Smooth Post-Stratification in Multiple Capture-Recapture
Thesis Proposal

Zachary Kurtz
Department of Statistics
Carnegie Mellon University
zkurtz@stat.cmu.edu
Version: August 27, 2012

Thesis Committee:

William F. Eddy (chair)
Steve Fienberg
Cosma Shalizi
Rebecca Steorts
Bruce Spencer (Northwestern University)

Abstract

Capture-recapture (CRC) is a way to estimate the size of a population by combining multiple incomplete lists of population units. For the two-list scenario, the oldest and simplest estimator is the Petersen estimator, which assumes that the event that a unit is captured on the first list is independent of the event that a unit is captured on the second list. Because this assumption is usually false, the Petersen estimator is biased for most applications.

Literature on overcoming the bias in the Petersen estimates tends to fall into one of two groups. The first group of models expresses capture probabilities as functions of capture pattern in ways that allow for complex interactions between lists in the aggregate – without explicitly modeling covariate effects. The second group of models regresses capture probabilities conditional on covariates. Post-stratification is a discrete way to condition on covariates. However, continuous generalizations (i.e., *smooth* post-stratification models) typically assume a strong form of independence between lists which is not optimal in the case of three or more lists.

We combine these two lines of work to produce an estimator that models complex list interactions locally. Our procedure begins with estimating the conditional distribution of capture pattern as a smooth function of the covariates. We extend this estimated conditional distribution to the unobserved capture pattern (no captures) by applying a log-linear model locally. A Horvitz-Thompson style population estimator is then immediate. We intend to demonstrate our method by combining a list of records from the 2010 Census, the 2010 Census Coverage Measurement (CCM) survey, and the American Community Survey (ACS).

Contents

1 Introduction

1.1 Problem Statement

Capture-recapture (CRC) is a way to estimate the size of the population by combining information from multiple incomplete lists of population units. A list is a collection of units (i.e., people, or animals). We refer to the act of generating a list as a *capture*. In the simplest CRC setting, we are given two lists of units, List 1 and List 2. Assuming that units captured on both lists are perfectly matched across lists, it is possible to find the cross-classification of units according to list membership as displayed in Table 1.

Table 1:

		List 2	
		yes	no
List 1	yes	c_{11}	c_{10}
	no	c_{01}	c_{00}

Each term c_{ij} denotes the count of units that have capture pattern (i, j). For example, c_{10} is the number of units that appear on List 1 but do not appear on List 2. The number of units that are not observed on either list, c_{00}, is not observable, so estimating the population size is the same as estimating c_{00}. With three lists, the task is to estimate c_{000}. This problem becomes more interesting as the number of lists grows and as covariates for observed units become available.

For the two-list scenario, the oldest estimator is the Petersen estimator, which assumes that the event that a unit is captured on the first list is independent of the event that a unit is captured on the second list. Because this assumption is usually false, the Petersen estimator is biased for most applications.

Literature on overcoming the bias in the Petersen estimates tends to fall into one of two groups. The first group of models expresses capture probabilities as functions of capture pattern in ways that allow for complex interactions between lists in the aggregate – without explicitly modeling covariate effects. The second group of models regresses capture probabilities conditional on covariates. Post-stratification is a discrete way to condition on covariates. However, continuous generalizations (i.e., *smooth* post-stratification models) typically assume a strong form of independence between lists which is not optimal in the case of three or more lists.

We combine these two lines of work to produce an estimator that models complex list interactions locally. Our procedure begins with estimating the conditional distribution of capture pattern as a smooth function of the covariates. We extend this estimated conditional distribution to the unobserved capture pattern (no captures) by applying a log-linear model locally. A Horvitz-Thompson style population estimator is then immediate. We intend to demonstrate our method by combining a list of records from the 2010 Census, the 2010 Census Coverage Measurement (CCM) survey, and the American Community Survey (ACS).

The remainder of this introduction discusses the importance of CRC for the U.S. Census, presents general notation, defines some common assumptions, describes our high-level conceptualization of the problem, and concludes with a simulation example to illustrate fundamental CRC concepts.

1.2 U.S. Census

Assessing the accuracy – or *coverage* – of the U.S. Census is an important application of CRC theory. The U.S. Census Bureau conducted a formal census coverage evaluation after every decennial census starting in 1980. The name of this formal evaluation changed with each new census; for the 2010 Census, it was called the Census Coverage Measurement (CCM).

Several sources of error affect the accuracy of a census. Most of these error sources can be understood as a contribution to either an *overcount* or an *undercount*. An example of an overcount is when a college student is counted both at college and at the parents' home. On the other hand, an undercount occurs when people are missed by the census. Both the overcount and the undercount rates in the 2010 Census were estimated to be around 5%, and the CCM program concluded that the 2010 census was in error by less than 0.1% overall, although the error rate was higher for specific demographic subgroups [1].

Estimation of the undercount rate in the CCM (and in previous coverage evaluations) relied on CRC methods. In particular, the CCM combined two lists using a CRC estimator to generate estimates of the population in a cluster sample of census blocks. The undercount rate was taken as one minus the census count as a fraction of the CRC population estimate.

The two lists used for the CRC estimator are called the *E-sample* and the *P-sample*. Within a geographic cluster sample of census blocks, the collection of all census enumerations is analyzed to identify and remove erroneous enumerations [2]. The resulting edited list of census enumerations is the E-sample. The post-enumeration survey, or P-sample, is a fresh attempt at listing all people living in the same census blocks that were selected for the E-sample.

Individuals from the E-sample are matched to individuals from the P-sample, resulting in a table of counts of the observable capture patterns as in Table 1. Many kinds of CRC estimators exist for imputing the missing cell c_{00}, and they all require making a strong assumption about the relationship between the two lists. A particularly desirable condition is to have the two lists be independent, in the sense that the event that an individual appears in the E-sample is independent of the event that the same individual appears in the P-sample.

Therefore, the P-sample is specifically designed to be independent from the E-sample. A small random selection of Census block clusters defines the geographic area of the E-sample. Within these block clusters, a list of all housing units is generated independently of the main Census housing unit list. Interviewers visit each listed housing unit in the selected block clusters, generating the P-sample as an all-new census. The desire for independence between the E-sample and P-sample means that the timing of the two samples is a sensitive matter. Collecting the P-sample too soon risks introducing interaction effects between the E-sample and the P-sample. Waiting too long to collect the P-sample increases the effects of an open population, as people migrate, reproduce, and die in between the two samples. The 2010 Census targeted April 1 as the E-sample survey date, and the 2010 P-sample survey took place some months afterwards, from August to October.

A major shift in the Census' methodology for coverage evaluation occurred between 2000 and 2010. Prior to 2010, the primary tool was the Petersen estimator in conjunction with post-stratification (Sections 1.5 and 2.3). The 2010 CCM was the first formal coverage evaluation to apply logistic regression, a smooth generalization of post-stratification (Sections 2.3 and 2.4).

Coverage evaluations in 2020 and beyond may incorporate more than two lists. A key reason for including a third list is to reduce bias resulting from the assumption of independence between the E-sample and P-sample. (Despite the emphasis on independence in generating the P-sample, it is generally not possible to verify that independence between the lists is achieved.) The Census Bureau conducted a "dress-rehearsal" study in St. Louis in 1988 to prepare for the 1990 Census coverage evaluation. In this study, the P-sample was supplemented with an *A-sample*, a compilation of records based on Employment Security, driver's license, Internal Revenue Service, Selective Service, and Veteran's Administration registrants [3]. Although the A-sample was originally viewed only as a supplement to the P-sample, several authors throughout the 1990's explored triple system estimation, viewing the A-sample as a list in its own right [4] [5] [6].

Much CRC theory began in the context of estimating animal populations in biological studies. For application to human populations, CRC is often referred to as dual system estimation (DSE) or triple system estimation in the case of three lists.

1.3 Notation

1.3.1 Data

A CRC experiment produces k different lists $L_1, ..., L_k$ of units from a population of size n. Let $i = 1, ..., n_c$ index the set of units that are captured at least one time, $\cup_j L_j$. Let $\mathcal{N} = \{1, ..., n\}$. For each $i \in \mathcal{N}$, let $m_i := I(i \in \cup_j L_j)$ so that $n_c = \sum_{i=1}^{n} m_i$. We do not distinguish units from their indices when discussing the lists; the ith unit is in list L_j if and only if $i \in L_j$.

For each unit i and list L_j, let $y_{ij} = I(i \in L_j)$. Then $y_{i\cdot} = (y_{i1}, ..., y_{ik})$, and $y_{\cdot\cdot}$ is the $n \times k$ matrix with ith row $y_{i\cdot}$. The vector $y_{i\cdot}$ is called the *capture pattern* of the ith unit. Let $x_{i\cdot}$ denote a $1 \times q$ vector of covariates associated with the ith unit, and $x_{\cdot\cdot}$ is the $n \times q$ matrix with ith row $x_{i\cdot}$. For each $i > n_c$, the pair $(x_{i\cdot}, y_{i\cdot})$ is not observed. If $x^c_{\cdot\cdot}$ is the matrix formed by the first n_c rows of $x_{\cdot\cdot}$, and $y^c_{\cdot\cdot}$ is the matrix formed by the first n_c rows of $y_{\cdot\cdot}$, then the [observable] data consists of the pair of matrices $(x^c_{\cdot\cdot}, y^c_{\cdot\cdot})$. We will refer to the pair $(x_{\cdot\cdot}, y_{\cdot\cdot})$ as the *extended* data.

Let $\mathcal{Y}_k$ denote the set of binary row vectors of length k. For example, $\mathcal{Y}_2 = \{(1,1), (1,0), (0,1), (0,0)\}$. Note that each $y_{i\cdot}$ is an element of $\mathcal{Y}_k$. For every $\mathbf{y} \in \mathcal{Y}_k$, define $c_{\mathbf{y}} := |\{i : y_{i\cdot} = \mathbf{y}\}|$. Then the array $\mathbf{c} := \{c_{\mathbf{y}}\}_{\mathbf{y} \in \mathcal{Y}_k}$ is the

contingency table of counts of units in the lists $L_1, ..., L_k$. In particular, $c_{\mathbf{0}} = n - n_c$, the number of units that are not observed, and any estimate $\hat{n}$ of n implies a prediction $\hat{c}_{\mathbf{0}}$ of $c_{\mathbf{0}}$ such that $\hat{n} = \hat{c}_{\mathbf{0}} + n_c$.

1.3.2 Subscripting by ω

Let $\mathcal{K} = \{1, ..., k\}$. Let Ω_k denote the power set of $\mathcal{K}$, excluding the empty set. For example, $\Omega_3 = \{\{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}$.

