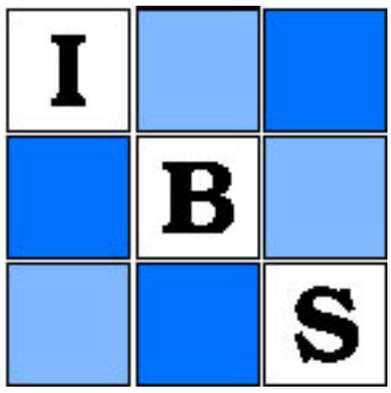

BIOMETRICS 46, 623–635
September 1990

Logistic Regression in Capture–Recapture Models

Juha M. Alho

Institute for Environmental Studies and Department of Statistics,
University of Illinois, 1101 W. Peabody Drive, Urbana, Illinois 61801, U.S.A.

SUMMARY

The effect of population heterogeneity in capture–recapture, or dual registration, models is discussed. An estimator of the unknown population size based on a logistic regression model is introduced. The model allows different capture probabilities across individuals and across capture times. The probabilities are estimated from the observed data using conditional maximum likelihood. The resulting population estimator is shown to be consistent and asymptotically normal. A variance estimator under population heterogeneity is derived. The finite-sample properties of the estimators are studied via simulation. An application to Finnish occupational disease registration data is presented.

1. Introduction

We consider the problem of estimating the size of a closed population based on a capture and a single recapture (e.g., Seber, 1982, Chap. 3; Seber, 1986, p. 273). In demography this is known as *dual-system estimation* (e.g., Ericksen and Kadane, 1985, pp. 102–103). Several authors have studied the problems caused by heterogeneity in capture probabilities (Seber, 1982, pp. 85–88). Burnham and Overton (1978, 1979) and Otis et al. (1978) postulated a model of unobservable heterogeneity in which the individual capture probabilities are a random sample from an unknown distribution. A jackknife estimator based on several recaptures is used to estimate the population size. This work has been extended by, e.g., Pollock and Otto (1983) and Chao (1987, 1988), who study the bias, variance, and robustness of alternative estimators. Rodrigues, Bolfarine, and Leite (1988) propose a Bayesian analysis using both noninformative and informative priors. Closer to our contribution is Pollock, Hines, and Nichols (1984), who introduced a logistic regression technique to account for observable population heterogeneity in the capture probabilities. In other words, the characteristics of the captured individuals are used to explain their probabilities of capture. To avoid problems connected with the unobservable part of the likelihood (due to those members of the population that are not captured at all), they categorized the independent variables to carry out the estimation (Pollock et al., 1984, p. 332). We circumvent these problems by conditioning (cf. Sanathanan, 1972; Bishop, Fienberg, and Holland, 1975, Chap. 6; Seber, 1982, pp. 489–490). This allows us to use independent variables without any grouping. Huggins (1989) has independently suggested a similar approach to the problem.

In Section 2 we generalize the classical estimator of population size to cover, for instance, the case in which all capture probabilities are different between individuals and captures. In Section 3 we develop the conditional maximum likelihood estimation procedures and establish sufficient conditions for the strong consistency and asymptotic normality of the proposed estimator. Part of the material is somewhat technical and can be skipped by a

Key words: Asymptotic theory; Capture–recapture models; Logistic regression; Simulation.

reader interested in applications. In Section 4 we generalize the variance formula of Sekar and Deming (1949) to cover our model. The finite-sample properties of our estimators are investigated via simulation in Section 5. We show that the analysis based on logistic regression corrects for the bias caused by observable population heterogeneity. However, in very small populations the estimator becomes unstable. Finally, in Section 6 we present an application to Finnish occupational disease data.

The notation used in capture–recapture/dual registration literature is not standard (cf. Cormack, 1968, p. 457), and the existing loose conventions (e.g., in Seber, 1982) differ from those used in many other parts of statistics. An attempt is made here to formulate the results in a language familiar in the capture–recapture literature.

2. Population Heterogeneity

2.1 *Classical Model*

Suppose we sample twice from a closed population of unknown size N. Let n_1 be the number of individuals captured the first time, n_2 the number captured the second time, and m the number captured twice. Let p_1 be the probability of capture on the first occasion, p_2 the probability of capture on the second occasion, and p_{12} the probability of being captured twice. We assume that the captures are *independent*, so that $p_{12} = p_1 p_2$.

Define $u_1 = n_1 - m$, $u_2 = n_2 - m$, and $M = n_1 + n_2 - m$. Assuming that different individuals are registered independently of each other, we have a *multinomial* model (e.g., Seber, 1986, p. 274)

$$(u_1, u_2, m, N - M) \sim \text{Mult}(N; p_1(1 - p_2); (1 - p_1)p_2; p_1 p_2; 1 - \phi),$$

where $\phi = p_1 + p_2 - p_1 p_2$. The classical model is often phrased conditionally on the values of n_1 and n_2. This leads to a *hypergeometric* distribution for m, with the well-known maximum likelihood estimators

$$\hat{p}_1 = \frac{m}{n_2}, \quad \hat{p}_2 = \frac{m}{n_1}, \quad \hat{N} = \frac{n_1 n_2}{m},$$

(e.g., Feller, 1968, pp. 45–46). Sekar and Deming (1949, pp. 114–115) derived an estimator for the asymptotic variance of $\hat{N}$ in the hypergeometric setting:

$$V_1 = \frac{n_1 n_2 u_1 u_2}{m^3}.$$

We shall show that $\hat{N}$ is maximum likelihood, and derive V_1 under the multinomial model as a special case in Examples 3.1 and 4.2.

We shall now extend the model to allow for variation in the individual probabilities of registration, by treating each individual as a separate stratum. Define indicator variables u_{1i}, u_{2i}, and m_i, for $i = 1, \ldots, N$,

$$u_{ji} = \begin{cases} 1, & \text{if individual } i \text{ is captured on occasion } j \text{ only, } j = 1, 2; \\ 0, & \text{otherwise;} \end{cases}$$

$$m_i = \begin{cases} 1, & \text{if individual } i \text{ is captured twice;} \\ 0, & \text{otherwise.} \end{cases}$$

Let $n_{ji} = u_{ji} + m_i$ $(j = 1, 2)$, $M_i = u_{1i} + u_{2i} + m_i$, and define for each individual the probabilities of being registered as $p_{ji} = \text{E}[n_{ji}]$ $(j = 1, 2)$, and $p_{12i} = \text{E}[m_i]$. Assume that these probabilities are strictly between 0 and 1. We shall complete the definition of the

model allowing for population heterogeneity, by assuming that the registers operate *independently* on the individual level, or $p_{12i} = p_{1i}p_{2i}$, and that the multinomial vectors

$$(u_{1i}, u_{2i}, m_i, 1 - M_i) \sim \text{Mult}(1; p_{1i}(1 - p_{2i}); (1 - p_{1i})p_{2i}; p_{1i}p_{2i}; 1 - \phi_i),$$

where $\phi_i = p_{1i} + p_{2i} + p_{1i}p_{2i}$, are *independent* for $i = 1, \ldots, N$. As we shall see below, this model allows for a population-level correlation between the captures.

