Advance Access publication March 22, 2011

Political Analysis (2011) 19:173–187
doi:10.1093/pan/mpr006

Reliable Inference in Highly Stratified Contingency Tables: Using Latent Class Models as Density Estimators

Drew A. Linzer
Department of Political Science, Emory University, 327 Tarbutton Hall, 1555 Dickey Drive,
Atlanta, GA 30322
e-mail: dlinzer@emory.edu

Contingency tables are among the most basic and useful techniques available for analyzing categorical data, but they produce highly imprecise estimates in small samples or for population subgroups that arise following repeated stratification. I demonstrate that preprocessing an observed set of categorical variables using a latent class model can greatly improve the quality of table-based inferences. As a density estimator, the latent class model closely approximates the underlying joint distribution of the variables of interest, which enables reliable estimation of conditional probabilities and marginal effects, even among subgroups containing fewer than 40 observations. Though here focused on applications to public opinion, the procedure has a wide range of potential uses. I illustrate the benefits of the latent class model–based approach for greatly improved accuracy in estimating and forecasting vote preferences within small demographic subgroups using survey data from the 2004 and 2008 U.S. presidential election campaigns.

1 Introduction

Contingency tables are among the most basic and widespread statistical tools available to analyze multivariate categorical data. When investigating patterns of association between two or more categorical variables, cross-tabulation provides the joint, marginal, and conditional distributions of variables of interest, describing the probabilities of various outcomes and the effects of changes in one or more variables on the probabilities of another. These table-based statistics are so easily constructed and easily understood that they pervade not only academic quantitative research but also commercial and journalistic data analysis as well.

In practice, contingency table analysis is limited by the fairly restrictive constraint that cross-tabulating by additional variables rapidly increases the number of cells in the table, which, for fixed, finite samples, greatly reduces the frequency of observations per cell—ultimately, to zero. The greater the amount of stratification (i.e., variables in the contingency table), the more sparsely distributed the data will be across cells in the observed cross-tab, and the greater the sample-to-sample variability in estimates of cell percentages and conditional effects. This makes it nearly impossible to produce reliable inferences from sample to population, especially in moderate- to small-sized samples. Researchers are typically forced to limit stratification to at most three or four categorical variables, even when much of theoretical value could be learned from additional stratification—had only the sample size been larger.

This scenario is often encountered in the analysis of public opinion survey data, which are nearly always categorical (Asher 2007, 194), and which motivate the examples in this paper. As Heeringa, West, and Berglund (2010, 113) observe, ‘‘experience has shown that many survey analysts often ‘push the limits’ of survey design, focusing on rare or highly concentrated subclasses of the population.’’ What motivates this practice is an expectation of finding variation in the attitudes and behaviors of individuals, even within broader social or demographic subgroups: Many of the most interesting and useful intragroup differences only appear following repeated stratification. Campaign strategists, for example, study cross-tabular reports to help identify narrow segments of ‘‘swing voters’’ to target with persuasive appeals (e.g., Cillizza 2007; Jamieson 2009). Political reporters rely heavily on the descriptive analysis of contingency

Author's note: I am grateful to Cliff Carrubba, Tom Clark, Justin Esarey, and Stephen Jessee, as well as the editors and reviewers at *Political Analysis*, for helpful comments on earlier versions of this article. The data I use in this paper are all public, but I have archived my replication code at the PA Dataverse, study ID hdl:1902.1/15626.

tables to uncover patterns of mass support for policies and politicians (e.g., Balz and Johnson 2009; Todd and Gawiser 2009). Social scientists commonly use multivariate survey data to test hypotheses about critical subpopulations as in the studies by Leal et al. (2005), de la Garza and Cortina (2007), and Abrajano, Alvarez, and Nagler (2008) of Hispanic voters in the 2004 U.S. presidential election—a group comprising less than 10% of the national electorate (Taylor and Fry 2007). In each case, although contingency table analysis can be used to reveal multivariate relationships and investigate confounding, the data eventually ''run out.'' The generalized linear and hierarchical regression models that many social scientists employ to get around this obstacle may not be necessary or appropriate for applied practitioners—and may, depending upon the nature of the data and the hypothesis being investigated, be ill-advised or unhelpful in much academic research as well (Achen 2002, 2005). Still, the current practice of not reporting table-based statistics when cell sizes are small, due to the high degree of estimation uncertainty involved, is clearly suboptimal as long as the data set contains other information that might be exploited to reduce the variance of these estimates.

In this paper, I introduce a simple but reliable tool to address this basic, but important, problem. The solution is based on a statistical technique for density estimation in cross-classification tables known as latent class analysis. The latent class model is a finite mixture model most commonly used to identify clusters of similar observations in multivariate categorical data. I demonstrate that fitting a latent class model to an observed multiway contingency table prior to analysis results in an estimate of the underlying probability mass function that is both more stable, and closer, on average, to the true underlying distribution than the ''raw'' observed cell percentages. From these model-based cell percentage estimates, it is straightforward to produce estimates of both conditional probabilities and marginal effects. To illustrate how dramatically (and how easily) the method can improve the reliability of table-based inferences, I apply it to the problem of estimating features of—and forecasting vote preferences within—small demographic subgroups, using survey data collected from preelection and exit polls surrounding the 2004 and 2008 U.S. presidential elections.

The benefit to applied researchers is the possibility of obtaining precise estimates of characteristics of small subgroups—population prevalence, conditional probabilities, and marginal effects—even when repeated stratification leaves fewer than 40 observations in the subgroup of interest. Because latent class analysis is based on a parametric model of individuals' multivariate responses, it is particularly well suited for contingency tables of high dimension, containing a large proportion of empty cells (or ''sampling zeros'') for outcome variables that are dichotomous or polytomous and nominal or ordinal. The issue of how to improve small cell estimates in sparse contingency tables can also be handled using nonparametric kernel-based methods, but selection of the smoothing parameter becomes a crucial consideration, especially as tables increase in size (Aitchison and Aitken 1976; Titterington 1980; Hall 1981; Grund 1993). An extensive literature on Bayesian approaches to the estimation of multinomial cell probabilities offers another potential solution (see Agresti and Hitchcock 2005; Congdon 2005), although such techniques still require careful attention to the choice of prior distributions and model specification to control the amount of smoothing applied to each cell.

In contrast to these alternative approaches, the latent class model–based method has the further advantage of being relatively uncomplicated to implement and interpret. Software to estimate the latent class model, and to perform the necessary postestimation calculations, is freely available as part of poLCA, a package for polytomous variable latent class analysis implemented in the R statistical computing environment (Linzer and Lewis 2010; R Development Core Team 2010).

2 Latent Class Models as Density Estimators

Density estimation refers to a broad class of statistical procedures used for empirically approximating the distribution of variables in a population, given a sample drawn from that population. Although the underlying distribution is unobserved, the idea is that it can be recovered based upon observed patterns in the data. Finite mixture models are a parametric approach to density estimation that assumes that the unknown generating distribution for the observed data can be approximated as a weighted sum of a finite number of component distributions for which the functional form *is* known (McLachlan and Peel 2000; Fraley and Raftery 2002). The choice of component distributions is left to the researcher but typically depends on known features of the observed data: are they continuous or discrete, univariate or multivariate, and so forth. Once this decision has been made, estimation of the model consists of estimating all the parameters

https://doi.org/10.1093/pan/mpr006 Published online by Cambridge University Press

of the component distributions, as well as a vector of "mixing" proportions—summing to one—that represent the weights assigned to each component.

