Statistica Sinica **17**(2007), 301-316

EFFECTS OF MEASUREMENT ERROR AND CONDITIONAL SCORE ESTIMATION IN CAPTURE-RECAPTURE MODELS

Wen-Han Hwang, Steve Y. H. Huang and C. Y. Wang

National Chung Hsing University, Tamkang University and Fred Hutchinson Cancer Research Center

Abstract: Although the literature in covariate measurement error is rather rich, the focus is primarily on regression coefficient estimation; far less is known in the context of capture-recapture experiments. In this article, we provide justification for effects of measurement error on estimation in a capture-recapture model. When errors are present, the conventional approach is shown to have bias in parameter estimation and it may underestimate the population size. In fact, no consistent estimation has been proposed before, especially for the functional case. We propose a new conditional score estimation to adjust for measurement error in capture-recapture models. This approach estimates regression parameters and population size consistently without imposing any distributional assumption on the covariates. An example involving the bird species *Prinia flaviventris* is used to illustrate the effects, and intensive simulation studies are conducted to evaluate the performance of the proposed estimator along with other existing methods. Under most simulation scenarios, this new method is preferable since it has smaller biases and better coverage probabilities.

Key words and phrases: Capture-recapture, conditional score, Horvitz-Thompson, measurement error, population size, regression calibration.

1. Introduction

Estimation of population size is one of the most important issues in ecology studies. A variety of capture-recapture models have been proposed and some inference procedures have been developed to address this issue. These models were developed to incorporate variation due to capture time, behavior response, individual heterogeneity, or a combination of these factors. They have been extensively studied; for a review, see Seber (1982), Chao (2001), Borchers, Buckland and Zucchini (2002), and references therein.

Pollock, Hine and Nichols (1984) proposed a heterogeneity model that accommodates information of individual characteristics, such as sex, weight, and wing length, to model the capture probabilities of the animals. They only considered categorical covariates. Huggins (1989) and Alho (1990) extended the

case to continuous covariates. They developed a conditional likelihood model for inference of the regression coefficients associated with the capture probabilities, and then proposed a Horvitz-Thompson (HT) weighted estimator to estimate the population size. Pollock (2002) indicated that this conditional likelihood approach has become a standard technique today. Pollock (2002) also pointed out that the covariates in the approach of Huggins-Alho are assumed to be measured without error, which may not be always realistic.

Measurement error is an important and common problem in epidemiological, medical, and other disciplines. It is well known that measurement error may cause bias in regression analysis and subsequently lead to invalid statistical inference. There is a growing body of literature investigating measurement error problems. Modern statistical methods have been reviewed in Fuller (1987), Carroll, Ruppert and Stefanski (1995), and Cheng and Van Ness (1998). However, most studies in this area to date have their focus primarily on the analysis of regression coefficients in generalized linear models, or in failure time regression models (Huang and Wang (2000)). There has been very limited work done on adjusting for measurement error in capture-recapture models.

For the capture-recapture problem, Hwang and Huang (2003) discussed bias in estimating the population size when measurement errors are ignored. They applied a refined regression calibration (RRC) estimator to correct estimation of the regression coefficients in the capture probabilities and then used them in an adjusted HT population size estimator. Their analysis and simulation results suggest an explanation for the severe bias from the naive regression coefficient and population size estimators. However, there is no theoretical justification provided for the bias of naive population size estimators. Furthermore, their method is based on the assumption that the true covariate variables are normally distributed. These shortcomings and restrictions motivate the present work.

This paper is organized as follows. In Section 2, we briefly review the conventional heterogeneity capture-recapture model of Huggins (1989) and the RRC approach of Hwang and Huang (2003). In Section 3, we show that measurement error has a *dependence effect* on the capture of an animal given its surrogate variables. We show that the RRC may encounter bias even though the covariate variables are normally distributed. Moreover, we explain how a naive estimator may underestimate the population size of interest. In Section 4, following the idea of conditional score estimation (Stefanski and Carroll (1987)), we construct an unbiased conditional score estimating function when measurement errors are present. When there is no measurement error, the conditional score approach reduces to the traditional approach. Some results from intensive simulation studies are provided in Section 5. Section 6 presents an illustrative example using data on the bird species *Prinia flaviventris* in Hong Kong. Final conclusions and suggestions are given in Section 7.

2. Statistical Modeling and the RRC Approach

2.1. Model and conditional maximum likelihood

Assume the population of interest consists of N animals in a capture-recapture experiment. Let $j = 1, \ldots, t$ index the trapping samples, Y_{ij} be the indicator function for whether or not the ith animal is caught in the jth sample, and $Y_i = \sum_{j=1}^{t} Y_{ij}$ be the number of times the ith animal is caught during the experiment. Given covariates, animals are assumed to behave independently among individuals and across trapping samples. The probability that the ith animal is captured in any trapping sample is assumed to be

$$p_i = P(Y_{ij} = 1|\boldsymbol{X}_i) = H(\alpha + \boldsymbol{\beta}^{'}\boldsymbol{X}_i), \quad i = 1, \ldots, N,\ j = 1, \ldots, t, \tag{1}$$

where $H(u) = \{1 + \exp(-u)\}^{-1}$ is the logistic function, $\boldsymbol{X}_i$ denotes the covariate vector of the ith animal, and α and $\boldsymbol{\beta}^{'}$ are unknown parameters. This model was discussed in Pollock, Hine and Nichols (1984) and is a version of the heterogeneity model in Otis, Burnham, White and Anderson (1978).