It is well known both empirically (Seber, 1982, p. 565) and theoretically (e.g, Burnham and Overton, 1979, Table 4, pp. 931–932) that the classical estimator may be severely biased under population heterogeneity. The expected bias can be expressed in terms of the sample covariance of the pairs (p_{1i}, p_{2i}), $i = 1, \ldots, N$: $\hat{N}$ gives asymptotically an underestimate if the covariance is positive, and an overestimate if it is negative. Under zero covariance the estimator is consistent (cf. Sekar and Deming, 1949, pp. 105–106; Seber, 1982, p. 86).

2.2 *Estimation Under Observable Heterogeneity*

Suppose for the moment that we know the probabilities of being captured at least once, ϕ_i, and consider the estimator

$$\tilde{N} = \sum_{i=1}^{N} \frac{M_i}{\phi_i} = \sum_{M_i=1} \frac{1}{\phi_i}.$$

The summation on the right means summation over those indices i for which $M_i = 1$. $\tilde{N}$ is obviously an unbiased estimator of N. One can also show (using sufficiency and completeness) that it has minimum variance among such estimators. However, for our purposes it is more important to note that $\tilde{N}$ can be calculated when ϕ_i *is known only for those individuals that have been captured at least once.* This means that if we can estimate ϕ_i's from the data concerning the *captured* individuals, then we can replace $\tilde{N}$ by its estimator, and get an estimator of N (cf. Sanathanan, 1972, p. 144). This is precisely what we shall do in Section 3.

However, some regularity conditions must be imposed on the true underlying ϕ_i's to guarantee the consistency of $\tilde{N}$ in large samples. Despite the fact that $\tilde{N}$ is always unbiased, its variance may explode unless the ϕ_i's are bounded in some way. We shall now prove a technical result that gives a sufficient condition for consistency. The result will also be used in the proof of Proposition 3.2.

Proposition 2.1 Suppose there is a constant $q > 0$ such that $q \leqslant \phi_i$ for all $i = 1, 2, \ldots$. Then $\tilde{N}$ is strongly consistent for N,

$$\frac{\tilde{N}}{N} \to 1 \quad \text{a.s. as } N \to \infty.$$

If in addition $\phi_i \leqslant 1 - q$ for all $i = 1, 2, \ldots$, then $\tilde{N}$ has an asymptotic normal distribution.

Proof Since $\text{var}(M_i/\phi_i) = (1 - \phi_i)/\phi_i \leqslant (1 - q)/q$, the strong law of large numbers (e.g., Chung, 1974, Theorem 5.1.2, p. 103) implies that $\tilde{N}/N \to 1$ almost surely. As a sum of independent variables with variances bounded from above and away from zero, $\tilde{N}$ is asymptotically normal.

This gives a sufficient condition for the strong consistency of $\tilde{N}$. It does exclude certain practical situations from consideration, such as the case in which part of the population is effectively uncatchable, i.e., it has positive, but very small capture probabilities. [Seber (1982, p. 72) relates an example due to Ayre of an ant population for which such conditions

hold: Only a fraction of ants go foraging, for others remain in the ant hill and so are uncatchable.] Then $\tilde{N}$ is still unbiased, but it may not be consistent. Furthermore, by considering, e.g., $\phi_i = i^{-\delta}$ $(0 < \delta;\ i = 1, 2, \ldots)$, we have that $\text{var}(\tilde{N}/N) = O(N^{\delta-1})$. For $\delta \geq 1$ the variance does not vanish as $N \to \infty$, so we do not get even weak consistency. For $0 < \delta < 1$, the terms $\text{var}(M_i/\phi_i)i^{-2} = i^{\delta-2} - i^{-2}$ $(i = 1, 2, \ldots)$ form a convergent series, so the strong consistency follows from Kolmogorov's strong law of large numbers (Chung, 1974, Corollary, p. 125). These examples show that the existence of a bound $q > 0$ is not necessary for Proposition 2.1 to hold, but some bound for the frequency of extreme values of ϕ_i is essential.

In practice the ϕ_i's would not be known. Instead, we shall suppose that there is a parameter vector $\boldsymbol{\theta}$ such that $p_{1i} = p_{1i}(\boldsymbol{\theta})$ and $p_{2i} = p_{2i}(\boldsymbol{\theta})$. We shall write $\hat{p}_{1i}$, $\hat{p}_{2i}$, $\hat{\phi}_i$, and $\hat{N}$, for p_{1i}, p_{2i}, ϕ_i, and $\tilde{N}$, when $\boldsymbol{\theta}$ has been estimated by a maximum likelihood estimator $\hat{\boldsymbol{\theta}}$. Under population homogeneity

$$\hat{N} = \sum_{M_i=1} \frac{1}{\hat{\phi}_i} = \sum_{M_i=1} \frac{1}{\phi_i(\hat{\boldsymbol{\theta}})}$$

will agree with the classical estimator so that the same symbol can be used for the estimator we have introduced.

3. Logistic Analysis

We shall now derive a conditional maximum likelihood estimator of $\boldsymbol{\theta}$ under logistic regression. The iterative formulas for the solution of the likelihood equations are given below, before Example 3.1, so a reader interested in applications may want to proceed there. Assume we have vectors $\mathbf{X}_{1i} = (X_{1i1}, \ldots, X_{1ik})^{\mathrm{T}}$ and $\mathbf{X}_{2i} = (X_{2i1}, \ldots, X_{2ih})^{\mathrm{T}}$ of "explanatory" variables giving the characteristics of individual i relevant to capture, with $X_{1i1} = X_{2i1} = 1$, for $i = 1, \ldots, N$. We model the p_{ji}'s by letting $\log(p_{ji}/(1 - p_{ji})) = \mathbf{X}_{ji}^{\mathrm{T}}\mathbf{a}_j$ $(j = 1, 2)$, where $\mathbf{a}_1 = (a_{11}, \ldots, a_{1k})^{\mathrm{T}}$ and $\mathbf{a}_2 = (a_{21}, \ldots, a_{2h})^{\mathrm{T}}$ are vectors of parameters.

