STATISTICA, anno LXIII, n. 1, 2003

A DOUBLE-SAMPLING APPROACH FOR MAXIMUM LIKELIHOOD ESTIMATION FOR A POISSON RATE PARAMETER WITH VISIBILITY-BIASED DATA

J. D. Stamey, D. M. Young, M. Cecchini

1. INTRODUCTION

Suppose one is interested in estimating the rate of gallinule nests along a certain waterway. To thoroughly search a large area may not be possible, and if only a cursory search is undertaken, nests obstructed from view would be missed. Thus, a cursory search would very likely result in observations that are uncounted, therefore underestimating the nest rate. Such an observation is known as a false-negative observation in the count. A probabilistic model suffering from the presence of false negatives is said to contain visibility bias.

One method to correct for visibility bias is to use a double-sampling method. There are two double-sample methods in which we are interested. The first double-sample method is the use of a training sample, which occurs when both a cursory (fallible) and a thorough (infallible) search are utilized to estimate the rate of interest. A second double-sample procedure is the use of a calibration sample. In this approach one obtains information only on the probability of observation in the smaller sample. Under both the training-sample and the calibration-sample approaches, one combines the information contained in the infallible sample with information gleaned from a fallible sample for which only a cursory search is performed.

Different models and methods have been proposed to estimate parameters for counted data subject to visibility bias. Anderson, Bratcher, and Kutron (1994) consider a fully Bayesian analysis for estimating a Poisson rate parameter when the data is subject to visibility bias. In related papers Fader and Hardie (2000) apply an empirical Bayesian analysis to wine purchase behavior, and Whitemore and Gong (1991) and Sposto *et al.* (1992) use this model to estimate death rates in the presence of visibility bias in a categorical regression context.

In this paper we propose two double-sample approaches for formulating maximum likelihood estimators (MLEs) for the Poisson rate parameter of interest and the observation probability parameter. The remainder of this paper is organized in the following manner. In section 2 we derive the MLEs of the Poisson

rate parameter λ and probability of observation p under the two double-sample scenarios. We also obtain asymptotic variances of the MLEs. Section 3 is a limited simulation study for the two MLEs under the training sample and the calibration sample procedures. In section 4 we analyze a real data set using our proposed double-sampling paradigm. Section 5 provides a method to determine appropriate search sizes for the fallible and the infallible sample sizes. We conclude with some brief comments in section 6.

2. MAXIMUM LIKELIHOOD ESTIMATORS DERIVED FROM DOUBLE-SAMPLING SCHEMES

Suppose in a search of size A (often person-years or area), one determines T_u true occurrences such that Z of these true occurrences are uncounted. The subscript u is to denote that this is an unobservable quantity. Hence, we assume that $T_u \sim \text{Poisson}(A\lambda)$ and that $Z \mid T_u \sim \text{binomial}(t_u, p)$, where λ is the rate parameter and $(1 - p)$ is the false-negative misclassification parameter. Only Z is actually observed, so that, unconditionally, Z has density function

$$f(z \mid \lambda, p) = \frac{e^{-A\lambda p}}{z!}(A\lambda p)^z, \tag{1}$$

which is a Poisson distribution allowing for false-negative misclassifications. Only $\mu = \lambda p$ is identifiable from the observed data.

To estimate λ, we utilize a second sample in the form of a training sample or a calibration sample. For the training-sample scenario one collects data over the area a where t events occur and are observed by a thorough (infallible) search and $x \leq t$ events are observed by a cursory search. Therefore, in our model $t \sim \text{Poisson}(a\lambda)$ and, conditional on t, $x \sim \text{binomial}(t, p)$. The joint likelihood corresponding to the training sample scenario is

$$L(\lambda, p) \propto \lambda^{z+t} e^{-a\lambda - A\lambda p} p^{z+x} (1-p)^{t-x} \ . \tag{2}$$

