CARNEGIE MELLON UNIVERSITY

LOCAL LOG-LINEAR MODELS FOR CAPTURE-RECAPTURE

A DISSERTATION SUBMITTED TO THE GRADUATE SCHOOL IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY
In
STATISTICS

by

ZACHARY TODD KURTZ

Department of Statistics
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

January, 2014

Abstract

Capture-recapture (CRC) models use two or more samples, or lists, to estimate the size of a population. In the canonical example, a researcher captures, marks, and releases several samples of fish in a lake. When the fish that are captured more than once are few compared to the total number that are captured, one suspects that the lake contains many more uncaptured fish. This basic intuition motivates CRC models in fields as diverse as epidemiology, entomology, and computer science.

We use simulations to study the performance of conventional log-linear models for CRC. Specifically, we evaluate model selection criteria, model averaging, an asymptotic variance formula, and several small-sample data adjustments. Next, we argue that interpretable models are essential for credible inference, since sets of models that fit the data equally well can imply vastly different estimates of the population size. A secondary analysis of data on survivors of the World Trade Center attacks illustrates

this issue.

Our main chapter develops *local* log-linear models. Heterogeneous populations tend to bias conventional log-linear models. Post-stratification can reduce the effects of heterogeneity by using covariates, such as the age or size of each observed unit, to partition the data into relatively homogeneous post-strata. One can fit a model to each post-stratum and aggregate the resulting estimates across post-strata. We extend post-stratification to its logical extreme by selecting a local log-linear model for each observed point in the covariate space, while smoothing to achieve stability.

Local log-linear models serve a dual purpose. Besides estimating the population size, they estimate the rate of missingness as a function of covariates. Simulations demonstrate the superiority of local log-linear models for estimating local rates of missingness for special cases in which the generating model varies over the covariate space. We apply the method to estimate bird species richness in continental North America and to estimate the prevalence of multiple sclerosis in a region of France.

Preface

Thanks to William F. Eddy, my thesis adviser, for encouraging me in my studies, starting in the Summer of 2010, when he set me to work predicting when buses would arrive at bus stops (an impossible task, it turned out). His investment of attention to my work was a major motivating force which inspired me to continually improve my analyses. Dr. Eddy also deserves full credit for kindling my interest in the subject of this thesis. At the end of 2011, as I cast about for a thesis topic, he pointed out that his Census research group needed someone to work on capture-recapture. Although I had never heard of the topic, I was at that time sufficiently aimless to be easily convinced.

I owe much of my progress to Stephen E. Fienberg, a "founding father" of log-linear modeling. My work builds directly on his. Dr. Fienberg's encyclopedic knowledge of the CRC literature proved useful, as I often went to him with the question of where to read more about this or that approach.

Cosma Shalizi has been supportive of me as a "technical" adviser, taking the time to go over detailed equations, or to help me see the bigger picture that would lead to a solution. Rebecca Steorts often reviewed my

manuscripts and gave many valuable suggestions. I also wish to thank Bruce Spencer (Northwestern University) for serving as an external adviser.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background and Summary

Capture-recapture (CRC) is the science of estimating the size of a population by using multiple incomplete lists. A list is a collection of units from some population, and we refer to the act of generating a list as a *capture*. Examples of populations studied using CRC include various animal species (Odum and Pontin, 1961; Pollock et al., 1984), human populations (Chen et al., 2010), the set of websites on a given topic (Fienberg et al., 1999), and the set of computer coding errors in a body of code (Runeson and Wohlin, 1998), to name just a few. Table 2.1 in Chapter 2 gives a much larger list of CRC applications.

This thesis introduces new methods for the underlying statistical problem: How to estimate the unknown size n of some population from k different incomplete lists $L_1, ..., L_k$ of the population units. We review some basic background on CRC before introducing our methods. In the simplest

setting, we are given two lists of units, List 1 and List 2. Assume that units captured on each list can be perfectly identified across lists, such that the exact number of distinct observed units is known. Then it is possible to construct the cross-classification of units according to list membership as displayed in Table 1.1.

Table 1.1: Contingency table for a two-list experiment

		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 has many challenging variations that involve additional lists, auxiliary covariates, population dynamics, measurement error, and inter-list dependence structure.

For the two list case, the Petersen estimator is

$$\hat{c}_{00} = \frac{c_{10}c_{01}}{c_{11}}, \tag{1.1}$$

which can be formalized as a maximum-likelihood estimator under certain assumptions (Feller, 1968; Pollock, 1976). Perhaps the strongest of these assumptions is that the lists are independent; 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. However, two kinds of dependence between lists have been thoroughly examined. The first is *unit-level list dependence*, in which previous capture directly affects the probability of subsequent capture. The second kind of dependence arises indirectly as a consequence of *heterogeneity*, or variability in capture probabilities across units (Fienberg et al., 1999).

Both sources of dependence may depend on covariates such as age, and much of the CRC literature in the last three decades addresses this fact. Sekar and Deming (1949) described post-stratification, one of the earliest methods of using auxiliary covariates to account for heterogeneity. Huggins (1989) and Alho (1990) derived logistic regression models for heterogeneity for the two-list scenario, and Yip et al. (2001) extended the logistic model to include k lists and a simple respondent fatigue effect. Chen and Lloyd (2002) introduced nonparametric regression into CRC models for two lists. Chapter 2 reviews several more models that allow for smooth dependence on continuous covariates.

Broadly, the CRC literature can be divided into two categories. The first category of models considers only the cross-classification of population units according to list membership as in Table 1.1. The second category of models considers – in addition to the cross-classification array – auxiliary covariates such as age, size, or gender. This thesis addresses both types of models, in Chapter 4 and Chapter 5, respectively. Throughout, we restrict attention to closed populations, ignoring births, deaths, and migration.

Chapter 4 addresses the problem of model selection when no auxiliary covariates are available. We connect recent literature on nonidentifiability of

CRC models to the relatively well-known nonidentifiability of the highest-order interaction in log-linear models, and we argue that interpretable models (in contrast to algorithmic, black-box methods) are essential for credible inference.

Chapter 5 describes our main work, a local conditional likelihood approach that allows both heterogeneity-induced and unit-level list dependence to depend on unit-level covariates (such as age) in a very general way. Specifically, we estimate the relative frequency of the unobserved capture pattern (no captures) by applying log-linear models locally. We project the fitted local log-linear models onto the missing cell to produce unit-level estimates of the "rate of missingness." Summing the rate of missingness across all units gives an estimate of the total number of undetected units. The fact that we select a different model at each point in the covariate space distinguishes our method from similar previous methods.

Chapter 6 illustrates local log-linear models on real data. Our first example uses data from the North American Breeding Bird Survey to estimate the number of different bird species that can be observed in continental North America. To our knowledge, this is the first instance of using CRC to estimate bird species richness over a large region. Our second example uses the data of El Adssi et al. (2012) to estimate the prevalence of multiple sclerosis in the Lorraine region of France. Finally, we discuss the possible future application of our method to estimate the coverage of the U.S. Census.

Chapter 7 explores the consequences of record linkage error for CRC estimation. We define several possible kinds of record linkage error, and suggest simulation methods for incorporating uncertainty of record linkage

in the final CRC estimates.

The remainder of this introduction provides notation and terminology to be used throughout the thesis.

1.2 Notation

Notation for CRC varies widely across the literature. We propose a new system of notation to facilitate a clear and consistent discussion throughout our literature review and subsequent analyses. Part of the notation may seem bulky at first, but later proves useful for a rigorous statement of assumptions.

1.2.1 Notation for Describing 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 consist 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 $\mathcal{Y}$ denote the set of binary row vectors of length k. For example, with two lists, $\mathcal{Y} = \{(1,1),(1,0),(0,1),(0,0)\}$. Note that each row $y_{i\cdot}$ is an element of $\mathcal{Y}$. For every $y \in \mathcal{Y}$, define $c_y := |\{i : y_{i\cdot} = y\}|$. Then the array $\mathbf{c} := \{c_y\}_{y\in\mathcal{Y}}$ is the contingency table of counts of units in the lists $L_1, ..., L_k$. In particular, $c_{\mathbf{0}} = n - n_c$, the unknown 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 $\hat{n} = \hat{c}_{\mathbf{0}} + n_c$.

1.2.2 Subscripting by ω

Let $\mathcal{K} = \{1, ..., k\}$. Let Ω denote the power set of $\mathcal{K}$, excluding the empty set. For example, if there are $k = 3$ lists, then $\Omega = \{\{1\},\{2\},\{3\},\{1,2\},\{1,3\},\{2,3\},\{1,2,3\}\}$. Let $\omega \in \Omega$, 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 $\omega \in \Omega$. 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 $y = (y_1, ..., y_k) \in \mathcal{Y}$, let $y_\omega := (y_{\omega_{(1)}}, ..., y_{\omega_{(|\omega|)}})$. For the special case in which ω is a singleton $\{j\}$, we write $y_{i\{j\}} = y_{ij}$ and $y_{\{j\}} = y_j$. Take ω^c to be the complement of ω. For example, let $y = (1,1,0) \in \mathcal{Y}$. Then $y_{\{2,3\}} = (1,0)$, $y_1 = y_{\{1\}} = y_{\{2,3\}^c} = 1$, and $y_{\{1,2,3\}} = y$.

1.2.3 Notation for Probability Models

Each capture pattern $y_{i\cdot}$ is 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 statis-

tics $\mathbf{c}$ 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 thesis, unless specified otherwise, let $k > 1$, and let $j \in \mathcal{K}$, $i \in \mathcal{N}$, $y \in \mathcal{Y}$, and $\omega \in \Omega$ be arbitrary.

Let $p(i, y) = P(Y_{i\cdot} = y)$, the probability that unit i has capture pattern y. Then $p(i, y_{i\cdot}) = P(Y_{i\cdot} = y_{i\cdot})$. Similarly, let $p_\omega(i, y) = P(Y_{i\omega} = y_\omega)$. Define $\mathbf{p}(i, \mathcal{Y}) := \{p(i, y)\}_{y\in\mathcal{Y}}$. Hence, we may regard the the capture pattern $Y_{i\cdot}$ as a multinomial random variable with a single trial and multinomial probabilities $\mathbf{p}(i, \mathcal{Y})$, so $Y_{i\cdot} \sim \text{Multi}(1, \mathbf{p}(i, \mathcal{Y}))$.

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

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 lists 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 let $\phi_j := n^{-1} \sum_{i\in\mathcal{N}} \phi_j(i)$.

1.2.4 Regression Notation

Let $\mathcal{X}$ denote the covariate space, and let $x \in \mathcal{X}$ be arbitrary. A function $r(y|x)$ 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 $r(y|x)$, define $r_\omega(y|x) := \sum_{z\in\mathcal{Y}:z_\omega=y_\omega} r(z|x)$.

Given a function $r(y|x)$, define the detection function $\psi_\omega(x) = 1 - r_\omega(\mathbf{0}|x)$, which can be interpreted as the probability that a unit with covariates 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 $r(y|x)$ is a regression model, then $\psi(x_{i\cdot}) = \phi(i)$, $\psi_\omega(x_{i\cdot}) = \phi_\omega(i)$, and $\psi_j(x_{i\cdot}) = \phi_j(i)$.

1.3 General Terminology

We call a population *closed* if the population is fixed during the generation of the lists $L_1, ..., L_k$. This excludes births, deaths, and migration. Populations which are not closed are *open*. This thesis considers only closed population models.

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 (Fellegi and Sunter, 1969). Most of the CRC literature assumes that the lists are linked perfectly, so that the cross-classification counts $\mathbf{c}$ are all observable except for $c_{\mathbf{0}}$. This is called the *perfect record linkage* assumption.

A CRC experiment is called *homogeneous* if the capture probabilities are constant across units. To be precise, an experiment is homogeneous if $p(i_1, y) = p(i_2, y)$ for every pair of units i_1, i_2 and every $y \in \mathcal{Y}_k$. Many CRC papers seem to use *heterogeneity* to mean the absence of this specific kind of homogeneity, even though some related aspects of population units can be "heterogeneous" in a more general sense.

The term "independence" describes – rather ambiguously – the relation-

ships between population units or between captures across lists. We clarify matters by precisely stating four distinct notions of independence. The *lists are independent at the unit level* if

$$p(i, y) = p_\omega(i, y)p_{\omega^c}(i, y). \tag{1.2}$$

Marginally (i.e., when $\omega = \{j\}$, a specific list), list independence at the unit level implies that the event that unit i is on a specific list is independent of the event that unit i is on any combination of the other lists. In the context of a regression model $r(y|x)$, list independence at the unit level is equivalent to *conditional independence*:

$$r(y|x) = r_\omega(y|x)r_{\omega^c}(y|x) \tag{1.3}$$

For example, suppose $k = 2$ with $y = (y_1, y_2)$ and $\omega = \{1\}$. Then $\omega^c = \{2\}$, and conditional independence implies

$$r(y|x) = \left[\psi_1(x)^{y_1}(1 - \psi_1(x))^{1-y_1}\right]\left[\psi_2(x)^{y_2}(1 - \psi_2(x))^{1-y_2}\right]. \tag{1.4}$$

The *lists are independent* if

$$p(y) = p_\omega(y)p_{\omega^c}(y). \tag{1.5}$$

It is important to understand the difference between the assumptions 1.5 and 1.2: Heterogeneity can cause list dependence even if the lists are independent at the unit-level (Fienberg et al., 1999). Finally, *independence of units* (or

independence between units) means that the capture pattern of a unit does not depend on the capture pattern of other units.

The basic *multinomial sampling model* assumes homogeneity (at least formally) and independence between units to get

$$P(\mathbf{C} = \mathbf{c}) = \frac{n!}{\prod_{y \in \mathcal{Y}} c_y!} \prod_{y \in \mathcal{Y}} p(y)^{c_y}. \tag{1.6}$$

A *unit-level* or *regression* multinomial sampling model uses a regression model to relax the homogeneity assumption, so that the sampling distribution is multinomial at the unit level:

$$P(Y_{\cdot\cdot} = y_{\cdot\cdot} | x_{\cdot\cdot}) = \prod_{i \in (1,\ldots,n)} r(y_{i\cdot} | x_{i\cdot}). \tag{1.7}$$

Note that both of these multinomial sampling models require the assumption that units are independent. Both models fail if, for example, the inclusion of a child on a list of people depends on the inclusion of that child's caretaker. (However, it could be argued that a child-parent dependence matters less if the regression model r accounts for age.) A few authors have studied dependence between units in a CRC setting. Cowan and Malec (1986) modeled household-induced dependence in a two-list census. Anderson et al. (1994) estimated an overdispersion parameter to reflect dependence between units.

A commonly used convention, perhaps originated by Otis et al. (1978), is to denote a model as $M_{\text{subscripts}}$, where the possible subscripts are t, b, and h. Models that allow capture probabilities to vary between lists (for

a fixed population unit) are indexed by t, which stands for *time*; models that include unit-level list dependence (also known as a *behavioral* effect) are indexed by b; and models that allow for heterogeneity are indexed by h. In this paradigm, the most general model is M_{tbh}, and various submodels M_{tb}, M_{bh}, etc., result from imposing constraints. In particular, let M_0 denote the model in which a single capture probability applies to every unit and every list. This notation qualitatively defines a hierarchy of eight different kinds of CRC models.

Chapter 2

Literature Review

2.1 A Reviewer's Review

A journal-article length review could not hope to adequately summarize the full CRC literature. For most work prior to the year 2000, we defer to previous reviewers, allowing the present review to focus on recent developments. Our exposition relies heavily on the notation and terminology presented in Sections 1.2 and 1.3. We discuss only closed populations, although many of our sources include extensions to open populations.

Schaefer (1951) reviewed the early history of CRC, tracing the theory of the Petersen estimator (1.1) as far back as Laplace in 1783. Cormack (1968) published a major synthesis, followed by a more technical review (Cormack, 1979). Otis et al. (1978) is a highly-cited monograph. Seber (1982) and Seber (1986) gave two additional reviews.

At least five full-length review articles appeared in 1990's alone. Pollock (1991) was remarkably broad in scope. Fienberg (1992) provided a rather

exhaustive bibliography of CRC papers with minimal commentary. Seber (1992) reviewed methods from the animal populations setting. The International Working Group for Disease Monitoring and Forecasting (1995a,b) reviewed CRC models with an emphasis on applications to epidemiology. Schwarz and Seber (1999) updated previous reviews by Seber on CRC methods for animal populations. Fienberg et al. (1999) briefly summarized the Bayesian developments up to that time.

Chao (2001) reviewed closed population models including continuous-time CRC models, which treat every record as a new list. Pollock (2000) gave a brief literature review, followed by a second review on the use of auxiliary covariates to model heterogeneity (Pollock, 2002). Pledger and Phillpot (2008) summarized models for heterogeneity that do not rely on observable covariates. Huggins and Hwang (2011) reviewed conditional likelihood estimation, including models with covariates. Finally, a dissertation by Stoklosa (2012) thoroughly references most of the recent sources relevant for our work, including approaches that use covariates to model unit-level list dependence and heterogeneity.

We structure the remainder of this review as follows. Section 2.2 gives a brief history of models which do not directly incorporate covariate information. Section 2.3 describes some basic regression models. Section 2.4 traces the development of the likelihood function as it parallels the emergence of models that regress capture probabilities as functions of covariates. Section 2.5 broadly reviews a selection of the data sets and simulation environments which have been used to test estimators. Throughout, we characterize models by using the "M_{tbh}" notation introduced at the end of Section 1.3.

2.2 Models without Covariates

2.2.1 The Petersen Estimator (M_t)

In the 1890's, Petersen re-discovered and popularized an estimator that provides the fundamental inspiration for most modern models (Pollock, 1991; Petersen, 1895). 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 1.5. Independence implies that $p((1,1)) = \phi_1\phi_2$, and one may hypothesize 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 and behavioral effects cause the failure of the independence assumption, so the Petersen estimate is rarely optimal. In applications with more than two lists, 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 (Fienberg et al., 1999).

The expectation of the Petersen estimator is not defined since it can happen that $c_{11} = 0$. Conditioning on $c_{11} \neq 0$, the Petersen estimator is biased upwards for small sample sizes, even when all the standard assumptions hold. The Chapman estimator is a slight modification of the Petersen estimator that reduces the bias by adding a small positive number to each element of the cross-classification $\mathbf{c}$ (Chapman, 1951). Evans and Bonett (1994), Hook and Regal (1997), and Rivest and Lévesque (2001) proposed

generalizations of Chapman's modification for $k > 2$ lists.

2.2.2 Some Early Models (M_t, M_0, M_h)

Sequels to the Petersen estimator sought to generalize its applicability to scenarios with more than two lists. Schnabel (1938) presented some of these methods in the form M_t. Several other models, described in the following paragraphs, assume that the probability of capture for each unit is constant across lists.

The basic *removal method* (M_0), introduced by Moran (1951), requires that captured units are either literally removed, or are marked before release back into the population and are not 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.

Chapter 3 reviews log-linear models (Fienberg, 1972; Cormack, 1989), which are of primary interest for this thesis. Here we merely note that log-linear modeling remains one of the most popular CRC techniques, especially for experiments involving 3-6 capture occasions (see Table 2.1).

Burnham and Overton (1978) derived an estimator of type M_h based on the generalized jackknife method. In an experiment with k captures, let z_j denote the number of units that are captured exactly j times, $j = 1, ..., k$. The idea of the jackknife estimator is express the estimate $\hat{n}$ as a linear combination of the quantities z_j. Note that $n_c = \sum_j z_j$. For some constant α_1, the first order jackknife estimator is $\hat{n} = n_c + \alpha_1 z_1$, the second order

estimator takes the form $\hat{n} = n_c + \alpha_1 z_1 + \alpha_2 z_2$, and so on. The jackknife model is designed to function well under some types of heterogeneity.

Chao (1987) derived the "Chao's lower bound" estimator (M_h), 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. Later, Chao et al. (1992) relaxed the requirement that the probability of capture for each unit is constant across lists.

2.2.3 Recent Methods without Covariates

In the 1980's and early 1990's, several authors developed a model (M_{th}) that uses martingales (Pollock, 2002). Lloyd (1992) observed that the sampling distribution of martingale estimators can be somewhat unstable, and we observe that the use of martingales in CRC has not become mainstream. However, Huggins (2006) used a theory of martingales for new results in modeling open populations.

Chao et al. (1992) used a series of approximations and mathematical identities to produce an estimator of type M_{th} with heterogeneity in the form of random effects. This estimator is nonparametric in the sense that it makes no explicit assumption about the distribution of the random effects. To our knowledge, no one has synthesized Chao's approach with models for which assumptions about the distribution of capture probabilities are easily made explicit, such as the quasi-symmetry model in Darroch et al. (1993),

and it is difficult to characterize the populations for which Chao's estimator performs well without resorting to simulation.

Following in a flavor similar to the Rasch quasi-symmetry model, Norris and Pollock (1995, 1996a) assumed that capture probabilities are drawn from some unknown mixing distribution and developed a nonparametric maximum-likelihood estimator for that distribution. Related mixture models and latent class models have also seen recent interest (Pledger and Phillpot, 2008). Notably, Link (2003) showed a lack of identifiability for many of these kinds of models. We further address identifiability issues in Chapter 4.

Several Bayesian CRC models incorporate heterogeneity. Fienberg et al. (1999) compared a Bayesian hierarchical model against quasi-symmetry models and other log-linear models. Basu and Ebrahimi (2001) chose prior distributions to Bayesianize a log-linear model that accounts for heterogeneity and unit-level list dependence (M_{tbh}). Manrique-Vallier and Fienberg (2008) modeled heterogeneity with a Bayesian grade of membership model (based on a latent membership covariate) in conjunction with the assumption of conditional independence between lists (M_{th}).

2.3 Models with Covariates

The idea of incorporating unit-level covariates in CRC appeared as early as Howard (1948) in the form of post-stratification, which partitions observed units into a collection of post-strata according to some set of categorical covariates. One may construct the cross-classification array $\mathbf{c}$ separately on each post-stratum, and estimate of the number of units for each post-stratum

before aggregating across post-strata to estimate the total number of missing units. Perhaps the earliest systematic treatment of post-stratification is in Sekar and Deming (1949). The official 1970, 1980, and 2000 U.S. Census coverage evaluations represent some of the largest applications of post-stratification (Citro et al., 2004). The methods in the following subsections incorporate covariates in increasingly nuanced ways; Chen et al. (2010) compared the performance of some of these methods against the relatively simple post-stratification approach.

2.3.1 Logistic Regression (M_{th})

Pollock et al. (1984) was the first to apply logistic regression for CRC, by allowing the probability of capture on each list to vary as a logistic function of list features or unit-level covariates. Several years later, Huggins (1989) and Alho (1990) used a slightly different – and ultimately more popular – version of logistic regression to estimate the unit-level detection function $\psi(x)$ by assuming that the probability of capture of each unit on each list is a logistic function of unit features x, such as age or sex in a human population. Alho et al. (1993) applied logistic regression for an informal Census coverage evaluation. The 2010 Census Coverage Measurement relied on logistic regression, unlike the previous coverage measurement programs which incorporated covariates via post-stratification (Olson and Griffin, 2012).

The Huggins-Alho procedure relies on the Horvitz-Thompson (HT) es-

timator of the population size n, which takes the form

$$\tilde{n} = \sum_{i:M_i=1} \frac{M_i}{\psi(x_{i\cdot})} = \sum_{i:M_i=1} \frac{1}{\psi(x_{i\cdot})} \tag{2.1}$$

The HT estimator uses the detection probabilities $\psi(x_{i\cdot})$ only for the units that are observed. If ψ is known, the HT estimator has some nice asymptotic properties. It is easy to verify that $E\tilde{n} = n$. Moreover, $\tilde{n}$ is consistent and asymptotically normal if $\psi(x_{i\cdot})$ is uniformly bounded away from 0 and 1 for all $i \in \mathcal{N}$ (Alho, 1990).

In some study designs (not in CRC), $\psi(x_{i\cdot})$ is known. However, to use (2.1) for CRC, we must estimate the detection function ψ. The expression that results from replacing ψ with an estimate $\hat{\psi}$ in (2.1) is technically not a HT estimator; some CRC authors refer instead to (2.1) as a "Horvitz-Thompson style" estimator. We continue to refer to both forms with "HT" for brevity.

Alho estimated $\psi(x_{i\cdot})$ only for the two-list case. For $j = 1, 2$, let θ_j be a $q \times 1$ vector of parameters, and $\theta := (\theta_1, \theta_2)$. Assume that the conditional probability $P(Y_{i\cdot} = y_{i\cdot}|x_{i\cdot}, \theta, M_i = 1)$ is a logistic function of $x_{i\cdot}$ and parameters θ. Alho estimated θ by maximizing the conditional likelihood function

$$L(\theta|(x^c_{\cdot\cdot}, y^c_{\cdot\cdot})) = \prod_{i:m_i=1} P(Y_{i\cdot} = y_{i\cdot}|x_{i\cdot}, \theta, M_i = 1).$$

The probabilities in this likelihood become identifiable under the conditional

independence structure of (1.4), which implies the following regression model

$$r(y|x) := \left[logit^{-1}(x\theta_1 y_1)^{y_1}(1 - logit^{-1}(x\theta_1 y_1))^{1-y_1}\right] \times \left[logit^{-1}(x\theta_2 y_2)^{y_2}(1 - logit^{-1}(x\theta_2 y_2))^{1-y_2}\right].$$

The detection probability is then $\psi_j(x) = logit^{-1}(x\theta_j)$, for $j = 1, 2$, and the assumption of conditional independence gives $r((1,1)|x) = \psi_1(x)\psi_2(x)$ so that $\psi(x) = \psi_1(x) + \psi_2(x) - \psi_1(x)\psi_2(x)$. Therefore, Alho's maximum likelihood estimator $\hat{\theta}$ leads directly to an HT estimator (2.1).

2.3.2 Nonparametric Regression Methods (M_{th})

Logistic regression 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 (Chen et al., 2010). In such cases, a nonparametric approach might be more fitting [sic].

Chen and Lloyd (2000) developed a two-list method that is 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 (1.5), 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 $c_{\mathbf{0}} = n - n_c$, one can reparameterize the

multinomial likelihood implied by (1.6) as

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

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 (1.3). Note that conditional independence does not imply general independence between lists (1.5), and the parameter α quantifies dependence only in the latter sense.

Chen and Lloyd (2002) also proposed a simpler nonparametric approach with two lists. Suppose that $r(y|x)$ is a regression model for $(x_{\cdot\cdot}, y_{\cdot\cdot})$, and let $\psi(x)$ be the detection function. Let $\omega(j) = \mathcal{K} \setminus \{j\}$. Assume conditional independence (1.3). 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 $\hat{\psi}_j(x)$ for $j = 1, 2$. Finally, conditional independence implies an estimate $\hat{\psi}(x)$, and an HT estimator (2.1) is immediate.

2.3.3 Recent Methods with Covariates

Baker (1990), followed by Evans et al. (1994), incorporated covariates in a log-linear model. Yip et al. (2001) extended the Huggins-Alho regression framework to build a parametric regression model with time effects, hetero-

geneity effects, and a simple behavioral effect (M_{tbh}). In a series of papers, Hwang and Huggins developed a partially nonparametric generalization of Yip's model for closed populations by allowing capture probabilities and behavioral effects to vary as nonparametric functions of auxiliary covariates (Huggins and Hwang, 2007; Hwang and Huggins, 2007, 2011). The next section references several additional ways of treating covariates.

2.4 Evolution of the Likelihood Function

This section reviews a few of the many ways that likelihood functions have appeared in the CRC literature. We devote a whole section to the likelihood function because of its usefulness for understanding how authors successively developed CRC models to reflect increasingly nuanced assumptions.

Most probability models for the counts $\mathbf{c}$ of capture patterns fall into one of three categories: hypergeometric, multinomial, or Poisson. A generalized hypergeometric distribution was popular early on because of its obvious similarity to performing a sequence of capture events in a finite population.

Darroch (1958) proposed applying the multinomial distribution for CRC, noting that it seems more appropriate than the hypergeometric distribution when the capture probabilities for each capture occasion (and *not* the sample sizes) are reasonably assumed to be fixed *a priori*.

Finally, Sandland and Cormack (1984) treated $\mathbf{c}$ as set of independent Poisson draws. Here, the model is definitively incorrect, by assigning nonzero probability to the event that the sum of the counts adds up to more than the population size n. Despite this potential pedagogical twist, parameter esti-

mates are equivalent under the Poisson and multinomial likelihoods, while certain asymptotic expansions of the variance lead to slightly more conservative prediction intervals under the Poisson model (Sandland and Cormack, 1984). The Poisson model in this setting could be a good example of a model that is wrong and yet useful, in accordance with the mantra.

The remainder of this section will build on the multinomial framework. For any population with independence between units (i.e., where the capture of one unit does not depend on the event of capture of any other unit), a fully general multinomial likelihood function is

$$P(Y_{\cdot\cdot} = y_{\cdot\cdot}) = \prod_{i \in (1,\ldots,n)} p(i, y_{i\cdot}). \tag{2.2}$$

Note that (2.2) assumes nothing regarding list-independence or independence of lists at the unit-level. If $r(y|x)$ is a regression model for $p(i, y_{i\cdot})$, then (2.2) becomes (1.7). If $r(y|x)$ is constant in x, the equation (2.2) becomes

$$P(Y_{\cdot\cdot} = y_{\cdot\cdot}|x_{\cdot\cdot}) = \prod_{i \in (1,\ldots,n)} p(y_{i\cdot}) = \prod_{y \in \mathcal{Y}_k} p(y)^{c_y} = P(Y_{\cdot\cdot} = y_{\cdot\cdot}|\mathbf{p}(\mathcal{Y}_k)). \tag{2.3}$$

Hence, the likelihood of the multinomial parameters $\mathbf{p} := \mathbf{p}(\mathcal{Y}_k)$, given the extended data, is

$$L^*(\mathbf{p}|Y_{\cdot\cdot} = y_{\cdot\cdot}) \propto P(\mathbf{C} = \mathbf{c}|\mathbf{p}) = \frac{n!}{\prod_{y \in \mathcal{Y}_k} c_y!} \prod_{y \in \mathcal{Y}_k} p(y)^{c_y}.$$

In applications, the extended data $(y_{\cdot\cdot}, x_{\cdot\cdot})$ are not observed. However, the

full likelihood function can be written in terms of the observed data $(y^c_{\cdot\cdot}, x^c_{\cdot\cdot})$ by taking n to be a parameter of the likelihood:

$$L(n, \mathbf{p}|\mathbf{c} \setminus c_{\mathbf{0}}) = \frac{n!}{(n-n_c)\prod_{y\neq\mathbf{0}} c_y!} p(\mathbf{0})^{n-n_c} \prod_{y\neq\mathbf{0}} p(y)^{c_y}. \tag{2.4}$$

Estimating n by maximizing L results in a consistent estimator if the capture probabilities for unobserved units are the same as for observed units. Although such homogeneity may seem improbable, the bulk of CRC work rests on some version of this assumption, for lack of a better alternative. See Chapter 4 for a deeper discussion on this point.

