The Annals of Applied Statistics
2017, Vol. 11, No. 2, 680–697
DOI: 10.1214/16-AOAS1012

MODELING LOG-LINEAR CONDITIONAL PROBABILITIES FOR ESTIMATION IN SURVEYS

BY YVES THIBAUDEAU*, ERIC SLUD*,† AND ALFRED GOTTSCHALCK*

*U.S. Census Bureau** *and University of Maryland*†

The Survey of Income and Program Participation (SIPP) is a survey with a longitudinal structure and complex nonignorable design, for which correct estimation requires using the weights. The longitudinal setting also suggests conditional-independence relations between survey variables and early- versus late-wave employment classifications. We state original assumptions justifying an extension of the partially model-based approach of Pfeffermann, Skinner and Humphreys [*J. Roy. Statist. Soc. Ser. A* **161** (1998) 13–32], accounting for the design of SIPP and similar longitudinal surveys. Our assumptions support the use of log-linear models of longitudinal survey data. We highlight the potential they offer for simultaneous bias-control and reduction of sampling error relative to direct methods when applied to small subdomains and cells. Our assumptions allow us to innovate by showing how to rigorously use only a longitudinal survey to estimate a complex log-linear longitudinal association structure and embed it in cross-sectional totals to construct estimators that can be more efficient than direct estimators for small cells.

1. Introduction.

1.1. *Motivation and goal.* For the past three decades, survey statisticians have actively researched methods for estimating characteristics of "small area" populations [Ghosh and Rao (1994); Rao and Molina (2015)] that are individually too small for traditional direct methods. Lopez-Vizcaino, Lombardía and Morales (2015) and Molina, Saei and Lombardía (2007) used small area methods to model survey data and estimate cross-sectional labor force participation for small areas through multinomial regression involving demographic covariates and a random effect structure. We exploit multinomial structure in a related but different framework, cross-classifying data at higher geographic levels, such as the state level in the U.S. While fitting to larger areas, by primarily design-based methods, we propose modeling fine-grained classifications which incorporate a longitudinal aspect. Based on the models, we estimate aggregates corresponding to marginal cells coarser than the finest cross-classified cells. This is a nontraditional form of small-area estimation in which some marginal total estimates are design-based and others model-based, in terms of conditional probabilities.

Received May 2010; revised December 2016.
Key words and phrases. Log-linear model, conditional probability, Horvitz–Thompson estimator, model calibration.

The Survey of Income and Program Participation (SIPP) is an essential longitudinal survey in the U.S. federal statistical system, enabling researchers to study life changes and also to track many administrative variables [Haber (1985); Abowd et al. (2005)]. SIPP is invaluable because of these longitudinal and multivariate features. It also lends itself naturally to enhancements through integration of cross-sectional information, which is increasingly available through administrative lists. In this paper, we propose methodology for producing accurate and relatively bias free estimates of aggregates of interrelated small cells. SIPP estimates are important both as an ultimate goal and for demonstration of our method of estimating small cell totals.

In the SIPP context, we define a fine-grained classification of the population of a state by a time-dependent multivariate cross-sectional variable $\boldsymbol{e}$, jointly for labor force status and health insurance coverage, where $\boldsymbol{e}_i$ is the value of $\boldsymbol{e}$ for person i. The $\boldsymbol{e}$ classification is further refined by another vector variable $\boldsymbol{f}$ that includes the same classifying variables embedded in $\boldsymbol{e}$ at an earlier point in time, as well as some demographic variables. We view the coordinate variables of $\boldsymbol{f}$ as "post-stratifiers." Then $(\boldsymbol{e}_i, \boldsymbol{f}_i)$ defines cell membership for unit i, and w_i is the sampling weight or inverse inclusion probability for that unit. The finer classification $(\boldsymbol{e}, \boldsymbol{f})$ and its corresponding cells are termed the "working" classification, while the partition into $\boldsymbol{e}$-defined cells is called the "target classification." The finer cells are instances of "gross flows," which decompose the population according to participation in the labor force at multiple points in time.

This paper focuses on the accurate estimation of small interrelated target cells. The motivation and payoff for doing this is substantial. High-quality large-area statistics such as gross flows, which share the same state totals as other population decompositions such as those defined by health-care coverage, are of great interest for researchers and policy makers. The role of the log-linear model in this work is to parameterize cell probabilities in the working classification by identifying the salient interactions between $\boldsymbol{e}$ and $\boldsymbol{f}$ while at the same time integrating the design of the survey into the model. We advocate a "small-cell model," leading to a method for partially design-based analysis incorporating administrative data, using log-linear models for conditional probabilities [Fienberg (1980)].

1.2. *Related work.* Log-linear models in small-area estimation seem to have appeared first in structure preserving estimation (SPREE), a synthetic estimation method due to Purcell and Kish (1979, 1980), succinctly presented in Rao and Molina (2015). Early versions of SPREE relied on iterative proportional fitting (IPF) to combine an interaction structure derived from a census with main effects computed from a survey. The original idea behind SPREE is that log-linear interactions are more stable through time than main effects, which need updating from contemporary survey estimates. Purcell and Kish [(1979), Section 9.5] point to Purcell's thesis (1979) and a conference paper, Chambers and Feeney (1977),

as early examples of categorical data methods in small-area estimation, the latter proposing to apply the "association structure" model (what we term below the conditional-probability specification given the stratifiers) fitted by IPF on census data and assumed to persist in a later sample survey in synthetic estimates for the sample survey.

Noble, Haslett and Arnold (2002) cite Marker (1999) as the first to observe explicitly that SPREE can be formally expressed as a log-linear model, and they and Zhang and Chambers (2004) extend SPREE through log-linear and generalized linear models. But all of these authors preserve the SPREE idea of fitting the "association structure" model on census (or other aggregated) data and then applying it synthetically to a current sample survey.

Our approach is distinguished from this synthetic method by viewing the log-linear model as a way to specify and fit conditional probability relationships on the current sample survey from which small-area estimates are to be produced. We know of no other paper that uses log-linear models to generate small-area estimates from current survey data alone, without an external source and synthetic assumption. Our application of log-linear models to complex survey data exploits a likelihood structure factoring into a marginal saturated multinomial likelihood and a parametric-model likelihood for the conditional probabilities. The survey data (SIPP, in our case) are modeled directly, without assuming that they share an externally (census-) derived set of model parameters. However, complex surveys like SIPP do make use of current Population Estimates through their nonresponse-adjusted calibrated weights.

