likelihood function (5) is then maximized with respect to the β’s, γ’s, and σ’s.

3. Placing observations into latent groups

In order to place observations into groups, we need to choose the covariates, z, which define D, the group-defining variable. The choice of potential covariates should arise naturally from a researcher’s substantive understanding and interest in a problem. That is, there should be a theory that underlies the model. Indeed, this is King’s suggestion for choosing covariates (King, 1997, pp. 284–285). This is also the point where our methods sharply diverge. While I concur that the model should be theory-laden, and so the set of *potential* covariates should arise from a researcher’s substantive knowledge of a problem, my approach, unlike King’s method, does not end here. I do not advocate including a variable in the model simply because one can think of a plausible theory for why the inclusion of this variable might alleviate aggregation bias. Instead, at this crucial step, I propose a test for whether these beliefs are in fact borne out in the data. If the beliefs, as it may turn out, are incorrect, inclusion of such a variable will adversely affect the resulting estimates. Hence, while my method and King’s method allow for the role of theory in the selection of an initial pool of covariates, the ability to test competing theories is absent in King’s model. In that respect, King’s “solution” is missing *the* critical step in ecological inference.

My proposed method amounts to a test for parameter constancy or for a change point in the data. The statistical literature on changepoints is large and encompasses techniques for both cross-sectional data and time-series data (e.g. Quandt, 1960; Brown et al., 1975; Ferreira, 1975; Schulze, 1982; Ploberger et al., 1989; Ritov, 1990; Andrews, 1993). One method is to analyse the stability of the parameters across groups through examining the recursive residuals in a manner similar to that suggested by Brown et al. (1975).[6]

The basic regression model is

$$y_p=\boldsymbol{x}_p'\,\beta_p+e_p \quad p=1,\ldots, P \tag{10}$$

where p indexes the observations (assume that the unit here is the precinct), y_p is the value of the dependent variables in precinct p, x_p is the column vector of observations on k regressors, and β_p is the column vector of parameters for precinct p. The errors are assumed to be independent and distributed $N(0,\sigma^2)$. The null hypothesis is constancy throughout the sample

$$H_0\text{: } \beta_1=\beta_2=\ldots=\beta_p=\beta \tag{11}$$

$$\sigma_1^2=\sigma_2^2=\ldots=\sigma_p^2=\sigma^2. \tag{12}$$

Assume that H_0 is true. Let $\boldsymbol{b}_r=(\boldsymbol{X}_r'\boldsymbol{X}_r)^{-1}\boldsymbol{X}_r'\boldsymbol{Y}_r$ be the least squares estimate of β based on the first r observations. Define the recursive residuals by

$$w_r=\frac{y_r-\boldsymbol{x}_r'\boldsymbol{b}_{r-1}}{\sqrt{1+\boldsymbol{x}_r'(\boldsymbol{X}_{r-1}'\boldsymbol{X}_{r-1})^{-1}\boldsymbol{x}_r}} \quad r=k+1,\ldots,P \tag{13}$$

where $\boldsymbol{X}_{r-1}'=[\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{r-1}]$ and $\boldsymbol{Y}_r'=[y_1,\ldots,y_r]$.

In order to test whether a certain covariate z affects constancy in the sample, order the observations in increasing value of z.[7] The logic is that at some level of the covariate, the behavior changes. Herein is where a researcher can incorporate the idea of behavioral clustering and the long line of theories that have been developed in the political behavior literature. The idea is that clustering of behavior occurs and is a function of some set of observable and measurable covariates. The link between this theory and the statistical test is, for example, if wealthy precincts are different from poor precincts, a test for parameter constancy should detect that a change point exists.

In more empirical terms, if β_p is constant up to some value $p=p_0$ but differs from this value for $p_\mathrm{i}>p_0$, the recursive residuals, w_r's, will have zero means for r up to

[6] The Brown, Durbin, Evans technique is certainly not the only method of testing for parameter constancy. Many techniques have been derived. It is beyond the scope of this paper to detail all of these methods. The claim here is not that the Brown, Durbin, Evans technique is superior but rather that this class of models is useful for group discrimination and that the statistical literature on parameter constancy and changepoints is the obvious reference. Certainly, an extension of this paper might focus on how well the different techniques are able to discriminate between latent groups in aggregate data.

