Hindawi Publishing Corporation
Advances in Decision Sciences
Volume 2011, Article ID 571768, 8 pages
doi:10.1155/2011/571768

Research Article

Modelling Inverse Gaussian Data with Censored Response Values: EM versus MCMC

R. S. Sparks,[1] G. Sutton,[1] P. Toscas,[1] and J. T. Ormerod[2]

[1] *CSIRO Mathematical, Informatics, and Statistics, Locked Bag 17, North Ryde, NSW 1670, Australia*
[2] *School of Mathematics and Statatistics, University of Sydney, Camperdown, NSW 2006, Australia*

Correspondence should be addressed to R. S. Sparks, ross.sparks@csiro.au

Received 8 September 2010; Revised 13 January 2011; Accepted 4 April 2011

Academic Editor: Shelton Peiris

Low detection limits are common in measure environmental variables. Building models using data containing low or high detection limits without adjusting for the censoring produces biased models. This paper offers approaches to estimate an inverse Gaussian distribution when some of the data used are censored because of low or high detection limits. Adjustments for the censoring can be made if there is between 2% and 20% censoring using either the EM algorithm or MCMC. This paper compares these approaches.

1. Introduction

Partial missing data is common in environmental applications, where measurement systems have low or high detection limits. The measurement processes at times are incapable of measuring below certain values; for example, near zero counts of colony form units, below a certain lower limit, denoted c, are considered too unreliable when the measurement process involves diluting solution samples. An example of a high detection limit (c) occurs for flows when flow is measured based on the height of a river, yet the river height measurement cannot exceed the height of the bank. All that is known in such cases is that the measurement is below or above c, respectively. In this paper, accurate model fitting is the aim. We advocate the EM algorithm [1] to establish realistic penalised maximum likelihood estimates of regression parameters in such cases. Other approaches apply MCMC methods [2]. The next section introduces the EM algorithm approach for dealing with low/high detection limits. A simulation study for high detects (the case that is more unstable because it is unbounded on the high side) is used to demonstrate the capability of the methodology. Section 3 looks at the MCMC approach. Section 4 uses an example to demonstrate the EM methodology. Section 5 concludes with a discussion.

2. The EM Algorithm

The expectation step of the EM algorithm is used to replace censored values based on the current parameter estimates and the detection limit. These replaced values are, in turn, used to establish penalised maximum likelihood estimates of regression parameters. These two steps are applied iteratively until estimates converge.

In this paper, we will use the GAMLSS model to fit the location and spread for the data using penalised maximum likelihood estimation (see [3, 4]). The library GAMLSS has the capability of fitting models with low detects, but this is experimental, while our approach has been tested at least for the applications considered in this paper.

Let the location of the response y_i given the vector of explanatory variables x_i be denoted by $f(\mu_{y\cdot x_i}) = x_i^t\beta$, where f is the link function and $\mu_{y\cdot x_i} = \mathrm{E}(y_i \mid x_i)$ is the mean. We are interested in the inverse Gausian distribution using the usual logarithm link function giving

$$\mathrm{E}(y_i \mid x_i) = \mu_{y\cdot x_i} = e^{x_i^t\beta}. \tag{2.1}$$

The inverse Gaussian distribution is given by

$$h(y \mid \mu_{y\cdot x_i}, \phi) = \sqrt{\frac{\phi}{2\pi}}\, y^{-3/2} e^{-\phi(y-\mu_{y\cdot x_i})^2/(2\mu_{y\cdot x_i}^2 y)}, \tag{2.2}$$

where ϕ is the shape parameter and the variance σ_i^2 is given by $\mu_{y\cdot x_i}^3/\phi$. The GAMLSS model for ϕ is

$$\phi = e^{x_i^t\alpha}, \tag{2.3}$$

so that $\sigma_i^2 = e^{3x_i^t\beta - x_i^t\alpha}$. To simplify the notation, we drop subscripts; that is, let $\mu = e^{x^t\beta}$. When y is a low detect, we want to evaluate $\mathrm{E}(y \mid x, y < c)$ which is given by (see Appendix A)

$$\mathrm{E}(y \mid x, y < c) = \mu - 2\mu e^{2\phi/\mu} \frac{\Phi\left(-(c/\mu + 1)\sqrt{\phi/c}\right)}{H(c, \mu, \phi)}, \tag{2.4}$$

where $H(c, \mu, \phi) = \int_0^c h(y, \mu, \phi)dy$ and $\Phi(z) = \int_{-\infty}^{z} e^{(-1/2)x^2}/\sqrt{2\pi}\; dx$.