We begin by reviewing previous published uses of conditional probabilities in longitudinal surveys, and of survey weighting of sampled data assumed to satisfy a log-linear model. We then present the high-level features of our approach using log-linear-model-assisted design-based analysis. The dependence structure of the log-linear models we have in mind has been developed through Markov and semi-Markov modeling in an extensive body of work on the National Crime Victimization Survey (NCVS). Fienberg (1980) showed how to construct a longitudinal transition matrix for a discrete or continuous time Markov process to model victimization. Conaway and Lohr (1994) studied such processes to identify the most salient factors related to repeated victimization. In the same context, Saphire (1984) and Ybarra and Lohr (2002) proposed various estimators based on discrete-time conditional probabilities which we extend in greater generality.

Fienberg and Stasny (1983) presented conceptually simple assumptions for a model of a probabilistic longitudinal survey design measuring gross flows and allowing sampling to be outcome dependent. Stasny (1987) modeled both the transition probabilities and the nonresponse mechanism for gross flows. Her models go beyond the scope of this paper, in which we do not consider nonresponse or attrition. The framework of Fienberg and Stasny (1983) underlies our methods in this paper. However, these authors and the others cited above ignored pre- and

post-survey weighting, while we emphasize the use of survey weights and design-based analysis together with log-linear models. We regard as essential the following: (i) the definitions of sampling strata and analysis post-strata may not coincide, (ii) in general, the modeled conditional probabilities are not specified to be homogeneous between sampling strata or within analysis post-strata, and (iii) the final survey weights may embed exogenous information, such as information from a census, not reflected in the survey data. We state simple assumptions that allow us to simplify the framework of Fienberg and Stasny and take advantage of the weighting structure, and integrate parametric models and design information together into "hybrid" estimators.

2. Motivating the methodology. Pfeffermann, Skinner and Humphreys (1998) show how to merge conditional probabilities modeled through multiple separate logistic regressions and to weight aggregates to correct estimates for measurement error. They motivate their analysis via a comprehensive "noninformative sampling" assumption. Our somewhat different objective is to reduce the sampling variability of the estimators, which we do through the automatic sharing of parameters across cells that is inherent to hierarchical log-linear models for the frame population and the sample.

In our formulation, the population is regarded as an identically and independently distributed sample from a "superpopulation" [Fuller (2009), p. 128], [Fuller and Isaki (1981)] with categorical descriptors. Our version of the superpopulation model and noninformative sampling assumption of Pfeffermann, Skinner and Humphreys (1998) makes essential use of multidimensional categorical variables $\boldsymbol{e}_i$, $\boldsymbol{f}_i$ for sampled units. Our assumptions restrict the joint distributions of these variables.

A1. The finite population from which data are sampled, together with their sampling weights (inverse inclusion probabilities), is an *i.i.d.* sample of vectors $(\boldsymbol{e}_i, \boldsymbol{f}_i, w_i)$, and the categorical vectors $\boldsymbol{e}_i$, $\boldsymbol{f}_i$ jointly follow a log-linear model within which the marginal models of $\boldsymbol{e}_i$ and of $\boldsymbol{f}_i$ are each saturated, and specified interactions between $\boldsymbol{e}_i$, $\boldsymbol{f}_i$ are absent.

A2. The sampling weights w_i are conditionally independent of $\boldsymbol{e}_i$ given $\boldsymbol{f}_i$.

Under these assumptions, ignoring the sampling weights when deriving the maximum likelihood estimators (MLEs) of conditional probabilities from the log-linear model leads to approximately unbiased hybrid estimators for the subdomains of the population defined by the characteristics $\boldsymbol{e}_i$, $\boldsymbol{f}_i$, and for population ratio parameters. Throughout the paper, we refer to the subdomains of the population or sample defined by $\{i : \boldsymbol{f}_i = \boldsymbol{f}\}$ as *post-strata*.

Assumption A1 is restrictive in asserting that a model holds for the finite population based on suppression of a specified set of higher-order interactions of the saturated log-linear model. The suppressed interactions depend on the logic of the dependence structure (longitudinal, in our main example) that is plausible for a

particular survey. The restriction is imposed in settings where the survey sample is not large enough to allow detection and estimation of all high-order interactions within the log-linear model. Under the submodel, approximately unbiased estimators are found for the population sufficient statistics for the submodel, in turn enabling derivation of estimators of conditional probabilities and of totals based on them. Assumption A2 means that estimated conditional probabilities derived from the log-linear model are unrelated to the sampling weights within the post-strata. When A2 holds, the residual weight variability within the post-strata is ancillary and can be ignored.

Our approach is original in directly modeling the longitudinal conditional transition probabilities using a hierarchical log-linear model with parameters shared between different strata in conjunction with direct survey-weighted estimates of post-stratum sizes, generalizing the approach of Pfeffermann, Skinner and Humphreys (1998), whose separate multinomial regressions did not share parameters. The formulation of Assumptions A1 and A2 is also new, providing a setting where design-based and model-assisted use of a log-linear-submodel yield large-sample model-consistent inferences for small domains, with improved accuracy over the purely design-based Horvitz–Thompson (HT) estimator. It commonly arises that HT estimators are insufficiently precise, for example, when the contingency tables derived from the survey data are relatively sparse, and that conditional-independence assumptions justifying suppression of high-order interactions are tenable.

The hybrid approach is conceptually similar to post-stratification except for the important caveat that stratum sizes are unknown. The research of Reilly, Gelman and Katz (2001) on post-stratification with unknown stratum sizes involved combining several samples and modeling the stratum sizes as MA(1) processes to stabilize variances. Their approach brings the experimental situation closer to that of traditional post-stratification with known stratum sizes. Instead of modeling stratum sizes, we share parameters across strata with unknown sizes.

Section 3 provides notation for defining hybrid estimators. Section 4 provides background on SIPP. Section 4.2 introduces the main example involving labor force characteristics and health care coverage in SIPP. Section 5 presents the multinomial log-linear models for which we derive MLEs of the conditional probabilities that underlie our approach. Section 5.2 develops our evaluation methodology, based on Balanced Repeated Replication (BRR) for variance estimation, and presents the main numerical results for the SIPP data. Section 6 summarizes the results and draws conclusions regarding the benefits of our approach. We include a web supplement containing theoretical details on the estimability of conditional probabilities for our constrained log-linear models, along with simulation results supporting BRR as a tool for variance estimation.

