Modeling Log-Linear Probabilities

Yves Thibaudeau[*] Eric Slud [†] Alfred Gottschalk[‡]

Abstract

It is shown how to model conditional probabilities subject to log-linear constraints and construct estimators that are more efficient than the Horvitz-Thompson weighted sum estimator for estimating population sizes in the context of a poststratified survey with unknown stratum sizes. Our approach rests on computing an "hybrid" predictor, similar to that of Pfeffermann et al. (1998), and the expansion of a specific parameterization for conditional probabilities restricted by log-linear constraints, as proposed by Thibaudeau (2003). This parameterization facilitates the computation of MLE's and makes it possible to apply the method of Laplace for variance estimation.

Key Words: Bayesian, parametric, *p*-value, Fisher

1. Introduction

We present a method for finite population prediction based on modeling log-linear conditional probabilities in the context of poststratified survey data. We show that the proposed method performs well when survey units can be stratified according to a highly discriminating poststratification, even if the sizes of the populations represented by the poststrata are unknown.

Cochran (1977, 134-135) defines poststratification essentially as a special case of stratification. He motivates stratification (pp. 89-90) as a device to create more homogeneous subpopulations, relative to the whole population. Predictors built on these homogeneous poststrata are likely to be more accurate than those not involving any post-stratification. Identifying a discriminating poststratification is a way to achieve this homogeneity objective.

Cochran also states that to take full advantage of stratification, the stratum totals must be known. Indeed, many authors have shown how to develop successful strategies for prediction based on poststratifications with known stratum totals. Such strategies can be either model based (Gelman, 2007; Little, 2007; Lu, Gelman, 2003) or model assisted (Sarndal et al., pp. 269-271. The limiting factor in these approaches is to find useful poststrata of known sizes. Typically, poststratifying variables are confined to basic demographic items such as age, race and sex. In these cases, stratum sizes are known because these items are recorded in a census of the entire population. But we cannot expect poststratifications based on such variables always to be sufficiently discriminating with respect to all the variables under study.

This paper explores prediction methods involving useful poststratifications which may not have known cell totals. We substitute for the unknown sizes of the strata design-based estimates arising from the survey itself, assuming that the design-based estimated stratum totals are unbiased. Having estimated postratum population totals, we carry through the same analysis that could be done with known totals. We focus on modeling log-linear conditional probabilities in the context of strictly

[*] U.S. Census Bureau, Washington, DC 20233

[†] U.S. Census Bureau, Washington, DC 20233

[‡] U.S. Census Bureau, Washington, DC 20233

categorical data. Our approach leads to defining "hybrid predictors." The predictors are hybrid in the sense that they combine a design-based component, defined through weighted sums, and a model-based component, namely MLE's of log-linear conditional probabilities. We present an example involving the hybrid predictor based on the Survey of Income and Program Participation.

We explore variance estimation methods for the hybrid predictor. Because one component of the predictor is design-based and another component is model-based, various strategies for estimating the variance are available. One strategy is to use a strictly design-based variance estimation method. In the context of SIPP, one such method is balanced repeated replication (BRR) (Wolter, 1985, pp. 110-145). Alternatively, an hybrid variance estimation strategy that matches the hybrid nature of the predictor is available. In this context, the variability of the hybrid predictor is explicitly decomposed according to three sources: The first source of variability is the design-based component. The second source is the model-based component. The third source is the coupling between these two. The total variance of the predictor is the sum of the two variances corresponding to the design-based and model-based components and the covariance.

The hybrid variance estimation approach is to estimate the variance corresponding to the design-based component using BRR, but to use a model-based method to estimate the variance corresponding to the model-based component. We experiment with such a method: the method of Laplace (Tierney, Kass, Kadane 1989; Tierney, Kadane 1986). The nominal advantage of such a method is its asymptotic properties, namely convergence in probability. The drawback is the large sample properties may not hold if the model is misspecified. We conduct a simulation to understand the behavior of either variance estimation procedure and ultimately make a recommendation as to which one is better.

The next section describes the type of prediction problems we address, as well as a perspective on the methods commonly used to solve such problems, together with our proposed method. Section 3 discusses in detail an instance of our method in the context of a longitudinal survey, the Survey of Income and Program Participation. Section 4 presents the notation and theoretical underpinnings of our method in this instance. Section 5 gives results comparing our prediction method with more traditional predictors based on weighted totals in terms of variance estimates. Section 6 recalls the method of Laplace for estimating the variance of the model-based component of our proposed predictor, and section 7 concludes with a summary and discussion.

2. Log-Linear Modeling in Complex Surveys

The keystone of our prediction strategy is the estimation of conditional probabilities, possibly under restrictions defined by high-dimensional log-linear constraints (Agresti, p. 215.) We give rigorous definitions for valid classes of such estimators, and for the predictors they can support, in later sections. In this section, we place the predictors and estimators within the broader perspective of survey-based predictions based on cross-classification. We identify four general approaches to prediction: design-based, model-based, model-assisted, and the hybrid approach. The last of these is the focus of this paper. The classification for a given predictor depends on whether the sizes of the poststrata are assumed to be known, and whether the sample design is explicitly taken into account through estimation weights.