Let $\omega \in \Omega_k$, and suppose $|\omega|$ denotes the size of ω. Let $(\omega_{(1)}, ..., \omega_{(|\omega|)})$ denote the vector of elements of ω arranged in increasing order. Pick arbitrary $i \in \{1, ..., n\}$ and $\omega \in \Omega_k$. Define $y_{i\omega} := (y_{i\omega_{(1)}}, ..., y_{i\omega_{(|\omega|)}})$. To be clear, $y_{i\omega}$ is a vector with elements taken from the ith row of the matrix $y_{\cdot\cdot}$ as specified by ω.

More generally, for any vector $\mathbf{y} = (\mathbf{y}_1, ..., \mathbf{y}_k) \in \mathcal{Y}_k$, let $\mathbf{y}_\omega := (\mathbf{y}_{\omega_{(1)}}, ..., \mathbf{y}_{\omega_{(|\omega|)}})$. For the special case in which ω is a singleton $\{j\}$, we write $y_{i\{j\}} = y_{ij}$ and $\mathbf{y}_{\{j\}} = \mathbf{y}_j$. Take ω^c to be the complement of ω. For example, let $\mathbf{y} = (1, 1, 0) \in \mathcal{Y}_k$. Then $\mathbf{y}_{\{2,3\}} = (1, 0)$, $\mathbf{y}_1 = \mathbf{y}_{\{1\}} = \mathbf{y}_{\{2,3\}^c} = 1$, and $\mathbf{y}_{\{1,2,3\}} = \mathbf{y}$.

1.3.3 Probability

Each capture pattern $y_{i\cdot}$ may be assumed to be a realization of a random vector $Y_{i\cdot}$. Then, the matrix $y_{\cdot\cdot}$ is a realization of a random matrix $Y_{\cdot\cdot}$. The corresponding statistics $\mathbf{c}$ and m_i are realizations of the implied random quantities $\mathbf{C}$ and M_i. Subscripting for each of the random quantities works exactly analogously to subscripting for the fixed realizations. For the remainder of this section, fix $k > 1$, and let $j \in \mathcal{K}$, $i \in \mathcal{N}$, $\mathbf{y} \in \mathcal{Y}_k$, and $\omega \in \Omega_k$ be arbitrary.

Let $p(i, \mathbf{y}) = P(Y_{i\cdot} = \mathbf{y})$, the probability that unit i has capture pattern $\mathbf{y}$. Then $p(i, y_{i\cdot}) = P(Y_{i\cdot} = y_{i\cdot})$. Similarly, let $p_\omega(i, \mathbf{y}) = P(Y_{i\omega} = \mathbf{y}_\omega)$. Define $\mathbf{p}(i, \mathcal{Y}_k) := \{p(i, \mathbf{y})\}_{\mathbf{y} \in \mathcal{Y}_k}$.

Let $p_\omega(\mathbf{y}) = n^{-1} \sum_{i \in \mathcal{N}} p_\omega(i, \mathbf{y})$. If $\omega = \mathcal{K}$, then we have $p_\omega(\mathbf{y}) = p(\mathbf{y})$, the average probability that a unit has the capture pattern $\mathbf{y}$. Define $\mathbf{p}(\mathcal{Y}_k) := \{p(\mathbf{y})\}_{\mathbf{y} \in \mathcal{Y}_k}$.

Let $\mathbf{0}_\omega$ denote the zero vector of length $|\omega|$. Let $\phi_\omega(i) = 1 - P(Y_{i\omega} = \mathbf{0}_\omega)$, the probability that the ith unit is on at least one of the list indexed by ω. For brevity, define $\phi(i) := \phi_{\mathcal{K}}(i)$, and note that $\phi(i) = E(M_i)$, the probability that the ith unit appears on at least one list. Finally, if $\omega = \{j\}$ is a singleton, we have $\phi_j(i) := \phi_{\{j\}}(i)$, the probability that the ith unit appears on the jth list, and $\phi_j = n^{-1} \sum_{i \in \mathcal{N}} \phi_j(i)$.

1.3.4 Regression

Let $\mathcal{X}$ denote the covariate space, and let $\mathbf{x} \in \mathcal{X}$ be arbitrary. A function $r(\mathbf{y}, \mathbf{x})$ is called a *regression model* for $(x_{\cdot\cdot}, y_{\cdot\cdot})$ if it is assumed that $p(i, y_{i\cdot}) = r(y_{i\cdot}, x_{i\cdot})$ holds for all $i \in \mathcal{N}$. For any function $r(\mathbf{y}, \mathbf{x})$, define $r_\omega(\mathbf{y}, \mathbf{x}) := \sum_{\mathbf{z} \in \mathcal{Y}_k : \mathbf{z}_\omega = \mathbf{y}_\omega} r(\mathbf{z}, \mathbf{x})$.

Given a function $r(\mathbf{y}, \mathbf{x})$, define the detection function $\psi_\omega(\mathbf{x}) = 1 - r_\omega(\mathbf{0}, \mathbf{x})$, which can be interpreted as the probability that an individual with covariates $\mathbf{x}$ will appear in at least one of the lists indexed by ω. Notice that ψ is to ϕ as r is to p. In particular, if $r(\mathbf{y}, \mathbf{x})$ is a regression model, then $\psi(x_{i\cdot}) = \phi(i)$, $\psi_\omega(x_{i\cdot}) = \phi_\omega(i)$, and $\psi_j(x_{i\cdot}) = \phi_j(i)$.

1.4 Assumptions

We describe some common assumptions that are employed in the CRC literature.

(Closed population) We assume that the population is fixed during the generation of the lists $L_1, ..., L_k$. This excludes births, deaths, and migration.

(Perfect matching) It is often unclear whether a record on one list refers to the same unit as a record on another list, due to typographical errors or other anomalies. The field of record linkage addresses the problem of matching units between lists (see Fellegi and Sunter [7]). We proceed with the assumption that the lists are linked perfectly, so that the cross-classification counts $\mathbf{c}$ are all observable except for $c_{\mathbf{0}}$.

(Homogeneity) A CRC experiment is called homogeneous if the capture probabilities are constant across units. To be precise, the experiment is homogeneous if $p(i_1, \mathbf{y}) = p(i_2, \mathbf{y})$ for every pair of units i_1, i_2 and every $\mathbf{y} \in \mathcal{Y}_k$. *Heterogeneity* [in general] is taken to mean the absence of this specific kind of homogeneity.

(Independence) Several kinds of independence assumptions are commonly used. Let $\mathbf{y} \in \mathcal{Y}_k$, and $\omega \in \Omega_k \setminus \{\mathcal{K}\}$. The *lists are independent* if

$$p(\mathbf{y}) = p_\omega(\mathbf{y}) p_{\omega^c}(\mathbf{y}) \tag{1}$$

Let $i \in \{1, ..., n\}$. The *lists independent at the individual level* if

$$p(i, \mathbf{y}) = p_\omega(i, \mathbf{y}) p_{\omega^c}(i, \mathbf{y}). \tag{2}$$

Marginally, list independence at the individual level implies that the event that unit i is on a specific list is independent of the event that unit i is on any other list. In the context of a regression model $r(\mathbf{y}, \mathbf{x})$, list independence at the individual level is equivalent to *conditional independence*:

$$r(\mathbf{y}, \mathbf{x}) = r_\omega(\mathbf{y}, \mathbf{x}) r_{\omega^c}(\mathbf{y}, \mathbf{x}) \tag{3}$$

For example, if $k = 2$ with $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2)$ and $\omega = \{1\}$, then $\omega^c = \{2\}$ and conditional independence implies

$$r((\mathbf{y}_1, \mathbf{y}_2), \mathbf{x}) = \left[\psi_1(\mathbf{x})^{\mathbf{y}_1}(1 - \psi_1(\mathbf{x}))^{1-\mathbf{y}_1}\right] \left[\psi_2(\mathbf{x})^{\mathbf{y}_2}(1 - \psi_2(\mathbf{x}))^{1-\mathbf{y}_2}\right]. \tag{4}$$

Finally, *independence of units* (or *independence between individuals*) means that $P(Y_{i_1 \cdot} = \mathbf{y}) = P(Y_{i_1 \cdot} = \mathbf{y} | Y_{i_2 \cdot} = \mathbf{z})$ for all $i_1, i_2 \in \mathcal{N}$, $\mathbf{y}, \mathbf{z} \in \mathcal{Y}_k$. That is, the units are independent if the capture pattern of a unit does not depend on the capture pattern of other units. This assumption appears to be universal in the CRC literature.

(Sampling Model M1) Early CRC models tend to assume homogeneity (at least formally) and independence between units to get a multinomial sampling distribution:

$$P(\mathbf{C} = \mathbf{c}) = \frac{n!}{\prod_{\mathbf{y} \in \mathcal{Y}_k} c_\mathbf{y}!} \prod_{\mathbf{y} \in \mathcal{Y}_k} p(\mathbf{y})^{c_\mathbf{y}}. \tag{5}$$

(Sampling Model M2) Relatively recent CRC models occasionally incorporate a regression model to get a sampling distribution that is multinomial at the individual level:

$$P(Y_{\cdot \cdot} = y_{\cdot \cdot} | x_{\cdot \cdot}) = \prod_{i \in \mathcal{N}} r(y_{i \cdot}, x_{i \cdot}). \tag{6}$$

The sampling models M1 and M2 implicitly assume some things about the independence structure of the units and lists. Both models fail if, for example, the inclusion of a child on a list depends on the inclusion of that child's caretaker. However, it could be argued that the model M2 implicitly accounts for this child-parent dependence if the regression model accounts for Age.

1.5 The Petersen Estimator

In the 1890's, Petersen re-discovered and popularized an estimator that remains a fundamental building block for modern models [8]. We illustrate the Petersen estimator with an initial capture list L_1 and a single recapture list L_2. Let $c_{1+} = c_{10} + c_{11}$ and $c_{+1} = c_{01} + c_{11}$. The Petersen estimator takes the form $\hat{n} = \frac{c_{1+}c_{+1}}{c_{11}}$ and relies on the assumption of independence between lists as in equation 1. A consequence of this independence assumption is that $p((1,1)) = \phi_1 \phi_2$, and it is hypothesized that

$$\hat{n} := \frac{c_{1+}c_{+1}}{c_{11}} \approx= \frac{E(C_{1+})E(C_{+1})}{E(C_{11})} = \frac{n\phi_1 n\phi_1}{np((1,1))} = n\frac{\phi_1\phi_1}{p((1,1))} = n.$$

Heterogeneity typically implies failure of the independence assumption, so the Petersen estimate is rarely optimal. In applications with more than two data sets, individual Petersen estimates that are computed from selected pairs of lists may produce estimates that are drastically less than n_c, the number of observed units [9]. Most modern CRC estimators are adaptations of the Petersen estimator that account either for heterogeneity, conditional dependence, or both. For specific applications, including the CCM, models may account for open population effects and imperfect matching.