3. Modeling conditional probabilities in complex surveys. Our strategy is to estimate categorical conditional probabilities, which are easy to express in

log-linear models, under restrictions defined by linear constraints on the conditional log-odds of a multinomial distribution [Agresti (2013), page 345]. Similarly, Slavković, Zhu and Petrović (2015) formulated survey nondisclosure problems directly in terms of conditional rather than the more usual cell probabilities, and investigated the distributions of the ancillary statistics, conditional on the sufficient statistic, problems and methods beyond the scope of this paper. In the literature on log-linear modeling, as here, the exponential family structure ensures that estimates of conditional probabilities are based on sufficient count statistics.

In this section, we define notation for a specific class of estimators which we call hybrids because they combine a design consistent with a model-based component. Let $\boldsymbol{e} = (e_1, \ldots, e_A)'$ and $\boldsymbol{f} = (f_1, \ldots, f_B)'$ be vectors of indices or classifying variables, where $e_a = 1, \ldots, E_a$ is the ath variable used to define domains, and $f_b = 1, \ldots, F_b$ is the bth variable used to define post-strata, for $a = 1, \ldots, A$ and $b = 1, \ldots, B$. We want to estimate domain and subdomain totals within the frame population $\mathcal{U}$ from a probability sample survey with inverse inclusion probabilities as weights. To that end, consider the full classification of the population by cells $E_1 \times \cdots \times E_A \times F_1 \times \cdots \times F_B$, and let $Y_{\boldsymbol{ef}} = |\{i \in \mathcal{U} : \boldsymbol{e}_i = \boldsymbol{e}, \boldsymbol{f}_i = \boldsymbol{f}\}|$ be the population count for cell $e_1, \ldots, e_A, f_1, \ldots, f_B$. Then the population count for post-stratum $\boldsymbol{f}$ is

$$Y_{+\boldsymbol{f}} = \sum_{\boldsymbol{e} \in E} Y_{\boldsymbol{ef}}. \tag{3.1}$$

Similarly, let $E = E_1 \times \cdots \times E_A$ be the full set of values for the $\boldsymbol{e}$ index A-tuples, define the population domain total for domain $\boldsymbol{e}$ by

$$Y_{\boldsymbol{e}+} = \sum_{\boldsymbol{f} \in F} Y_{\boldsymbol{ef}}, \tag{3.2}$$

and let $F = F_1 \times \cdots \times F_B$ be the set of values for the $\boldsymbol{f}$ index B-tuples. Denote by $Y_{\boldsymbol{e}}(g)$ the population total for the subdomain restricting $\boldsymbol{e}$ to the specific value $f_B = g$ for the last post-stratum index, that is,

$$Y_{\boldsymbol{e}}(g) = \sum_{\boldsymbol{f} \in F(g)} Y_{\boldsymbol{ef}}, \tag{3.3}$$

where $F(g) = F_1 \times \cdots \times F_{B-1} \times \{g\}$. We also consider the case of a subdomain restricted by two stratification variables, $f_{B-1} = g$ and $f_B = h$, to obtain

$$Y_{\boldsymbol{e}}(g, h) = \sum_{\boldsymbol{f} \in F(g,h)} Y_{\boldsymbol{ef}}, \tag{3.4}$$

where $F(g, h) = F_1 \times \cdots \times F_{B-2} \times \{g\} \times \{h\}$. The subdomains whose totals are defined in (3.3) and (3.4) are the "target cells," and are defined by a combination of domain and post-stratum indices.

We use uppercase Y modified by subscripts to denote population counts in the frame population $\mathcal{U}$, while weighted sample subtotals are denoted by lowercase letter y with subscripts. Hats are used to designate estimators of population counts based on sampled data, a single hat (as in $\hat{Y}$ with subscripts) to denote Horvitz–Thompson (HT) weighted-total estimators and a double hat (as in $\hat{\hat{Y}}$ with subscripts) for our hybrid estimator. Throughout, subscripts of arrays Y and y are replaced by $+$'s to indicate that those subscripts are summed out, for example, $Y_{+k+mn} = \sum_{j,l=1}^{2} Y_{jklmn}$. Then the HT estimators $\hat{Y}_{+f}$, $\hat{Y}_{e+}$, $\hat{Y}_{e}(g)$ and $\hat{Y}_{e}(g,h)$ are defined by replacing the population cell totals Y_{ef} by their HT estimators $\hat{Y}_{ef}$ in (3.1)–(3.4).

Let $P_{e|f}$ be a conditional probability expressed in terms of the parameters of a log-linear probability model for the indices (e, f) of a randomly (equiprobably) sampled member of the population. Let $\hat{P}_{e|f}$ be the MLE for this conditional probability $P_{e|f}$, and define the hybrid estimators by

$$\hat{\hat{Y}}_{e+} = \sum_{f \in F} (\hat{P}_{e|f}) \hat{Y}_{+f}, \tag{3.5}$$

$$\hat{\hat{Y}}_{e+}(g) = \sum_{f \in F(g)} (\hat{P}_{e|f}) \hat{Y}_{+f}, \tag{3.6}$$

$$\hat{\hat{Y}}_{e+}(g,h) = \sum_{f \in F(g,h)} (\hat{P}_{e|f}) \hat{Y}_{+f}. \tag{3.7}$$

The accuracy of $\hat{\hat{Y}}_{e+}$, $\hat{\hat{Y}}_{e+}(g)$ and $\hat{\hat{Y}}_{e+}(g,h)$ relies on the (design and model) unbiasedness of $\hat{Y}_{+f}$, which is fundamental to our technique. This kind of estimator can be useful whenever log-linear models for $P_{e|f}$ have fewer degrees of freedom than the saturated model, and may lead to reduced variances of the estimators in (3.6) and (3.7) at the subdomain level. The hybrid estimators in (3.5)–(3.7) can be unweighted or weighted according to whether the estimator $\hat{P}_{e|f}$ is. When $\hat{P}_{e|f}$ is unweighted, it is the MLE for a log-linear model. The weighted version of $\hat{P}_{e|f}$ leads to a pseudo-MLE [Skinner (2011)].