We define a class of predictors that is the object of our study. Let $\boldsymbol{e} = (e_1, ..., e_Q)'$

and $\boldsymbol{f} = (f_1, ..., f_R)'$ be vectors of indices, where $e_q = 1, ..., E_q$ indexes categorical variable q and $f_r = 1, ..., F_r$ indexes categorical variable $Q + r$, for $q = 1, ..., Q$ and $r = 1, ..., R$. We consider the full cross-classification of $E_1 \times ... \times E_Q \times F_1 \times ... \times F_R$ multi-indexed categories. Let $y_{\boldsymbol{ef}}$ denote the population total of attribute y corresponding to category $(e_1, ..., e_Q, f_1, ..., f_R)'$. A subscript is replaced with the symbol '+', or a vector of subscripts with $+$, to indicate summation over all possible values of the corresponding index or vector of indices. For example, $y_{\boldsymbol{e}+} = \sum_{\boldsymbol{f}} y_{\boldsymbol{ef}}$ is the population total corresponding to marginal multi-indexed category $\boldsymbol{e}$. The objective is to predict $y_{\boldsymbol{e}+}$. The general form of the predictors considered in the paper is

$$\hat{y}_{\boldsymbol{e}+} = \sum_{\boldsymbol{f}} \left(\hat{P}_{\boldsymbol{e}|\boldsymbol{f}}\right) \hat{y}_{+\boldsymbol{f}} \tag{1}$$

In Eq. (1), $\hat{y}_{+\boldsymbol{f}}$ is either the known population total ($\hat{y}_{+\boldsymbol{f}} = y_{+\boldsymbol{f}}$) within category $\boldsymbol{f}$, or the Horvitz-Thompson (HT) weighted-sum estimator (Cochran, 1977, p. 259), for that same total. The term $\hat{P}_{\boldsymbol{e}|\boldsymbol{f}}$ is always an estimator for the proportion of the population in the domain defined by marginal category $\boldsymbol{f}$ that is also included in marginal category $\boldsymbol{e}$. This estimator may be restricted by log-linear constraints or not, and may be weighted or unweighted. If it is weighted, the weights are design-based, accounting for the complex probability sample. Table 1 gives a taxonomy, according to the literature on survey estimation, for the four cases where $\hat{y}_{+\boldsymbol{f}}$ is known or estimated and $\hat{P}_{\boldsymbol{e}|\boldsymbol{f}}$ is unweighted or weighted. This taxonomy applies to predictors that incorporate either unrestricted or restricted estimators.

The predictor represented in the bottom right corner cell of Table 1 integrates weights both in the estimation of the poststrata totals, $\hat{y}_{+\boldsymbol{f}}$, and in the estimation of the proportions $\hat{P}_{\boldsymbol{e}|\boldsymbol{f}}$. This is the "Info S" predictor discussed in Sarndal et al. (2005, pp. 53-56), in the context of compensating for nonresponse. This predictor is calibrated to the weighted sample, rather than to the population. In effect, it is an instance of implicitely estimating of the poststrata sizes to do prediction. We will focus on predictors with this feature in the context of the "hybrid" predictor, in the bottom left of Table 2. For this predictor weights are also used to estimate the sizes of the poststrata, but the model-based approach is applied for the estimation of the relational parameters.

It is useful to look at the same taxonomy in the situation where the realtional parameter is unrestricted, in which case it is only a ratio. In that case, a more specific taxonomy applies. (table 2). Then, the model-assisted predictor becomes the In the Known/Weighted cell of Table 2, The predictor (1) becomes the "post-stratified" or "separate ratio" estimator (Cochran p. 270), a model-assisted estimator based on the poststratification defined by marginal categories $\boldsymbol{f}$. In the Known/Unweighted (upper-left) cell of Table 2, (1) is the model-based ratio estimator. In the lower-right Unknown/Weighted cell of Table 2, (1) is the strictly design-based Horvitz-Thompson estimator.

This paper focuses on restricted hybrid predictors, as identified in Table 1. Pfeffermann et al. (1998) use a predictor of this type to forecast labor-force flows in the presence of measurement error. In their application, the model is logistic regression, while log-linear modeling is done here. Hybrid-type predictors may be practical when the population sizes $\hat{y}_{+\boldsymbol{f}}$ are not known and there is enough evidence of model validity to support model-based, unweighted estimates $\hat{P}_{\boldsymbol{e}|\boldsymbol{f}}$.

Table 1: Taxonomy for Predictors Based on Conditional Probabilities

	Parameter Estimation	
	Unweighted	Weighted
Poststrata Sizes Known	Model-Based	Model-Assisted
Poststrata Sizes Estimated	Hybrid	Design-Based

Table 2: Taxonomy for Predictors Based on Unrestricted Conditional Probabilities

	Parameter Estimation	
	Unweighted	Weighted
Poststrata Sizes Known	Model-Based	Separate Ratio
Poststrata Sizes Estimated	Hybrid	Horvitz-Thompson

With hybrid estimators, we want to know how much the estimator (1) is penalized for not knowing y_{+f} and using the HT estimator to approximate it. In later sections, we propose a method based on linearization to answer this question.

3. Application to the Survey of Income and Program Participation

For this expose we assume a nontrivial, but simple, structure to illustrate the necessity of estimated population totals and their contributions to variances. The Survey of Income and Program Participation (SIPP), a longitudinal survey conducted by the U.S. Census Bureau, is the source of our example.

SIPP measures the economic well being of the U.S. general population in relation to participation in government social programs, such as unemployment compensation and several programs of economic assistance. Participants in the survey are asked to report on their status in relation to several programs for each month they are in the cohort. But participants are interviewed only every four months in ”waves” or installments.

The longitudinal nature of SIPP makes it a prime candidate for specifying homogeneous poststrata: we expect relatively high within-stratum homogeneity when defining postrata through values of wave 1. For example, tables 3 and 4 show sample counts for the state of California for wave 1 and wave 2 of the 2004 panel. It gives a cross-classification of each screened respondent by two qualitative variables (employment status – employed vs. unemployed – and medical insurance coverage – insured vs. uninsured) for each wave.