Let $\mathbf{M} = (M_1, \ldots, M_N)^{\mathrm{T}}$, and define $\mathbf{u}_1$, $\mathbf{u}_2$, and $\mathbf{m}$ correspondingly. The *conditional likelihood* of $\boldsymbol{\theta} = (\mathbf{a}_1^{\mathrm{T}}, \mathbf{a}_2^{\mathrm{T}})^{\mathrm{T}}$, given $\mathbf{M}$, can (with some algebra) be shown to be

$$L(\boldsymbol{\theta} \mid \mathbf{u}_1, \mathbf{u}_2, \mathbf{m}; \mathbf{M}) = \prod_{M_i=1} \Pr(u_{1i}, u_{2i}, m_i \mid M_i = 1),$$

where

$$\Pr(u_{1i}, u_{2i}, m_i \mid M_i = 1) = \frac{\exp(u_{1i}\mathbf{X}_{1i}^{\mathrm{T}}\mathbf{a}_1 + u_{2i}\mathbf{X}_{2i}^{\mathrm{T}}\mathbf{a}_2 + m_i(\mathbf{X}_{1i}^{\mathrm{T}}\mathbf{a}_1 + \mathbf{X}_{2i}^{\mathrm{T}}\mathbf{a}_2))}{K_i(\boldsymbol{\theta})},$$

with

$$K_i(\boldsymbol{\theta}) = \exp(\mathbf{X}_{1i}^{\mathrm{T}}\mathbf{a}_1) + \exp(\mathbf{X}_{2i}^{\mathrm{T}}\mathbf{a}_2) + \exp(\mathbf{X}_{1i}^{\mathrm{T}}\mathbf{a}_1 + \mathbf{X}_{2i}^{\mathrm{T}}\mathbf{a}_2).$$

Consequently, we can write

$$L(\boldsymbol{\theta} \mid \mathbf{u}_1, \mathbf{u}_2, \mathbf{m}; \mathbf{M}) = \exp(\mathbf{T}_1^{\mathrm{T}}\mathbf{a}_1 + \mathbf{T}_2^{\mathrm{T}}\mathbf{a}_2) \prod_{M_i=1} K_i(\boldsymbol{\theta})^{-1},$$

where

$$\mathbf{T}_j = \sum_{M_i=1} n_{ji}\mathbf{X}_{ji}, \quad j = 1, 2.$$

This shows that the conditional distribution of $\mathbf{u}_1$, $\mathbf{u}_2$, and $\mathbf{m}$, given $\mathbf{M}$, belongs to the exponential family of degree $k + h$, and the vectors $\mathbf{T}_1$ and $\mathbf{T}_2$ form the minimal sufficient

statistic for $\boldsymbol{\theta}$ (Andersen, 1980, p. 28). This is also a generalized linear model with the natural link function (Nelder and Wedderburn, 1972, p. 372; Fahrmeir and Kaufmann, 1985, p. 345). It follows (cf. Andersen, 1980, Theorem 3.1, p. 56, for the i.i.d. case) that the likelihood equations are

$$\mathrm{E}[\mathbf{T}_j \mid \mathbf{M}] = \mathbf{t}_j, \quad j = 1, 2,$$

where $\mathbf{t}_j$ is the observed value of $\mathbf{T}_j$ ($j = 1, 2$). Let us write

$$\mathbf{T} = \begin{bmatrix} \mathbf{T}_1 \\ \mathbf{T}_2 \end{bmatrix}, \quad \mathbf{t} = \begin{bmatrix} \mathbf{t}_1 \\ \mathbf{t}_2 \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 \end{bmatrix},$$

where $\mathbf{X}_j^{\mathrm{T}} = (X_{j1}, \ldots, X_{jN})$, $j = 1, 2$. Write also $\mathbf{Y} = (n_{11}, \ldots, n_{1N}, n_{21}, \ldots, n_{2N})^{\mathrm{T}}$, and note that $\mathbf{T} = \mathbf{X}^{\mathrm{T}}\mathbf{Y}$. The likelihood equations can now be written as

$$\mathbf{t} - \mathbf{X}^{\mathrm{T}}\mathrm{E}[\mathbf{Y} \mid \mathbf{M}] = \mathbf{0},$$

where for $i = 1, \ldots, N$ we have $\mathrm{E}[Y_i \mid M_i = 1] = \Pr(1, 0, 0 \mid M_i = 1) + \Pr(0, 0, 1 \mid M_i = 1)$, with $\Pr(u_{1i}, u_{2i}, m_i \mid M_i = 1)$ as above; for $i = N + 1, \ldots, 2N$, we have $\mathrm{E}[Y_i \mid M_i = 1] = \Pr(0, 1, 0 \mid M_i = 1) + \Pr(0, 0, 1 \mid M_i = 1)$. Note that for all i, we have $\mathrm{E}[Y_i \mid M_i = 0] = 0$, so the $\mathbf{X}_{ji}$'s belonging to unobserved individuals do not enter into the likelihood equations even through they are formally included in the formulas.

To solve the equations we need $\mathrm{cov}(\mathbf{T} \mid \mathbf{M}) = \mathbf{X}^{\mathrm{T}}\mathrm{cov}(\mathbf{Y} \mid \mathbf{M})\mathbf{X}$ (cf. Andersen, 1980, proof of Lemma 3.3, p. 59, and Theorem 3.4, p. 65, for the i.i.d. case). Components of $\mathbf{Y}$ relating to different individuals are independent by assumption. However, conditionally on M_i, Y_i is dependent on Y_{N+i}, $i = 1, \ldots, N$. It follows that $\mathrm{cov}(\mathbf{Y} \mid \mathbf{M}) \equiv \mathbf{W}$ is of the form

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_1 & \mathbf{W}_3 \\ \mathbf{W}_4 & \mathbf{W}_2 \end{bmatrix},$$

where $\mathbf{W}_j$'s are diagonal $N \times N$ matrices. To define their elements, denote $\tilde{\mathbf{W}}_j = (\tilde{w}_{ik}^j)$ ($j = 1, \ldots, 4$; $i, k = 1, \ldots, N$), where

$$\tilde{w}_{ii}^j = \mathrm{var}(n_{ji} \mid M_i = 1) = \frac{p_{ji}}{\phi_i} - \frac{p_{ji}^2}{\phi_i^2}, \quad j = 1, 2,$$

$$\tilde{w}_{ii}^3 = \tilde{w}_{ii}^4 = \mathrm{cov}(n_{1i}, n_{2i} \mid M_i = 1) = \frac{p_{1i}p_{2i}}{\phi_i} - \frac{p_{1i}p_{2i}}{\phi_i^2}.$$

Now define the elements of $\mathbf{W}_j = (w_{ik}^j)$ by taking $w_{ik}^j = M_i M_k \tilde{w}_{ik}^j$.