The MLEs result from straight-forward optimization of likelihood (2) and are:

$$\hat{\lambda} = \frac{A(t-x) + a(t+z)}{a(a+A)} \tag{3}$$

and

$$\hat{p} = \frac{a(x+z)}{A(t-x) + a(t+z)}. \tag{4}$$

The estimator $\hat{\lambda}$ given in (3) has an intuitive interpretation. Consider the following alternative representation of (3):

$$\hat{\lambda} = \alpha_1(t - x)/a + \alpha_2(t + z)/a, \tag{5}$$

where $\alpha_1 = A/(a + A)$ and $\alpha_1 + \alpha_2 = 1$. From (5) one can see that $\hat{\lambda}$ is a weighted average of the number of observations missed in the small area, $(t - x)$, and the total number observed in both samples, $(t + z)$. Thus, the first component of (5) simply "adds back" the number of observations missed proportionally to the size of A. We note that (5) is an improvement over the posterior mean of Anderson et al. (1994) and the empirical Bayes estimator of Fader and Hardie (2000) in terms of the computational demand and the ease of interpretation.

The inverse of Fisher's information matrix gives the asymptotic covariance matrix

$$\mathrm{Cov}(\hat{\lambda}, \hat{p}) \approx \begin{bmatrix} \dfrac{(a + A(1-p))\lambda}{a(a+A)} & \dfrac{-Ap(1-p)}{a(a+A)} \\ \dfrac{-Ap(1-p)}{a(a+A)} & \dfrac{p(1-p)(a+Ap)}{a(a+A)\lambda} \end{bmatrix}. \tag{6}$$

The diagonal elements of (6) yield the asymptotic variance expressions

$$\mathrm{Var}(\hat{\lambda}) = \frac{(a + A(1-p))\lambda}{a(a+A)} \tag{7}$$

and

$$\mathrm{Var}(\hat{p}) \approx \frac{p(1-p)(a+Ap)}{a(a+A)\lambda}. \tag{8}$$

One can easily show that (7) is the variance of $\hat{\lambda}$ for all sample sizes a and A.

We can estimate the variance expressions (7) and (8) by substituting the estimators (3) and (4) into the variance expressions (7) and (8) for the corresponding parameters. We also note the following properties of $\hat{\lambda}$:

i) $\hat{\lambda}$ is unbiased;

ii) If $p = 1$, $\mathrm{Var}(\hat{\lambda}) = \dfrac{\lambda}{a + A}$, which is the variance of the Poisson rate MLE for λ without misclassified data;

iii) For small values of p, $\mathrm{Var}(\hat{\lambda}) \approx \dfrac{\lambda}{a}$; that is, essentially all information for estimating λ is from the training sample when p is small.

We next formulate MLEs for λ and p under a different sample-use scenario. Suppose that instead of using a training sample, one has only information on the probability of observance p through a second sample. We refer to this estimation paradigm as a calibration sample. For instance, Whittemore and Gong (1991) analyze a similar model in the regression context to estimate cervical cancer death rates for varying age groups and geographic regions. In Whittemore and Gong (1991) a case history is provided to a sample of physicians, and each is asked to complete a death certificate. The probability of observance p is estimated by the number of correctly classified counts divided by the calibration sample size. Thus, in the context we are considering, there are t known occurrences and we use only the fallible search and observe x occurrences.

We can combine the information gained via the calibration sample with the larger fallible search through the likelihood function in a fashion similar to the training-sample case described above. The joint likelihood function corresponding to the calibration sample is

$$L(\lambda, p) \propto \lambda^{z} e^{-A\lambda p} p^{z+x} (1-p)^{t-x}. \tag{9}$$

One can readily show that the MLEs for p and λ derived from (9) are $\tilde{p} = \frac{x}{t}$ and $\tilde{\lambda} = \frac{z}{A\tilde{p}}$, respectively, and their asymptotic covariance matrix is

$$\mathrm{Cov}(\tilde{\lambda}, \tilde{p}) \approx \begin{bmatrix} \frac{\lambda(t + A(1-p)\lambda)}{Apt} & \frac{-\lambda(1-p)}{t} \\ \frac{-\lambda(1-p)}{t} & \frac{p(1-p)}{t} \end{bmatrix}. \tag{10}$$