Latent class models are a type of finite mixture model appropriate for producing density estimates of the joint distribution of two or more categorical variables.[1] Such distributions are commonly expressed as multiway cross-classification tables over the observed (or "manifest") variables of interest.[2] The latent class model fits the observed contingency table by "mixing" together component distributions that are themselves cross-classification tables of identical dimension to the observed table; but within each component table, all variables are assumed to be statistically independent. This assumption is referred to as "conditional" or "local" independence. Estimating this model is equivalent to assuming that any confounding among the manifest variables can be explained by stratifying the observed multiway table by an unobserved (latent) nominal categorical variable. The inferred value of this category for each observation is its latent "class."

An intuitive way to conceptualize the latent class model is to think of a population as being comprised of a finite number of "types" of individuals. Within each type—or class—individuals produce responses to the manifest variables in a consistent manner. In political terms, "conservatives" might respond to survey questions in one way, whereas "liberals" all respond differently. The classification of types, however, is not directly observable—and may not even be this clear-cut. The ability of the latent class model is to identify and separate out clusters of similar individuals based upon the observed pattern of responses to a series of categorical variables. The latent groupings are characterized by the probabilities with which individuals of each type provide each of the possible responses; and because the model is probabilistic, it accounts for a certain amount of sampling variability. The latent class model then "reassembles" the classes to produce an overall summary—the density estimate—of the joint distribution of the manifest variables in the population.

2.1 *Specification of the Latent Class Model*

Assume that individuals $i = 1, \ldots, N$ produce a series of responses on manifest variables indexed $j = 1, \ldots, J$, each of which contains a finite number of outcomes, $K_{j.}$ These variables form a J-way contingency table containing a total of $C = \prod_{j=1}^{J} K_j$ cells. Denote as Y_{ijk} the observed data, such that $Y_{ijk} = 1$ if individual i produces the kth outcome on the jth variable, and $Y_{ijk} = 0$ otherwise. Let R represent the number of latent classes in the model, which is fixed by the analyst prior to estimation. Then $r = 1, \ldots, R$ indexes the component cross-classification tables, with p_r representing the R mixing proportions such that $\sum_r p_r = 1$. The parameters estimated by the model are the mixing proportions p_r, and the conditional probabilities that an individual in class r produces the kth outcome on the jth manifest variable, denoted $\pi_{jrk.}$

Following the assumption of local independence, the probability of individual i producing its particular set of J responses on the manifest variables, assuming it belongs to class r, is

$$f(Y_i; \pi_r) = \prod_{j=1}^{J} \prod_{k=1}^{K_j} (\pi_{jrk})^{Y_{ijk}}, \tag{1}$$

the product of the respective class-conditional marginal percentages. The probability mass function across all R latent classes is then

$$\Pr(Y_i; \pi, p) = \sum_{r=1}^{R} p_r \prod_{j=1}^{J} \prod_{k=1}^{K_j} (\pi_{jrk})^{Y_{ijk}}. \tag{2}$$

[1] The technique of latent class analysis was first set forth by Lazarsfeld (1950) under the name latent structure analysis and expanded upon by Goodman (1974a, 1974b), among others. Hagenaars and McCutcheon (2002) provide a broad and useful overview of recent advances in latent class modeling. Also see Bartholomew et al. (2008, Chapter 10).

[2] Latent class models can be modified to accommodate manifest and latent variables that are either ordered or unordered, but it is sufficient here to use the simple classical latent class model that treats both types of variables as unordered.

https://doi.org/10.1093/pan/mpr006 Published online by Cambridge University Press

The parameters of the latent class model may be estimated by maximizing the log-likelihood function

$$\ln L = \sum_{i=1}^{N} \ln \sum_{r=1}^{R} p_r \prod_{j=1}^{J} \prod_{k=1}^{K_j} (\pi_{jrk})^{Y_{ijk}} \tag{3}$$

with respect to p_r and $\pi_{jrk.}$ Bayesian estimation of the latent class model is also feasible, given properly specified prior distributions, as in Garrett and Zeger (2000).

Estimating the latent class model with a sufficient number of latent classes produces a fully parameterized probability mass function that closely approximates the unobserved underlying joint distribution of the variables of interest in the population. The question of how many latent classes are "sufficient" is of some importance, as this determines the amount of smoothing that is applied to the observed contingency table. Setting $R = 1$ is equivalent to assuming that all J variables are statistically independent; larger values of R produce fits closer to the observed data and hence greater sample-to-sample variability. Selection of an "optimal" number of latent classes to identify clusters present in the data set has been considered by Bandeen-Roche et al. (1997), Garrett and Zeger (2000), Huang (2005), and Nylund, Asparouhov, and Muthén (2007), among others. For the purposes of density estimation, however, slightly underfitted models have the benefit of reducing the variance of cell percentage estimates. The evidence that I present below suggests that $R = 2$ is an appropriate and useful choice.

The expected percentage of the population in each cell of the fitted J-dimensional table is easily calculated by inserting estimates $\hat{p}_r$ and $\hat{\pi}_{jrk}$ into equation (2). Denote as y_c the sequence of J outcomes corresponding to the cth cell in the fitted contingency table, such that $y_{cjk} = 1$ if cell c contains the kth response on the jth variable, and $y_{cjk} = 0$ otherwise. Then, the estimated probability mass function produced by the latent class model is

$$\tilde{P}(y_c) = \sum_{r=1}^{R} \hat{p}_r \prod_{j=1}^{J} \prod_{k=1}^{K_j} (\hat{\pi}_{jrk})^{y_{cjk}} \tag{4}$$

for cells $c = 1, \ldots, C$. The model-based estimate $\tilde{P}(y_c)$ of the population percentage $P(y_c)$ in each cell is the weighted sum (by $\hat{p}_r$) of the products of a cell's estimated marginal probabilities in each component table.

By comparison, the maximum likelihood estimate (MLE), $\hat{P}(y_c)$, of the population cell percentage is equal to the observed number of cases in cell c divided by the total number of observations, N. For high-dimensional contingency tables—especially with many outcomes per manifest variable and small-to-moderate sample sizes—a nonnegligible proportion of cells will contain zero observations. For those cells, $\hat{P}(y_c) = 0$, even though it is unlikely that in the underlying population, the "true" percentage of cases with the set of characteristics y_c is ever exactly zero. Agresti and Hitchcock (2005, 298) remark simply that "for a cell with a sampling zero, 0.0 is usually an unappealing estimate." The cell percentage estimate $\tilde{P}(y_c)$ based on the latent class model, however, will always be greater than zero (if only slightly).

As a density estimator, the latent class model is effectively "filling in" the (problematic) zero cells in the sample cross-classification table, whereas at the same time "smoothing out" some of the sampling variability in both the zero cells and the nonzero cells. The "meanings" of the latent classes, which in applications focusing on clustering or scaling would be revealed by interpreting the estimated $\hat{\pi}_{jrk}$ parameters, are of no special importance when using latent class models for density estimation; a point emphasized by Vermunt et al. (2008, 377–378).