1.6 A Dogmatic Bifurcation

It is intuitively appealing to view a CRC estimator as the sum of two conceptual pieces. Each of these pieces is (or should be) an active area of research, and so we think of CRC has having two distinct frontiers. Having little external justification, this bifurcation constitutes a dogma that motivates our research objectives.

Researchers on the first – and oldest – frontier start with the cross-classification table $\mathbf{c}$ and make various assumptions to derive an estimator for $\mathbf{p}(\mathcal{Y}_k)$. Note that $\mathbf{p}(\mathcal{Y}_k) = \{p(\mathbf{y})\}_{\mathbf{y} \in \mathcal{Y}_k}$ is merely a collection of 2^k probabilities; covariates are ignored. When $k = 2$, researchers almost universally assume list independence as in Equation 1. As we increase the number of lists, the number of possible relationships among the lists grows. Log-linear models (Section 2.2) may

represent the completion of this research frontier in terms of achieving generality, but the surface of the essential problem of model selection has scarcely been scratched.

The second frontier uses covariates $x^c_{\cdot\cdot}$, such as Age and Race, to model the relative probabilities of the observable capture patterns. Post-stratification, applied in all Census coverage evaluations prior to 2010, was the earliest strategy to represent $\mathbf{c}$ as a function of covariates. Dividing observations according to S different post-strata $\{x^c_{\cdot\cdot}(s)\}_{s=1}^S$ leads to S partial cross-classifications $\{\mathbf{c}(s)\}_{s=1}^S$. Applying a first-frontier estimator on each partial cross-classification $\mathbf{c}(s)$ leads to S estimates, one for each post-stratum, which can be summed to get a population estimate $\hat{n}$. Continuous generalizations of post-stratification have been prominent in the literature since the mid 1980s.

Starting with Section 2.3, every CRC method that we review combines an idea from the first frontier with an idea from the second frontier, although the distinction between frontiers seems to have previously never been emphasized. Indeed, most proposed estimators that incorporate $x^c_{\cdot\cdot}$ were derived only for the two-list scenario, and these estimators tend to rely on the conditional independence assumption, a rather trivial product of the first frontier.

Remarkably, the CRC literature has proposed little that combines cutting edge methods on both frontiers. This gap in the literature is most evident when there are more than two lists. Triple system estimators that combine an advanced first-frontier method such as log-linear modeling (Section 2.2) with an advanced second-frontier method such as logistic regression (Section 2.4) have been considered only recently (Section 2.6).

In terms of our bifurcation of CRC methods into two frontiers, our thesis has three parts. The first part is to advance the first frontier by providing a regression approach that readily generalizes to three or more lists. The second part is to pioneer a *local* log-linear modeling framework for imputing multinomial capture probabilities at the individual level rather than only at the population level. The third part is to "marry" an advanced method from the first frontier with an advanced method from the second frontier in a way that leads to unprecedented accuracy in CRC estimation. Section 3 makes these goals more concrete and provides initial results.

1.7 Simulation Example

We use a simulation experiment to illustrate two kinds of list dependence. A population of 2000 individuals has ages assigned through a random process that is consistent with a baby-boomer generation and decreasing numbers of individuals as Age increases (see the first panel in Figure 1). Three lists of individuals are drawn from the population sequentially. The probability of an individual appearing on each list is a function of Age and capture history; an individual that is captured on one list becomes less likely to be captured on subsequent lists, consistent with a "respondent fatigue" effect.

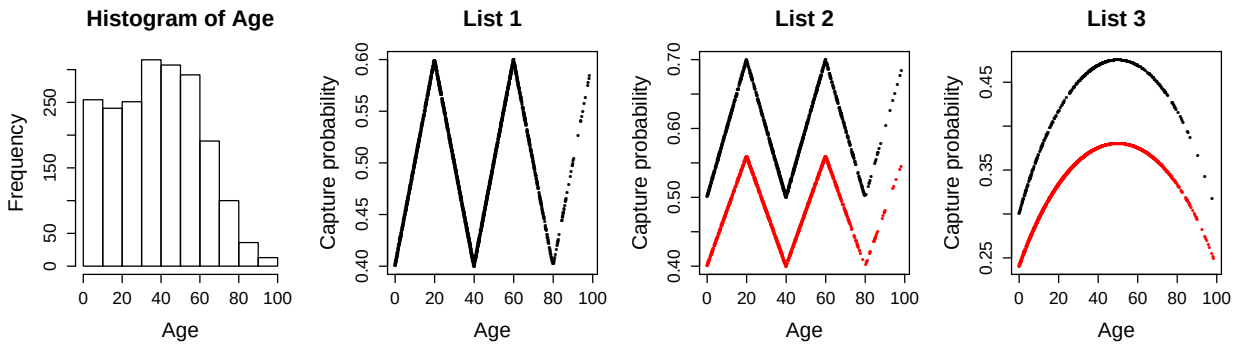

Figure 1: These panels illustrate a simulation with three lists and complex list dependence. The first panel shows the distribution of the simulated population over Age. The second panel shows the capture probabilities that were used to generate the first list. The zig-zag pattern was chosen rather arbitrarily for the purpose of creating a population with heterogeneous capture probabilities. The third panel shows the capture probabilities used to generate the second list. These probabilities are a function of both Age and capture history, and capture probabilities follow two separate curves. Points on the lower curve represent individuals that were captured on List 1. These points are about 0.1 lower than the points on the upper curve, and this difference can be thought of as a "respondent fatigue" effect. Finally, the capture probabilities for the third list have a different kind of dependence on Age in addition to a respondent fatigue effect.

Figure 2 displays box plot summaries for a hundred replicates of each of the three pairwise Petersen estimates. For each pair of lists, the Petersen estimator significantly overestimates the population size. The fact the the Petersen estimator is biased upwards (and not downwards) in this example is not obvious merely from looking at Figure 1, even for an experienced researcher, because the simulation involves two kinds of list dependence that affect the Petersen estimates in conflicting ways.

We will discuss this conflict for the first and second capture events (see the second and third panels in Figure 1). The first form of list dependence is the respondent-fatigue effect, which says that $P(Y_{i2} = 1|Y_{i1} = 1) < P(Y_{i2} = 1)$. The Petersen estimator can be rewritten as $\hat{n} = \frac{c_1+c_{+1}}{c_{11}} = c_{11} + c_{10} + c_{01} + \frac{c_{10}c_{01}}{c_{11}}$, and simple reasoning on this expanded expression shows that the respondent fatigue effect induces an upward bias in the Petersen estimator, consistent with the simulation results summarized by the first box plot in Figure 2.

The second form of list dependence results from the structure of the heterogeneity in List 1 and List 2. When capture probabilities are variable and positively correlated between lists, the Petersen estimator is typically biased downwards. Indeed, if this simulation is repeated, removing the respondent fatigue effect but preserving the basic heterogeneity structure between List 1 and List 2, the corresponding Petersen estimator becomes biased downwards.

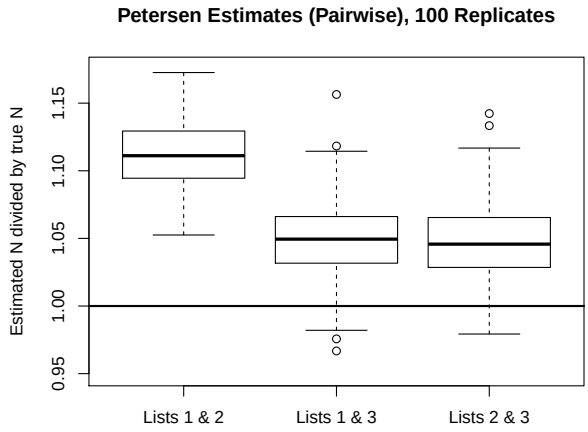

Figure 2: Petersen estimates are typically biased in the presence of list interactions. The vertical axis measures the ratio $\hat{n}/n$; when this ratio is greater than one, $\hat{n}$ is an overestimate of n.

2 Literature Review

More than a century of CRC research has produced many approaches to the problem. The `R` package `Rcapture` by Baillargian and Rivest fits some of the models discussed below [10].

2.1 Removal Methods, the Jackknife, and Chao's Lower Bound

Removal methods differ from Petersen estimates by assuming that the probability of capture for each unit is constant across lists. The basic removal method, introduced by Moran, requires that captured units are either literally removed, or are returned to the population after being identified such that they do not get counted in subsequent captures. Given a finite population, the sequence of counts of "new" units identified in each capture should converge towards zero in a roughly geometric fashion. Removal methods attempt to fit parameters to the observed terms of the sequence, and consequently infer the population size [11].

Burnham and Overton derive an estimator based on the generalized jackknife method [12]. In an experiment with k captures, let f_j denote the number of units captured exactly j times, $j = 1, ..., k$. The idea of the jackknife estimator is to generate a population size estimate $\hat{n}$ as a linear combination of the quantities f_j. Note that $n_c = \sum_j f_j$. For some constant α_1, the first order jackknife estimator is $\hat{n} = n_c + \alpha_1 f_1$, the second order estimator takes the form $\hat{n} = n_c + \alpha_1 f_1 + \alpha_2 f_2$, and so on. The model allows for heterogeneity, but assumes that for each unit, the probability of capture on each list is constant.

Chao derived the "Chao's lower bound" population estimator, which is closely related to the jackknife estimator for populations with heterogeneous capture probabilities. Chao showed that her estimator may perform better than the jackknife estimator when the number of capture events k is large and the capture probabilities are "severely" heterogeneous such that many units are captured substantially less frequently than the rest of the population [13].

2.2 Log-linear Models

Fienberg applied log-linear models to the CRC setting in 1972, offering maximal flexibility [14]. Building off of Sampling Model M1 (Equation 5), Fienberg's approach assumes only that the highest-order interaction between lists is negligible.

Given any function $f : \mathcal{Y}_k \to \mathbb{R}$ with $\sum_{\mathbf{y}} \exp f(\mathbf{y}) = 1$, one can use a log-linear model to reparameterize the multinomial capture probabilities as a function of capture pattern: $\log p(\mathbf{y}) = f(\mathbf{y})$. The form of f is completely arbitrary, and a log-linear parameterization always exists to exactly fit any multinomial probability array $\mathbf{p}(\mathcal{Y}_k)$.