Table 3: No College Degree

		Wave 2			
		Employed		Unemployed	
Wave 1		Covered	Not Covered	Covered	Not Covered
Employed	Covered	1252	44	8	4
	Not Covered	82	604	2	13
Unemployed	Covered	20	2	13	1
	Not Covered	4	40	1	25

Table 4: College Degree

		Wave 2			
		Employed		Unemployed	
Wave 1		Covered	Not Covered	Covered	Not Covered
Employed	Covered	1451	22	10	2
	Not Covered	38	181	0	6
Unemployed	Covered	12	2	11	0
	Not Covered	2	14	1	11

Modeling is expected to yield superior estimators in this example because each qualitative variable tends to be consistent across wave, and because the models will incorporate information from both waves in predicting wave 2 characteristics. By contrast, the HT estimator, which is the benchmark “traditional” design-based estimator, is based strictly on wave 2 information. In the notation of Section 2, the example has the following structure:

$$Q = 2\,,\quad R = 3\,,\quad \boldsymbol{e} = (j,k)\,,\quad \boldsymbol{f} = (l,m,n)$$

$$y_{\boldsymbol{e}\boldsymbol{f}} = y_{jklmn}\,,\quad y_{\boldsymbol{e}+} = y_{jk+++}\,,\quad y_{+\boldsymbol{f}} = y_{++lmn}\,,\quad \hat{P}_{\boldsymbol{e}|\boldsymbol{f}} = P_{jk\,|\,lmn}$$

Here y_{jklmn} denotes the population count of individuals with education n, labor status m at wave 1, coverage l at wave 1, labor status k at wave 2, and coverage j at wave 2. Educational level n takes value 1 for no college degree and 2 for college degree. Labor force 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. As before, a single “hat” over a population total $\hat{y}$ with appropriate subscripts denotes the Horvitz-Thompson estimator for that total.

For the example in this paper, the generic predictor (1) becomes:

$$\hat{\hat{y}}_{jk+++} = \sum_{l=1}^{2} \sum_{m=1}^{2} \sum_{n=1}^{2} \left(\hat{P}_{jk\,|\,lmn}\right) \hat{y}_{++lmn} \tag{2}$$

We also consider predictors obtained by confining the predictor in (1) to domains defined by education, defining

$$\tilde{y}_{jk++n} = \sum_{l=1}^{2} \sum_{m=1}^{2} \left(\hat{P}_{jk\,|\,lmn}\right) \hat{y}_{++lmn} \tag{3}$$

As defined in the next Section, $\hat{P}_{jk\,|\,lmn}$ is a model based estimator for $P_{jk\,|\,lmn}$, the proportion of individuals with coverage/labor status j,k at wave 2 among the population with education n and coverage/labor status l,m at wave 1. We will compare $\hat{\hat{y}}_{jk+++}$ and $\tilde{y}_{jk++n}$ with their HT counterparts, $\hat{y}_{jk+++}$ and $\hat{y}_{jk++n}$, respectively.

4. A Likelihood for Conditional Probabilities under Log-linear Constraints

We define a specific likelihood for the cross-classified survey counts with respect to parameters $P_{jk\;|\;lmn}$ for the purpose of deriving a maximum likelihood estimate,

$\hat{P}_{jk\,|\,lmn}$. The model likelihood is expressed in terms of "reference likelihoods" for the individual cell probabilities of the individual cells of tables 3 and **??**.

Let $\boldsymbol{x} = (x_{11111}, x_{21111}, ..., x_{22222})'$ be the vector of the counts, assumed multinomial, for the 32 cells defined by education and by medical coverage and labor force status at both waves in the sample, in the same order of subscripts as before. Denote the corresponding cell probabilities by $\boldsymbol{\pi} = (\pi_{11111}, \pi_{21111}, ..., \pi_{22222})'$, and let $\boldsymbol{\tau} = (\tau_{111}, \tau_{211}, ...\tau_{222})'$ be the vector of joint marginal probabilities for health insurance coverage and labor-force at wave 1, along with education . Also denote by $\boldsymbol{P} = \left(P_{11\,|\,111}, P_{21\,|\,111}, ..., P_{22\,|\,222}\right)$ the conditional probabilities of coverage and labor-force status at wave 2 conditional on coverage and labor-force status at wave 1 and education. Formally,

$$\tau_{lmn} \;=\; \pi_{++lmn} \;=\; \sum_{j,k=1}^{2} \pi_{jklmn} \tag{4}$$

and

$$P_{jk\,|\,lmn} \;=\; \pi_{jklmn} \,/\, \tau_{lmn} \tag{5}$$

With the object of modeling $\boldsymbol{P}$ realistically, we model $\boldsymbol{\pi}$ explicitly: the multinomial cell probabilities are assumed to follow the constraints of a standard hierarchical log-linear model (Bishop, Fienberg, Holland p. 34) with certain higher-order interactions suppressed. Doing so, we implicitly constrain $\boldsymbol{P}$. Denote by $\Omega_{\boldsymbol{\pi}}$, $\Omega_{\boldsymbol{\tau}}$, $\Omega_{\boldsymbol{P}}$ the respective parameter spaces for $\boldsymbol{\pi}$, $\boldsymbol{\tau}$, $\boldsymbol{P}$ Then we define $\Omega_{\boldsymbol{\pi}}$ explicitly, corresponding to a log-linear model that includes second-order and lower-order interactions for each pair of variables except pairs involving employment status and coverage between different waves. The other parameter spaces $\Omega_{\boldsymbol{\tau}}$, $\Omega_{\boldsymbol{P}}$ then inherit their meaning from $\Omega_{\boldsymbol{\pi}}$ and equations (4)–(5).