2.2 *Demonstration of Density Estimation*

To illustrate how the latent class model bridges the gap between an observed sample distribution and the underlying (unobserved) population distribution, I reanalyze exit poll data collected following the 2004 U.S. presidential election by the National Election Pool, Edison Media Research, and Mitofsky International (2004). Because the sample was so large—a total of 13,719 voters were interviewed—I treat the observed cell percentages as the "known" population percentages. Tabulating responses to six questions measuring respondents' race, sex, income, age, marital status, and political ideology produces a six-way

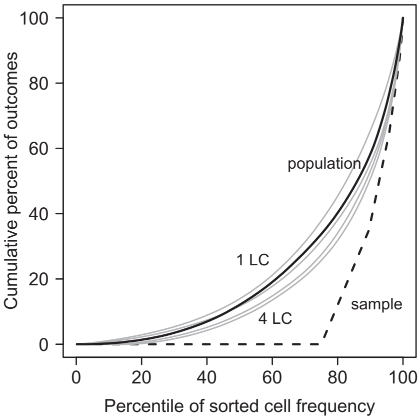


Fig. 1 Cumulative distribution of cell percentages in a 6-variable, 480-cell cross-classification table, using data from the 2004 U.S. presidential election national exit poll. The solid black curve represents the population distribution $P(y_c)$, whereas the dashed line represents cell percentages $\hat{P}(y_c)$ obtained from a random sample of 200 respondents. Gray lines reflect latent class model–based estimates $\tilde{P}(y_c)$ assuming $R = 1$ through $R = 4$ latent classes: Fewer latent classes produce greater smoothing, whereas more latent classes produce a fit closer to the observed data.

table with 480 cells corresponding to each potential six-response sequence.[3] Demographic categories such as these are regularly used by political researchers to yield information of interest about public opinion and voting behavior.

I now draw a random subsample of just 200 of the original 13,719 respondents and tabulate the same six variables. In the subsample shown in Fig. 1, 75% of the cells contain zero observations, and another 15% have just one observation. The solid black curve represents the cumulative distribution function of the population cell percentages, sorted by relative frequency from lowest to highest. The dashed line represents the cumulative distribution function of MLEs $\hat{P}(y_c)$ for the sample of 200. The population and sample distributions are clearly highly dissimilar.

Yet, there remains enough structure even in the small 200-person sample to estimate a latent class model that recovers something close to the original ‘‘population’’ distribution. Fitting a latent class model to the same sample of 200 observations, and calculating model-based cell percentages $\tilde{P}(y_c)$, vastly improves the estimate of the underlying population distribution—regardless of the choice of number of latent classes.[4] The gray lines in Fig. 1 represent the cumulative distribution of estimated cell frequencies based on models assuming one through four latent classes. A latent class model that assumes too few components R will result in an underfitted model with the gray line slightly above the solid line, whereas a model with too great an R will be overfitted and fall somewhat further beneath the solid line. But in all cases, the latent class model–based estimates are much closer to the population distribution than are the MLEs.

As an additional illustration, I plot the observed cell percentages in the 200-person sample against their corresponding known population cell percentages (Fig. 2). The horizontal ‘‘stripes’’ in the left-hand plot correspond to sample cells with zero through six observations; the sample is not large enough to estimate the cell percentages with any greater precision. In contrast, the model-based estimates of the cell percentages, following fitting by a two-class latent class model, are much more consistently close to the true values. This is the key insight from which the improvements to various methods of contingency table analysis discussed in this article all follow.

[3] Outcome categories for each variable are coded as follows. Race: white and nonwhite. Sex: male and female. Yearly household income: $0–30,000, $30–50,000, $50–75,000, $75–100,000, and >$100,000. Age in years: 18–29, 30–44, 45–64, and older than 65. Marital status: married and unmarried. Ideology: conservative, liberal, and moderate.

[4] All latent class models used in this paper are estimated by maximum likelihood using the R package poLCA (Linzer and Lewis 2010; R Development Core Team 2010). The poLCA package can automatically compute the model-based cell percentages $\tilde{P}(y_c)$ following estimation of a latent class model.

https://doi.org/10.1093/pan/mpr006 Published online by Cambridge University Press

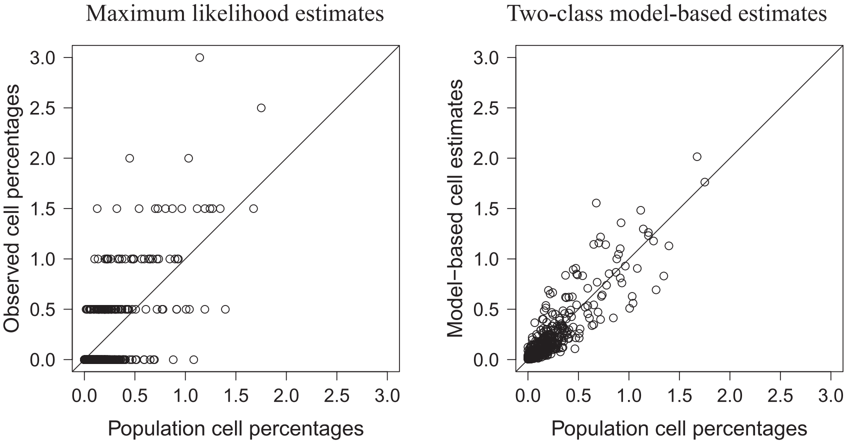


Fig. 2 Comparison of small-sample maximum likelihood cell percentage estimates (left) and two-class latent class model–based cell percentage estimates (right) to population cell percentages. Each point corresponds to one cell in the 480-cell table. The model-based estimates tend to be much closer to the true values as evidenced by their greater proximity to the 45° line.

2.3 *From Cell Percentages to Conditional Probabilities*

In many studies, the quantity of interest is a conditional probability rather than the unconditional cell percentage $P(y_c)$. For example, researchers may wish to estimate the percentage of white males with incomes over $100,000 who are also political conservatives. But this is just a function of two cell percentages: the population prevalence of individuals who are white, male, wealthy, *and* conservative, divided by the percentage of the population that is white, male, and wealthy. The conditional probability is easily calculated:

$$\Pr(\text{conservative} \mid \text{white, male, wealthy}) = \frac{\Pr(\text{white, male, wealthy, conservative})}{\Pr(\text{white, male, wealthy})}.$$

The term in the denominator, Pr(white, male, wealthy), is equal to the sum Pr(white, male, wealthy, liberal) + Pr(white, male, wealthy, moderate) + Pr(white, male, wealthy, conservative). Any conditional probability of interest can be produced in this manner, given an estimate—either the MLE, $\hat{P}(y_c)$, or the model-based $\tilde{P}(y_c)$—of the joint distribution of the selected categorical variables.

A variety of other model-based methods also exist to estimate conditional probabilities within subgroups, without estimating the corresponding cell percentages. One approach is to fit a generalized linear model (GLM) to the observed data, using an appropriate link function to model the (categorical) dependent variable as a linear combination of a set of predictor variables as well as some or all the higher-order interactions between those variables at the researcher's discretion (Maddala 1983; Long 1997). The conditional probabilities of interest may be calculated from the coefficients estimated by the model for specified values of the covariates.

In applications where observations are clustered into higher-level units, more sophisticated generalized linear mixed models or multilevel models (MLMs) may also be applied (Agresti et al. 2000; Agresti 2002; Gelman and Hill 2007). These models are particularly useful when each unit contains a small number of observations, akin to the case of small cell sizes in high-dimension contingency tables. By assuming cluster random effects, MLMs allow the within-unit conditional probability estimates to ''shrink'' toward the grand mean, improving the resulting estimates by reducing their sampling variability. This is just another form of smoothing in which more pooling is applied to estimates from units containing fewer observations (Gelman and Hill 2007, 258). The technique is closely related to statistical approaches to small area estimation (Jackson 1989; Ghosh and Rao 1994; Rao 2003). In similar fashion, the aim is to use partial pooling to improve estimates of the proportion of individuals in a population who possess some property of interest, conditional upon a combination of (typically) geographic, and demographic factors. Although neither the GLM nor the MLM produce estimates of $P(y_c)$, recent advances in multilevel regression with

https://doi.org/10.1093/pan/mpr006 Published online by Cambridge University Press

poststratification (Park, Gelman, and Bafumi 2004) have been shown to produce accurate estimates of public opinion at the state level (Lax and Phillips 2009). More generally, this technique is best applied when small cell sizes arise not from repeated stratification, but rather from manifest variables containing large numbers of outcome categories. In such cases, MLM-based methods may be preferred to the approach described in this paper. Of course, the structure of the observed data is not always conducive to a random-effects specification.