However, given that the $c_{\mathbf{0}}$ cell of $\mathbf{c}$ is not observed, any saturated parameterization for $\mathbf{p}(\mathcal{Y}_k)$ is not identifiable. Indeed, the number of free parameters in the full multinomial model $\mathbf{p}(\mathcal{Y}_k)$ is $2^k - 1$, with the "-1" due to the restriction that $\sum_{\mathbf{y}} p(\mathbf{y}) = 1$. Meanwhile, the number of known values in $\mathbf{c}$ is $2^k - 1$, and the constraint is that $\sum_{\mathbf{y} \neq \mathbf{0}} c_{\mathbf{y}} = n_c$.

Therefore, given the missingness in the data, the saturated model has $2^k - 2$ free parameters, providing a perfect fit for the probabilities $\mathbf{p}(\mathcal{Y}_k \setminus \{\mathbf{0}\})$. Fitting models with even fewer than $2^k - 2$ free parameters provides degrees of freedom for testing model fit, adds bias, and reduces the variance of the corresponding population size estimate $\hat{n}$. That is, minimizing prediction risk involves (as usual!) a bias-variance trade-off, and selection of an appropriate model is not straightforward. Fienberg outlined a model selection strategy based on likelihood-ratio tests, restricting attention to hierarchical log models [14].

We emphasize that hierarchical modeling requires the assumption that the highest-order interaction is negligible. This assumption fails even in basic simulation scenarios (see Appendix A). Therefore, even a hierarchical log-linear model with many degrees of freedom and the appearance of excellent model fit may lead to disastrously incorrect predictions.

2.3 Smooth Post-Stratification via Horvitz-Thompson

While many of the models discussed above allow for significant flexibility in the structure of the multinomial capture probabilities $\mathbf{p}(\mathcal{Y}_k)$ as in Equation 5, none of these models explicitly use individual-level characteristics as in Equation 6. An elementary way to incorporate covariate information is to fit a model separately on each of a collection of post-strata. Petersen estimates based on each of several hundred post-strata formed the basis for official census coverage evaluations for the 1980, 1990, and 2000 censuses [15]. As a simple example of post-stratification, consider partitioning the E-sample and P-sample census coverage evaluation data into four categories: white and under 40 years old; white and over 40 years old; non-white and under 40 years old; and non-white and over 40 years old. Each of these four post-strata gives a separate cross-classification table $c(Age, Race) = c(x^c_{\cdot\cdot})$ like the one displayed in Table 1. The four Petersen estimates corresponding to the four post-strata may be summed to provide an estimate of the population:

$$\hat{n} = \sum_{Age, Race} \hat{n}(Age, Race) \tag{7}$$

Suppose that $r(\mathbf{y}, \mathbf{x})$ is a regression model for $(x_{\cdot\cdot}, y_{\cdot\cdot})$, and $\psi(\mathbf{x})$ is the implied detection function. Hence $\phi_i = \psi(x_{i\cdot}) = P(M_i = 1)$. Post-stratification implies a discrete approximation $\hat{r}$ of the regression function r, while a smooth approximation may be preferable. The Horvitz-Thompson estimator generalizes post-stratification to a [optionally] smooth analogue:

$$\widetilde{N} = \sum_{i:M_i=1} \frac{M_i}{\psi(x_i)} = \sum_{i:M_i=1} \frac{1}{\psi(x_i)} \tag{8}$$

It is easy to verify that $E\widetilde{N} = n$. Alho showed that $\widetilde{N}$ is consistent and asymptotically normal if ϕ_i is uniformly bounded away from 0 and 1 for all $i \in \mathcal{N}$ [16].

Now, any detection function estimator $\hat{\psi}(\mathbf{x})$ implies a Horvitz-Thompson estimator. For example, consider the post-stratification example of Equation 7. Let $n_c(Age, Race)$ denote the number of individuals observed in a particular stratum and $n(Age, Race)$ denote the corresponding true number of individuals, and let $\hat{\psi}(Age, Race) =$

$n_c(Age, Race)/\hat{n}(Age, Race)$. Then

$$\hat{n} = \sum_{Age,Race} \hat{n}(Age, Race) = \sum_{Age,Race} \frac{n_c(Age, Race)}{\frac{n_c(Age,Race)}{\hat{n}(Age,Race)}} = \sum_{Age,Race} \frac{n_c(Age, Race)}{\hat{\psi}(Age, Race)} = \sum_{i:m_i=1} \frac{1}{\hat{\psi}(x_i)} \tag{9}$$

Hence, Equation 8 is more general than Equation 7, giving us the option of replacing the discrete estimator $\hat{\psi}$ with a smooth one. We refer to every Horvitz-Thompson estimator that on relies on a smooth estimate $\hat{\psi}$ of the detection function as a *smooth post-stratification* method.

Just as the Petersen estimates rely on the assumption of list independence within each post-strata, most of the smooth estimators of the detection function that have been proposed in the literature for the two-list scenario make the conditional independence assumption as in equations 3 and 4, and few of these proposed estimators have been generalized for a three-list scenario.

2.4 Logistic Regression

Logistic regression was perhaps the earliest smooth estimator of the detection function. Pollock, Hines, and Nichols in 1984 were the first to apply logistic regression in the CRC context [17]. Alho made additional contributions and applied logistic regression for a Census coverage evaluation in the 1990's [16] [18]. The CCM was the first official Census coverage evaluation to rely on logistic regression instead of post-stratification. (For more details on the 2010 CCM, see Section 2.7.)

Alho proceeded approximately as follows with $k = 2$. For $j = 1, 2$, let θ_j be a $q \times 1$ vector of parameters, and $\theta := (\theta_1, \theta_2)$. Define

$$r(\mathbf{y}, \mathbf{x}) := \left[logit^{-1}(\mathbf{x}\theta_1\mathbf{y}_1)^{\mathbf{y}_1}(1 - logit^{-1}(\mathbf{x}\theta_1\mathbf{y}_1))^{1-\mathbf{y}_1}\right]\left[logit^{-1}(\mathbf{x}\theta_2\mathbf{y}_2)^{\mathbf{y}_2}(1 - logit^{-1}(\mathbf{x}\theta_2\mathbf{y}_2))^{1-\mathbf{y}_2}\right],$$

and assume that $r(\mathbf{y}, \mathbf{x})$ is a regression model. Then $\psi_j(\mathbf{x}) = logit^{-1}(\mathbf{x}\theta_j)$, for $j = 1, 2$, and conditional independence holds. In particular, $r((1, 1), \mathbf{x}) = \psi_1(\mathbf{x})\psi_2(\mathbf{x})$, so that

$$\psi(\mathbf{x}) = \psi_1(\mathbf{x}) + \psi_2(\mathbf{x}) - \psi_1(\mathbf{x})\psi_2(\mathbf{x}) \tag{10}$$

Therefore, any estimate $\hat{\theta}$ implies an estimate of the detection function $\hat{\psi}$. In turn, $\hat{\psi}$ can be used to generate a Horvitz-Thompson estimate as in Equation 8.

Alho estimated θ by maximizing a conditional likelihood function. Recall that $\mathbf{y}_{ij}$ is the indicator that the ith individual appears on the jth list. A few lines of algebra shows that

$$P(Y_{i\cdot} = y_{i\cdot}|x_{i\cdot}, \theta, M_i = 1) = \frac{e^{(x_{i\cdot}\theta_1 y_{i1} + x_{i\cdot}\theta_2 y_{i2})}}{e^{x_{i\cdot}\theta_1} + e^{x_{i\cdot}\theta_2} + e^{(x_{i\cdot}\theta_1 + x_{i\cdot}\theta_2)}} \tag{11}$$

The conditional likelihood function is then $\prod_{m_i=1} P(Y_{i\cdot} = y_{i\cdot}|x_{i\cdot}, \theta, M_i = 1)$.

2.5 Kernel Density Estimation

The logistic regression method of Alho may require high polynomial orders in the covariates to fit the data. For example, as a function of Age, capture probabilities for the Census tend to be very nonlinear, with a dip around ages 18 to 29 as children leave their parents' residences for school or work [19]. In such cases, a nonparametric approach might be more fitting [sic].

Chen and Lloyd [20] developed a two-list nonparametric estimation framework centered around estimating the "dependence parameter" α satisfying $\alpha\phi_1\phi_2 = p((1, 1))$. Note that taking $\alpha = 1$ is the same as assuming list independence (1), and $\alpha > 1$ is consistent with positive list dependence. If α is known, a simple maximum likelihood estimation leads directly to a population estimate. Specifically, with $c_{\mathbf{0}} = n - n_c$, the multinomial likelihood implied by Equation 5 can be re-parameterized as

$$L(\phi_1, \phi_2, n|\mathbf{c}, \alpha) \propto \frac{n!}{(n - n_c)!}\left(1 - \phi_1 - \phi_2 + \alpha\phi_1\phi_2\right)^{n-n_c}(\alpha\phi_1\phi_2)^{c_{11}}(\phi_1 - \alpha\phi_1\phi_2)^{c_{10}}(\phi_2 - \alpha\phi_1\phi_2)^{c_{01}}.$$

Chen and Lloyd estimated α externally (prior to performing maximum likelihood for the remaining parameters) using a rather bulky nonparametric kernel density estimation framework that relied on the assumption of conditional independence.

Chen and Lloyd [21] proposed a far simpler nonparametric approach using two lists. Suppose that $r(\mathbf{y}, \mathbf{x})$ is a regression model for $(x_{..}, y_{..})$, and let $\psi(\mathbf{x})$ be the detection function. Let $\omega(j) = \mathcal{K} \backslash \{j\}$. Assume conditional independence (Equation 3). Then $\psi_j(x_{i\cdot}) = P(Y_{ij} = 1 | x_{i\cdot}) = P(Y_{ij} = 1 | Y_{i\omega(j)}, x_{i\cdot})$. In particular, if $I_{(-j)} = \cup_{\ell=1,\ldots,k:\ell\neq j} L_\ell$ is the set of units that appear on at least one list *excluding* the jth list, then $\psi_j(x_{i\cdot}) = E(Y_{ij} | i \in I_{(-j)}, x_{i\cdot})$. Therefore, regressing Y_{ij} on $x_{i\cdot}$ for only the observed units $i \in I_{(-j)}$ provides an estimate $\hat{\psi}_j(\mathbf{x})$ for $j = 1, 2$. Finally, conditional independence implies an estimate $\hat{\psi}(\mathbf{x})$ as in equation 10, and a Horvitz-Thompson estimator is immediate.

2.6 Joint Estimation of Covariate Effects and List Interactions

A relatively small and recent body of work deals with the problem of simultaneously modeling covariate effects and list interactions. In terms of our dogmatic bifurcation (Section 1.6), these methods provide a link between the first and second frontier.

In 1990, Baker introduced a method for fitting a log-linear model jointly across a collection of post-strata in a two-list scenario [22]. In 2000, Pledger presented a logistic-linear approach that admits more than two lists and allows capture probabilities to depend on time, capture histories, and post-strata in a unified way (i.e., not fitting a separate model for each post-strata) [23].