It turns out (Thibaudeau, 2003) that this model for $\boldsymbol{\pi}$ leaves $\boldsymbol{\tau}$ unconstrained, i.e.

$$\Omega_{\boldsymbol{\tau}} \;=\; \{\mathbf{v} \in \mathbf{R}^8 :\; v_1, \ldots, v_8 \geq 0 \quad \text{and} \quad v_1 + \cdots + v_8 = 1\}$$

while the vector of conditional probabilities $\boldsymbol{P}$ in (5) range freely over a 7-dimensional set regardless of τ. The suppression of interactions thus has the effect of reducing the 32-dimensional probability vector $\boldsymbol{\pi}$ which would contain 31 free parameters, to a 14-dimensional space, in such a way that $\Omega_{\boldsymbol{\pi}}$ is in one-to-one correspondence with $\Omega_{\boldsymbol{\tau}} \times \Omega_{\boldsymbol{P}}$. The likelihood of $\mathbf{x}$ with respect to parameters $\boldsymbol{\pi}$ factors:

$$L(\boldsymbol{x};\boldsymbol{\pi}) = \prod_{j,k,l,m,n=1}^{2} (\pi_{jklmn})^{x_{jklmn}}\, I_{[\boldsymbol{\pi}\in\Omega_\pi]} \;=\; L_\tau(\boldsymbol{x_f};\boldsymbol{\tau})\, L_P(\boldsymbol{x}\,|\,\boldsymbol{x_f}\,;\,\boldsymbol{P}) \tag{6}$$

where

$$L_\tau(\boldsymbol{x};\boldsymbol{\tau}) \;=\; \prod_{l,m,n=1}^{2} (\tau_{lmn})^{x_{++lmn}}\, I_{[\boldsymbol{\tau}\in\Omega_\tau]} \tag{7}$$

and

$$L_P(\boldsymbol{x};\boldsymbol{P}) \;=\; \prod_{j,k,l,m,n=1}^{2} \left(P_{jk\,|\,lmn}\right)^{x_{jklmn}}\, I_{[\boldsymbol{P}\in\Omega_P]} \tag{8}$$

To specify the parametric restrictions on $P_{jk\,|\,lmn}$ more formally, we develop a non-singular parameterization for $\boldsymbol{P}$ that reflects the log-linear constraints imposed through Ω_P. One source of non-degenerate parameterizations for log-linear models is Liu, Massam, Dobra (2009): they propose parametric expressions driven by the structure of the interactions involved in the log-linear model. Such parameterizations have meaningful interpretations, but do not naturally provide for a hierarchy of conditionality among the the variables of the model, as would be appropriate in this case. Instead we follow the parametrization of Thibaudeau (2003) specifically designed to reflect the order of conditioning for cell probabilities subject to log-linear constraints. This parameterization leads to a set of algebraically independent conditional probabilities for each "layer" of conditioning in the log-linear model. In our application there are two layers of conditioning, the top layer corresponding to coverage at wave 2 conditional on everything else, and a second layer for employment status at wave 2, conditional on everything else except coverage at wave 2. The parameters for the top layer of conditional probabilities in this setting are:

$$\begin{aligned}
\gamma_1 &= P_{11\,|\,111} \,/\, (P_{11\,|\,111} + P_{21\,|\,111}) \\
\gamma_2 &= P_{12\,|\,111} \,/\, (P_{12\,|\,111} + P_{22\,|\,111}) \\
\gamma_3 &= P_{11\,|\,211} \,/\, (P_{11\,|\,211} + P_{21\,|\,211}) \\
\gamma_4 &= P_{11\,|\,112} \,/\, (P_{11\,|\,112} + P_{21\,|\,112})
\end{aligned}$$

The parameters for the second layer of conditional probabilities are:

$$\begin{aligned}
\gamma_5 &= P_{11\,|\,111} \,/\, (P_{11\,|\,111} + P_{12\,|\,111}) \\
\gamma_6 &= P_{11\,|\,121} \,/\, (P_{11\,|\,121} + P_{12\,|\,121}) \\
\gamma_7 &= P_{11\,|\,112} \,/\, (P_{11\,|\,112} + P_{12\,|\,112})
\end{aligned}$$

The model parameterization for $\boldsymbol{P}$ is defined by these seven equations, with all γ_a ranging freely in the unit interval $[0,1]$. These seven degrees of freedom fully describe the allowed conditional probabilities for second-wave status given first-wave status and the education covariate. Appendix A gives the explicit mapping (12) from $\gamma = \{\gamma_a\}_{a=1}^{7}$ to $\boldsymbol{P}$ Substituting for $P_{jk\,|\,lmn}$ in terms of γ using (12), we can re-write down the likelihood factor (8) in the closed exponential-family form

$$\begin{aligned}
L_P(\boldsymbol{x};\boldsymbol{P}) &= L_\gamma\,(\boldsymbol{x};\boldsymbol{\gamma}) \\
&= \prod_{i=1}^{7} (\gamma_i)^{z_i}\,(1-\gamma_i)^{t_i-z_i} \prod_{l,m,n=1}^{2} (\Gamma_{lmn})^{-z_{l+2m+4n+1}}\, I_{[\boldsymbol{\gamma}\in[0,1]^7]}
\end{aligned}$$

where Γ_{lmn} is defined in (12)-(13) in Appendix A, and where $\boldsymbol{z} = (z_1, z_2, \ldots, z_{15})' \equiv \boldsymbol{A}'\boldsymbol{x}$ and $\boldsymbol{t} = (t_1, t_2, \ldots, t_7)' = \boldsymbol{B}'(z_8, z_9, \ldots, z_{15})'$ are defined in Appendix B. Here $\boldsymbol{z}$ is a sufficient statistic (for $\boldsymbol{P}$) defined from the cell counts $\boldsymbol{x}$, and $\boldsymbol{z}$ has full rank in the sense that the 32×15 matrix $\boldsymbol{A}$ does.