With these notations Newton's method yields the recursion

$$\boldsymbol{\theta}_{s+1} = \boldsymbol{\theta}_s + (\mathbf{X}^{\mathrm{T}}\mathbf{W}_s\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}(\mathbf{Y} - \mathrm{E}_s[\mathbf{Y} \mid \mathbf{M}]), \quad s = 0, 1, \ldots,$$

where $s = 0$ corresponds to an initial value we use to start the iteration, and $\mathbf{W}_s$ and $\mathrm{E}_s[\mathbf{Y} \mid \mathbf{M}]$ contain the estimates of $\mathbf{W}$ and $\mathrm{E}[\mathbf{Y} \mid \mathbf{M}]$ based on $\boldsymbol{\theta}_s$. Using the "working variate"

$$\mathbf{g}_s = \mathbf{X}\boldsymbol{\theta}_s + \mathbf{W}_s^{-1}(\mathbf{Y} - \mathrm{E}_s[\mathbf{Y} \mid \mathbf{M}])$$

renders the recursion into the familiar regression form

$$\boldsymbol{\theta}_{s+1} = (\mathbf{X}^{\mathrm{T}}\mathbf{W}_s\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}_s\mathbf{g}_s.$$

This can be implemented by running regressions with any statistical package that allows weighting of observations in regression or that has matrix algebra.

An estimator for the covariance matrix of $\hat{\boldsymbol{\theta}}$ is

$$\widehat{\text{cov}}(\hat{\boldsymbol{\theta}} \mid \mathbf{M}) = (\mathbf{X}^{\mathrm{T}}\hat{\mathbf{W}}\mathbf{X})^{-1},$$

where $\hat{\mathbf{W}}$ is the matrix $\mathbf{W}_s$ corresponding to $\boldsymbol{\theta}_s = \hat{\boldsymbol{\theta}}$.

Example 3.1 Assume $k = h = 1$, i.e., there is no population heterogeneity. Then $t_1 = n_1$ and $t_2 = n_2$. For $i = 1, \ldots, N$, we have $\mathrm{E}[Y_i \mid M_i = 1] = p_1/\phi$, and for $i = N + 1, \ldots, 2N$, we have $\mathrm{E}[Y_i \mid M_i = 1] = p_2/\phi$. Conditionally on M, the maximum likelihood estimators of these are n_1/M and n_2/M because they satisfy the likelihood equations. Solving for p_1 and p_2 gives us $\hat{p}_1 = m/n_2$ and $\hat{p}_2 = m/n_1$ as the maximum likelihood estimators. The estimator for $1/\phi$ is $(n_1/M)(n_2/m)$, and consequently $\hat{N} = n_1 n_2/m$, or the classical estimator.

Example 3.2 Suppose $k = 2$ and $h = 1$, i.e., there is no heterogeneity at second capture time. Two of the three likelihood equations are

$$n_1 = \sum_{M_i=1} \frac{p_{1i}}{\phi_i}, \quad n_2 = \sum_{M_i=1} \frac{p_2}{\phi_i} = p_2\tilde{N},$$

when we write $p_{2i} = p_2$ for all i. Since $p_{1i} = (\phi_i - p_2)/(1 - p_2)$, the first equation becomes $n_1 = (M - p_2\tilde{N})/(1 - p_2)$. Solving these equations gives $\hat{p}_2 = m/n_1$ and $\hat{N} = n_1 n_2/m$. Consequently, the classical estimator $\hat{p}_2$ is conditionally maximum likelihood even when there is heterogeneity of the probabilities of n_{1i}'s. Similarly, our proposed estimator for N agrees with the classical one.

The existence, consistency, and asymptotic normality of the maximum likelihood estimator follow from the theory of generalized linear models under suitable conditions. One set of sufficient conditions is given below. The proofs can be found in the Appendix.

Proposition 3.1 Assume that the elements of $\mathbf{X}$ are bounded, and that $\mathbf{X}^{\mathrm{T}}\mathbf{X}/N$ converges to a positive-definite matrix, as $N \to \infty$. Then $\hat{\boldsymbol{\theta}} \to \boldsymbol{\theta}$ a.s. and $\hat{\boldsymbol{\theta}}$ has an asymptotic normal distribution that does not depend on $\mathbf{M}$ a.s.

Proposition 3.2 Under the conditions of Proposition 3.1, $\hat{N}$ is strongly consistent for N,

$$\frac{\hat{N}}{N} \to 1 \quad \text{a.s. as } N \to \infty,$$

and it has an asymptotic normal distribution.

An advantage of the logistic analysis over the stratified analyses proposed by Sekar and Deming (1949, pp. 106–107) is that we can use continuous explanatory variables (such as age) in our models. Second, we have the standard theory of exponential families at our disposal for inference. Even in the situation of Example 3.2 the logistic analysis may be helpful in understanding the underlying capture mechanisms. The logistic analysis provides also a simple means of estimating the distribution of explanatory variables in the population of interest, such as age distribution in a human population, or, say, the distribution of "size" (weight, length, . . .) in a fish population.

4. Variance Estimation

We shall first derive an estimator of the conditional asymptotic variance of $\hat{N}$, given $\mathbf{M}$. This estimator does not account for the variability in $\mathbf{M}$ itself. Then we present an approximation to the unconditional variance. The resulting estimator can be thought of as

a generalization of V_1 introduced by Sekar and Deming (1949). We saw in Section 3 that the conditional maximum likelihood estimators combined with the estimator $\tilde{N}$ of Section 2 result in estimators of N that are consistent and asymptotically normal. The unconditional variance estimator derived below will allow us to present *unconditional* confidence intervals for N under population heterogeneity even though a conditional likelihood was used in the estimation of $\boldsymbol{\theta}$.

