Comparison of methods for dealing with censored cost data

Pelham Barton and James Raftery

Health Economics Facility
University of Birmingham
Park House
40 Edgbaston Park Road
Birmingham
B15 2RT

This paper has been prepared for discussion at the Health Economists' Study Group meeting in Nottingham, July 2000.

Work in progress - please do not quote without permission

1. Background

A perennial problem in analysis of data from trials and similar studies is the problem of missing data. One form of this is where individuals are lost to follow-up before reaching the end-point of the study. Data from such individuals may be complete up to a given point of time, and completely missing thereafter. Such data is said to be *censored* (Armitage and Berry 1987 p. 421).

This work was prompted by the two papers on cost censoring given at HESG Newcastle in January 2000 (Young and Ratcliffe 2000, Raikou and McGuire 2000). These papers each applied a variety of methods to some actual data and compared the results. While the use of actual data ensures that the results are realistic, it has the disadvantage that the correct answer is unknown.

We decided to work with a complete data set, and apply various degrees of censoring to it. Simulation approaches have previously been used by Lin *et al* (1997), who used a fairly simple data set, with two levels of censoring, and Etzioni *et al* (1999), with only a single level of censoring. The benefit of using a complete data set is that the amount and form of censoring can be systematically varied, as shown below.

The results presented here are very much preliminary in nature. We welcome suggestions as to how the work can be extended.

2. The data set used

The data set used was generated from a simulation model concerned with the treatment of dyspepsia. This model was developed as part of an HTA-funded systematic review (HTA project code 96/37/01). An earlier version of the model was reported by Barton *et al* (1999).

Dyspepsia is a chronic condition which has a number of distinct causes, some of which are associated with the presence in the patient's stomach of the bacterium *Helicobacter pylori*. Various types of medication are available; their effectiveness depends on the type of dyspepsia from which the patient is suffering.

Patients enter the model on first consulting a GP with dyspepsia. The GP can apply one of a number of strategies for managing the patient. Strategies may involve medication alone, endoscopy or tests for *H. pylori*.

The model allows a large number of patient histories to be generated. All costs incurred during the patient's lifetime can be recorded, together with the exact times at which they were incurred. Costs occurring more than one year after initial consultation were discounted by 6 per cent per complete year from initial consultation. A data set of 5000 patients was generated using PPIs as the main form of medication. The benefit of the simulation approach is seen here in that large amounts of data may be generated at very little cost.

3. Forms of censoring

Given a data set with costs and times, various forms of censoring can be imposed. The simplest is to assume that patients have a constant risk of being lost to study throughout the duration of the study (in this case, the patient's lifetime). We simulated this by sampling for each patient from a standard negative exponential distribution and multiplying the resulting time by a factor m. The conditional probability that a patient will be censored by time $T + t,$ given that he has not been censored by time T is then approximately equal to $\frac{t}{m}$ for small values of t.

For a given value of m, a censored data set was obtained in the following way. First, the time to censoring for each patient was determined. If the time to censoring was greater than the difference between age at first consultation and age at death, then no censoring was deemed to have occurred and the complete patient record was kept. If the time to censoring was less than the difference between age at first consultation and age at death, then the patient data was recorded as censored and all costs after the censoring time were discarded. Note that in the example used here, a patient record may have been recorded as censored even though all the costs were in fact included.

The data set described in Section 2 were converted into a range of censored data sets in the following way. First, the data set was divided into ten sets of 500 patients each. For each such set, a sufficiently large value of m was found so that applying the

procedure in the previous paragraph led to no censoring at all; the value of m was then systematically reduced. Every time the number of censored entries was increased, the censored data set was analysed using the various methods described in the next section. The process was repeated with a different interpretation of the original random numbers.

We have used one form of censoring in which the risk is constant in time. Other forms may apply - on which we would welcome suggestions.

4. Methods of analysis

The main methods of analysis have been described elsewhere, in particular by Young and Ratcliffe (2000) and Raikou and McGuire (2000). Descriptions are given here mainly to clarify the terminology used in this paper.