[7] Although the Brown, Durbin, Evans technique is primarily for time-series data, if we order the observations by some ordering, say the increasing value of a covariate, we are able to detect whether there is a departure from constancy of parameters in cross-sectional data at some value of a covariate. This use of the Brown, Durbin, Evans technique in a cross-sectional context has been discussed by Goldfeld and Quandt (1973). Alternatively, we may adopt other techniques for detecting change points in cross-sectional data. The insight here is simply that we can adopt the change point literature.

p_0 but have non-zero means thereafter. In order to test whether the mean has significantly deviated from zero, we examine the plot of the cusum quantity

$$W_r = \frac{1}{\hat{\sigma}} \sum_{k+1}^{r} w_j \tag{14}$$

against the values of r for $r = k+1, \ldots, P$ where $\hat{\sigma}^2 = S_P/(P-k)$ is the estimated variance. To assess the significance of the departure of the sample path of W_r from its mean value line $E(W_\mathrm{r})=0$, we may calculate the probability that w_r crosses one or both of two lines around $W_\mathrm{r}=0$ at a determined significance level α. Specifically, we are interested in the pairs of straight lines through the points $(k, \pm a\sqrt{(P-k)})$, $(P, \pm 3a\sqrt{(P-k)})$ where a is a parameter that is chosen according to the desired significance level.[8] Hence, we reject significance at the 0.05 level if the sample path travels outside of the region bounded by the two lines, $(k, \pm 0.948\sqrt{(P-k)})$, $(P, \pm 3*0.948\sqrt{(P-k)})$.

In addition to the cusum test, additional or supporting information is obtained from the cusum of squares test which utilizes w_r^2, the square of the recursive residual.[9] We plot

$$s_r = \left(\sum_{j=k+1}^{r} w_j^2 \right) \Bigg/ \left(\sum_{j=k+1}^{P} w_j^2 \right) = S_r / S_P, \quad r = k+1, \ldots, P \tag{15}$$

against the ordering of the observations. If the null hypothesis of constancy throughout the sample holds, it can be shown that s_r has a beta distribution with mean $(r-k)/(P-k)$. Hence, we test for significant deviations from the mean value line $E(s_\mathrm{r}) = (r-k)/(P-k)$. If we draw the pair of lines $s_\mathrm{r} = \pm c_0 + (r-k)/(P-k)$ around the mean value line, the probability that the sample path crosses either of these lines is then a chosen significance level α where the value of c_0 differs for each point and depends on whether the value of $(P-k)$ is odd or even. The value $\frac{1}{2}(P-k)$ is closely related to Pyke's (1959) modified form of the Kolmogorov–Smirnov statistic, C_n^+, where $n = \frac{1}{2}(P-k)-1$. Hence, given a significance level α, we may obtain the value of c_0 from the table of significance values for the modified Kolmogorov–Smirnov stat-

[8] Known results in Brownian motion theory yield the following pairs of values for a and α: for a=1.143, α=0.01; for a=0.948, α=0.05; for a=0.850, α=0.10. For a fuller exposition on this point, see Brown et al. (1975) or Harvey (1981, p. 152). For a short summary, see Kmenta (1986).

[9] Both the cusum and the cusum of squares test provide evidence of parameter instability. Evidence for departure from constancy is obviously stronger if it is indicated in both tests. However, since they measure instability differently, only one test needs to be significant to signify parameter instability. For instance, evidence in the cusum squared plot but not the cusum plot might indicate that the instability is the result of changes in the variance of the residuals rather than a shift in the parameter values. Kmenta states that type of change is "haphazard" rather than "systematic" (Kmenta, 1986, p. 578), i.e. there is a stochastic component that does not follow a white noise process. Alternatively, the source of "systematic" change (identified by the cusum test) might be a shift in parameters. A closer examination of the data would be needed to make this assessment. In either case, regardless of the underlying reason for the shift in parameters, if there is such a shift at all, the model specification should reflect these shifts.

istic.[10] If the value of $(P-k)$ is even, the value of c_0 is the entry that corresponds to $n=\frac{1}{2}(P-k)-1$ and $\frac{1}{2}\alpha$. If $(P-k)$ is odd, we linearly interpolate between the values for $n=\frac{1}{2}(P-k)-\frac{1}{2}$ and $n=\frac{1}{2}(P-k)-\frac{3}{2}$ with $\frac{1}{2}\alpha$.

4. Empirical examples

These tests will be illustrated by applying them to two sets of real data. Both sets consist of survey data that have been aggregated to precinct-level sums. Consequently, we have individual-level responses and can use these to assess whether the results from our aggregate data model conform with the underlying unaggregated data. The first data set describes a telephone poll from the 1984 general election in California in which minority populations were oversampled. The survey includes 574 Latinos, 335 blacks, 308 Asians, and 317 non-Hispanic whites.