Let $\mathcal{C}_i$ be the event that the ith animal is caught at least once during the experiment and $D = \sum_{i=1}^{N} \mathrm{I}(\mathcal{C}_i)$ be the number of distinct animals captured, where $\mathrm{I}(\cdot)$ is the usual indicator function. Without loss of generality, we assume $\mathrm{I}(\mathcal{C}_i) = 1$ for $i = 1, \ldots, D$ and $\mathrm{I}(\mathcal{C}_j) = 0$ for $j = D + 1, \ldots, N$. The conditional maximum likelihood estimate (Huggins (1989)) of $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}')'$ solves

$$\sum_{i=1}^{D} \begin{pmatrix} 1 \\ \boldsymbol{X}_i \end{pmatrix} \left\{ Y_i - \frac{tp_i}{1 - (1 - p_i)^t} \right\} = 0. \tag{2}$$

Note that $1-(1-p_i)^t$ equals $P(\mathcal{C}_i)$, the probability of being captured at least once. Huggins (1989) proposed two types of HT estimators to estimate the population size. We refer to them as the first and second HT estimators, and there are given by

$$\widehat{N}_1 = \sum_{i=1}^{N} \frac{\mathrm{I}(\mathcal{C}_i)}{1 - \{1 - p_i(\widehat{\boldsymbol{\theta}}_M)\}^t} = \sum_{i=1}^{D} \frac{1}{1 - \{1 - p_i(\widehat{\boldsymbol{\theta}}_M)\}^t}, \tag{3}$$

$$\widehat{N}_2 = \sum_{i=1}^{N} \frac{Y_i}{tp_i(\widehat{\boldsymbol{\theta}}_M)}, \tag{4}$$

where $\widehat{\boldsymbol{\theta}}_M$ is the root of (2). These estimators are consistent because the left-hand side of (2) has mean 0, and both (3) and (4) have mean N when evaluated at the true parameter. Two cautions are noted here. First, the consistency of a population size estimator $\widehat{N}$ is defined as $(\widehat{N} - N)/N \xrightarrow{p} 0$; see Huggins (1989) for details. Second, although the difference between $\widehat{N}_1$ and $\widehat{N}_2$ is not much

from our simulation experience, it is easy to show that the asymptotic variance of the first-type HT estimator is smaller than the second-type HT estimator for a known parameter $\boldsymbol{\theta}$ when there is no measurement error. Thus, the first-type HT estimator is preferable whenever it is available.

2.2. RRC estimator for normal covariate variables

In this subsection, we briefly review the RRC estimator proposed by Hwang and Huang (2003). When the vector $\boldsymbol{X}_i$ is measured with random errors, we denote the observed surrogate measurement for $\boldsymbol{X}_i$ by $\boldsymbol{W}_i$ and write $\boldsymbol{W}_i = \boldsymbol{X}_i + \boldsymbol{U}_i$. For any component of $\boldsymbol{X}_i$, the corresponding component of $\boldsymbol{U}_i$ is 0 if it is not measured with error. The RRC method generally works satisfactorily under the following assumptions.

(R1) $\boldsymbol{U}_i$ is $N(0, \boldsymbol{\Sigma}_u)$ distributed, where $\boldsymbol{\Sigma}_u$ is a known matrix that can be singular when not all components of $\boldsymbol{X}$ are measured with errors.
(R2) $\boldsymbol{U}_i$ and $(\boldsymbol{X}_i, Y_i)$ are stochastically independent.
(R3) $P(Y_{ij} = 1|\boldsymbol{X}_i, \boldsymbol{W}_i) = P(Y_{ij} = 1|\boldsymbol{X}_i)$ for $j = 1, \ldots, t$, $i = 1, \ldots, N$.
(R4) $\boldsymbol{X}_i$ is $N(\mu_x, \boldsymbol{\Sigma}_x)$ distributed, where $\boldsymbol{\Sigma}_x$ is a full-rank matrix.

The assumptions here are essentially equivalent to those in Hwang and Huang (2003). Because some individual covariates such as gender and age (adult vs. juvenile) do not have a measurement error problem, the variance matrix $\boldsymbol{\Sigma}_u$ may contain some 0 entries. We assume $\boldsymbol{\Sigma}_u$ is known for simplicity; in practice, it usually can be estimated by repeated surrogate measurements. Repeat measurements are possible in a capture-recapture experiment since individual covariates could be measured on every capture occasion. Assumption (R4) is not needed for the new estimation method in Section 4.

The RRC method uses $H(\alpha^* + \boldsymbol{\beta}^{*'}\boldsymbol{W}_i)$ to approximate $P(Y_{ij} = 1|\boldsymbol{W}_i)$, where

$$(\alpha^*, \boldsymbol{\beta}^{*'})' = \{1 + \frac{\boldsymbol{\beta}' \mathrm{Var}\,(\boldsymbol{X}|\boldsymbol{W})\boldsymbol{\beta}}{2.89}\}^{-\frac{1}{2}} \times \left(\alpha + \boldsymbol{\beta}'\mu_x - \boldsymbol{\beta}'\boldsymbol{\Sigma}_x(\boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_u)^{-1}\mu_x,\ \boldsymbol{\beta}'\boldsymbol{\Sigma}_x(\boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_u)^{-1}\right)',$$

and $\mathrm{Var}\,(\boldsymbol{X}|\boldsymbol{W})$ is the conditional variance of $\boldsymbol{X}$ given $\boldsymbol{W}$.

If $p_{Ri} = H(\alpha^* + \boldsymbol{\beta}^{*'}\boldsymbol{W}_i)$, an estimating equation for $(\alpha, \boldsymbol{\beta}')$ is

$$\sum_{i=1}^{D} \binom{1}{\boldsymbol{W}_i} \left\{ Y_i - \frac{t p_{Ri}}{1 - (1 - p_{Ri})^t} \right\} = 0. \tag{5}$$

The solution to (5) is denoted by $\widehat{\boldsymbol{\theta}}_R = (\widehat{\alpha}_R, \widehat{\boldsymbol{\beta}}_R')'$, which can be derived (approximately) by replacing $\alpha + \boldsymbol{\beta}'\boldsymbol{X}_i$ with $\{1 + \boldsymbol{\beta}'\mathrm{Var}\,(\boldsymbol{X}|\boldsymbol{W})\boldsymbol{\beta}/2.89\}^{-1/2} E((\alpha + \boldsymbol{\beta}'\boldsymbol{X}_i) \mid \boldsymbol{W}_i)$ in (2). Hence we can refer to it as an RRC estimator. Hwang and Huang

(2003) noted that the first HT population size estimator based on the RRC estimator of $\boldsymbol{\theta}_R$ is essentially the same as the first HT naive estimator, and they proposed the following adjusted population size estimator based on the second HT estimator:

$$\widehat{N}_R = \frac{1}{t}\sum_{i=1}^{N} \frac{Y_i}{H\{\widehat{\alpha}_R + \widehat{\boldsymbol{\beta}}_R' \boldsymbol{W}_i + \frac{1}{2}\widehat{\boldsymbol{\beta}}_R' \boldsymbol{\Sigma}_u \widehat{\boldsymbol{\beta}}_R\}}. \tag{6}$$

The RRC estimator given above provides a convenient way to improve the naive estimators when measurement error is present. However, it is still not consistent because there is indeed some bias in (5). This bias will be confirmed in the following section.