Since L has too many parameters to be identifiable (see Chapter 4), it is common to replace $\mathbf{p}$ with a parameterization $\mathbf{p}(\theta)$ of a dimension that is small enough to allow identifiability. Hence $p(y) = p(y|\theta)$. Let $\pi(y|\theta) = p(y|\theta)/(1-p(\mathbf{0}|\theta))$. Following Sanathanan (1972a), let

$$L_1(\theta|\mathbf{c} \setminus c_{\mathbf{0}}) = \frac{n_c!}{\prod_{y\neq 0} c_y!} \prod_{y\neq\mathbf{0}} \pi(y|\theta)^{c_y},$$

and

$$L_2(n|n_c, \theta) = \frac{n!}{n_c!(n-n_c)!} p(\mathbf{0}|\theta)^{n-n_c}(1-p(\mathbf{0}|\theta))^{n_c}. \tag{2.5}$$

Here, L_1 is called the conditional likelihood function because it conditions on the observed part of the data cross-classification $\mathbf{c} \setminus c_{\mathbf{0}}$, and L_2 is called the marginal likelihood. Note that

$$L(n, \theta|\mathbf{C} = \mathbf{c}) = L_1(\theta|\mathbf{c} \setminus c_{\mathbf{0}}) L_2(n|n_c, \theta). \tag{2.6}$$

Let n_L denote the estimate of n obtained by jointly maximizing $L(n, \theta|\mathbf{C} = \mathbf{c})$. Solving this optimization is not always easy, and a common method of approximating n_L is to maximize L_1 and L_2 in sequence, as follows: Let θ_{L_1} denote the maximizer of L_1, and let $n_{L_1L_2}$ denote the estimate of n that is obtained by plugging θ_{L_1} into L_2 and maximizing over n. Sanathanan (1972a) showed that, under mild conditions, the asymptotic distributions of n_L and $n_{L_1L_2}$ are equal. Thus,

$$n_{L_1L_2} \approx n_L \tag{2.7}$$

when the true population size n is large, and several authors cited this result to justify the use of $n_{L_1L_2}$ in the context of log-linear models. We refer to (2.7) simply as Sanathanan's Theorem, although her original result was considerably more nuanced.

Beginning with equation (2.3), we developed the equations above with the assumption of homogeneity (i.e., $r(y|x)$ is a regression model that is constant in the covariates x). Of course, the point of having a regression model is to allow $r(y|x)$ to vary in x, thereby modeling heterogeneity. In this vein, several recent authors proposed maximum-likelihood estimation for parametric regression models, such as the logistic model of Huggins/Alho. These approaches require a more nuanced look at the likelihood function. Returning to (2.2), assume that $r(y|x) = r(y|x, \theta)$ is a parametric regression model. The analogue of (2.4) for the regression context is

$$L(n, \theta|(x^c_{\cdot\cdot}, y^c_{\cdot\cdot})) = \prod_{i\in(1,\ldots,n)} r(y_{i\cdot}|x_{i\cdot}, \theta). \tag{2.8}$$

Note that $L(n, \theta|(x^c_{\cdot\cdot}, y^c_{\cdot\cdot}))$ cannot be evaluated, since $x_{i\cdot}$ is not observed for all $i > n_c$. Nevertheless, let $\pi(y_{i\cdot}|x_{i\cdot}, \theta) := r(y_{i\cdot}|x_{i\cdot}, \theta)/(1 - r(\mathbf{0}|x_{i\cdot}, \theta))$, and define

$$L_1(\theta|(x^c_{\cdot\cdot}, y^c_{\cdot\cdot})) = \prod_{i \in (1,\ldots,n_c)} \pi(y_{i\cdot}|x_{i\cdot}, \theta)$$

and

$$L_2(n|n_c, \theta) = \left(\prod_{i \in (1,\ldots,n_c)} (1 - r(\mathbf{0}|x_{i\cdot}, \theta)) \right) \left(\prod_{i \in (n_c+1,\ldots,n)} r(\mathbf{0}|x_{i\cdot}, \theta) \right).$$

It is easy to verify that (2.8) decomposes as the product $L_1 L_2$. Having obtained θ_{L_1} as the MLE of L_1, there is no clear role of L_2 in estimating n, since $x_{i\cdot}$ is not observed for all $i > n_c$. (This situation differs from the previous $L_1 L_2$ decomposition in that $p(i, \mathbf{0})$ may now vary over i, via dependence on covariates x_i, instead of being fixed at $p(\mathbf{0}|\theta)$.)

Since L_2 is not directly estimable, we seek a method of estimation that relies only on the conditional likelihood L_1 and its corresponding maximum likelihood estimate $\hat{\theta}$. The most common solution is to plug $1 - r(\mathbf{0}|x_{i\cdot}, \theta)$ in for $\psi(x_{i\cdot})$ in the HT estimator (2.1). Huggins and Hwang (2011) characterized the conditional likelihood in terms of a generalized linear model, for computational convenience in likelihood maximization. A related approach is to estimate θ based on the *partial likelihood* instead of the conditional likelihood, as proposed by Stoklosa et al. (2011) and Stoklosa (2012).

Whereas most HT estimators assume that each $x_{i\cdot}$ is fixed, some non-HT models treat $x_{i\cdot}$ as a random variable. Following the Bayesian data augmentation scheme of Royle et al. (2007) and Royle (2009), let the θ

in (2.8) be renamed as θ_1, and assume that the rows of $x_{\cdot\cdot}$ come from a sequence of I.I.D. random variables from some distribution $g(x, \theta_2)$, for some parameter vector θ_2. Then (2.8) becomes

$$L(n, \theta|(x^c_{\cdot\cdot}, y^c_{\cdot\cdot})) = \prod_{i\in(1,\ldots,n)} r(y_{i\cdot}|x_{i\cdot}, \theta_1)g(x_{i\cdot}|\theta_2). \tag{2.9}$$

Pledger (2000) maximized (2.9) *directly*, treating $x_{\cdot\cdot}$ as a latent covariate indicating random assignment to one of several possible probability models in a mixture.

2.5 Application and Simulation

Table 2.1 summarizes a selection of previously analyzed data sets. Far from being exhaustive, the table provides only a sample from the wide range of settings in which CRC is relevant. In fact, if another author publishes a similar list of applications independently, the resulting pair of lists could be amenable to a CRC study.

Table 2.2 summarizes a large fraction of the simulation experiments that have been published to date. (Recall that $\phi_j(i)$ denotes the probability of the ith unit being captured on the jth list; the column with this heading gives approximate ranges of $\phi_j(i)$ over all i, j when this can be easily deduced from the source.) Assuming that we did not miss a large chunk of the literature, it is clear from the table that simulations have not been particularly broad in scope. Few simulations have incorporated covariates, only one simulation involved a population larger than 2000, and few simulations

Table 2.1: A selection of previously analyzed data sets

Reference	**Population**	**k**	**Covariates**	**Model idea**
Odum and Pontin (1961)	Ants in selected ant colonies near Oxford, England	2	none	First principles
Fienberg (1972)	Children with a congenital anomaly in Massachusetts, 1966	5	?	Log-linear
Pollock et al. (1984)	Lobsters at Port Maitland, Canada, 1950-1951	13	Temp., Effort	Logistic-linear regression
Bruno et al. (1992)	People with *diabetes mellitus* in Northern Italy, 1988	4	Sex, Age	Post-stratification
Leyland et al.	Female Street-Working Prostitution in Glasgow	53	HIV & drug stat.	Log-linear, chaining
Abeni et al. (1994)	HIV-1 infected people in the Lazio region of Italy	4	?	Log-linear
Runeson and Wohlin (1998)	Errors in a body of computer code	8	10 error types	Ad-hoc
Fienberg et al. (1999)	Websites related to a specific search query	6	none	Log-linear/Bayes
Yip et al. (2001)	Peromyscus maniculatus at East Stuart Gulch, Colorado, pre-1978	6	Sex, Age, Weight	Logistic-linear regression
Chen and Lloyd (2002)	First Nations members in a region of Canada, 1998-99	2	Age	Kernel regression
Aaron et al. (2003)	Lesbians in Allegheny County, Pennsylvania	4	?	Log-linear
Silver et al. (2004)	Jaguars in several areas of Bolivia and Belize	2?	Sex	CAPTURE software
et. al. (2006)	Deaths among HIV-Infected Adults in France in 2000	3	Age, sex, etc.	Log-linear
Baillargeon and Rivest (2007)	Snowshoe hares (*Lepus americanus*) on six consecutive days	6	none	various
Baillargeon and Rivest (2007)	Meadow voles	30	none	various
Malec and Maples (2008)	People in the United States, 2000	2	Age, Sex, etc.	Bayesian
Manrique-Vallier and Fienberg (2008)	Political assassinations/disappearances, Peru, 1980-2000	6	none	Grade of Membership
Malec and Maples (2008)	People in the United States, 2000	2	Sex, Age, Race	A Bayesian analysis
Kelly et al. (2009)	Prevalence of Opiate Use in Ireland	3	Age, sex, etc.	Log-linear
Murphy (2009)	People in the World Trade Center on September 11, 2001	3	none	Log-linear
Royle (2009)	Mallard ducks in the United States and Canada, 2005	2	Cluster Size	A Bayesian analysis
Royle (2009)	Meadow Voles in Laurel, Maryland, 1981	5	Body Mass	A Bayesian analysis
Chen et al. (2010)	People in the United States, 2000	2	Sex, Age, Race	Kernel regression
Stoklosa et al. (2011)	Mountain Pygmy Possum in Victoria, Australia, 2003	5	Body Weight	Logistic-linear regression
Stoklosa et al. (2011)	Harvest Mice in Shei-Pa National Park, Taiwan, 2008	14	Body Weight	Logistic-linear regression
Stoklosa et al. (2011)	*Prinia flaviventris* at the Mai Po Bird Sanctuary, Hong Kong,1993	17	Wing Length	Logistic-linear regression
El Adssi et al. (2012)	Multiple Sclerosis in the Lorraine region in France, 2006	3	Age, Sex	Log-linear

Table 2.2: A selection of simulation experiments

Reference	**k**	**n** ($\approx$)	$\phi_j(i)$	**Covariates**	**Type**
Chao (1987)	5-10	200-400	0.05-0.10	NA	M_{th}
Yip (1991)	5	150	0.2-0.4	NA	M_t
Chao et al. (1992)	5-10	100-400	0.2-0.5	NA	M_{th}
Lloyd (1992)	5-10	50-500	-	NA	-
Evans and Bonett (1994)	2-4	25-100	0.3-0.9	NA	M_{tb}
Norris and Pollock (1995)	10	50	0.6	NA	M_{bh}
Norris and Pollock (1996a)	10-20	50-100	-	NA	M_{bh}
Chao and Tsay (1998)	3	200	0.7	NA	M_{tbh}
Fienberg et al. (1999)	6	2000	0.7	NA	M_{th}
Pledger (2000)	10-20	50-100	0.1-0.5	NA	M_h
Chen and Lloyd (2000)	2	500-1000	0.2-0.6	norm. mix.	M_{th}
Yip et al. (2001)	5	100	0.5	normal	M_{bh}
Basu and Ebrahimi (2001)	4	1000	-	-	-
Dorazio and Royle (2003)	5-10	50-2000	0.6	NA	M_h
Pledger (2005)	6	100	0.1-0.9	NA	M_h
Manrique-Vallier and Fienberg (2008)	4	2000	0.035-0.7	mix. memb.	M_{th}
Chen et al. (2010)	2	281×10^6	0.95	many	M_{th}
Stoklosa et al. (2011)	7-10	100-400	-	unif., bern.	M_h
Hwang and Huggins (2011)	6	200	-	unif., bern.	M_{tbh}

have used exactly three lists.

Of particular note is the simulation by Chen et al. (2010) which attempted to reflect the population of the United States on April 1, 2000, based on population characteristics that were consistent with the 2000 census count. The simulation is large and complex, involving 281,421,906 persons, 51 subregions, and 5 covariates.

Chapter 3

Log-linear Models

This chapter introduces traditional log-linear models that do not include auxiliary covariates. Our method of extending traditional log-linear models to incorporate auxiliary covariates does not appear until Chapter 5.

3.1 Brief Review

Given any function from the set of capture patterns into the real numbers, $f : \mathcal{Y} \to \mathbb{R}$, with $\sum_y \exp f(y) = 1$, one can model the multinomial capture probabilities in terms of the capture pattern: $\log p(y) = f(y)$. If f is a linear function of a vector of parameters u, then $\exp(f(y|u))$ is a log-linear parameterization. A log-linear model exists to exactly fit any multinomial probability array $\mathbf{p}(\mathcal{Y})$. Given a vector of parameters $u = (u_0, u_1, u_2, u_{12})$, a simple log-linear model is

$$\log p(y|u) \quad = \quad u_0 + u_1 y_1 + u_2 y_2 + u_{12} y_1 y_2 \ (y \in \mathcal{Y}), \tag{3.1}$$

where y_j denotes the jth element of the vector y ($j = 1, ..., k$). The parameters u_1, u_2 describe list effects, and u_{12} represents the interaction between the first and second list. If there are more than two lists, additional parameters may describe the other list interactions. For example, with k lists, $u_{1\cdots k}$ denotes the highest-order list interaction.

One way to estimate parameters is to maximize the likelihood function corresponding to the assumption that $\mathbf{c}$ is a realization of a multinomial random variable with n trials from the probability array $\{p(y|u)\}_{y\in\mathcal{Y}}$. The multinomial conditional likelihood (2.4) with the dependence on the log-linear parameters made explicit (and switching notation by replacing θ with u) is

$$L_c(u|\mathbf{c} \setminus c_{\mathbf{0}}) = \frac{n_c!}{\prod_{y\neq\mathbf{0}} c_y!} \prod_{y\neq\mathbf{0}} \pi(y|u)^{c_y},$$

where

$$\pi(y|u) := p(y|u)/(1 - p(\mathbf{0}|u)). \tag{3.2}$$

Maximizing L_c gives parameter estimates $\hat{u}$ which we plug into the marginal likelihood (2.5) and maximize over n, following Sanathanan's Theorem (2.7). The result is approximately equal (within a rounding error) to

$$\hat{n} := n_c + n_c\pi(\mathbf{0}|\hat{u}). \tag{3.3}$$

As an alternative to using the multinomial distribution, one may assume that $\mathbf{c}$ is a realization of a collection of independent Poisson distribu-

tions. The parameter estimates following Sanathanan's Theorem (2.7) using a multinomial likelihood are the same as the estimates under the Poisson model, but asymptotic variance expressions differ substantially (Sandland and Cormack, 1984). Our discussion of log-linear models for the remainder of this chapter relates only the conditional multinomial approach unless stated otherwise, and accordingly we build models conditionally, in terms of $\pi(\cdot)$ instead of $p(\cdot)$ as in (3.2). Cormack and Jupp (1991) elaborate on the generally minor differences between conditional multinomial, unconditional multinomial, and Poisson modes of inference.

The cross-classification $\mathbf{c}$ has only $2^k - 1$ observable cells, and a unique maximizer of the likelihood can exist only if the model contains at most $2^k - 1$ parameters. Thus, an identifiable model with exactly $2^k - 1$ parameters is called saturated, providing a perfect fit for the observed relative frequencies in the cells of $\mathbf{c}$. With only two lists, one may take $u_{12} := 0$ in (3.1) to get a saturated hierarchical log-linear model, and maximizing the conditional and marginal likelihoods leads to the Petersen estimator for the missing cell, $\hat{c}_{00} := \hat{n} - n_c = c_{10}c_{01}/c_{11}$.

For three lists, the most general log-linear model presented in Fienberg (1972), in our notation, is

$$\begin{aligned} \log \pi(y|u) =& u_0 + u_1 y_1 + u_2 y_2 + u_3 y_3 + u_{12} y_1 y_2 \\ &+ u_{13} y_1 y_3 + u_{23} y_2 y_3 + u_{123} y_1 y_2 y_3 \ (y \in \mathcal{Y}). \end{aligned} \tag{3.4}$$

For any capture pattern $y \in \mathcal{Y}$ and $\omega \in \Omega$, define the product indicator $I_\omega(y) = \prod_{j \in \omega} y_j$, where $I_\emptyset(y) := 1$. We succinctly generalize the log-linear

models (3.1) and (3.4) for any number of lists as

$$\log \pi(y|u) = \sum_{\omega \in \Omega} u_\omega I_\omega(y). \tag{3.5}$$

Like model (3.1) for two lists, model (3.4) for three lists has one too many parameters; it is necessary to fix one of the u-terms as constant prior to estimating the remaining terms to have a unique maximum conditional likelihood solution. The standard choice (when $k = 3$) is to take $u_{123} := 0$ in model (3.4) and allow the other seven parameters to vary subject only to the probability constraint $\sum_y p(y|u) = 1$, or, equivalently, $\sum_{y \neq \mathbf{0}} \pi(y|u) = 1$. Then (3.4) is saturated in the sense that a vector $u = (u_0, u_1, u_2, u_3, u_{12}, u_{23}, u_{13})$ exists with $\pi(y|u)$ being exactly equal to $\pi(y) := p(y)/(1 - p(\mathbf{0}))$ $(y \neq \mathbf{0})$ for every possible choice of multinomial cell probabilities $\{p(y)\}_{y \in \mathcal{Y}}$.

The classification of log-linear models into the form M_{tbh} or one of its sub-forms is partially a matter of interpretation and context. When $k = 2$, the saturated hierarchical model has only enough terms for a separate effect for each list (M_t). When $k \geq 3$, we can include interaction terms. The interpretation of the interaction terms is ambiguous because interactions may result from list dependence induced by heterogeneity *or* from unit-level list dependence. It is impossible to distinguish between these two causes of interaction without additional assumptions or information.

3.2 Several Important Log-linear Models

Removing terms from (3.5) gives several submodels. For example, removing all interaction terms results in an independence model for k lists, so that the probability of capture on each list is independent of the event of capture on any other list:

$$\log \pi(y|u) = u_0 + u_1 y_1 + \cdots + u_k y_k. \tag{3.6}$$

Requiring that the list effects are all equal to a single parameter, $u_\Sigma := u_1 = \cdots = u_k$, further constrains the independence model (3.6) to get a model of independence with equal catch-ability across lists, an extremely sparse model with only one free parameter:

$$\log \pi(y|u) = u_0 + u_\Sigma \sum_{j=1}^{k} y_j. \tag{3.7}$$

A version of (3.7) appeared in Moran (1951), and perhaps most recently as model M_0 in Rivest and Lévesque (2001).

For any submodel of (3.5) with the highest-order interaction term set to zero (i.e., $u_{1\cdots k} := 0$), eliminating all of the explicit u-terms from the model equations leads to the identity

$$\pi(\mathbf{0}|u) = \frac{\prod_{y \in O} \pi(y|u)}{\prod_{z \in E} \pi(z|u)}, \tag{3.8}$$

where O is the set of capture patterns with entries summing to an odd number, and E is the set of nonzero capture patterns summing to an even number

(Fienberg, 1972). To use (3.8) to estimate the missing cell, we substitute the maximum likelihood estimate $\hat{u}$ in for u. Section 3.6 gives a simple proof of (3.8).

The saturated model (3.5) is particularly convenient because the model always fits the observable data exactly, removing the need to estimate nuisance parameters. Combined with (3.3), the formula (3.8) becomes

$$\hat{c}_{\mathbf{0}} = \frac{\prod_{y \in O} c_y}{\prod_{z \in E} c_z}. \tag{3.9}$$

(An early variant of this formula appears in Bartlett (1935)). Unfortunately, the resulting estimate of the population size can be extremely unstable due to overfitting, since any of the denominator terms approaching zero causes the estimator to blow up.

To obtain stability in a heuristic way that does not require selection of a specific parsimonious submodel, we propose mixing the saturated model (3.5) with the two-parameter independence model (3.7), as follows. Let $\Pi(\hat{u}) = [\pi\{y|\hat{u}\} : y \neq \mathbf{0}]$ denote the array of conditional multinomial probabilities implied by model (3.7) given parameter estimates $\hat{u} = (\hat{u}_0, \hat{u}_\Sigma)$. Let $\nu = min_{y \neq 0} \mathbf{c}_y$. Define a mixing constant $\alpha \in (0, 1)$ as

$$\alpha = \frac{\nu}{1 + \nu},$$

and define a weighted average

$$\Pi(\hat{u}, \alpha) := (1 - \alpha)\Pi(\hat{u}) + \alpha\widehat{\Pi}, \tag{3.10}$$

where the linear combination of arrays is evaluated element-wise, and $\widehat{\Pi} = \mathbf{c}/n_c$ the empirical frequencies of the capture patterns. Plugging $\Pi(\hat{u}, \alpha)$ into the right-hand side of (3.8) gives an estimate for $\pi(\mathbf{0}|u)$. Note that the constant α is small, putting greater weight on the sparse fit $\Pi(\hat{u})$, when sample size n_c is not large enough to stabilize the smallest element of the saturated fit $\widehat{\Pi}$. We call (3.10) the adjusted saturated model. We do not study this model in detail in this thesis, suggesting it only as a topic for future study. It may be fruitful to experiment with modified definitions of α, and to contrast results with the comparably ad-hoc "robust" estimation adjustment presented in Stoklosa and Huggins (2012).

Finally, we mention a log-linear treatment of heterogeneity that arises from the educational testing model of Rasch (1960). Sanathanan (1972b) applied the Rasch model in a CRC setting in which the lists are assumed to be independent at the unit level but dependent in the aggregate due to heterogeneity. The model is built around the equation

$$logit(p_{ij}) = t_i + \beta_j, \tag{3.11}$$

where p_{ij} is the probability that the ith unit is captured on the jth list, t_i is an individual effect, and β_j is a list effect. Treating the individual effects and list effects as fixed or as random in various ways leads to diverse log-linear models that are all grounded in (3.11). Fienberg and Meyer (1983) gives a good introduction to these approaches, and Fienberg et al. (1999) develops the Rasch model from a Bayesian perspective.

3.3 Variance

Extensive derivations of asymptotic variance formulas pertaining to log-linear estimates of population size are in Darroch (1958), Fienberg (1972), Sanathanan (1972a), and Sandland and Cormack (1984), to name a just few. This section does not contain any new formulas. Instead, we examine an existing formula for insights into the causes of high variance, and we test this formula with a simulation experiment.

To be clear, the variance that we wish to estimate is conditional on n but not conditional on n_c. Specifically,

$$Var(\hat{n}) := Var(\hat{n}|n) = E(Var(\hat{n}|n_c, n)) + Var(E(\hat{n}|n_c, n)),$$

where n is fixed and n_c is random. Fienberg (1972) used the delta method to derive several asymptotic variance formulas for specific hierarchical log-linear models. In our notation, his approximation for the variance in the saturated model is

$$\widehat{Var}(\hat{n}) = \widehat{Var}(n_c + \hat{c}_{\mathbf{0}}) \approx \hat{c}_{\mathbf{0}}^2 \left(\frac{1}{\hat{c}_{\mathbf{0}}} + \sum_{y \neq \mathbf{0}} \frac{1}{c_y} \right) \tag{3.12}$$

Formulas for the variances of the estimators corresponding to the various hierarchical submodels of the saturated log-linear model are similar except for having different sets of denominators. Fienberg's derivation of (3.12) relies on the assumption that the estimate of the population size is correct, i.e., $n = n_c + \hat{c}_{\mathbf{0}}$.

The variance approximation (3.12) is undefined when the observed count

c_y for any of the capture patterns is zero. In fact, due to the potential for zeros in the denominator of (3.9), the expectation and variance of the estimator $\hat{n}$ is technically not defined. Thus, in all of our variance computations, we implicitly condition on the event that there is at least one observation of every capture pattern. This condition is not of much importance when each of the multinomial probabilities $p(y)$ is not close to zero and n is large.

We use (3.12) to deduce several qualitative statements regarding $Var(\hat{n})$. First, it is clear that the size of the variance depends strongly upon the estimate of the missing cell $\hat{c}_{\mathbf{0}}$. A smaller missing cell is associated with a smaller variance. Second, holding $\hat{c}_{\mathbf{0}}$ fixed, the count c_y in any cell approaching 0 causes the variance to explode. Third, let z denote a nonzero capture pattern, and hold c_y fixed for every other nonzero capture pattern $y \neq z$. Then, the effect on $Var(\hat{n})$ (and on $\hat{c}_{\mathbf{0}}$) of letting c_z approach zero depends upon whether z is an "even" or "odd" capture pattern, through the formula (3.9). Specifically, $Var(\hat{n})$ and $\hat{c}_{\mathbf{0}}$ both explode if z is an even capture pattern, and they tend toward zero if z is an odd capture pattern.

The variance approximation (3.12) is only a first order Taylor approximation, and so we are interested in evaluating its performance via simulation. We study the variance of the saturated-model estimator $\hat{n}$ for the three-list case when certain multinomial probabilities approach zero under several

basic constraints:

$$\sum_y p(y) = 1 \tag{3.13}$$

$$p(000) = \frac{p(111)p(100)p(010)p(001)}{p(110)p(101)p(011)} \tag{3.14}$$

$$p(000) = 0.2 \tag{3.15}$$

The first constraint (3.13) is the basic probability constraint. We require (3.14) so that the saturated log-linear model is appropriate for our simulation as in (3.8). We chose the size of $p(000)$ in (3.15) arbitrarily; the decision to hold $p(000)$ constant across simulations is so that no variability in the overall detection rate $1-p(000)$ can confound the effect of varying the other multinomial probabilities.

For a "base case", let $p(111) = p(000)$, and take $p(y) = 0.1$ for all other capture patterns, and observe that these choices satisfy the constraints (3.13)-(3.15). If we set $p(110) = a$ with $0 < a < 0.1$, the constraints are no longer satisfied, unless we also shift probability mass across the other possible capture patterns in ways that satisfy the constraints. Doing this in a general way analytically turns out to be hard, as the constraints are both additive (3.13) and multiplicative (3.14). However, noting that $0.1 + (0.1 - a)/2 = 0.15 - a/2$, it easy to verify that the following simple modification scheme satisfies the constraints:

$$p(000) = p(111)\frac{a(0.15 - a/2)(0.15 - a/2)}{a(0.15 - a/2)(0.15 - a/2)}. \tag{3.16}$$

That is, we can set $p(110) = a$, a small number, if we also take $p(100) = a$

and set the remaining probabilities equal to $0.1 + (0.1 - a)/2$.

Our simulation looks at every combination of the following sets of specifications:

1. Let n equal 500 or 1000.

2. a ranges over 0.005 to 0.07 in intervals of 0.005.

For each combination of n and a, we simulate a three list CRC experiment according to the multinomial probabilities determined by the constraints and (3.16). For each simulation, we use the saturated log-linear model to estimate $c_{\mathbf{0}}$ and the corresponding variance approximation (3.12). We run 10000 replications of the experiment, generating estimates $\hat{n}^{(1)}...\hat{n}^{(10000)}$ and $\widehat{Var}(\hat{n})^{(1)}...\widehat{Var}(\hat{n})^{(10000)}$. We compute the sample variance of the population size estimates to obtain a simulation estimate of $Var(\hat{n})$; the dashed curve in each panel of Figure 3.1 indicates the square root of this estimate, for the various values of the small cell probability a. The smoothness of this curve hints that the simulation result is rather accurate at 10000 replications.

To evaluate the usefulness of (3.12), we compute the 25th and 75th percentiles of the 10000 simulation draws from the sampling distribution of $\widehat{Var}(\hat{n})$, and subsequently take square roots to convert to the standard deviation scale. The vertical bar in each panel of Figure 3.1 indicates the interquartile range of the estimates $\widehat{Var}(\hat{n})^{(1)}...\widehat{Var}(\hat{n})^{(10000)}$ for each small cell probability. Based on the proximities of the dashed curved to the centers of the vertical bars, the distribution of the square root of (3.12) appears to be centered near the truth when $p(100) = p(110)$ is moderately small,

but less well-centered for extremely small values of $p(100) = p(110)$. By comparing the second panel against the first panel, it appears that a large population size (i.e., $n = 1000$ versus $n = 500$) improves the centeredness of the distribution of (3.12) around our simulation estimate of the true variance.

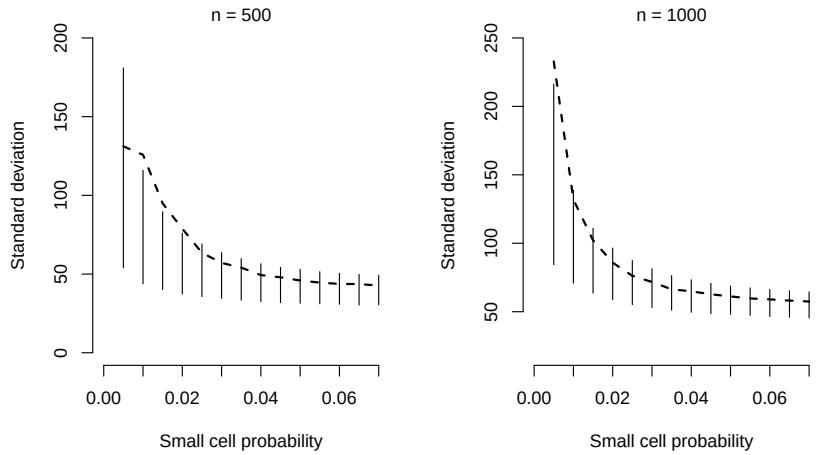

Figure 3.1: The vertical line segments indicate the interquartile range of the square root of the variance approximation (3.12) over 10000 simulations at each selected small cell probability. The dashed curve reflects the empirical standard deviation of $\hat{n}$ as estimated from the saturated log-linear formula (3.9).