The simplest of all methods is to ignore the issue of censoring altogether and simply calculate the mean of the recorded costs. This will be referred to as the "Full Sample" estimator.

Next is to omit the censored cases and calculate the mean cost for uncensored cases alone. This will be called the "Uncensored" esimator.

Fenn *et al* (1995) proposed the use the Kaplan-Meier method on the cost scale, treating costs as if they were survival times. This will be called the "Kaplan-Meier" estimator.

Lin *et al* (1997) have developed two estimators which use survival time as well as the distribution of costs. These estimators require dividing the total time interval under consideration into a number of (not necessarily equal) subintervals. Suppose that the i-th individual incurs costs c during the j-th time interval. Note that c will be zero for all time intervals after the death of the i-th individual, but may be non-zero for the interval during which the i-th individual dies. Then (assuming a complete data set), the average cost is

$$\frac{1}{N}\sum_i\sum_j c \ ,$$

where N is the number of individuals. This sum can be rearranged in a number of ways, two of which allow for censored data. The first rearrangement works by observing that

$$\frac{1}{N}\sum_i c \ = \frac{N_j}{N}\cdot\frac{1}{N_j}\sum_i c \ = S_j\bar{c}_j,$$

where

N_j is the number surviving into the j-th time interval,

S_j is the proportion surviving into the j-th time interval,

$\bar{c}_j$ is the average cost incurred during the j-th time interval, where the average is taken among those who survived into the j-th time interval.

In the presence of censored data, S_j may be estimated by the usual Kaplan-Meier estimator, while $\bar{c}_j$ may be estimated in one of two ways:

either the average cost incurred during the j-th time interval by all those known to survive into the j-th time interval,

or the average cost incurred during the j-th time interval by those known to survive into the j-th time interval and not censored before the end of the interval.

It can be seen that with only one time interval, this method reduces to the Full Sample or Uncensored estimator, depending on which choice is made for estimating $\bar{c}_j$. The version that reduces to the Full Sample estimator will be referred to here as “Lin M1”.

The possible disadvantage of Lin M1 is that it requires the full cost histories for all individuals, or at least the breakdown of costs into the time intervals used. An alternative which does not require this breakdown will now be described.

Suppose that the *i*-th individual has total costs C_i. Let T_j be the total cost of those who die during the *j*-th time interval. Then, again assuming complete data,

$$\frac{1}{N}\sum_i\sum_j c \;= \frac{1}{N}\sum_i C_i = \frac{1}{N}\sum_j T_j = \sum_j \frac{n_j}{N}\cdot\frac{1}{n_j}T_j = \sum_j \frac{N_j - N_{j+1}}{N}\cdot t_j = \sum_j \left(S_j - S_{j+1}\right)\cdot t_j,$$

where n_j is the number dying during the *j*-th time interval,

t_j is the average total cost incurred by those dying during the *j*-th time interval,

N_j and S_j are as before.

In the presence of censored data, Kaplan-Meier estimators can be used for S_j, while t_j can be estimated as the average costs of those known to die during the *j*-th time interval. This estimator will be referred to as “Lin M2”. Note that the Lin M2 estimator does not use the cost data from censored cases.

If follow-up is for a specific length of time rather than until death, the formula needs interpreting for the last interval. In that case, S_{j+1} is taken as zero and t_j is the average total cost among those who either die during the last time interval or survive to the end of follow-up.

A further point about the Lin estimators is that they require the time of censoring to be known. If all that is known is the date at which the last cost was occurred, then it is reasonable to assume that censoring occurred immediately after the last cost. When this assumption has been made, the estimators will be referred to as “Lin M1 A” and

5. Results

The data set using PPIs as the main form of medication was split into ten parts and each part was subjected to time-dependent censoring as described in Section 3, using two interpretations of the random numbers generated for each data set. The proportion of cases censored was allowed to go as high as 0.7. The process of increasing the number of cases censored gave an average of about 190 out of a possible 350 data sets in each case. The resulting censored data sets were then analysed using the seven methods described in Section 4. The results for each method were expressed as a percentage of the true value for the appropriate uncensored data set. Each method thus gave a range of percentages for a given number of cases censored. The two curves for each method, shown in Figures 1 to 5, represent the upper and lower outlines of a scatterplot of the results.