The major drawback to the regression-based approach is that the conditional probability estimates are highly sensitive to the choice of model specification. Again consider the problem of estimating the percentage of white males with incomes over $100,000 who are also political conservatives. A fully specified GLM would include as covariates indicator variables for race, sex, and each category of income, as well as all the second- and third-order interactions. Because the independent variables are neither continuous nor interval, little is gained by placing them in a regression framework; as I demonstrate below, models that include all the higher-order interactions produce estimates that are nearly identical to those calculated directly from $\hat{P}(y_c)$. Yet, omitting particular higher-order interactions implicitly assumes that those effect modifiers are zero. As discussed by Berry, DeMeritt, and Esarey (2010), decisions by applied researchers about which interaction terms to include as covariates in a GLM are most often made idiosyncratically and on an incorrect basis. The latent class model makes all these model specification issues moot by automatically capturing the full set of interrelationships between the variables of interest in the density estimate.

3 Small-Cell Probabilities and Conditional Effects

Estimating characteristics of population subgroups using tabular data becomes extremely imprecise in small samples or following repeated stratification. Even if an overall sample is very large, in a subgroup with 100 observations, the theoretical margin of error for an estimated proportion will be as much as ±10% at a 95% level of confidence. For a subgroup with 50 observations, the maximum margin of error jumps to nearly ±15%. With 20 observations, the maximum margin of error is well over ±20%, rendering any inferences from the sample to the population essentially uninformative. It is for this reason that tabular analyses rarely stratify beyond three or four categorical variables at once before running out of cases upon which to reliably base parameter estimates.

Populations corresponding to groups with so few observed cases can be quite large and substantively important. In a typically sized public opinion survey of approximately 1000 respondents, 50 observations represents 5% of the population; in the United States (as of 2010), that translates to groups comprising over 15 *million* individuals—larger than the states of Pennsylvania or Illinois. Or consider the demographic category of black males aged 18–34; a group of nearly 5 million individuals but only 1.5% of the national population (U.S. Census Bureau 2008). In a sample of 1000, this works out to just 15 individuals on *average*; of course, in any particular sample, the number of young black males interviewed will vary and could be potentially much less than 15. Sampling variability alone prevents researchers from obtaining meaningful estimates of the characteristics of any groups this size or smaller. Estimating the conditional effect of sex (or any other independent variable) on a chosen dependent variable for blacks aged 18–34 will be even more imprecise because it also requires estimating the percentage of interest among black *females* aged 18–34—another small subgroup.

The obvious and best solution to this problem would be to collect larger samples. In most research situations, however, this option is prohibitively expensive, time consuming, or even impossible once a study has been completed. Instead, I show that preprocessing an observed cross-classification table by fitting a latent class model can be an efficient and reliable approach to estimating small-cell probabilities and conditional effects. As a rule of thumb, the method outperforms the MLE when repeated stratification of categorical data produces cell sizes of 40 observations or fewer. Simulation evidence indicates that the method may continue to outperform the MLE in even larger cell sizes, when estimating characteristics of subgroups that comprise a larger share of the population or when estimating marginal effects across multiple small subgroups.

3.1 *Latent Class Model-Based Estimation*

The MLE $\hat{P}(y_c)$ is an unbiased estimate of $P(y_c)$, but it is subject to a large amount of sampling variation when the number of observations is small. Applying a latent class model to the observed table, and using the model-based $\tilde{P}(y_c)$ from equation (4) as an estimate of $P(y_c)$, produces a multivariate density estimate

https://doi.org/10.1093/pan/mpr006 Published online by Cambridge University Press

of the underlying population distribution that is highly consistent from sample-to-sample. The result is a much more precise estimate of $P(y_c)$, which leads directly to more reliable estimates of conditional probabilities and conditional effects when looking at the relationships between two or more variables.

The latent class model–based technique has a range of practical advantages as well. Unlike the MLE, the model-based technique can estimate conditional probabilities even when the number of individuals sampled from the subgroup of interest is zero. In addition, the latent class model frees the researcher from the large number of often arbitrary (and potentially erroneous) modeling assumptions that go into the specification of a GLM or related model. Nor does the latent class model require the data to possess a hierarchical structure, as in an MLM. The result is that the latent class model is easier to implement and interpret, less technically demanding, and far less time consuming to the researcher than other model-based estimators.

3.2 *Improving Inference by Reducing Sampling Variability*

To demonstrate each of these points, I again treat the complete 2004 U.S. presidential election exit poll as the known population and computationally simulate smaller random samples drawn from that population. The quantity of interest will be the conditional probability that white *and* nonwhite male voters aged 18–29 voted for John Kerry for president. These subgroups are chosen because they represent a suitably small proportion of the overall sample: White males aged 18–29 are just 6% of the original set of interviewees, whereas similarly aged nonwhite males make up less than 2.5%.

For repeated samples of sizes varying from 50 to 1000 individuals, I calculate the cell percentages $\hat{P}(y_c)$ of white and nonwhite, young, male, Kerry voters; that is, the MLE based on a four-way cross-tabulation of the variables vote choice, race, sex, and age. I then fit a series of latent class models assuming one through four latent classes to the vote choice, race, sex, and age variables in each simulated sample and produce model-based estimates $\tilde{P}(y_c)$ of the same quantities using equation (4). In this manner, no other data are needed beyond what was used to produce the MLEs. From both sets of cell percentage estimates, I calculate the conditional probabilities that voters in each demographic group voted for Kerry:

$$\Pr(\text{Kerry}\,|\,\text{white},\ \text{male},\ 18-29) = \frac{\Pr(\text{white},\ \text{male},\ 18{-}29,\ \text{Kerry})}{\Pr(\text{white},\ \text{male},\ 18{-}29)}$$

and

$$\Pr(\text{Kerry}\,|\,\text{nonwhite},\ \text{male},\ 18-29) = \frac{\Pr(\text{nonwhite},\ \text{male},\ 18{-}29,\ \text{Kerry})}{\Pr(\text{nonwhite},\ \text{male},\ 18{-}29)}.$$

I use these quantities to further estimate the marginal effect of race on voting for Kerry for males aged 18–29: Pr(Kerry | white, male, 18–29) − Pr(Kerry | nonwhite, male, 18–29).

Finally, I apply a multinomial logistic regression model to the sample data, with vote choice as the dependent variable (including three categories for Bush, Kerry, and other), and race, sex, and age—recoded as indicator variables—as the independent variables along with each of their higher-order interaction terms.[5] Of the 13,660 exit poll respondents stating their vote choice, 182, or 1.3%, voted for a candidate other than Bush or Kerry. Retaining this small but not insignificant segment of voters tests the modeling assumptions of both the latent class model and the GLM and permits estimation of Kerry's actual vote share rather than just his share of the two-party vote. From the resulting coefficient estimates, I again calculate the predicted percentage of white and nonwhite young males who voted for Kerry.

