The Stata Journal

The *Stata Journal* publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go "beyond the Stata manual" in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers of interest to those who teach, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

The *Stata Journal* is indexed and abstracted by *CompuMath Citation Index*, *Current Contents/Social and Behavioral Sciences*, *RePEc: Research Papers in Economics*, *Science Citation Index Expanded* (also known as *SciSearch*), *Scopus*, and *Social Sciences Citation Index*.

For more information on the *Stata Journal*, including information for authors, see the webpage

http://www.stata-journal.com

Subscriptions are available from StataCorp, 4905 Lakeway Drive, College Station, Texas 77845, telephone 979-696-4600 or 800-STATA-PC, fax 979-696-4601, or online at

http://www.stata.com/bookstore/sj.html

Subscription rates listed below include both a printed and an electronic copy unless otherwise mentioned.

U.S. and Canada		Elsewhere	
Printed & electronic		**Printed & electronic**	
1-year subscription	$115	1-year subscription	$145
2-year subscription	$210	2-year subscription	$270
3-year subscription	$285	3-year subscription	$375
1-year student subscription	$ 85	1-year student subscription	$115
1-year institutional subscription	$345	1-year institutional subscription	$375
2-year institutional subscription	$625	2-year institutional subscription	$685
3-year institutional subscription	$875	3-year institutional subscription	$965
Electronic only		**Electronic only**	
1-year subscription	$ 85	1-year subscription	$ 85
2-year subscription	$155	2-year subscription	$155
3-year subscription	$215	3-year subscription	$215
1-year student subscription	$ 55	1-year student subscription	$ 55

Back issues of the *Stata Journal* may be ordered online at

http://www.stata.com/bookstore/sjj.html

Individual articles three or more years old may be accessed online without charge. More recent articles may be ordered online.

http://www.stata-journal.com/archives.html

The *Stata Journal* is published quarterly by the Stata Press, College Station, Texas, USA.

Address changes should be sent to the *Stata Journal*, StataCorp, 4905 Lakeway Drive, College Station, TX 77845, USA, or emailed to sj@stata.com.

The *Stata Journal* (ISSN 1536-867X) is a publication of Stata Press. Stata, Stata, Stata Press, Mata, Mata, and NetCourse are registered trademarks of StataCorp LP.

The Stata Journal (2015)
15, Number 3, pp. 698–711

Estimation of mean health care costs and incremental cost-effectiveness ratios with possibly censored data

Shuai Chen
Department of Biostatistics and Medical Informatics
University of Wisconsin–Madison
Madison, WI
schen264@wisc.edu

Jennifer Rolfes
T-Mobile
Seattle, WA

Hongwei Zhao
Department of Epidemiology and Biostatistics
School of Rural Public Health
Texas A&M Health Science Center
College Station, TX

Abstract. In this article, we describe the `hcost` program for estimating mean health care costs and incremental cost-effectiveness ratios with possibly censored data. `hcost` estimates the mean survival time and the mean costs, as well as their variances and covariance, for a given time horizon. If the group variable is specified, `hcost` will report the differences between two groups, as well as the incremental cost-effectiveness ratio and its confidence interval (optional). `hcost` can estimate the mean costs using two methods corresponding to different types of data: the Bang and Tsiatis (2000, *Biometrika* 87: 329–343) estimator using only the total costs or the Zhao and Tian (2001, *Biometrics* 57: 1002–1008) estimator when cost-history data are available. The `hcost` program can also be used to specify the annual discounting rates for survival time and costs.

Keywords: st0399, hcost, mean costs, censored data, cost history, cost-effectiveness analysis

1 Introduction

Recently, the estimation of health care costs in economic evaluations of new treatments has received a lot of attention. In an environment of extremely high health care costs and limited resources, cost-effectiveness analysis to devise treatment strategies that offer a greater health benefit without imposing large economic burdens on society has become common.

In clinical trials and observational studies, survival time and accumulated health costs frequently are censored because not all patients can be observed until terminal events (for example, death) occur. Censoring poses a unique problem for cost estimation because of the "induced informative censoring" challenge, first noted by Lin et al.

(1997). As a result, traditional methods for handling censored survival data, such as the Kaplan–Meier estimator or Cox proportional hazards regression model, are no longer valid for analyzing censored cost data.

Because of the presence of censoring, the marginal distribution of cost may not be identifiable without making some parametric assumptions (Huang 2002). Thus most methods of estimating mean costs focus on restricted medical costs—that is, on costs accumulated within a time limit L, where L is chosen such that a reasonable number of subjects are still available at that time. Consequently, further cost-effectiveness analysis concentrates on time-restricted costs and effectiveness.