One oddity in the simulation results is in the left-most endpoint of the first panel of Figure 3.1, where there appears to be an "elbow" in the empirical variance. We attribute this to the occurrence of zero-cell outcomes. Specifically, we discarded every experiment that had $c_y = 0$ for any nonzero capture pattern y, since (3.12) is undefined. These discards were most frequent for the smaller population size ($n = 500$) and when $p100 = p110$ were particularly small. Here, the discard rate was on the order of 15%.

3.4 Automated Model Selection

An appropriately parsimonious model provides degrees of freedom for testing model fit and reduces the variance of the corresponding population size estimate $\hat{n}$. Selecting the best model from the set of submodels of (3.5) is challenging when the number of lists k is large, as there is a very large number of candidate models. Even worse, the "best" log-linear model is not necessarily identifiable in the most general setting, as we discuss in Chapter 4. We proceed in this section by assuming that the highest-order interaction $u_{1...k}$ is zero.

We adopt two common conventions for simplifying the model selection problem. The first is to restrict attention to hierarchical models. The second is to put a lower bound on model complexity by considering only models which include all main effects $u_1, ..., u_k$.

This section focuses on automated model selection methods because of their necessity in our subsequent work in Chapter 5. In contrast with automated methods, Fienberg (1972) outlined a model selection strategy based on likelihood-ratio tests, restricting attention to hierarchical models. Fienberg's approach was nuanced and contained subjective elements that were not necessarily amenable to automated model search. Nevertheless, Burnham et al. (1995) automated a likelihood-ratio test model selection strategy and compared it, using different critical levels, against several information-theoretic strategies, concluding that information criteria are generally preferable to likelihood ratio tests for model selection.

3.4.1 Akaike Information Criterion

Model scoring criteria, such as the Akaike Information Criterion (AIC), provide an especially simple way automate model selection among [potentially non-nested] log-linear models. Several recent CRC studies, such as Aaron et al. (2003), Murphy (2009), and El Adssi et al. (2012), used the AIC. Despite its common use, some papers have argued that the AIC tends to overfit. Burnham and Anderson (2004) reviews the AICc, a "corrected" version of the AIC that is intended to avoid overfitting; they seem to state that the AICc is always at least as good as the original AIC, and therefore should be used by default.

It is easy to err in the implementation of the AIC or AICc, due to some confusing notation in the literature. We will review the formulas for both criteria here while attempting to clarify the confusing points. Let K denote "the number of estimable parameters in the approximating model" (Burnham and Anderson, 2002). Then we have

$$AIC = -2 \log L + 2K, \tag{3.17}$$

where L is the likelihood function, and

$$AICc = -2 \log L + 2K + \frac{2K(K+1)}{N-K-1}, \tag{3.18}$$

where N is the number of data points (i.e., $N = n_c$ for our CRC context). Akaike formally derived the AIC as an estimate of the Kullback-Leibler distance between the approximating model and a hypothetical true generating

model. Intuitively, a model is good when the negative log-likelihood is small, and so we prefer models with a small AIC score. The $+2K$ term in the AIC acts as a penalty on overfitting, and the additional term in the AICc is a extra penalty that disappears as N becomes large relative to K.

The first potential source of confusion is that, although the independence model (3.6) formally has $k+1$ parameters, only k of these parameters are "estimable" because the first parameter is entirely determined by the constraint that the multinomial sampling probabilities must sum to 1. Thus, we have $K = k$ for model (3.6).

The second potential source of confusion stems from a change in notation that occurred around 1997. Early discussions of the AICc followed the original development by Hurvich and Tsai (1989), which was in the context of a linear regression. Here, m was introduced as the "dimensionality of the approximating model," but (to the surprise of this author) $m \neq K$. The reason is that the variance of the residuals was included by default as an estimated parameter, but not counted in m, so that $K = m + 1$. Thus, for the reader that is not alerted to the difference between K and m, the final penalty term in the AICc can [incorrectly] seem to be

$$\frac{2(K+1)(K+2)}{N-K-2}$$

in many sources prior to 1997.

The AIC/AICc is only one of a large number of model scoring criteria, including the criteria known as the QAICc, TIC, BIC, MDL, and HQ (Anderson et al., 1994; Anderson and Burnham, 1999). This thesis primarily

applies the AICc, as it seems to improve upon the AIC, which is already prevalent in the literature. More importantly, the AICc performed well in the studies of Burnham and Anderson and in our simulations in Section 3.4.4.

3.4.2 Bayesian Information Criterion

The Schwarz information criterion (Schwarz, 1978), or Bayesian information criterion (BIC), is similar in form to the AIC:

$$BIC = -2 \log L + K \log n_c \tag{3.19}$$

Section 4 of Burnham and Anderson (2004) argues that the AIC and BIC are fundmentally similar, with the key difference being that the BIC uses a flat prior over the candidate models, while the AIC uses an informative prior. Hook and Regal (1997) used "internal validity analysis" to compare the usefulness of the AIC (evidently not the AICc) versus the BIC in a CRC setting, finding the AIC to be slightly preferable.

Draper (1995) suggested, on the basis of a technical point in the derivation of the BIC, that another reasonable criterion is

$$BIC_\pi = -2 \log L + K \log \frac{n_c}{2\pi}. \tag{3.20}$$

We hypothesize that no simple criterion will be optimal for all situations; the performance of each criterion depends on the simulation design. Section 3.4.4 contrasts the AICc, BIC, and BICπ with the goal of characterizing the

scenarios in which each criterion performs best.

3.4.3 Model Averaging

Model selection as a discrete process, either including or excluding each parameter, is somewhat unsatisfying. It intuitively seems preferable to include model selection uncertainty in the population estimate by somehow weighting the estimates of all reasonable models according to their information criterion scores, instead of picking a single best model. A simple model averaging method using the AICc, BIC, or BICπ appears in Hook and Regal (1997), Burnham and Anderson (2002), and Wagenmakers and Farrell (2004). We briefly state the basic method.

Suppose m models are under consideration. Let $s(IC)$ denote the vector of length m that contains the model scores based on some information criterion. Let $\nabla s(IC) = s(IC) - \min s(IC)$ denote the vector of differences between each entry of $s(IC)$ and its least entry. In ordinary use of the information criterion, the entry of $\nabla s(IC)$ that is equal to 0 corresponds to the preferred model. To generalize beyond this to an average across models, define a vector of model weights with ith entry

$$W_i = \frac{exp(-\frac{1}{2}\nabla s(IC)_i)}{\sum_{i'=1}^{m} exp(-\frac{1}{2}\nabla s(IC)_{i'})} \ (i = 1, ..., m)$$

The weights clearly sum to 1. If $\hat{c}_{\mathbf{0}}(i)$ is the estimate of the missing population size for the ith model, then the model average estimate of the population

size is

$$\hat{c}_{\mathbf{0}}(IC) = \sum_{i=1}^{m} W_i \hat{c}_{\mathbf{0}}(i). \tag{3.21}$$

3.4.4 Information Criterion Performance

The performance of model selection methods is difficult to measure because any specific method tends to perform well in certain situations but not others, and there are infinitely many situations to consider. We use simulation to assess the performance of the AICc, BIC, and the BICπ across a carefully chosen set of generating models with $k = 3$ lists.

We define several generating models in terms of a vector of six log-linear coefficients $u = (u_1, u_2, u_3, u_{12}, u_{13}, u_{23})$, as in (3.4), omitting the highest-order interaction term u_{123}. Given u, we typically leave unstated the intercept term u_0 in (3.4). For example, the vector $u = (1, -0.5, -1, 2, 0, -2)$ represents the model

$$\log p(y|u) = u_0 + y_1 - 0.5y_2 - y_3 + 2y_1y_2 - 2y_2y_3,$$

where the intercept is uniquely determined by the constraint that the multinomial probabilities sum to 1 (here, we found $u_0 \approx -2.92$).

Figure 3.2 explores the performance of information criterion in terms of the sampling distribution of the ratio $R := \hat{c}_{\mathbf{0}}/c_{\mathbf{0}}$, the estimated number of missing units to the actual number of missing units. We use each of several generating models, indicated by u at the top of each panel, to conduct 1000 CRC experiments. For the data generated in each experiment, we select and

fit the best model from the eight hierarchical submodels of (3.4) that contain all main effects (i.e., u_1, u_2, u_3) and no highest-order interaction term. We do the model selection and estimation in six ways for each experiment, corresponding to the AICc, BICπ, BIC, weighted AICc, weighted BICπ, and weighted BIC respectively, where each weighted version is a model average (3.21). Figure 3.2 shows the mean and interquartile range of R for each model selection method and each of several population sizes.

We chose the set of generating models to "span" a large subspace of the set of feasible generating models. The first column of panels in Figure 3.2 looks at several different sets of interaction terms: no interactions, one interaction, two interactions, all interactions, and a mix of positive and negative interactions. In each of these cases, the main effects satisfy $u_1 = u_2 = u_3 = a_1$, and we used numerical techniques to pick the value of a_1 such that $p(\mathbf{0}|u) = 0.2$, or equivalently, $u_0 = \log(0.2)$. The second column of panels is the same except that the main effects are no longer constant; $(u_1, u_2, u_3) = (a_2, 2a_2, 3a_2)$, where again we chose a_2 to satisfy $p(\mathbf{0}|u) = 0.2$.

Figure 3.2 contains important lessons. The first is that the quality of estimates typically improves with increasing population size. For small populations, the estimates are extremely right skewed, with the median value of R often well below 1 even as the bias is substantially positive in some cases. The second lesson is that the performance of each model selection criterion depends strongly upon the generating model. The median of the ratio R is closest to 1 for the AICc (first vertical line) with small populations in the panels that are second and third from the top. The BIC is the clear winner in terms of having an average value of R that is closest to 1 in the panels

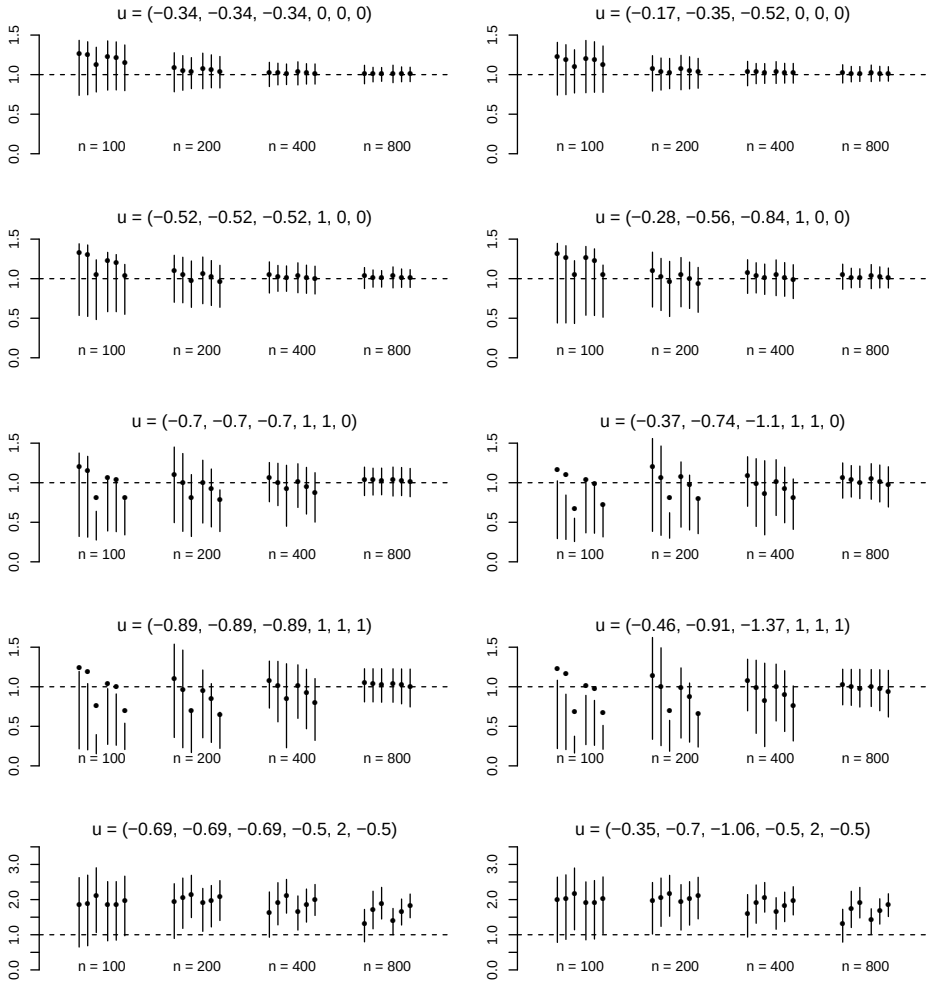

Figure 3.2: Each panel illustrates aspects of model selection performance for a specific generating model. The analyses is done at each of several population sizes; $n = 100, ..., 800$. Each vertical axis indicates the ratio $\hat{c}_{\mathbf{0}}/c_{\mathbf{0}}$, so that a "perfect" estimate would fall on the horizontal dashed line at 1 in each panel. For a given generating model u and population size n, the analysis is summarized with a cluster of six vertical lines with corresponding dots. Each vertical line represents the interquartile range for a population estimate across 1000 CRC experiments. Similarly, the black dot corresponding to each vertical line marks the mean ratio $\hat{c}_{\mathbf{0}}/c_{\mathbf{0}}$ over 1000 experiments. The six vertical lines for each n within each panel correspond to the AICc, BICπ, BIC, weighted AICc, weighted BICπ, and weighted BIC respectively.

that are first and second from the top. The BICπ typically produces estimates between the AICc and BIC. The bottom panels show seriously biased estimates for every model selection method, with $R \approx 2$ even for large n; note here that the vertical axis is on a larger scale to allow space to display the estimates.

The third lesson of Figure 3.2 is potentially surprising. Model averaging typically produces only a slight improvement over the estimate from the best model. We see this by comparing the 4th - 6th vertical lines in each group against the 1st - 3rd. This finding is somewhat consistent with the findings of Hook and Regal (1997).

The results in the bottom panels of Figure 3.2 are alarmingly poor. Upon closer study, it turns out that the multinomial probabilities that correspond to this model include two especially small values corresponding to "even" capture patterns. We speculate that variation in the estimates of these small values tends to inflate the population size estimates, since these are denominator terms in (3.8).

To explore the importance of small entries of $\mathbf{c}$, Figure 3.3 illustrates the performance of the AICc for three different generating models, with population sizes ranging from $\exp(5) \approx 150$ to $\exp(10) \approx 22000$. From the first to third panel of Figure 3.3, the generating models have progressively more extreme values of log-linear coefficients, such that the set of multinomial probabilities includes entries of diminishing size. The third panel corresponds to the most extreme generating model. Here, the smallest multinomial probabilities are so small that there were no occurrences of at least one of the capture patterns for most of the simulation experiments with $n < \exp(8)$.

We discarded all simulations that resulted in a zero in $\mathbf{c}$, because the maximum likelihood estimates for the saturated log-linear model do not exist in these cases.

A fascinating observation from Figure 3.3 is that the selection of an incorrect model (i.e., a model that does not match the generating model as far as the choice of nonzero u-terms) is not related in an obvious way with the quality of the resulting estimates, as the red points are mixed evenly with the black points in the figure. This observation must be interpreted carefully, however, as we do not see the counter-factual. We expect that the estimates based on the generating model will, on average, be better than the estimates corresponding to red dots.

As a case study, we examine a particular data set that was generated according to the model in the third panel of Figure 3.3. With a population size $n = 1709$, there were observed 703 units, with capture patterns summarized in Table 3.1. This case is troubling because every model selection method that we test (and, in addition, the AIC, uncorrected) points to the model with only the interaction term u_{23} (in addition to the main effects), corresponding to the estimate $\hat{c}_{\mathbf{0}} = 10088$, approximately 10 times bigger than the truth. Upon careful examination, we find that the u_{23} model is indeed nearly a perfect fit to the data, with goodness of fit test statistic

$$\sum_{y \neq \mathbf{0}} \frac{(c_y - \hat{c}_y)^2}{\hat{c}_y} \approx 0.1.$$

The true model, $u_{12} + u_{13}$, has a slightly better fit, but one cannot "blame" the information criterion for opting for the more parsimonious model, given

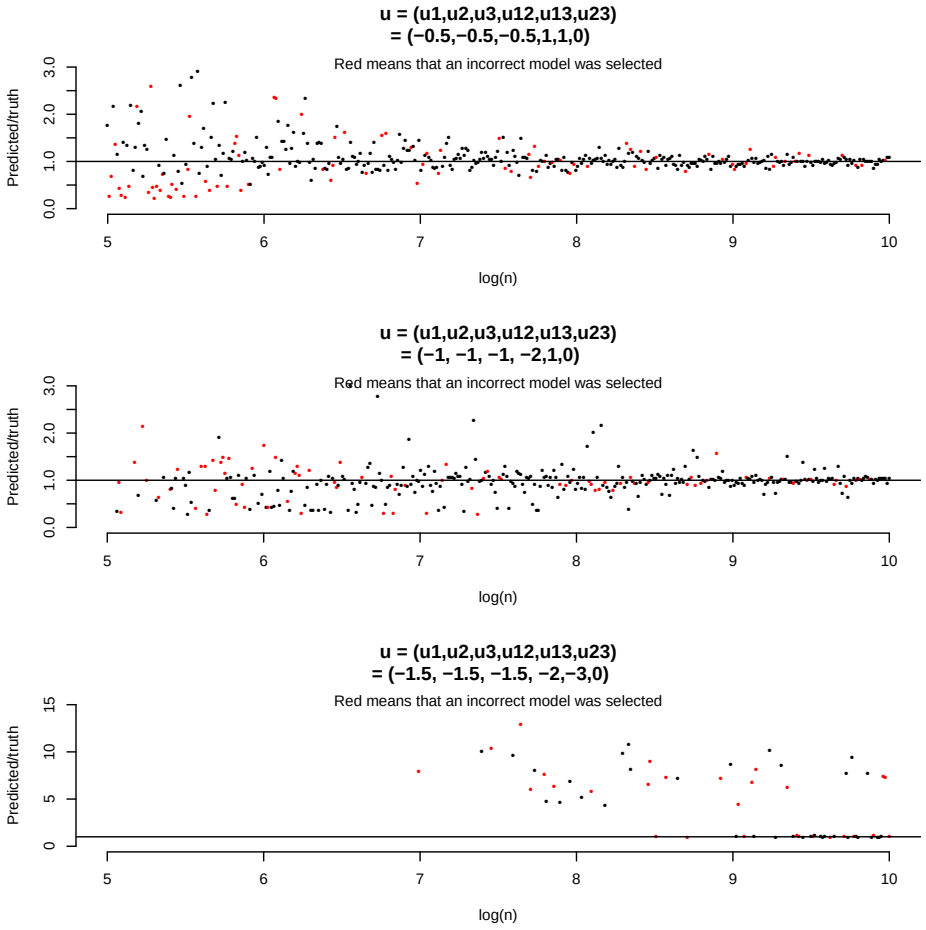

Figure 3.3: The title of each panel indicates the coefficients of the log-linear model that we used to generate the simulated data. Each point in each plot corresponds to one entire simulated capture-recapture experiment with subsequent model selection. The vertical axis shows the ratio $\hat{c}_{\mathbf{0}}/c_{\mathbf{0}}$. With increasing population size (horizontal axis), the accuracy of both model selection and estimation of the missing cell increases. However, notice that the volatility depends not only on the population size, but also on the size of the smallest frequencies in the set $\{\pi(y|u)\}_{y\neq\mathbf{0}}$ corresponding to the generating model. These smallest values are especially small for the center plot, and still smaller in the bottom plot, resulting in the extreme volatility. This hints that the AICc is vulnerable to overfitting or misfitting in special circumstances.

its excellent fit.

Any shock value of the preceding example must be tempered with the observation that it is an extreme case. Under the assumed generating model, an outcome with a nonzero c_{111} cell occurs only about 7 percent of the time (see parenthetical numbers in Table 3.1). For the other 93% of the time, $c_{111} = 0$, removing the option of using maximum likelihood to estimate the entire standard set of log-linear models.

Table 3.1: Simulated data from a single replication of the experiment in the third panel of Figure 3.3. The values in parentheses reflect the expected values for each cell according to the generating model, given that the true population size was 1709.

		In List 3	Not in List 3
In List 2	In List 1	1 (0.074)	5 (6.7)
	Not in List 1	55 (49)	215 (221)
Not in List 2	In List 1	4 (2.5)	208 (221)
	Not in List 1	215 (221)	1006 (989)

3.4.5 The AICc Gets Rasch

A major potential weakness of log-linear models is that they assume a homogeneous population. In practice, it is often reasonable to suppose that the probabilities of capture vary widely across population units. The Rasch model (3.11) offers a mechanism for including certain types of heterogeneity within a log-linear framework. This section deals with a very specific development of the basic Rasch framework. Specifically, Darroch et al. (1993)

derived the model

$$\log \pi(y|u) = u + u_1 y_1 + u_2 y_2 + u_3 y_3 + u_\Sigma \left(\sum_{j=1}^{k} y_j \right)^2 . \tag{3.22}$$

If u_Σ is constrained to be greater than zero in accordance with basic log-moment inequalities, then (3.22) can be interpreted as the intersection of the Rasch model with the assumption of no highest-order interaction. In Darroch et al. (1993), the no highest-order interaction assumption is valid if the distribution of unit-level effects $\{t_i\}$ in (3.11), conditional on the corresponding units not being captured, is Gaussian. Although other Rasch models exist, we use "Rasch" to refer only the specific model described above for the remainder of this section.

We wish to explore (a) the utility of the Rasch model, above and beyond basic log-linear models, and (b) the ability of the AICc to detect when the Rasch model is appropriate. To do this, we simulate data in a way that is approximately consistent with the Rasch model, and subsequently apply the AICc to select the "best" model from the standard set of eight log-linear models plus the Rasch model.

Simulating data that is exactly consistent with the Rasch model depends on knowing a probability distribution F that satisfies the following relationship: If the unit-level effects $\{t_i\}$ are draws from the distribution F, then, conditional on avoiding capture on all three lists, their distribution is Gaussian. Unfortunately we do not know of any such F (although we conjecture that one would be derived easily by a relatively competent theoretician). Fortunately, it turn out that the Gaussian distribution is a good approxi-

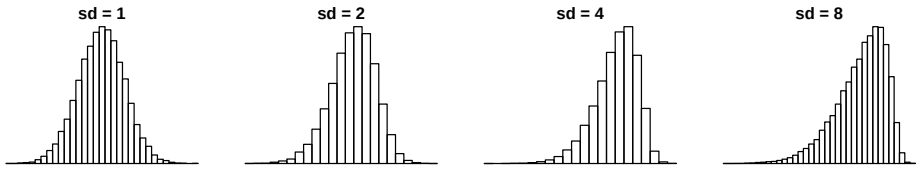

Figure 3.4: We simulate four CRC experiments, each having three lists drawn independently from a population of size $n = 100000$. In the first experiment, the probability of capture for each unit (a constant across the three lists) is the inverse logit of a draw from the $N(0, sd = 1)$ distribution. The left histogram shows the distribution of the logits of the capture probabilities for those units that remain uncaptured. The subsequent histograms are the same except that the standard deviation (sd) of the generating normal distribution is increased, leading to greater skewness in the distribution of individual effects among units that are never captured.

mation to a solution F as long as the standard deviation of the Gaussian is not large, and Figure 3.4 illustrates this fact. When the distribution of $\{t_i\}$ is the standard normal, the distribution of the set of capture probabilities associated with units that were never captured is visually indistinguishable from a normal distribution. The non-normality is pronounced only when the standard deviation of the Gaussian distribution is large. Even then, the skewed bell shape in the right-side panels of Figure 3.4 is arguably as close to the normal distribution as one should reasonably expect in practical applications.

We simulate a three-list CRC experiment according to (3.11), with the $\{t_i\}$ drawn as standard normal random variables and list effects $\{\beta_j\}$ that are identically equal to 0. We replicate the simulation 2000 times for populations of size $100, 200, 400, 800$, and 1600. At each replication, we perform model selection using the AICc on two different sets of feasible models. The first

set is the eight standard log-linear models. The second set is the union of the first set and the Rasch model. For each resulting estimate, we compute the percentage error as $100(\hat{c}_{\mathbf{0}} - c_{\mathbf{0}})/c_{\mathbf{0}}$.

(An important constraint in the Rasch model for three lists (3.22) is that the parameter u_Σ must be greater than zero. For these simulations, we did not explicitly enforce this constraint. However, with $n \geq 400$, we found that the maximum likelihood estimate $\hat{u}_\Sigma > 0$ with extremely high probability.)

Figure 3.5 shows the results. Interestingly, the Rasch model is not substantially less-biased than the saturated log-linear model for large populations. In principle, a key advantage may be that the Rasch model has two fewer parameters than the saturated model, which presumably reduces the variance of the estimates. However, we computed the root mean square error for the case $n = 1600$, obtaining 66.0 when the Rasch model is allowed and 66.5 when the Rasch model is not allowed, a rather trivial improvement.

The second panel of Figure 3.5 shows that the AICc is "smart enough to know" when the data are consistent with the Rasch model most of the time for large populations. We find intriguing that the AICc consistently chooses the Rasch model even as the Rasch model affords minimal improvement in estimates.

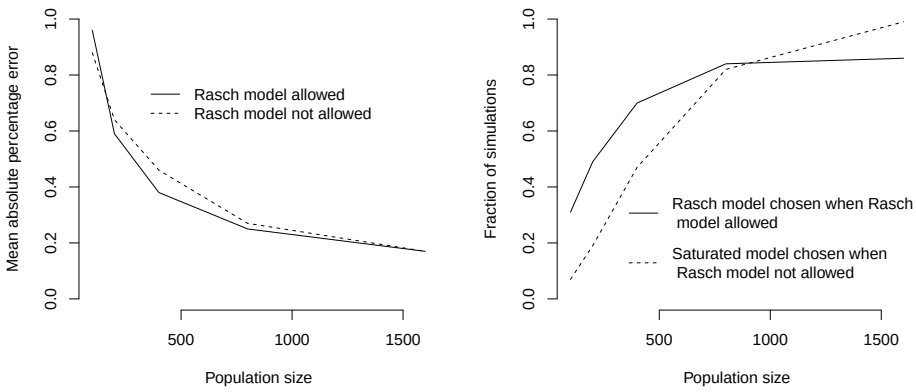

Figure 3.5: For the simulations defined in Section 3.4.5, theory suggests that the Rasch-inspired model (3.22) is a superior candidate model. In fact, the left panel suggests that the Rasch model is not substantially better than the next-best standard log-linear model, at least for larger populations. Despite the lack of substantial superiority in the Rasch estimates, the second panel shows that the AICc is somehow "smart enough" to know that the Rasch model is in fact the true generating model as much as 90% of the time for large populations.

3.5 Small Sample Adjustments

The analyses illustrated in Figures 3.2 and 3.3 show that population size estimates are often biased, particularly when there are capture pattern probabilities that approach zero and when the population size is small. A common way to address this problem is to adjust the data $\mathbf{c}$ before fitting a model.

The estimator of Chapman (1951) for $k = 2$ lists is a variation of the Petersen estimate, $\hat{c}_{\mathbf{0}} = c_{01}c_{10}/c_{11}$. The Chapman estimator is $c_{01}c_{10}/(c_{11} + 1)$, which we view simply as applying the Petersen estimator to adjusted data, where the adjustment consists of adding 1 the c_{11} cell. The Chapman estimator often outperforms the Petersen estimator both in terms of bias

and variance for small populations (Evans and Bonett, 1994).

For log-linear models with at least three lists, several data adjustments have been proposed. Evans and Bonett (1994) suggested adding $(0.5)^{k-1}$ to each cell in $\mathbf{c}$; we call this the EB adjustment. Hook and Regal (1997) suggested adding 1 to each element of $\mathbf{c}$ that appears in the denominator of (3.9); we call this the HR adjustment. Rivest and Lévesque (2001) derived first-order bias corrections for several specific log-linear models, but it is not clear how to apply these in automated model selection of general log-linear models.

The EB and HR adjustments seem to be ad hoc. We propose our own ad hoc data adjustment for comparison with the existing adjustments. Let $\hat{\mathbf{c}}$ denote the fitted values corresponding to the maximum likelihood estimates for the single-parameter independence model (3.7), and let

$$\alpha = \frac{1}{1 + n_c/(2^k - 1)}.$$

Note that α is large only if the average number of observations per cell, $n_c/(2^k - 1)$, is small. Define the adjusted data as

$$\mathbf{c}' := (1 - \alpha)\mathbf{c} + \alpha\hat{\mathbf{c}}. \tag{3.23}$$

We call this the ZK adjustment, using the initials of the current author.

We compare the EB, HR, and ZK adjustments by using the same CRC experiments that are described in Section 3.4.4 and the corresponding figure 3.2. Figure 3.6 shows that the performance of each adjustment varies

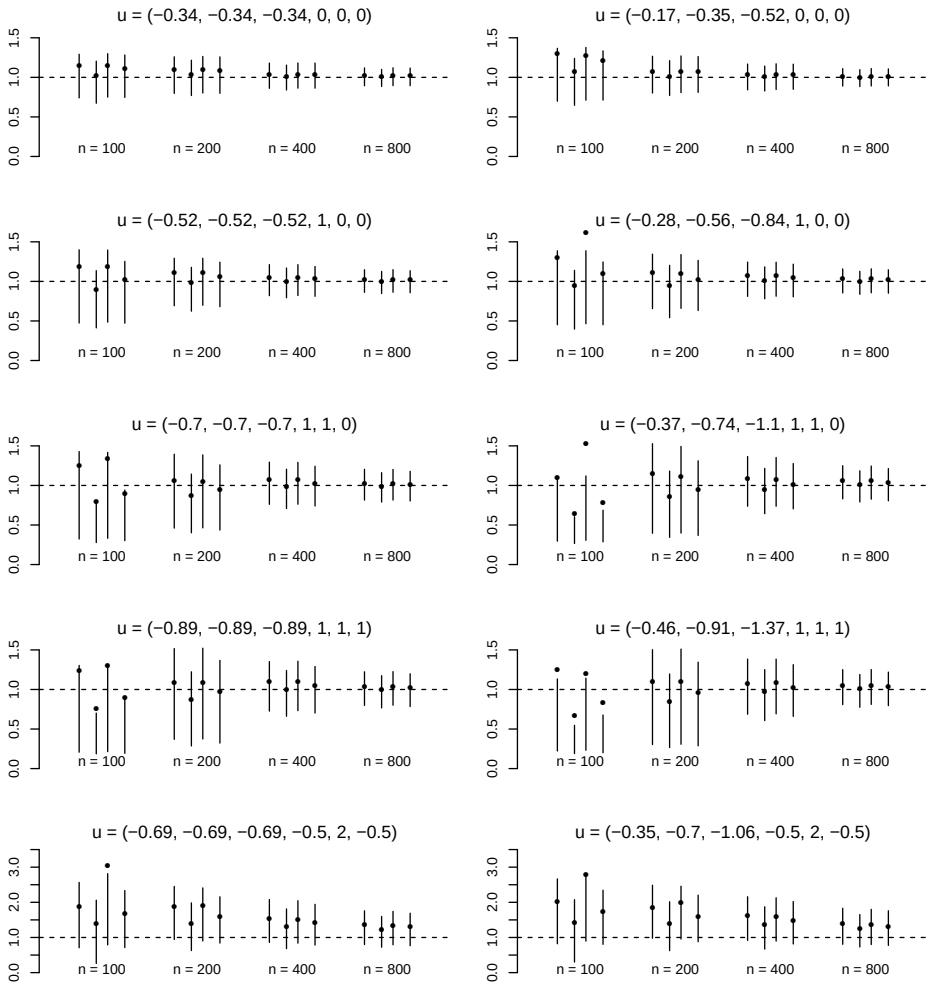

Figure 3.6: The basic interpretation of the lines and dots is the same as in Figure 3.2. The four vertical lines for each n within each panel correspond to no adjustment, HR, EB, and ZK, respectively.

with the generating model and the population size. In general, the effect of using an adjustment is not dramatic. However, using a data adjustment is convenient because it allows us to apply automated log-linear model search to arbitrary data sets without first establishing special rules to handle data that contains zeros.