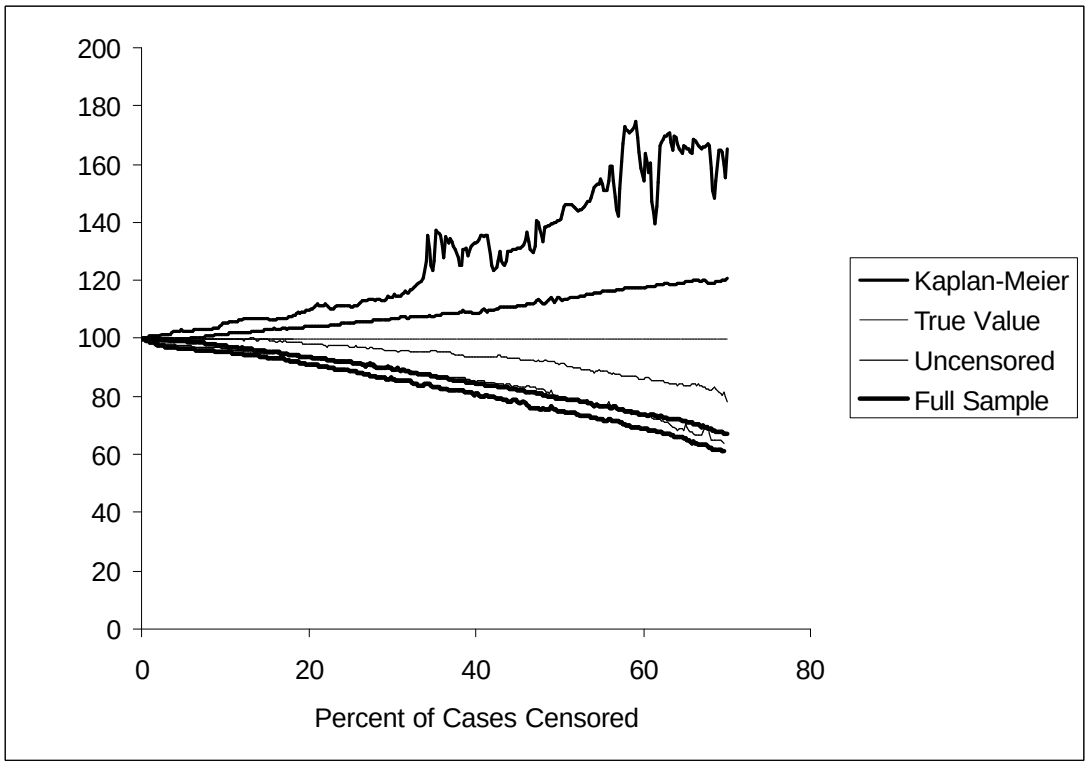

Figure 1. Simple estimators

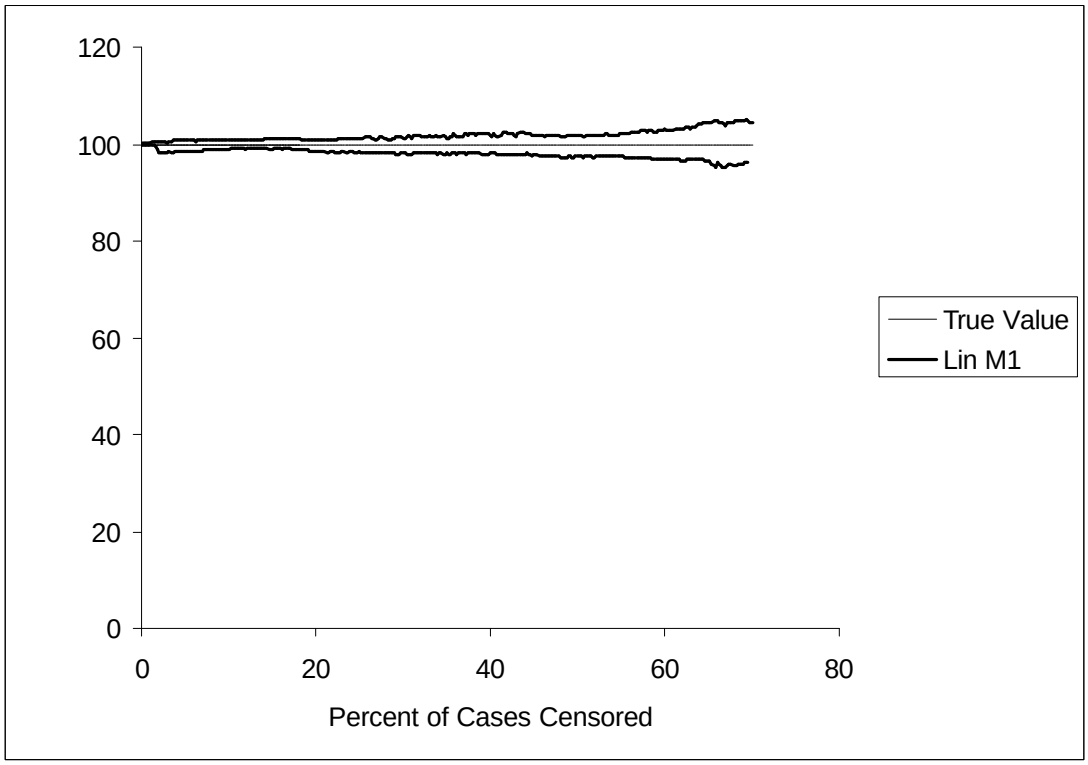

Figure 2. The Lin M1 estimator

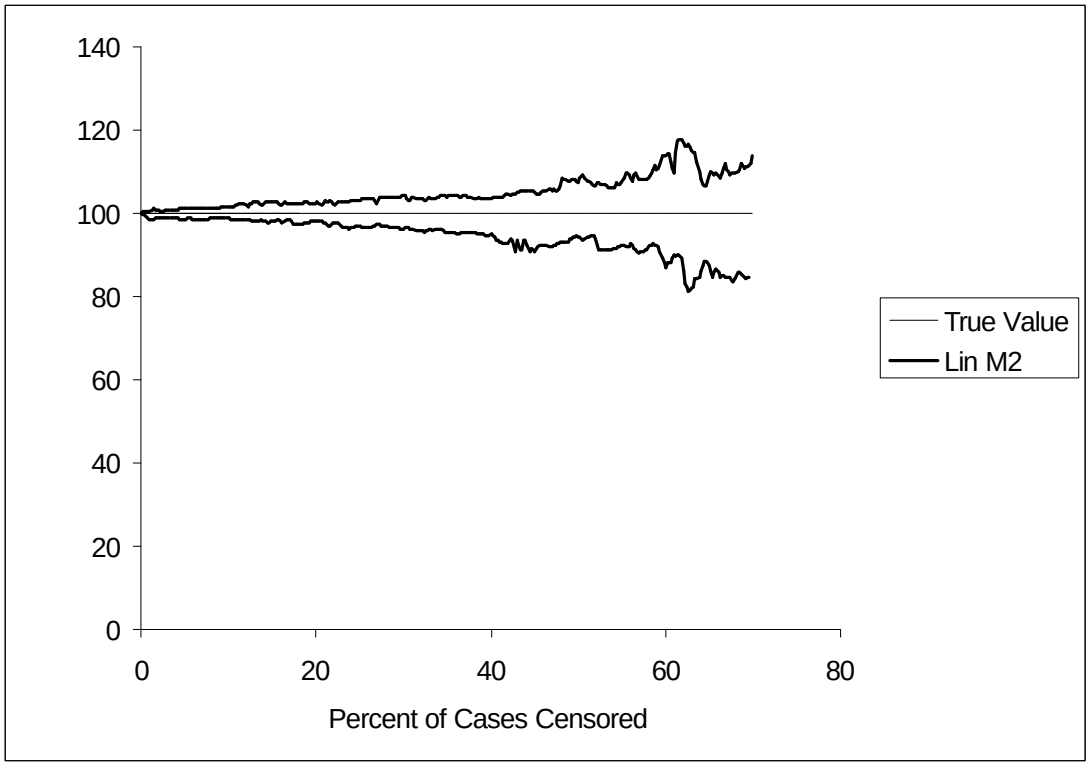

Figure 3. The Lin M2 estimator

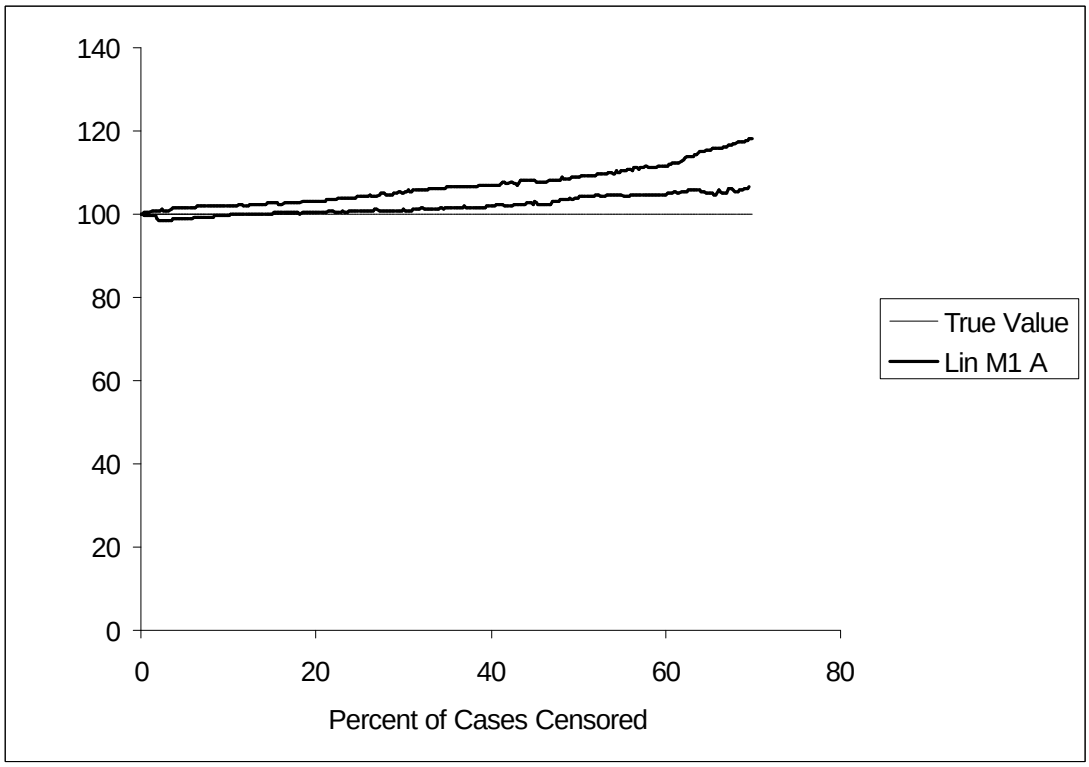

Figure 4. The Lin M1 A estimator

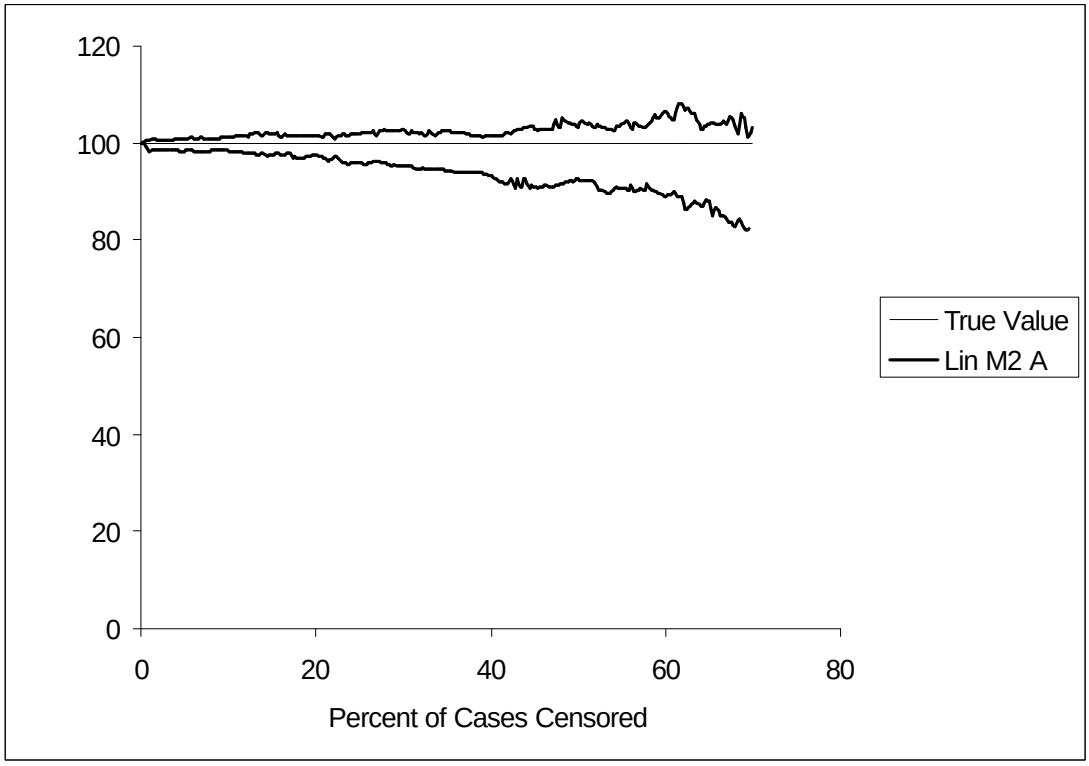

Figure 5. The Lin M2 A estimator

For the particular data set analysed here, the two simplest estimators give considerable underestimates of cost, while the Kaplan-Meier estimator is a serious overestimate. The two estimators of Lin *et al* do very well when the censoring dates are known, but slightly less well when an assumption about the censoring date has to be made. Note that Lin M1 A gives