4. Application to survey of income and program participation.

4.1. *Background on SIPP.* SIPP collects information on the income and well-being of the U.S. population through longitudinal panels with a complex design. From 1983 to 2013, SIPP was a panel survey consisting of 4-month waves of data collection, each wave staggered over four calendar months. The sample was divided into 4 rotating groups approximately equal in size. At each of the 4 calendar months in a wave, data for one of the 4 rotating groups was collected. The reference period for each monthly data collection was the 4 previous months. In 2013,

SIPP was redesigned and now has yearly installments, with 12 rotating groups. Although the redesigned survey is simpler, weights reflecting the probability sample design, post-survey adjustments and oversampling of low-income strata continue to be produced every month, and the methods of this paper still apply. An important feature of the SIPP design is a low-income stratum defined from information collected through the 2000 Census long and short forms [U.S. Bureau of the Census (2009)]. In the 2004 panel, the low-income stratum was oversampled at a 1.48 rate relative to the high income stratum. Thus the design is "informative," and analyses that ignore it are generally incorrect. The (cross-sectional) analysis weights (or "final weights") computed by the Census Bureau for each month of SIPP convey rich information beyond the data collected. The computation of these weights involves several nonresponse adjustments [Scott (2005)] to the design "base weights." Furthermore, after the first-stage nonresponse adjustments mentioned above, the SIPP weights are raked to demographic control totals derived from the 2000 decennial Census and administrative information. The process of producing final weights for the calendar months of wave 2 and beyond differed from that of wave 1 through further nonresponse adjustments. Again, this means that the weights contain necessary information, and discarding them may directly lead to estimation errors.

We illustrate with data reported from the first two waves of the 2004 SIPP panel. For purpose of illustration we simplify, retaining only the data describing the most recent reference month relative to the interview for each of the 4 rotating groups at wave 1 and 2. The implicit "within wave stationarity" assumption is that, for $j = 1, \ldots, 4$, the conditional probabilities for variables in month j of the second wave given variables in month j of the first wave do not depend on j.

We model only the population within the labor force in both waves, thus restricting to a "closed universe," as did Pfeffermann, Skinner and Humphreys (1998). An "open universe" would account for persons entering and leaving the labor force or population, as proposed by Fienberg and Stasny (1983). We use the same wave 1 weights for computing the HT estimators as for our hybrids. We are interested only in the population in the labor force initially identified at wave 1 that remains in wave 2. This means that the weighted sufficient statistics involved in the log-linear models will coincide with HT estimators of the wave-specific population domains (those defined by variables $\boldsymbol{e}$ or $\boldsymbol{f}$ in the notation of Section 3).

4.2. *Example*: *Gross flows and health care coverage for California.* Our example involves the sample counts for the state of California for waves 1 and 2 of the 2004 panel (from public-use data available at www.census.gov/sipp). Table 1 is a cross-classification of the surveyed persons within the labor force both at waves 1 and 2 by two qualitative variables, employment status (employed vs. unemployed) and medical insurance coverage (covered vs. not covered) for each wave. These

TABLE 1
Cross-Classification of SIPP 2004 *Sample Units for California*

		Wave 2			
		Employed		Unemployed	
Wave 1		**Covered**	**Not covered**	**Covered**	**Not covered**
		No college degree			
Employed	Covered	1252	44	8	4
	Not covered	82	604	2	13
Unemployed	Covered	20	2	13	1
	Not covered	4	40	1	25
		College degree			
Employed	Covered	1451	22	10	2
	Not covered	38	181	0	6
Unemployed	Covered	12	2	11	0
	Not covered	2	14	1	11

respondents were screened to remove cases subject to unit nonresponse at either wave.

In the notation of Section 3, with $\boldsymbol{e} = (jk)$, $\boldsymbol{f} = (lmn)$ and $\boldsymbol{ef} = (jklmn)$, $Y_{(jk)(lmn)}$ or Y_{jklmn} denotes the frame-population count of individuals with education n, labor status m and coverage l at wave 1, and labor status k and coverage j at wave 2. Values of educational level n are 1 for no college degree and 2 for college degree. Labor status values are 1 for employment and 2 for no employment, and medical insurance levels are 1 for coverage and 2 for lack of coverage. We have $A = 2$, $B = 3$, and the other terms in formulas (3.5)–(3.7) are given by

$$Y_{+\boldsymbol{f}} = Y_{++lmn} = \sum_{j,k=1}^{2} Y_{jklmn},$$

$$Y_{\boldsymbol{e}+} = Y_{jk+++} = \sum_{l,m,n=1}^{2} Y_{jklmn},$$

$$Y_{\boldsymbol{e}}(g) = Y_{jk}(g) = Y_{jk++g} = \sum_{l,m=1}^{2} Y_{jklmg},$$

$$Y_{\boldsymbol{e}}(g,h) = Y_{jk}(g,h) = Y_{jk+gh} = \sum_{l,l=1}^{2} Y_{jklgh},$$

$$P_{\boldsymbol{e}|\boldsymbol{f}} = P_{jk|lmn}.$$

The population counts $Y_{\boldsymbol{e}}(g)$, $Y_{\boldsymbol{e}}(g,h)$ and corresponding estimators denote all the necessary population and estimated target classifications.

5. Log-linear models for estimation.

5.1. *Models.* In this section, we present two log-linear model parameterizations in order to define conditional probabilities on a finite population. These models, indexed by j, k, l, m, n as above, are specifically tailored to the SIPP example. The second log-linear model is a parametric reduction or submodel of the first, with a lower-dimensional sufficient statistic.

Consider the following ANOVA-like log-linear multinomial model for r multinomially sampled observations classified into cells with the vector of cell probabilities $\boldsymbol{\pi} = (\pi_{11111}, \pi_{21111}, \ldots, \pi_{22222})'$, subject to explicitly parameterized interactions as follows:

$$\begin{aligned}\log(\pi_{jklmn}) = \alpha + \beta_{lmn} + \delta_j^J + \delta_k^K + \eta_{jk}^{JK} + \eta_{jl}^{JL} + \eta_{jn}^{JN} \\ + \eta_{jm}^{JM} + \eta_{km}^{KM} + \eta_{kn}^{KN} + \upsilon_{jkm}^{JKM},\end{aligned} \tag{5.1}$$

where the unknown parameters on the right-hand side satisfy constraints

$$\sum_{l,m,n=1}^{2} \beta_{lmn} = 0, \tag{5.2}$$

$$\sum_{j=1}^{2} \delta_j^J = \sum_{k=1}^{2} \delta_k^K = 0, \tag{5.3}$$

$$\begin{aligned}&\sum_{j=1}^{2} \eta_{jk}^{JK} = \sum_{k=1}^{2} \eta_{jk}^{JK} = \sum_{j=1}^{2} \eta_{jl}^{JL} = \sum_{l=1}^{2} \eta_{jl}^{JL} = \sum_{j=1}^{2} \eta_{jm}^{JM} = \sum_{m=1}^{2} \eta_{jm}^{JM} = 0, \\ &\sum_{j=1}^{2} \eta_{jn}^{JN} = \sum_{n=1}^{2} \eta_{jn}^{JN} = \sum_{k=1}^{2} \eta_{km}^{KM} = \sum_{m=1}^{2} \eta_{km}^{KM} = \sum_{k=1}^{2} \eta_{kn}^{KN} = \sum_{n=1}^{2} \eta_{kn}^{KN} = 0,\end{aligned} \tag{5.4}$$

$$\sum_{j=1}^{2} \upsilon_{jkm}^{JKM} = \sum_{k=1}^{2} \upsilon_{jkm}^{JKM} = \sum_{m=1}^{2} \upsilon_{jkm}^{JKM} = 0. \tag{5.5}$$