3.6 Proof of the "Odd-Even" Formula

We state and prove a generalization of the formulas (3.8) and (3.9), which are equivalent to a formula in Fienberg (1972). Let $odd(\mathcal{Y}) = \{y \in \mathcal{Y} : \sum_i y_i \text{ is odd}\}$ and define $even(\mathcal{Y})$ analogously. Following the notation in Section 1.2.2, let $\Omega' = \Omega \setminus \mathcal{K}$.

Theorem 1. *For any collection of real-valued u-terms, $u = \{u_\omega\}$, define an arbitrary function f of the capture pattern $y \in \mathcal{Y}$ as in 3.5 with no highest-order term, i.e.,*

$$\log f(y) = \sum_{\omega \in \Omega'} u_\omega I_\omega(y). \tag{3.24}$$

Then

$$f(\mathbf{0}) \;=\; \frac{\prod_{\mathbf{y} \in odd(\mathcal{Y})} f(y)}{\prod_{z \in even(\mathcal{Y}) \setminus \{\mathbf{0}\}} f(z)}.$$

Proof. Let $n(u_\omega, odd)$ denote the number of elements $y \in \mathcal{Y}$ such that $I_\omega(y) = 1$ and $y \in odd(\mathcal{Y})$. Define $n(u_\omega, even)$ analogously. For each $\omega \in \Omega'$, it is easy to verify that $n(u_\omega, odd) = n(u_\omega, even)$. Note that this equality occurs only for $\omega \in \Omega'$; if $\omega = \mathcal{K}$, then $n(u_\omega, odd) = 0 \neq 1 = n(u_\omega, even)$ if

k is even, and $n(u_\omega, odd) = 1 \neq 0 = n(u_\omega, even)$ if k is odd. Thus,

$$\begin{aligned}
\sum_{y \in odd(\mathcal{Y})} \log f(y) &= \sum_{y \in odd(\mathcal{Y})} \sum_{\omega \in \Omega'} u_\omega I_\omega(y) \\
&= \sum_{\omega \in \Omega'} u_\omega \sum_{y \in odd(\mathcal{Y})} I_\omega(y) \\
&= \sum_{\omega \in \Omega'} n(u_\omega, odd) u_\omega \\
&= \sum_{\omega \in \Omega'} n(u_\omega, even) u_\omega \\
&= \sum_{y \in even(\mathcal{Y})} \log f(y).
\end{aligned}$$

The first and last terms of this equation can easily be re-arranged to obtain the desired result. □

Note that setting any of the $u-$terms in (3.24) equal to zero has no effect in the argument. Therefore, the theorem applies equally well for both hierarchical and non-hierarchical models.

Chapter 4

Identifiability

4.1 Overview

Link (2003) showed that the population size is nonidentifiable across certain types of CRC models. To address the general problem, we argue that (a) one must use information that is external to the data to narrow the set of feasible models, and that (b) interpretable models are essential for making use of such auxiliary information. A secondary analysis of data on survivors of the World Trade Center attacks illustrates the central issues.

The sampling mechanism in CRC is an extreme case of convenience sampling; each list represents an exhaustive enumeration of all population units that are detectable by some observer. Thus, detectability defines the sample, potentially causing the prediction set (all of the unobserved units) to have a substantially different character than the training set (all of the observed units). The implication is that inference depends profoundly on the assumptions that relate the observable data to the unobservable data.

This exaggerated model dependency makes CRC a prime case study for an infamous debate on statistical modeling cultures. Breiman (2001) argued that algorithmic methods with no obvious interpretation are often superior to models which explicitly incorporate various scientific features in an interpretable way. In a response, Cox (2001) included the following comment:

> Often the prediction is under quite different conditions from the data ... That is, it may be desired to predict the consequences of something only indirectly addressed by the data available for analysis. As we move toward such more ambitious tasks, prediction, always hazardous, without some understanding of underlying process and linking with other sources of information, becomes more and more tentative.

Indeed, the observed cells in Table 1.1 address only indirectly the unobserved cell, making estimation especially hazardous. Although we support Breiman's thesis in many other settings, CRC is a special case that requires careful attention to the underlying processes and alternative information sources.

We discuss general principles of model selection before returning to CRC. Suppose $\mathcal{M}$ is the set of all available models, and consider the following general model selection strategies.

$\mathcal{S}_1$: Choose the best model in $\mathcal{M}$ according to a data-driven criterion that includes both a measure of goodness of fit and a penalty on model complexity to avoid overfitting, such as the AIC.

$\mathcal{S}_2$: Define a subset $\mathcal{M}' \subset \mathcal{M}$ consisting of models that are especially plausible based on prior scientific beliefs about the mechanisms that generated the data, independent of the actual data values. Use a scoring criterion as in $\mathcal{S}_1$ to select a model from $\mathcal{M}'$.

$\mathcal{S}_3$: Use a data-driven criterion to score the models in $\mathcal{M}$ as in $\mathcal{S}_1$, but augment these scores to favor plausible or intuitive models, i.e., the kinds of models that are in $\mathcal{M}'$. Normalize the augmented scores, and use them as weights in an average of the models in $\mathcal{M}$. This average is not necessarily a standard Bayesian model average.

This chapter advocates the use of $\mathcal{S}_2$ or $\mathcal{S}_3$ as a necessary alternative to $\mathcal{S}_1$. Both of these strategies incorporate information external to the data, including researcher expertise and basic facts about how the data was collected. We refer to these external sources of information collectively as the *data context*. A modern emphasis on automated model selection has perhaps led many researchers to undervalue the data context. In advocating attention to data context, we provide an insightful discussion rather than a fundamentally new result. Indeed, attention to data context is as old as model selection itself.

These issues become more concrete in thinking about specific things that one might desire in a model selection strategy. One prerequisite for successful model selection is that $\mathcal{M}$ must contain at least one good model. For a high-level discussion of model goodness, we defer to Kass (2011). A second prerequisite is the ability to rank the goodness of the models in $\mathcal{M}$. We will review mathematical results that show that the data values

alone, without consideration of the data context, are not always sufficient to discriminate between competing models that imply vastly different inference. When this inability to discriminate persists asymptotically, we say that the model selection problem is nonidentifiable, or that $\mathcal{M}$ is nonidentifiable.

The strategy $\mathcal{S}_2$ may succeed as an alternative to $\mathcal{S}_1$ if $\mathcal{M}'$ is identifiable. Further, we propose $\mathcal{S}_3$ as a refinement of $\mathcal{S}_2$ that incorporates model selection uncertainty into the final estimates. Both $\mathcal{S}_2$ and $\mathcal{S}_3$ suffer from their dependence on the subjectivity of researcher expertise (an element of the data context). On the other hand, dependence on researcher expertise is fundamentally unavoidable when $\mathcal{M}$ is nonidentifiable. Explicitly representing a researcher's biases in the modeling process is the most honest approach.

Nonidentifiability of model selection is closely related to the ordinary nonidentifiability of parameters that we occasionally observe in likelihood function maximization. To view nonidentifiability of model selection in a parametric way, somewhat analogous to nonidentifiability in likelihood maximization, we parameterize the model selection problem within a supermodel, as follows. For simplicity, assume that the set of models $\mathcal{M}$ is finite, and index these models using the integers $1, ..., m$. Define a supermodel by using a list of parameter vectors $(\theta_1, ..., \theta_m, q)$, where each θ_i $(i = 1, ..., m)$ denotes the parameter vector of the ith model in $\mathcal{M}$, and q denotes the integer-valued parameter whose unknown true value is the index of the best model in $\mathcal{M}$. The problem of model selection is now a problem of estimating the parameter q. We discuss the nonidentifiability of q for a family of log-linear models in Section 4.3.

The remainder of this chapter proceeds as follows: Section 4.2 reviews previous work on model selection for CRC, Section 4.3 discusses several aspects of nonidentifiability, particularly from the perspective of log-linear models; Section 4.4 explores ways to structure log-linear models to reflect the data context; and Section 4.5 concludes with an example on the population of survivors of the World Trade Center attacks.

4.2 The Rise of Nonidentifiability in CRC

Fienberg (1972) considered only hierarchical models to preserve the interpretability of model parameters. The focus on interpretability could have begun to seem antiquated as a steady stream of newer and relatively complex models promised improvements in prediction accuracy. Newer models include the jackknife estimator (Burnham and Overton, 1978), the estimators of Chao et al. (1992), the martingale-inspired estimators (Lloyd, 1992), a nonparametric mixture model by Norris and Pollock (1996a), and several parametric mixture models summarized by Pledger and Phillpot (2008).

As the number of models grew, the use of $\mathcal{S}_1$ – also known as "model fishing" – became increasingly common; authors would fit several different models to the data, and pick the model with the best fit by some model scoring tool such as the Akaike information criterion (AIC). For a recent example of this, see Table 1 and Table 3 in Dorazio and Royle (2003).

Link (2003) summarily rejected the uncritical use of $\mathcal{S}_1$ by showing that the population size is not identifiable across the space of several popular families of models. Link stated that "it is likely that an analyst will be

unable to distinguish between reasonable descriptions of the heterogeneity, even when these alternative descriptions lead to vastly different inferences about population sizes" and that this problem is "inherent to all attempts to model heterogeneity in detection probability." Link went on to contrive fake data sets for which several heterogeneity models lead to substantially different inferences despite fitting the data equally well. Heated discussion ensued. Link (2006) addressed counterpoints by Pledger (2005) and Holzmann et al. (2006). Mao (2007, 2008) bolstered Link's findings on theoretical grounds, with the caveat that a lower bound on the population size may be identifiable.

The specific form of nonidentifiability discussed by Link and Mao arises in binomial mixture models. Consider an experiment that involves k lists, and let Z_i denote the number of lists that include the ith population unit. Let p_i denote the probability of capture of the ith unit on each list, and assume unit-level list independence such that the probability of the ith unit appearing on all of the lists is p_i^k. In a binomial mixture model, Z_i is a binomial random variable with k trials and success probability p_i, where p_i comes from some common mixing distribution F. The choice of mixing distribution F can have a large effect on the resulting population estimates. Link (2003) showed that F is nonidentifiable among several popular parametric models, and Mao (2007) showed that F is nonidentifiable nonparametrically.

Oddly, the nonidentifiability explored by Link and Mao has never (to our knowledge) been explicitly tied to the nonidentifiability in log-linear models for CRC that has been understood ever since Fienberg (1972). Section 4.3 unifies these two forms of nonidentifiability.

4.3 Nonidentifiability in Log-linear Models

4.3.1 Log-linear Expression of Binomial Mixture Models

Log-linear models (Chapter 3) are substantially more general than the binomial mixture models of Link and Mao. These binomial mixture models assume that, for a fixed population unit, the probability of capture on each list is equal to a single number p_i, and there are no explicit interaction effects between lists. By contrast, log-linear models can include individual list effects ($u_2 \neq u_3$, for example) or interactions between lists ($u_{12} \neq 0$, for example).

In one sense, log-linear models are more restrictive than binomial mixture models, since log-linear models formally assume homogeneity of capture probabilities across units. We emphasize the term "formally" here because log-linear models can perfectly fit any multinomial capture pattern array (such as Table 1.1) despite the homogeneity assumption. The form (3.5) is fully general in the sense that u-terms must exist that satisfy $p(y) \equiv p(y|u)$. This means that the log-linear model is compatible with any configuration of relative expected cell-count frequencies for the cross-classification array c, regardless of any underlying heterogeneity or dependency structures. Appendix A of Darroch et al. (1993) provides several details on the precise implications of assuming a homogeneous multinomial distribution for data that arise from heterogeneous capture probabilities.

We can express any binomial mixture model as a log-linear model. Consider the assumptions common to binomial mixture models: For each unit, the probability of capture is constant across the lists, and the event of cap-

ture on each list is independent of capture on the other lists. These two assumptions induce symmetries in (3.5) that lead to a reparameterization with only $k+1$ parameters $v = (v_0, ..., v_k)$,

$$\log p(y|v) = \sum_{j=0}^{k} v_j S_j(y), \tag{4.1}$$

where $S_j(y) := I(\sum_{t=1}^{k} y_t \geq j)$ is the indicator that the capture pattern includes at least j captures. Since $k+1 \leq 2^k - 1$ for $k \geq 2$, one might hope that v is identifiable. However, consistent with the notion that a supermodel of binomial mixture models can be nonidentifiable, the log-linear generalization (4.1) is also nonidentifiable, because the columns of the design matrix that correspond to v_0 and v_1 are identical for the observable data.

With $S'_j(y) := I(\sum_{t=1}^{k} y_t = j)$, a simple alternative parameterization $w = (w_0, ..., w_k)$ of (4.1) is

$$\log p(y|w) = \sum_{j=0}^{k} w_j S'_j(y). \tag{4.2}$$

Parameter vectors w (or v) exist such that (4.2) (or (4.1)) does not correspond to any binomial mixture model. Cressie and Holland (1983) provide a set of constraints on w that must hold in order for (4.2) to reflect a binomial mixture model (in fact, their setting is slightly more general).

4.3.2 Relevance of the Highest-Order Interaction

For notational simplicity, denote the highest-order interaction as $q := u_{1\cdots k}$. Aside from the objective of maintaining interpretability of lower-order pa-

rameters in models such as (3.1) and (3.4), the choice of zero for q is rather arbitrary, as we may fix q to be any real number. Thus, the highest order interaction q can be viewed as an index on an infinite class of saturated models $\mathcal{M} = \{M_q\}_{q \in \mathbb{R}}$, where $\mathbb{R}$ is the set of real numbers. The parameters in each of the models in $\mathcal{M}$ are internally identifiable, and yet there is no sense in mathematically trying to distinguish between M_{q_1} and M_{q_2} for $q_1 \neq q_2$ because both models fit the observable data perfectly, with equal degrees of freedom. Thus, log-linear models for CRC suffer from nonidentifiability within a supermodel as defined at the end of Section 7.1.

Our strong interest in the highest-order interaction may seem odd to anyone who is versed in the traditional application of linear models. A long history of model selection builds around the notion that setting high-order interactions to zero is a good way to manage the bias-variance tradeoff, consistent with Occam's razor. In fitting a linear model to estimate the various treatment effects, interactions that are higher than second-order are not commonly used, and inclusion of the highest-order interaction may be out of the question. This approach is not always appropriate for CRC, however. We will discuss the issues in terms of the following two examples:

Example A (Traditional): The cells of a contingency table contain crop yields under a variety of conditions, averaged across several growing seasons. The goal is to predict crop yields for future replications of a specific combination of conditions that appeared in the data.

Example B (CRC): The cells of a contingency table provide the counts of the population units having each capture pattern. The goal is to

estimate $c_{\mathbf{0}}$, which is missing by design.

Leaving out useful explanatory variables in a linear model for example (A) does not typically degrade the coverage probabilities of prediction intervals, provided that the distribution of errors is well-behaved, although the prediction intervals may widen. In fact, beyond the assumption that future crop yields have the same distribution as past yields, one need not assume a linear (or any other specific) relationship between variables whatsoever. One can take as the starting point the average crop yield within each cell. These averages are unbiased predictors of future crop yields.

In Example B differs sharply from Example A in that no direct estimate exists for the quantity of interest since there are no data from the unobservable cell. The only way to obtain an estimate – besides simply guessing – is to assume some set of relationships among the capture patterns. As we discussed above in the language of identifiability, no purely data-driven argument can motivate a particular set of assumptions. But one must start with something. If the data are not sufficient, what may qualify as a reasonable basis for a model?

One answer appears as though handed down from statistical tradition. Borrowing from classical problems such as Example A, we can begin with the assumption that $u_{1\cdots k} = 0$, and then build on this assumption to proceed with estimation as usual. Whether a practitioner is satisfied with this approach depends in part on the kinds of data sets that she has studied. We will demonstrate how the assumption $u_{1\cdots k} = 0$ works well in some cases, but leads to bizarre results in others.

Section 4.4 interprets the highest-order interaction $u_{1\cdots k}$, and Section 4.5 gives an example of using the data context to create a sort of prior distribution for $u_{1\cdots k}$. The use of a prior distribution on $u_{1\cdots k}$ deviates from the existing literature. Although Madigan and York (1997) and others have proposed Bayesian ways of including prior information or allowed for model selection uncertainty in other ways, all of these (to our knowledge) invoke the assumption that the highest-order interaction is zero, in contrast to our proposal.

4.4 Interpretable Models

The ability to interpret a model allows us to incorporate data context into a model selection criterion as in $\mathcal{S}_2$ or $\mathcal{S}_3$. Interpretability is itself a term that is open to interpretation. For our purposes, broadly speaking, a model is interpretable to the extent that simple relationships exist to connect the model and the data context. Although many different types of models are interpretable, we restrict attention to log-linear models, for simplicity. We begin by reviewing several observations on interpretation by Fienberg (1972), and conclude with a discussion of log-linear models with less-obvious interpretations.

Hierarchical models are often easy to interpret. A submodel of (3.5) is called hierarchical if $u_\omega := 0$ implies that $u_{\omega'} := 0$ whenever $\omega \subset \omega'$. One of the simplest hierarchical log-linear models is the independence model,

$$\log p(y|u) = u_0 + u_1 y_1 + \cdots + u_k y_k. \tag{4.3}$$

Exponentiating (4.3) reveals a direct correspondence to the product rule for independent events,

$$\begin{aligned} p(y|u) &= e^{u_0} e^{u_1 y_1} \cdots e^{u_k y_k} \\ &= e^{\frac{u_0}{k} + u_1 y_1} \cdots e^{\frac{u_0}{k} + u_k y_k} \\ &= pr(Y_1 = y_1) \cdots pr(Y_k = y_k). \end{aligned}$$

Thus, under the independence assumption, each of the coefficients u_j ($j = 1, ..., k$) in (4.3) controls the marginal probability that a random unit appears on the jth list.

Conditional on the inclusion of the lower order terms $u_1 y_1, u_2 y_2$, an additional term $u_{12} y_1 y_2$ is the interaction between the first and second list. This interaction at face value represents, for each population unit, an association between being captured on the first list and being captured on the second list. Alternatively, when the model is fitted to data from a heterogeneous population (in contradiction to the homogeneity assumption) the interaction may represent an association between the events of capture on the first and second lists, mixed across units (not for each unit individually). Higher-order interactions have similar interpretations in hierarchical models.

Nonhierarchical models may also have interesting interpretations. The model

$$\log p(y|u) = u_0 + u_{12} y_1 y_2 \tag{4.4}$$

straightforwardly encodes the situation in which the expected values in Table

1.1 must all take on the same value except for the $(1,1)$ cell, which has its own parameter. On the other hand, while the preceding sentence "interprets" the model into English, there is still no explicit connection to a data context. Thus, it may be said that (4.4) is "usefully interpreted" only when some scientific reason exists to support that units are equally alloted to each of the capture patterns except for the $(1,1)$ cell.

For every log-linear model with the highest order term $u_{1\cdots k}$ set equal to zero, eliminating all of the explicit u-terms from the model equations leads to

$$\frac{\prod_{y\in O} p(y|u)}{\prod_{z\in E} p(z|u)} = 1, \tag{4.5}$$

where O is the set of capture patterns with entries summing to an odd number, and E is the set of capture patterns summing to an even number. For $k = 3$ lists, (4.5) is equivalent to there being a constant odds radio for the expected values in any two disjoint 2×2 half-arrays of the full $2 \times 2 \times 2$ cross-classification array. This formula is in essence the same as (3.9).

Formula (4.5) holds for nonhierarchical log-linear models, and can also hold with $u_{1\cdots k} \neq 0$ in conjunction with certain constraints. For example, Darroch et al. (1993) explored a non-hierarchical log-linear representation for a heterogeneity model of a form similar to (4.2). The corresponding general parameterization (3.5) has $u_{1\cdots k} \neq 0$. Aside from this example, we know of no further effort to interpret models with $u_{1\cdots k}$ not equal to zero. Indeed, $u_{1\cdots k}$ is identifiable only if we set a lower-order term equal to zero, in which case the meaning of $u_{1\cdots k}$ is perhaps too subtle.

Interesting interpretations exist, however, if we pick a value for $u_{1\cdots k}$, perhaps as a draw from some prior distribution, before estimating the lower-order parameters. Some algebra clarifies this point: Add $\log n$ to both sides of (3.5) to get

$$\log E(C_y) = \sum_{\omega \in \Omega_k} u_\omega I_\omega(y), \tag{4.6}$$

where $\log n$ has been absorbed into u_0 on the right-hand side, and observe that $E(C_y) = np(y|u)$, where C_y is the assumed random variable corresponding to the observed cross-classification count c_y. Next, rearrange (4.6) slightly by moving the $u_{1\cdots k}$ term to the left-hand side:

$$\log E(C_y) - u_{1\cdots k} I_{1\cdots k}(y) = \sum_{\omega \in \Omega_k \setminus \{1,\ldots,k\}} u_\omega I_\omega(y). \tag{4.7}$$

Let $\mathbf{1} \in \mathcal{Y}_k$ denote the capture pattern indicating capture on every list. Let C' denote the cross-classification array C with a modification to the $\mathbf{1}$ cell as $C'_{\mathbf{1}} := e^{u_{1\cdots k}} C_{\mathbf{1}}$. Then (4.7) becomes

$$\log E(C'_y) = \sum_{\omega \in \Omega_k \setminus \{1,\ldots,k\}} u_\omega I_\omega(y), \tag{4.8}$$

a model that is hierarchical with respect to C'. Thus, the prescribed value for $u_{1\cdots k}$ is an adjustment to the $\mathbf{1}$ cell of the cross-classification data array c. The data context could justify a negative adjustment (i.e., $u_{1\cdots k} > 0$) if the analyst believes that some number of population units are structurally (i.e., with probability one) included in all of the lists, since structurally

included units typically do not belong in any probability model that is used to impute $c_{\mathbf{0}}$. Conversely, a positive adjustment to $c_{\mathbf{1}}$ may be appropriate if some number of units are structurally excluded from appearing on all of the lists simultaneously. We experiment with this type of data adjustment in the example in Section 4.5.

4.5 Example: World Trade Center Survivors

4.5.1 Background

Murphy (2009) estimated the number of people who were in the World Trade Centers (WTC) and survived after the first attack on September 11, 2001, by using three separate lists. Table 4.1 shows Murphy's estimate of the patterns of overlap between the three lists. The lists included $n_c = 8965$ distinct people in total. We begin by exposing some problems in Murphy's analysis and try to improve upon it by using $\mathcal{S}_3$ to motivate the model selection approach.

Table 4.1: Cross-classification by WTC list membership

		In List 3	Not in List 3
In List 2	In List 1	174	88
	Not in List 1	750	270
Not in List 2	In List 1	1658	1702
	Not in List 1	4323	$c_{\mathbf{0}}$

The World Trade Center Health Registry (WTCHR) gathered the three lists from their respective sources. The lists originally included covariates such as age and sex. Such auxiliary covariates can inform estimates of the

population size (Pollock, 2002). However, in personal correspondence, administrators at the WTCHR said that their detailed data sets were expunged at the conclusion of Murphy's study. This restricts us to a relatively basic CRC analysis based on Table 4.1. We build on Murphy's analysis of the data, first by considering only the numbers of Table 4.1, and subsequently by incorporating elements of the data context.

4.5.2 Analysis of Table 4.1

Murphy considered the seven hierarchical models that include an intercept term, a coefficient for each list, and any combination of the pairwise list interactions, but excluding the saturated model. Murphy fitted each of these seven models and compared them using the AIC (3.17) to pick a winner:

$$\log E(C_y) = u_0 + u_1 y_1 + u_2 y_2 + u_3 y_3 + u_{12} y_1 y_2 + u_{23} y_2 y_3.$$

Using a shorthand model notation, we assume the presence of all individual list effects, and write only the interaction terms; the above model is then "u12+u23." Murphy's resulting estimate was $\hat{n} = 13400$ with a 95% confidence interval $(13064, 13745)$.

Murphy's confidence interval computation appears to rest on the assumption that model u12+u23 is the correct model. Unfortunately, this assumption may produce a confidence interval that is too narrow by ignoring the variability in the model selection algorithm. Norris and Pollock (1996b) presented three bootstrap variance estimates that include model selection uncertainty. We use the method referred to in their paper as "Method 2,"

as follows.

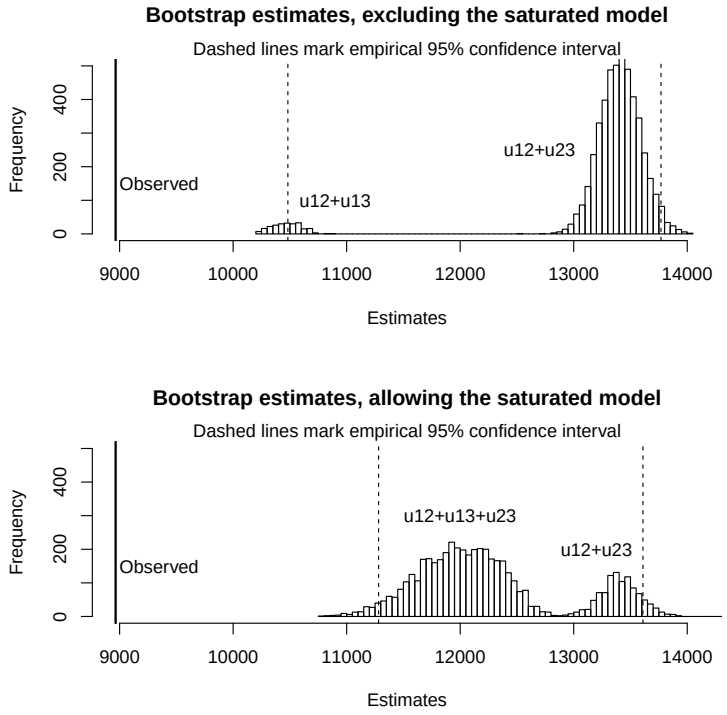

Figure 4.1: The top panel displays a histogram of 5000 bootstrap estimates based on consideration of Murphy's seven models with the best model selected by AIC. The model u12+u13 was selected roughly 5% of the time, resulting in a much lower 95% confidence interval bound. The bottom panel replicates the top panel, but with the addition of the saturated model u12+u13+u23. The vertical bold line at the left side in each plot indicates the observed number of people, $n_c = 8965$.

Assume that the true population is $n = 13400$, the estimate obtained under model $u12 + u23$. Hence the number of units not observed is $c_0 = 13400 - 8965 = 4435$. Simulate a multinomial with 13400 trials based on direct estimates $\hat{p}(y) := c_y/n$ $(y \in \mathcal{Y})$ of the probabilities for the eight multinomial outcomes. Treating the resulting nonzero capture pattern out-

comes as a bootstrap data set, select a new model (possibly not the same as $u12 + u23$) using the AIC, and use the selected model to estimate the population.

The top panel of Figure 4.1 illustrates 5000 replications of the bootstrap procedure. This procedure selects a different model, u12+u23, roughly 5% of the time. The result is a drastically wider confidence interval, with a lower bound of about 10500, in contrast with Murphy's 13064. Oddly, Murphy excluded the saturated model u12+u13+u23 in his analysis, even though this model attains the best AIC score for the original data set as well as most of the simulated data sets. In the second panel of Figure 4.1, we repeat the bootstrap simulation experiment with the saturated model included, and find a somewhat narrower confidence interval. This demonstrates the importance of including a sufficiently broad class of models in the model selection procedure.

In fact, many more log-linear models exist that are not in our set $\mathcal{M}$ of eight models considered here. One way to expand $\mathcal{M}$ is to pick several nonzero values of the highest-order interaction term u_{123} instead of only using zero, the default value. The fundamental problem with this proposal is that u_{123} is nonidentifiable (see Section 4.3.2). We suggest that appealing to the data context to set a prior on u_{123} is a good way to acknowledge the uncertainty of our model selection procedure.

4.5.3 Incorporating the Data Context

The concept of data context is vague, as one could philosophize that the entire universe is part of any context. Our goal in appealing to context

is fortunately somewhat less ambitious; if we can merely gather a bit of evidence from the context – even something so small or vague as the opinion of an "expert" – then there is hope to salvage a credible estimate from an otherwise-nonidentifiabe set of models.

We begin by simply listing several facts surrounding the data, with quotations taken from Murphy (2009):

- List 1 consisted of "individuals who volunteered by Web site or telephone" to complete a WTCHR study.
- List 2 was "supplied by businesses ... with office space in the [WTC] towers identifying employees who were present."
- List 3 was obtained from local government institutions and "included individuals with security access to the towers."
- Each individual on each list was confirmed in follow up communication to have been in the WTC at the time of the attacks.