5. Using Conditional Probabilities for Prediction

To illustrate our method and some of its relative advantages, we study three specific predictors in the setting above. These predictors are $\tilde{y}_{12++1}$, $\hat{\hat{y}}_{12+++}$, as defined in Section 3, and $\tilde{\tilde{y}}_{2++++}$, where

$$\tilde{\tilde{y}}_{2++++} = \hat{\hat{y}}_{21+++} + \hat{\hat{y}}_{22+++} \tag{9}$$

We compare the variances of $\tilde{y}_{12++1}$, $\hat{\hat{y}}_{12+++}$, $\tilde{\tilde{y}}_{2++++}$ to those of their Horvitz-Thompson counterparts, $\hat{y}_{12++1}$, $\hat{y}_{12+++}$, $\hat{y}_{2++++}$. To do so, we apply Balance

Table 5: Hybrid and Horvitz-Thompson Estimation of Domain Totals

	Estimator			
	Horvitz-Thompson		Hybrid	
Domain	Estimate	Variance	Estimate	Variance
Not Covered - Total	3518150	2.823×10^{10}	3551468	2.552×10^{10}
Covered and Unemployed	178762	3.890×10^{8}	179133	4.031×10^{8}
Covered, Unemployed w/o College	84073	2.208×10^{8}	84038	1.697×10^{8}

Repeated Replication (Rao, Shao, 1996; Wolter, 1985), a common variance estimation procedure which the SIPP survey has been designed to enable. BRR is a nonparametric variance estimation procedure which does depend on parametric model assumptions for its validity. Our simulation in appendix C indicates BRR is relatively free of bias in this context. But at the same time, our simulation shows BRR is inacurate when the number of its degrees of freedom –the number of estimation strata– is small. In the case of the state of California, SIPP has only 23 estimation strata to produce variance estimators through replication. This is too small a number for comfort about the accuracy of BRR for estimating a given variance.

We are interested not so much in the variance of each predictor as in the difference between the variance of the hybrid and that of the HT estimator. Certainly, the difference between the BRR estimates of these two estimates is an unbiased estimator for the actual difference. Furthermore the difference between the two BRR estimates will be substantially more accurate than the BRR estimates themselves if the later are substantially correlated. Assuming that is the case, we retain the difference between the BRR variance estimate of the hybrid and that of the HT estimator as a measure for the relative performance of the hybrid and the HT estimator.

There appears not to be a substantial difference between hybrid predictors and their corresponding HT estimators in terms of variance, for the first two populations. However, the difference between hybrid and HT is large for prediction confined to the domain of non-college-degree holders. This is no surprise as hybrid prediction makes use of information retrieved from all poststrata, including from the poststrata involving college-degree holders. The HT estimator, on the other hand, does not benefit from a model to exploit the information available from all the post-strata, using only information retrieved from the post-strata involving non-college-degree holders.

Beyond the gain in efficiency from using a model to predict the sizes of small domain populations, our method is useful in a variety of context. To show this we decompose the variance of the hybrid predictor defined in (1) into the sum of a model-based component and a design-based component and a covariance. The model-based variance component stems from the uncertainty about the model parameters, as estimated by their MLE's. The design-based variance component reflects the sampling variance inherent to the design and its impact on the HT weighted totals involved in the hybrid predictor. The total variance of the hybrid predictor is approximately the sum of these two variance components discounted by the coupling between them.

This decomposition of the variance can be carried through using BRR to estimate each variance component, as exhibited in table 5. This decomposition of the

Table 6: Linearization of the Variance of the Hybrid Predictor

	Replication		
Variance Component	Linearized	Direct	Method of Laplace
Model-Based	1.553×10^8	N/A	2.268×10^8
Design-Based	2.039×10^7	N/A	N/A
Covariance	-1.863×10^6	N/A	N/A
Total Variance	1.720×10^8	1.697×10^8	N/A

variability is meaningful as the variability of the design-based component provides a measure of the "cost" of having to estimate the strata totals with weighted sums. In our example, the variability component of the variance due to design-based estimation is less than Ten percent of the total variance. What's more, the covariance between the model-based and design-based components is negative. This suggests not much is lost from not knowing the sizes of the poststrata.

6. Model-Based Variance Estimation and Model Selection with the Method of Laplace

The shortcomings of BRR leaves us with a desire for another type of technique for variance estimation. The method of Laplace is a model-based technique that can be used for estimating the variance of the model-based component of a predictor. The following result is an extension of Theorem 1 in Tierney, Kass and Kadane (1989).