In 2001, Yip, Wan, and Chan applied logistic regression in a way that admits more than two lists and allows capture probabilities to depend continuously on covariates, time, and capture history. To our knowledge, this is the first instance of Horvitz-Thompson estimation using more than two lists [24].

2.7 CRC in the 2010 CCM

An internal Census memorandum discusses the modeling process used in the CCM [25]. The Census estimates several adjustment factors for each observed individual. Denote these adjustments collectively as A_i. For each observed individual, a logistic regression model is used to produce an estimate $\hat{\psi}_1(\mathbf{x})$ of $\psi_1(\mathbf{x})$, the probability that an individual with covariates $\mathbf{x}$ is enumerated in the Census. Let D denote an estimation domain. For example, D could denote the set of individuals in the state of Pennsylvania. Then the CCM estimate for $n(D)$, the population in domain D, is given as an adjusted Horvitz-Thompson style estimator:

$$\hat{n}(D) = \sum_{i \in D : y_{i1} = 1} \frac{A_i}{\hat{\psi}_j(x_{i\cdot})}. \tag{12}$$

Note that Equation 12 differs from the standard Horvitz-Thompson estimator by including using $\hat{\psi}_1$ instead of $\hat{\psi}$ in the denominator. The reason for this difference is that if D is a domain that does not overlap with the P-sample, then the probability of detection is equal to the probability of enumeration in the Census.

3 Research Proposal and Initial Results

3.1 A List of Goals

Our research goals are strongly motivated by the "dogmatic bifurcation" introduced in Section 1.6. Broadly, we see the field of CRC as having two distinct frontiers, and we aim to advance each frontier separately with an eye towards combining the most advanced methods from each frontier into a single algorithm. To achieve this, we intend to pursue the following research objectives:

- **G1**: Derive a simple framework that subsumes every post-stratification and smooth post-stratification approach, including logistic regression as in Alho [16] and non-parametric regression as in Chen and Lloyd [21]. The new framework should incorporate three or more lists in a way that does not impose any specific assumptions regarding the structure of the multinomial probabilities $\mathbf{p}(\mathcal{Y}_k)$. This is done in Section 3.2.
- **G2**: Develop a *local* log-linear modeling strategy that performs model selection for each observed unit based on smooth post stratification with more than two lists, and use the fitted models to estimate detection probabilities for Horvitz-Thompson style estimation. Section 3.5 will demonstrate how to apply a saturated log-linear model for every unit. However, a non-saturated model may be more appropriate. Explore selecting a separate log-linear model for each unit or group of units.
- **G3**: Hierarchical log-linear models may introduce significant bias by requiring that the highest-order list interaction ρ is zero. Such an assumption is fundamentally unavoidable, but using $\rho = 0$ is not necessarily the most natural

choice. In particular, the population estimate is some function of ρ. Rather than setting $\rho = 0$, use a moderately informative prior distribution for ρ to propagate uncertainty into the population size confidence [or credible] intervals.

G4: Write a simulation package for testing CRC estimators in `R`, and use this package to compare our proposed methods against existing methods. Since validation data typically does not exist in applications, empirical risk estimation is not feasible. Simulation may represent the best way to assess the sensitivity of CRC estimators to the model assumptions (see Section 4.1).

G5: Evaluate the accuracy of the 2010 Census by combining the ACS with the E- and P-samples which were used in the CCM. Because the overlap between the geographic domains of the ACS and the CCM is small, the triple-system data set will likely be too small to make use of many covariates. Therefore, we may estimate the Census capture probability function $\psi_c(\mathbf{x})$ in a way that incorporates information from both the ACS/E-sample/P-sample triple-system and the E-sample/P-sample dual system. An intuitive starting point is to produce a hybrid estimator $\hat{\psi}_c(\mathbf{x}) = (1 - \alpha(\mathbf{x}))\hat{\psi}_{c_d}(\mathbf{x}) + \alpha(\mathbf{x})\hat{\psi}_{c_t}(\mathbf{x})$, where $\hat{\psi}_{c_d}(\mathbf{x})$ is a dual system estimator, $\hat{\psi}_{c_t}(\mathbf{x})$ is a separate triple system estimator, and $0 < \alpha(\mathbf{x}) < 1$ is a weighting function that depends on the variance of $\hat{\psi}_{c_t}(\mathbf{x})$.

G6: Our algorithm (see Section 4.3) strings together three separate estimation techniques. Regression is used to estimate the relative frequencies of observable capture patterns, log-linear models are used to estimate the relative frequency of the missing cell, and the result is plugged into a Horvitz-Thompson estimator. Explore how these techniques relate to each other, and develop a more unified estimation approach.

Throughout, the scope of our study is restricted by the following assumptions unless explicitly stated otherwise.

A1: Records are perfectly matched between lists.

A2: The population is closed.

A3: Given $x_{\cdot,\cdot}$, the sampling distribution of $Y_{\cdot\cdot}$ is multinomial as in Equation 6. In particular, assume that $r(\mathbf{y}, \mathbf{x})$ is a regression model for $(x_{\cdot\cdot}, y_{\cdot\cdot})$ such that smooth estimators $\hat{\psi}(x)$ converge to $\psi(x)$.

A4: For each $x_{i\cdot}$, the array of local multinomial probabilities $\{r(\mathbf{y}, x_{i\cdot})\}_{\mathbf{y}\in\mathcal{Y}_k}$ can be parameterized by a hierarchical log-linear model using only $2^k - 2$ free parameters.

3.2 A New Smooth Post-Stratification Framework

We begin by deriving an estimator of the detection function for two or three lists that relies on the assumption of conditional independence (3). Observe that

$$\begin{aligned}\psi_j(x_{i\cdot}) = P(Y_{ij} = 1|x_{i\cdot}) &= P(Y_{ij} = 1|M_i = 1, x_{i\cdot})\psi(x_{i\cdot}) + P(Y_{ij} = 1|M_i = 0, x_{i\cdot})(1 - \psi(x_{i\cdot})) \\ &= P(Y_{ij} = 1|M_i = 1, x_{i\cdot})\psi(x_{i\cdot})\end{aligned}$$

Combining this result with conditional independence gives

$$\begin{aligned}\psi(x_{i\cdot}) &= 1 - P(M_i = 0|x_{i\cdot}) \\ &= 1 - \prod_j (1 - P(Y_{ij} = 1|x_{i\cdot})) \\ &= 1 - \prod_j (1 - P(Y_{ij} = 1|M_i = 1, x_{i\cdot})\psi(x_{i\cdot}))\end{aligned}$$

Suppose there exist functions $\{\pi(j, \mathbf{x})\}_{j\in\mathcal{K}}$ that are smooth in $\mathbf{x}$ with $\pi(j, x_{i\cdot}) = P(Y_{ij} = 1|M_i = 1, x_{i\cdot})$ for $i = 1, ..., n_c$, $j \in \mathcal{K}$. Note that each $\pi(j, \mathbf{x})$ is conditioned on observability (i.e., $M_i = 1$), so $\pi(j, \mathbf{x})$, can be estimated directly from the data $(x^c_{\cdot\cdot}, y^c_{\cdot\cdot})$ using any kind of binary regression. The expression above becomes

$$\psi(x_{i\cdot}) = 1 - \prod_j (1 - \pi(j, x_{i\cdot})\psi(x_{i\cdot})) \tag{13}$$

Now, for each $i = 1, ..., n_c$, the detection probability $\psi(x_{i\cdot})$ can be estimated numerically. If $k = 2$,

$$\psi(x_{i\cdot}) = \frac{\pi(1, x_{i\cdot}) + \pi(2, x_{i\cdot}) - 1}{\pi(1, x_{i\cdot})\pi(2, x_{i\cdot})} \tag{14}$$

Suppose $k > 2$. Fix i for brevity, and let $\pi_j := \pi(j, x_{i\cdot})$. It is not immediately clear whether Equation 13 admits a unique feasible root $\psi(x_{i\cdot})$ for all possible combinations of values $\pi_1, ..., \pi_k$ that satisfy the necessary constraints

$$1 \geq \pi_j \geq 0; \tag{15}$$

$$\sum_j \pi_j \geq 1. \tag{16}$$

If $k = 3$, Equation (13) gives

$$\psi(x_{i\cdot}) = (\pi_1 + \pi_2 + \pi_3)\psi(x_{i\cdot}) - (\pi_1\pi_2 + \pi_1\pi_3 + \pi_2\pi_3)\psi(x_{i\cdot})^2 + \pi_1\pi_2\pi_3\psi(x_{i\cdot})^3 \tag{17}$$

We are interested only in solutions in the interval $(0, 1)$. After excluding the trivial solution, $\psi(x_{i\cdot}) = 0$, two quadratic roots remain. If exactly one of these two roots is feasible for all individuals, the solutions can be applied in a Horvitz-Thompson estimator. However, finding a feasible root of Equation 17 is of dubious value when $k > 2$ because this equation rests on the conditional independence assumption, whereas a more complex relationship between the multinomial probabilities may be appropriate, as discussed in 3.4. On the other hand, the assumption of conditional independence between three or more lists is not unprecedented (see the grade-of-membership model of Manrique and Fienberg [26]). Implementing the Horvitz-Thompson estimator based on Equation 17 could lead to insightful contrasts with existing estimators.

We now derive a more general framework for smooth post-stratification that subsumes both equations (14 and 17). Define functions $\Pi := \{\pi(\mathbf{y}, \mathbf{x})\}_{\mathbf{y} \in \mathcal{Y}_k}$ via

$$\pi(\mathbf{y}, \mathbf{x}) := \frac{r(\mathbf{y}, \mathbf{x})}{\sum_{\mathbf{z} \neq \mathbf{0}} r(\mathbf{z}, \mathbf{x})} = \frac{r(\mathbf{y}, \mathbf{x})}{\psi(\mathbf{x})}. \tag{18}$$

Hence $r(\mathbf{y}, \mathbf{x}) = \psi(\mathbf{x})\pi(\mathbf{y}, \mathbf{x})$, and

$$\psi(\mathbf{x}) = \frac{\sum_{\mathbf{y} \neq \mathbf{0}} r(\mathbf{y}, \mathbf{x})}{\sum_{\mathbf{y}} r(\mathbf{y}, \mathbf{x})} = \frac{\sum_{\mathbf{y} \neq \mathbf{0}} \psi(\mathbf{x})\pi(\mathbf{y}, \mathbf{x})}{\sum_{\mathbf{y}} \psi(\mathbf{x})\pi(\mathbf{y}, \mathbf{x})} = \frac{\sum_{\mathbf{y} \neq \mathbf{0}} \pi(\mathbf{y}, \mathbf{x})}{\pi(\mathbf{0}, \mathbf{x}) + \sum_{\mathbf{y} \neq \mathbf{0}} \pi(\mathbf{y}, \mathbf{x})} \tag{19}$$

(Clearly, $\sum_{\mathbf{y} \neq \mathbf{0}} \pi(\mathbf{y}, \mathbf{x}) = 1$, so the final expression reduces to $\frac{1}{\pi(\mathbf{0}, \mathbf{x}) + 1}$. However, the form of Equation 19 turns out to be convenient.)