These simple descriptions of the data context seem rather unhelpful. How can we connect the context to the model to hint at the best value for u_{123}? If it were known (or even likely) that only a fraction of employees could have security access, or conversely that only a fraction of those with security access could be employees, such information would suggest that $c_{\mathbf{1}} = 174$ is the result of a structurally constrained process, and that adjusting this cell upwards before fitting a model could improve inference. We give little weight to this possibility here, mentioning it only as an example of how

additional information on the data context (i.e., the policies that were in place for security access) could be relevant.

A more general point is that heterogeneity of capture probabilities seems to induce a positive highest-order interaction u_{123}, provided that the form of the heterogeneity is approximately the same for each list. We base this claim on the positive highest-order interaction in the quasi-symmetric log-linear model for heterogeneity in Darroch et al. (1993). Further, it is plausible that such a pattern of heterogeneity is in the WTC lists. Without knowing many details, it seems that individuals may have been required to personally verify their presence in the towers to qualify to be on each list. Different types of people have varying interest in cooperating with survey studies, and such variation of interest may plausibly persist across the three lists to a small degree. Therefore, we speculate that $u_{123} = 0.1$. This corresponds to multiplying $c_{\mathbf{1}}$ by $e^{-0.1}$, a downward adjustment by approximately 10%, as in equation (4.7).

The thoughts in the previous paragraph are almost – but not quite – pure speculation. The crucial point, however, is that failing to take such thoughts into account is perhaps an even worse form of speculation, since no clear justification exists for concluding that $u_{123} = 0$. Equally important is our ability to indicate uncertainty in our proposed value of u_{123}. Thus, we take u_{123} as a draw from a $N(0.1, 0.3^2)$ normal distribution. The choice of distributional form (and the variance) is again speculation, and should be interpreted only as a starting point. Others who follow this analysis may find justification for a completely different distribution for u_{123}.

Adjusting $c_{\mathbf{1}}$ based on $u_{123} = 0.1$ as in equation (4.7) and proceeding

with model selection by AIC using the adjusted data, we obtain the point estimate $\hat{n} = 11815$. To estimate the sampling distribution of this point estimate, including the assumption that $u_{123} \sim N(0.1, 0.3^2)$, we again follow "Method 2" of Norris and Pollock (1996a) as in Section (4.5.2). Specifically, we treat the point estimate as truth to simulate Table 4.1 as a multinomial. Next, we adjust the simulated table by multiplying the $c_{\mathbf{1}}$ cell by $e^{u_{123}}$, where u_{123} is drawn as a $N(0.1, 0.3^2)$ random variable. For the resulting table, we use the AIC to select a hierarchical log-linear model (with highest-order interaction nominally equal to zero) and generate a new estimate of the missing cell $\hat{c}_{\mathbf{0}}$, which we add to $n_c = 8965$ to get a simulated estimate.

Figure 4.2 shows the simulated sampling distribution of the estimator. The multi-modality of the histogram reflects that several different models were frequently selected over the 5000 replications. Since this simulation incorporates a probability distribution in a flavor similar to a Bayesian prior, but is otherwise frequentist, we blend the words confidence and credible to describe the empirical 95% interval in the first panel of Figure 4.2 as a confidible interval. The 95% confidible interval is $(10610, 15855)$. Compared to Murphy's confidence interval, our confidible interval is somewhat more compatible with previous estimates by USA Today and the National Institutes of Standards and Technology that used completely different methodologies, resulting in estimates ranging from about 9800 to 16500 (Murphy, 2009).

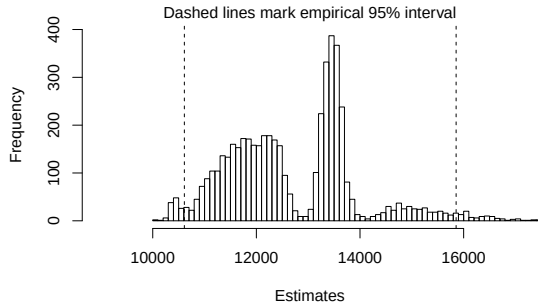

Figure 4.2: The histogram summarizes 5000 simulated population estimates based on model selection by AIC after fixing the highest-order interaction u_{123} as a draw from the $N(0.1, 0.3^2)$ distribution.

4.6 Simulation Example

A simple simulation experiment demonstrates how an important highest-order interaction can, in principle, arise as a side-effect of heterogeneity. Suppose that a population contains two types of units. Each list contains Type A units with probability 1, and each list contains Type B units with probability 0.2. Let $n = 1000$ be the size of the population. Assume that 200 units are of Type A, and 800 units are of Type B. We generate three lists according to the probabilities stated above, with independence between units as well as unit-level independence across lists. We cross-tabulate the results to get a CRC array $\mathbf{c}$.

From our omniscient position as the designers of the simulation, it is straightforward to compute the probability of every capture pattern including $\mathbf{0}$. We solve the system of 8 log-linear equations to determine the unique set of 8 log-linear parameters in (3.4) for the true generating model, obtain-

ing (approximately) the non-hierarchical model

$$\log p(y) = -0.89 - 1.39y_1 - 1.39y_2 - 1.39y_3 + 3.47y_1y_2y_3.$$

In this totally rigged simulation, the highest order interaction is clearly important. Whether the conditions described above are likely to occur in practice is a valid question.

Despite the enormity of u_{123}, one might hope that log-linear models with no highest-order interaction might still perform satisfactorily. For instance, it seems possible that the collection of interaction terms u_{12}, u_{13}, u_{23} in the saturated hierarchical model (which excludes u_{123}) may be able to indirectly compensate for the absence of u_{123}. Alternatively, the Rasch heterogeneity log-linear model (3.22) is specifically designed to account for certain types of heterogeneity within the log-linear structure (although the authors warn that the Rasch model does not account well for bimodal heterogeneity, as we have in the present simulation). To explore these possibilities, we simulate the population estimates for the saturated log-linear model and the Rasch model, and compare these against the true generating model in Figure 4.3. Both of the alternative models typically overestimate the population size by more than a factor of 10, demonstrating that an unacknowledged highest-order interaction can cause major problems.

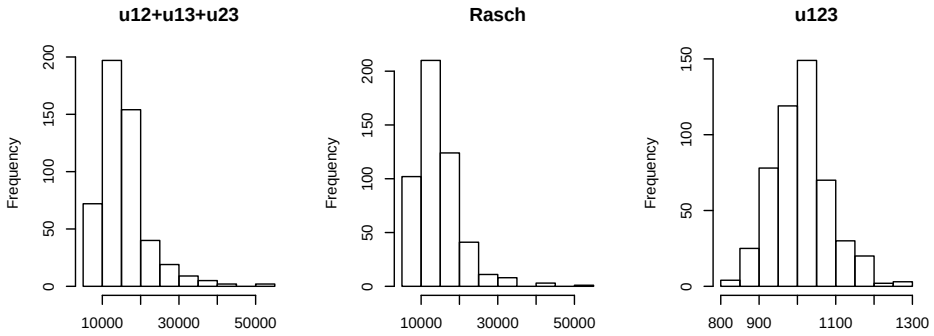

Figure 4.3: The histograms summarize 500 simulation experiments. The non-hierarchical model that includes u_{123} in the third panel performs well, while both alternative models (first and second panels) produce estimates that err by an order of magnitude.

4.7 Discussion

Our approach to model selection addresses two problems. The first problem is that of identifiability: How should a researcher proceed when confronted with a nonidentifiable likelihood function? The second problem is in the sampling design: How can a researcher build a model that is useful for prediction when the training data set is not a representative random sample? For both problems, we find it self-evident that the solution (if one exists) is to incorporate the data context – information that is external to the data.

The problems of identifiability and data representativeness seem closely related. We conjecture that nonidentifiability (within a likelihood, or across a class of models $\mathcal{M}$) tends to arise most frequently when the training data are not directly representative of the prediction set. In fact, any model can be made nonidentifiable by including a coefficient β for the indicator that a

data point has been observed. Including such a term is preposterous from the point of view of using the data to make an extrapolation, as the data contain zero information regarding β. On the other hand, a researcher who believes that the prediction set is systematically different from the training set would be irresponsible not to propose an adjustment to the predictions of the model. As the data are irrelevant for estimating β, the only way that this could be done is to appeal to the data context.

One common reaction to nonidentifiability in a likelihood function is to fuss with the model until identifiability is restored. Our study of CRC suggests that this response is not always optimal, since a nonidentifiable parameter (such as $u_{1\cdots k}$) can be used to model a systematic difference between the observed data and the unobserved data. Thus, a model search that automatically disqualifies nonidentifiable models may introduce a major source of error.

Even if we use frequentist methods to fit and assess the goodness-of-fit of each model, the strategy $\mathcal{S}_3$ is Bayesian in spirit, since the inclusion of the data context in the model selection criterion is a lot like putting a prior on the model index q. Unlike a standard Bayesian prior, the importance of the data context does not necessarily diminish as the sample size grows. In a nonidentifiable comparison of two models $M_1, M_2 \in \mathcal{M}$, the likelihood function evaluations may be equal, providing no update to the relative prior probabilities of the two models, necessitating continued reliance on the data context. Important properties of Bayesian inference with nonidentifiable likelihood functions are not yet completely understood; see Gustafson (2005) and the large number of comments that followed. See also

Neath and Samaniego (1997).

We consider $\mathcal{S}_3$ to be an ideal approach, but a difficult one, and we did not fully attain this ideal in our example (Section 4.5). We used the data context in a rather weak way to include u_{123}, and we used the AIC to perform sub-model selection conditional on u_{123}. Ideally, we would like to use the data context also to augment the AIC-based comparison of submodels, potentially modifying the distribution displayed in Figure 4.2.

Chapter 5

Local Log-linear Models

This chapter presents our main work, a method of building log-linear models locally, or fitting a separate model for each observed point in the covariate space. Throughout, we assume that the population is closed and that there are no errors of record linkage.

5.1 Introduction

Heterogeneity of capture probabilities can cause bias in log-linear models (Darroch et al., 1993; Fienberg et al., 1999). One way to reduce heterogeneity bias is to post-stratify on auxiliary covariates. In a human population, individuals may be grouped by age, such as 0-10 years, 11-20 years, etc. After fitting a model to each age group, we sum the resulting estimates across groups to estimate the population size. For instance, the coverage evaluation of the United States 2000 Census used more than 400 post-strata (Citro et al., 2004).

A fundamental challenge in post-stratification is to determine the optimal number of strata. To remove within-stratum heterogeneity, it is desirable to have as many strata as possible. At the same time, we must maintain a minimum stratum size to control the variance of estimates. We address this trade-off by applying log-linear models to individuals, the smallest possible strata, while "borrowing strength" to maintain adequate effective sample size.

Specifically, we select and fit a local log-linear model for the capture pattern associated with each individual. To illustrate this with a human population, suppose that we observe exactly one person of age 19. This person constitutes a post-stratum of size 1. A single observation is, of course, not enough to select a log-linear model, but we can get an adequate effective sample size by using a locally weighted average of the capture patterns of people with ages close to 19. Thus, we fit each local log-linear model to a local average.

Our approach is closely related to several existing methods. Huggins (1989) and Alho (1990) developed logistic regression models that allow capture probabilities to vary with auxiliary covariates when there are only two lists. Yip et al. (2001) extended their method to include certain list interactions when there are more than two lists. Chen and Lloyd (2002) used "local post-stratification," which essentially replaces the Alho-Huggins logistic regression with a nonparametric regression. Zwane and van der Heijden (2004) fit log-linear models that used penalized splines to express dependence on a continuous covariate. With less of an emphasis on interactions between lists, Hwang and Huggins (2011) used a semi-parametric logistic regression

involving local polynomials to model the effect of a continuous covariate, and Stoklosa and Huggins (2012) took a similar approach with penalized splines instead of local polynomials. Our treatment differs from those above by allowing the form – and not only the fitted values – of the model to vary over the covariate space. This generality can meaningfully improve estimates, as we demonstrate via simulation.

Many models treat heterogeneity as a latent feature, without using covariates (Darroch et al., 1993; Manrique-Vallier and Fienberg, 2008; Pledger and Phillpot, 2008). Such models are especially important when the auxiliary covariates are noninformative or unavailable. However, when informative covariates are available, their inclusion adds a significant dimension to the value of a capture-recapture study by enabling estimation of the rate of missingness, or the number of unobserved units divided by the number of observed units, at each point in the covariate space. Thus we learn about not only the population size, but also its composition. Many existing methods relate covariates to the detection probabilities, but our approach is exceptionally specialized to this task, since we build a full log-linear model at each observed covariate vector.

5.2 Basic Framework

We make a key assumption that the literature on auxiliary-covariate models of heterogeneity often leaves unstated. Namely, we assume the existence of a function $r(y|x)$ that is piecewise smooth in x and satisfies $p(i, y_{i\cdot}) = r(y_{i\cdot}|x_{i\cdot})$ $(i = 1, ..., n)$. This is a strong assumption, requiring

that the covariates x fully explain any variation in the capture probabilities.

Recall that $\mathbf{0}$ denotes the row vector of k zeros. Define the detection function $\psi(x) = 1 - r(\mathbf{0}|x)$, which is the probability that a unit with covariates x appears in at least one of the lists. The [pseudo] HT estimator (2.1 is a convenient way to relate the regression function r to the population size. In practice, using (2.1) requires us to estimate the detection function ψ. Our estimator will join several existing capture-recapture estimators that take this route in essence, including those of Alho (1990), Chen and Lloyd (2002), and Zwane and van der Heijden (2004). We begin by putting (2.1) into a different form. Define a function

$$\pi(y|x) := \frac{r(y|x)}{\sum_{z \neq \mathbf{0}} r(z|x)} = \frac{r(y|x)}{\psi(x)}. \tag{5.1}$$

For each nonzero $y \in \mathcal{Y}$, $\pi(y|x)$ is the conditional probability that a unit with covariates x has capture pattern y, given that the unit appears on at least one list.

Plugging an estimate $\hat{\pi}(\mathbf{0}|x)$ of $\pi(\mathbf{0}|x)$ into (5.1) and expanding (2.1) in terms of (5.1) leads us to an estimator involving the sum of the unit-level estimates:

$$\hat{n} := n_c + \hat{c}_{\mathbf{0}}, \text{ where } \hat{c}_{\mathbf{0}} := \sum_{i=1}^{n_c} \hat{\pi}(\mathbf{0}|x_{i\cdot}). \tag{5.2}$$

Thus, our challenge is to derive useful estimates $\hat{\pi}(\mathbf{0}|x_{i\cdot})$ $(i = 1, ..., n_c)$, which will be done in Section 5.3.

We propose the use of local as well as global measures of model per-

formance. Specifically, let $n_c(x)$ denote the number of units observed at covariate x, let $c_{\mathbf{0}}(x)$ denote the corresponding number of missing units, and let $\hat{c}_{\mathbf{0}}(x)$ denote the corresponding estimate, computed as $n_c(x)\hat{\pi}(\mathbf{0}|x)$. Define the local root-mean-square error as

$$RMSE(x) = \sqrt{E[\{\hat{c}_{\mathbf{0}}(x) - c_{\mathbf{0}}(x)\}^2]}. \tag{5.3}$$

Similarly, we use the root-mean-square error $\sqrt{E(\hat{c}_{\mathbf{0}} - c_{\mathbf{0}})^2}$ as a global measure of model performance. These measures can be estimated only in simulations, as validation data typically do not exist in real applications.

Define an (arbitrary) ordering of the capture patterns $\mathcal{Y}$ so that $\mathcal{Y}_j$ denotes the jth nonzero capture pattern ($j = 1, ..., 2^k - 1$). Let Π and A be matrices of dimension $n_c \times (2^k - 1)$ with elements $\Pi_{ij} = \pi(\mathcal{Y}_j|x_{i\cdot})$ and $A_{ij} = I(y_{i\cdot} = \mathcal{Y}_j)$. Thus, the ith row A_i of A indicates the multinomial outcome corresponding to the multinomial probabilities given in the ith row Π_i of Π.

The vector average $\sum_{i=1}^{n_c} A_i/n_c$ contains the empirical relative frequencies of the nonzero capture patterns. This average, together with n_c, contains the same information as $\mathbf{c}$, and so is sufficient for a traditional log-linear model, as in Fienberg (1972). To include heterogeneity that is associated with covariates, we fit local log-linear models to local averages of the form

$$\widehat{\Pi}_i = \sum_{t=1}^{n_c} w_t^i A_t \ (i = 1, ..., n_c), \tag{5.4}$$

where each w^i is a normalized vector of nonnegative weights of length n_c. For

example, any basic kernel smoother or weighted k-nearest-neighbors regression can be expressed as (5.4). Although a local average $\widehat{\Pi}_i$ is an estimator of Π_i in its own right (hence, the notational similarity), we use $\widehat{\Pi}_i$ only as the "data" for building a local log-linear model. The ith fitted local model, in turn, implies an estimate $\hat{\pi}(\mathbf{0}|x_{i\cdot})$ as needed for (5.2).

One can specify a vector of weights $w^i (i = 1, ..., n_c)$ by stating a standard kernel and picking a smoothing bandwidth based on subjective researcher expertise. Although data-driven methods exist for bandwidth selection in local regression, we caution against the careless use of such methods in our specific context for reasons that we discuss in Section 5.3.3. We take the weights as known and fixed.

For each local average $\widehat{\Pi}_i$, we define the effective sample size of the ith row as

$$\eta_i := \frac{\left(\sum_{t=1}^{n_c} w_t^i\right)^2}{\sum_{t=1}^{n_c} (w_t^i)^2} = \frac{1}{\sum_{t=1}^{n_c} (w_t^i)^2} \ (i = 1, ..., n_c). \tag{5.5}$$

Section 5.3.3 motivates this definition, which may have originated with Kish (1965), page 259.

5.3 Estimating $\pi(\mathbf{0}, x)$

We use a local log-linear model $\mathcal{M}_i$ for each ith row $\widehat{\Pi}_i$ to estimate the missing cell $\pi(\mathbf{0}|x_{i\cdot})$ $(i = 1, ..., n_c)$. While having as many as n_c models for n_c points looks like overfitting in the extreme, it is important to notice that the models are highly correlated across the covariate space, since $\widehat{\Pi}_i$ is

continuous in $x_{i\cdot}$. If the difference between $x_{i_1\cdot}$ and $x_{i_2\cdot}$ is small, than the difference between the local averages $\widehat{\Pi}_{i_1}$ and $\widehat{\Pi}_{i_2}$ is also small, effectively constraining $\mathcal{M}_{i_1}$ to be similar to $\mathcal{M}_{i_2}$. Specifically, $x_{i_1\cdot} = x_{i_2\cdot}$ implies that $\mathcal{M}_{i_1} = \mathcal{M}_{i_2}$.

5.3.1 *Local* Log-linear Models, a Special Case

Refer to Chapter 3 for a review of log-linear models.

Fix i in $1, ..., n_c$. Let $\mathcal{W}_i$ denote the set of indices of all nonzero entries of the vector of weights w^i, and let $n_i = |\mathcal{W}_i|$, the number of nonzero entries. We describe a local log-linear model $\mathcal{M}_i$ for the smoothed data $\widehat{\Pi}_i$. Throughout this section we consider only the special case that is defined by the following two assumptions:

- Boxcar assumption: $w^i_t = 1/n_i$ for all $t \in \mathcal{W}_i$.
- Local homogeneity: $r(y|x_{t_1\cdot}) = r(y|x_{t_2\cdot})$ for all indices $t_1, t_2 \in \mathcal{W}_i$ and $y \in \mathcal{Y}$.

The boxcar assumption says simply that the nonzero weights are uniform, which holds for any boxcar kernel smoother or k-nearest-neighbors regression with uniform weights. The local homogeneity assumption requires that the capture probabilities are constant over $\mathcal{W}_i$. Homogeneity is a standard assumption (at least formally) for classical log-linear models.

By the definition (5.5), the boxcar assumption gives $\eta_i = n_i$, and the

vector

$$\eta_i \widehat{\Pi}_i \;=\; \eta_i \sum_{t \in \mathcal{W}_i} \frac{1}{n_i} A_t = \sum_{t \in \mathcal{W}_i} A_t \tag{5.6}$$

is a sum of multinomials. Local homogeneity implies that the η_i terms in the sum are identically distributed. We have already assumed independence between units, and it follows that $\eta_i \widehat{\Pi}_i$ is a multinomial random variable with η_i trials and probabilities Π_i. With $k = 3$, we apply the saturated local log-linear model (3.5) to $\eta_i \widehat{\Pi}_i$, replacing the parameter vector u with a local parameter vector u^i. Since the entries of $\eta_i \widehat{\Pi}_i$ are the elements of the set $\{\eta_i \hat{\pi}(y|x_{i\cdot})\}_{y \neq \vec{0}}$, the local likelihood function is

$$L_i(u^i | \eta_i \widehat{\Pi}_i) = \frac{\eta_i!}{\prod_{y \neq \vec{0}} \{\eta_i \hat{\pi}(y|x_{i\cdot})\}!} \prod_{y \neq \vec{0}} \pi(y|u^i)^{\eta_i \hat{\pi}(y|x_{i\cdot})}. \tag{5.7}$$

Let $\hat{u}^i$ denote the parameter estimates found by maximizing (5.7). An important special case is when the kernel is infinite, or when all the units have equal weight so that each local average (5.4) coincides with the global average. Then $\eta_i = n_i = n_c$, and the local likelihood (5.7) coincides with the global likelihood (3.2).

We obtain various submodels of the saturated model (3.5) by removing terms. The independence model for three lists encodes the assumption that the probability of capture on each list is independent of the event of capture on any other list:

$$\log \pi(y|u^i) = \;\; u_0^i + u_1^i y_1 + u_2^i y_2 + u_3^i y_3. \tag{5.8}$$

We emphasize that one can do model selection locally. If $i_1 \neq i_2$, the models $\mathcal{M}_{i_1}$ and $\mathcal{M}_{i_2}$ may be of completely different forms. For example, the parameter vector u^{i_1} need not be of the same dimension as u^{i_2}.

Given estimated log-linear parameters $\hat{u}^i$, we estimate $\pi(\vec{0}|x_{i\cdot})$ by projecting the corresponding log-linear model onto the missing cell,

$$\hat{\pi}(\vec{0}|x_{i\cdot}) := \pi(\vec{0}|\hat{u}^i) = \exp(\hat{u}^i_0), \tag{5.9}$$

and this is all that is needed to construct the population size estimate (5.2).

5.3.2 Local Log-linear Models, the General Case

We derived the likelihood (5.7) from the boxcar and local homogeneity assumptions. Removing either of these assumptions makes $\eta_i \widehat{\Pi}_i$ a nontrivial mixture of multinomials, such that (5.7) need not be exactly equal to the probability of the data $\eta_i \widehat{\Pi}_i$ given the parameters u^i. Relaxing the boxcar assumption means that the local effective sample size η_i does not equal the number of nonzero weights n_i. If η_i is not integer-valued, exact evaluation of (5.7) requires a continuous generalization of the factorial function, the Gamma function. Relaxing local homogeneity means that some local heterogeneity in capture probabilities may occur. Then $\widehat{\Pi}_i$ is a mixture of not-identically distributed multinomials. Provided that the true regression function $r(y|x)$ is sufficiently smooth in the covariates x, and provided that the bandwidth of the smoother is sufficiently narrow, the corrupting effects of local heterogeneity may be limited.

Conditional on having selected a model $\mathcal{M}_i$, maximizing L_i is equivalent

to maximizing

$$\begin{aligned}\sum_{y\neq\vec{0}} \hat{\pi}(y|x_{i\cdot}) \log \pi(y|u^i) &= \sum_j \hat{\pi}(\mathcal{Y}_j|x_{i\cdot}) \log \pi(\mathcal{Y}_j|u^i) \\ &= \sum_{t=1}^{n_c} \sum_j w_t^i \log\{\pi(\mathcal{Y}_j|u^i)^{I(y_{t\cdot}=\mathcal{Y}_j)}\},\end{aligned}$$

where the final term is a weighted likelihood function, much like the standard weighted likelihood functions for local polynomial regression (Loader, 1999). That is, the approximations involved in relaxing the boxcar and local homogeneity assumptions to use L_i for parameter estimation are analogous to the approximations involved in the use of standard of weighted likelihoods.

5.3.3 Local Model Selection

Although the Bayesian model averaging approach of Madigan and York (1997) is attractive, building the model separately for each observed unit requires superior computational speed for even a dataset of moderate size. Saving a Bayesian implementation for future work, we adapt the AICc (3.18) as a practical way to facilitate fast automated local model selection. For the ith unit, define the local AICc,

$$AICc_i := -2 \log L_i(\hat{u}(x_{i\cdot})) + 2K_i + \frac{2K_i(K_i+1)}{\eta_i - K_i - 1}, \tag{5.10}$$

where $\hat{u}$ is the local maximum likelihood estimate, K_i is the number of free parameters in the local log-linear model for the ith unit, and η_i is defined in (5.5).

For the special case in which $\widehat{\Pi}$ is generated from a boxcar kernel, $\eta_i\widehat{\Pi}_i$ is integer-valued, and the $AICc_i$ corresponds to the usual AICc. For other kinds of kernels, with non-uniform weights, we wish to continue to work with the AICc and the pseudo-likelihood L_i for model selection and parameter estimation, even though (as we mentioned previously), the vector of locally smoothed capture pattern frequencies $\eta_i\widehat{\Pi}_i$ is not exactly multinomial. In this sense, the use of the $AICc_i$ for local model selection is intrinsically ad hoc when the weights come from any kernel (not only a boxcar kernel). Our intent is for the $AICc_i$ to serve as an *approximate* generalization of the AICc to the case of nonuniform weights.

We motivate the definition (5.5) of η_i in terms of the $AICc_i$. Heuristically, the importance of η_i is clear. If η_i is too large, the $AICc_i$ will tend to select models with too many parameters, and if η_i is too small, the $AICc_i$ may select models with too few parameters.

The definition (5.5) is a consequence of the following conditions:

- C1: When the kernel is a boxcar kernel, η_i must equal n_i, the number of units with nonzero weights.

- C2: Up to rounding error, the element-wise variances of $\eta_i\widehat{\Pi}_i$ should be the same as the corresponding variances of a multinomial random variable with η_i trials and conditional outcome probabilities Π_i. (Rounding error does not corrupt this comparison in the special case in which all of the elements of $\eta_i\widehat{\Pi}_i$ are integer-valued.)

We now derive (5.5) by applying the conditions C1 and C2. We begin with C2. Suppose that $\widehat{\Pi}_i(y)$ is the entry of the vector $\widehat{\Pi}_i$ that corresponds

to the capture pattern y. Then

$$\begin{aligned} Var(\eta_i \widehat{\Pi}_i(y)) &= Var\left(\eta_i \sum_{t=1}^{n_c} w_t^i I(Y_t. = y)\right) \\ &= \eta_i^2 \pi(y|x_i)(1-\pi(y|x_i)) \sum_{t=1}^{n_c} (w_t^i)^2 \end{aligned}$$

Next, if η_i is an integer, let $B_y(i)$ denote a binomial random variable with η_i trials and success probability $\pi(y|x_i)$. Then

$$Var(B_y(i)) = \eta_i \pi(y|x_i)(1-\pi(y|x_i)).$$

To satisfy condition C2, the equality $Var(B_y(i)) = Var(\eta_i \widehat{\Pi}_i(y))$ must hold. That is, we need

$$\eta_i \pi(y,x_i)(1-\pi(y,x_i)) = \eta_i^2 \pi(y,x_i)(1-\pi(y,x_i)) \sum_{t=1}^{n_c} (w_t^i)^2,$$

and the definition (5.5) merely states the nonzero solution.

Finally, it is easy to see that the definition (5.5) also satisfies condition C1. If we construct $\widehat{\Pi}_i$ from a boxcar kernel, there exist n_i nonzero weights $w_{(1)}^i, ..., w_{(n_i)}^i$ that are each equal to $1/n_i$, and all other weights are zero, and

$$\eta_i := \frac{1}{\sum_{t=1}^{n_c} (w_t^i)^2} = \frac{1}{n_i (1/n_i)^2} = n_i.$$

The choice of kernel bandwidth for the weights in the local average $\widehat{\Pi}_i$ is intimately related to the selection of local log-linear models. If the bandwidth for $\widehat{\Pi}_i$ is too small, then η_i is small, and the criterion IC_i tends to

favor an extremely sparse local model even if important high-order interactions are present. When the bandwidth is large, the criterion IC_i tends to favor a local model with many parameters. The fact that increasing the bandwidth can correspond to increasing [local] model complexity, the usual bias-variance tradeoff that traditionally guides bandwidth selection is of questionable value. In addition to bias and variance, one must consider the tradeoff between a bandwith that is small enough to reduce heterogeneity and yet large enough to facilitate the selection of local log-linear models with enough parameters to capture dependencies between lists. This intuition guides our ad-hoc choices of bandwidth, and we leave data-driven bandwidth selection methods for future work.

One aspect of our criterion (5.10) bears sharp contrast with previous uses of information criteria for problems involving data smoothing. In a typical nonparametric regression problem, the sample size is unambiguous, and the trace of the "hat" matrix approximates the number of parameters in the model. Our emphasis on local model selection leads to the converse: we can count the number of parameters directly, but we compute η_i using the ith row of the "hat" matrix to approximate the sample size.

5.4 Bootstrap Variance Estimation

Treating $\eta_i \widehat{\Pi}_i$ as an approximately multinomial random variable, the asymptotic variance formulas of Fienberg (1972) could be relevant for specific local log-linear models. However, Norris and Pollock (1996b) emphasized the importance of including model uncertainty in the variance estimate. Norris

and Pollock (1996b) proposed several bootstrap methods that include model uncertainty in capture-recapture settings; we adopt their "Method 2" to estimate the unconditional variance of the population size estimate $\hat{c}_0$ defined in (5.2). We describe the method in detail, including modifications to deal with the auxiliary covariates.

The first step is to simulate covariate vectors of the unobserved units. Together with the observed covariates $x^c_{\cdot\cdot}$, these new covariate vectors define a population that is consistent with the model. The second step is to randomly assign a capture pattern for each unit, discarding all units that are assigned the $\mathbf{0}$ capture pattern. The third step is to select and fit local log-linear models to the simulated data to obtain a bootstrap estimate $\hat{c}^{boot}_{\mathbf{0}}$. Replicating the bootstrap B times gives a set $\{\hat{c}^{boot}_{\mathbf{0}}(1), ..., \hat{c}^{boot}_{\mathbf{0}}(B)\}$, and the variance of this set is the bootstrap estimate of $Var(\hat{c}_{\mathbf{0}})$. The following subsections provide details on the first two steps.