Each of these estimates are compared with the "true" percentage of voters of each type who voted for Kerry according to the full sample: 46.3% of white males aged 18–29 and 71.5% of nonwhite males aged

[5] The variables sex and race each have two outcome categories, and the age variable has four categories. Expressing these independent variables as indicators, and omitting one outcome category for each, produces five first-order terms, seven second-order terms (all two-way interactions between sex, race, and age), and three third-order terms (all three-way interactions)—for a total of 15 terms on the right-hand side (or 16, including the constant).

https://doi.org/10.1093/pan/mpr006 Published online by Cambridge University Press

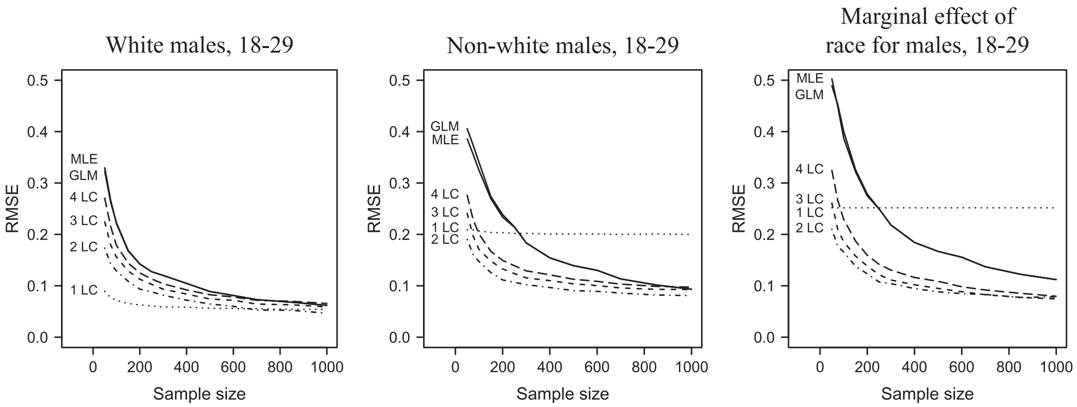


Fig. 3 RMSE of maximum likelihood, GLM, and latent class model–based estimates of the conditional probability that young white and nonwhite males voted for Kerry for president as well as the marginal effect of race on vote choice. Results are based on 2000 simulated samples of size 50–1000.

18–29. The ‘‘true’’ marginal effect of race on vote choice for young males is thus approximately 25 percentage points. Based on these values, I calculate the root mean squared error (RMSE) of each estimator over repeated simulation of subsamples of varying sizes drawn from the original large data set. Smaller values of the RMSE indicate that the estimator is producing parameter estimates that are closer, on average, to the true value in the population.

The reduction in RMSE due to smoothing by the two or more class latent class model is most pronounced in small samples where the variance of the observed cell percentages is greatest (Fig. 3). Even in samples of 200, the average number of young white males is only approximately 12; and young nonwhite males are fewer than half that number. Using the model-based density estimate in place of the estimate taken directly from the observed contingency table can reduce the RMSE of estimates of conditional probabilities by *one-half or more* in samples with fewer than 200 observations. The model-based estimator particularly excels when estimating marginal effects—in this case, the effect of race on vote choice for males aged 18–29—because this depends on prior estimation of two conditional probabilities, both of which are subject to their own sampling error.

Fitting a basic model with just two latent classes produces the best estimates of conditional probabilities and marginal effects in both small and moderate sample sizes. Models with more than two latent classes fit the observed tables more closely; as a result, there is less smoothing, greater sample-to-sample variation in the density estimate and an RMSE closer to that of the observed cell percentages. For these reasons, it is recommended to estimate the two-component latent class model when preprocessing the observed data.

As a conservative rule of thumb, the latent class model–based estimates of conditional probability are superior to the MLE in subgroups containing 40 observations or fewer. They may outperform the MLE in larger subgroups as well. Figure 3 demonstrates that in samples up to size 1000, the model-based estimates have a smaller RMSE than the MLE. Beyond 1000 observations, the estimators are effectively equivalent. As sample size increases, the RMSE of the MLE eventually falls below that of the model-based estimates. This is because, as discussed in Section 3.3, the model-based estimates sacrifice a small amount of unbiasedness in exchange for a considerable reduction in the variance of the estimator. Additional simulations (not shown) indicate that for estimating the probability that nonwhite males aged 18–29 voted for Kerry, the MLE does not overtake the two-class estimator until $N = 1500$. Recall that this demographic subgroup represents just 2.5% of the U.S. population. Multiplying 0.025 by 1500 observations gives us the (approximate) 40-observation benchmark. The reason this is a conservative rule of thumb is because in subgroups that make up a larger share of the population—for example, among white males aged 18–29—the model-based estimator may continue to have a lower RMSE than the MLE in even larger samples.

The one-class latent class model should not be used, though it is instructive to note the superiority of this estimator in the very smallest samples. This model assumes that the conditional cell percentages are equal to the marginal cell percentages—that is, in this example,

$$\Pr(\text{Kerry}) = \Pr(\text{Kerry} \mid \text{white},\ \ \text{male},\ \ 18-29) = \Pr(\text{Kerry} \mid \text{nonwhite},\ \ \text{male},\ \ 18-29),$$

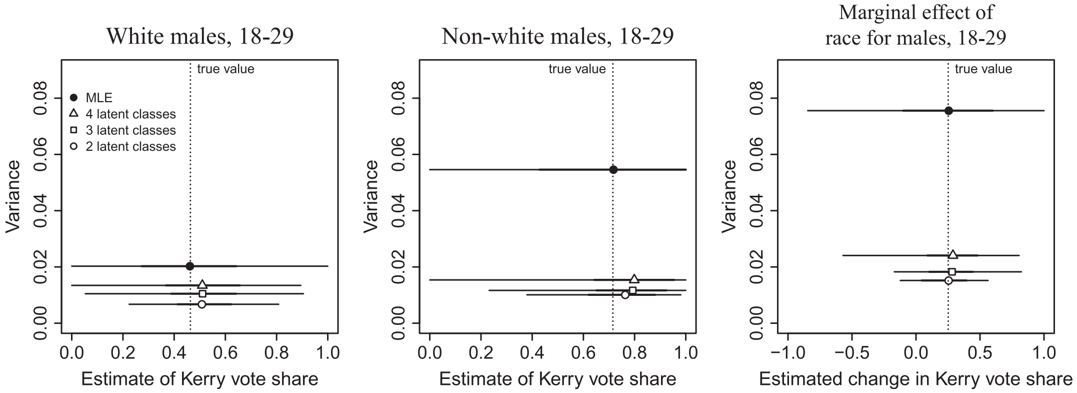


Fig. 4 Model-based estimates introduce a small amount of bias to achieve a large reduction in variance. Points denote the mean and variance of observed conditional probabilities (MLE), as well as following density estimation by latent class models with two through four latent classes, across 2000 simulated data sets of 200 observations. Thick bars span 80% of the simulated estimates; thin bars denote the range of estimates from minimum to maximum.

or that the percentage of white *and* nonwhite males aged 18–29 who voted for Kerry are both equal to the overall percentage of survey respondents who voted for Kerry: 51.7%. Depending upon the application, this may or may not be an accurate assumption; here, clearly, it is not. But because the sample-to-sample variance in the estimated marginal Pr(Kerry) is so small (due to a much larger sample size) compared with the variance of the small-cell conditional estimate, the marginal probability will be a better estimate, on average, in very small samples and as long as the true conditional and marginal probabilities are not too dissimilar.[6] The problem, of course, is that there is no way of knowing a priori how dissimilar the two population values actually are. Additionally—and unhelpfully—any estimates of conditional effects will necessarily equal zero as is apparent in the right-hand plot in Fig. 3.