4.1. Example 1. Predicting education by race

Suppose that the goal is to estimate the percentages of college graduates among blacks and whites given only precinct-level data on education levels and racial composition. The model is

$$(\%\ \text{COLLEGE})_p=(1-\%\ \text{BLACK})_p\,\beta^W+(\%\ \text{BLACK})_p\,\beta^B+\epsilon_p.$$

The dependent variable is the proportion of people in the precinct who have at least some college education. The independent variable is the proportion of the precinct that is black, so the coefficient, β^{B}, is the percentage of college-educated blacks while β^{W} represents the percentage of college-educated whites. We first examine models that do not incorporate the latent groups. For instance, OLS imposes the constancy assumption, which amounts to assuming that the state electorate is a single homogeneous group. King’s model makes a slightly weaker assumption by positing that the parameters vary according to a truncated bivariate normal distribution. He does not impose the constancy assumption, but his random coefficient model incorporates an analogous and strong “similarity assumption” through the truncated bivariate normal distribution.

The estimates from these models for the percentage of blacks and whites who have at least some college education are displayed in Table 1. Neither of these methods is able accurately to capture the situation at hand. While both methods estimate the percentage of college-educated whites fairly well, neither is able to produce an accurate estimate of the percentage of college-educated blacks. Both significantly underestimate this quantity and both return very large standard errors that do not allow one to render substantively interesting inferences. Apparently, there is aggregation

[10] This table is found in J. Durbin’s *Biometrika* (Durbin, 1969) article, “Tests for serial correlation in regression analysis based on the periodogram of least-squares residuals”.

Table 1
Predicting percentage of college-educated individuals[a]

	College-educated blacks	College-educated whites
Truth	0.5343	0.5126
Models not accounting for latent groups		
OLS	0.2322	0.6042
	(0.2223)	(0.0584)
Random coefficients	0.3404	0.5747
	(0.3993)	(0.0845)
Latent group models		
Switching regression	0.5483	0.5250
Pure groups: Income	(0.2615)	(0.0523)
Switching regression	0.5359	0.5307
Hybrid groups: Income	(0.0893)	(0.0233)
Random coefficients	0.5060	0.5360
Income	(0.0551)	(0.0137)

[a] Source: 1984 general election data from California.
Standard errors in parentheses.

bias and both the constancy assumption as well as the similarity assumption are violated by these data.

Since latent groups apparently exist in the data, the next set of models attempts to account for the latent grouping. To do this, we must find the variables that underlie the grouping. Here, again, one first relies on beliefs and the theory underlying the research. Of course, there is no novelty at all in the advice to start with theory-one is rarely urged to select topics or variables randomly.[11] What is critical is a means to adjudicate between rival theoretically-justifiable specifications.

In the current example, income seems likely to divide the data since a reasonable theory that has been verified empirically is that neighborhoods are often defined by income levels, and neighborhoods with different income levels are likely to differ on educational attainment as well (e.g. Miller, 1977; Logan and Collver, 1983; Verba et al., 1995). A college education does not directly translate into a high income bracket and a nice neighborhood, since people who are college-educated often display varying abilities and/or desires to translate their education into affluence and comfort, but the hypothesis is worthy of exploring.

In order to test whether income is a reasonable covariate, we sort the precincts by increasing income levels. Next, we compute the cusum of the recursive residuals

[11] Surprisingly, in King’s lengthy book, he provides no guidance on how to move beyond “use your theory”, with the possible exception of a passage in which he recommends “walking around some of these neighborhoods, or standing by polling places, or reading the local press, or going to the supermarkets in the area” (King, 1997, p. 281). This recommendation is in the spirit of a “covariate test”, though it is clearly different from the proposed formal covariate test presented here.

as well as the cusum of the squares of the recursive residuals. These quantities are plotted against the newly-ordered precincts. Fig. 2 displays the plots of the cusum of recursive residuals and the cusum of the squares of the recursive residuals. Significance lines are plotted to indicate where deviations from the mean value line occur. In the first plot of the recursive residuals, a significance level of 0.01 yields the value 1.143 for the parameter *a*. The significance lines in the upper plot are thus two lines through the sets of points