Different methods have been proposed for estimating the time-restricted mean costs. Lin et al. (1997) proposed estimators based on survival-probability weighting using partitioned time intervals; Bang and Tsiatis (2000) proposed several consistent estimators, including the Bang and Tsiatis (BT) estimator and its Bang and Tsiatis partitioned version (BTP) using the inverse-probability weighting technique; and Zhao and Tian (2001) proposed an efficient Zhao and Tian (ZT) estimator for when cost history is available. Later, Zhao et al. (2007) described the conditions under which the BT estimator is equivalent to the Lin T estimator (when cost history is not available) and the condition under which the ZT estimator is equivalent to the BTP and the two estimators Lin A and B proposed by Lin et al. (1997) (when cost history is available). Specifically, the equality occurs when the partition boundaries coincide with those censoring times. The BTP, the Lin T estimator, and the Lin A and B estimators may change with different ways of partitioning, while the BT and ZT estimators do not depend on any type of partitioning.

In this article, we describe how to implement the BT estimator (using the total costs only) and the ZT estimator (when cost history is available) in Stata to estimate mean health care costs with possibly censored data. If the data come from two treatment groups, a cost-effectiveness analysis can also be conducted, which estimates the incremental cost-effectiveness ratio and its confidence interval based on Fieller's method (Fieller 1954; Zhao and Tian 2001). We also discuss methods that allow discounting for costs and survival times.

2 Estimation of mean costs and mean survival time with censored data

For the ith individual in the study, $i = 1, 2, \ldots, n$, we define T_i as the survival time from study enrollment until the occurrence of some event (for example, death or disease relapse). Because some individuals are often still living when the study ends, these are considered censored observations. The censoring time for the ith individual is denoted as C_i. We can observe either the survival time or the censoring time, whichever is shorter; that is, we observe the follow-up time, $X_i = \min(T_i, C_i)$, and the death indicator variable, $\Delta_i = I(T_i \leq C_i)$. We define $M_i(t)$ as the accumulated costs for patient i from time 0 to t. For some applications, we observe only the total costs, $M_i = M_i(X_i)$. However, in other studies, we may know the entire cost history, $\{M_i(t), 0 < t < X_i\}$.

Here we briefly describe the methods for estimating the mean costs accumulated over time L with censored data, where L can be any time limit such that a reasonable number of subjects are still being observed at that time. The restricted survival time must be defined as $T_i^L = \min(T_i, L)$. Accordingly, the follow-up time becomes $X_i^L = \min(T_i^L, C_i)$, and the death indicator variable becomes $\Delta_i^L = I(T_i^L \leq C_i)$. We want to estimate the mean costs accumulated within time L, $\mu^M = E\{M(T_i^L)\}$. For clarity, we omit the superscript L in subsequent sections of this article.

Bang and Tsiatis (2000) proposed a consistent BT estimator based on the inverse-probability weighting technique.

$$\widehat{\mu}^M_{\mathrm{BT}} = \frac{1}{n}\sum_{i=1}^{n} \frac{\Delta_i M_i}{\widehat{K}(T_i)} \tag{1}$$

Here $\widehat{K}(T_i)$ is the Kaplan–Meier estimator for the survival function of the censoring time C, $K(u) = \Pr(C > u)$. The BT estimator is consistent and asymptotically normally distributed with a variance that can be estimated consistently (Bang and Tsiatis 2000) by

$$\widehat{\mathrm{Var}}(\widehat{\mu}^M_{\mathrm{BT}}) = \frac{1}{n^2}\sum_{i=1}^{n} \frac{\Delta_i (M_i - \widehat{\mu}^M_{\mathrm{BT}})^2}{\widehat{K}(T_i)} + \frac{1}{n^2}\sum_{i=1}^{n} \frac{1-\Delta_i}{\widehat{K}(C_i)^2}\{G_i(M^2) - G_i^2(M)\}$$

where

$$G_i(Z) = \frac{1}{n\widehat{S}(C_i)}\sum_{j=1}^{n} \frac{\Delta_j}{\widehat{K}(T_j)} Z_j I(T_j \geq C_i)$$

for any random variable Z and $\widehat{S}(u)$ is the Kaplan–Meier estimator for $S(u) = \Pr(T > u)$, which is the survival distribution of survival time T at time u.