The term β_{lmn} is saturated like the vector of probabilities of a cell partition, with eight index levels and seven degrees of freedom, since it satisfies the single constraint (5.2). Thus, the cell counts based on the $\boldsymbol{f}$ indices (l, m, n) obey a fully saturated multinomial model, while the $\boldsymbol{ef}$ cell counts obey a conditional model given $\boldsymbol{f}$, with high-order interactions suppressed. Most of the third-order and some second-order interactions (those between J and M) have been suppressed in the model (5.1). The parameter dimension of this model is 16, with the $\boldsymbol{e}|\boldsymbol{f}$ conditional probability parameters accounting for 9. This entails considerable sharing of degrees of freedom across conditional cell probabilities since the saturated model for these $\boldsymbol{e}|\boldsymbol{f}$ probabilities has dimension $31 - 7 = 24$.

The validity of the population model (5.1)–(5.5), a specific instance of assumption A1, is ultimately an empirical question. It fits the SIPP data well, as we will show, and will be assessed along with A2 for the fitted model by comparisons of observed versus expected population margins. In particular, the third order interaction υ_{jkm}^{JKM} turns out to be crucial to a good fit, implying that the associations between coverage and employment status at each wave are not homogeneous [Agresti (2013), page 344] and the entire historical employment sequence is informative for insurance coverage status. We will refer to this model in the rest of the paper as *Model 1*.

A simplified form of Model 1 suppresses two of the terms in (5.1):

$$\upsilon_{111}^{JKM} = \eta_{11}^{JM} = 0. \tag{5.6}$$

This model (5.1)–(5.6), which we refer to as *Model 2*, reduces the dimension of the $\boldsymbol{e}|\boldsymbol{f}$ parameter to 7, and imposes additional conditional independence, saying that conditional on labor force status at wave 2, health coverage at wave 1, and education, coverage at wave 2 and labor force status at wave 1 are independent (conditional on indices k, l and n, indices j and m are independent). This restriction allows for much simpler algebraic development.

As with all log-linear models, the likelihoods of Model 1 and Model 2 can be shown to have an exponential family form, and the MLEs and pseudo-MLEs of the conditional probabilities in (3.5)–(3.7) can be obtained through traditional methods, such as iterative proportional [Bishop, Fienberg and Holland (1975), p. 83], [Winkler (1993)] or raking. This ensures an existing strong theoretical basis for the methods we are discussing.

The full set of unknown parameters in Model 2, (5.1)–(5.6), is given by

$$\boldsymbol{\beta} = (\beta_{111}, \beta_{211}, \beta_{121}, \beta_{221}, \beta_{112}, \beta_{212}, \beta_{122}) \tag{5.7}$$

and

$$\boldsymbol{\theta} = (\delta_1^J, \delta_1^K, \eta_{11}^{JK}, \eta_{11}^{JL}, \eta_{11}^{JN}, \eta_{11}^{KM}, \eta_{11}^{KN}), \tag{5.8}$$

and in Model 1, (5.1)–(5.5), the $\boldsymbol{\theta}$ parameter has the additional two components η_{11}^{JM}, η_{111}^{JKM}.

In the Web Supplement we show that, in both models, $\boldsymbol{\theta}$ is a smooth one-to-one reparameterization of the conditional $\boldsymbol{e}|\boldsymbol{f}$ probabilities under the constraints of the model. The joint MLEs $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\beta}}$ both enter the expressions for estimators of the desired conditional $\boldsymbol{e}|\boldsymbol{f}$ probabilities. It is possible to extend log-linear models to the situation where the joint MLE does not exist, but certain subvectors such as $\hat{\boldsymbol{\theta}}$ do [Fienberg and Rinaldo (2012)].

This representation for the likelihood is also useful in deriving the weighted variant of our hybrid estimator. Design-based survey estimators of superpopulation parameters are often constructed as maximizers of a sample-weighted estimator of the frame-level *census* log-likelihood. With dummy attribute-vectors $\mathbf{z}_i$ denoting the factor levels j, k, l, m, n for unit i in the frame population $\mathcal{U}$, the

census log-likelihood is $\sum_{i\in\mathcal{U}} \log p(\mathbf{z}_i, \beta, \theta)$, which could be defined by partitioning the frame population into cross-classified cells C_{jklmn} (or, equivalently, $\boldsymbol{ef}$ cells). Now, in terms of a *weighted* sample, the HT-type estimate of the census log-likelihood becomes

$$\sum_{i\in\mathcal{S}} \sum_{j,k,l,m,n} I_{[i\in C_{j,k,l,m,n}]}\omega_i \log \pi_{jklmn} = \sum_{j,k,l,m,n} \tilde{y}_{jklmn} \log \pi_{jklmn}, \tag{5.9}$$

where now the former array $\boldsymbol{y}$ of unweighted sample counts is replaced by the weighted version $\tilde{\boldsymbol{y}}$ based on sample weights ω_i,

$$\tilde{y}_{jklmn} = \sum_{i\in\mathcal{S}} I_{[i\in C_{jklmn}]}\omega_i. \tag{5.10}$$

The survey-estimated census log-likelihood, also called a *pseudo-log-likelihood* [Binder (1983)], has exactly the same form as the multinomial log-likelihood, except for the replacement of the unweighted array $\{y_{jklmn}\}$ of sample counts by the weighted sample totals (5.10). Again, any convergent IPF solution of the method-of-moments equations setting marginal weight totals equal to weighted totals of theoretical cell probabilities, which will always exist when the marginal totals are all positive, is necessarily a pseudo-MLE, and is unique by concavity of the log-likelihood as a function of the parameters. The pseudo-MLEs obtained in this way, for the model and example of the paper, are also called *weighted MLEs* of $(\boldsymbol{\beta}, \boldsymbol{\theta})$.