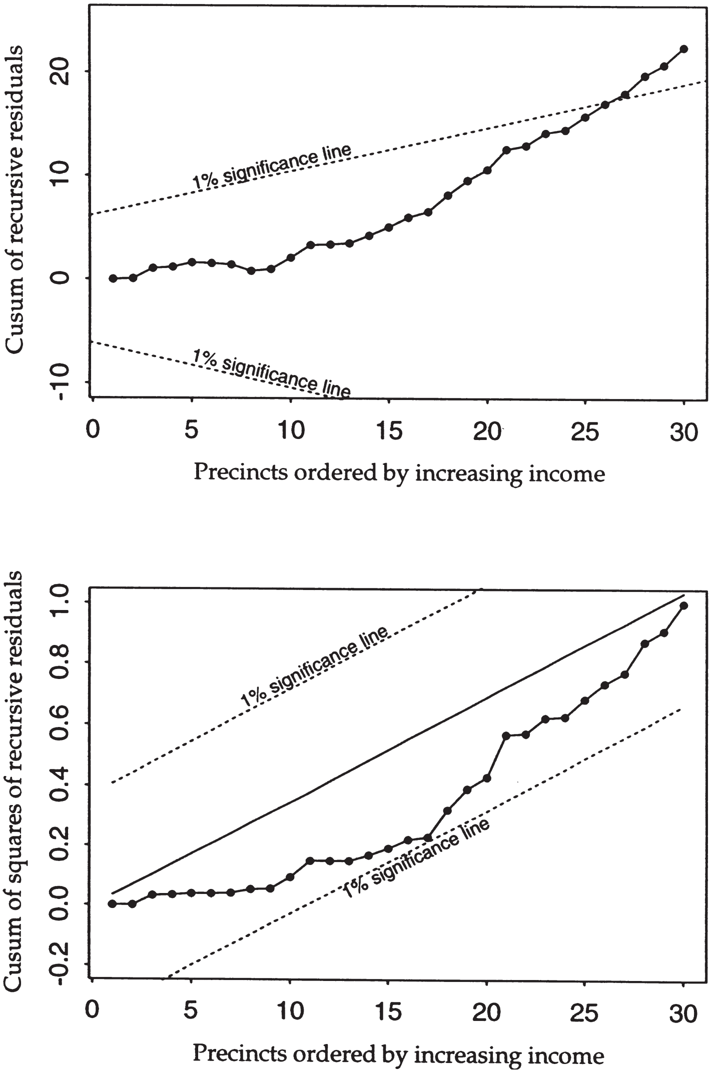

Fig. 2. Example 1. Income variable. Cusum plots, forward recursion. Observations are ordered in increasing value of the income variable.

$$(k,\pm a\sqrt{(P\text{–}k)}),\ (P,\pm 3a\sqrt{(P\text{–}k)})=(1,6.16),\ (30,18.47),\ (1,-6.16),\ (30,-18.47).$$

The set of parallel significance lines for the cusum of squares of recursive residuals is displayed in the lower plot of Fig. 2. The significance value for c_0 corresponding to a 0.01 significance level is obtained by finding the modified Kolmogorov–Smirnov value corresponding to $n = 13\frac{1}{2}$ and $\frac{1}{2}\alpha = 0.005$. In this case, we find $c_0 = 0.37176$. The lines are plotted accordingly.

In both the recursive residual plot as well as the squares of the recursive residuals plot, there is evidence of parameter instability. Hence, we should include the income variable in our likelihood function as one of the covariates, z, for the group-defining variable, D. If we wanted to adopt a pure group framework, we would note that the instability seems to manifest itself at the 17th observation point.[12] This alternative framework yields similar results. Moreover, the random coefficient model with the income covariate yields a similar answer as well.[13]

There is, thus, some evidence to validate the role of income in the aggregation process. Other hypotheses about different covariates are also viable. In particular, perhaps the age variable plays a key role in the aggregation process. Just as one might make a credible qualitative argument for why groups might be based on income level, so too might we make a credible qualitative argument for the age variable. Age separates time periods and defines generations. Baby Boomers, children of the 60s, and those who came of age during the Reagan era all bear distinctive profiles and tendencies. Further, the claim that American has become more educated over time is well-documented (Brody, 1978). The age hypothesis, thus, is believable on several fronts. To test the age hypothesis, we re-order the observations by increasing value of the age variable before we compute the recursive residuals as well as their squares. The plots of these quantities are displayed in Fig. 3. As we can see, in contrast to the income variable, no evidence of instability is realized. Neither sample path crosses the significance lines. Both hover near the mean value line. Hence, despite the fact that we can make qualitative arguments for including both the income and age variables, our quantitative assessments verify the inclusion of income as a group-defining variable but assert the exclusion of the age variable. Indeed, including age as a covariate produces poor results. The random coefficient model with the age

[12] While the recursive residual line does not cross the significance line until the 27th observations, the 17th observation point is where the line seems to begin its upward slope. This is consistent with the cusum of squares of recursive residuals line.

[13] One may notice that the results from the pure groups specification have considerably larger standard errors than the hybrid groups specification. It is difficult to determine the exact cause for this difference. One reason may be that the demarcation of pure groups may be too sharp for these data, i.e. the groups are not as well-defined as the specification that is being imposed.

In addition, in choosing between the two types of group specifications, one might consider that the hybrid group framework requires less user interaction. In the pure group framework, the researcher must still determine the point at which the different groups diverge. This can be a subjective process. Under the hybrid model, the researcher needs only to choose the covariates. The maximum likelihood function is not subjective.