The lesson for applied researchers is this: If a subgroup is so small that the observed small-cell (conditional) probability will be a worse estimate of the population value than the corresponding marginal probability, then the most prudent choice is to use *neither*. Fortunately, a minimal amount of smoothing using a two-class latent class model can be effective in recovering the small-cell probability in the underlying population.

Estimates based on a fully specified multinomial logit model are nearly indistinguishable from simply using the cell percentages directly observed in the cross-classification table. It is possible that a multinomial logit model specified with fewer than the complete set of interactions among the independent variables might have produced estimates with a lower RMSE—but which among the myriad potential specifications to choose? The latent class model–based estimation procedure sidesteps the issue altogether.

3.3 *Bias in Latent Class Model-Based Estimates*

Although the latent class model improves the precision of estimated conditional probabilities and marginal effects, the smoothing does introduce a small amount of bias. Simonoff (1995, 48) captures this trade-off succinctly: "One way to view the use of smoothing methods is as an attempt to balance the low bias of undersmoothing with the low variability of oversmoothing." The question is by how *much* variability is reduced and at what cost in bias. Figure 4 separates out the bias and variance of each estimator based upon 2000 random samples of 200 individuals drawn from the complete 2004 election data set. For white males aged 18–29 (left panel), the model-based estimates tend to overestimate the true percent voting for Kerry by approximately 5% on average. For nonwhite males (center panel), the overestimation is approximately 8%. Meanwhile, the MLE is unbiased across the 2000 simulated samples.

In practice, however, we only observe one sample—not 2000. And in any one sample, the model-based estimates are much more reliable than the observed cell percentages. Among random samples of 200, the number of white males aged 18–29 varied from a minimum of 2 to a maximum of 26. The sampled number

[6] This logic is highly similar to that of the benefits of shrinkage estimators in general; for example, in the context of a Bayesian hierarchical model, where the RMSE is reduced by adjusting within-unit estimates toward the global mean.

https://doi.org/10.1093/pan/mpr006 Published online by Cambridge University Press

Table 1 Sample sizes of four CBS News polls prior to Election Day, 2008, as well as numbers of respondents in various subgroups

	Date of poll			
	10/28–30	*10/29–31*	*10/29–11/1*	*10/31–11/2*
Complete sample	833	1390	1167	1051
Whites	716	1192	995	876
White males	293	488	402	357
White male independents	105	169	141	126
and, income under $15,000	9	13	10	4
and, income $15–30,000	13	19	10	8
and, income $30–50,000	14	28	24	27
and, income $50–75,000	25	34	31	18
and, income $75–100,000	20	27	19	15
and, income over $100,000	17	40	38	44

of nonwhite males 18–29 ranged from 0 to just 13. In the former group, preprocessing the observed cross-classification table using a four-class model reduces the variance of the estimator by nearly *half* compared with the MLE. The variance of the two-class model-based estimate is less than *one-third* that of the MLE.

Moving to the even smaller group of nonwhites, the variance of the MLE is much larger, but the variance of the model-based estimates stays almost unchanged. Now, the effect of preprocessing the data is an 80% reduction in variance for the two-class model compared with the MLE. Whereas the range of estimates from the two-class model is 38–98%, the range of estimates taken from the observed cross-classification table is 0–100%, with 100% being the modal MLE, appearing in over one-quarter of all samples. In an additional 1% of samples, no nonwhite young males were present in the sample; the latent class model can still produce estimates of the percentage of nonwhite young males voting for Kerry, but the MLE cannot.

The latent class model–based estimates are minimally biased but also much more tightly distributed around their expected value in repeated samples. With fewer latent classes—and hence, greater smoothing—the variance of the sampling distribution of $\tilde{P}(y_c)$ decreases. The minor amount of bias in the model-based estimates is easily tolerable from an applied perspective, and well worth the substantial decrease in variance (and RMSE) that the density estimator provides.

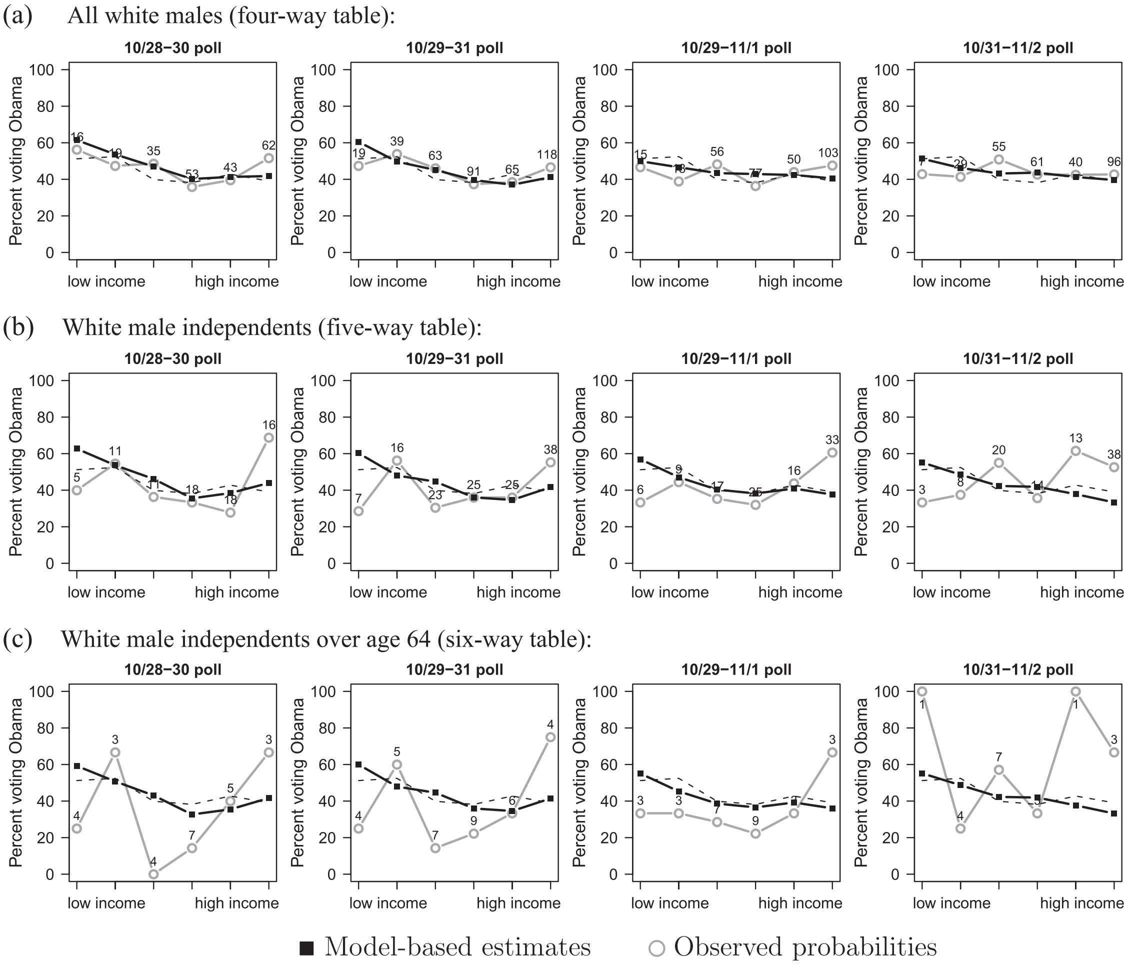
4 Application: Election Forecasting in Small Subgroups

A key question leading up to the 2008 U.S. presidential election was whether political independents would ‘‘break’’ toward Democrat Barack Obama or Republican John McCain on Election Day. The question focused especially on white independents, as it was understood that just as Republican voters would overwhelmingly support McCain, Democrats, and nonwhites would be largely voting for Obama. Observers also debated what effect income would play on vote choice, as well as if a ‘‘gender gap’’ would persist with men less likely than women to vote for Obama.