Let us make the dependency of $\hat{N}$ on $\boldsymbol{\theta}$ explicit by writing $\hat{N} = \tilde{N}(\hat{\boldsymbol{\theta}}) = \tilde{N}_1(\hat{\boldsymbol{\theta}}) + \cdots + \tilde{N}_N(\hat{\boldsymbol{\theta}})$, where $\tilde{N}_i(\hat{\boldsymbol{\theta}}) = M_i/\phi_i(\hat{\boldsymbol{\theta}})$. We shall calculate the conditional asymptotic variance of $\tilde{N}(\hat{\boldsymbol{\theta}})$, given $\mathbf{M}$. For $i = 1, \ldots, N$, define the vectors

$$\mathbf{V}_i(\boldsymbol{\theta}) = \left(\frac{\partial \tilde{N}_i}{\partial \theta_1}(\boldsymbol{\theta}), \ldots, \frac{\partial \tilde{N}_i}{\partial \theta_{k+h}}(\boldsymbol{\theta})\right)^{\mathrm{T}},$$

and let $\mathbf{V}(\boldsymbol{\theta}) = \mathbf{V}_1(\boldsymbol{\theta}) + \cdots + \mathbf{V}_N(\boldsymbol{\theta})$. The first-degree Taylor approximation gives the asymptotic variance of $\tilde{N}(\hat{\boldsymbol{\theta}})$, given $\mathbf{M}$, as

$$\operatorname{var}(\tilde{N}(\hat{\boldsymbol{\theta}}) \mid \mathbf{M}) = \sum_{i,j=1}^{N} \mathbf{V}_i(\boldsymbol{\theta})^{\mathrm{T}} \operatorname{cov}(\hat{\boldsymbol{\theta}} \mid \mathbf{M}) \mathbf{V}_j(\boldsymbol{\theta}) = \mathbf{V}(\boldsymbol{\theta})^{\mathrm{T}} \operatorname{cov}(\hat{\boldsymbol{\theta}} \mid \mathbf{M}) \mathbf{V}(\boldsymbol{\theta}).$$

A straightforward calculation shows that, for $i = 1, \ldots, N$, and $j = 1, \ldots, k$,

$$\frac{\partial \tilde{N}_i}{\partial \theta_j}(\boldsymbol{\theta}) = -X_{1ij}\psi_i(\boldsymbol{\theta}),$$

where $\psi_i(\boldsymbol{\theta}) = M_i \exp(\mathbf{X}_{1i}^{\mathrm{T}}\mathbf{a}_1)(1 + \exp(\mathbf{X}_{2i}^{\mathrm{T}}\mathbf{a}_2))/K_i(\boldsymbol{\theta})^2$. Similarly, for $j = k + 1, \ldots, k + h$, we have

$$\frac{\partial \tilde{N}_i}{\partial \theta_j}(\boldsymbol{\theta}) = -X_{2i,j-k}\psi_{N+i}(\boldsymbol{\theta}),$$

where $\psi_{N+i}(\boldsymbol{\theta}) = M_i \exp(\mathbf{X}_{2i}^{\mathrm{T}}\mathbf{a}_2)(1 + \exp(\mathbf{X}_{1i}^{\mathrm{T}}\mathbf{a}_1))/K_i(\boldsymbol{\theta})^2$. Let $\boldsymbol{\psi}(\boldsymbol{\theta}) = (\psi_1(\boldsymbol{\theta}), \ldots, \psi_{2N}(\boldsymbol{\theta}))^{\mathrm{T}}$. Then we can write

$$\mathbf{V}(\boldsymbol{\theta}) = -\mathbf{X}^{\mathrm{T}}\boldsymbol{\psi}(\boldsymbol{\theta}),$$

and our formula for the estimator V_2 of the conditional asymptotic variance $\operatorname{var}(\tilde{N}(\hat{\boldsymbol{\theta}}) \mid \mathbf{M})$ becomes

$$V_2 = \boldsymbol{\psi}(\hat{\boldsymbol{\theta}})^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\hat{\mathbf{W}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{\psi}(\hat{\boldsymbol{\theta}}),$$

where $\hat{\mathbf{W}}$ is as in Section 3.

To show that conditioning on $\mathbf{M}$ reduces the variability of $\tilde{N}$, let us consider the case of no population heterogeneity.

Example 4.1 We saw in Example 3.1 that under population homogeneity, our point estimators coincide with the classical ones. In this case $\mathbf{X}_1 = \mathbf{X}_2 =$ vector of N ones. A direct substitution into the formula for V_2 yields after some algebra that

$$V_2 = V_1 \frac{u_1 + u_2}{M},$$

or $V_2 < V_1$.

We shall now derive an estimator for the unconditional asymptotic variance of $\hat{N}$. Conditioning on $\mathbf{M}$, we have

$$\operatorname{var}(\hat{N}) = \mathrm{E}[\operatorname{var}(\hat{N} \mid \mathbf{M})] + \operatorname{var}(\mathrm{E}[\hat{N} \mid \mathbf{M}]).$$

We estimate $E[\text{var}(\hat{N} \mid \mathbf{M})]$ by V_2. To estimate the latter term, note that

$$E[\hat{N} \mid \mathbf{M}] = \sum_{i=1}^{N} M_i E\left[\frac{1}{\hat{\phi}_i} \middle| \mathbf{M}\right].$$

We shall first approximate $E[1/\hat{\phi}_i \mid \mathbf{M}]$ by its limit value $1/\phi_i$, so

$$\text{var}(E[\hat{N} \mid \mathbf{M}]) \approx \sum_{i=1}^{N} \frac{\phi_i(1 - \phi_i)}{\phi_i^2}.$$

Estimating the first ϕ_i by M_i, and the remaining ϕ_i's by $\hat{\phi}_i$, gives us an estimator V_3, which can be written as

$$V_3 = \sum_{M_i=1} \frac{1 - \hat{\phi}_i}{\hat{\phi}_i^2}.$$

Combining the results, we get the unconditional estimator V_0 of $\text{var}(\hat{N})$ as

$$V_0 = V_2 + V_3.$$

Example 4.2 A direct calculation shows that under population homogeneity,

$$V_3 = M\left(1 - \frac{mM}{n_1 n_2}\right)\left(\frac{n_1 n_2}{mM}\right)^2 = V_1 \frac{m}{M}.$$

Together with Example 4.1, this shows that $V_0 = V_2 + V_3 = V_1$, so that our unconditional estimator agrees with the classical one under population homogeneity.

5. Finite-Sample Properties

We conducted a simulation study to see how the logistic analysis compares with the classical one in terms of bias, variance, and accuracy of confidence intervals, in small samples. First, we assumed that the probability mechanism generating population heterogeneity was correctly specified. Then the case of missing covariates was considered.