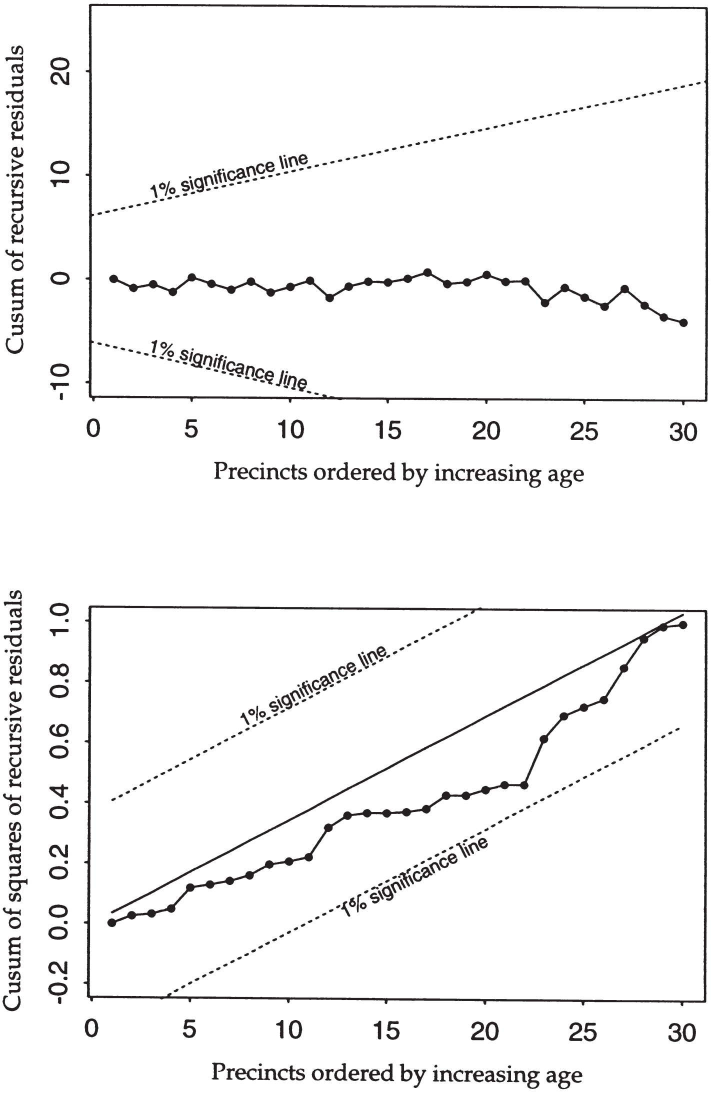

Fig. 3. Example 1. Age variable. Cusum plots, forward recursion. Observations are ordered in increasing value of the age variable.

covariate gives estimates that are far from their true values: 16.6% for blacks and 62.1% for whites.

Clearly, both OLS and random coefficient models provide reasonable estimates only when the specification is correct. Choosing income as a covariate was the critical decision, not whether one should adopt a switching regression or random coefficient model. The models adopting age as a covariate all produced poor estimates as did the models that adopted no covariates. There is, then, strong empirical evidence that the crucial decision concerns hypotheses regarding how individuals came to be aggregated into the observed precincts. The efficiency that the bounds provide in the random coefficients model is subsumed by the errors that result from a poor specification.

4.2. Example 2. Predicting proposition 209 vote by party

The data for a second example describe a poll conducted in California for the 1996 general election which included Proposition 209, the California Civil Rights Initiative. The poll included a total of 1500 respondents with an oversampling of minority groups. There were 262 respondents self-identifying as Asian American, 167 as black, 416 as Latino, and 427 as white. One of the goals of this survey was to examine contextual effects, specifically whether the effect from living in densely minority areas differed from the effect of living in non-majority minority areas. Hence, this poll includes identifiers for these contextual variables which aided in the proper aggregation of the voters back into their original precincts.

Suppose the goal is to assess support for Proposition 209 among partisans. The model is

$$\begin{aligned}&(\text{Pro PROPOSITION 209 VOTE})_p = (1-\%\ \text{DEMOCRAT})_p\,\beta^R\\ &\quad+(\%\ \text{DEMOCRAT})_p\,\beta^D+\epsilon_p\end{aligned}$$

where the dependent variable is the proportion of the vote in favor of Proposition 209, and the independent variable is the proportion of the precinct that is registered with the Democratic party. Since Proposition 209 was a vote on affirmative action, one conjecture that arises immediately is that districts that have higher proportions of minorities are likely to have differing levels of support for the Proposition than districts that are predominantly white. Also along this same line of reasoning, one might consider including gender as a covariate. Although Proposition 209 was specifically targeted at race and not gender biases, gender is often an issue in affirmative action legislation and thus one might reasonably suspect that its proponents would also be sympathetic here. As well, the support for Bob Dole as a presidential candidate may also be a dividing factor since Dole took a principled position in favor of Proposition 209. It is not difficult to develop some theoretical grounding for proposing that a certain set of covariates would define behavioral clustering and thus might be candidates for alleviating aggregation bias. Instead of simply adopting one's guesses without any type of empirical verification, it is a good idea to test whether these factors affected the overall support of Democrats for the Proposition. In general, as with all social science models, the role of theory and the ability to include substantive knowledge are important aspects of building a good model.