In the final week of the campaign, the CBS News organization conducted four national surveys asking voters if they intended to support Obama or McCain as well as a battery of other attitudinal and demographic items (CBS News 2008a, 2008b, 2008c, 2008d). Although the overall sample sizes of these polls were large, stratifying by race, sex, party identification, and family income level leaves relatively few observations from which to draw reliable inferences about vote intentions within each subgroup (Table 1). The number of white male independents with family income below $15,000, for example, ranged from only 4–13 respondents across the four polls. As a result, we expect a large amount of sampling variation from poll-to-poll in the observed percentage of white male independents at each level of income who intend to vote for Obama. In the case of white male independents earning $75–100,000, estimates of Obama’s vote share ranged from 28% to 62% across the four polls—a difference of 34 percentage points. For the purpose of forecasting the election outcome, results from any single poll will be at best uninformative; and at worse, misleading.

To assess the accuracy of predictions based upon cross-tabulating the ‘‘raw’’ survey data versus model-based predictions produced by a two-class latent class model, I compare both sets of estimates with the results

https://doi.org/10.1093/pan/mpr006 Published online by Cambridge University Press


(a) All white males (four-way table):

(b) White male independents (five-way table):

(c) White male independents over age 64 (six-way table):

Fig. 5 Estimated effect of income on vote choice for (a) all white males, (b) white male independents, and (c) white male independents over age 64 in four separate CBS News polls leading up to the 2008 U.S. presidential election. Gray lines connect observed conditional probabilities, with cell sizes labeled adjacent to each point. Black lines denote latent class model–based estimates. The actual proportion of white males voting for Obama, according to the 2008 presidential election exit poll, is indicated by the dashed line. Income is measured across six categories, from under $15,000 to over $100,000, as in Table 1.

of the very large, 18,018-respondent postelection exit poll fielded by the National Election Pool (2008). Applying the latent class model to the preelection CBS survey data consistently and correctly forecasts the actual election result in each of the four polls at multiple levels of stratification. The latent class model recovers information that exists in each survey data set but is not utilized in a basic contingency table.

I begin by examining the voting intentions of white males at each level of family income. Estimates of Obama's vote share based upon the four-way table of race, sex, income, and vote choice fluctuate in a 5–15 percentage point range across the four polls. To attempt to improve upon these estimates, I apply a two-class latent class model to these four variables in each survey data set. I then calculate the model-based estimated percent voting for Obama in each subgroup using equation (4). Results are shown in Fig. 5a. The latent class model–based estimates demonstrate less sample-to-sample variability than the observed percentages, but the difference is minimal. The dashed lines in Fig. 5 show the actual percentage of white males at each level of income who reported voting for Obama in the exit poll.[7] The latent class model–based estimates match the results from the exit poll almost identically in both the trend and

[7] Percentages from the exit poll are obtained following weighting by the survey's officially released sampling weights. The survey included responses from 5502 white males.

the level of support for Obama at different levels of income. At the very least, no harm is being done by the model-based estimator, as it does not appear that the model-based estimates are systematically biased above or below the observed cell percentages.

The power of the latent class model–based estimation procedure becomes more strongly evident in smaller subgroups. Stratifying further by party identification, Fig. 5b shows the observed percent voting for Obama among white male *independents* at various levels of income, as well as the latent class model–based estimates following application of a two-class model to the five-way table in each CBS poll. The model-based estimates consistently indicate that white male independents earning over $100,000 are approximately 20 percentage points less likely to vote for Obama than those earning <$15,000. From poll-to-poll, each set of model-based estimates reveals almost *exactly the same* relationship. The observed cell percentages, in contrast, fluctuate markedly across the four surveys, masking the finding—borne out by the election results—that greater income is associated with decreased support for Obama.[8]

Finally, I stratify the table once more, this time by age, and consider the vote choices of *elderly* white male independents—those over age 64. The number of such individuals in each of the four polls is, respectively, 29, 40, 33, and 26—and these are then further subdivided across the six income categories. As shown in Fig. 5c, the observed cell percentages of support for Obama among elderly white male independents at different levels of income vary wildly between the four polls, failing to convey any consistent or meaningful information. Instead, I once again reestimate the two-class latent class model, adding in the four-category age variable to the variables for race, sex, party identification, family income, and vote choice. The latent class model–based estimates still succeed in revealing the underlying pattern in the relationship in all four of the preelection polls.

5 Conclusion

This paper has proposed a solution to the problem posed by small cell sizes when estimating table-based statistics in finite samples and following stratification by multiple categorical variables. In research based on public opinion survey data, this is a barrier frequently confronted by academics and applied practitioners alike. In large contingency tables, having too few observations per cell leads to highly unreliable estimates of cell percentages, conditional probabilities, and marginal effects. Most often, researchers simply decline to report these quantities once cell sizes become too small, recognizing that their estimates are subject to extremely large amounts of sampling error.

I show that for small subgroups arising from repeated stratification, improved estimates of conditional probabilities and marginal effects can be produced by preprocessing the observed data with a latent class model. As a density estimator, the latent class model can closely approximate the underlying joint distribution of the variables of interest based upon the observed multivariate data. This density estimate will be relatively stable from sample-to-sample. As a result, substituting model-based cell percentage estimates $\tilde{P}(y_c)$ for the observed cell percentages, $\hat{P}(y_c)$, can significantly reduce the mean squared error of estimates of conditional probabilities and marginal effects—especially for subgroups containing 40 observations or fewer.

One of the most appealing features of the method described in this paper is its ease of use. Unlike when estimating a GLM or MLM, the latent class model–based approach does not require the researcher to make a large number of potentially arbitrary or idiosyncratic decisions about how to specify the model. Two latent classes are generally sufficient to reliably estimate conditional probabilities and marginal effects. Software to perform the necessary calculations is freely available as part of the R statistical computing

[8] The observed cell sizes shown in Fig. 5 are smaller than those given in Table 1 because a small number of respondents in each subcategory declined to state their vote intention.

https://doi.org/10.1093/pan/mpr006 Published online by Cambridge University Press

program. The procedure has foreseeable benefits in studies of public opinion, marketing, demography, epidemiology, and other areas in which multivariate contingency table analysis is common and sample sizes are moderate to small.

References

Abrajano, Marisa A., R. Michael Alvarez, and Jonathan Nagler. 2008. The Hispanic Vote in the 2004 presidential election: insecurity and moral concerns. *The Journal of Politics* 70:368–82.

Achen, Christopher H. 2002. Toward a new political methodology: microfoundations and ART. *Annual Review of Political Science* 5:423–50.

Achen, Christopher H. 2005. Let's put garbage-can regressions and garbage-can probits where they belong. *Conflict Management and Peace Science* 22:327–39.

Agresti, Alan. 2002. *Categorical data analysis*. 2nd ed. Hoboken, NJ: John Wiley & Sons.

Agresti, Alan, James G. Booth, James P. Hobert, and Brian Caffo. 2000. Random-effects modeling of categorical response data. *Sociological Methodology* 30:27–80.