In general, the weighted version should be used to compute estimates of the population MLEs from the sample. However, if the population model is correct and our conditional independence assumption A2 holds, the weights can be dropped for estimating the conditional probabilities in the hybrids.

5.2. *Precision of hybrid estimators via balanced repeated replication.* To measure the standard error (SE) of the HT estimator and of the hybrid estimators based on the model of Section 5, we use Balanced Repeated Replication [BRR, Judkins (1990); Wolter (1985)]. BRR is approximately design-unbiased for the variance of estimators which are smooth functions of survey-weighted totals. However, the numerical stability of BRR depends on the number of degrees of freedom (df's, roughly proportional to the number of strata) being large [Krewski and Rao (1981)]. In California, SIPP supports only 23 estimation df's, too small a number for comfort. In Section 6 of the Web Supplement, we present a simulation study based on multinomial observations using the results of IPF for estimates of the multinomial probabilities [Thibaudeau, Slud and Gottschalck (2017)]. The results indicate that BRR is relatively free of bias, but highly variable with only 23 df's. However, the difference between two BRR estimates will have much smaller standard error than the BRR estimates themselves if they are highly correlated. Assuming that is the case, we employ the difference between the BRR variance estimate of the hybrid and that of the HT estimator as a performance measure for comparing the hybrid and HT estimators.

TABLE 2
Gross Flows in Units $\times 10^4$ *SIPP* 2004 *CA Data—College Degree*

Labor force status		Weighted			Unweighted	
Wave 2	**Wave 1**	**HT**	**Model 1**	**Model 2**	**Model 1**	**Model 2**
			Estimator			
Employed	Employed	663.4	663.1	662.5	663.5	662.8
Employed	Unemp	10.83	11.20	11.76	11.69	12.36
Unemp	Employed	7.090	7.466	8.026	7.049	7.723
Unemp	Unemp	9.590	9.213	8.654	8.727	8.052
			SE			
Employed	Employed	23.20	23.17	23.14	23.04	23.00
Employed	Unemp	1.316	1.539	1.527	1.571	1.562
Unemp	Employed	2.325	1.906	1.919	1.680	1.720
Unemp	Unemp	1.711	1.822	1.737	1.557	1.469

Tables 2 and 3 illustrate target classifications. Table 2 gives estimates for the size of subdomains identified by employment status at wave 1 and 2 (employed or unemployed) for the college educated subpopulation. Table 3 gives estimates for employment status at wave 2, insurance coverage at wave 2 and education for the same subpopulation. The estimators are computed using five different versions of the hybrids: the HT estimator, and both the unweighted and weighted versions of the hybrid estimator based on the two models discussed in Section 5. In some cases, weighting the hybrids appears to reduce the relative bias. Table 2 shows this clearly for the subdomain of unemployed in both waves with a college education. For both Models 1 and 2, weighting the hybrids reduces the gap between the hy-

TABLE 3
Health Coverage in Units $\times 10^4$ *SIPP* 2004 *CA Data—College Degree*

Wave 2		Weighted			Unweighted	
Coverage	**Labor force**	**HT**	**Model 1**	**Model 2**	**Model 1**	**Model 2**
			Estimator			
Covered	Employed	596.0	595.8	595.6	595.1	594.9
Uncov	Employed	78.24	78.48	78.62	80.09	80.24
Covered	Unemp	9.468	9.714	9.855	9.147	9.308
Uncov	Unemp	7.211	6.966	6.824	6.629	6.467
			SE			
Covered	Employed	23.33	23.28	23.25	22.98	22.95
Uncov	Employed	4.596	4.554	4.531	4.743	4.725
Covered	Unemp	1.920	1.958	2.031	1.775	1.827
Uncov	Unemp	1.851	1.773	1.708	1.563	1.520

brids and HT substantially. However, there is a price to pay for this bias reduction: with weighting, the variance is larger.

Our overarching goal has been to produce subdomain estimators more accurately than the direct HT method, especially in cases where HT estimates are egregiously inaccurate due to small sample sizes. We quantify our success through key ratios associated with the target classifications, and their relative errors. Tables 4–5 exhibit ratios of subdomain sizes in the numerators over domain sizes in the denominators. The subdomains are the cells of a 3-way classification given by all three variables in the tables, while the denominator domains are the 2-way marginal cells, as indicated in Tables 4 and 5.

Define the relative error of a hybrid ratio estimator as its standard deviation divided by the HT ratio estimator. The denominator in the relative error is chosen to be HT, instead of the hybrid ratio estimator, because the ratio in the latter case would be the coefficient of variation. With this choice, the difference in relative errors between two hybrid ratio estimators will not arise solely from a lack of stability of the ratio estimates. Because they involve the same 3-way classification as the hybrid in Table 2, we expect the hybrid ratio estimators in Table 4 to behave similarly to the hybrid estimator in Table 2 with respect to relative error.

In the comparison of Table 4 among weighted and unweighted hybrid estimators, the unweighted hybrid based on Model 1 arguably performs best. With no apparent increase in bias, it cuts down the largest relative errors associated with HT from 32% to 23% for the first ratio of unemployed, 23% to 18% for the second, and 18% to 14% for the insurance coverage ratio of unemployed with no college degree. Again, assuming the bias is negligible, this means substantially shorter confidence intervals when estimating the population ratios using the preferred model. Overall, the relative errors of three of the four unemployment ratios and one coverage ratio are cut substantially. Most importantly, the ratios with

TABLE 4

Unemployment Rate at Wave 2: *Estimators and Relative Errors*, *from the SIPP* 2004 *CA data*

Wave 1		**Weighted**			**Unweighted**	
Labor status	**Education**	**HT**	**Model 1**	**Model 2**	**Model 1**	**Model 2**
		Wave-2 unemployment rate $\times 10^{-2}$				
Employed	College	1.05	1.11	1.19	1.05	1.15
Employed	No Col	1.24	1.19	1.12	1.29	1.23
Unemp	College	46.9	45.1	42.3	42.7	39.4
Unemp	No Col	36.7	37.7	39.2	38.0	39.4
		Relative error				
Employed	College	0.325	0.266	0.267	0.232	0.236
Employed	No Col	0.228	0.169	0.180	0.178	0.185
Unemp	College	0.113	0.135	0.129	0.117	0.110
Unemp	No Col	0.146	0.128	0.123	0.126	0.121