As we can see in Table 2, the results from models that do not incorporate latent grouping are not close to the truth, yielding estimates which are in excess of a standard error from the true values. The substantive interpretation of these numbers overstates Democratic support and understates the non-Democratic support. Appar-

Table 2
Predicting the Proposition 209 vote[a]

	Democrats	Non-Democrats		
Truth	0.3080	0.5460		
	Models not accounting for latent groups			
	OLS		Random coefficients	
	0.2015 (0.0495)	0.6958 (0.0758)	0.2199 (0.0395)	0.6792 (0.0630)
	Latent group models			
Covariates	Switching regression (Hybrid model)		Random coefficients	
Proposition 187	0.2887 (0.0453)	0.5427 (0.0716)	0.2856 (0.0618)	0.5618 (0.0985)
Proposition 187, % white	0.3152 (0.0446)	0.5024 (0.0707)	0.3034 (0.0809)	0.5333 (0.1291)
Proposition 187, % white, age	0.2360 (0.0483)	0.5676 (0.0719)	*Program failed to converge*	

[a] Source: 1996 California statewide survey.
Standard errors in parentheses.

ently, neither the constancy assumption nor the similarity assumption hold. The latent group models perform more admirably, demonstrating again the importance of determining which variables define the proper underlying grouping. A series of the cusum of squares of the recursive residuals plots for various orderings of the data are displayed in Fig. 4 and Fig. 5. As we can see from Fig. 4, the parameter values change as the first four variables, percent white, vote for Proposition 187, the presidential vote, and age, rise. Hence, we surmise that the level of support within the Democratic party varies between voters who reside in predominately minority communities and those who reside in primarily non-minority communities. Not surprisingly, the parameters do not remain constant as the balance of race grows increasingly disproportionate. A similar distinction can be made for precincts where the presidential vote or the vote on Proposition 187 was lopsided. On the other hand, Fig. 5 displays some variables which do not affect the stability of the system. In particular, gender and ideological identification do not inject instability into the system.

Hence, in parameterizing the group-defining variable, D, in our model, we should include the four variables, percent white, vote for Proposition 187, presidential vote, and age. Due to numerical complications, the maximization procedure for the switching regression model with all four covariates did not converge.[14] The other switching

[14] A lingering challenge is devising some method for choosing the optimal number of covariates. Cer-

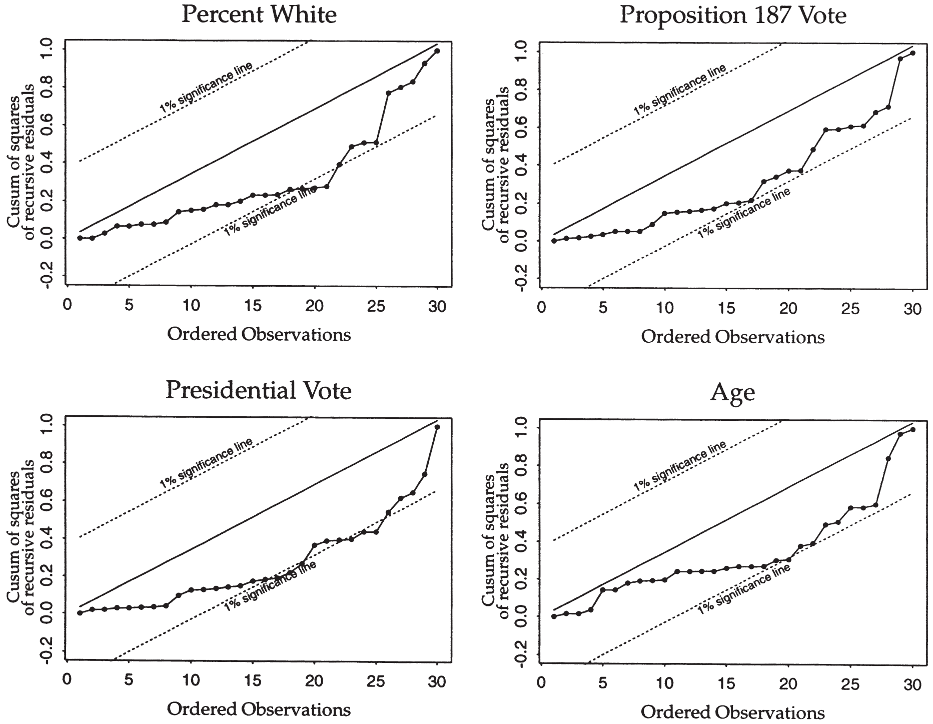

Fig. 4. Example 2. Cusum of Squares of Recursive Residual plots, forward recursion.