5.4.1 Simulating Unobserved Units for the Bootstrap

The covariate matrix $x^c_{\cdot\cdot}$ contains a row for the covariate vector of each observed unit. We simulate approximately $\hat{c}_{\mathbf{0}}$ additional covariate vectors to represent the unobserved units. According the fitted local model, $\hat{\pi}(\mathbf{0}|x_{i\cdot})$ is the number of unobserved units with covariate vector $x_{i\cdot}$.

However, $\hat{\pi}(\mathbf{0}|x_{i\cdot})$ is not generally an integer, and it is not clear how to interpret non-integer numbers of units. Much like Zwane and van der Heijden (2003), we use random rounding to replace $\hat{\pi}(\mathbf{0}|x_{i\cdot})$ with a whole number, as follows. Decompose each $\hat{\pi}(\mathbf{0}|x_{i\cdot})$ into its integer and decimal parts, $\hat{\pi}^{int}_i$ and $\hat{\pi}^{dec}_i$, such that $\hat{\pi}(\mathbf{0}|x_{i\cdot}) = \hat{\pi}^{int}_i + \hat{\pi}^{dec}_i$. Let $\tilde{\pi}_i$ denote the result of randomly

rounding $\hat{\pi}(\mathbf{0}|x_{i\cdot})$, where one rounds up to the next larger integer $\hat{\pi}_i^{int}+1$ with probability $\hat{\pi}_i^{dec}$, and rounds down to the next smaller integer π_i^{int} with probability $1-\hat{\pi}_i^{dec}$.

Let $c_{\mathbf{0}}^{sim}=\sum_i \tilde{\pi}_i$, noting that $E(c_{\mathbf{0}}^{sim})=\hat{c}_{\mathbf{0}}$. Let $x_{\cdot\cdot}^{sim}$ denote a $c_{\mathbf{0}}^{sim}\times q$ matrix of covariate row vectors that are copied from $x_{\cdot\cdot}^{c}$ according to the nonzero elements of $\{\tilde{\pi}_i : i=1,...,n_c\}$. For example, if $\tilde{\pi}_7=2$, then $x_{\cdot\cdot}^{sim}$ contains two rows which are replicates of the 7th observed covariate vector $x_{7\cdot}$. Finally, let $x_{\cdot\cdot}^{c+sim}$ denote the $(n_c+c_{\mathbf{0}}^{sim})\times q$ matrix of covariates formed by appending $x_{\cdot\cdot}^{sim}$ to the bottom of $x_{\cdot\cdot}^{c}$. The matrix $x_{\cdot\cdot}^{c+sim}$ represents the full population to be used for the bootstrap. The number of new units $c_{\mathbf{0}}^{sim}$ tends to be close, but not generally equal, to $\hat{c}_{\mathbf{0}}$. Thus, random rounding introduces some variability that is not part of the modeling process, and this may slightly inflate the bootstrap variance, leading to conservative confidence intervals.

5.4.2 Assigning Capture Patterns for the Bootstrap

Building on the definition (5.1), define estimates for $r(y|x_{i\cdot})$ as

$$\hat{r}(y|x_{i\cdot}) := \hat{\pi}(y|x_{i\cdot})\hat{\psi}(x_{i\cdot}) = \frac{\hat{\pi}(y|x_{i\cdot})}{1+\hat{\pi}(\mathbf{0}|x_{i\cdot})} \ (i=1,...,n_c+c_{\mathbf{0}}^{sim}; y\in\mathcal{Y}),$$

where $\hat{\pi}(y|x_{i\cdot})$ is an element of $\widehat{\Pi}_i$ if $y\neq\mathbf{0}$, and $\hat{\pi}(\mathbf{0}|x_{i\cdot})$ is defined in (5.9). Draw the capture pattern for the ith unit from the set $\mathcal{Y}$ of possible capture patterns according to the multinomial probabilities $\{\hat{r}(y|x_{i\cdot}) : y\in\mathcal{Y}\}$.

5.5 Performance Evaluation

5.5.1 Simulation I

We used simulation to compare local log-linear models against the additive multinomial logit model (Zwane and van der Heijden, 2004) when the form of the generating model varies over a single continuous covariate. In each simulation, we assign a population of size n = 5000 uniformly over the integers $x = 1, ..., 100$. Assigning many (i.e, 50) units to each unique covariate lightens the computational burden for local log-linear models, since it suffices to select a single local model at each of the 100 possible values of x instead of selecting a model for each of several thousand observed units. We define the generating model for simulating the capture patterns in terms of two log-linear models for three lists,

$$\begin{aligned} \log p_v(y) &= v_0 + v_1 y_1 + v_2 y_2 + v_3 y_3 + v_{12} y_1 y_2 + v_{13} y_1 y_3 + v_{23} y_2 y_3 \\ \log p_w(y) &= w_0 + w_1 y_1 + w_2 y_2 + w_3 y_3 + w_{12} y_1 y_2 + w_{13} y_1 y_3 + w_{23} y_2 y_3. \end{aligned}$$

Let

$$\begin{aligned} v &= (v_1, v_2, v_3, v_{12}, v_{13}, v_{23}) = (1, 1, 1, 0, 0, 0) \\ w &= (w_1, w_2, w_3, w_{12}, w_{13}, w_{23}) = (-1.5, -1.5, 1, 1.5, 2) \end{aligned}$$

Note that the first log-linear model has no list interactions, and the second model has interactions between every pair of lists. We define the generating model for the simulation in terms of v and w, beginning with model v and

transitioning smoothly to model w, as displayed in the first panel of Figure 5.1. The figure shows the relative frequencies $\{\pi(y|x)\}_{y\in\mathcal{Y}}$, defined in terms of $p(y|x)$ as in (5.1), in a stacked form.

To obtain the smooth transition shown in the first panel of Figure 5.1, let Φ denote the cumulative distribution function of a standard normal random variable, and (somewhat arbitrarily) let

$$T(x) = \frac{\Phi\{5(x-10)/20\} - \Phi(-5*9/20)}{\Phi(2.5) - \Phi(-2.25)} \ (x = 1, ..., 20).$$

Thus, $T(x)$ is an s-shaped "transition" curve that is 0 at $x = 1$ and is 1 at $x = 20$. For the generating model we take $u(x) = \{u_1(x), u_2(x), u_3(x), u_{12}(x), u_{13}(x), u_{23}(x)\}$ as

$$u(x) = \begin{cases} v & (x = 1, ..., 40) \\ \{1 - T(x)\}v + T(x)w & (x = 41, ..., 60) \\ w & (x = 61, ..., 100). \end{cases}$$

We define the multinomial capture probabilities at each x according to the saturated log-linear model of the same form as w or v but with parameter values of u. The requirement that the multinomial probabilities must sum to 1 uniquely determines the value of $u_0(x)$.

We simulated 2000 replications of the capture-recapture experiment, with the capture patterns drawn as independent multinomials according to the relative capture pattern frequencies illustrated in the first panel of Figure 5.1. On each replication of the experiment, we performed local log-linear modeling, with weights defined using the Epanechnikov kernel with a

bandwidth of 12, producing local effective sample sizes on the order of 600. The exact choice of bandwidth was arbitrary, but the order of magnitude was motivated by relatively basic (i.e., without covariates) simulation experiments that seemed to suggest that sample sizes of several hundred are needed to produce reasonably stable model selection results for log-linear models with three lists. In a post-hoc analysis, we repeated the simulations with the kernel bandwidth set to 10 and 14, resulting in slightly worse and significantly better performance, respectively.

For comparison, we replicated the implementation of the additive multinomial logit model as in Zwane and van der Heijden (2004), who used the `VGAM` package (Yee, 2010) in `R`. In addition, we partitioned the data into post-strata of approximately equal size and selected a log-linear model on each stratum using the Akaike information criterion with the small sample correction of Hurvich and Tsai (1989). Table 5.1 summarizes the performance of each model for the task of estimating $c_{\mathbf{0}}$, the number $c_{\mathbf{0}}$ of units that were not captured in each simulation. The post-stratification with 5 post-strata had the best results among several numbers of strata that we tried. Table 5.1 shows all three methods performing comparably well. To put the biases into perspective, $c_{\mathbf{0}}$ was typically around 1300 in these experiments.

Table 5.1: Simulation on the performance of local log-linear models

Model	RMSE	bias	s. dev	95% interval width
Local log-linear	149	-8	149	579
Additive multinomial logit	152	-62	139	547
Log-linear, 5 post-strata	153	14	152	593

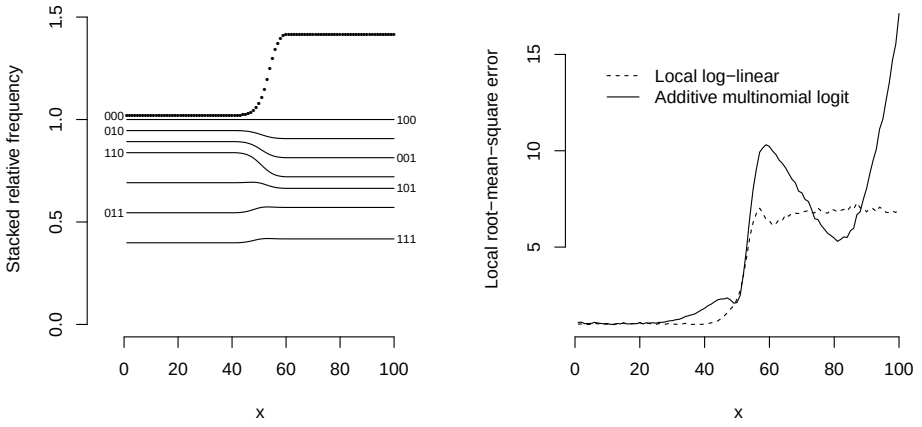

Figure 5.1: The first panel illustrates the probability structure of the generating model. The relative frequencies of the capture patterns (i.e., "111", "011", ...) are plotted as functions of x in a stacked form. For example, the curve labeled "011" represents the sum $\pi\{111|x\} + \pi\{011|x\}$. The relative frequencies of observable capture patterns sum to 1, the horizontal line, labeled "100". Above this horizontal line, the estimates $\hat{\pi}(\vec{0}|x_i)$ are plotted as $\hat{\pi}(\vec{0}|x) + 1$. The dotted curve, labeled "000" indicates the rate of missingness at each value of x for the generating model. For example, when $x < 40$, the rate of missingness is less than 0.05, and for $x > 60$ the rate of missingness is approximately 0.45. The second panel shows the local root-mean-square error (5.3) for local log-linear models and for the additive multinomial logit model for 2000 simulation replications.

The local root-mean-square error (5.3) is an important aspect of model performance that is not reflected in Table 5.1. We compute the empirical local root-mean-square error at each x across the 2000 simulation replications and plot the result in the second panel of Figure 5.1. We conclude that local log-linear models can outperform the additive multinomial logit model in terms of the local error in certain settings.

5.5.2 Simulation II

The simulation in the previous section describes an extremely unique scenario. To broaden our understanding of model performance, we devise yet another probability structure on which to run a new set of simulations. The first panel of Figure 5.2 shows the generating model, and the second panel shows the local RMSE curves for the local log-linear and additive multinomial logit approaches. The generating model is defined in terms of three different log-linear models, with transitions between the models guided by a function that is analogous to the "transition" curve $T(x)$ in the previous section. We omit the details. The second panel of Figure 5.2 shows that, in this instance, the additive multinomial logit model performs better than local log-linear models in terms of the local RMSE (5.3).

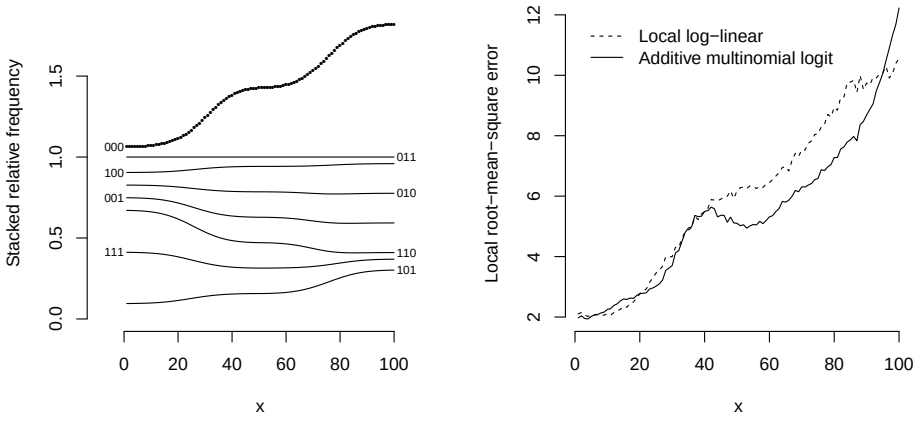

Figure 5.2: The structure and interpretation of this figure is exactly analogous to that of Figure 5.1, but with the new generating model in the left panel.

5.6 Sampling Distribution of Estimate

This section performs a sanity check to ensure that the sampling distribution of the population size estimate from local log-linear models is similar to the sampling distribution for ordinary log-linear models if we hold the effective sample size (5.5) constant and enforce homogeneity.

For $k = 3$, we apply the saturated hierarchical log-linear maximum likelihood estimator (3.8) for two separate cases. In both cases, we draw a fixed number of units according to the conditional cell probabilities $\pi(y|x) = \pi(y) = 1/(2^k - 1) = 1/7$ for all $y \neq \mathbf{0}$. That is, the data generating process is homogeneous with respect to both the units and the lists. In case (a), we take $n_c = 50$ and define $\hat{\pi}(y) = (1/50) \sum_{t=1}^{50} I(y_{t\cdot} = y)$. In case (b), we take $n_c = 141$ and define $\hat{\pi}(y) = \sum_{t=1}^{200} w_t I(y_{t\cdot} = y)$, where

we define the vector of weights w as the standard normal density evaluated over 141 evenly-spaced points in the interval $(-5, 5)$. We choose $n_c = 141$ because the resulting effective sample size turns out to be approximately equal to 50, so that cases (a) and (b) are, in some sense, comparable (see Section 5.3.3).

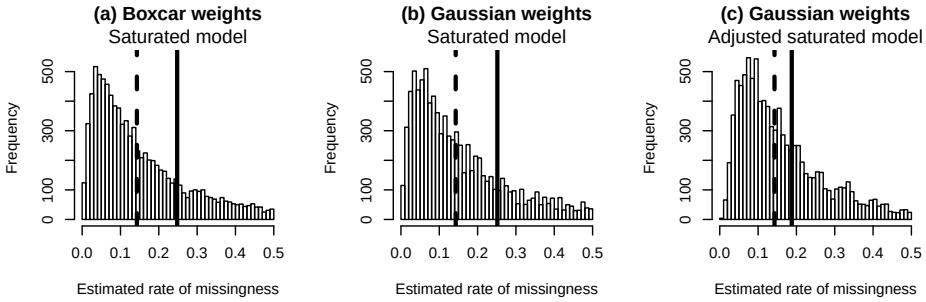

Figure 5.3: Simulated sampling distributions (truncated at 0.5) are displayed for three modeling scenarios when the true conditional distribution of capture patterns is uniform across units and lists. The dashed vertical line represents the true (expected) rate of missingness, which is $\pi(\mathbf{0}) = 1/7$. The solid vertical line marks the empirical mean of the sampling distribution for the corresponding estimator. In the first case (a), we use (3.8) on ordinary multinomial data $\hat{\Pi} = \mathbf{c}/n_c$ to estimate the missing cell $\hat{\pi}(\mathbf{0})$. The second case (b) differs from (a) only in that $\hat{\Pi}$ is obtained as a multinomial mixture with Gaussian weights as in (5.4). Finally, (c) shows the estimates from the adjusted saturated model as specified by (3.10).

Panels (a) and (b) of Figure 5.3 summarize 10000 replications of this simulation. The solid vertical lines mark the empirical mean of each sampling distribution, and the dashed vertical lines mark the true relative frequency of the missing cell, which is 1/7 according to the saturated hierarchical log-linear model. As desired, the sampling distribution of the estimate using Gaussian weights (b) is similar to the distribution of the traditional estimates

which implicitly use uniform weights as in a boxcar kernel (a). Although this is only a special case, the result is consistent with the notion that removing the boxcar assumption (Section 5.3.1) does not cause substantial bias.

The estimates in both (a) and (b) are notably biased upwards. This is related to the small effective sample sizes in conjunction with use of the most complex model. In the third panel (c), we examine the performance of the adjusted saturated model (3.10). The reduced bias is encouraging, although much more work is needed to fully understand the properties of that proposed adjustment.

5.7 Alternative Weighting Schemes

There are many ways to estimate the columns of Π, i.e., the functions $\pi(y|x)$ as functions of x. We have discussed only estimators of the form (5.4), which are local averages with non-negative weights, because these averages are particularly convenient for the subsequent discussion of the likelihood function and local model selection criterion. In principle, it may be possible to use relatively sophisticated methods to estimate Π in conjunction with local log-linear models.

One such advanced method is the additive multinomial logit model employed by Zwane and van der Heijden (2004), which used vector splines to model the effects of continuous variables. Another is the nonparametric conditional density estimator by Hall et al. (2004), which we briefly describe here. Let $m_i = 1$ if the ith unit is captured at least once, and $m_i = 0$ otherwise. Assume that each m_i is the outcome of a Bernoulli variable M_i.

From (5.1), we have $\pi(y|x_{i\cdot}) = P(Y_{i\cdot} = y|M_i = 1, x_{i\cdot})$ for each $y \neq \mathbf{0}$. Suppose that each vector $x_{i\cdot}$ is a realization of some random variable X. Let $f_M(x_{i\cdot}) := P(X = x_{i\cdot}|M_i = 1)$ and $g_M(y, x) := \pi(y|x)f_M(x)$. Then,

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

One can estimate g_M and f_M directly from the observable data (i.e., units with $m_i = 1$), and the conditional density of any nonzero capture pattern y given $X = x$ is

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

The `np` package (Hayfield and Racine, 2008) in the `R` statistical software (R Core Team, 2012) implements this approach. Note that although a nonparametric estimator of g_M uses smoothing parameters for both x and for the multinomial outcome y, we recommend setting the y bandwidth to zero (no smoothing), since the local log-linear model effectively smooths over y.

Whenever a multinomial regression method can be written in the form $\widehat{\Pi} = HA$, where H is an $n_c \times n_c$ projection matrix, also known as the "hat" matrix, one can read off the weights w^i for the ith unit from the ith row of H so that projection by H is equivalent to (5.4). These weights are often negative, with absolute values that do not sum to unity. Thus, building local log-linear models on top of such smoothers would require, at minimum, a slight generalization of the effective sample size 5.5.

5.8 Discussion

Local log-linear models point to several avenues of future work. First, in Section 5.3.3 we identified the desirability of a data-driven model selection criterion that simultaneously optimizes the local averaging bandwidth and the complexity of the local models. Second, the information criterion (5.10) needs a more detailed theoretical basis. In particular, one could explore alternative definitions for the effective local sample size. Third, one could improve our approach by using the variations and refinements to traditional log-linear models that have been suggested by Cormack (1989), Darroch et al. (1993), and Rivest and Lévesque (2001). Fourth, it is straightforward to apply local log-linear model averaging (Section 3.4.3), and Section 6.2 does exactly this.

Our simulations to compare local log-linear models against the additive multinomial logit model suggests that the two approaches are similar, with each model performing better than the other in special circumstances. Local models may have unique ability to accurately estimate rates of missingness in a large and diverse population, such as the human population of a nation, in which the basic relationships between lists may vary across age and socioeconomic group. Reliably estimating a large set of unique local models obviously requires large sample sizes. However, even when the sample size is large, there may be scientific reasons to believe that a single log-linear relationship should hold across all strata, and in this case we expect the additive multinomial logit to be superior.

Computing local averages of the form (5.4) and subsequently selecting a

local log-linear model at each observed unit is computationally demanding. We discuss this problem in the section on multiple sclerosis data in the next chapter.

Chapter 6

Applications

6.1 Bird Species Richness

We estimate the number of bird species using the North American Breeding Bird Survey for continental North America north of Mexico (Sauer et al., 2011). The purpose of this section is to illustrate; we take the liberty to make several modeling choices that lead to pedagogically useful behavior while potentially compromising the validity of our inference.

Table 6.1: Cross-classification of species observed over three years

		In 2011	Not in 2011
In 2010	In 2009	581	13
	Not in 2009	10	10
Not in 2010	In 2009	11	18
	Not in 2009	21	c_0

Table 6.1 displays **c**, the cross-classification of species observed in the years 2009 - 2011, treating each year as a separate list. For example, exactly

581 species were observed in all three years, and 18 species were observed only in 2009. Define a covariate x as the reverse of the rank ordering of the observed species based on the total number of times that each species was observed. For example, the species that was observed most often over the three years has covariate $x = 664$, as 664 distinct species were observed. The obvious interpretation of x is that species with a high value of x are easy to observe. Compared to covariates used previously to model heterogeneity in the detectability of birds, such as wingspan, our covariate appears to be a relatively direct proxy for species detectability.

We estimate the $\pi(\vec{0}|x_i)$ only for $i = 1, ..., 150$, corresponding to the 150 least-observed species, since the species that are difficult to observe are the only ones for which significant numbers of species can have gone missing. Weights from a Gaussian kernel define the local averages (5.4). We set the bandwidth at 45, but increase it near the boundary such that the number of nonzero weights is constant across units. Figure 6.1 shows the local averages $\{\hat{\pi}(y|x)\}_{y \neq 0}$ in a stacked form as seven smooth functions of x.

We minimize the information criterion (5.10) to select a local log-linear model for each observed species. We plot the resulting log-linear estimates for $\pi(\vec{0}|x)$ as individual points above the horizontal line at 1 in Figure 6.1. These estimates appear to follow a curve that has discontinuities at the points at which a different model is selected. It is trivial to replace the point estimates with model averages as in Section 3.4.3 to smooth the discontinuities in Figure 6.1.

We interpret each estimate $\hat{\pi}(\vec{0}|x_i)$ as a rate of missingness. For example, at $x = 1$, corresponding to the least frequently observed species, the

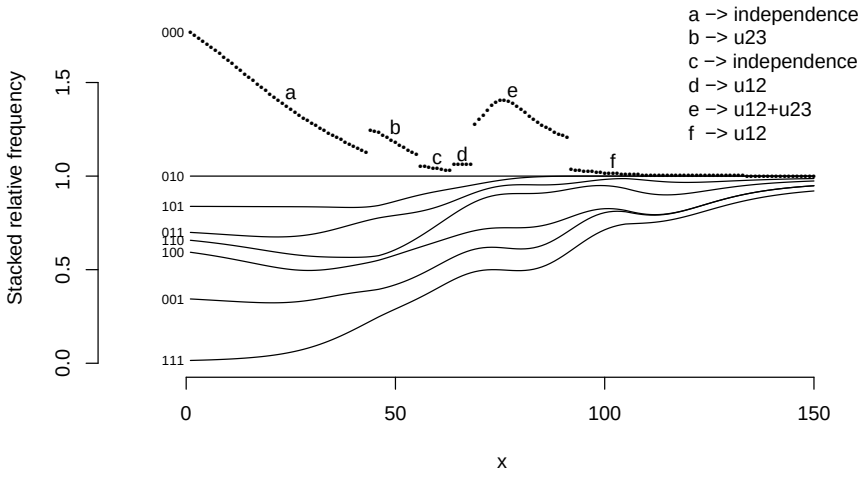

Figure 6.1: Our proxy for species detectibility, x, is on the horizontal axis. The relative frequencies of the capture patterns (i.e., "111", "001", ...) are plotted as functions of x in a stacked form. For example, the curve labeled "001" represents the sum $\pi\{111|x\} + \pi\{001|x\}$. The relative frequencies of observable capture patterns sum to 1, the horizontal line, labeled "010". Above this horizontal line, the estimates $\hat{\pi}(\vec{0}|x_i)$ are plotted as $\hat{\pi}(\vec{0}|x_i) + 1$ $(i = 1, ..., 150)$. Each continuous section of that curve corresponds to a specific log-linear model, as indicated by the labels *a* - *f*. The "independence" model is (3.6), and "u23" is shorthand for the log-linear model that includes the interaction coefficient u_{23} in addition to the main effects.

distance between the uppermost point (labelled "000") and the horizontal line below it is approximately 0.7, indicating an estimated rate of missingness $\hat{\pi}(\vec{0}|x = 1) \approx 0.7$. For all units with $x \geq 100$, the rate of missingness is nearly zero.

Applying (5.2) gives $\hat{c}_0 = 28.8$, or $\hat{n} = 692.8$. The bootstrapped 95% confidence interval for $\hat{c}_0$, based on 500 replications is $(12.4, 30986)$. Two

issues contribute to the large right tail of the confidence interval. The first is numerical instability: When the resampled data in the bootstrap leads to a zero in some row of the local average $\widehat{\Pi}$, the MLE for the corresponding local log-linear model does not always exist. (A data adjustment as in Section 3.5 may be a quick fix.) The second issue is a point made by Alho (1990): the Horvitz-Thompson sum is unstable when the detection probability $\psi(x)$ approaches 0, even if $\psi(x)$ were known. Indeed, the detection probabilities in the left tail of the distribution of x in Figure 6.1 may be too low to estimate accurately. We suspect that the latter issue deserves more attention in many capture-recapture studies, including previous studies using Breeding Bird Survey data such as Boulinier et al. (1998) and Dorazio and Royle (2003).

The Breeding Bird Survey data exists for many years prior to 2009, and so our use of only three years of data raises an obvious question: Why not extend the model to incorporate all available years? However, the assumption of a closed population may fail over long spans of time, as certain species go extinct, and new species evolve or change their geographic region of preference. Population size estimation on a 3-year moving window could, in principle, reveal changes in species richness over time. A separate consideration is that not using data earlier than 2009 allows us to use the previous years of data as a partial validation of our method. The data from 1997 to 2008 includes 40 additional species, which is on a similar order of magnitude as our point estimate.

6.2 Prevalence of Multiple Sclerosis in France

6.2.1 Background

El Adssi et al. (2012), hereafter referred to simply as "El Adssi", estimated the prevalence of multiple sclerosis (MS) in the Lorraine region of France by analyzing three separate lists of subjects that were believed to suffer from MS. We briefly review their analysis before applying local log-linear models. Table 6.2 displays the number of people with each capture pattern. Here, "LR" stands for the Lorraine registry of MS, which is the largest list of subjects. "RHIS" refers to records from the Regional Health Insurance System, the second-largest source of subjects. Finally, "MRD" refers to the aggregated results of contacting the medical records departments in 119 hospitals, identifying 64 subjects that were not already identified on at least one other source. Altogether, the data contains records on 4001 distinct subjects.

Table 6.2: Cross-classification of subjects by list membership

		In LR	Not in LR
In RHIS	In MRD	474	42
	Not in MRD	1343	199
Not in RHIS	In MRD	393	64
	Not in MRD	1486	$c_{\mathbf{0}}$

El Adssi fitted every hierarchical log-linear model that includes all main effects to the data in Table 6.2, and chose the saturated model based on the [uncorrected] AIC. The resulting point estimate of the number of undetected MS cases is 404.7, and a 95% confidence interval of (260.5, 628.7)

follows from the assumption that the estimated intercept term $\hat{u}_0$ has a $N(u_0, se(\hat{u}_0)^2)$ sampling distribution. Alternatively, we obtain the slightly wider interval $(195.0, 645.0)$ by the nonparametric bootstrap procedure of Norris and Pollock (1996b) (see also Section 4.5.2 for a detailed example of this bootstrap).

6.2.2 Applying Local Log-linear Models

We use local log-linear models to reanalyze the MS data, taking advantage of the several variables: Age, sex, and zip code. All of our local log-linear methods in this section incorporate the "EB" adjustment as in Section 3.5 and use the AICc (5.10) for local model selection.

Age follows a bell-shaped distribution with a peak near 50 years and a minimum and maximum of 16 and 89 years, respectively. The zip code has four possible values. The first column of Table 6.3 contains the counts of individuals for each combination of sex and zip code, aggregated over age. For example, 1174 females were observed with zip code 57.

Fix x as one of the observed covariates vectors. To define the weights for the rows of the local average $\widehat{\Pi}$ that correspond to x, we proceed as follows:

1. Choose a smoothing parameter λ in the interval $(0, 1]$, which represents the approximate fraction of the data that will be included in the support of a kernel smoother.

2. Define a metric on the covariate space. We use a kind of Euclidean distance, where differences in age are scaled by 0.2, while differences in the categorical variables, sex and zip code, are scaled by $\sqrt{2}$. For

example, if $x' = (\text{age, sex, zip}) = (35, 1, 57)$ and $x'' = (50, 0, 88)$, then the square of the distance between x' and x'' is

$$(0.2 * 15)^2 + (\sqrt{2} * 1)^2 + (\sqrt{2} * 1)^2.$$

The scaling factors $(0.2, \sqrt{2}, \sqrt{2})$ are subjective, reflecting the author's prior belief about the relative usefulness of the variables for explaining heterogeneity of capture probabilities. To encode the belief that age is not a useful predictor, one might set the scale factor for age to be much smaller than 0.2.

3. Identify the minimum distance, say $md(x)$, for which the set of units that lie no further than $md(x)$ away from x makes up at least $100 \times \lambda$ percent of the data. Let $D(x)$ denote the set of units that lie in the closed ball of radius $md(x)$ centered at x.

4. For all units in $D(x)$, scale the set of distances by $1/(md(x)\sqrt{1.01})$, and compute the Epanechnikov kernel weights as $(1 - \text{distance}^2)$. For all units not in $D(x)$, define the corresponding weights to be zero. Finally, normalize the full vector of weights corresponding to x. The $\sqrt{1.01}$ in the denominator of the previous expression is an ad hoc way to give nonzero weight to units on the boundary of the support $D(x)$ of the kernel.

Let lll.yy denote our local log-linear method with the smoothing parameter λ set to $0.yy$. For example, the "lll.20" method uses $\lambda = 0.20$. For several values of λ, Table 6.3 shows the average estimated rates of missingness for

units in each of the sex/zip code categories. The lll.20.av method implements model averaging as in Section 3.4.3 locally.