TABLE 5
Health Insurance Coverage Rate at Wave 2: *Estimators and Relative Errors, from the SIPP* 2004 *CA data*

Wave 2	**Wave 1**	**Weighted**			**Unweighted**	
Labor status	**Education**	**HT**	**Model 1**	**Model 2**	**Model 1**	**Model 2**
		Wave-2 health-insurance coverage rate $\times 10^{-2}$				
Employed	College	88.3	88.3	88.3	88.1	88.1
Employed	No Col	68.1	68.1	68.2	67.9	67.9
Unemp	College	56.7	58.2	59.0	57.9	59.0
Unemp	No Col	36.3	35.1	34.7	35.18	35.0
		Relative error				
Employed	College	0.008	0.008	0.008	0.008	0.008
Employed	No Col	0.019	0.019	0.019	0.020	0.020
Unemp	College	0.117	0.108	0.108	0.109	0.110
Unemp	No Col	0.189	0.139	0.139	0.147	0.145

the largest relative errors yield the most substantial reductions in relative errors, a desirable outcome. The improved efficiency of ratios arises from the improved estimates of subdomain totals in Tables 2 and 3.

6. Discussion. Our log-linear estimation can be viewed as a kind of survey regression in terms of post-stratification variables $\boldsymbol{f}$ and classifiers $\boldsymbol{e}$. The setting here is analogous to that of Gelman's (2007) treatment of ordinary linear regression in a post-stratified setting if his sampling fractions are replaced by survey weights and linear regression is replaced by conditional logistic regression. The conditional logistic specification has the form, for Model 2 in (5.1)–(5.6), when $2 < j + k < 5$ and $1 < l, m, n < 3$,

$$\begin{aligned}\log\left(\frac{P_{jk|lmn}}{P_{11|lmn}}\right) &= \delta_j^J - \delta_1^J + \delta_k^K - \delta_1^K + \eta_{jk}^{JK} - \eta_{11}^{JK} + \eta_{jl}^{JL} - \eta_{1l}^{JL} \\ &\quad + \eta_{jn}^{JN} - \eta_{1n}^{JN} + \eta_{km}^{KM} - \eta_{1m}^{KM} + \eta_{kn}^{KN} - \eta_{1n}^{KN} \\ &= \psi_j^J + \psi_k^K + \xi_{jk}^{JK} + \phi_{jl}^{JL} + \phi_{jn}^{JN} + \phi_{km}^{KM} + \phi_{kn}^{KN}.\end{aligned} \tag{6.1}$$

Such a model would generally be used for sparse multinomial cross-classifications. If the corresponding log-linear model has a unique MLE estimable through IPF, as is true for Models 1 and 2, then the MLEs of the parameters of (6.1) derive directly from the multinomial log-linear model MLEs.

Log-linear models are applicable within surveys where the population can be regarded as a stratified simple random sample from an *i.i.d.* superpopulation, with certain survey variables plausibly satisfying additional conditional independence relations. We are interested in assessing the fit of the models in order to avoid

biased estimation of the target cells. Our hybrid model consists of the model (6.1) for the conditional probabilities of the small cells superimposed on the saturated model estimated in design-based fashion. The likelihood-ratio goodness-of-fit statistic G^2 for the model reflects the fit only of the conditional probabilities, since the marginal probabilities for the post-strata $\boldsymbol{f}$ follow the saturated model. According to (6.1), the conditional probabilities for $\boldsymbol{e}|\boldsymbol{f}$ have parameter dimension reduced from 24 in a fully saturated model to 7 in Model 2 so that the goodness-of-fit test has $24-7=17$ degrees of freedom. The likelihood is easily maximized via IPF and the goodness-of-fit (likelihood ratio) test yields $G^2=29.1$ ($p\approx 0.033$), which indicates a mediocre fit at best.

Model 1 has a representation similar to (6.1). The parameter dimension is reduced from the conditional probabilities of the saturated model from 24 down to 9 in Model 1, leaving 15 degrees of freedom for the goodness-of-fit test. In this case G^2 is calculated to be 15.5 ($p\approx 0.41$), qualifying Model 1 as statistically adequate. While Model 2 appears to perform well enough in Tables 4 and 5, its poor G^2 fit largely disqualifies it. We strongly advocate the use only of statistically adequate hybrid models in making estimation of the sort studied here, which supports although does not guarantee unbiasedness relative to HT estimates.

We showed that the log-linear Model 1 is compatible with the design due to its saturated structure at the post-stratum level, arises naturally from the population model due to the longitudinal nature of SIPP through a logical suppression of interactions, and nearly allows further suppression of interactions that would lead to the stricter parametric constraints of Model 2, which might have shown adequate fit if the sample size had been smaller.

Goodness-of-fit tests of Rao and Scott (1984) could also have been used to assess model fit from a design-based point of view, but the sampling variability inherent in cross-classifications with small cells will generally give these tests low power. Noisy survey weights overall will handicap weighted hybrids and will lead the practitioner to prefer unweighted hybrids, as in the example of the paper. But such a replacement needs to be motivated and assessed anew in each application.

As pointed out by a referee, nonparametric variance estimation through BRR is essential to our approach to assessing the hybrid models. However, BRR will suffer from insufficient degrees of freedom for small states. Our recommendation is to model several states together to increase degrees of freedom. Then state would be used as an additional stratifier. The hope is again that some of the parameters defining the log-linear conditional probabilities could be shared between strata.

Acknowledgments. Any views expressed on statistical methodology, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau. The authors are grateful to the Editors and Associate Editor for their guidance throughout this process, as well as to a referee for important suggestions.

SUPPLEMENTARY MATERIAL

Supplement to "Modeling log-linear conditional probabilities for estimation in surveys" (DOI: 10.1214/16-AOAS1012SUPP; .pdf). Technical Supplement contains three sections: Conditional Likelihood Representation; Conditional Probability Parameterization; and Simulation to evaluate BRR vs. the Laplace Method.

REFERENCES

ABOWD, J., STEPHENS, B., VILHUBER, L., ANDERSSON, F., MCKINNEY, K., ROEMER, M. and WOODCOCK, S. (2005). The LEHD Infrastructure Files and the Creation of the Quarterly Workforce Indicators. Technical Paper TP 2006-01. Available at lehd.ces.census.gov/doc/technical_paper/tp-2006-01.pdf.

