BIOMETRICS 55, 904–908
September 1999

Estimation of Population Size Based on Additive Hazards Models for Continuous-Time Recapture Experiments

Paul S. F. Yip,[1,*] Yong Zhou,[2] D. Y. Lin,[3] and Xiang-Zhong Fang[4]

[1]Department of Statistics, The University of Hong Kong, Hong Kong
[2]Insitute of Applied Mathematics, Academia Sinica, Beijing, China
[3]Department of Biostatistics, University of Washington, Box 357232, Seattle, Washington 98195, U.S.A.
[4]Department of Probability and Statistics, Peking University, Beijing, China
* *email:* sfpyip@hkucc.hku.hk

SUMMARY. We use the semiparametric additive hazards model to formulate the effects of individual covariates on the capture rates in the continuous-time capture–recapture experiment, and then construct a Horvitz–Thompson-type estimator for the unknown population size. The resulting estimator is consistent and asymptotically normal with an easily estimated variance. Simulation studies show that the asymptotic approximations are adequate for practical use when the average capture probabilities exceed .5. Ignoring covariates would underestimate the population size and the coverage probability is poor. A wildlife example is provided.

KEY WORDS: Additive risk model; Capture–recapture experiment; Counting process; Horvitz–Thompson estimator; Nonhomogeneous Poisson process.

1. Introduction

The estimation of the size of a population is an important problem in ecology, epidemiology, and other areas. We consider this estimation problem based on capture–recapture sampling. As usual, we assume that the population is closed during the experimental period and that the identity of previously captured subjects is known with certainty. Furthermore, we assume that the experiment is conducted in continuous time and that subjects are captured and recaptured according to nonhomogeneous Poisson processes.

The problem described above was previously studied by Yip, Huggins, and Lin (1996), who modelled the effects of covariates on the subject's capturability by the Cox (1972) proportional hazards model. In this paper, we use the additive hazards model (Cox and Oakes, 1984, p. 74; Breslow and Day, 1987, pp. 122–131) instead. Because temporal effects are assumed to be additive rather than proportional for each covariate, the additive hazards model characterizes certain patterns of temporal influence of covariates more adequately than the Cox model. Buckley (1984) pointed out that the additive hazards model is biologically more plausible than the proportional hazards model, while O'Neill (1986) demonstrated that the use of the proportional hazards model may introduce serious bias when the true model is additive.

The additive hazards model has not been widely used in survival analysis due to the lack of satisfactory semiparametric inference procedures. Recently, Lin and Ying (1994) provided some simple semiparametric methods for estimating the regression parameters and survival probabilities for the additive hazards model. In this paper, we extend the results of Lin and Ying to the context of capture–recapture experiments to estimate the unknown size of a heterogeneous population. The resultant estimator is shown to be consistent and asymptotically normal with an easily estimated variance. Simulation studies show that the asymptotic approximations are adequate for practical use. A wildlife example is provided for illustration.

2. Methods

Let ν be the size of a closed population and τ the duration of the experiment. For $i = 1, 2, \ldots, \nu$, let $\tilde{N}_i(t)$ count the number of times the ith subject has been caught by time t and let $Z_i(t) = (Z_{1i}(t), \ldots, Z_{pi}(t))'$ denote a $p \times 1$ vector of possibly time-varying covariates for the ith subject. Suppose that $\{\tilde{N}_i(t), Z_i(t); t \in [0, \tau]\}$ $(i = 1, \ldots, \nu)$ are independent and identically distributed.

Assume that the counting processes $\tilde{N}_i(t)$ have the intensity functions

$$\lambda(t; Z_i) = \lambda_0(t) + \beta' Z_i(t), \qquad i = 1, \ldots, \nu, \tag{1}$$

where $\lambda_0(\cdot)$ is an unspecified positive function and β is a p vector of unknown parameters. We will estimate β and $\Lambda_0(t) = \int_0^t \lambda_0(u)du$ from the captured subjects.