regression models did converge and yielded similar answers. An important note is that the random coefficient models yielded estimates that were similar and statistically equivalent to the estimates from the switching regression models. Hence, we can see that the crucial determinant is the choice of covariates and not whether one chooses to employ a random coefficient framework. The random coefficient model yields a small degree of efficiency in exchange for an extremely large degree of

tainly, in regression analysis, there are a number of measures which can be employed as criteria for subset selection. In this context, one might be able to employ the logic from general model building to the standardized logit formulation. In particular, in criteria-based subset selection, one might use criteria based on prediction errors. As a measure of fit, then, one might employ, for example, measures such as Mallows's C_p, Adjusted R^2, Akaike Information Criterion (AIC) or cross-validated predicted residuals. These methods may allow a researcher to determine which subset of covariates comprise the most explanatory power. These ideas need to be explored further, since the resulting fit from added covariates does not necessarily translate into a better model for the aggregate data. The relationship is much more complex and so these values may be misleading (Achen and Shively, 1995; Cho, 1998; Cho and Gaines, 2000). The reasoning behind these methods is certainly a promising start. If one were able to devise a method for choosing the optimal number and ideal subset of covariates in the context of aggregate data analysis, this method would obviously be very valuable.

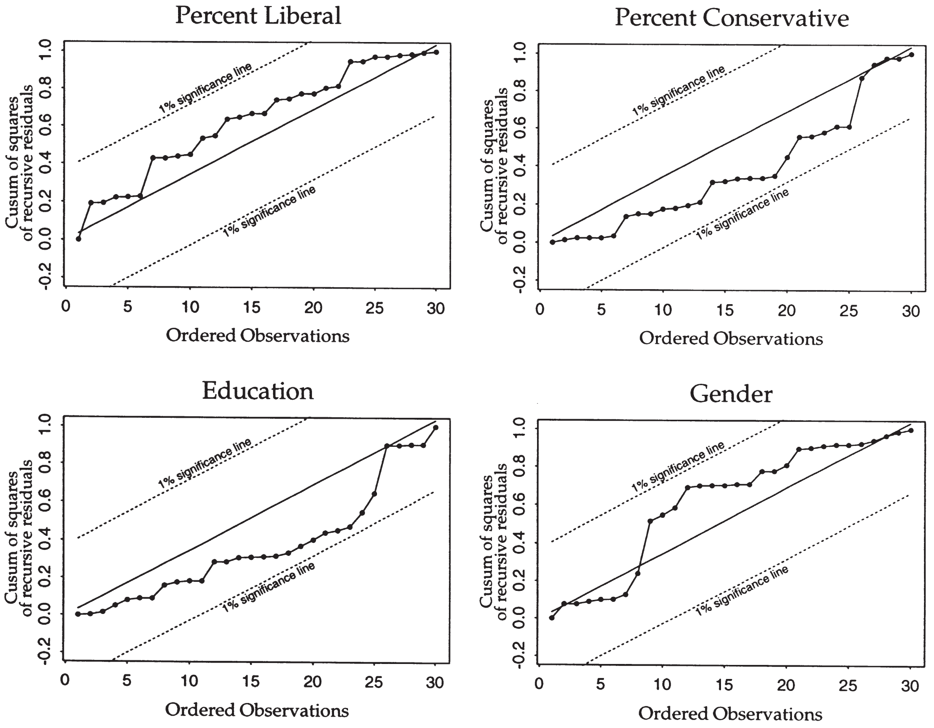

Fig. 5. Example 2. Cusum of Squares of Recursive Residual plots, forward recursion

unnecessary complexity. Moreover, because the switching regression framework is less computationally intensive and complex, it converges more often. This is clearly a virtue. Note that the algorithm for the random coefficient model did not converge with three covariates while the switching regression model did.

In addition, my model has provided us with additional insight into voting patterns and behavioral clustering. It has allowed us to estimate the support rates of Democrats for Proposition 209 while also providing us with insight into why there might be differing levels of support from within the Democratic party.

5. Conclusion

Ecological regression relies on the constancy assumption. If group behavior is not constant in the entire data set, ecological regression will produce unreliable parameter estimates. Random coefficient models without covariates will perform equally poorly. The possible exception is when the chosen distribution is such that the parameters are independent of the regressors. This would be uncommon and unknown ex ante. A useful insight is to note that the data are more reasonably believed to arise from some set of unobservable groups where behavior is similar within the group. In order to adopt this framework, a method for identifying the latent groups is necessary. A method has been proposed for testing hypotheses about latent groups in the aggregate data. While this method may not be foolproof, it provides a formal test for critical hypotheses about individual-level behavior that underlie aggregate data, and is it helpful in the context where practical considerations demand answers to an unsolvable problem.