3. Dependence Effect and Bias

In this section, we show that measurement error will affect the independence structure of $Y_{ij}, j = 1, \ldots, t$, and the naive estimator $\widehat{N}_1$ will generally underestimate the population size. To illustrate the bias of the estimating equation of the naive estimator, we consider the case when the X_i is a univariate normal random variable. Recall that $p_i = P(Y_{ij} = 1 \mid X_i) = H(\alpha + \beta X_i)$ as defined in (1). The random variables $Y_{i1}, \ldots, Y_{it}$ are i.i.d. Bernoulli random variables with mean p_i when conditioned on X_i. Given W_i, $Y_{i1}, \ldots, Y_{it}$ are Bernoulli variables with parameter p_{Ri} (approximately), but they no longer carry the independence property. This can be seen by noting that

$$E(Y_{ij}Y_{ik} \mid W_i) = E\{H^2(\alpha + \boldsymbol{\beta} X_i) \mid W_i\} \neq E(Y_{ij} \mid W_i)E(Y_{ik} \mid W_i),$$

for $1 \leq j \neq k \leq t$. Consequently, the quantity $1 - \{1 - P(Y_{i1} = 1 \mid W_i)\}^t$ may not be a good approximation to $P(Y_i \geq 1 \mid W_i)$. Furthermore, since $1 - (1 - H)^t$ is a concave function of H on $(0, 1)$, we have

$$\begin{aligned} P(Y_i \geq 1 \mid W_i) &= E[E\{I(Y_i \geq 1) \mid X_i, W_i\} \mid W_i] \\ &= E[1 - \{1 - H(\alpha + \beta X_i)\}^t \mid W_i] \leq 1 - [1 - E\{H(\alpha + \beta X_i) \mid W_i\}]^t \\ &\approx 1 - \{1 - H(\alpha^* + \beta^* W_i)\}^t = 1 - (1 - p_{Ri})^t. \end{aligned}$$

For example, if W is a positive variable, then it is approximately true that

$$\sum_{i=1}^{D} \binom{1}{W_i} \left\{ Y_i - \frac{t p_{Ri}}{1 - (1 - p_{Ri})^t} \right\} \geq \sum_{1}^{D} \binom{1}{W_i} \{ Y_i - \frac{t p_{Ri}}{P(Y_i \geq 1 \mid W_i)} \}. \tag{7}$$

Because the right-hand side of (7) has mean 0, the left-hand side of (7) has a positive expectation when evaluated at (α^*, β^*). As a consequence, the estimator

solving (5) does not converge to (α^*, β^*). Instead, it converges to $(\alpha^{**}, \beta^{**})$, where $(\alpha^{**}, \beta^{**})$ satisfies

$$E\begin{pmatrix} 1 \\ W_i \end{pmatrix}\left[Y_i - \frac{tH(\alpha^{**} + \beta^{**}W_i)}{1 - \{1 - H(\alpha^{**} + \beta^{**}W_i)\}^t}\right] = 0. \tag{8}$$

Furthermore, since $H(s)/[1-\{1-H(s)\}^t]$ is an increasing function of s, it follows that either $\alpha^{**} > \alpha^*$, $\beta^{**} > \beta^*$, or both inequalities hold provided that W is positive. This implies that solving (5) generally will not yield a consistent estimator of $(\alpha^*, \boldsymbol{\beta}^{*'})'$, and hence the estimator of $(\alpha, \boldsymbol{\beta}')'$ is not consistent. This explains the bias of the RRC and hence the naive estimators for estimating regression parameters in the capture model.

For the estimation of the population size, if the two inequalities $\alpha^{**} > \alpha^*$ and $\beta^{**} > \beta^*$ both hold, then p_{Ri} is less than $H(\alpha^{**} + \beta^{**}W_i)$, and hence

$$\sum_{i=1}^{N} \frac{\mathrm{I}(\mathcal{C}_i)}{P(Y_i \geq 1 \mid W_i)} \geq \sum_{i=1}^{N} \frac{\mathrm{I}(\mathcal{C}_i)}{1 - (1 - p_{Ri})^t} \geq \sum_{i=1}^{N} \frac{\mathrm{I}(\mathcal{C}_i)}{1 - \{1 - H(\alpha^{**} + \beta^{**}W_i)\}^t}.$$

The first term has mean N. The last term is exactly the naive population size estimator in (3) evaluated at parameter $(\alpha^{**}, \beta^{**})$, which is the limit of the naive estimator $(\widehat{\alpha}_M, \widehat{\beta}_M)$ as N goes to infinity. This explains why the population size will generally be underestimated by the naive estimator.

Briefly, the above heuristic arguments exhibit two major difficulties for measurement errors for capture-recapture models. One is the change of regression coefficients between response and observed covariates, while the other is the correlation of the binary responses $Y_{i1}, \ldots, Y_{it}$. The RRC approach provided a solution to the first problem, but not the second. Hence there is still bias in estimating the regression parameters and the population size.