Let $Y_i(t)$ indicate, by the value one versus zero, whether or not the ith subject has been captured before time t and let $N_i(t)$ count the number of times a captured subject i has been recaptured before time t. Then, according to Lin and Ying (1994), β and $\Lambda_0(t)$ can be consistently estimated, respectively, by

$$\hat{\beta} = \left[\sum_{i=1}^{\nu} \int_0^{\tau} Y_i(t)\big\{Z_i(t) - \bar{Z}(t)\big\}^{\otimes 2} dt\right]^{-1} \times \left[\sum_{i=1}^{\nu} \int_0^{\tau} \big\{Z_i(t) - \bar{Z}(t)\big\} dN_i(t)\right],$$

and

$$\hat{\Lambda}_0(\hat{\beta}, t) = \sum_{i=1}^{\nu} \int_0^{t} \frac{\big\{dN_i(u) - Y_i(u)\hat{\beta}' Z_i(u) du\big\}}{\bar{Y}(u)},$$

where $\bar{Z}(t) = \Sigma_{i=1}^{\nu} Y_i(t) Z_i(t)/\bar{Y}(t)$, $\bar{Y}(t) = \Sigma_{i=1}^{\nu} Y_i(t)$, $a^{\otimes 2} = aa'$.

Remark 1. The calculations of $\hat{\beta}$ and $\hat{\Lambda}_0$ require that the entire covariate histories be available on all the subjects who have been caught. This requirement is often met for time-invariant covariates, such as gender and territory location, as well as for time-dependent covariates that vary over time in a deterministic manner, such as age, and for environmental factors, such as temperature and rainfall. For subject-specific random processes with continuous sample paths, such as weight, it would be impractical to obtain the exact measurements over the entire experimental period.

Let δ_i indicate, by the value one versus zero, whether or not the ith subject has ever been caught during the experiment. Under model (1), the probability $\Pr(\delta_i = 1)$ is

$$p_i = 1 - \exp\left\{-\Lambda_0(\tau) - \beta' \int_0^{\tau} Z_i(u) du\right\}.$$

If $\Lambda_0(\tau)$ and β were known, then the Horvitz–Thompson-type estimator for ν would be $\Sigma_{i=1}^{\nu} \delta_i/p_i$. Thus, we estimate ν by

$$\hat{\nu} = \sum_{i=1}^{\nu} \frac{\delta_i}{\hat{p}_i} = \sum_{i=1}^{n} \frac{1}{\hat{p}_i},$$

where $\hat{p}_i = 1 - \exp\{-\hat{\Lambda}_0(\hat{\beta}, \tau) - \hat{\beta}' \int_0^{\tau} Z_i(u) du\}$. Note that each of the estimators here or after that use ν as the upper bound could use n, the observed number of distinct individuals captured during the course of the experiment. The terms for the unobserved $\nu - n$ are all zero. We show in the Appendix that $\hat{\nu}$ is approximately normal with mean ν and variance

$$\hat{S}^2 = \sum_{i=1}^{\nu} \frac{\delta_i(1 - \hat{p}_i)}{\hat{p}_i^2} + \hat{\nu}\hat{K}'\hat{A}^{-1}\hat{B}\hat{A}^{-1}\hat{K} + \hat{\nu}\hat{h}^2\hat{\psi} + 2\hat{\nu}\hat{h}\hat{K}'\hat{A}^{-1}\hat{G}, \tag{2}$$

where

$$\hat{A} = \hat{\nu}^{-1} \sum_{i=1}^{\nu} \int_0^{\tau} Y_i(t)\big\{Z_i(t) - \bar{Z}(t)\big\}^{\otimes 2} dt,$$

$$\hat{B} = \hat{\nu}^{-1} \sum_{i=1}^{\nu} \int_0^{\tau} \big\{Z_i(t) - \bar{Z}(t)\big\}^{\otimes 2} dN_i(t),$$