Agresti, Alan, and David B. Hitchcock. 2005. Bayesian inference for categorical data analysis. *Statistical Methods & Applications* 14:297–330.

Aitchison, J., and C. G. C. Aitken. 1976. Multivariate binary discrimination by the kernel method. *Biometrika* 63:413–20.

Asher, Herbert. 2007. *Polling and the public: What every citizen should know.* 7th ed. Washington, DC: CQ Press.

Balz, Daniel J., and Haynes Johnson. 2009. *The battle for America 2008: The story of an extraordinary election*. New York: Viking.

Bandeen-Roche, Karen, Diana L. Miglioretti, Scott L. Zeger, and Paul J. Rathouz. 1997. Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association* 92:1375–86.

Bartholomew, David J., Fiona Steele, Irini Moustaki, and Jane I. Galbraith. 2008. *Analysis of multivariate social science data*. 2nd ed. Boca Raton, FL: Chapman & Hall.

Berry, William, Jacqueline H. DeMeritt, and Justin Esarey. 2010. Testing for interaction in binary logit and probit models: is a product term essential? *American Journal of Political Science* 54:248–66.

CBS News. 2008a. *CBS News Monthly Poll #2, October 2008 (Computer file). ICPSR26826-v1*. Ann Arbor, MI: Inter-University Consortium for Political and Social Research [distributor] January 29, 2010. doi:10.3886/ICPSR26826.

CBS News. 2008b. *CBS News Monthly Poll #3, October 2008 [Computer file]. ICPSR26827-v1*. Ann Arbor, MI: Inter-University Consortium for Political and Social Research [distributor] January 29, 2010. doi:10.3886/ICPSR26827.

CBS News. 2008c. *CBS News Monthly Poll #4, October 2008 (Computer file). ICPSR26832-v1*. Ann Arbor, MI: Inter-University Consortium for Political and Social Research [distributor] January 04, 2010. doi:10.3886/ICPSR26832.

CBS News. 2008d. *CBS News Monthly Poll #5, October 2008 (Computer file). ICPSR26828-v1*. Ann Arbor, MI: Inter-University Consortium for Political and Social Research [distributor] December 14, 2009. doi:10.3886/ICPSR26828.

Cillizza, Chris. 2007. *Romney's data cruncher.* The Washington Post, 5 July.

Congdon, Peter. 2005. *Bayesian models for categorical data*. Chichester, UK: John Wiley & Sons.

de la Garza, Rodolfo O., and Jeronimo Cortina. 2007. Are Latinos Republicans but just don't know it? The Latino Vote in the 2000 and 2004 presidential elections. *American Politics Research* 35:202–23.

Fraley, Chris, and Adrian E. Raftery. 2002. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* 97:611–31.

Garrett, Elizabeth S., and Scott L. Zeger. 2000. Latent class model diagnosis. *Biometrics* 56:1055–67.

Gelman, Andrew, and Jennifer Hill. 2007. *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.

Ghosh, M., and J. N. K. Rao. 1994. Small area estimation: an appraisal. *Statistical Science* 9:55–76.

Goodman, Leo A. 1974a. The analysis of systems of qualitative variables when some of the variables are unobservable. Part I—a modified latent structure approach. *The American Journal of Sociology* 79:1179–259.

Goodman, Leo A. 1974b. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* 61:215–31.

Grund, B. 1993. Kernel estimators for cell probabilities. *Journal of Multivariate Analysis* 46:283–308.

Hagenaars, Jacques A., and Allan L. McCutcheon. 2002. *Applied latent class analysis*. New York: Cambridge University Press.

Hall, Peter. 1981. On nonparametric multivariate binary discrimination. *Biometrika* 68:287–94.

Heeringa, Steven G., Brady T. West, and Patricia A. Berglund. 2010. *Applied survey data analysis*. Boca Raton, FL: Chapman and Hall.

Huang, Guan-Hua. 2005. Selecting the number of classes under latent class regression: a factor analytic analogue. *Psychometrika* 70:325–45.

Jackson, John. 1989. An errors-in-variables approach to estimating models with small area data. *Political Analysis* 1:157–80.

Jamieson, Kathleen Hall 2009. *Electing the President, 2008: The insiders' view.* Philadelphia, PA: University of Pennsylvania Press.

Lax, Jeffrey R., and Justin H. Phillips. 2009. How should we estimate public opinion in the states? *American Journal of Political Science* 53:107–21.

Lazarsfeld, Paul F. 1950. The logical and mathematical goundations of latent structure analysis. In *Measurement and prediction*, ed. Samuel A. Stouffer, 362–412. New York: John Wiley & Sons.

Leal, David L., Matt A. Barreto, Jongho Lee, and Rodolfo O. de la Garza. 2005. The Latino Vote in the 2004 Election. *PS: Political Science & Politics* 38:41–9.

https://doi.org/10.1093/pan/mpr006 Published online by Cambridge University Press

Linzer, Drew A., and Jeffrey Lewis. 2010. *poLCA: Polytomous variable latent class analysis*. R package version 1.2. http://userwww.service.emory.edu/~dlinzer/poLCA.

Long, J. Scott. 1997. *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage Publications.

Maddala, G. S. 1983. *Limited-Dependent and Qualitative Variables in Econometrics*. New York: Cambridge University Press.

McLachlan, Geoffrey J., and David Peel. 2000. *Finite mixture models*. New York: John Wiley & Sons.

National Election Pool. 2008. *Poll #2008-NATELEC: National Election Day Exit Poll (USMI2008-NATELEC)*. ABC News/Associated Press/CBS News/CNN/Fox News/NBC News.

National Election Pool, Edison Media Research, and Mitofsky International. 2004. *National Election Pool General Election Exit Polls (Computer file)*. ICPSR version. Somerville, NJ: Edison Media Research/New York, NY: Mitofsky International [producers], 2004. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor].

Nylund, Karen L., Tihomi Asparouhov, and Bengt O. Muthén. 2007. Deciding on the number of classes in latent class analysis and growth mixture modeling: a Monte Carlo simulation study. *Structural Equation Modeling* 14:535–69.

Park, David K., Andrew Gelman, and Joseph Bafumi. 2004. Bayesian multilevel estimation with poststratification: state-level estimates from national polls. *Political Analysis* 12:375–85.

R Development Core Team. 2010. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0. http://www.R-project.org.

Rao, J. N. K. 2003. *Small area estimation*. Hoboken, NJ: John Wiley & Sons.

Simonoff, Jeffrey S. 1995. Smoothing categorical data. *Journal of Statistical Planning and Inference* 47:41–69.

Taylor, Paul, and Richard Fry. 2007. *Hispanics and the 2008 Election: A swing vote?* Washington, DC: Pew Hispanic Center.

Titterington, D. M. 1980. A comparative study of kernel-based density estimates for categorical data. *Technometrics* 22:259–68.

Todd, Chuck, and Sheldon Gawiser. 2009. *How Barack Obama won: A state-by-state guide to the historic 2008 presidential election*. New York: Vintage Books.

U.S. Census Bureau. 2008. *American Community Survey 1-year estimates*. http://www.census.gov/acs/www (accessed May 13, 2010).

Vermunt, Jeroen K., Joost R. Van Ginkel, L. Andries Van der Ark, and Klaas Sijtsma. 2008. Multiple imputation of incomplete categorical data using latent class analysis. *Sociological Methodology* 38:369–97.

https://doi.org/10.1093/pan/mpr006 Published online by Cambridge University Press