higher values than Lin M1 when censoring dates are known accurately; this is intuitively reasonable as time when no costs occur has been lost. This problem does not affect the Lin M2 estimator in the same way as costs from the censored cases are not used in that estimator.

6. Discussion

Various methods have been proposed for analysis of censored cost data. This paper shows a comparison between the best-known methods using a complete data set generated from a simulation model. When testing the methods, it is useful to have a complete data set on which to work, so that the correct answer is known. The methods of Lin *et al*, which are reasonably easy to calculate, have shown up well on the data set used here. However, the suggestion that these methods are preferred must remain tentative because

- only one type of data set has been used;
- only one form of censoring has been imposed.

It is highly desirable that the methods are tested on a wide variety of different (and realistic) data sets and that a range of realistic methods of censoring are employed. It may be that one method proves to be clearly the best in all cases, or that it is possible to find characteristics of data sets to choose an appropriate method. This is the subject of continuing work. We particularly welcome suggestions on other functional forms of censoring.

References

Armitage P and G Berry (1987) *Statistical Methods in Medical Research*, 2nd edition, Blackwell Scientific Publications, Oxford.

Barton PM, BC Delaney and P Moayyedi (1999) Management Strategies for Dyspepsia in Primary Care. Presentation to the Dyspepsia Trials Collaborators' Group. Copies available from the authors.

Etzioni RD, EJ Feuer, SD Sullivan, D Lin, C Hu and SD Ramsey (1999) On the use of survival analysis techniques to estimate medical care costs. *Journal of Health Economics* 18, 365-380.

Lin DY, EJ Feuer, R Etzioni and Y Wax (1997) Estimating medical costs from incomplete follow-up data. *Biometrics* 53, 113-128.

Raikou M and A McGuire (2000) Analysing Censored Cost Data in Trials. HESG Newcastle.

Young, Tracey and Julie Ratcliffe (2000) The Importance of Cost Censoring - The Case of Transplantation. HESG Newcastle.