$$\hat{h} = -\hat{\nu}^{-1} \sum_{i=1}^{\nu} \frac{\delta_i(1 - \hat{p}_i)}{\hat{p_i}^2},$$

$$\hat{\psi} = \hat{\nu} \sum_{i=1}^{\nu} \int_0^{\tau} \frac{dN_i(t)}{\bar{Y}^2(t)},$$

$$\hat{K} = -\hat{\nu}^{-1} \sum_{i=1}^{\nu} \frac{\delta_i(1 - \hat{p}_i)}{\hat{p}_i^2} \int_0^{\tau} \big\{Z_i(t) - \bar{Z}(t)\big\} dt,$$

$$\hat{G} = \sum_{i=1}^{\nu} \int_0^{\tau} \frac{\{Z_i(u) - \bar{Z}(u)\} dN_i(u)}{\bar{Y}(u)}.$$

Remark 2. Let $F_i(t)$ denote the probability that the ith subject has been captured by time t. Then it is natural to estimate $F_i(t)$ by

$$\hat{F}_i(t) = 1 - \exp\left\{-\hat{\Lambda}_0(\hat{\beta}, t) - \int_0^{t} \hat{\beta}' Z_i(u) du\right\}. \tag{3}$$

Obviously $\hat{p}_i = \hat{F}_i(\tau)$. As it stands, the estimator (3) may not always be monotone in t. However, a simple modification suggested by Lin and Ying (1994) can be used to ensure monotonicity while preserving the given asymptotic properties. Specifically, we can obtain a more stable estimator for p_i by using

$$\hat{p}_i = \sup_{t \le \tau} \hat{F}_i(t),$$

which will be used in our numerical calculations.

3. Numerical Results

3.1 *Simulation Studies*

A series of simulation studies were carried out to assess the performance of the proposed methods. The capture times were generated from the model $\lambda(t; Z_1, Z_2) = 1 + (0.3Z_1 - 0.02Z_2)$, where Z_1 corresponds to sex, with half of the subjects assigned to each sex, and Z_2 corresponds to weight, with a normal distribution of mean eight and variance four. This model implies that males are more catchable than females and the catchability declines with weight. Six combinations of population size and capture period were considered, corresponding to $\nu = 100$ and 200 and $\tau = .5, 1, 2$, and 4. The overall probabilities of being captured are .39, .63, .86, and .98 for $\tau = .5, 1, 2$, and 4, respectively. There were 1000 simulation samples for each combination of ν and τ.

The simulation results are summarized in Table 1. For $\tau = .5$, a significant proportion of the simulated values failed to provide reasonable values for β, the estimator didn't perform well, and the coverage was not satisfactory. The simulation results were based on successful simulation. However, for τ of one or larger, both the mean and median of $\hat{\nu}$ are close to ν. In addition, the mean of the standard error estimator $\hat{S}$ is close to the true standard error of $\hat{\nu}$. The 95% confidence interval $(\hat{\nu} - 1.96\hat{S}, \hat{\nu} + 1.96\hat{S})$ has proper coverage probabilities. Also, ignoring the covariates would cause a negative bias and a smaller standard error for the population size estimate. For example, for a given hazard, $\lambda(t; Z_1, Z_2) = 1 + (0.8Z_1 - 0.02Z_2)$. The values of ν and τ are 400 and 4, respectively. Table 2 gives the results of ignoring one, two, or both covariates. The effect is more serious if the more significant covariate, Z_1, is neglected. Ignoring the

Table 1
Summary of simulation results

	$\nu = 100$				$\nu = 200$			
	$\tau = .5$	$\tau = 1$	$\tau = 2$	$\tau = 4$	$\tau = .5$	$\tau = 1$	$\tau = 2$	$\tau = 4$
Median	79	102	101	100	137	200	202	201
Mean	82	103	101	100	145	200	203	201
SE($\hat{\nu}$)	17.5	15.7	6.7	1.9	19.3	22.7	14.9	5.3
Mean($\hat{s}$)	22.4	16.6	6.6	1.8	27.0	26.0	15.4	5.0
Coverage	.80	.94	.94	.95	.47	.95	.95	.94

covariates underestimates the population size and the coverage is very poor.