High detect limits are less common in the application considered later but do occur in practice. In this situation, it can be shown that the expectation step is given by (see [5] for a different result that is often numerically equivalent to the result below)

$$E(y \mid x, y > c) = \mu + 2\mu e^{2\phi/\mu} \frac{\Phi\left(-(c/\mu + 1)\sqrt{\phi/c}\right)}{(1 - H(c, \mu, \phi))}. \tag{2.5}$$

A simulation study, of 1000 runs per table entry, generated data from the inverse Gaussian distribution with density $f(x, \mu = 2, \phi = 1)$ and truncated values at high detection

Table 1: EM simulation results for fitting Inverse Gaussian models with high detects.

	mle estimate ± standard error			
c	3	4	5	6
missing	19%	13%	10%	7%
	mle estimate ± standard error			
Parameters	**1000 observations**			
μ	2.01 ± 0.12	1.99 ± 0.10	2.00 ± 0.11	2.00 ± 0.10
ϕ	1.01 ± 0.04	1.01 ± 0.05	1.01 ± 0.05	1.00 ± 0.05
Parameters	**300 observations**			
μ	2.02 ± 0.22	2.02 ± 0.22	2.03 ± 0.22	2.02 ± 0.20
ϕ	1.02 ± 0.08	1.02 ± 0.11	1.01 ± 0.10	1.01 ± 0.11
Parameters	**100 observations**			
μ	2.06 ± 0.41	2.06 ± 0.43	2.03 ± 0.37	2.05 ± 0.37
ϕ	1.03 ± 0.17	1.04 ± 0.15	1.04 ± 0.18	1.03 ± 0.18
Parameters	**50 observations**			
μ	2.15 ± 0.71	2.08 ± 0.59	2.07 ± 0.60	2.07 ± 0.54
ϕ	1.06 ± 0.23	1.07 ± 0.28	1.07 ± 0.27	1.06 ± 0.28

limits $c = 3, 4, 5$, and 6. The application of the EM algorithm is outlined in Whitemore [5]. This was used to find the maximum likelihood estimates of hidden values μ and ϕ. The results are reported in Table 1. Similar results were found for inverse Gaussian distributions with low detects.

It is clear from the standard errors in Table 1 that the EM algorithm is unstable for 50 observations with standard errors more than 25% of the parameter estimated values. However, for this few observations, we are unlikely to diagnose the data as Inverse Gaussian, and any methodology is going to be suspicious.

If the presence of high detects is ignored so that data with actual values greater than c are set equal to c, then the resulting parameter estimates are biased. For $c = 3, 4, 5$, and 6, the 1000 observations example with high-truncated values in Table 1, produced average estimates of μ and ϕ, without correction for the truncation, equal to ($\widehat{\mu}$ =) 1.374, 1.528, 1.639, and 1.716 (actual $\mu = 2$) and ($\widehat{\phi}$ =) 1.257, 1.169, 1.119, and 1.092 (actual $\phi = 1$), respectively. These parameter estimates are significantly biased. A simulation exercise for low detects was carried out but is not presented, because the results are less extreme (e.g., if $c \ll \mu$, then little is lost by replacing the low detects by $c/2$). This will be demonstrated in the application in Section 4.

3. Markov Chain Monte Carlo

In this section, the results of a Markov chain Monte Carlo analysis of the simulation study of high detects are presented (see previous section for details). The mean and dispersion parameters are estimated using Gibbs sampling (e.g., [6]). Defining the mean parameter as $\mu = e^{\beta}$, the MCMC analysis is undertaken with a uniform prior distribution for β and a gamma prior distribution for ϕ, with shape parameter equal to 0.001 and scale parameter equal to 1,000 (i.e., prior variance = 1000). That is, the mean of the prior distribution for ϕ is equal to 1. Based on these prior distributions, the conditional distribution for ϕ is Gamma $(n/2 + 0.001, \sum_{i=1}^{n} (y_i - \mu)^2/(2\mu^2 y_i) + 0.001)$, so samples of ϕ can be drawn from

Table 2: MCMC simulation results for fitting inverse Gaussian models with high detects.