It is worthwhile noting that as the magnitude of the measurement errors becomes large, the surrogates provide less information on captures, and the conditional probability of being captured would likely shrink to a constant – the average capture percentage. In such a situation, the model looks more like a homogeneous model rather than a heterogeneous model. Analysis is then very similar to treating a heterogeneous model as if it were a homogeneous model. Thus the estimator derived in the way of analyzing a homogeneous model would underestimate the population size in general when the true model is indeed heterogeneous; see Burnham and Overton (1978) and Chao, Lee and Jeng (1992). Recently, Hwang and Huggins (2005) have provided a theoretical justification for this phenomenon.

4. Conditional Score Estimation

We now present the conditional score (CS) method, which is a semiparametric estimation and does not require Assumption (R4). It is applicable in

certain functional measurement error models (Carroll, Ruppert and Stefanski (1995, Chap.6)), especially for the generalized linear model with natural parameter being a linear function of covariates.

We pursue the conditional distribution of Y_i when appropriate statistics are given. This conditional distribution is made to be independent of the unknown $\boldsymbol{X}_i$ given some sufficient statistics. Here, $\boldsymbol{X}_i$ is treated as a parameter and hence no distribution assumption on $\boldsymbol{X}_i$ is needed.

4.1. Conditional scores

Consider the conditional distribution of Y_i given $\boldsymbol{X}_i$ and $\mathcal{C}_i$:

$$P(Y_i = y_i|\boldsymbol{X}_i, \mathcal{C}_i) = t!\{y_i!(t-y_i)!\}^{-1} \exp\{y_i(\alpha + \boldsymbol{\beta}'\boldsymbol{X}_i) + \mathcal{D}(\eta_i)\}, \quad y_i = 1, \ldots, t,$$

which belongs to the generalized linear model (McCullagh and Nelder (1989)) with linear predictor $\eta_i = \alpha + \boldsymbol{\beta}'\boldsymbol{X}_i$ and $\mathcal{D}(\eta_i) = \log\{(1-p_i)^t/(1-(1-p_i)^t)\}$. From (R2) and (R3), the joint distribution of Y_i and $\boldsymbol{W}_i$, conditional on $\boldsymbol{X}_i$ and $\mathcal{C}_i$, is given by

$$f(y_i, \boldsymbol{w}_i|\boldsymbol{X}_i, \mathcal{C}_i) = t!\{y_i!(t-y_i)!\}^{-1} \exp\{y_i(\alpha + \boldsymbol{\beta}'\boldsymbol{X}_i) + \mathcal{D}(\eta_i) - \frac{1}{2}(\boldsymbol{w}_i - \boldsymbol{X}_i)'\boldsymbol{\Sigma}_u^{-1}(\boldsymbol{w}_i - \boldsymbol{X}_i)\}.$$

Denoting $\boldsymbol{W}_i + Y_i\boldsymbol{\Sigma}_u\boldsymbol{\beta}$ by $\boldsymbol{\Delta}_i$, it is clear that $\boldsymbol{\Delta}_i$ is a sufficient and complete statistic of $\boldsymbol{X}_i$ if $\boldsymbol{\beta}$ is known (Stefanski and Carroll (1987)). By sufficiency, the conditional distribution of Y_i given $\boldsymbol{\Delta}_i$ and $\mathcal{C}_i$ is independent of $\boldsymbol{X}_i$, and can be shown to be

$$P(Y_i = y_i|\boldsymbol{\Delta}_i, \mathcal{C}_i) = t!\{y_i!(t-y_i)!\}^{-1} \exp\{y_i\eta_i^* - \frac{1}{2}y_i^2\boldsymbol{\beta}'\boldsymbol{\Sigma}_u\boldsymbol{\beta} - \log S_i(\eta^*, \boldsymbol{\beta})\}, \quad (9)$$

for $y_i = 1, \ldots, t$, where $\eta_i^* = \alpha + \boldsymbol{\beta}'\boldsymbol{\Delta}_i$ and $S_i(\eta^*, \boldsymbol{\beta}) = \sum_{y_i=1}^{t} t!\{y_i!(t-y_i)!\}^{-1}$ $\exp\{y_i\eta_i^* - y_i^2\boldsymbol{\beta}'\Sigma_u\boldsymbol{\beta}/2\}$. The conditional score (CS) estimator of $\boldsymbol{\theta}$, denoted by $\widehat{\boldsymbol{\theta}}_C$, solves

$$G(\boldsymbol{\theta}) = \sum_{i=1}^{D} g_i(\boldsymbol{\theta}) = \sum_{i=1}^{D} \begin{pmatrix} 1 \\ \boldsymbol{\Delta}_i^* \end{pmatrix} \{Y_i - E(Y_i \mid \boldsymbol{\Delta}_i, \mathcal{C}_i)\} = 0, \quad (10)$$

where $\boldsymbol{\Delta}_i^* = E(\boldsymbol{W}_i|\boldsymbol{\Delta}_i, \mathcal{C}_i) = \boldsymbol{\Delta}_i - E(Y_i|\boldsymbol{\Delta}_i, \mathcal{C}_i)\boldsymbol{\Sigma}_u\boldsymbol{\beta}$ and $E(Y_i|\boldsymbol{\Delta}_i, \mathcal{C}_i) = \sum_{k=1}^{t}$ $kP(Y_i = k|\boldsymbol{\Delta}_i, \mathcal{C}_i)$. It is easy to see that the conditional expectation of $g_i(\boldsymbol{\theta})$ given $\boldsymbol{\Delta}_i$ and $\mathcal{C}_i$ is 0. Therefore the expectation of $G(\boldsymbol{\theta})$ is 0, and hence $\widehat{\boldsymbol{\theta}}_C$ can be shown to be consistent, and the estimate of the variance-covariance of $\widehat{\boldsymbol{\theta}}_C$ can be derived by the sandwich method.