3.2 *Mai Po Recapture Data*

We now apply the proposed methods to a capture–recapture experiment from the Mai Po Bird Sanctuary in Hong Kong. The data set pertains to the captures of *Prinia inornata* (brown wren-warble) at Mai Po in the year 1992. Mai Po is one of the few wetlands in Hong Kong. The Mai Po reserve provides a vitally important feeding ground for migrants and winter visitors. There are also a number of resident birds. *Prinia inornata* is a very common territorial species and is a permanent resident of Mai Po.

Prinia inornata mainly inhabit the reed beds in the swamp. The birds were captured in mist nets in the early mornings. The nets were set in the reed beds and were approximately 2 m high. Thus, the catchable population consists of those birds that were active in the reed beds. For this illustration, we only make use of the gender information. Only about one quarter of the captured birds had the gender information because the gender of the bird is not routinely determined by the ringers. We cannot make use of other covariates, such as weight, age, tail length, and fat level, because they were only measured at the time of capture and cannot be assumed to be constant during the study period.

To be specific, a total of 74 distinct birds with known gender were captured in 1992 (46 males and 28 females). The time was measured in days since January 1, 1992, and standardized by 365. The capture frequency distribution is as follows: $f_1 = 53, f_2 = 14, f_3 = 6$, and $f_4 = 1$, where f_i denotes the number of birds that were caught i times in 1992.

The regression coefficient β for the gender is estimated at 1.089×10^{-3} with an estimated standard error of $.521 \times 10^{-3}$. Thus, the gender effect is significant. The positive coefficient implies that male birds were more capturable than their female counterparts. This observation was confirmed by the Mai Po Bird Sanctuary. The higher capturability of male birds can be explained by the fact that female birds spend more time in the nests while male birds are more active defending their territories and searching for food.

Table 2
Summary of simulation results of ignoring the covariates with $\nu = 400$, $\tau = 4$, and the hazard $\lambda(t; Z_1, Z_2) = 1 + (0.8Z_1 - 0.02Z_2)$

	Covariates			
	Z_1 and Z_2	Z_1 only	Z_2 only	None
Median	400.7	399.2	386.2	385.9
Mean	401.0	399.1	386.0	385.6
SE($\hat{\nu}$)	6.2	5.8	4.6	4.6
Mean($\hat{s}$)	6.4	5.7	2.6	2.5
Coverage	.96	.93	.03	.02

Our estimated population size based on the 74 birds with gender information is 201 with an estimated standard error of 51.4. If one ignores the gender effect, then the corresponding estimate of the population size is 174 with an estimated standard error of 29.6. Assuming that the gender information was missing completely at random, our estimate of the total population size is 804, which was also confirmed by the Mai Po Sanctuary to be a realistic estimate. The average capture probability of all the distinct birds is 0.43.

4. Conclusion

The semiparametric additive hazard model provides a useful alternative to the Cox proportional hazards model in formulating the effects of covariates on the capture probability. In practice, we recommend that both models be used. Although these models are very flexible and can accommodate any type of covariates, the use of time-dependent covariates has the practical limitation of requiring the entire covariate history be known throughout the study period.

Our simulation results suggest that the estimation procedure works well as long as the overall capture proportion is not too small, say greater than 50%. When the capture proportion is very low, the proposed estimator is not stable. This is also the case with any existing methods allowing for heterogeneity (Otis et al., 1978; Huggins, 1989, 1991). Also, ignoring heterogeneity in catchability would cause a negative bias (Carothers, 1973). In the example, the simulation studies confirmed the negative bias in estimation. It is also important to note that, with a fixed hazard function, increasing τ will increase the average capture probabilities. Generally, it is the average capture probability and its distribution over the individuals in the population for a given τ that will determine the performance of the estimator.