Observe that $\pi(\mathbf{y}, x_{i\cdot}) = P(Y_{i\cdot} = \mathbf{y} | M_i = 1, x_{i\cdot})$ for each $\mathbf{y} \neq \mathbf{0}$. That is, $\pi(\mathbf{y}, x_{i\cdot})$ represents the probability that the ith unit has capture pattern $\mathbf{y}$ *conditional* on the ith unit being observed in at least one of the lists. Therefore, each $\pi(\mathbf{y}, x_{i\cdot})$, $\mathbf{y} \neq \mathbf{0}$, can be estimated directly from the data using any binary regression method including logistic regression, nonparametric generalized additive models, or conditional density estimation (see Section 3.3).

Let $\Pi^* := \{\pi(\mathbf{y}, \mathbf{x})\}_{\mathbf{y} \neq 0}$. Any set of estimates $\hat{\Pi}^* = \{\hat{\pi}_{\mathbf{j}}(x)\}_{\mathbf{j} \neq 0}$ can subsequently be used to impute the function $\pi(\mathbf{y}, \mathbf{x})$ using any method. If there are only two lists, one example of an imputation strategy is to assume conditional independence, which implies $\pi((0, 0), \mathbf{x}) = \frac{\pi((1,0),\mathbf{x})\pi((0,1),\mathbf{x})}{\pi((1,1),\mathbf{x})}$. For three lists, note that $\pi_j(\mathbf{x}) = \sum_{\mathbf{y}:\mathbf{y}_j=1} \pi_{\mathbf{y}}(\mathbf{x})$ for $j = 1, 2, 3$. Then Equation 17 can be applied to $\hat{\Pi}^*$ to provide an estimate for $\psi(\mathbf{x})$, which in turn implies an estimate of $\pi(\mathbf{y}, \mathbf{x})$ as $\pi(\mathbf{0}, \mathbf{x}) = \frac{1}{\psi(\mathbf{x})} - 1$. Alternatively, one can discard conditional independence and apply a saturated hierarchical log-linear model as demonstrated in Section 3.5. For the most general case, we propose deriving a local log-linear fitting procedure for imputing $\pi(\mathbf{0}, \mathbf{x})$ with any number of lists in Section 3.4.

We emphasize that Equation 19 provides the primary motivation for breaking CRC estimation into a two step process as outlined in Section 1.6. The first step is to generate the estimates $\hat{\Pi}^* := \{\hat{\pi}(\mathbf{y}, \mathbf{x})\}_{\mathbf{y} \neq 0}$, while the second step is to impute $\pi(\mathbf{0}, \mathbf{x})$. This bifurcated approach is not necessarily optimal; we see it as a platform which may give rise to a unified estimation framework as suggested in goal G6.

Returning to the two-list scenario, it is easy to verify that Equation 14 is a special case of Equation 19. Moreover, in the logistic regression framework of Alho (see Section 2.4), we have $\pi(\mathbf{y}, x_{i\cdot}) = P(Y_{i\cdot} = \mathbf{y} | x_{i\cdot}, \theta, M_i = 1)$, and substituting $\psi_j(\mathbf{x}) = logit^{-1}(\mathbf{x}\theta_j)$ into Equation 10 gives exactly the same expression obtained by using Equation 11 to substitute into Equation 19. Hence, Alho's logistic regression framework is a special case of our framework. Kernel regression as in Chen and Lloyd [21] fits into our framework using similar reasoning.

3.3 Conditional Frequency Estimation

In this section we review a tool for estimating the functions Π^*, discuss a slight generalization of Equation 19, and use a simulation to demonstrate this generalization.

Suppose that each vector $x_{i\cdot}$ is a realization of some random variable $\mathbf{X}$. Suppose $f_M(\mathbf{x})$ is a function that satisfies $f_M(x_{i\cdot}) = P(\mathbf{X} = x_{i\cdot}|M_i = 1)$, and let $g_M(\mathbf{y}, \mathbf{x}) := \pi(\mathbf{y}, \mathbf{x}) f_M(\mathbf{x})$. Then, for all $i \in \mathcal{N}$,

$$g_M(y_{i\cdot}, x_{i\cdot}) = P(Y_{i\cdot} = y_{i\cdot}|X = x_{i\cdot}, M_i = 1)P(X = x_{i\cdot}|M_i = 1) = P(y_{i\cdot}, x_{i\cdot}|M_i = 1).$$

Note that g_M and f_M each can be estimated directly from the observable data (i.e., units with $M_i = 1$), and from the definition of g_M we can express the conditional density of capture pattern $\mathbf{y}$ given $\mathbf{X} = \mathbf{x}$ as

$$\pi(\mathbf{y}, \mathbf{x}) \quad = \quad \frac{g_M(\mathbf{y}, \mathbf{x})}{f_M(\mathbf{x})}.$$

Hall, Racine, and Li propose a nonparametric conditional density estimator that selects bandwidths by cross-validation [27]. At present, the implementation of their algorithm in R is too slow to use for data sets near the scale of Census applications, but the existence of at least a slow solution suggests that conditional density estimation is not a dead-end for the CRC context.

As an alternative to conditional density estimation, one could fit a separate binary regression to estimate each function in Π^*. A reason not to take this alternative approach is that fitting each π separately may lead to $\sum_{\mathbf{y} \neq \mathbf{0}} \pi(\mathbf{y}, \mathbf{x}) \neq 1$ due to local estimation error. However, a convenient feature of the final expression in Equation 19 is that this fraction does not change if we scale all the π's by a constant. To be precise, suppose that the imputation model for $\pi(\mathbf{0}, \mathbf{x})$ is of the form $\pi(\mathbf{0}, \mathbf{x}) = f_{\text{imp}}(\Pi^*)$ for some function f_{imp}. We say that f_{imp} is *scale-invariant* if $f_{\text{imp}}(t\Pi^*) = t f_{\text{imp}}(\Pi^*)$ for all $t > 0$. For example, imputation under conditional independence is scale-invariant, since

$$t f_{\text{imp}}(\Pi^*) = t \frac{\pi((1,0), \mathbf{x})\pi((0,1), \mathbf{x})}{\pi((1,1), \mathbf{x})} = \frac{t\pi((1,0), \mathbf{x}) t\pi((0,1), \mathbf{x})}{t\pi((1,1), \mathbf{x})} = f_{\text{imp}}(t\Pi^*).$$

Therefore, with any scale-invariant imputation method, Equation 19 may still lead to a valid estimator for the detection function $\psi(\mathbf{x})$ even if we replace all the π's with estimates that break the constraint $\sum_{\mathbf{y} \neq \mathbf{0}} \pi(\mathbf{y}, \mathbf{x}) \equiv 1$.

We illustrate estimating Π^* using three separate regressions in a simulation example with two lists. Each replication involves a population of 1000 individuals with selection probabilities a function of Age as shown in the first panel of Figure 3. The selection probabilities are the same for both lists.

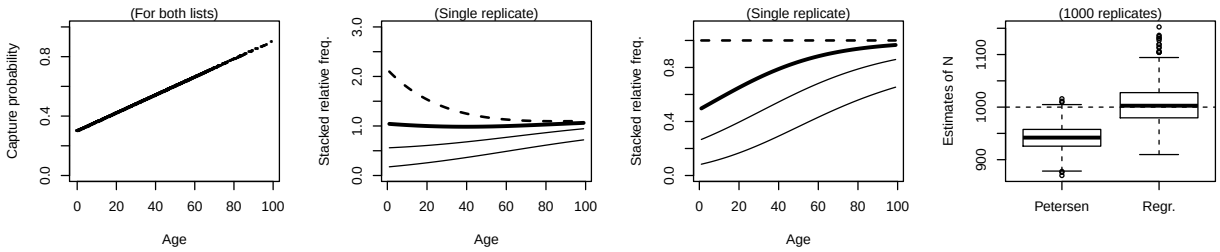

Figure 3: This figure illustrates the simulation described in Section 3.3. First panel: Capture probabilities as a function of Age. Second panel: Stacked estimates $\hat{\pi}_{\mathbf{j}}$ are illustrated as solid curves, and the imputation $\hat{\pi}(\mathbf{0}, \mathbf{x})$ that is implied by conditional independence is added on top as the dashed curve. Third panel: normalized π's. Fourth panel: A thousand replicates of Petersen estimates (left) and regression estimates (right).

We estimated $\pi((1,1), x)$ by estimating $E(I(Y_{i\cdot} = (1,1))|M_i = 1, Age = x)$ using the `gam` function with the logit link function via the `mgcv` package in `R`. This algorithm fits a nonparametric logistic kernel regression on Age. We estimated the other two π's in Π^* analogously. The solid black curves in the second panel of Figure 3 illustrate the three regression curves stacked on top of each other. That is, the bottom curve is $\hat{\pi}((1,1), x)$; the second curve is $\hat{\pi}((1,1), x) + \hat{\pi}((1,0), x)$, and the third (bold) curve is $\hat{\pi}((1,1), x) + \hat{\pi}((1,0), x) + \hat{\pi}((0,1), x)$.

Since this simulation does not incorporate a respondent fatigue effect, unlike the simulation in Section 1.7, the only cause for list dependence is heterogeneity. Therefore, the simulated lists are independent conditional on Age, and we use the conditional independence assumption to impute an estimate $\hat{\pi}((0,0), x)$, plotted on top of the other curves in the second panel of Figure 3 as the dashed curve. The fact that the bold solid curve is not identically equal to unity reveals error in the estimates $\hat{\Pi}^*$, but normalizing those estimates prior to imputing $\hat{\pi}((0,0), x)$ will have no effect on the resulting detection function estimate $\hat{\psi}(x)$ because our imputation method is scale-invariant.

The π's in the second panel of Figure 3 are normalized and re-plotted in the third panel. Here, the bold curve represents an estimate $\hat{\psi}(x)$ according to Equation 19, and this curve corresponds to a single Horvitz-Thompson estimate of the population size. A thousand replicates of this regression estimator are summarized in the second box-plot in the fourth panel of Figure 3. For comparison, the first box-plot summarizes a thousand replicates of a Petersen estimator.