4.2. Estimation of N

Since we have a conditional distribution (9) that does not involve any unknown $\boldsymbol{X}_i$, it is natural to develop a first-type HT estimator based on it. The population size can be consistently estimated by $\sum_{i=1}^{N}\{\mathrm{I}(\mathcal{C}_i)/P(\mathcal{C}_i|\boldsymbol{\Delta}_i)\}$. Therefore, the proposed CS estimate for the population size is

$$\widehat{N}_C = \sum_{i=1}^{N} \frac{\mathrm{I}(\mathcal{C}_i)}{\widehat{P}(\mathcal{C}_i|\boldsymbol{\Delta}_i)}, \tag{11}$$

where $\widehat{P}(\mathcal{C}_i|\boldsymbol{\Delta}_i)$ is $P(\mathcal{C}_i|\boldsymbol{\Delta}_i)$ evaluated at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}_C$ and $\boldsymbol{\Delta}_i = \boldsymbol{W}_i + Y_i\boldsymbol{\Sigma}_u\widehat{\boldsymbol{\beta}}_C$. This CS estimator can be shown to be consistent in the sense that $(\widehat{N}_C - N)/N \xrightarrow{p} 0$ as $N \to \infty$. By a Taylor series expansion, the asymptotic variance of $\widehat{N}_C$ can be shown to be

$$\widehat{\mathrm{Var}}\,(\widehat{N}_C) = \Big\{ \sum_{i=1}^{D} \frac{1 - P(\mathcal{C}_i|\boldsymbol{\Delta}_i)}{P(\mathcal{C}_i|\boldsymbol{\Delta}_i)^2} + (\frac{\partial \widehat{N}_C}{\partial \boldsymbol{\theta}})'\widehat{\mathrm{Var}}\,(\widehat{\boldsymbol{\theta}}_C)(\frac{\partial \widehat{N}_C}{\partial \boldsymbol{\theta}}) \Big\} \,|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_C} .$$

The proof is straightforward and hence is omitted here.

4.3. Some notes on the CS method

A crucial condition for the conditional score method being applicable is that a sufficient statistic of $\boldsymbol{X}_i$ can be found. This would be easy if the joint distribution of $(Y_i, \boldsymbol{W}_i)$ given $\boldsymbol{X}_i$ is in the exponential family with $\boldsymbol{X}_i$ being the canonical parameter. In such a situation, a complete and sufficient statistic for $\boldsymbol{X}_i$ can be identified immediately. Thus if the regression models of Y_i given $\boldsymbol{X}_i$ and $\boldsymbol{W}_i$ given $\boldsymbol{X}_i$ are both generalized linear models, then we should require the link functions to be natural links. The frequently used logit link function and the Binomial distribution assumption of Y_i given $\boldsymbol{X}_i$, as well as normality assumption of U_i, clearly meet the above requirements. To check the adequacy of the logistic regression assumption, one can compute estimates of the first two moments for the distribution in (9), and compute a test statistic

$$T \equiv \sum_{i=1}^{D} \frac{\{Y_i - \widehat{E}(Y_i \mid \boldsymbol{\Delta}_i, \mathcal{C}_i)\}^2}{\widehat{\mathrm{Var}}\,(Y_i \mid \boldsymbol{\Delta}_i, \mathcal{C}_i)}$$

that has an approximate chi-square distribution with degree of freedom $D-p-1$ under the assumptions of CS method.

The CS method is also applicable to the random effect model Consider a random effect model

$$P(Y_{ij} = 1 \mid \boldsymbol{X}_i, R_i) = H(\alpha + \boldsymbol{\beta}'\boldsymbol{X}_i + R_i),\ i = 1, \ldots, N,\ j = 1, \ldots, t, \tag{12}$$

where the R_i are $N(0, \sigma^2)$ distributed for all i and are independent. The model is similar to a model in Coull and Agresti (1999), except that their model contains no individual covariate. A simple calculation shows that

$$P(Y_i = y_i|\boldsymbol{\Delta}_i, R_i, \mathcal{C}_i) \propto t!\{y_i!(t-y_i)!\}^{-1} \times \exp\{y_i(\alpha + \boldsymbol{\beta}'\boldsymbol{\Delta}_i + R_i) - \frac{1}{2}y_i^2\boldsymbol{\beta}'\boldsymbol{\Sigma}_u\boldsymbol{\beta}\}, \tag{13}$$

which is similar to (9) but with a different normalizing constant. Following the arguments of Coull and Agresti (1999), and based on the conditional distribution (13), we can estimate $\boldsymbol{\beta}$ consistently. However, the performance of the CS estimator in this setup requires further investigation.

5. Simulation Study

A simulation study was conducted to evaluate the performance of the proposed conditional score estimator. We considered population sizes $N = 400$ and $N = 1,000$, and there were five capture samples in every experiment. The probability of being captured was set to be $p_i = H(\alpha + \beta X_i)$, where the unobserved covariates X_i were generated from two distributions, the $N(0,1)$ and a mixture normal distribution. The mixture normal was from two normal variables with means $(-2/\sqrt{5}, 2/\sqrt{5})$, variances $(1/5, 1/5)$, and the mixture percentage was fixed at 0.5 so that the mixture normal distribution also has mean 0 and variance 1. The observed surrogates W_i, $i = 1, \ldots, N$, were generated by $W_i = X_i + U_i$, where the U_i were from $N(0, \Sigma_u)$. The variance Σ_u had three levels, 0, 0.5, and 1. Finally, we generated Y_{ij} from a Bernoulli distribution with mean $H(\alpha + \beta X)$ and $(\alpha, \beta) = (-1, 1)$. For all parameter combinations, 1,000 simulation samples were generated. For each sample, we computed the following estimates and their standard errors: (i) $\widehat{\boldsymbol{\theta}}_N = (\widehat{\alpha}_N, \widehat{\beta}_N)'$, the naive conditional maximum likelihood estimate which replaces X_i by W_i in solving (2); (ii) $\widehat{\boldsymbol{\theta}}_R = (\widehat{\alpha}_R, \widehat{\beta}_R)'$, the RRC estimate that solves (5); (iii) $\widehat{\boldsymbol{\theta}}_C = (\widehat{\alpha}_C, \widehat{\beta}_C)'$, the conditional score estimate that solves (10); (iv)$\widehat{N}_1$, the first-type HT estimates that use $\widehat{\boldsymbol{\theta}}_N$ in (3); (v) $\widehat{N}_R$, the adjusted HT estimates based on $\widehat{\boldsymbol{\theta}}_R$ in (6); (vi) $\widehat{N}_C$, the proposed HT estimates based on $\widehat{\boldsymbol{\theta}}_C$ in (11).