Acknowledgements

The authors would like to thank the useful suggestions by two referees, an associate editor, and the editor. They are also grateful to the Mai Po Bird Sanctuary for the use of their data set. This research was supported by the Research Grant

Council of Hong Kong (PSFY), a postdoctoral fellowship of the University of Hong Kong (YZ), and the U.S. National Institutes of Health (DYL).

Résumé

Nous utilisons le modèle semi-paramétrique à risques additifs pour formaliser les effets de covariables individuelles sur les taux de capture dans une expérience de capture-recapture à temps continu, et pour construire un estimateur de type Horvitz-Thomson pour la taille inconnue de la population. L'estimateur obtenu est convergent, asymptotiquement gaussien et de variance aisément estimable. Des études de simulation montrent que les approximations asymptotiques sont adaptées en pratique pour des probabilités moyennes de capture supérieures à 0.5 . La non-prise en compte des covariables conduirait à sous-estimer la taille de la population, et la robustesse serait alors faible. On présente un exemple sur une espèce sauvage.

References

Breslow, N. E. and Day, N. E. (1987). *Statistical Methods in Cancer Research,* Volume II, *The Design and Analysis of Cohort Studies.* Lyon: IARC.

Buckley, J. D. (1984). Additive and multiplicative models for survival rates. *Biometrics* **40,** 51–62.

Carothers, A. D. (1973). The effects of unequal catchability on Jolly–Seber estimates. *Biometrics* **29,** 79–100.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34,** 187–220.

Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data.* London: Chapman and Hall.

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

Huggins, R. M. (1991). Some practiced aspects of a conditional likelihood approach to capture experiments. *Biometrics* **47,** 725–732.

Lin, D. Y. and Ying, Z. (1994). Semiparametric analysis of the additive risk model. *Biometrika* **81,** 61–71.

O'Neill, T. J. (1986). Inconsistency of the misspecified proportional hazards model. *Statistics and Probability Letters* **4,** 219–222.

Otis, D. L., Burnham, K. P., White, G. C., and Anderson, D. R. (1978). *Statistical Inference from Capture Data on Closed Animals Populations.* Wildlife Monograph 62.

Yip, P. S. F., Huggins, R. M., and Lin, D. Y. (1996). Inference for capture–recapture experiments in continuous time with variable capture rates. *Biometrika* **83,** 477–483.

Received January 1997. *Revised April and October* 1998.
Accepted October 1998.

Appendix

Asymptotic Properties of $\hat{\nu}$

Clearly,

$$\begin{aligned}
&\nu^{-\frac{1}{2}}(\hat{\nu}-\nu)\\
&\quad = \nu^{-\frac{1}{2}} \sum_{i=1}^{\nu} \left[\frac{\delta_i}{1-\exp\left\{-\hat{\Lambda}_0(\hat{\beta},\tau) - \hat{\beta}' \int_0^\tau Z(u)du\right\}} - 1 \right]\\
&\quad = \nu^{-\frac{1}{2}} \sum_{i=1}^{\nu} \left[\frac{\delta_i}{1-\exp\left\{-\hat{\Lambda}_0(\hat{\beta},\tau) - \hat{\beta}' \int_0^\tau Z_i(u)du\right\}} \right.\\
&\qquad\qquad \left. - \frac{\delta_i}{1-\exp\left\{-\hat{\Lambda}_0(\beta,\tau) - \beta' \int_0^\tau Z_i(u)du\right\}} \right]\\
&\qquad + \nu^{-\frac{1}{2}} \sum_{i=1}^{\nu} \left[\frac{\delta_i}{1-\exp\left\{-\hat{\Lambda}_0(\beta,\tau) - \beta' \int_0^\tau Z_i(u)du\right\}} \right.\\
&\qquad\qquad \left. - \frac{\delta_i}{1-\exp\left\{-\Lambda_0(\tau) - \beta' \int_0^\tau Z_i(u)du\right\}} \right]\\
&\qquad + \nu^{-\frac{1}{2}} \sum_{i=1}^{\nu} \left[\frac{\delta_i}{1-\exp\left\{-\Lambda_0(\tau) - \beta' \int_0^\tau Z_i(u)du\right\}} - 1 \right].
\end{aligned} \tag{A.1}$$