Three unknown population sizes, $N = 100$, $N = 300$, and $N = 1{,}000$, were considered. One covariate $X \sim N(0, 1)$ was used to generate the observations under two models.

Under Model I the n_{1i}'s were generated by taking $\text{logit}(p_{1i}) = .5 + .8X_i$, and the n_{2i}'s by taking $\text{logit}(p_{2i}) = 1.5 + .4X_i$, $i = 1, \ldots, N$. This implies that we have $E[n_1] = .608N$, $E[n_2] = .810N$, $E[m] = .503N$, and $E[M] = .915N$. In other words, about 92% of the unknown population are expected to be registered by at least one of the registers. The distributions of the p_{1i}'s and p_{2i}'s are slightly skewed to the left with standard deviations .18 and .07, respectively.

Under Model II we took $\text{logit}(p_{1i}) = -.5 + .8X_i$ and $\text{logit}(p_{2i}) = -1.0 + .4X_i$. Then we have $E[n_1] = .392N$, $E[n_2] = .276N$, $E[m] = .121$, and $E[M] = .547N$. This time the distributions of p_{1i}'s and p_{2i}'s are slightly skewed to the right with standard deviations .17 and .08, respectively.

Under both models the pairs (p_{1i}, p_{2i}) are essentially perfectly correlated. Based on the bias results referred to in Section 2.1, we expect the classical estimator to be about 2.5% downward biased under Model I and about 11.2% downward biased under Model II.

The simulations were carried out as follows. (1) N independent observations from $N(0, 1)$ were generated. (2) n_{1i} and n_{2i} were generated independently under Model I for $i = 1, \ldots, N$. (3) The classical estimator for N and its variance were calculated. (4) Newton's method was used to estimate the logistic regression models, as outlined in Section 3, and an unconditional variance estimator was calculated using formulas of Section 4.

Steps (1)–(4) were repeated 600 times for $N = 100$, $N = 300$, and $N = 1{,}000$. After that the same was done with Model II. The calculations were carried out using MINITAB on a personal computer.

Table 1 presents results from these simulations. Under both Models I and II the classical estimator $\hat{N}$ is downward biased but slightly less so than the first-order asymptotics would imply. Its standard deviation is adequately estimated by $\sqrt{V_1}$, and plots (not shown here) indicate that its distribution is very close to normal. Due to the bias, the purported "95% confidence intervals" for N do not come close to reaching the nominal level of coverage.

In contrast, the "95% confidence intervals" based on the *logistic estimator*, denoted by $\hat{N}'$, are very nearly adequate for both models and all three values of N. However, the way this is accomplished needs a closer look.

The logistic analysis corrects for the bias of $\hat{N}$. We pay for this in the increased variance of the estimator. Under Model I, the classical estimator $\hat{N}$ and the proposed logistic estimator $\hat{N}'$ have approximately the same mean squared errors (MSE) for $N = 300$. For $N = 100$ the MSE of $\hat{N}'$ is larger, and for $N = 1{,}000$ it is smaller than that for $\hat{N}$. For Model II the break-even point of MSEs is further between $N = 300$ and $N = 1{,}000$. A comparison of the mean and the median of $\hat{N}'$ for $N = 100$ under Model II indicates that its distribution is highly skewed to the right. As N increases, the skewness decreases.

The results of Table 1 do not give a direct indication of how well the *method* based on logistic regression might perform in a real situation, in which one would test whether the coefficients a_{12} and a_{22} vanish. If either one would be deemed not to be significantly different from 0, then, in view of Example 3.2, a classical analysis could be performed. Any reasonable testing strategies should give results that fall between those obtained using exclusively $\hat{N}$ or $\hat{N}'$, both in terms of bias and the coverage of confidence intervals.

Additional simulations (600 repetitions) were carried out with Model II and $N = 1{,}000$ to study the effect of model misspecification. The situation was modified by replacing X by $(X' + X'')/\sqrt{2}$, where X' and X'' are independent $N(0, 1)$ variables, and by assuming that only X' was observable. From the point of view of classical estimation this situation

Table 1

Small-sample properties of the classical population size estimator $\hat{N}$ and the estimator applying logistic regression $\hat{N}'$, based on 600 *simulations of Models* I *and* II *for* $N = 100, 300, 1{,}000$.

		Model I		Model II	
	N	$\hat{N}$	$\hat{N}'$	$\hat{N}$	$\hat{N}'$
Average of estimates divided by N	100	.981	1.011	.931	1.434
	300	.980	1.004	.898	1.044
	1,000	.980	1.002	.894	1.010
Median of estimates divided by N	100	.982	1.002	.896	1.021
	300	.980	1.003	.887	1.006
	1,000	.981	1.001	.891	.999
Standard deviation of estimator divided by N	100	.037	.063	.207	2.723
	300	.022	.030	.096	.191
	1,000	.012	.016	.053	.088
Average of estimated standard deviations of estimator divided by N	100	.035	.058	.185	1.082
	300	.020	.031	.095	.183
	1,000	.011	.016	.051	.087
Coverage probability of "95% confidence intervals"	100	.845	.933	.802	.933
	300	.798	.953	.703	.945
	1,000	.582	.975	.432	.955

is identical to Model II. However, the proposed logistic regression procedure performed worse than under Model II, since only 50% of the population heterogeneity was observable: average/1,000 = .948; median/1,000 = .943; standard deviation/1,000 = .071; average of estimated standard deviations/1,000 = .067; coverage of "95% confidence intervals" = .802.

A further effect of model misspecification is that the parameter estimates become biased. Under the original Model II we had $a_{12} = .8$ and $a_{22} = .4$. When the model was correctly specified, the average of $\hat{a}_{12}$'s was .809, and the average of $\hat{a}_{22}$'s was .399. Under the modified model we had $a_{12} = .8/\sqrt{2} = .566$, and $a_{22} = .4/\sqrt{2} = .283$. In this case the average of $\hat{a}_{12}$'s was .524, and the average of $\hat{a}_{22}$'s was .265. Their empirical standard errors were .005 and .004, respectively. This bias appears analogous to the effect of "errors in explanatory variables" in ordinary regression.