The averages of the resulting 1,000 parameter estimates and standard error estimates are given in Tables 1−2. We present the relative bias (RB), the empirical SE (standard error), the average of estimated SE, the sample root mean squared error (RMSE), and the sample coverage percentage (CP) of 95% confidence intervals. We also report the average number of distinct individuals that were captured ($\bar{D}$) and the total number of captures ($\overline{\sum Y_i}$) in the experiments.

Table 1. Comparison of estimator performance, where X is normally distributed, $N = 400$ (upper panel) and $N = 1,000$ (lower panel).

Σ_u	Method	Average Estimate	RB (%)	Empirical SE	Average SE	RMSE	CP (%)
$N = 400$,	$\bar{D} = 294$, $\overline{\Sigma Y_i} = 606$						
$\Sigma_u = 0$	$\widehat{N}_1$	403	0.9	27.6	26.4	27.9	94.9
$\Sigma_u = 0.5$	$\widehat{N}_1$	362	-9.4	17.6	15.7	41.6	34.6
	$\widehat{N}_R$	388	-2.8	30.3	32.1	32.4	87.1
	$\widehat{N}_C$	406	1.6	38.1	35.2	38.6	92.9
$\Sigma_u = 1$	$\widehat{N}_1$	349	-12.5	14.7	12.8	52.2	8.6
	$\widehat{N}_R$	383	-4.0	37.0	38.9	40.3	82.2
	$\widehat{N}_C$	410	2.6	54.0	46.2	55.0	91.5
$N = 1,000$,	$\bar{D} = 734$, $\overline{\Sigma Y_i} = 1517$						
$\Sigma_u = 0$	$\widehat{N}_1$	1002	0.3	40.3	40.8	40.4	95.2
$\Sigma_u = 0.5$	$\widehat{N}_1$	902	-9.7	28.3	24.5	101.5	7.4
	$\widehat{N}_R$	964	-3.5	46.3	48.7	58.4	81.4
	$\widehat{N}_C$	1003	0.3	54.8	51.2	54.9	93.8
$\Sigma_u = 1$	$\widehat{N}_1$	872	-12.7	23.6	19.8	130.0	0.2
	$\widehat{N}_R$	950	-4.9	53.1	57.5	72.7	74.6
	$\widehat{N}_C$	1009	1.0	68.0	63.5	68.7	93.2

Table 2. Comparison of estimator performance, where X is mixture normally distributed, $N = 400$ (upper panel) and $N = 1,000$ (lower panel).

Σ_u	Method	Average Estimate	RB (%)	Empirical SE	Average SE	RMSE	CP (%)
$N = 400$,	$\bar{D} = 290$, $\overline{\Sigma Y_i} = 613$						
$\Sigma_u = 0$	$\widehat{N}_1$	404	1.0	29.5	28.7	29.8	94.7
$\Sigma_u = 0.5$	$\widehat{N}_1$	352	-11.8	16.6	14.7	50.3	18.2
	$\widehat{N}_R$	370	-7.3	26.0	26.7	39.2	69.1
	$\widehat{N}_C$	407	1.9	40.3	38.4	41.0	93.4
$\Sigma_u = 1$	$\widehat{N}_1$	339	-15.1	14.0	11.5	62.3	2.2
	$\widehat{N}_R$	361	-9.5	28.1	30.0	47.6	57.2
	$\widehat{N}_C$	407	1.9	49.1	47.7	49.8	91.7
$N = 1,000$,	$\bar{D} = 726$, $\overline{\Sigma Y_i} = 1534$						
$\Sigma_u = 0$	$\widehat{N}_1$	1002	0.2	45.0	44.1	45.1	94.4
$\Sigma_u = 0.5$	$\widehat{N}_1$	876	-12.3	26.1	22.7	125.9	0.9
	$\widehat{N}_R$	918	-8.1	39.1	40.8	89.9	43.6
	$\widehat{N}_C$	1008	0.8	59.9	57.4	60.5	95.3
$\Sigma_u = 1$	$\widehat{N}_1$	848	-15.1	22.6	18.0	153.3	0.0
	$\widehat{N}_R$	900	-9.9	40.2	45.3	107.5	36.4
	$\widehat{N}_C$	1013	1.3	71.9	69.6	73.2	93.4

Note that we say an estimator of the population size or regression coefficient has failed if its value is 5 times larger than D, or if it does not reach a stable value after 1,000 iterations. If an estimator failed in a sample, then we discarded that sample and generated another one. The simulation study was carried out until 1,000 non-failure estimates were derived. From the various simulation replicates, there was one failure from the RRC estimator and two failures from the CS estimator under the setup that X is from a mixture normal and N equals 400. When the sample size decreases or the measurement error increases, the CS estimator will likely have more divergences.

We omit regression coefficients estimation in Tables 1 and 2 since the findings are similar in many regression models, and our interest here is primarily in population size estimation. Briefly, for the estimation of the regression coefficients, the naive estimates $\widehat{\alpha}_N$ and $\widehat{\beta}_N$ deteriorate as the measurement errors become large; it is also clear that there is an attenuation effect in $\widehat{\beta}_N$. The RRC estimates $\widehat{\alpha}_R$ and $\widehat{\beta}_R$ perform well in the normal case, especially when Σ_u is small, but for the mixture normal cases, the RRC is less satisfactory when compared to the conditional score estimator. The proposed conditional score approach generally performs best in terms of RB, RMSE, and CP among all three methods. The asymptotic variance estimator based on the sandwich method also works well.

Concerning population size inference, Table 1 (the normal case) show that the naive estimator $\widehat{N}_1$ exhibits a downward tendency when the measurement error variance Σ_u increases. This agrees with our remarks in Section 3. This downward tendency still holds for nonnormal X in Table 2, although it is not verified in general.