AGRESTI, A. (2013). *Categorical Data Analysis*, 3rd ed. Wiley, Hoboken, NJ. MR3087436

BINDER, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Int. Stat. Rev.* **51** 279–292. MR0731144

BISHOP, Y. M. M., FIENBERG, S. E. and HOLLAND, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, MA. MR0381130

CHAMBERS, R. L. FEENEY, G. A. (1977). Log linear models for small area estimation. Paper presented at the Joint Conference of the CSIRO Division of Mathematics and the Australian of the Biometrics Society, Newcastle, Australia, Biometrics Abstract 2655.

CONAWAY, M. and LOHR, L. (1994). A longitudinal analysis of factors associated with reporting violent crimes to the police. *J. Quant. Criminol.* **10** 23–39.

FIENBERG, S. E. (1980). The measurement of crime victimization: Prospect for panel analysis of a panel survey. *J. Roy. Statist. Soc. Ser. D* **29** 313–350.

FIENBERG, S. E. and RINALDO, A. (2012). Maximum likelihood estimation in log-linear models. *Ann. Statist.* **40** 996–1023. MR2985941

FIENBERG, S. E. and STASNY, E. (1983). Estimating monthly gross flows in labour force participation. *Surv. Methodol.* **9** 78–101.

FULLER, W. (2009). *Sampling Statistics*. Wiley, New York.

FULLER, W. and ISAKI, C. (1981). Survey design under superpopulation models. In *Current Topics in Survey Sampling* (D. Krewski, J. N. K. Rao and R. Platek, eds.). Academic Press, San Diego.

GELMAN, A. (2007). Struggles with survey weighting and regression modeling. *Statist. Sci.* **22** 153–164. MR2408951

GHOSH, M. and RAO, J. N. K. (1994). Small area estimation: An appraisal. *Statist. Sci.* **9** 55–93. MR1278679

HABER, S. (1985). Applications of a Matched File Linking the Bureau of the Census Survey of Income Program Participation and Income Data. The Survey of Income and Program Participation WP 3.

JUDKINS, D. (1990). Fay's method for variance estimation. *J. Off. Stat.* **6** 223–239.

KREWSKI, D. and RAO, J. N. K. (1981). Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods. *Ann. Statist.* **9** 1010–1019. MR0628756

LOPEZ-VIZCAINO, E., LOMBARDÍA, M. and MORALES, D. (2015). Small area estimation of labour force indicators under a multinomial model with correlated time and area effects. *J. Roy. Statist. Soc. Ser. A* **178** 535–575.

MARKER, D. (1999). Organization of small area estimates estimators using a generalized linear regression framework. *J. Off. Stat.* **15** 1–24.

MOLINA, I., SAEI, A. and LOMBARDÍA, M. J. (2007). Small area estimates of labour force participation under a multinomial logit mixed model. *J. Roy. Statist. Soc. Ser. A* **170** 975–1000. MR2408988

NOBLE, A., HASLETT, S. and ARNOLD, G. (2002). Small area estimation via generalized linear models. *J. Off. Stat.* **18** 45–60.

PFEFFERMANN, D., SKINNER, C. and HUMPHREYS, K. (1998). The estimation of gross flows in the presence of measurement error using auxiliary variables. *J. Roy. Statist. Soc. Ser. A* **161** 13–32.

PURCELL, N. J. and KISH, L. (1979). Estimation for small domains. *Biometrics* **35** 365–384. MR0535774

PURCELL, N. J. and KISH, L. (1980). Postcensal estimates from local areas (or domains). *Int. Stat. Rev.* **48** 3–18.

PURCELL, N. J. (1979). Efficient Small Domain Estimation: A Categorical Data Approach. Unpublished Ph.D. Thesis, Univ. Michigan.

RAO, J. N. K. and MOLINA, I. (2015). *Small Area Estimation*, 2nd ed. Wiley, Hoboken, NJ. MR3380626

RAO, J. N. K. and SCOTT, A. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *Ann. Statist.* **12** 46–60. MR0733498

REILLY, C., GELMAN, A. and KATZ, J. (2001). Poststratification without population level information on the poststratifying variable with application to political polling. *J. Amer. Statist. Assoc.* **96** 1–11. MR1973779

SAPHIRE, D. (1984). *Estimation of Victimization Prevalence Using Data from the National Crime Survey*. Springer, New York.

SCOTT, J. (2005). SIPP2004+: Cross-Sectional Weighting Specifications for Wave 1. Memorandum WGT-20. U.S. Department of Commerce, U.S. Census Bureau, Washington, DC.

SKINNER, C. (2011). Log-linear Modelling with Complex Survey Data. Proceedings 58th World Statistical Congress, 2011, Dublin (Session IPS056).

SLAVKOVIĆ, A., ZHU, X. and PETROVIĆ, S. (2015). Fibers of multi-way contingency tables given conditionals: Relation to marginals, cell bounds and Markov bases. *Ann. Inst. Statist. Math.* **67** 621–648. MR3357932

STASNY, E. (1987). Some Markov-chain models for nonresponse in estimating Gross labor force flows. *J. Off. Stat.* **3** 359–373.

THIBAUDEAU, Y., SLUD, E. and GOTTSCHALCK, A. (2017). Supplement to "Modeling log-linear conditional probabilities for estimation in surveys." DOI:10.1214/16-AOAS1012SUPP.

U.S. BUREAU OF THE CENSUS (2009). SIPP User Guide's Sample Design and Interview Procedures Chapter 2. Available at http://www2.census.gov/programs-surveys/sipp/guidance/SIPP_2008_USERS_Guide_Chapter2.pdf.

WINKLER, W. (1993). On Dykstra's iterative fitting procedure. *Ann. Probab.* **18** 1410–1415. MR1062075

WOLTER, K. M. (1985). *Introduction to Variance Estimation*. Springer, New York. MR0799715

YBARRA, L. and LOHR, S. (2002). Estimates of repeat victimization using the national crime victimization survey. *J. Quant. Criminol.* **18** 1–21.

ZHANG, L.-C. and CHAMBERS, R. L. (2004). Small area estimates for cross-classifications. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **66** 479–496. MR2062389

U.S. CENSUS BUREAU
WASHINGTON, DC 20233
USA
E-MAIL: yves.thibaudeau@census.gov
eric.v.slud@census.gov
Alfred.O.Gottschalk@census.gov