Table 6.3: Percentage rates of missingness by sex and zip code for various estimation methods. Rates are averaged across age within each class. The reference categories are sex = female and zip code = 57. The leftmost column shows the number of observed units in each class. The rightmost column shows the rates of missingness as estimated by the additive multinomial logit (AML) model. The columns beginning with "lll" show the rates of missingness for various local log-linear models.

count	male	zip54	zip88	zip55	lll.20	lll.20.av	lll.30	lll.15	lll.05	AML
394	1	1	0	0	5.2	4.9	5.4	5.2	4.5	3.2
1043	0	1	0	0	5.6	5.7	6.0	5.4	4.3	3.2
78	1	0	0	1	5.9	5.6	5.8	5.9	4.7	3.1
156	1	0	1	0	6.3	5.8	5.9	6.3	4.9	4.7
427	0	0	1	0	6.7	6.7	6.9	6.3	4.4	4.9
505	1	0	0	0	6.7	7.1	6.7	7.1	6.9	10.0
219	0	0	0	1	8.3	7.6	7.7	9.4	4.8	2.9
1174	0	0	0	0	8.7	8.3	8.5	8.7	8.5	9.2

Fitting a local log-linear model for each of the 4001 population units is computationally demanding. When fitting n models, the time to fit the models is only of order n, but the time required for our [admittedly primative] method of generating the underlying local modeling weights, described above in this section, is of order $n^2 \log n$. Variance estimation following the method in Section 5.4 further multiplies the required computation time by the number of bootstrap replications.

To reduce computation time, we simplify the covariate space by rounding ages to the nearest whole number. The result is that there are only 473 distinct points in the covariate space (age×sex×zip, with some cells empty), and so it suffices to fit only 473 distinct models, a task which takes less than a minute (or several hours, including replications for variance estimation)

on a typical laptop computer manufactured circa the year 2010.

The question of choosing the optimal smoothing parameter λ is perhaps more complicated than it may first appear, as we discussed near the end of Section 5.3.3. Based on the simulation studies illustrated in Figures 3.2 and 3.6, we want to choose λ to be large enough so that the effective sample sizes at each of the 473 models typically falls between 400 and 800. Combined with this criterion, Figure 6.2 suggests that $\lambda = 0.2$ is a reasonable choice. Tables 6.3 and 6.4 suggest that the estimated rates of missingness are not extremely sensitive to the choice of smoothing parameter.

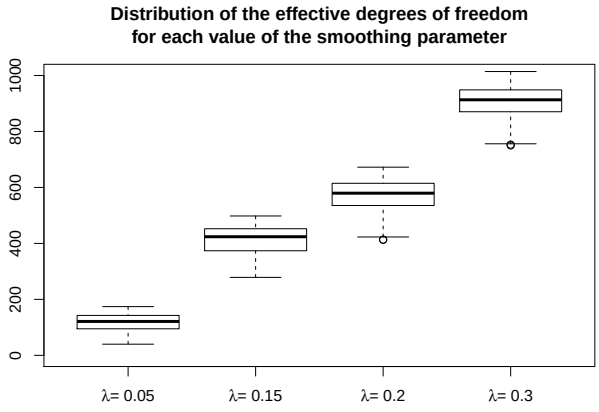

Figure 6.2: The distribution of effective sample size for several values of the smoothing parameter λ.

The model lll.05, with the extremely small smoothing parameter $\lambda = 0.05$, illustrates the consequences of choosing a λ that is almost certainly too small. Table 6.4 shows that the upper end of the 95% confidence interval for lll.05 is lower than for the other local log-linear models. This is potentially counterintuitive because undersmoothing tends to produce wider

confidence intervals in a typical nonparametric regression.

Table 6.4

Model	$\hat{c}_{\mathbf{0}}$	95% confidence interval
AML	241	(209, 281)
lll.20	276	(204, 486)
lll.20.av	271	(204, 465)
lll.30	277	(190, 496)
lll.15	276	(206, 518)
lll.05	238	(204, 413)
basic.AICc	410	(207, 631)
basic.BICpi	206	(183, 618)
basic.BIC	206	(176, 628)

Table 6.5, which shows the frequency of each local log-linear model at each of several values of λ, provides insight into the narrowness of the confidence interval corresponding to model lll.05. Specifically, the AICc selects the independence model, the most parsimonious model, relatively often when λ is small. Naturally, the AICc tends to disfavor complex models when the effective sample size is extremely small, regardless of the complexity of the "true" model.

6.2.3 Comparison with Other Methods

In addition to the local log-linear models ("lll.xx") and additive multinomial logistic (AML) model referenced in Table 6.3, we used the information criteria AICc, BICπ, and BIC to select ordinary log-linear models ("basic.xIC") globally, i.e., without including any of the covariates. Table 6.4 shows point estimates of $c_{\mathbf{0}}$ and bootstrapped 95% confidence intervals for each model. Our local log-linear estimates of $c_{\mathbf{0}}$ are substantially more conservative than

Table 6.5: Frequency table for sets of interaction terms appearing in the lll.30 , lll.20, and lll.05 local log-linear models for the multiple sclerosis data.

	lll.30	lll.20	lll.05
independence	905	945	1906
u_{23}	1442	1539	1213
u_{12}	203	342	289
u_{13}	165	233	345
$u_{13} + u_{23}$	490	481	124
$u_{12} + u_{23}$	304	227	87
$u_{12} + u_{13}$	325	176	12
$u_{12} + u_{13} + u_{23}$	162	53	20

El Adssi's results (the basic.AICc model, in essence).

The AML model stands out in Table 6.4 and in Figure 6.3 for its exceptionally narrow confidence bands. This is enticing for the researcher in search of certainty. Unfortunately, the narrowness may reflect a lack of flexibility, since the AML model uses only a single set of log-linear interaction terms for the entire covariate space. The AICc selected only the interaction term u_{23} in addition to main effects for the AML model, while local-log linear models can vary. Interestingly, the u_{23} model is also the most frequently selected local log-linear model (see Table 6.5).

A second – and likely more important – reason for the narrowness of the AML confidence interval is that we created it using the parametric bootstrap of Zwane and van der Heijden (2003). The confidence intervals for all of the other models in Figure 6.3 rely on the nonparametric bootstrap in Section 5.4. The parametric bootstrap assumes that the chosen log-linear model is the correct one, leading to narrower confidence intervals.

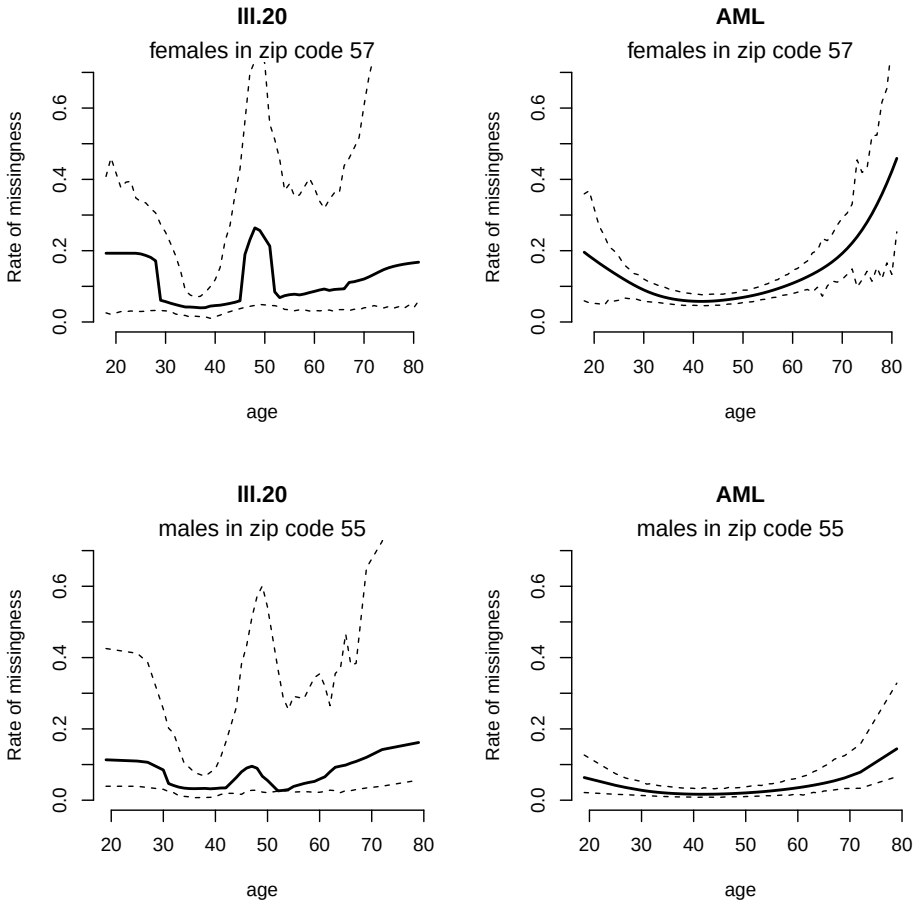

Figure 6.3: The top panels describe the set of females with MS in zip code 57, and the bottom panels describe the set of males with MS in zip code 55. In each panel, the solid curve represents the estimated rate of missingness as a function of age. The dashed curves mark the edges of a 95% confidence interval obtained by the bootstrap. The estimates in the left-side panels come from local log-linear models with smoothing parameter $\lambda = 0.20$, while the right-side panels use the additive multinomial logit (AML) model.

The curves in Figure 6.3 are smoother for the AML model (right-side panels) than for log-linear models (left-side panels). One can (and should)

easily improve both the smoothness and the local root-mean-square error for the local log-linear models by using model averaging as in Section 3.4.3.

In Chapter 4, we emphasized the importance of including contextual information when putting weights on the various models. The present analysis does less of this than we would like, as we were not directly involved in the collection of the data. However, it is interesting to speculate on possible reasons for the potentially high rates of missingness for subjects in the age range 40-55 as suggested in our local log-linear analysis in Figure 6.3.

The incidence (appearance of new cases) of MS depends on age. Suppose that x denotes a specific age for which the incidence rate is elevated. Next, suppose that there is, on average, a t-year time lag before incident cases of MS are diagnosed. In such a scenario, the rate of missingness in the age interval $(x, x + t)$ could be also elevated. Thus, it is plausible that the highest rates of missingness should approximately coincide with the highest incidence rates, with perhaps a delay of between 0 and t years. The second panel of Figure 1 in Ligouri et al. (2000) shows incidence rates as a function of age based on data from Italy. The incidence appears to peak around the ages of 25-30 years, at least 10 years ahead of the midlife peak in the rate of missingness that is suggested in Figure 6.3. We find it unlikely that the time until diagnosis is typically as large as 10 years, and so our speculative theory does not seem to be valid, unless the average age of MS incidence in France is several greater than the average for Italy.

6.3 Census Coverage

Assessing the accuracy – or *coverage* – of the U.S. Census (hereafter *the Census*) is one of largest and most significant applications of CRC in terms of the size of the target population (over 300 million U.S. residents), the complexity of the estimation problem, and the amount of research funding. Relevant data were not available to us, so this section only identifies ways in which one could apply local log-linear models to the coverage estimation problem.

The U.S. Census Bureau used CRC theory in formal census coverage evaluations after every decennial census starting no later than 1980. The name of these formal evaluations changed with each new census; for the 2010 Census, the evaluation was called the Census Coverage Measurement (CCM). Several sources of error affect the accuracy of a census. We categorize most of these error sources as a contribution to either an *overcount* or an *undercount*. An example of an overcount is when the census counts a college student both at college and at the parents' home. On the other hand, an undercount occurs when the census misses people. The CCM estiimated both the overcount and the undercount rates of around 5% for 2010, and the CCM concluded that the 2010 census was in error by less than 0.1% overall, although the error rate was higher for specific demographic subgroups (Mule, 2012).

Census data are organized into the following hierarchy: Person, Housing Unit, Census Block, Block Group, Census Tract, County, State, Division, Region, Nation. Sampling of persons for the CCM is complex, but can be

understood approximately as a geographic cluster sample using clusters of Census blocks. Census blocks are the smallest geographic unit in the hierarchy, corresponding approximately with physical city blocks.

6.3.1 The Census/P-sample Dual System

The CCM used two lists – or a dual system – to generate an estimate of the population size. The two lists are called the *Census* and the *P-sample*. The Census is the collection of all Census enumerations, as you might expect from its name. The P-sample is an independent census on a small region. The P-sample region is a random sample of housing segments, or geographically connected groups of houses, within selected census blocks (Hogan, 2003).

Separate from the Census and P-sample, the *E-sample* is a geographically clustered subset of Census enumerations that coverage evaluations use to estimate certain error rates associated with the Census and P-sample. In the context of duals system estimation, some authors, such as Chen et al. (2010), unfortunately do not clearly distinguish the roles of the Census and E-sample. In particular, the E-sample is not one of the two lists that comprise the "dual system".

The CCM matches individuals from the P-sample to the Census enumerations. Let N_P denote the size of the P-sample and let N_{CP} denote the number of P-sample individuals that are matched to a Census enumeration. Nationally, the CCM estimates the rate of undercount as $\hat{r}_u = 1 - N_{CP}/N_P$. This rate can also be computed *locally*. If x is a covariate vector that defines a post-stratum, then $\hat{r}_u(x) = 1 - N_{CP}(x)/N_P(x)$ is a post-stratum-specific estimate of the undercount rate. If the Census and P-sample are indepen-

dent, then $1 - \hat{r}_u(x_i)$ is a good estimator of the detection function $\psi(x_i)$ for the ith recorded individual, and the CCM constructs the national population size estimate as an HT estimate (2.1).

To minimize the bias of $\hat{r}_u$, the CCM designs the P-sample to be independent from the Census. Within the block clusters that are selected for the P-sample, 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 Census 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 Census and the P-sample. Waiting too long increases the effects of an open population, as people migrate, reproduce, and die between the two samples. The 2010 Census targeted April 1 as the Census 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. The 2010 CCM was the first formal coverage evaluation to apply logistic regression as inspired by Alho (1990). Olson and Griffin (2012) describes how the CCM applied logistic regression. In an independent analysis, Chen et al. (2010) applied kernel regression as "local post-stratification" to estimate the coverage of the 2000 census.

6.3.2 Local Log-linear Models for Census Coverage

Local log-linear models are potentially relevant for measuring Census coverage if three or more lists are available that all have auxiliary covariates such as age and sex. Taking the Census as the main list, there are many possible candidates for a second and third list. These include the American Community Survey (ACS), records from the Internal Revenue Service (IRS), and drivers license records for each state. Auxiliary covariates that are typically available with such lists include race, age, sex, and geographic location.

Several practical challenges stand in the way of the "easy" application of local log-linear models for Census coverage estimation:

1. The P-sample, traditionally used as the second list for dual system estimation, employs a sampling frame without clear geographic boundaries. Although census blocks are clearly defined geographically, the P-sample involves sub-sampling within census blocks. The sub-sampling scheme is a function of the P-sample housing list, and does not rely on explicit geographic boundaries. The result is that it is difficult (if not impossible) to characterize the geographic region of overlap between the Census and P-sample. Thus, there is no obvious way to count (or even to estimate) the number c_{10} in Table 1.1, i.e., the number of persons captured in the Census but missed by the P-sample within the P-sample target region.

2. Apart from the Census itself, no available list represents a credible attempt to survey the entire nation. The ACS covers only a small fraction of the population. IRS records omit much of the population

that does not earn enough income to file taxes. Drivers license records omit most of the population that does not drive, and are not currently available in a centralized database for all states. The P-sample includes only about 0.1% of the population.

3. The various lists are collected across a span of time, introducing open-population effects. The matching between the Census and P-sample dual system estimates involves an elaborate set of criteria to keep track of people who move between the Census date and the P-sample person interviews. Implementing comparable criteria at a national level with a third list is sure to be difficult.

4. Record linkage across three or more lists is a hard problem (see Chapter 7), and especially hard on the scale of the Census, where there are more than 100 million records. (In dual system estimation, it is desirable to match every P-sample record against the entire set of about 300 million Census enumerations. However, in practice, the P-sample records are matched against only a subset of Census enumerations where there is prior reason to believe that a match could exist. Restricting the matching process in this way reduces the computational burden of the matching as well as reducing the potential for "false positive" linkage errors.)

Despite these obvious challenges, the experimental use of a third list in Census coverage evaluation has precedent. The Census Bureau conducted a "dress-rehearsal" study in St. Louis in 1988 to prepare for the 1990 Census coverage evaluation. This study supplemented the P-sample with the

A-sample, a compilation of records based on Employment Security, driver's license, Internal Revenue Service, Selective Service, and Veteran's Administration registrants (Wolfgang, 1989). Several authors throughout the 1990's explored triple system estimation, viewing the A-sample as a list in its own right (Zaslavsky and Wolfgang, 1993; Darroch et al., 1993; Chao and Tsay, 1998).

None of the studies mentioned above used auxiliary covariates in any way more sophisticated than post-stratification. Supposing that a coherent triple system can be assembled with (a) minimal open-population activity between samplings, (b) minimal record linkage error, and (c) high-quality auxiliary covariates exist for all individuals that are included on each list, it should then be reasonably straightforward to apply local-log linear models as illustrated on the multiple sclerosis dataset in Section 6.2.

As Chapter 4 emphasized, CRC models are most reliable when grounded in a detailed knowledge of the data sources and auxiliary information. Chapter 10 in Alho and Spencer (2005) analyzes census data while giving quite a lot of attention to these practical matters. For instance, they adjust the dual-system estimates when the ratio of males to females differs substantially from prior beliefs about this ratio based on alternative data sources.

6.3.3 Political Context of Census Estimation

The quality of the Census is political charged because the allocation of governmental resources to various geopolitical bodies (such as states, counties, and municipalities) often depends directly on population estimates. In addition, the constitution requires the allocation of Congressional seats to

reflect Census counts. The overall fraction of people missed in the Census, the *undercount rate*, is of some importance, but the crucial question of how Census errors may affect the distribution of power across the nation rests on the *differential* undercount, or differences in the undercount rate across various demographic groups. The CCM typically estimates the undercount rate to be highest among certain minority groups, contributing to real or perceived injustices.

The Census Bureau studied the issue of differential undercount as early as the 1940's, and gradually acquired the knowledge and infrastructure for a careful Census coverage evaluation. Leading up the 1990 Census, plans were in place to statistically adjust the official 1990 Census count, but the Secretary of Commerce prevented the adjustments from being used for legislative redistricting. A coalition of political groups including New York City filed a lawsuit to compel the Census Bureau to use its statistical adjustments. In 1996 the Supreme Court ruled that the original decision of the Commerce Department to exclude the adjustments was not illegal. Thus, the official 1990 Census counts did not include the adjustments. Similar legal battles preceded the 2000 Census, and in 1999 the Supreme Court ruled that using statistical adjustments for legislative redistricting is unconstitutional. Despite restrictions on using statistical adjustments for Congressional redistricting, the Census Bureau continues to publish statistically adjusted counts for other purposes.

The Census adjustment controversy is largely a political story, involving a great deal of miscommunication and misunderstanding woven together with bits of truth. Rather than trudge through the details, we refer the in-

terested reader to a handful of key sources which document much of the substance and tone of the debate. Freedman et al. (1993) argued that the assumptions underlying the Census coverage evaluation are too strong, rendering the coverage evaluation project more a source of confusion than a source of clarity. Wachter and Freedman (2000) discussed how a hypothetical extreme form of heterogeneity could cause bias in standard Census coverage estimation methodology. Anderson and Fienberg (2000) documented several rebuttals to these kinds of negative findings. In a terse follow-up article, Brunell (2002) accused Andersen and Fienberg of using "twisted language."

Full disclosure: The Fienberg referenced above is on the committee overseeing the present thesis.

Chapter 7

Record-linkage Error

7.1 Introduction

Estimates of population size from CRC models rely on accurate record linkage across multiple lists. Any attempt to estimate the correct record linkage structure is likely to include errors when the records contain errors. We use simulation and theory to explore how various types of record linkage errors can propagate into CRC estimates.

One source of motivation to pursue this topic comes from the Census coverage evaluation problem. A specific criticism of CRC in the Census context is that record linkage errors tend to propagate through CRC estimates. To our knowledge, no major Census study has specifically addressed the propagation of linkage error, and, in general, this problem has not been extensively studied. In personal correspondence, Richard A. Griffin of the U.S. Census Bureau reported preliminary simulation results which suggest that even modest error rates in record linkage can lead to a situation in which

triple system estimates of population size are less accurate than dual system estimates.

Record linkage errors come in two kinds: false matches and false nonmatches. A false nonmatch is the absence of a link between two records that refer to the same population unit. A false match a link between two records that do not refer to the same population unit. Both false matches and false nonmatches may occur when matching records across lists or when deduplicating records within a list.

The classic Petersen formula estimates few unobserved records if the overlap between lists is large and estimates many unobserved records if the overlap between lists is small. This basic intuition of how the population estimate must depend on the overlap between lists is relevant for all CRC models and for any number of lists. The record linkage completely determines the relative sizes of the various patterns of overlap. This motivates us to investigate how different kinds of record linkage error influence estimates of population size.

The field of record linkage is large. Fellegi and Sunter (1969) provided one of the first probabilistic theories for record linkage. Herzog et al. (2010) discuss practical aspects of the Fellegi-Sunter approach in modern applications. Sadinle and Fienberg (2012) recently extended two-list record linkage to a joint model for more than two lists. Previous work at the intersection of CRC and record linkage includes a Bayesian CRC model that incorporates record linkage uncertainty for two lists (Tancredi and Liseo, 2011).

7.2 Foundations

7.2.1 Truth

The notation in this chapter differs slightly from the other chapters in this thesis. Suppose k lists of records $L_1, ..., L_k$ come from some population of unknown size n. Let m_j denote the number of records on list L_j. Let r_{ji} denote the ith record on the jth list, so that $L_j = \{r_{j1}, ..., r_{jm_j}\}$. In a typical application with a population of people, a specific record r_{ji} may consist of a vector of name, age, and date of birth, for example.

Index the population units by the numbers $1, ..., n$ so that "the ith unit" has a clear physical meaning. Assume that every record is generated in a coherent way, in the sense that each record – at least in principle – corresponds to exactly one unit of the population. Define the "truth index assignment" function $A^* : \cup_j L_j \to \{1, ..., n\}$ as an oracle function that assigns each record to the index for the unit that the record describes. Thus $A^*(r_{ji}) = t$ if the record r_{ji} "belongs to" the tth unit.

We assume that each list is *nominally* deduplicated. That is, $i_1 \neq i_2$ implies that $r_{ji_1} \neq r_{ji_2}$ $(j = 1, ..., k)$. Note, however, that distinct records may refer to the same unit; it is possible to have $r_{ji_1} \neq r_{ji_2}$ with $A^*(r_{ji_1}) = A^*(r_{ji_2})$. If the truth assignment function A^* were known, one could construct a "truth table" to show the relationship between the population and the lists. Table 7.1 is a hypothetical example of a truth table for three lists drawn from a population with only 7 units:

In Table 7.1, records r_{11}, r_{21}, and r_{13} all belong to unit 1. The record r_{15} is the only record that belongs to unit 5. Records r_{13} and r_{14} are duplicates,

Table 7.1: Example of a truth table for record linkage

Population index	L_1	L_2	L_3
1	r_{11}	r_{21}	r_{31}
2	r_{12}	r_{22}	-
3	-	r_{23}	(r_{32}, r_{33})
4	(r_{13}, r_{14})	-	-
5	r_{15}	-	-
6	-	-	-
7	-	-	-

as both belong to unit 4. Similarly, r_{32} and r_{33} are duplicates. Units 6 and 7 are not recorded on any of the lists.

7.2.2 Discrete Linkage

We define a discrete linker function, or simply a linker, as any function from the set of records $\cup_j L_j$ into the set of positive integers. The truth index assignment function is an example of a linker. Given an arbitrary discrete linker A, any two records r, r' are said to be A-*linked* (or simply *linked* if the linker is clear from the context) if $A(r) = A(r')$.

If a linker $\widehat{A}$ is an estimate of the truth index assignment function, then we call $\widehat{A}$ a linkage estimator. For our purposes, $\widehat{A}$ is a perfect estimate of A^* if the set of $\widehat{A}$-linked record pairs $\{(r, r') | r, r' \in \cup_j L_j$ with $\widehat{A}(r) = \widehat{A}(r')\}$ is equal to the set of A^*-linked record pairs $\{(r, r') | r, r' \in \cup_j L_j$ with $A^*(r) = A^*(r')\}$.

All linkers are transitive: Given any three records $r, r', r'' \in \cup_j L_j$ and a linker A, the statements $A(r) = A(r')$ and $A(r') = A(r'')$ together imply that $A(r) = A(r'')$. Transitivity has strong implications. If D is a metric on the space of records and $t > 0$ is some threshhold, the statements $D(r, r') < t$

and $D(r', r'') < t$ do not, in general, imply that $D(r, r'') < t$. Thus, transitivity requires careful attention when designing a linkage estimator.

7.2.3 Discrete Linkage Errors

The truth assignment function is typically not observable, and so we must estimate a linker. To illustrate some of the ways in which record linkage errors may arise, we give Table 7.2 as a hypothetical example of an estimated linker $\widehat{A}$ for the lists in Table 7.1.

Table 7.2: A hypothetical estimated record linkage structure

Linkage index	(Population index)	L_1	L_2	L_3
1	1	-	r_{21}	r_{31}
2	1	r_{11}	-	-
3	2,3	r_{12}	(r_{22}, r_{23})	(r_{32}, r_{33})
4	4	r_{13}	-	-
5	5	(r_{14}, r_{15})	-	-

Table 7.2 demonstrates several record linkage errors. The first error is the failure to recognize that r_{11} belongs to the same unit as r_{21} and r_{31}. This failure to link records is called a *false nonmatch*. The second error is *false match*, merging the records of unit 2 and unit 3. The final error is simultaneously a false match and a false nonmatch, as record r_{13} is unlinked from r_{14} and linked to r_{15}.

A false match or nonmatch may be caused by (a) a generically suboptimal feature in the design of the estimator $\widehat{A}$, or by (b) recording error, or by some mix of (a) and (b). These two sources of error have an interesting relationship. Even a low-quality record linkage algorithm may produce the truth table when no recording error exists. Conversely, a linker $\widehat{A}$ that is

based on an extremely coherent and well-designed linkage algorithm may contain many errors if substantial recording error exists. That is, trusting an estimated linker $\widehat{A}$ requires some degree of confidence in both the record linkage algorithm and the quality of the underlying records.

7.3 Linkage uncertainty

7.3.1 A Distribution Over Linkers

Linkage uncertainty propagates into CRC estimates through the sufficient statistics $\mathbf{c}$ of the CRC model. To be more explicit, let $\mathbf{c}(A)$ denote the result of using the linker A to construct the array $\mathbf{c}$ of capture pattern counts.

Let $\mathcal{A}(\cup_j L_j)$, or simply $\mathcal{A}$, denote the set of linkers. Let $n_L = |L_1| + \cdots + |L_k|$, the sum of the number of records on the lists. Since a linker is a partition of the set $\{1, ..., n_L\}$, the number of linkers is the Bell number of the set, an astronomical number even for small n_L.

Let $\mathcal{F}$ be some probability distribution over $\mathcal{A}$. Suppose that A is a random draw from the distribution $\mathcal{F}$ so that $\sum_{\alpha \in \mathcal{A}} P(A = \alpha) = 1$. Then the expectation of the cross-classification of capture patterns is

$$E(\mathbf{c}(A)) = \sum_{\alpha \in \mathcal{A}} P(A = \alpha) * \mathbf{c}(A).$$

The expectation is with respect the the distribution of A, and assumes that the lists $L_1, ..., L_k$ are fixed.

Let $\hat{n}(A)$ denote the estimate (7.1) of n that results from using linker A. An estimate of n that takes into account the uncertainly in record linkage

is

$$\hat{n}(\mathcal{A}) = \sum_{\alpha \in \mathcal{A}} P(A = \alpha) * \hat{n}(A).$$

The variability of this estimate depends on (a) the variability of the multinomial capture pattern counts $\mathbf{c}$ and (b) the variability in the estimate of the distribution $\mathcal{F}$ over linkers, which may be substantial if the record-generating process is noisy.

The discrete set $\mathcal{A}$ is extremely large, and difficult to handle analytically. In a working paper, Steorts et al. propose a Bayesian approach that uses a Gibbs sampler to simulate draws from $\mathcal{F}$. This sampler produces a sequence of linkers $A_1, ..., A_q$ from a posterior distribution $\widehat{\mathcal{F}}$. From these draws, one could estimate the expected capture-pattern counts $\mathbf{c}$ as

$$E(\mathbf{c}(A)) \approx \frac{1}{q} \sum_{i=1}^{q} \mathbf{c}(A_i).$$

7.3.2 Dissection of a Simple Case

Suppose there are just two lists. Let $r \in L_1$ and $r' \in L_2$. Consider the contribution of the two records to the true capture pattern counts. There are two possibilities. If the records belong to different units, then each record contributes a '1' to the count of observed units as in Table 7.3.

Table 7.3: Contribution of two records of two units to the true capture pattern counts

	In L_2	Not in L_2
In L_1	0	1
Not in L_1	1	0

Table 7.4 illustrates the result if the records belong to the same unit:

The two records are merged into one.

Table 7.4: Contribution of two records of one unit to the true capture pattern counts

	In L_2	Not in L_2
In L_1	1	0
Not in L_1	0	0

Let p denote the probability that both records belong to the same unit, i.e., $p = P(A^*(r) = A^*(r'))$. Table 7.5 shows the weighted element-wise average of the tables 7.3 and 7.4.

Table 7.5: Contribution of two records with undetermined linkage status

	In L_2	Not in L_2
In L_1	p	$1-p$
Not in L_1	$1-p$	0

An alternative way to think about Table 7.5 is in terms of the individual contribution of each record to the table. Table 7.5 is the elementwise sum of Tables 7.6 and 7.7. In the case that both records refer to the same unit, the contribution p to the $(1,1)$ cell of the table is shared between two records – hence the division by two.

Table 7.6: Expected contribution of a record from L_1

	In L_2	Not in L_2
In L_1	$p/2$	$1-p$
Not in L_1	0	0

Table 7.7: Expected contribution of a record from L_2

	In L_2	Not in L_2
In L_1	$p/2$	0
Not in L_1	$1-p$	0

7.4 Simulation

7.4.1 A CRC Model

We use the simplest CRC independence model (3.7) for a brief study on how record linkage errors may propagate into $\hat{n}$, the estimate of the population size. We use the model twice – once for simulating fake lists (in conjunction with an additional simulation step to provoke linkage errors), and once for estimating the number of units that are missing on all of the simulated lists. Let $i = 1, ..., n_c$ be some estimated linkage index as in Table 7.2. With n the true population size, assume that $n \geq n_c$ (this assumption could be violated if the estimated linkage is particularly erroneous).