Recall that t is a constant in the calibration sample case. The diagonal elements of (10) yield the asymptotic variances

$$\mathrm{Var}(\tilde{\lambda}) \approx \frac{\lambda(t + A(1-p)\lambda)}{Apt} \tag{11}$$

and

$$\mathrm{Var}(\tilde{p}) = \frac{p(1-p)}{t}. \tag{12}$$

We note, however, that (12) is the variance of $\tilde{p}$ for all sample sizes and that $\tilde{\lambda}$ does not exist if $x = 0$.

3. A SIMULATION STUDY

We now compare the efficacy of the training-sample and the calibration-sample MLE procedures derived and discussed in section 2 using a simulation study. We use the rate of occurrence $\lambda = 20$ for all simulation configurations. To put the study in a context, one could view $\lambda = 20$ as the rate of an illness per 1000 person-years. For the larger fallible sample, we consider sizes of $A = 2{,}500$, 10,000, 50,000. For the training sample we consider sizes of $a = 1000$, 2000. Note that the sample sizes are in terms of thousands. For the calibration sample we use values of $t = 20$, 40, and we use the cases where $p = .4$, .7, and .9 for both the training and the calibration sample. The average mean square errors (MSE) for the two simulation experiments are given in tables 1 and 2.

We performed the simulation using S-Plus. For each parameter configuration 1000 simulations of 1000 data sets were generated. Thus, each configuration MSE value is the result of 1000 simulations of 1000 iterations each. The reported values are the average MSEs and their estimated standard errors over the 1000 simulations.

TABLE 1

MSE ($\hat{\lambda}$) *for the training sample scenario*

	$a = 1, A = 2.5$	$a = 1, A = 10$	$a = 1, A = 50$
$p = .4$	14.33 (.021)	12.75 (.020)	12.18 (.017)
$p = .7$	10.00 (.015)	7.27 (.011)	6.29 (.009)
$p = .9$	7.14 (.011)	3.64 (.005)	2.35 (.004)
	$a = 2, A = 2.5$	$a = 2, A = 10$	$a = 2, A = 50$
$p = .4$	7.77 (.011)	6.67 (.010)	6.16 (.010)
$p = .7$	6.12 (.009)	4.16 (.006)	3.28 (.005)
$p = .9$	5.01 (.007)	2.50 (.004)	1.35 (.002)

TABLE 2

MSE($\tilde{\lambda}$) *for the calibration sample scenario*

	$t = 20, A = 2.5$	$t = 20, A = 10$	$t = 20, A = 50$
$p = .4$	102.10 (.592)	80.80 (.522)	75.73 (.526)
$p = .7$	23.64 (.050)	14.42 (.036)	11.91 (.032)
$p = .9$	11.64 (.020)	4.84 (.009)	3.03 (.006)
	$t = 40, A = 2.5$	$t = 40, A = 10$	$t = 40, A = 50$
$p = .4$	44.23 (.108)	27.15 (.072)	22.58 (.068)
$p = .7$	16.74 (.028)	7.84 (.014)	5.48 (.010)
$p = .9$	10.17 (.015)	3.43 (.005)	1.64 (.003)

Although the corresponding MSE values in the two tables are not directly comparable, we gain some information by comparing the corresponding MSEs of the two double-sampling paradigms. The calibration sample sizes t in table 2 are the expected number of true occurrences in the training sample for a rate of $\lambda = 20$ in a training sample of sizes $a = 1000$, 2000, respectively. For all combinations of parameters considered here, the MSE of the rate estimator involving a training

sample is less than the MSE of the rate estimator using the calibration sample. As one might anticipate, this difference in corresponding MSEs decreases as the sample sizes a and A and probability of observance p increase.

Generally, the training-sample approach is more expensive to implement than the calibration sample. Because of the large difference in MSEs, we recommend the use of a training sample when p is small. However, if p is thought to be relatively large, the fallible sample size A is large, and resources are not available to easily implement the training-sample scheme, the calibration-sample approach can be used with only a relatively small increase in MSE over the training-sample scenario for the configurations considered here.