	MCMC estimate ± standard error			
c	3	4	5	6
missing	19%	13%	10%	7%
Parameters	**1000 observations**			
μ	2.02 ± 0.13	2.01 ± 0.11	2.01 ± 0.11	2.01 ± 0.11
ϕ	1.00 ± 0.05	1.00 ± 0.05	1.00 ± 0.05	1.00 ± 0.05
Parameters	**300 observations**			
μ	2.04 ± 0.24	2.03 ± 0.21	2.03 ± 0.20	2.02 ± 0.19
ϕ	1.00 ± 0.09	1.00 ± 0.09	1.00 ± 0.09	1.00 ± 0.09
Parameters	**100 observations**			
μ	2.18 ± 0.58	2.13 ± 0.46	2.10 ± 0.40	2.08 ± 0.34
ϕ	1.00 ± 0.16	1.00 ± 0.15	1.00 ± 0.15	1.00 ± 0.15
Parameters	**50 observations**			
μ	2.55 ± 1.07	2.43 ± 0.58	2.36 ± 0.78	2.31 ± 0.74
ϕ	1.00 ± 0.25	1.00 ± 0.24	1.01 ± 0.23	1.01 ± 0.23

this distribution, while samples for β, and hence μ, are drawn using rejection sampling (e.g., [7, Chapter 3]).

The results of the MCMC analysis of the simulations are presented in the Table 2 below. These results are generated by using MCMC to draw 7,000 parameter samples for each of the 1,000 simulated data sets of the 16 different combinations of cutoff and number of observations in Table 2. That is, a total of 16,000 MCMC analyses are run and each of these is run for 7,000 sample iterations. The first 2,000 samples of each MCMC run are discarded, and of the remaining 5,000 samples, every fifth sample is saved and used to estimate ϕ and μ. The medians of these 1,000 realizations is taken as the estimate of ϕ and μ for that simulated data set. in Table 2 presents summary statistics for the 1,000 median estimates for μ and ϕ from the 1,000 simulated data sets for each cutoff and number of observations combinations. An examination of the results in this table and those for the EM analysis in Table 1 reveals that the MCMC estimates are generally less biased for the dispersion parameter ϕ than the corresponding EM estimates, especially for the smaller sample sizes. For the mean parameter μ the means of the 1,000 median estimates for μ are generally more biased than the EM estimates.

4. Application

The data include analytes (e.g., total nitrogen, chlorophyll-a, e. coli, turbidity, etc.) and flows routinely measured at 35 different sites in the waterways that supply Sydney with drinking water for a period going back 20 years. These measures are used to assess the quality of water in the various reservoirs, rivers, and dams that are used to supply Sydney and greater Sydney with drinking water. Some measurements are taken daily (temperature, dissolved oxygen, turbidity, and pH), while others are collected roughly either fortnightly or monthly. The fitted models were used to assess the risk that particular analytes pose to the drinking water supply to Sydney residents.

Table 3: Fitted models replacing low detects with $\min(y)/2$, $c/2$, c or the EM E-step.

	Model parameters for the trend (μ)					
Missing value replacement	Const.	Flow on day t	Day/Time t	$\sin(2\pi t/365.25)$	$\cos(2\pi t/365.25)$	$\sin(2\pi t)$
$\min(y)/2$	−6.66	0.000644	0.00057	−0.0256	0.171	0.142
$c/2$	−6.71	0.000616	0.00060	−0.0248	0.189	0.167
c	−6.27	0.000809	0.00042	−0.0049	0.115	0.092
EM E-step	−6.70	0.000617	0.00059	−0.0287	0.182	0.116
			Model parameters for the spread (ϕ)			
Missing value replacement	$\cos(2\pi t)$		Const.	Day/Time$t$	$\sin(2\pi t)$	$\cos(2\pi t)$
$\min(y)/2$	0.771		0.82	0.000271	0.084	0.585
$c/2$	0.809		0.61	0.000341	−0.045	0.714
c	0.546		1.76	−0.000017	0.102	0.215
EM E-step	0.810		0.61	0.000336	0.147	0.660

We used AIC to select the distribution and model which was most appropriate for several water quality variables. The distributions considered were normal, log-normal, inverse gaussian, and zero-adjusted inverse Gaussian. The analytes that appeared to select the inverse Gaussian as the preferred distribution in more than 75% of the sites were firstly phosphorus filterable, nitorogen ammoniacal, and e. coli. Other analytes that where selected less frequently as inverse Gaussian distributed were entrococci, toxic cyanobacterial count, chloride, nitrogen oxide, and phosphorus total. The number of measurements ranged from several thousand to a few hundred with different number of low detects. All of these were successfully fitted by the approach advocated in the paper. As an example, we model phosphorus total(mg/L) at Bendeela Pondage in the Sydney water catchment supply area. This example is illustrated below.