By Taylor expansions, the first term on the right side of (A.1) is asymptotically

$$K'A^{-1}\nu^{-\frac{1}{2}} \sum_{i=1}^{\nu} \int_0^\tau \left\{Z_i(t) - \bar{Z}(t)\right\} dM_i(t), \tag{A.2}$$

where

$$K = -\lim_{\nu\to\infty} \nu^{-1} \sum_{i=1}^{\nu} \frac{\delta_i(1-p_i)}{p_i^2} \int_0^\tau \left\{Z_i(t) - \bar{Z}(t)\right\} dt,$$

$$A = \lim_{\nu\to\infty} \nu^{-1} \sum_{i=1}^{\nu} \int_0^\tau Y_i(t) \left\{Z_i(t) - \bar{Z}(t)\right\}^{\otimes 2} dt,$$

and

$$M_i(t) = N_i(t) - \int_0^t Y_i(u) \left\{\lambda_0(u) + \beta' Z_i(u)\right\} du.$$

Similarly, the second term on the right side of (A.1) is asymptotically

$$h\nu^{\frac{1}{2}} \sum_{i=1}^{\nu} \int_0^\tau \frac{dM_i(t)}{\bar{Y}(t)}, \tag{A.3}$$

where

$$h = -\lim_{\nu\to\infty} \nu^{-1} \sum_{i=1}^{\nu} \frac{\delta_i(1-p_i)}{p_i^2}.$$

The third term on the right side of (A.1) is

$$\nu^{-\frac{1}{2}} \sum_{i=1}^{\nu} \frac{\delta_i - p_i}{p_i}. \tag{A.4}$$

Note that (A.2) and (A.3) are martingale integrals, while (A.4) is a sum of independent and identically distributed random variables with zero mean and finite variance. Thus, it follows from the martingale and classical central limit theorem that the random variable $\nu^{-1/2}(\hat{\nu}-\nu)$ is asymptotically zero-mean normal. By standard martingale arguments, the limiting variances of (A.2) and (A.3) are, respectively, $K'A^{-1}BA^{-1}K$ and $h^2\psi$, where

$$B = \lim_{\nu\to\infty} \nu^{-1} \sum_{i=1}^{\nu} \int_0^\tau \left\{Z_i(t) - \bar{Z}(t)\right\}^{\otimes 2} dN_i(t)$$

and

$$\psi = \lim_{\nu \to \infty} \nu \sum_{i=1}^{\nu} \int_0^{\tau} \frac{dN_i(t)}{\bar{Y}^2(t)}.$$

In addition, the limiting covariance between (A.2) and (A.4) is $hK'A^{-1}G$, where

$$G = \lim_{\nu \to \infty} \sum_{i=1}^{\nu} \int_0^{\tau} \frac{\left\{Z_i(t) - \bar{Z}(t)\right\} dN_i(t)}{\bar{Y}(t)}.$$

Furthermore, the covariances between (A.2) and (A.4) and between (A.3) and (A.4) are both zero because $\delta_i M_i(t) = M_i(t)$. Finally, the variance of (A.4) is

$$\nu^{-1} \sum_{i=1}^{\nu} \frac{1 - p_i}{p_i}.$$

Replacements of all the unknown quantities in the limiting variance and covariance terms yield $\hat{v}^{-1}\hat{S}^2$, where $\hat{S}^2$ is given in (2).