4. AN APPLICATION

We analyze data from Anderson et al. (1994) for which the parameter of interest is the rate of gallinule nests along the water of Lacassine National Wildlife Refuge in southern Louisiana. A cursory (fallible) search along the waterway is conducted along with a thorough (infallible) search by airboat over a smaller area. The fallible search is over the area A = 4300 linear feet, and the infallible search is applied over the smaller area a = 500 linear feet. In the thorough search, 11 nests are spotted, 7 of which were in using a cursory search over the same area.

For the larger area for which only a cursory search was applied, 21 nests were spotted. Using (3), one gets the resulting estimate of the Poisson rate to be $\hat{\lambda}$ = 19.21 nests per 1000 linear feet with an estimated standard error of 5.40. In addition, the estimated probability of observance is $\hat{p} = 0.27$ with an estimated standard error of .084.

If one uses only the infallible data to estimate the nesting rate, one gets the estimate λ^* = 22 nests per 1000 linear feet with an estimated standard error of 6.63. For this data we note that using the fallible data in conjunction with the training data yields only a modest 19% reduction in the estimated standard error of $\hat{\lambda}$. This modest reduction in the dispersion of $\hat{\lambda}$ is caused by the relatively small value of the probability of observance p. We also note that $\hat{\lambda}$, which is an unbiased estimator, corrects for the bias in λ^* due to misclassification.

5. SAMPLE SIZE DETERMINATION

A possible concern prior to implementing our proposed training-sample-based MLE for λ is determining appropriate sample sizes of a and A. The quantity $\mathrm{Var}(\hat{\lambda})$ is a function of both the fallible sample size A and the infallible sample size a. We consider the case where the cost per unit sampled from a is c_a units and the cost per unit sampled from A is c_A units.

One might wish to determine the sample-size allocation that minimizes the variance of $\hat{\lambda}$ when the researcher has a budget of C dollars. The sample-size determination problem can be stated explicitly as the constrained minimization problem

$$\text{minimize } \frac{(a + \mathcal{A}(1-p))\lambda}{a(a+\mathcal{A})} \text{ subject to } c_a a + c_{\mathcal{A}} \mathcal{A} \le C .$$

We use the method of Lagrange multipliers to determine the optimal sample-size allocations

$$a = \frac{(c_a - c_{\mathcal{A}})C(1-p) + \sqrt{(c_a - c_{\mathcal{A}})c_{\mathcal{A}} C^2 p(1-p)}}{(c_a - c_{\mathcal{A}})\left[c_{\mathcal{A}} - c_a(1-p)\right]}$$

for the infallible sample size and (13)

$$\mathcal{A} = \frac{c_a\left[c_{\mathcal{A}} C - \sqrt{(c_a - c_{\mathcal{A}})c_{\mathcal{A}} C^2 p(1-p)}\right] - c_{\mathcal{A}}^2 C}{(c_a - c_{\mathcal{A}})\left[c_{\mathcal{A}} - c_a(1-p)\right]}$$

for the fallible sample size. One drawback to this sample-size determination approach is that the allocations are functions of the reporting probability p that is generally unknown. However, a researcher can possibly have an *a priori* estimate p_0. An alternative approach is to take a pilot sample of size a_0 to estimate p and then use the allocation formulas (13) to determine the opportunity sizes a and $\mathcal{A}$.

Consider the following as an example application of the sample-size determination method derived in section 4. Suppose the 500 linear feet in the gallinule example is a pilot study; then, $\hat{p}$ = .34. Also, suppose that the cost to infallibly search the area is \$75 per 1000 linear feet, the cost to fallibly search the area is \$8 per 1000 linear feet, and we have a budget of \$750. Evaluating the allocation formulas (13) for these values results in the sample sizes a = 8.97 and $\mathcal{A}$ = 9.66. Consequently, the allocation formula proposes that one should use approximately the same number of infallible data as fallible data in the case where p is small. For comparison purposes, if $\hat{p}$ had been .85, then the infallible and fallible sample sizes would have been a = 6.14 and $\mathcal{A}$ = 36.17, respectively.