Table 3 presents the result of the fitted model for phosphorus total (mg/L) at Bendeela Pondage. This was measured for the past 15 years with frequency of just over one per month (giving 213 measurements). There are seven low detects with one being below 0.001 and six being below 0.01. Not adjusting for this can adversely influence the model fit, because 0.01 is close to the mean value. Applying the above EM algorithm to this example results in a non-homogeneous mean and variance, which was fitted in R [8] by GAMLSS with family equal to the inverse Gaussian. The EM algorithm ran in a fraction of the time of the MCMC approach.

In addition, qq-plots of the theoretical values versus the quantile residuals were closer to each other after convergence using EM algorithm. In Table 3, setting the low detects equal to $c/2$ produces a fitted model that is closest to the EM fit. However, even with the propotionally few low detects in the sample, there are significant differences in some of the time-of-the-day harmonic coefficients for the trend and spread (i.e., sin / $\cos(2\pi t)$ coefficients).

5. Discussion

The paper offers a way of improving models with the modelled response having an inverse Gaussian distribution when the response variable has low or high detects. Although only one example is reported in the paper, the methodology has been successfully applied to many more examples for the Sydney Catchment Authority. Problems with convergence were

encountered in a few applications. The MCMC approach was computationally too intensive to run on all applications, but little appears to be lost by applying the EM algorithm. We used the AIC criterion to select the family from normal, log-normal, inverse Gaussian, or zero-adjusted inverse Gaussian (for which low detects were assumed zeros). This latter model is only realistic when the low detects are below all other measured values, hence, it is inappropriate in the first example in Section 4, where some low detects (i.e., 0.01) are close to the average response value. The high detection limit is seldom an issue in applications, but its theory was included, because it was encountered on a few occasions in the Sydney Catchment Authority work. The EM algorithm in this paper makes no attempt to achieve a maximum likelihood estimate for the variance; however, this limitation is of little consequence, since the variance estimate is reasonably comparable to the MCMC approach, which does properly account for the variance.

Numerical instabilities in the approaches were experienced in simulations, but results can be achieved by small adjustments to the estimated variances.

Appendix

Our aim in this appendix is to find $\mathbf{E}(\mathbf{x} \mid \mathbf{x} < \mathbf{c})$ and $\mathbf{E}(\mathbf{x} \mid \mathbf{x} > \mathbf{c})$ using the moment generating function (mgf) of $\mathbf{x}$ given $\mathbf{x} < \mathbf{c}$, denoted $\mathbf{M}_{\mathbf{x}|\mathbf{x}<\mathbf{c}}(\mathbf{t})$ or $\mathbf{M}_{\mathbf{x}|\mathbf{x}>\mathbf{c}}(\mathbf{t})$, respectively; for example, $\mathbf{E}(\mathbf{x} \mid \mathbf{x} < \mathbf{c}) = (\partial \mathbf{t}\mathbf{M}_{\mathbf{x}|\mathbf{x}<\mathbf{c}}(\mathbf{t}))/(\partial \mathbf{t})|_{\mathbf{t}=\mathbf{0}}$.

A. Inverse Gaussian Distribution

The inverse Gaussian distribution for random variable x is given by density

$$h(x,\mu,\phi) = \sqrt{\frac{\phi}{2\pi}}\, x^{-3/2} e^{-\phi(x-\mu)^2/(2\mu^2 x)} \tag{A.1}$$

and has distribution function $H(c,\mu,\phi) = \int_0^c h(x,\mu,\phi)dx = \Phi(((c/\mu) - 1)\sqrt{\phi/c}) + e^{2\phi/\mu}\Phi(-((c/\mu) + 1)\sqrt{\phi/c})$, where $\Phi(c) = \int_{-\infty}^{c} \sqrt{(1/2\pi)} e^{-z^2/2} dz$.