Let $M(\hat{\boldsymbol{\gamma}};\boldsymbol{x})$ be a smooth function of all its arguments. An approximation for the variance of the model-based component of $M(\hat{\boldsymbol{\gamma}};\boldsymbol{x})$ is

$$V^L\left[M(\boldsymbol{\gamma};\boldsymbol{x})\,|\boldsymbol{x}\right] = \lim_{c\to\infty} c\left|M(\hat{\boldsymbol{\gamma}}_c;\boldsymbol{x}) - M(\hat{\boldsymbol{\gamma}};\boldsymbol{x})\right| \tag{10}$$

where $\hat{\boldsymbol{\gamma}}$ is the MLE of $\boldsymbol{\gamma}$ and

$$\hat{\boldsymbol{\gamma}}_c = \max_{\boldsymbol{\gamma}} \arg\left[\left(M(\boldsymbol{\gamma};\boldsymbol{x}) + c\right) L_\gamma(\boldsymbol{x};\boldsymbol{\gamma})\right] \tag{11}$$

The approximation in (10) is a low order asymptotic approximation derived from the version of the method of Laplace for approximating posterior variances in a Bayesian context, as propoposed in Tierney and Kadane (1986). The approximation for the posterior variance of in turns serves as an approximation for the frequentist variance of an functional derived from maximum likelihood estimation.

A relatively large discrepancy between the BRR variance estimates of the parameters and the estimates obtained through the method of Laplace would suggest the model is not appropriate. Based on our simulation results in Table 9 (appendix C) , the discrepancies between BRR and the method of Laplace in Table 7 are not unexpectedly large. Again, the simulation suggests that the culprit is BRR. The standard deviations for the BRR variance estimate are substantially larger than for the method of Laplace, when the model is correct.

The method of Laplace can also serve to test specific hypotheses about the parametric structure of the model, thereby serving as a tool for model selection. One such hypothesis of interest is whether or not the information on education is useful to predict health coverage, given coverage and employment status at the previous wave are known, and unemployment status at the current wave is also known. This is equivalent to testing the hypothesis: $H_0 : \gamma_4 - \gamma_1 = 0$. Similarly,

Table 7: Model Parameters: Estimates and Variances

		Variance		
Parameter	Estimate	Replication	Method of Laplace	Significant
γ_1	.9665	1.785×10^{-5}	2.080×10^{-5}	–
γ_2	.9168	4.143×10^{-4}	3.672×10^{-4}	–
γ_3	.1142	8.869×10^{-5}	1.174×10^{-4}	–
γ_4	.9806	6.756×10^{-6}	7.861×10^{-6}	–
γ_5	.9895	2.407×10^{-6}	4.312×10^{-6}	–
γ_6	.7976	2.252×10^{-3}	1.624×10^{-3}	–
γ_7	.9882	5.187×10^{-6}	5.060×10^{-6}	–
$\gamma_4 - \gamma_1$	.01414	1.633×10^{-5}	1.2228×10^{-5}	Yes
$\gamma_7 - \gamma_5$	-.001357	6.125×10^{-6}	7.957×10^{-6}	No
Hybrid - HT	-35	2.765×10^{4}	–	No

we can test whether or not education is useful to predict employment status, given the rest of the variables at wave 1 and 2, by testing $H_{00} : \gamma_7 - \gamma_5 = 0$.

Table 7 gives the variance obtained through the method of Laplace for the test statistics $\hat{\gamma_4} - \hat{\gamma_1}$, and $\hat{\gamma_7} - \hat{\gamma_5}$ for H_0 and H_{00} respectively. The results suggest that education does significantly improve the prediction of coverage, even in presence of the previous wave information and the current employment status. But, it appears there is little basis for including an interaction term between education and employment status when the rest of the variables in the model, and their interactions with employment status as given, are present.

7. Discussion

The paper investigated the properties of predictors integrating model-based features and estimated strata sizes. The sizes are etimated from the same survey the predictors are derived from. This approach is driven by the goal of selecting poststrata primarily on the basis of their discriminatory power, rather than on the basis of extraneous knowledge of poststrata sizes.

Our results suggest that a good model and a good design-based weighted-total estimation —both of these together— is a combination that is hard to beat for prediction. Table 5 indicates a substantial variance reduction when using the hybrid approach. In addition, the decomposition of the variance in Table 5 suggests the cost of estimating the totals involved in the hybrid approach is small, in terms of total variance.

These results indicate the desirability of having known poststrata totals to derive predictors based on poststratification may be overplayed. Of course not every prediction situation is as easy to model as that in the paper. But, the cost of estimating poststrata totals may be modest in other situations as well. Then the statistician should entertain more discriminatory poststratifications, as compared to rigid postratifications guided only by the insistance on knowing the sizes of the poststrata.

We also note that in the context of the specific predictions considered in the paper we need not have full model validity, in the sense of the best fitting model. Our basic requirement is that the predictors involving components derived from a model are unbiased. In the case of predicting the small domain, we were able to test agains biases from our model-based estimator, relative to the Horvitz Thompson estimator (table 7). The absence of bias, along with a lower variance than that of Horvitz-Thompson estimator guarantees the model-based estimator is desirable.

We explored two different approach for estimating the model-based variance. That is the variance of the sampling noise filtered through the maximum likelihood estimators. A simulation suggests BRR and the method of Laplace are unbiased for the model-based variance, when the model is correct (table 8). But, for a moderately large sample, BRR is substantially less stable than the method of Laplace (table 9). The natural question is do these results extend to cases when the model in incorrect. In other words, is the method of Laplace robust to some model specifications? This shall be the object of additional research.