The Petersen estimator tends to be biased downward because it does not account for the heterogeneity of capture probabilities. By comparison, the regression estimator tends to have a small positive bias. Based on the 1000 replications, a 95% asymptotic normal C.I. for the mean estimate is approximately $(1003, 1007.6)$. This upward bias in the regression estimator has at least two likely causes. The first cause is bias in the detection probability estimates $\hat{\psi}(x)$, particularly at the end points, where Age $= 0$ or 100. The second cause is that variance in the estimator $\hat{\psi}$ tends to induce an upward bias in the Horvitz-Thompson estimator, even if $\hat{\psi}(x)$ is unbiased.

While the Petersen estimator has a larger bias, the regression estimator has a larger variance. This can be understood in terms of the bias-variance trade-off in model fitting: The Petersen model uses only three parameters, whereas the [nonparametric] regression model uses up a large number of [effective] degrees of freedom in accounting for Age. A related point is that the present simulation does not address the question of whether post-stratification in combination with the Petersen estimator could potentially outperform the regression estimator. In fact, we do not anticipate an appreciable difference between the two approaches in a two-list scenario for most applications; we expect the benefits of our smooth post-stratification approach to become more pronounced with three or more lists.

3.4 Local Log-linear Models

With three lists, the saturated hierarchical log-linear model for the multinomial cross-classification Y can be parameterized as

$$\log P(Y_{i\cdot} = \mathbf{y}) \quad = \quad u + u_1\mathbf{y}_1 + u_2\mathbf{y}_2 + u_3\mathbf{y}_3 + u_{12}\mathbf{y}_1\mathbf{y}_2 + u_{13}\mathbf{y}_1\mathbf{y}_3 + u_{23}\mathbf{y}_2\mathbf{y}_3. \tag{20}$$

As discussed in Section 2.2, including a coefficient ρ for the highest-order interaction $\mathbf{y}_1\mathbf{y}_2\mathbf{y}_3$ would make the model not identifiable. The subject of goal G3 is to explore the level of bias caused by excluding ρ.

Fienberg [14] stated a maximum likelihood solution for the saturated model as

$$\hat{c}_{000} = e^{\hat{u}} \quad = \quad \frac{c_{111}c_{001}c_{010}c_{100}}{c_{011}c_{110}c_{101}}. \tag{21}$$

Setting $\hat{\pi}(\mathbf{y}) = c_{\mathbf{y}}/n_c$ for all $\mathbf{y}$ and substituting into Equation 21, every n_c cancels to give

$$\hat{\pi}((0,0,0)) \quad = \quad \frac{\hat{\pi}((1,1,1))\hat{\pi}((0,0,1))\hat{\pi}((0,1,0))\hat{\pi}((1,0,0))}{\hat{\pi}((0,1,1))\hat{\pi}((1,1,0))\hat{\pi}((1,0,1))}. \tag{22}$$

One can view Equation 22 as an imputation method that is applicable in the context of Equation 19. This log-linear imputation method provides our first example of a *local* log-linear model solution if we simply replace each $\hat{\pi}(\mathbf{y})$ by an estimate that is allowed to depend on covariates:

$$\hat{\pi}((0,0,0),\mathbf{x}) \quad = \quad \frac{\hat{\pi}((1,1,1),\mathbf{x})\hat{\pi}((0,0,1),\mathbf{x})\hat{\pi}((0,1,0),\mathbf{x})\hat{\pi}((1,0,0),\mathbf{x})}{\hat{\pi}((0,1,1),\mathbf{x})\hat{\pi}((1,1,0),\mathbf{x})\hat{\pi}((1,0,1),\mathbf{x})}. \tag{23}$$

In fact, Equation 20 can be rewritten in local form as a model for the relative frequency of each capture pattern with a [optionally] separate model for each unit:

$$\log \pi(\mathbf{y}, x_{i\cdot}) \quad = \quad u(x_{i\cdot}) + u_1(x_{i\cdot})\mathbf{y}_1 + u_2(x_{i\cdot})\mathbf{y}_2 + \cdots + u_{23}(x_{i\cdot})\mathbf{y}_2\mathbf{y}_3. \tag{24}$$

To avoid overfitting, we wish to test the importance of each coefficient in search of an adequate submodel, and this is the subject of goal G2. At least two popular model select strategies exist. Fienberg demonstrated how to use a sequence of likelihood ratio test statistics to include or exclude model terms in a stepwise fashion [14]. Another criteria for a stepwise model search is to optimize the AIC score, as implemented in the `glm` function in `R`.

Local log-linear modeling involves several closely related issues. The first issue is that trying to avoid over-fitting by selecting a separate log-linear sub-model for each unit appears to introduce overfitting in its own right. The second issue is that the collection of locally-fitted log-linear models will necessarily be highly correlated across units due to the smoothed nature of the functions $\hat{\Pi}^*$ which determine the log-linear models. Hence, log-linear models for "units" are actually models for groups of units, even though it is not clear how to think about membership (or degree-of-membership) of each unit to each group. While a vague concept of group membership may be acceptable, it is at

least necessary to define and bound an *effective* count of units underlying each set of relative frequency estimates $\hat{\Pi}_i^* := \{\pi(\mathbf{y}, x_{i\cdot})\}_{\mathbf{y}\neq 0}$ to facilitate variance estimation.

Finally, we have no reason to expect that locally-selected log-linear models are consistent with a smooth imputation function $\pi(\mathbf{0}, \mathbf{x})$. Discontinuity may result from the discrete nature of the standard approach to model selection: each coefficient is either included in the model or it is not. We will reserve smooth generalizations of model selection for future work.

3.5 Estimation for Three Lists: Simulation Example

We continue with the simulation illustrated in Figure 1. Using Bernoulli draws according to the capture probabilities illustrated in Figure 1, we generate the extended data $(y_{\cdot\cdot}, x_{\cdot\cdot})$. Next, we use the `np` package in `R` (see Hall and Racine [27]) to estimate the conditional probability functions Π^*. These conditional probability functions are stacked in Figure 4 as the curved black lines.

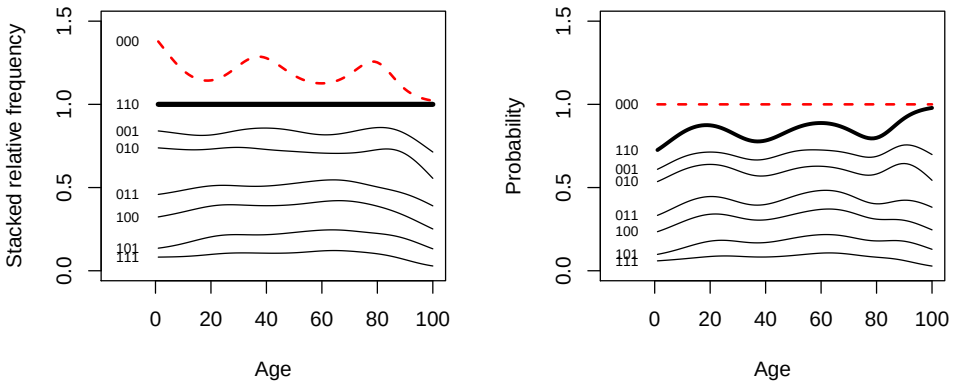

Figure 4: Stacked conditional probability functions. For example, the curve labeled "100" represents the sum $\hat{\pi}((1,0,0), Age) + \hat{\pi}((1,0,1), Age) + \hat{\pi}((1,1,1), Age)$. The dashed curve represents the imputed value of $\hat{\pi}((0,0,0), Age)$. The curves are normalized in the right panel so that the bold solid curve represents the estimated detection probability.

For each $i = 1, ..., n_c$, we fit a log-linear model to the set of values $\hat{\Pi}_i^*$ as in Equation 24. The optimal choice of parameters to include in the log-linear model is complex (see Section 3.4); for our example we simply apply the saturated hierarchical log-linear model's maximum likelihood estimate as in Equation 23. This imputation is plotted as the dashed curve in Figure 4, stacked on top of the other conditional probabilities. The bold curve in the right panel of Figure 4 represents the estimate of the detection function $\hat{\psi}(Age)$ implied by Equation 19, and a Horvitz-Thompson estimate is immediate.

Our result in the simulation above is substantially biased. In 25 replications, population size estimates ranged from 2005 to 2204 with a median of 2101 (recall that the true population size is 2000). We suggest that a primary reason for the bias is that the saturated hierarchical log-linear model is not adequate. In particular, the log-linear model does not include the highest-order interaction coefficient ρ, which is significant. The problem of how to incorporate our uncertainty regarding ρ into our estimates of population size is the subject of goal G3.

One way to reduce the role of highest-order interaction in our simulation setting is to remove the respondent fatigue effects, so that the third and fourth panels in Figure 1 contain only the upper curve. We apply this simplification and generate 25 replications of our Horvitz-Thompson estimator. This time, the median was exactly 2000 (!), the minimum was 1927, and the maximum was 2056. We provide a more thorough examination of model performance in Appendix A.

4 Discussion and Next Steps

4.1 Model Validation

In a textbook regression problem, a fixed set of covariates and a single response with many replicates make possible cross-validation and clear measures of predictive power under only mild assumptions. By contrast, CRC methods push the limits of what can be accomplished by principled inference. CRC requires extremely strong assumptions, offers nothing particularly compelling in the way of model checking (it's hard to quantify the likelihood of a vast school of "wily trout"), and, even under favorable conditions, the prediction intervals may be too wide to be of much value. [1]

The CRC problem is fundamentally a missing data problem. The missing quantity of interest is n, the population size, but, perhaps more relevantly, the covariate values $x_{(n_c+i)\cdot}$ are missing for $i = 1, ..., n - n_c$. The nature of this missingness is arguably of the worst possible kind, because it is reasonable to suppose that the units which are not observed are not observed precisely because they are *different* from the observed units. While differences between the training data and the test data are always a concern in prediction problems, it is not generally the case that the prediction data is different from the training data *a priori*.

The absence of validation data (the missing data) makes simulation an especially important tool for CRC modeling. However, the possibilities for simulation are endless. The variability of simulation settings means that finding the "best" CRC estimator is like shooting at a moving target. In particular, comparisons between different CRC algorithms do not typically provide a clear winner. Rather than seek an optimal estimator, our goal is more vague: to generate reasonable estimators and understand their behavior across a spectrum of simulation scenarios.

4.2 A Multi-level Bias-Variance Tradeoff

Our estimator involves three distinct estimation steps, each of which has its own bias-variance tradeoff. The first step is estimating the functions Π^*, and this involves a variable selection and/or bandwidth selection problem. The second bias-variance tradeoff occurs in the local selection of log-linear models, which may emphasize parsimony to a greater or lesser degree. Finally, biasing the estimated detection function $\hat{\psi}(\mathbf{x})$ upwards may reduce the variance of the resulting Horvitz-Thompson estimator.