The mgf for the Inverse Gaussian distribution is given by

$$M_x(t) = \int_0^\infty e^{xt} h(x,\mu,\phi)dx = e^{(\phi/\mu)(1-\sqrt{1-2\mu^2 t/\phi})}. \tag{A.2}$$

By completing the square for the exponent terms in (A.1) and setting $\mu_* = \mu/\sqrt{1-2\mu^2 t/\phi}$, we have

$$\begin{aligned} -\frac{\phi(x-\mu)^2}{(2\mu^2 x)} + xt &= -\phi\frac{(x-\mu_*)^2}{(2\mu_*^2 x)} + \frac{\phi}{\mu}\left(1 - \sqrt{1 - \frac{2\mu^2 t}{\phi}}\right) \\ &= -\phi\frac{(x-\mu_*)^2}{(2\mu_*^2 x)} + \frac{\phi}{\mu} - \frac{\phi}{\mu_*}, \end{aligned} \tag{A.3}$$

so that for *low detects*, where we know that $x < c$, the mgf becomes

$$\begin{aligned} M_{x|x<c}(t) &= \int_0^c e^{xt} \frac{h(x,\mu,\phi)dx}{H(c,\mu,\phi)} \\ &= e^{\phi/\mu - \phi/\mu_*} \int_0^c \frac{h(x,\mu_*,\phi)dx}{H(c,\mu,\phi)} \\ &= e^{\phi/\mu - \phi/\mu_*} \frac{H(x,\mu_*,\phi)}{H(c,\mu,\phi)}. \end{aligned} \tag{A.4}$$

Since $\mu_* = \mu$ when $t = 0$, we have

$$\begin{aligned} E(x \mid x < c) &= \left\{ \frac{(\partial\mu_*/\partial t)\partial\left[e^{\phi/\mu - \phi/\mu_*} H(x,\mu_*,\phi)\right]/\partial\mu_*}{H(c,\mu,\phi)} \right\}_{t=0} \\ &= \left\{ \left(\frac{\mu_*^3}{\phi}\right) e^{\phi/\mu - \phi/\mu_*} \frac{\left[\phi H(x,\mu_*,\phi)/\mu_*^2 + \partial H(x,\mu_*,\phi)/\partial\mu_*\right]}{H(c,\mu,\phi)} \right\}_{t=0}. \\ \frac{\partial H(c,\mu,\phi)}{\partial\mu} &= \frac{\partial\left[\Phi\left((c/\mu - 1)\sqrt{\phi/c}\right) + e^{2\phi/\mu}\Phi\left(-(c/\mu + 1)\sqrt{\phi/\mu}\right)\right]}{\partial\mu} \\ &= -2\mu e^{2\phi/\mu}\Phi\left(-\left(\frac{c}{\mu} + 1\right)\sqrt{\frac{\phi}{c}}\right). \end{aligned} \tag{A.5}$$

Therefore

$$E(x \mid x < c) = \mu - 2\mu e^{2\phi/\mu} \frac{\Phi\left(-(c/\mu + 1)\sqrt{\phi/c}\right)}{H(c,\mu,\phi)}. \tag{A.6}$$

Similarly, for *high detects* with $x > c$,

$$E(x \mid x > c) = \mu + 2\mu e^{2\phi/\mu} \frac{\Phi\left(-(c/\mu + 1)\sqrt{\phi/c}\right)}{(1 - H(c,\mu,\phi))}. \tag{A.7}$$

See also Whitemore [5] which seems to often agree numerically with the result.

References

[1] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[2] P. Toscas, "MCMC and data augmentation for paramter estimation of censored data," CSIRO Technical Report, CMIS 07/92, 2007.

[3] R. A. Rigby and D. M. Stasinopoulos, "Generalized additive models for location, scale and shape," *Applied Statistics*, vol. 54, part 3, pp. 507–554, 2005.

[4] D. M. Stasinopoulos, R. A. Rigby, and C. Akantziliotou, "Instructionson how to use the GAMLSS package in R. Accompanying documentation in the current GAMLSS help files," 2006, http://www.londonmet.ac.uk/gamlss.
[5] G. A. Whitmore, "A regression method for censored inverse-Gaussian data," *The Canadian Journal of Statistics*, vol. 11, no. 4, pp. 305–315, 1983.
[6] G. Casella and E. George, "Explaining the Gibbs sampler," *The American Statistician*, vol. 46, no. 3, pp. 167–174, 1992.
[7] M. A. Tanner, *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, Springer, New York, NY, USA, 2nd edition, 1993.
[8] R Development Core Team, "R: a language and environment for statistical computing," Tech. Rep., 2005, http://www.R-project.org .