6. COMMENTS

In this paper we have derived MLEs of the rate parameter of a Poisson model subject to visibility bias of the counted data. These estimators compare favorably with two other estimators previously presented in the literature. However, the fully Bayesian estimators given in Anderson et al. (1994) can be computationally

demanding for large sample sizes. These estimators also do not have a simple and intuitive interpretation as the MLE estimators. The empirical Bayes estimation approach of Fader and Hardie (2000) suffers the same difficulty.

An additional problem of the empirical Bayes approach is the identifiability of parameters. By using the data to estimate the prior parameters, Fader and Hardie (2000) overlook a fundamental issue of parameter identifiability. For instance, consider the case with when $\lambda = 10$, $p = .2$ and when $\lambda = 4$, $p = .5$. For both of these parameter configurations we have that $\mu = \lambda p = 2$. Thus, two different sets of parameters could generate essentially equivalent observable data. Consequently, Fader and Hardie's approach could result in roughly the same estimates for the two different parameter configurations. The training sample approach we suggest overcomes these difficulties.

Last, we note that we can readily construct confidence intervals for λ and p for both the training-sample and calibration-sample double-sampling paradigms using the properties of MLEs and estimates of the asymptotic variances given in (7) and (8) and in (11) and (12).

Department of Mathematics and Statistics JAMES D. STAMEY
Stephen F. Austin University

Department of Statistical Sciences DEAN M. YOUNG
Baylor University

Department of Mathematics MARTITA CECCHINI
Northwestern State University

REFERENCES

C. ANDERSON, T. BRATCHER, K. KURTRAN, (1994), *Bayesian estimation of population density and visibility*, "The Texas Journal of Science", 46, pp. 7-12.

P.S. FADER, B.G.S. HARDIE, (2000), *A note on modeling underreported Poisson counts*, "Journal of Applied Statistics", 8, 953-964.

R. SPOSTO, D.L. PRESTON, Y. SHIMIZU, K. MABUCHI, (1992), *The effect of diagnostic misclassification on non-cancer and cancer mortality dose response in A-bomb survivors*, "Biometrics", 48, pp. 605-617.

A.S. WHITTEMORE, G. GONG, (1991), *Poisson regression with misclassified counts: application to cervical cancer mortality rates*, "Applied Statistics", 40, pp. 81-93.

RIASSUNTO

Un metodo di campionamento doppio per la stima di massima verosimiglianza del parametro di una distribuzione di Poisson quando una parte della popolazione non è visibile all'osservatore

Nel lavoro si propone un modello poissoniano che usa sia dati precisi che dati imprecisi, ovvero soggetti ad errata classificazione in quanto parte della popolazione non è visibile all'osservatore (*visibility bias*). Vengono quindi derivati gli stimatori di massima verosimiglianza del parametro della distribuzione poissoniana e del parametro di errata classifica-

zione in presenza di due scenari differenti. Sono inoltre derivate analiticamente le matrici di informazione e le varianze asintotiche degli stimatori di massima verosimiglianza di entrambi i parametri. Infine, il modello proposto viene analizzato sulla base di un esperimento di simulazione e quindi applicato ad un problema concreto.

SUMMARY

A double-sampling approach for maximum likelihood estimation for a Poisson rate parameter with visibility-biased data

We propose a Poisson-based model that uses both infallible data and fallible data subject to misclassification in the form of false negatives that yield visibility bias. We then derive maximum likelihood estimators for the Poisson rate parameter of interest and the misclassification parameter under two different sampling scenarios. We also derive expressions for the information matrices and the asymptotic variances of the maximum likelihood estimators for the rate parameter and the maximum likelihood estimators for the false-negative parameter. Finally, we also study our new models via a simulation experiment and then apply our new estimation procedures to a real data set.