A. Parameterization for Log-Linear Conditional Probabilities

The conditional probabilities $P(jk \,|\, lmn)$ are uniquely determined by the parameters γ_a, $a = 1, \ldots, 7$, defined in Section 4, according to the following scheme developed in Thibaudeau (2003), with $\{(jk)\}_{j,k=1}^{2} = \{11, 21, 12, 22\}$:

$$
\begin{aligned}
\{P_{jk\,|\,111}\}_{j,k=1}^{2} &= \Big(\gamma_1\gamma_2\gamma_5,\, (1-\gamma_1)\gamma_2\gamma_5,\, \gamma_1\gamma_2(1-\gamma_5),\, \gamma_1(1-\gamma_2)(1-\gamma_5)\Big)\,/\,\Gamma_{111}\\
\{P_{jk\,|\,211}\}_{j,k=1}^{2} &= \Big((1-\gamma_1)\gamma_2\gamma_3\gamma_5,\, (1-\gamma_1)\gamma_2(1-\gamma_3)\gamma_5,\, (1-\gamma_1)\gamma_2\gamma_3(1-\gamma_5),\\
&\quad \gamma_1(1-\gamma_2)(1-\gamma_3)(1-\gamma_5)\Big)\,/\,\Gamma_{211}\\
\{P_{jk\,|\,121}\}_{j,k=1}^{2} &= \Big(\gamma_1\gamma_2\gamma_6,\, (1-\gamma_1)\gamma_2\gamma_6,\, \gamma_1\gamma_2(1-\gamma_6),\, \gamma_1(1-\gamma_2)(1-\gamma_6)\Big)\,/\,\Gamma_{121}\\
\{P_{jk\,|\,221}\}_{j,k=1}^{2} &= \Big((1-\gamma_1)\gamma_2\gamma_3\gamma_6,\, (1-\gamma_1)\gamma_2(1-\gamma_3)\gamma_6,\, (1-\gamma_1)\gamma_2\gamma_3(1-\gamma_6),\\
&\quad \gamma_1(1-\gamma_2)(1-\gamma_3)(1-\gamma_6)\Big)\,/\,\Gamma_{221}\\
\{P_{jk\,|\,112}\}_{j,k=1}^{2} &= \Big((1-\gamma_1)\gamma_2\gamma_4\gamma_7,\, (1-\gamma_1)\gamma_2(1-\gamma_4)\gamma_7,\, (1-\gamma_1)\gamma_2\gamma_4(1-\gamma_7),\\
&\quad \gamma_1(1-\gamma_2)(1-\gamma_4)(1-\gamma_7)\Big)\,/\,\Gamma_{112}\\
\{P_{jk\,|\,212}\}_{j,k=1}^{2} &= \Big((1-\gamma_1)^2\gamma_2\gamma_3\gamma_4\gamma_7,\, \gamma_1(1-\gamma_1)\gamma_2(1-\gamma_3)(1-\gamma_4)\gamma_7,\, (1-\gamma_1)^2\gamma_2\gamma_3\gamma_4(1-\gamma_7),\\
&\quad \gamma_1^2(1-\gamma_2)(1-\gamma_3)(1-\gamma_4)(1-\gamma_7)\Big)\,/\,\Gamma_{212}\\
\{P_{jk\,|\,122}\}_{j,k=1}^{2} &= \Big((1-\gamma_1)\gamma_2\gamma_4(1-\gamma_5)\gamma_6\gamma_7,\, (1-\gamma_1)\gamma_2(1-\gamma_4)(1-\gamma_5)\gamma_6\gamma_7,\\
&\quad (1-\gamma_1)\gamma_2\gamma_4\gamma_5(1-\gamma_6)(1-\gamma_7),\, \gamma_1(1-\gamma_2)(1-\gamma_4)\gamma_5(1-\gamma_6)(1-\gamma_7)\Big)\,/\,\Gamma_{122}\\
\{P_{jk\,|\,222}\}_{j,k=1}^{2} &= \Big((1-\gamma_1)^2\gamma_2\gamma_3\gamma_4(1-\gamma_5)\gamma_6\gamma_7,\, \gamma_1(1-\gamma_1)\gamma_2(1-\gamma_3)(1-\gamma_4)(1-\gamma_5)\gamma_6\gamma_7,\\
&\quad (1-\gamma_1)^2\gamma_2\gamma_3\gamma_4\gamma_5(1-\gamma_6)(1-\gamma_7),\\
&\quad \gamma_1^2(1-\gamma_2)(1-\gamma_3)(1-\gamma_4)\gamma_5(1-\gamma_6)(1-\gamma_7)\Big)\,/\,\Gamma_{222}
\end{aligned}
\tag{12}
$$

where Γ_{lmn} is defined implicitly through (12) and for all values of l, m, n,

$$
\sum_{j,k=1}^{2} P_{jk\ |\ lmn} = 1 \tag{13}
$$

B. Sufficient Statistic

The sufficient statistics z and totals t appearing in the likelihood (9) are defined here, again following the approach of Thibaudeau (2003), as

$$\boldsymbol{z} = (z_1, z_2, \ldots, z_{15})' = \boldsymbol{A}'\boldsymbol{x} \quad , \qquad \boldsymbol{t} = (t_1, t_2, \ldots, t_7)' = B'(z_8, z_9, \ldots, z_{15})'$$

where

$$\boldsymbol{A} = \begin{pmatrix}
1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
2 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{pmatrix} \qquad (14)$$

and

$$\boldsymbol{B} = \begin{pmatrix}
1 & 1 & 0 & 0 & 1 & 0 & 0 \\
1 & 1 & 1 & 0 & 1 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 1 & 0 \\
1 & 1 & 1 & 0 & 0 & 1 & 0 \\
1 & 1 & 0 & 1 & 0 & 0 & 1 \\
2 & 1 & 1 & 1 & 0 & 0 & 1 \\
1 & 1 & 0 & 1 & 1 & 1 & 1 \\
2 & 1 & 1 & 1 & 1 & 1 & 1
\end{pmatrix} \qquad (15)$$