The conditional maximum likelihood estimate of the parameter $\hat{u}_0$ gives an estimate of the population size as

$$\hat{n} = n_c + \pi(\vec{0}|\hat{u}_0) = n_c + \exp(\hat{u}_0). \tag{7.1}$$

CRC models of greater complexity are of interest for specific applications, but it would be difficult to simulate across the a representative set of models for even three lists. The simplicity of model (3.7) is a reasonable base case, or reference scenario.

7.4.2 Single-linkage Clustering

Single-linkage clustering provides a relatively straightforward method of record linkage:

1. Define a metric d on the space of records. The metric may or may not include special conditions to treat inter-list comparisons differently than intra-list comparisons.

2. Build the minimal spanning tree (MST) over the full set of records $\cup_j L_j$ using the metric d.

3. Declare a threshold t, and delete all edges from the MST which correspond to distances greater than t.

4. Index the connected components of the graph, and define $A_{SL(d,t)}$ to be the linker that assigns each record to the index of the connected component to which it belongs.

We refer to this algorithm as the $\mathrm{SL}(d, t)$ linker. Samuel L. Ventura briefly described the $\mathrm{SL}(d, t)$ linker in his Carnegie Mellon University doctoral thesis proposal, but this linker does not yet seem to be well known in the record linkage literature.

7.4.3 Effects of Linkage Error

We perform the following simulation for various values of the threshold t:

1. Define a population of size $n = 1000$.

2. Define covariates for each unit. The "continuous" covariate x_1 is simply twice the population index. The "categorical" covariate x_2 is a random draw from $\{0, 1\}$, i.e., a separate draw for each unit, such that approximately half of the units end up in category 0 and the rest end up in category 1.

3. Generate three lists L_1, L_2, L_3 using uniform capture probability $p = 0.6$. Let $L = \cup_j L_j$.

4. For each record $r \in L$, generate a duplicate record with probility 0.5.

5. Let $x_1(r)$ denote the value of x_1 for record r. For each record r, replace $x_1(r)$ with a draw from the normal distribution with mean $x_1(r)$ and variance 0.4^2.

6. Let $x_2(r)$ denote the value of x_2 for record r. For each record r, replace $x_2(r)$ with $1 - x_2(r)$ with probability 0.04.

7. Apply the $A_{SL(d,t)}$ linker with d taken as the Euclidean distance.

8. Compute the capture pattern counts $\hat{c} = c(A_{SL(d,t)})$, and use (7.1) to estimate the population size.

Note that the purpose of steps (4)-(6) is to introduce ambiguity that leads to errors in the record linkage step (7). The duplication introduced in step (4) creates the potential for a false nonmatch in the linked data, while the covariate errors introduced in (5) and (6) directly cause some amount of false matches and false nonmatches. Figure 7.1 shows the results, which seem to indicate that the population estimate $\hat{n}$ can be sensitive to the record linkage

approach. In terms of population size estimation, the conservative linker is the linker that errs with more false matches than with false nonmatches.

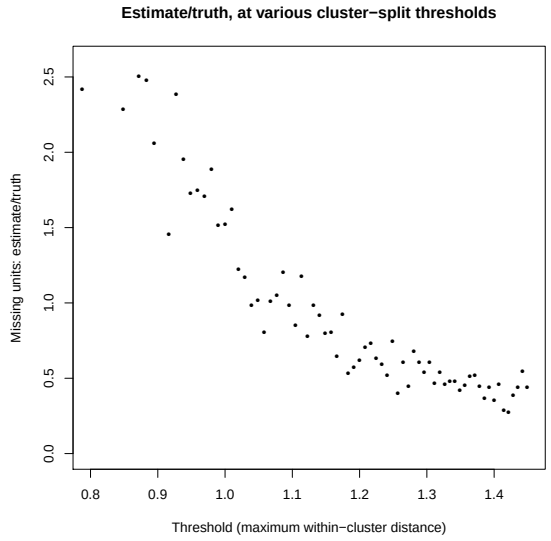

Figure 7.1: The CRC estimate of $c_{\mathbf{0}}$ divided by the true value $c_{\mathbf{0}}$ is ideally equal to one. In this simulation with three lists, it is clear that the choice of threshold t in the SL(d, t) linker underlying the capture-pattern counts $\mathbf{c}$ is of great importance. Using a low threshold leads to many false nonmatches, shrinking the overlap between lists, increasing the number of evidently distinct records on each list, and, ultimately, inflating the population size estimate. The opposite occurs when t is too large.

We conclude with several remarks. The CRC model that we selected for this section is clearly only a simple case. Similarly, the record linker we chose is only one of many possible linkage algorithms. Many more combinations exist to be tested. In particular, one could look at what happens with more than three lists. We suspect that the same basic intuition will hold in the many-list case, but it could be worthwhile to test this.

Chapter 8

Conclusion

8.1 Overview

The CRC problem is on one level a missing data problem. The missing quantity of interest is $c_{\mathbf{0}}$, the number of missing units, but, perhaps equally relevantly, the covariate values $x_{i\cdot}$ are missing for $i > n_c$. The nature of this missingness is of the worst possible kind when it is reasonable to suppose that the units which are not observed are not observed precisely because they are different from the observed units, not only in the distribution of covariates but also in how capture probabilities depend on covariates. This difficulty sets apart CRC as an exceptionally risky enterprise. Chapter 4 discussed these issues in depth, and argued that setting the highest-order interaction equal to zero a priori is potentially unwise, as this choice may gloss over biases in the sampling mechanism.

Chapter 5 used auxiliary covariates to build local log-linear models, motivated by the idea that local models are less vulnerable to heterogeneity

bias. While the use of covariates to model heterogeneity is not new, local models give a unique way to model exceptionally complex CRC data in which the form of the generating model varies over the covariate space. Covariates are useful when they explain much of the heterogeneity in capture probabilities. Unfortunately, we have no way to ensure that the observable covariates have good explanatory power; strong forms of heterogeneity may exist and simply not be significantly associated with observed covariates.

Even if the observable covariates have excellent explanatory power, the ability to extract useful information from a covariate depends on the size of the sample. The required sample size grows quickly in the number of lists. In an experiment with only two lists, only three conditional probabilities (one for each capture pattern) must be estimated over the covariate space. With k lists, we must model $2^k - 1$ conditional multinomial probabilities. When the number of cells is large relative to the local effective sample size, a standard log-linear analysis may not be appropriate, and much less a local log-linear analysis. In these situations, we suggest consideration of the relatively parsimonious model of Stoklosa and Huggins (2012), or the class of models applied by Dorazio and Royle (2003).

Regardless of whether covariates are included, an important theme emerging from Alho (1990), Link (2003), and Mao (2008) is that estimating n becomes increasingly difficult when the detection probability $\psi(x)$ is close to zero for much of the covariate space. Mao suggested that units captured moderately frequently give enough information to imply the existence of a least a few unobserved units with high probability, but not enough information to deny the existence of a large set of unobserved units with capture

probabilities very close to zero (Mao, 2008).

8.2 Review of Assumptions

This thesis developed tools for CRC analyses that may work well under several strong assumptions. We paraphrase the assumptions here, with brief commentary:

1. No record linkage error: Units can be identified in the sense that they can be properly linked across lists. Chapter 7 considered the performance of a simple CRC model under the failure of this assumption. The results were consistent with the intuitive notion that false matches (nonmatches) tend to lead to under- (over-) estimates of n.

2. Independence between units: The probability that a unit is on a list does not depend on whether a different unit is on the lists. We suspect that failure of this assumption is common, leading to confidence intervals that are too narrow, while having a relatively small effect in terms of bias.

3. Multinomial sampling distribution: The capture pattern of a unit is a realization of a draw from a multinomial distribution. For this assumption, the question of correctness is potentially irrelevant for those who accept the notion that "all models are wrong, but some are useful."

4. Homogeneity: The multinomial capture pattern probabilities are constant across units, at least for units within the same post-stratum. This assumption certainly fails almost always, to varying degrees. The

effect of the assumption's failure depends importantly on whether the pattern of heterogeneity for capture probabilities on one list is correlated with the pattern of heterogeneity on another list. The Rasch model (3.22) is especially interesting when the set of capture probabilities of units on each list is positively correlated between every pair of lists.

5. Closed populations: No births, deaths, or migration of units.

Population closure is a subtle problem when the target population is not defined with perfect clarity. In the species richness example (Section 6.1), is a species that tends to live in Mexico, but that occasionally ventures north of the border, part of the target population? In the multiple sclerosis example, is the diagnosis of multiple sclerosis always a clear-cut outcome (certainly not!). When the population is not clearly defined, one must ask whether estimated rates of missingness reflect (a) units that are undetected due to random sampling or (b) units that lie in the gray area, at the fringes of the set of units that qualify to be part of the target population.

One way to address this question in future studies would be to include a covariate that indicates the certainty with which – or, degree to which – an observed unit belongs to the target population. For example, multiple sclerosis is a disease with varying sub-diagnoses and degrees of severity. Suppose a covariate were available that provided a physician's rating of the severity of each case. If this covariate were used in post-stratification, or in local log-linear modeling, one might observe a relationship between the estimated rate of missingness and the severity of cases. A high estimated rate

of missingness in a group with extremely low severity could be an indication of indecisiveness in diagnosis instead of an indication of missing cases, while a high estimated rate of missingness in high-severity group might be a more trustworthy indication of unobserved cases.

The sheer number and strength of the various assumptions is worth lengthly contemplation. In short, any CRC analysis deserves careful scrutiny, as the failure of any of the assumptions can cause substantial bias.

In addition, our development of local log-linear models rests crucially on log-linear models. Chapter 3 reviewed log-linear models and explored their usefulness through simulation. We found that log-linear models perform well much of the time, but also have the potential to be substantially biased in special cases (see the bottom panels of Figures 3.2 and 3.3), even when all of the basic assumptions (above) are satisfied. A key unknown is the frequency with which applications tend to fall into the "special" cases.

8.3 Future work

Future work may improve upon our presentation of local log-linear models in countless ways. Here are several:

- A key problem is to devise data-driven methods for selecting the best smoothing parameter for the local weights. This is a multi-dimensional problem when multiple auxiliary covariates are available. The cross-validation method of Hall et al. (2004) holds promise, but we found the existing implementation in the `np` software package (Hayfield and Racine, 2008) to be unacceptably slow for populations greater than a

couple of thousand.

- The smoothing parameter that minimizes prediction risk over observable capture-patterns does not necessarily minimize the risk involved in extrapolation to the unobservable cell under local model selection. As starting point to explore this issue, see Theorem 2.9 in Stoklosa (2012).

- Scalability of local log-linear models is a problem because the computing time grows quickly in the number of lists and the number of distinct points in the covariate space. With three lists, there are eight basic hierarchical log-linear models to choose from; with four lists, there are 113 such models, and with 5 lists, there are nearly 7000 models. To reduce the number of model selection procedures, one could emulate Loader (1999) by fitting local models only on a course grid of points, and use interpolation to derive estimates for all observations that lie between the points of the grid.

- One may incorporate into local log-linear models some of the various devices that exist for modeling open populations. Important work on open populations appears in Cormack (1964), Jolly (1965), and Seber (1965). See also the recent frequentist analysis of Pledger et al. (2010) and a Bayesian adaptation by Royle and Dorazio (2012).

A final area of future work is in building a simulation package for CRC studies. Several software packages exist for applying specific CRC estimators, but we are not aware of a package that is dedicated to simulating realistic

populations for testing arbitrary estimators. If Table 2.2 is any indication, simulation studies to date have been rather limited. An important use of simulation in CRC is to test the sensitivity of models to their strongest assumptions. The simulations described in this thesis are a starting point for such an endeavor.

Bibliography

D J Aaron, Y-F Chang, N Markovic, and R E LaPorte. Estimating the lesbian population: A capture-recapture approach. *Journal of Epidemiology and Community Health*, 57:207–209, 2003.

Damiano D. Abeni, Giovanna Brancato, and Carlo A. Perucci. Capture-recapture to estimate the size of the population with Human Immunodeficiency Virus Type 1 Infection. *Epidemiology*, 5(4):410–414, 1994.

Juha Alho and Bruce Spencer. *Statistical demography and forecasting*. Springer, 2005.

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

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.

David R Anderson and Kenneth P Burnham. Understanding information

criteria for selection among capture-recapture or ring recovery models. *Bird Study*, 46(S1):S14–S21, 1999.

DR Anderson, KP Burnham, and GC White. AIC model selection in overdispersed capture-recapture data. *Ecology*, 75(6):1780–1793, 1994.

Margo Anderson and Stephen E. Fienberg. Partisan politics at work: Sampling and the 2000 census. *PS: Political Science and Politics*, 33(4):795–799, 2000.

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

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

Maurice S Bartlett. Contingency table interactions. *Supplement to the Journal of the Royal Statistical Society*, 2(2):248–252, 1935.

Sanjib Basu and Nader Ebrahimi. Bayesian capture-recapture methods for error detection and estimation of population size: Heterogeneity and dependence. *Biometrika*, 88(1):269–279, 2001.

Thierry Boulinier, James D. Nichols, John R. Sauer, James E. Hines, and K. H. Pollock. Estimating species richness: The importance of heterogeneity in species detectability. *Ecology*, 79(3):1018–1028, 1998.

Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231, 2001.

Thomas L. Brunell. Why there is still a controversy about adjusting the census. *Political Science and Politics*, 35:85–85, 2002.

G. Bruno, G. Bargero, A. Vuolo, E. Pisu, and G. Pagano. A population-based prevalence survey of known diabetes mellitus in Northern Italy based upon multiple independent sources of ascertainment. *Diabetologia*, 35:851–856, 1992.

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.

Kenneth P Burnham and David R Anderson. *Model selection and multimodel inference: a practical information-theoretic approach*. Springer, 2002.

Kenneth P Burnham and David R Anderson. Multimodel inference understanding aIC and bIC in model selection. *Sociological methods & research*, 33(2):261–304, 2004.

Kenneth P Burnham, Gary C White, and David R Anderson. Model selection strategy in the analysis of capture-recapture data. *Biometrics*, pages 888–898, 1995.

Anne Chao. Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, 43(4):783–791, 1987.

Anne Chao. An overview of closed capture-recapture models. *Journal of Agricultural, Biological, and Environmental Statistics*, 6(2):158–175, 2001.

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.

Anne Chao, S-M Lee, and S-L Jeng. Estimating population size for capture-recapture data when capture probabilities vary by time and individual animal. *Biometrics*, 48(1):201–216, 1992.

D.G. Chapman. *Some properties of the hypergeometric distribution with applications to zoological sample censuses.* Berkeley, University of California Press, 1951.

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.

Song Xi Chen and Chris J. Lloyd. Estimation of population size from biased samples using non-parametric binary regression. *Statistica Sinica*, 12:505–518, 2002.

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, 2010.

Constance F. Citro, Daniel L. Cork, and Janet L. Norwood. *The 2000 Census: Counting Under Adversity.* National Academies Press, 2004. Panel to Review the 2000 Census.

R. M. Cormack. The statistics of capture-recapture methods. *Oceanography and Marine Biology: An Annual Review 6*, pages 455–506, 1968.

R. M. Cormack. *Models for Capture-Recapture Studies, in Statistical Ecology: Sampling Biological Populations*. International Co-operative Publishing House, Fairland, Maryland, 1979.

Richard M Cormack. Log-linear models for capture-recapture. *Biometrics*, pages 395–413, 1989.

RM Cormack. Estimates of survival from the sighting of marked animals. *Biometrika*, 51(3/4):429–438, 1964.

RM Cormack and PE Jupp. Inference for Poisson and multinomial models for capture-recapture experiments. *Biometrika*, 78(4):911–916, 1991.

Charles D Cowan and Donald Malec. Capturerecapture models when both sources have clustered observations. *Journal of the American Statistical Association*, 81(394):347–353, 1986.

DR Cox. Comment-Statistical modeling: The two cultures. *Statistical Science*, 16(3):216–217, 2001.

Noel Cressie and Paul W Holland. Characterizing the manifest probabilities of latent trait models. *Psychometrika*, 48(1):129–141, 1983.

John N. Darroch. The multiple-recapture census: I. Estimation of a closed population. *Biometrika*, 45(3/4):343–359, 1958.

John N. Darroch, Stephen E. Fienberg, Gary F. V. Glonek, and Brian W. Junker. A three-sample multiple-recapture approach to census popula-

tion estimation with heterogeneous catchability. *Journal of the American Statistical Association*, 88(423):1137–1148, 1993.

Robert M. Dorazio and J. Andrew Royle. Mixture models for estimating the size of a closed population when capture rates vary among individuals. *Biometrics*, 59:351–364, 2003.

David Draper. Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 45–97, 1995.

Haritina El Adssi, Marc Debouverie, and Francis Guillemin. Estimating the prevalence and incidence of multiple sclerosis in the Lorraine region, France, by the capture-recapture method. *Multiple Sclerosis Journal*, 18 (9):1244–1250, 2012.

C. Lewden et. al. Number of deaths among HIV-infected adults in France in 2000, three-source capture-recapture estimation. *Epidemiology and Infection*, 134(6):1345–1352, 2006.

Marc A. Evans and Douglas G. Bonett. Bias reduction for multiple-recapture estimators of closed population size. *Biometrics*, 50(2):388–395, 1994.

Marc A. Evans, Douglas G. Bonett, and Lyman L. McDonald. A general theory for modeling capture-recapture data from a closed population. *Biometrics*, 50(2):396–405, 1994.

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

William Feller. *An Introduction to Probability Theory and Its Applications*, volume I. Wiley, January 1968.

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

Stephen E. Fienberg. Bibliography on capture-recapture modelling with application to census undercount adjustment. *Survey Methodology*, 18(1): 143–154, 1992.

Stephen E Fienberg and Michael M Meyer. Loglinear models and categorical data analysis with psychometric and econometric applications. *Journal of Econometrics*, 22(1):191–214, 1983.

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.

David A. Freedman, Kenneth W. Wachter, Daniel C. Coster, D. Richard Cutler, and Stephen P. Klein. Adjusting the census of 1990: The smoothing model. *Evaluation Review*, 17(4):371–443, 1993.

Paul Gustafson. On model expansion, model contraction, identifiability and prior information: two illustrative scenarios involving mismeasured variables. *Statistical Science*, 20(2):111–140, 2005.

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.

Tristen Hayfield and Jeffrey S. Racine. Nonparametric econometrics: The `np` package. *Journal of Statistical Software*, 27(5), 2008. URL `http://www.jstatsoft.org/v27/i05/`.

Thomas H Herzog, Fritz Scheuren, and William E Winkler. Record linkage. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(5):535–543, 2010.

Howard Hogan. The accuracy and coverage evaluation: Theory and design. *Survey Methodology*, 29(2):129–138, 2003.

Hajo Holzmann, Axel Munk, and Walter Zucchini. On identifiability in capture-recapture models. *Biometrics*, 62:934–939, 2006.

Emest B Hook and Ronald R Regal. Validity of methods for model selection, weighting for model uncertainty, and small sample adjustment in capture-recapture estimation. *American Journal of Epidemiology*, 145(12):1138–1144, 1997.

Gerald V. Howard. *A Study of the Tagging Method in the Enumeration of Sockeye Salmon Populations*. International Pacific Salmon Fisheries Commission, 1948. Bulletin II.

R. Huggins and W.H. Hwang. Non-parametric estimation of population size from capture-recapture data when the capture probability depends

on a covariate. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 56(4):429–443, 2007.

Richard Huggins. On the statistical analysis of capture experiments. *Biometrika*, 76(1):133–140, 1989.

Richard Huggins. Semiparametric estimation of animal abundance using capture-recapture data from open populations. *Biometrics*, 62:684–690, 2006.

Richard Huggins and Wen-Han Hwang. A review of the use of conditional likelihood in capture-recapture experiments. *International Statistical Review*, 79(3):385400, 2011.

Clifford M Hurvich and Chih-Ling Tsai. Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307, 1989.

Wen-Han Hwang and Richard Huggins. Application of semiparametric regression models in the analysis of capture-recapture experiments. *Australian & New Zealand Journal of Statistics*, 49(2):403–421, 2007.

Wen-Han Hwang and Richard Huggins. A semiparametric model for a functional behavioural response to capture in capture-recapture experiments. *Australian & New Zealand Journal of Statistics*, 53(4):191–202, 2011.

George M Jolly. Explicit estimates from capture-recapture data with both death and immigration-stochastic model. *Biometrika*, 52(1/2):225–247, 1965.

Robert E Kass. Statistical inference: The big picture. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 26(1):1, 2011.

Alan Kelly, Conor Teljeur, and Marlen Carvalho. Prevalence of opiate use in Ireland 2006: A 3-source capture recapture study. *Small Area Health Research Unit, Department of Public Health & Primary Care, Trinity College Dublin*, 2009. Published by the Stationery Office.

Leslie Kish. *Survey Sampling*. New York: Wiley, 1965.

A.H. Leyland, M. Barnard, and N. McKegancy. The use of capture recapture methodology to estimate and describe covert populations: an application to female streetworking prostitution in glasgow. Paper presented to the International Conference on Social Science Methodology, University of Trento, Italy, 22-26 June 1992.

M Ligouri, MG Marrosu, M Pugliatti, F Giuliani, F De Robertis, E Cocco, GB Zimatore, P Livrea, and M Trojano. Age at onset in multiple sclerosis. *Neurological Sciences*, 21(2):S825–S829, 2000.

W. A. Link. Nonidentifiability of population size from capture-recapture data with heterogeneous detection probabilities. *Biometrics*, 59:1123–1130, 2003.

W. A. Link. Reply to a paper by Holzmann, Munk, and Zucchini. *Biometrics*, 59:1123–1130, 2006.

Christopher J. Lloyd. Modified martingale estimation for recapture experi-

ments with heterogeneous capture probabilities. *Biometrika*, 79:833–836, 1992.

Clive Loader. *Local regression and likelihood*, volume 47. springer New York, 1999.

David Madigan and Jeremy C York. Bayesian methods for estimation of the size of a closed population. *Biometrika*, 84(1):19–31, 1997.

Donald Malec and Jerry Maples. Small area random effects models for capture/recapture methods with applications to estimating coverage error in the U.S. decennial census. *Statistics in Medicine*, 27:4038–4056, 2008.

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

Chang Xuan Mao. Estimating population sizes for capture-recapture sampling with binomial mixtures. *Computational Statistics and Data Analysis*, 51:5211–5219, 2007.

Chang Xuan Mao. On the nonidentifiability of population sizes. *Biometrics*, 64:977–981, 2008.

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

Thomas Mule. DSSD 2010 Census Coverage Measurement Memorandum Series #2010-g-01, May 2012. U.S. Census Bureau.

Joe Murphy. Estimating the World Trade Center tower population on September 11, 2001: A capture-recapture approach. *American Journal of Public Health*, 99(1):65–67, 2009.

Andrew A Neath and Francisco J Samaniego. On the efficacy of Bayesian inference for nonidentifiable models. *The American Statistician*, 51(3): 225–232, 1997.

James L. Norris and Kenneth H. Pollock. A capture-recapture model with heterogeneity and behavioural response. *Environmental and Ecological Statistics*, 2:305–313, 1995.

James L. Norris and Kenneth H. Pollock. Nonparametric MLE under two closed capture-recapture models with heterogeneity. *Biometrics*, 52:639–649, 1996a.

James L. Norris and Kenneth H. Pollock. Including model uncertainty in estimating variances in multiple capture studies. *Environmental and Ecological Statistics*, 3:235–244, 1996b.

Eugene P. Odum and A. J. Pontin. Population density of the underground ant, lasius flavus, as determined by tagging with p32. *Ecology*, 42(1): 186–188, 1961.

Douglas Olson and Richard Griffin. DSSD 2010 Census Coverage Measurement Memorandum Series #2010-g-10, May 2012. U.S. Census Bureau.

D. L. Otis, K. P. Burnham, G. C. White, and D. R. Anderson. Statisti-

cal inference from capture data on closed animal populations. *Wildlife Monographs*, 62:1–135, 1978.

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

S. Pledger, K.H. Pollock, and J.L. Norris. Open capture–recapture models with heterogeneity: Ii. jolly–seber model. *Biometrics*, 66(3):883–890, 2010.

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

Shirley Pledger. The performance of mixture models in heterogeneous closed population capture-recapture. *Biometrics*, 61:868–876, 2005.

Shirley Pledger and Polly Phillpot. Using mixtures to model heterogeneity in ecological capture-recapture studies. *Biometrical Journal*, 50(6):1022–1034, 2008.

Kenneth H. Pollock. Building models of capture-recapture experiments. *Journal of the Royal Statistical Society*, 25(4):253–259, 1976.

Kenneth H. Pollock. Modeling capture, recapture, and removal statistics for estimation of demographic parameters for fish and wildlife populations: Past, present, and future. *Journal of the American Statistical Association*, 86(413):225–238, 1991.

Kenneth H. Pollock. Capture-recapture models. *Journal of the American Statistical Association*, 95(449):293–296, 2000.

Kenneth H. Pollock. The use of auxiliary variables in capture-recapture modelling: An overview. *Journal of Applied Statistics*, 29(1-4):85–102, 2002.

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.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. URL `http://www.R-project.org/`. ISBN 3-900051-07-0.

Georg Rasch. Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests. 1960.

Louis-Paul Rivest and Tina Lévesque. Improved log-linear model estimators of abundance in capture-recapture experiments. *Canadian Journal of Statistics*, 29(4):555–572, 2001.

J. Andrew Royle. Analysis of capture-recapture models with individual covariates using data augmentation. *Biometrics*, 65:267–274, 2009.

J Andrew Royle and Robert M Dorazio. Parameter-expanded data augmentation for bayesian analysis of capture–recapture models. *Journal of Ornithology*, 152(2):521–537, 2012.

J. Andrew Royle, Robert M. Dorazio, and William A. Link. Analysis of multinomial models with unknown index using data augmentation. *Journal of Computational and Graphical Statistics*, 16(1):67–85, 2007.

Per Runeson and Claes Wohlin. An experimental evaluation of an experience-based capture-recapture method in software code inspections. *Empirical Software Engineering*, 3:381–406, 1998.

Mauricio Sadinle and Stephen E Fienberg. A generalized fellegi–sunter framework for multiple record linkage with application to homicide record–systems. *Journal of the American Statistical Association*, (just-accepted), 2012.

Lalitha Sanathanan. Estimating the size of a multinomial population. *Annals of Mathematical Statistics*, 43(1):142–152, 1972a.

Lalitha Sanathanan. Models and estimation methods in visual scanning experiments. *Technometrics*, 14(4):813–829, 1972b.

RL Sandland and RM Cormack. Statistical inference for poisson and multinomial models for capture-recapture experiments. *Biometrika*, 71(1):27–33, 1984.

J. R. Sauer, J. E. Hines, J. E. Fallon, K. L. Pardieck, Jr. D. J. Ziolkowski, and W. A. Link. The North American Breeding Bird Survey, Results and Analysis 1966 - 2010. 2011. Version 12.07.2011 USGS Patuxent Wildlife Research Center, Laurel, MD.

Milner B. Schaefer. Estimation of size of animal populations by marking experiments. *Fishery Bulletin*, 52(69):191–203, 1951.

Z. E. Schnabel. The estimation of the total fish population of a lake. *American Mathematical Monthly*, 45:348–352, 1938.

Carl J. Schwarz and George A. F. Seber. Estimating animal abundance: Review III. *Statistical Science*, 14(4):427–456, 1999.

Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

G. A. Seber. *The Estimation of Animal Abundance and Related Parameters*. Griffin, 1982. 2nd ed.

George A. F. Seber. A review of estimating animal abundance. *Biometrics*, 42(2):267–292, 1986.

George A. F. Seber. A review of estimating animal abundance ii. *International Statistical Review*, 60(2):129–166, 1992.

George AF Seber. A note on the multiple-recapture census. *Biometrika*, 52 (1/2):249–259, 1965.

C Chandra Sekar and W Edwards Deming. On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association*, 44(245):101–115, 1949.

Scott C. Silver, Linde E. T. Ostro, Laura K. Marsh, Leonardo Maffei, Andrew J. Noss, Marcella J. Kelly, Robert B. Wallace, Humberto Gomez, and Guido Ayala. The use of camera traps for estimating jaguar *panthera onca* abundance and density using capture/recapture analysis. *Oryx*, 38 (2), 2004.

Rebecca C Steorts, Rob Hall, and Stephen E Fienberg. Bayesian parametric and nonparametric inference for multiple record linkage.

Jakub Stoklosa. *Modern Statistical Methods for the Analysis of Capture-Recapture Data*. The University of Melbourne, 2012. Doctoral thesis.

Jakub Stoklosa and Richard M. Huggins. A robust P-spline approach to closed population capture-recapture models with time dependence and heterogeneity. *Computational Statistics & Data Analysis*, 56(2):408 – 417, 2012.

Jakub Stoklosa, Wen-Han Hwang, Sheng-Hai Wu, and Richard Huggins. Heterogeneous capture-recapture models with covariates: A partial likelihood approach for closed populations. *Biometrics*, 67:1659–1665, 2011.

Andrea Tancredi and Brunero Liseo. A hierarchical bayesian approach to record linkage and population size problems. *The Annals of Applied Statistics*, 5(2B):1553–1585, 2011.

Kenneth W Wachter and David A Freedman. The fifth cell correlation bias in us census adjustment. *Evaluation Review*, 24(2):191–211, 2000.

Eric-Jan Wagenmakers and Simon Farrell. AIC model selection using Akaike weights. *Psychonomic bulletin & review*, 11(1):192–196, 2004.

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.

Thomas W Yee. The VGAM package for categorical data analysis. *Journal of Statistical Software*, 32(10):1–34, 2010.

Paul Yip. A martingale estimating equation for a capture-recapture experiment in discrete time. *Biometrics*, 47(3):1081–1088, 1991.

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.

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.

E.N. Zwane and P.G.M. van der Heijden. Implementing the parametric bootstrap in capture-recapture models with continuous covariates. *Statistics & Probability Letters*, 65:121–125, 2003.

E.N. Zwane and P.G.M. van der Heijden. Semiparametric models for capture-recapture studies with covariates. *Computational Statistics & Data Analysis*, 47:729–743, 2004.