The RRC estimator $\widehat{N}_R$ has a small negative bias in all cases, but it is less than $\widehat{N}_1$. The CS estimator, $\widehat{N}_C$ shows a slight positive bias, this is probably due to a finite sample property of the Horvitz-Thompson type estimator, see Hwang and Huang (2003). The RB here reduces as N increases from 400 to 1,000, which agrees with the fact that it is a consistent estimator of N. Even if it has slightly larger standard errors than the others, its coverage probabilities are very close to the nominal level. Reducing bias at the cost of increasing variance is a common phenomenon in measurement error.

Although we applied robust sandwich estimates, some biases were observed in the variance estimates for all population size estimators. Note that variance estimation for regression coefficient estimation has less bias, and the bias reduces generally when the sample size increases. The explanation for variance estimation bias for population size estimators is more complicated than that for regression coefficients given that increasing the sample size actually means increasing the parameter of interest. A more accurate variance estimation for the CS population size estimator is an interesting topic to be pursued.

In summary, the conditional score estimator is preferable to the existing estimators, $\widehat{N}_1$ and $\widehat{N}_R$, due to its smaller biases, better coverage probabilities and robustness.

6. Numerical Example: A Sensitivity Analysis

We consider measurement error analyses of the bird species *Prinia flaviventris* data, collected by the Hong Kong Bird Society in 1993 at Mai Po Bird Sanctuary of Hong Kong. After excluding one bird which did not have its covariate, wing length, the data set to be analyzed consists of capture histories of 164 distinct birds that were caught in 206 total captures on 17 trapping occasions. The average observed wing length is 45.24 mm, with sample variance 1.61.

Suppose the capture probability is modeled through a logistic function $H(\alpha+\beta X_i)$, where X_i is the exact wing length of the ith individual. We conducted some sensitivity analyses of the effect of measurement error on estimation by analyzing the data under various reliability ratio R values, where R is defined as $Var(X)/Var(W)$. We considered $R = 100\%, 75\%$, and 50%, which corresponded to high, medium, and poor qualities of measurement instruments or techniques.

Table 3 shows the results of the various estimators discussed in the simulation study under different levels of reliability. The standard error estimates of the RRC and the conditional score estimates are large because the data is sparse: there are only 42 recaptures out of 164 birds during the experiment. Since this example is sparse, we also conducted some bootstrap samples to obtain SE estimates and bootstrap percentile confidence intervals (BPCI). There are 1,000 bootstrap replications, and we discarded those bootstrap samples for which the regression coefficients estimates diverged or yielded size estimates larger than 1,640 (10 times D). The bootstrap SE's are only a little smaller than asymptotic SE, but the BPCI's are much shorter than the conventional one.

It is seen that the naive estimator may underestimate the population size by 60 and 260 animals when $R = 75\%$ and 50%, respectively, compared to the conditional score approach. The differences are so large that we believe it is plausible in real data that the degree of quality in measuring variables affects the resulting estimates substantially. We conclude that the measurement error approach should be considered when an investigator has a reason to question the precision of the measurement instruments.

7. Conclusion

We have demonstrated the effect of measurement error in the capture-recapture model. When errors are present, the naive approach estimates regression coefficients with considerable bias and usually underestimates the true population size. The RRC approach mitigates the bias and the underestimation problem.

Table 3. Comparison of estimator performance for *Prinia flaviventris* data under different assumptions of reliability. When the reliability $R = 1$, $\theta = (\alpha, \beta)$s and N are estimated by the naive approach. When the reliability $R < 1$, θs and N are estimated by the RC and conditional score approach.

Reliability	Method	Estimate	Estimate SE*	BPCI
Regression Coefficients				
$R = 100\%$	$\widehat{\alpha}$	-21.13	5.279 (4.400)	-29.64 ∼ -12.05
	$\widehat{\beta}$	0.38	0.115 (0.096)	0.18 ∼ 0.56
$R = 75\%$	$\widehat{\alpha}_R$	-27.33	7.341 (6.155)	-41.34 ∼ -15.90
	$\widehat{\alpha}_C$	-28.95	6.802 (8.248)	-51.48 ∼ -16.77
	$\widehat{\beta}_R$	0.52	0.160 (0.134)	0.26 ∼ 0.82
	$\widehat{\beta}_C$	0.55	0.148 (0.180)	0.28 ∼ 1.03
$R = 50\%$	$\widehat{\alpha}_R$	-40.31	12.021 (11.594)	-66.81 ∼ -22.49
	$\widehat{\alpha}_C$	-48.52	17.027 (27.243)	-79.12 ∼ -26.21
	$\widehat{\beta}_R$	0.80	0.262 (0.253)	0.41 ∼ 1.36
	$\widehat{\beta}_C$	0.98	0.372 (0.597)	0.49 ∼ 1.64
Population size				
$R = 100\%$	$\widehat{N}_1$	511	95.3 (108.3)	388 ∼ 798
	$\widehat{N}_2$	508	91.7 (107.8)	383 ∼ 786
$R = 75\%$	$\widehat{N}_R$	545	120.3 (147.2)	387 ∼ 966
	$\widehat{N}_C$	572	125.7 (174.6)	411 ∼ 1094
$R = 50\%$	$\widehat{N}_R$	626	195.0 (217.5)	420 ∼ 1282
	$\widehat{N}_C$	769	305.0 (264.0)	462 ∼ 1455

*Values in parentheses represent the bootstrap standard error estimates.