Table 8: Simulation: Variance Estimators for MLE's

Parameter	Actual	BRR	Method of Laplace
$\gamma 1$	1.978×10^{-5}	2.048×10^{-5}	1.925×10^{-5}
$\gamma 2$	3.456×10^{-4}	3.844×10^{-4}	3.731×10^{-4}
$\gamma 3$	1.205×10^{-4}	1.199×10^{-4}	1.190×10^{-4}
$\gamma 4$	8.199×10^{-6}	8.253×10^{-6}	8.179×10^{-6}
$\gamma 5$	3.109×10^{-6}	3.476×10^{-6}	3.426×10^{-6}
$\gamma 6$	2.243×10^{-3}	2.373×10^{-3}	2.346×10^{-3}
$\gamma 7$	4.026×10^{-6}	4.286×10^{-6}	4.247×10^{-6}

Table 9: Simulation: Standard Deviations of BRR and the Method of Laplace

Parameter	BRR	Method of Laplace	Difference
$\gamma 1$	6.388×10^{-6}	2.036×10^{-6}	6.210×10^{-6}
$\gamma 2$	2.144×10^{-4}	1.644×10^{-4}	1.356×10^{-4}
$\gamma 3$	3.74×10^{-5}	1.201×10^{-5}	3.504×10^{-5}
$\gamma 4$	3.063×10^{-6}	1.576×10^{-6}	2.560×10^{-6}
$\gamma 5$	1.485×10^{-6}	9.579×10^{-7}	1.110×10^{-6}
$\gamma 6$	8.348×10^{-4}	4.133×10^{-4}	7.271×10^{-3}
$\gamma 7$	1.775×10^{-6}	1.065×10^{-6}	1.372×10^{-6}

C. Simulation

We conduct a simulation based on the model described in Section 4. The MLE's derived from the SIPP data is substituted as the "true parameters" in the model for purpose of simulation. Our main goals is to get an idea of the biases and variances of both BRR and the method of Laplace when used to estimate the variance of the MLE's. 1000 samples of the same size as the observed data (3800) were drawn from the multinomial probability function prescribed by the model.

Table 8 shows the expected value of both methods against the true value of the variances for the MLE's of each conditional probability. Both method are fairly bias free and there is no winner or loser. However, Table 9 shows the standard deviations for both variance estimators. The method of Laplace is the clear winner, as BRR has standard deviations anywhere from 30 percent to 200 percent larger. This should not come as a surprise since the simulation implicitely assumes the model is correct. How robust the method of Laplace is under model misspecification remains an object of speculation. In revenge, BRR is nonparametric and in theory is unbiased regardless of any model assumptions.

Another purpose of the simulation was to validate the discrepancies between BRR and the method of Laplace, as observed in Table 7. In that respect, Table 9 displays the standard deviation of the differences between BRR and the method of Laplace. Under that light, the discrepancies observed in Table 7 are not excessive. Again, the results of the simulation suggest the main culprit for these discrepancies is the instability of the BRR estimates. Essentially, the variability of the method of Laplace is the order of a constant relative to the variability of BRR.

REFERENCES

Agresti, A. (2007), An Introduction to Categorical Data Analysis, 2nd Ed. Wiley.

Binder, D. (1983), On the Variances of Asymptotically Normal Estimators from Complex Surveys.

International Statistical Review. 51.

Bishop Y., Fienberg, S., Holland, P. (1975), Discrete Multivariate Analysis. MIT Press.

Cochran, W. (1977), Sampling Techniques, 3rd Ed. Wiley.

Fienberg, S. (2009), JSM Short Course: Analysis of Cross-Classified Categorical Data - Some Examples of Contingency Tables.

Gelman, A. (2007), Struggles with Survey Weighting and Regression Modeling. Statistical Science, 22, 2.

Little, R. (2007), Post-Stratification: A Modeler's Perspective. Journal of the American Statistical Association, 88, 423.

Lu, H., Gelman, A. (2003), A Method for Estimating Design-based Sampling Variances for Surveys with Weighting, Poststratification and Raking. Journal of Official Statistics, 19, 2.

Massam, H., Liu, J., Dobra, A. (2009), A Conjugate Prior for Discrete Hierarchical Log-Linear Models. Annals of Statistics, 37, 6A.

Pfeffermann, D., Skinner, C., Humphreys, K. (1998), The Estimation of Gross Flows in the Presence of Measurement Error Using Auxiliary Variables. JRSS A, 161, 1.

Sarndal, C.-E., Swensson, B., Wretman, J. (1992), Model Assisted Survey Sampling. Springer.

Sarndal, C.-E., Lundstrom, L. (2005), Estimation in Surveys with Nonresponse. Wiley.

Slud, E. V., Thibaudeau, Y. (2010), Simultaneous Calibration and Nonresponse Adjustment. Research Report 2010/03, Statistical Research Division, U.S. Census Bureau.

Thibaudeau, Y. (2003), An Algorithm for Computing Full Rank Minimal Sufficient Statistics with Applications to Confidentiality Protection. Monographs of Official Statistics, Work Session on Statistical Data Confidentiality, Luxembourg, 7 to 9 April 2003, Part 1. pp. 45-58.

Thibaudeau, Y. (2002), Model Explicit Item Imputation for Demographic Categories. Survey Methodology, 28, 2.

Tierney, L., Kass, R., Kadane, J. (1989), Fully Exponential Laplace Approximations to Expectation and Variances of Nonpositive Functions. Journal of the American Statistical Association, 84, 407.

Tierney, L., Kadane, J. (1986), Accurate Approximations for Posterior Moments and Marginal Densities. Journal of the American Statistical Association, 81, 393.

Wolter, K. (1985), Introduction to Variance Estimation. Springer-Verlag.