So far, all our simulations have been concerned with the case of positive correlation between the captures. Since a negative correlation is also a possibility, we generated additional data in the set-up where $N = 1{,}000$ and Model II has been modified to have $\text{logit}(p_{2i}) = 1.0 - .4X_i$. In other words, the sign of the coefficient of X has been reversed. Based on Section 2.1, the classical estimator was expected to be about 14% *upward* biased. In 600 simulations this turned out to be the case. The logistic estimator was again nearly unbiased in terms of both mean and median. The coverage probabilities of the "95% confidence intervals" were .652 for the classical method and .948 for the proposed method. Interestingly, in this case $\sqrt{\text{var}(\hat{N})} = 86.7 > \sqrt{\text{var}(\hat{N}')} = 74.7$. The estimators of both variances were slightly downward biased. In this case both the bias *and* the variance of the classical estimator are larger than those of the logistic estimator.

6. An Application to Occupational Disease Registration

We illustrate the logistic analysis by an application to occupational disease data from Finland in 1981. The Finnish Register of Occupational Diseases was founded in 1964. At that time it was agreed that accident insurance companies would report all new cases of occupational disease to the Register, irrespective of the compensation decision.

From 1975 every physician has been required to report all new cases of occupational disease directly to a government agency, which reports the cases to the Register. Despite the legal obligation, physicians neglect to report a large number of cases, presumably because such a report causes paperwork, but does not directly benefit the patient. Unfortunately, all cases are not reported via the other channel either.

The Finnish Register of Occupational Diseases can, thus, be thought of as a dual registration system. However, the probabilities of registration are not constant over different types of cases, nor are the two information channels independent (on a population level). Alho (paper presented at 11th Nordic Conference on Mathematical Statistics, Uppsala, 1986) has shown previously that diagnosis has a strong impact on the probabilities of registration. Using data from 1981, we found $M = 5{,}231$. Stratifying the data into four groups of diagnoses, (1) noise-induced hearing loss, (2) diseases of the musculo-skeletal system caused by repetitive or monotonous work, (3) skin diseases, and (4) other diseases, yielded $\hat{N} = 8{,}258$, as opposed to $\hat{N} = 7{,}232$ obtained without stratification.

Using the logistic techniques, we can check whether these findings hold when we control for the possible effect of age. Age at which a disease is diagnosed may be correlated with the severity of the case, making cases of older workers more likely to be reported through either channel. On the other hand, one may argue that older workers are less likely to have their diseases diagnosed as occupational ones due to fear of losing one's job at an advanced age. A logistic analysis with age as an explanatory variable might, thus, reveal additional population heterogeneity within the groups of diagnoses. This turned out to be the case for

noise-induced hearing loss and the "other" category, but in both cases the heterogeneity pertained to only one channel, so it did not influence $\hat{N}$.

As an example, let us look at noise-induced hearing loss, which had $M = 1{,}854$ in 1981. Let p_1 = probability that a case of noise-induced hearing loss is reported to the Register from insurance companies, and p_2 = the corresponding probability for the other channel. Let X = age. We estimated the model $\text{logit}(p_1) = a_{11} + a_{12}X$, and $\text{logit}(p_2) = a_{21} + a_{22}X$, but found $\hat{a}_{22}$ not to be significant at 5% level. Taking $a_{22} = 0$ gave $\hat{a}_{11} = -1.543$, $\hat{a}_{12} = .0438$ (estimated standard error = .0020), and $\hat{a}_{21} = .0409$. The range of p_{1i}'s was from .359 (corresponding to $X = 22$) to .881 (for $X = 81$), with a roughly normal distribution with mean .663 and standard deviation .0870. It is likely that the cases of noise-induced hearing loss that are diagnosed at a late age are more severe, due to a longer exposure time, than the ones diagnosed at an early age. Consequently, the likelihood of positive compensation decision probably increases with age.

In this case the classical analysis and the logistic analysis give $\tilde{N} = 2{,}218$, with $\sqrt{V_1} = 33.3$.

7. Discussion

We have introduced a conditional logistic estimation procedure that allows us to analyze capture–recapture data using individual-level covariate information. It is worth noting that the distributions of the covariates typically are *not* the same in the observed and unobserved segments of the population (cf. Cormack, 1989, p. 412). Our work can be viewed as an extension of the classical procedures and a model introduced by Pollock et al. (1984). Both asymptotic and finite-sample properties of the proposed estimator have been studied. This indicates that the model may be of wide use when the required covariate information exists. The model can be generalized to a multiple-recapture situation. It may be possible to formulate the problem in terms of a nonparametric logistic regression model. Perhaps a Bayesian approach could be used to reduce the instability of the estimator in small samples, if suitable prior information exists.

Acknowledgements

The author would like to thank D. Simpson for comments on an early version of the paper. The helpful suggestions of an associate editor and the referees are also gratefully acknowledged.

Résumé

L'effet d'une hétérogénéité de la population dans les modèles de capture–recapture on de double enregistrement est discuté. Un estimateur de l'effectif inconnu de la population basé sur un modèle de regression logistique est obtenu. Le modèle permet des probabilités de capture différentes entre individus et entre dates de capture. Les probabilités sont estimées à partir des données observées en utilisant le maximum de vraisemblance conditionnel. On montre que l'estimateur résultant de l'effectif de la population est convergent et asymptotiquement normal. Un estimateur de la variance sous l'hypothèse d'hétérogénéité de la population est obtenu. Les propriétés des estimateurs pour des tailles d'échantillon finies sont étudiées par simulation. Une application à des données d'enregistrement de maladies professionnelles en Finlande est présentée.

References

Andersen, E. B. (1980). *Discrete Statistical Models with Social Science Applications*. Amsterdam: North-Holland.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete Multivariate Analysis*. Cambridge, Massachusetts: MIT Press.