Acknowledgements

Thanks to Chris Achen, Henry Brady, Bruce Cain, David Freedman, Brian Gaines, Doug Rivers, Julie Shaffer, Phil Spector and Stanley Wasserman for helpful comments. I would also like to thank seminar participants at the Institute of Governmental Studies at U.C. Berkeley, the Departments of Psychology and Statistics at the University of Illinois at Urbana-Champaign, and at the University of Rochester for lively and helpful discussion. I also gratefully acknowledge support from the National Science Foundation (SBR — 9806448). An earlier version of this work appeared as my Master’s Essay in the Statistics Department at U.C. Berkeley, Tam (1997).

References

Achen, C.H., Shively, W.P., 1995. Cross-Level Inference. University of Chicago Press, Chicago.

Andrews, D.W.K., 1993. Test for parameter instability and structural change with unknown change point. Econometrica 61 (4), 821–856.

Ansolabehere, S., Rivers, D., 1997. Bias in ecological regression estimates. (In preparation).

Berelson, B.R., Lazarsfeld, P.F., McPhee, W.N., 1954. Voting. University of Chicago Press, Chicago.

Brody, R.A., 1978. The puzzle of political participation in America. In: King, A. (Ed.), The New American Political System. American Economic Institute Press, Washington, DC, pp. 287–324.

Brown, R.L., Durbin, J., Evans, J.M., 1975. Techniques for testing the constancy of regression relationships over time. Journal of the Royal Statistical Society B37 (2), 149–192.

Cho, W.K.T., 1998. Iff the assumption fits...: a comment on the King ecological inference solution. Political Analysis 7, 143–163.

Cho, W.K.T., Gaines, B.J., 2000. Reassessing the study of split-ticket voting. Manuscript.

Durbin, J., 1969. Test for serial correlation in regression analysis based on the periodogram of least-squares residuals. Biometrika 56 (1), 1–15.

Ferreira, P.E., 1975. A Bayesian analysis of a switching regression model: known number of regimes. Journal of the American Statistical Association 70 (350), 370–374.

Freedman, D., Klein, S., Sacks, J., Smyth, C., Everett, C., 1991. Ecological regression and voting rights. Evaluation Review 15 (6), 673–711.

Freedman, D.A., Klein, S.P., Ostland, M., Roberts, M.R., 1998. Review of A Solution to the Ecological Inference Problem. Journal of the American Statistical Association 93 (444), 1518–1522.

Freedman, D.A., Ostland, M., Roberts, M.R., Klein, S.P., 1999. Response to King’s comments. Journal of the American Statistical Association 94 (445), 355–357.

Goldfeld, S.M., Quandt, R.E., 1973. The estimation of structural shifts by switching regressions. Annals of Economic and Social Measurement 2, 475–485.

Goodman, L.A., 1953. Ecological regressions and behavior of individuals. American Sociological Review 18 (6), 663–664.

Harvey, A.C., 1981. The Econometric Analysis of Time Series. Halsted Press, New York.

Huckfeldt, R.R., 1979. Political participation and the neighborhood social context. American Journal of Political Science 23 (3), 579–592.

King, G., 1997. A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data. Princeton University Press, Princeton.

Kmenta, J., 1986. Elements of Econometrics, second ed. Macmillan Press.

Logan, J.R., Collver, O.A., 1983. Residents' perceptions of suburban community differences. American Sociological Review 48 (3), 428–433.

Miller, W.L., 1977. Electoral Dynamics in Britain Since 1918. St Martin's Press, New York.

Ploberger, W., Krämer, W., Kontrus, K., 1989. A new test for structural stability in the linear regression model. Journal of Econometrics 40 (2), 307–318.

Putnam, R.D., 1966. Political attitudes and the local community. American Political Science Review 60 (3), 640–654.

Quandt, R.E., 1960. Tests of the hypothesis that a linear regression system obeys two separate regimes. Journal of the American Statistical Association 55 (290), 324–330.

Ritov, Y., 1990. Asymptotic efficient estimation of the change point with unknown distributions. The Annals of Statistics 18 (4), 1829–1839.

Robinson, W.S., 1950. Ecological correlations and the behavior of individuals. American Sociological Review 15 (3), 351–357.

Schulze, U., 1982. Estimation in segmented regression: known number of regimes. Mathematische Operationsforschung und Statistik, Series Statistics 13 (2), 295–316.

Tam, W.K., 1997. Structural Shifts and Deterministic Regime Switching in Aggregate Data Analysis. Master's Thesis. Department of Statistics. University of California at Berkeley.

Theil, H., 1971. Principles of Econometrics. John Wiley and Sons Inc, New York.

Verba, S., Schlozman, K.L., Brady, H.E., 1995. Voice and Equality: Civic Voluntarism in American Politics. Harvard University Press, Cambridge.