In our present approach, each of these three bias-variance tradeoffs is optimized separately, whereas a joint optimization could lead to very different results. For example, it could conceiveably be optimal to undersmooth the conditional density estimates, overfit the log-linear models, and restrict the variance only in the final step by imposing an upward bias in the detection probabity estimates. This kind of joint optimization is the subject of our research goal G6.

Notably, Yip et. al. made significant progress towards unifying the first two steps – estimating Π^* and imputing $\pi(\mathbf{0}, \mathbf{x})$ [24]. However, Yip's fundamental approach is different than ours, relying on a conditional maximum likelihood approach that fits all list interactions and covariate effects jointly only on the observed portion of the data. In effect, Yip fits list interactions using only the relative frequencies for observable capture patterns Π^*. By contrast, we fit the list-interaction (i.e., log-linear modeling) component to the full set of functions Π locally, acknowledging that one function is missing.

4.3 Proposed Algorithm Overview

Here is a high-level review of our proposed CRC program:

1. **Preprocess:** The inputs are the lists $L_1, ..., L_k$ and associated covariates $x^c_{\cdot\cdot}(L_j)$ that come with the jth list, $j = 1, ..., k$. The first pre-processing step is to generate the cross-classification table $\mathbf{c}$ of counts of each observable capture pattern. The second pre-processing step is to identify covariates that are common across all k covariates $x^c_{\cdot\cdot}(L_j)$ and merge them into a single set of covariates $x^c_{\cdot\cdot}$.
2. **Model the relative capture pattern frequencies as a function of the covariate space:** Use a conditional density estimation or regression technique to estimate the functions in Π^*. Evaluate these functions at each $x_{i\cdot}$ for all observed units $i = 1, ..., n_c$ to generate n_c sets $\hat{\Pi}^*_i$.
3. **Impute the missing cell:** Fit parsimonious log-linear models $\mathcal{M}_i$ to the estimates $\hat{\Pi}^*_i$. Use the fitted models to impute the relative frequencies of the unobserved capture pattern $\{\pi(\mathbf{0}, x_{i\cdot})\}_i$. Note that the log-linear model selection may involve estimating the effective counts n_i underlying each Π^*_i.

[1] These obstacles make CRC an ideal topic for a Ph.D. dissertation: No matter how much progress we make, we will never arrive at a tidy solution, leaving endless avenues of marginal utility for future work.

4. **Horvitz-Thompson population estimation:** Estimate the detection function as in Equation 19 and compute the Horvitz-Thompson estimator.

4.4 Timeline for Next Steps

Section 3.1 presented a list of research goals. The foundational result proposed by G1 was essentially completed in Section 3.2. Goals G2 and G3 were more fully introduced in Sections 3.4 and 3.5. Our simulation examples throughout provided a starting point for goal G4. We have not yet deeply explored goal G5, as we do not expect access to the data for at least several months. Finally, goal G6 was discussed in Section 4.2.

We propose the following timeline for completion of tasks related to each of the goals:

Tomorrow: Replicate the log-linear model selection approach of Fienberg [14] for a simple three-list simulation example. Modify the simulation so that no highest-order interaction exists – while including a respondent fatigue effect – and verify that the proposed algorithm (Section 4.3) produces reasonable results (goals G2, G4).

Early September 2012: Implement a pseudo-hierarchical log-linear model by including a moderately informative prior for the highest-level interaction term as discussed in Section 3.4 (goal G3). Test the accuracy of the new confidence intervals by using simulations that involve second-order list interaction (goal G4).

October 2012: For a particular smoother $\hat{\Pi}^*$, find a lower bound for the local effective counts (see Section 4.3) and apply a pseudo-hierarchical log-linear model locally, using simulation to test the accuracy (goals G2, G3, G4).

January 2013: Complete goal G4.

April 2013: In terms of bias and variance, compare our estimator against existing approaches.

July 2013: Demonstrate our methods by combining a list of records from the 2010 Census, the 2010 Census Coverage Measurement (CCM) survey, and the American Community Survey (ACS) (goal G5).

October 2013: Pursue goal G6 as far as possible.

December 2013: Dissertation defense.

References

[1] Patrick J. Cantwell. DSSD 2010 census coverage measurement memorandum series #2010-g-01, May 2012. U.S. Census Bureau.

[2] Howard Hogan. The 1990 post-enumeration survey: Operations and result. *Journal of the American Statistical Association*, 88(423):1047–1060, 1993.

[3] G.S. Wolfgang. Using administrative lists to supplement coverage in hard-to-count areas of the post-enumeration survey for the 1988 census of st. louis. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pages 669–674, 1989.

[4] Anne Chao and P. K. Tsay. A sample coverage approach to multiple-system estimation with application to census undercount. *Journal of the American Statistical Association*, 93(441):283–293, 1998.

[5] John N. Derroch, Stephen E. Fienberg, Gary F. V. Glonek, and Brian W. Junker. A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability. *Journal of the American Statistical Association*, 88(423):1137–1148, 1993.

[6] Alan M. Zaslavsky and Glenn S. Wolfgang. Triple-system modeling of census, post-enumeration survey, and administrative-list data. *Journal of Business and Economic Statistics*, 11(3):279–288, 1993.

[7] Ivan P. Fellegi and Alan B. Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210, 1969.

[8] C. G. Joh. Petersen. The yearly immigration of young plaice into the limfjord from the german sea. *The Danish Biological Station*, 1895.

[9] Stephen E. Fienberg, Matthew S. Johnson, and Brian W. Junker. Classical multilevel and bayesian approaches to population size estimation using multiple lists. *Journal of the Royal Statistical Society*, 162(3):383–405, 1999.

[10] Sophie Baillargeon and Louis-Paul Rivest. Rcapture: Loglinear models for capture-recapture. *Journal of Statistical Software*, 19(5), 2007.

[11] P. A. P. Moran. A mathematical theory of animal trapping. *Biometrika*, 38(3/4), 1951.

[12] K. P. Burnham and W. S. Overton. Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika*, 65(3):625–633, 1978.

[13] Anne Chao. Estimating the population size for capturerecapture data with unequal catchability. *Biometrics*, 43(4):783–791, 1987.

[14] Stephen E. Fienberg. The multiple recapture census for closed populations and incomplete 2^k contingency tables. *Biometrika*, 59(3):591, 1972.

[15] Panel to Review the 2000 Census. *The 2000 Census: Counting Under Adversity*. National Academies Press, 2004.

[16] Juha M. Alho. Logistic regression in capture-recapture models. *Biometrics*, 46(3):623–635, 1990.

[17] Kenneth H. Pollock, James E. Hines, and James D. Nichols. The use of auxiliary variables in capture-recapture and removal experiments. *Biometrics*, 40(2):329–340, 1984.

[18] Juha M. Alho, Mary H. Mulry, Kent Wurdeman, and Jay Kim. Estimating heterogeneity in the probabilities of enumeration for dual-system estimation. *Journal of the American Statistical Association*, 88(423):1130–1136, 1993.

[19] Song Xi Chen, Cheng Yong Tang, and Jr. Vincent T. Mule. Local post-stratification in dual system accuracy and coverage evaluation for the u.s. census. *Journal of the American Statistical Association*, 105(489):105–119, 2012.

[20] Song Xi Chen and Chris J. LLoyd. A nonparametric approach to the analysis of two-stage mark-recapture experiment. *Biometrika*, 87(3):633–649, 2000.

[21] Song Xi Chen and Chris J. LLoyd. Estimation of population size from biased samples using non-parametric binary regression. *Statistica Sinica*, 12, 2002.

[22] Stuart G. Baker. A simple em algorithm for capture-recapture data with categorical covariates. *Biometrics*, 46(4):1193–1200, 1990.

[23] Shirley Pledger. Unified maximum likelihood estimates for closed capture-recapture models using mixtures. *Biometrics*, 56(2):434–442, 2000.

[24] Paul S. F. Yip, Emmy C. Y. Wan, and K. S. Chan. A unified approach for estimating population size in capture-recapture studies with arbitrary removals. *Journal of Agricultural, Biological, and Environmental Statistics*, 6(2):183–194, 2001.

[25] Patrick J. Cantwell. DSSD 2010 census coverage measurement memorandum series #2010-g-10, May 2012. U.S. Census Bureau.

[26] Daniel Manrique-Vallier and Stephen E. Fienberg. Population size estimation using individual level mixture models. *Biometrical Journal*, 50(6):1–13, 2008.

[27] Peter Hall, Jeff Racine, and Qi Li. Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association*, 99(468):1015–1026, 2004.

A Continuation of Section 3.5

To compare the performance of our algorithm performs is to compare it with the performance of an existing method, we consider estimating the population size using a log-linear model *globally*, without taking Age into account. We apply a saturated log-linear model globally for the simulation illustrated in Figure 1, as well as the variation on this simulation that excludes the respondent fatigue effects. Figure 5 describes the results.

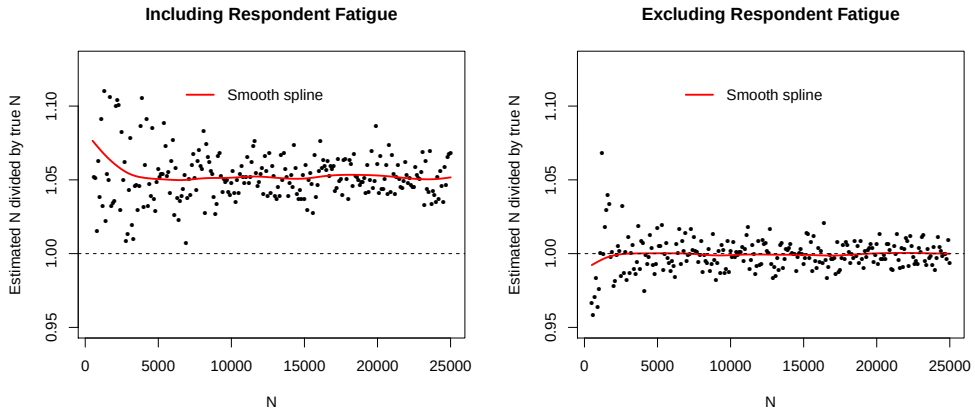

Figure 5: The log-linear model applied globally behaves in a way that is remarkably similar to the performance of our local log-linear approach near $n = 2000$. In the left panel, estimates are biased upwards by about 5%. After eliminating respondent fatigue, the estimates are essentially unbiased (right panel). In addition, the mean and variance of the estimates at $N = 2000$ do not differ dramatically from the local log-linear approach.