Burnham, P. K. and Overton, W. S. (1978). Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika* **65,** 625–633.
Burnham, P. K. and Overton, W. S. (1979). Robust estimation of population size when capture probabilities vary among animals. *Ecology* **60,** 927–936.
Chao, A. (1987). Estimating the population size for capture–recapture data with unequal catchability. *Biometrics* **43,** 783–791.
Chao, A. (1988). Estimating animal abundance with capture frequency data. *Journal of Wildlife Management* **52,** 295–300.
Chung, K. L. (1974). *A Course in Probability Theory*, 2nd edition. New York: Academic Press.
Cormack, R. M. (1968). The statistics of capture–recapture methods. *Oceanography and Marine Biology Annual Review* **6,** 455–506.
Cormack, R. M. (1989). Log-linear models for capture–recapture. *Biometrics* **45,** 395–413.
Ericksen, E. P. and Kadane, J. B. (1985). Estimating the population in a census year. *Journal of the American Statistical Association* **80,** 98–109.
Fahrmeir, L. and Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *Annals of Statistics* **13,** 342–368.
Feller, W. (1968). *An Introduction to Probability Theory and Its Applications. Vol.* I, 3rd edition. New York: Wiley.
Huggins, R. M. (1989). On the statistical analysis of capture experiments. *Biometrika* **76,** 133–140.
Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A* **135,** 370–384.
Otis, D. L., Burnham, K. P., White, G. C., and Anderson, D. R. (1978). Statistical inference from capture data on closed animal populations. *Wildlife Monographs* No. 62.
Pollock, K. H., Hines, J. E., and Nichols, J. D. (1984). The use of auxiliary variables in capture–recapture and removal experiments. *Biometrics* **40,** 329–340.
Pollock, K. H. and Otto, M. C. (1983). Robust estimation of population size in closed animal populations from capture–recapture experiments. *Biometrics* **39,** 1035–1049.
Rodrigues, J., Bolfarine, H., and Leite, J. G. (1988). A Bayesian analysis in closed animal populations from capture–recapture experiments with trap response. *Communications in Statistics—Simulation and Computation* **17,** 407–430.
Sanathanan, L. (1972). Estimating the size of a multinomial population. *Annals of Mathematical Statistics* **43,** 142–152.
Seber, G. A. F. (1982). *The Estimation of Animal Abundance*, 2nd edition. New York: Griffin.
Seber, G. A. F. (1986). A review of estimating animal abundance. *Biometrics* **42,** 267–292.
Sekar, C. C. and Deming, W. E. (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association* **44,** 101–115.

Received November 1987; *revised December* 1988 *and August* 1989; *accepted September* 1989.

APPENDIX

A.1 Proof of Proposition 3.1

The boundedness of the elements of $\mathbf{X}$ implies that $M \to \infty$ a.s. as $N \to \infty$. Moreover, under our hypothesis the conditions of Fahrmeir and Kaufmann's (1985, p. 355) Corollary 1 are easily satisfied. It follows that $\hat{\boldsymbol{\theta}}$ is strongly consistent for θ and has an asymptotic normal distribution, conditionally on $\mathbf{M}$. The unconditional strong consistency follows from $\Pr(\| \hat{\boldsymbol{\theta}} - \boldsymbol{\theta} \| > \varepsilon \text{ i.o.}) = \mathrm{E}[\Pr(\| \hat{\boldsymbol{\theta}} - \boldsymbol{\theta} \| > \varepsilon \text{ i.o.} \mid \mathbf{M})] = 0$, where "i.o." means infinitely often in N (cf. Chung, 1974, Theorem 4.2.2, p. 73). Consider the normality now.

The asymptotic covariance matrix of $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is given by $(\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X}/N)^{-1}$. We shall show that $\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X}/N$ converges a.s. to a matrix (say) $\boldsymbol{\Sigma}$ that does not depend on $\mathbf{M}$. Partition $\boldsymbol{\Sigma}$ into four submatrices corresponding to those of $\mathbf{W}$. For instance, take the upper left-hand corner to be the limit of $\mathbf{X}^{\mathrm{T}}\mathbf{W}_1\mathbf{X}/N$. Note that the (i, j) element of this can be written as

$$\frac{1}{N} \sum_{k=1}^{N} M_k \tilde{w}^1_{kk} X_{1ki} X_{1kj}.$$

By the boundedness of the summands this converges to its asymptotic expectation a.s. Hence, it does not depend on the particular realization $\mathbf{M}$ a.s. For the other elements we reason analogously. It follows that $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \sim N(\mathbf{0}, \boldsymbol{\Sigma}^{-1})$ asymptotically, conditionally on $\mathbf{M}$.

We can uncondition by noting that the normality of $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is equivalent to the condition that for every vector $\boldsymbol{\nu} \neq \mathbf{0}$, $\sqrt{N}\boldsymbol{\nu}^{\mathrm{T}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \sim N(\mathbf{0}, \boldsymbol{\nu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\nu})$ asymptotically. This, in turn, is the same as $\mathrm{E}[f(\sqrt{N}\boldsymbol{\nu}^{\mathrm{T}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}))\,|\,\mathbf{M}] \to \mathrm{E}[f(\boldsymbol{\nu}^{\mathrm{T}}\boldsymbol{\xi})]$ as $N \to \infty$ for $\boldsymbol{\xi} \sim N(\mathbf{0}, \boldsymbol{\Sigma}^{-1})$ and any continuous function f that vanishes outside a compact set (Chung, 1974, Theorem 4.4.1, p. 87). Using the dominated convergence theorem, we get that

$$\mathrm{E}[f(\sqrt{N}\boldsymbol{\nu}^{\mathrm{T}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}))] \to \mathrm{E}[f(\boldsymbol{\nu}^{\mathrm{T}}\boldsymbol{\xi})] \quad \text{as } N \to \infty.$$

This proves the unconditional asymptotic normality.

A.2 Proof of Proposition 3.2

The boundedness of the elements of $\mathbf{X}$ implies that ϕ_i's are bounded away from both 0 and 1. Hence, by Proposition 2.1, $(\hat{N} - N)/N \to 0$ a.s. Another consequence is that there is a neighborhood U of the true value $\boldsymbol{\theta}$ such that the first derivatives S_i, and the second derivatives H_i, of $1/\phi_i$, are bounded in it. Write the Taylor series expansion

$$\frac{\hat{N} - \hat{N}}{\sqrt{N}} = R_1 + R_2,$$

where

$$R_1 = \left(\frac{1}{N}\sum_{i=1}^{N} M_i S_i(\boldsymbol{\theta})^{\mathrm{T}}\right)\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}),$$

and

$$R_2 = \frac{1}{2}\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^{\mathrm{T}}\left(\frac{1}{N^{3/2}}\sum_{i=1}^{N} M_i H_i(\boldsymbol{\eta})\right)\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}),$$

where $\boldsymbol{\eta}$ is between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}$. The boundedness of H_i's ensures that $R_2 \to 0$ a.s. as $N \to \infty$. The sum in R_1 converges to its asymptotic expectation a.s. It follows from Proposition 3.1 that $R_1/\sqrt{N} \to 0$ a.s., so that $\hat{N}$ is strongly consistent. Second, R_1 has an asymptotic normal distribution that does not depend on $\mathbf{M}$ a.s. It follows that $(\hat{N} - N)/\sqrt{N}$, which is a function of $\mathbf{M}$, and $(\hat{N} - \hat{N})/\sqrt{N}$ are asymptotically independent. As a sum of two asymptotically independent normal variables $(\hat{N} - N)/\sqrt{N}$ is asymptotically normal.