However, the RRC estimator may not be suitable when the measurement errors are large, in the case of high heterogeneity, or when missing covariates are not normally distributed. That is, the RRC estimation only provides a simple but approximate solution. In contrast, the CS approach has the following advantages. First, it does not require a distributional assumption on the covariates. Second, rather than approximation, it provides an exact inference procedure. In addition to the aforementioned advantages, the CS estimator offers a more direct approach to population size estimation. Apparently, existing methods for covariate measurement error can be applied to correct the estimation of regression coefficients in the capture-recapture probability model, but it often takes significant efforts to carry this adjustment forward to the inference of population size estimation. For example, the previous work of Hwang and Huang (2003) requires further calculation using the distribution of $Y_i|\boldsymbol{W}_i$, which involves some tedious approximations. Also, the population size estimator of Hwang and Huang was based on the second-type HT estimator $\widehat{N}_2$. That approach could not reduce

bias if applied to the well-known first-type HT weighted estimator $\widehat{N}_1$. In contrast, the idea of the proposed CS estimator is to use the distribution of $Y_i|\boldsymbol{\Delta}_i$, which does not require approximation and can be used to construct a first-type HT estimator. These are significant and exclusive features of the CS estimation procedure.

Although the CS method provides consistent estimation for both of the regression coefficients and the population size, some small sample performance cautions are noted. When information on the capture probability is insufficient, the CS method may be not work well and there may not exist a root for the regression coefficient estimation. For example, if the population size N equals 100 and the other parameters are the same as those in our simulation, then the RRC estimator is better than the CS method. However, in this case, if we increase the trapping samples from 5 to 10, the results are similar to those reported in Section 5. Furthermore, the superiority of the CS estimator over other estimators is more significant if the total population size is increased. Similar findings were noted in Wang and Huang (2001) in the context of joint modeling of logistic regression and longitudinal covariate data.

The phenomenon of underestimation of a population size was verified by Chen (1998) in a line transect model when sighting distances were measured with errors. It was also observed by Gould, Stefanski and Pollock (1997) in a catch-effort model via a simulation study. For the capture-recapture model, we have explained this downward bias under the conditions of normal covariates and normal measurement errors. Effects of measurement error on population size estimation under other distributions remain unknown.

Finally, estimation bias of regression coefficients and underestimation of the population size can result not only from the presence of measurement error, but also from omitting important covariates, even when all covariates are measured precisely. The reason is that if one overlooks a covariate related to the capture probability, then there might be correlation between two capture samples, and this would cause underestimation of the HT estimates. This follows by the same reasoning as that for the underestimation of the RC and naive methods discussed in Section 3.

Acknowledgements

The authors are grateful to the editor, associate editor and the referee for the constructive suggestions which have greatly improved the paper. This research was supported by the National Science Council of Taiwan (Hwang and Huang), and the National Institutes of Health grants AG15026 (Wang), and CA53996 (Wang).

References

Alho, J. M. (1990). Logistic regression in capture-recapture models. *Biometrics* **46**, 623-635.

Borchers, D. L., Buckland, S. T. and Zucchini, W. (2002). *Estimating Animal Abundance: Closed Populations.* Springer, London.

Burnham, K. P. and Overton, W. S. (1978). Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika* **75**, 625-633.

Carroll, R. J., Ruppert, D. and Stefanski, L. A. (1995). *Measurement Errors in Nonlinear Models*. Chapman and Hall, London.

Chao, A. (2001). An overview of closed capture-recapture models. *J. Agric. Biol. Environ. Stat.* **6**, 158-175.

Chao, A., Lee, S. M. and Jeng, S. L. (1992). Estimating population size for capture-recapture data when capture probabilities vary by time and individual animal. *Biometrics* **48**, 210-216.

Chen, S. X. (1998). Measurement errors in line transect surveys. *Biometrics* **54**, 899-908.

Cheng, C. L. and Van Ness, J. W. (1998). *Statistical Regression with Measurement Error*. Oxford University Press, London.

Coull, B. A. and Agresti, A. (1999). The use of mixed logit models to reflect heterogeneity in capture-recapture studies. *Biometrics* **55**, 294-301.

Fuller, W. A. (1987). *Measurement Error Models*. John Wiley and Sons, New York.

Gould, W. R., Stefanski, L. A. and Pollock, K. H. (1997). Effects of measurement error on catch-effort estimation. *Canad. J. Fish. Aquat. Sci.* **54**, 898-906.

Huang, Y. and Wang, C. Y. (2000). Cox regression with accurate covariates unascertainable: a nonparametric-correction approach. *J. Amer. Statist. Assoc.* **95**, 1209-1219.

Huggins, R. M. (1989). On the statistical analysis of capture experiments. *Biometrika* **76**, 133-140.

Hwang, W. and Huang, S. Y. H. (2003). Estimation in capture-recapture models when covariates are subject to measurement errors. *Biometrics* **59**, 1115-1124.

Hwang, W. H. and Huggins, R. M. (2005). An examination of the effect of heterogeneity on the estimation of population size using capture-recapture data. *Biometrika* **92**, 229-233.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Second edition. Chapman and Hall, New York.

Otis, D. L., Burnham, K. P., White, G. C. and Anderson, D. R. (1978). Statistical inference for capture-recapture data on closed animal populations. *Wildlife Monogr.* **62**, University of Kentucky, Louisville.

Pollock, K. H. (2002). The use of auxiliary variables in capture-recapture modeling: an overview. *J. Appl. Statist.* **29**, 85-102.

Pollock, K. H., Hine, J. E. and Nichols, J. D. (1984). The use of auxiliary variables in capture-recapture and removal experiments. *Biometrics* **40**, 329-340.

Seber, G. A. F. (1982). *The Estimation of Animal Abundance and Related Parameters.* Charles Griffin, London.

Stefanski, L. A. and Carroll, R. J. (1987). Conditional scores and optimal scores for generalized linear measurement-error models. *Biometrika* **74**, 703-716.

Wang, C. Y. and Huang, Y. (2001). Functional methods for logistic regression on random-effect-coefficient-covariates for longitudinal measurements. *Statist. Probab. Lett.* **53**, 347-356.

Department of Applied Mathematics, National Chung Hsing University, Taichung, Taiwan.
E-mail: wenhan@amath.nchu.edu.tw

Department of Mathematics, Tamkang University, Taipei, Taiwan.
E-mail: huang@math.tku.edu.tw

Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, P.O. Box 19024, Seattle, WA 98109-1024, U.S.A.
E-mail: cywang@fhcrc.org

(Received September 2004; accepted September 2005)