With cost history available, the BT estimator is not efficient, because it does not use the cost information from censored observations. A more efficient estimator is proposed by Zhao and Tian (2001). The simplified form of the ZT estimator (Pfeifer and Bang 2005) is

$$\widehat{\mu}^M_{\mathrm{ZT}} = \frac{1}{n}\sum_{i=1}^{n} \frac{\Delta_i M_i}{\widehat{K}(T_i)} + \frac{1}{n}\sum_{i=1}^{n} \frac{(1-\Delta_i)\{M_i(C_i) - \overline{M(C_i)}\}}{\widehat{K}(C_i)} \tag{2}$$

where $\overline{M(C_i)} = \sum_{j=1}^{n} I(X_j \geq C_i) M_j(C_i) / \sum_{j=1}^{n} I(X_j \geq C_i)$, which is the average of accumulated costs at time C_i of those subjects who are living at C_i.

The ZT estimator is consistent and asymptotically normally distributed with variance that can be estimated consistently (Zhao and Tian 2001) by

$$\begin{aligned}\widehat{\mathrm{Var}}(\widehat{\mu}^M_{\mathrm{ZT}}) &= \frac{1}{n^2}\sum_{i=1}^{n}\frac{\Delta_i\left(M_i-\widehat{\mu}^M_{\mathrm{ZT}}\right)^2}{\widehat{K}(T_i)}+\frac{1}{n^2}\sum_{i=1}^{n}\frac{1-\Delta_i}{\widehat{K}(C_i)^2}\{G_i(M^2)-G_i^2(M)\}\\ &\quad -\frac{2}{n^2}\sum_{i=1}^{n}\frac{1-\Delta_i}{\widehat{K}(C_i)^2}[G_i\{MM(C_i)\}-G_i(M)G_i\{M(C_i)\}]\\ &\quad +\frac{1}{n^2}\sum_{i=1}^{n}\frac{1-\Delta_i}{\widehat{K}(C_i)^2}\left\{\overline{M(C_i)^2}-\overline{M(C_i)}^2\right\}\end{aligned}$$

where G_i and $\overline{M(C_i)}$ are defined previously, and

$$\overline{M(C_i)^2}=\frac{\sum_{j=1}^{n}I(X_j\geq C_i)M_j(C_i)^2}{\sum_{j=1}^{n}I(X_j\geq C_i)}$$

Meanwhile, the mean survival time to time L can be obtained by the area under the survival function; that is, $\widehat{\mu}^T=\int_0^L\widehat{S}(u)du$, where $\widehat{S}(u)$ is the Kaplan–Meier estimator for $S(u)=\Pr(T>u)$. This estimator can be obtained more conveniently (Satten and Datta 2001; Zhao and Tian 2001) by

$$\widehat{\mu}^T=\frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i T_i}{\widehat{K}(T_i)} \tag{3}$$

Following Zhao and Tian (2001), we can estimate its variance consistently by

$$\widehat{\mathrm{Var}}(\widehat{\mu}^T)=\frac{1}{n^2}\sum_{i=1}^{n}\frac{\Delta_i(T_i-\widehat{\mu}^T)^2}{\widehat{K}(T_i)}+\frac{1}{n^2}\sum_{i=1}^{n}\frac{1-\Delta_i}{\widehat{K}(C_i)^2}\{G_i(T^2)-G_i^2(T)\}$$

We can also estimate the covariance of $\widehat{\mu}^M_{\mathrm{BT}}$ and $\widehat{\mu}^T$ (Zhao and Tian 2001) by

$$\begin{aligned}\widehat{\mathrm{Cov}}\left(\widehat{\mu}^M_{\mathrm{BT}},\widehat{\mu}^T\right) &= \frac{1}{n^2}\sum_{i=1}^{n}\frac{\Delta_i M_i T_i}{\widehat{K}(T_i)}-\frac{1}{n^3}\sum_{i=1}^{n}\frac{\Delta_i M_i}{\widehat{K}(T_i)}\sum_{i=1}^{n}\frac{\Delta_i T_i}{\widehat{K}(T_i)}\\ &\quad +\frac{1}{n^2}\sum_{i=1}^{n}\frac{1-\Delta_i}{\widehat{K}(C_i)^2}\{G_i(TM)-G_i(M)G_i(T)\}\end{aligned}$$

We can estimate the covariance of $\widehat{\mu}^M_{\rm ZT}$ and $\widehat{\mu}^T$ (Zhao and Tian 2001) by

$$\begin{aligned}\widehat{\rm Cov}(\widehat{\mu}^M_{\rm ZT}, \widehat{\mu}^T) &= \frac{1}{n^2}\sum_{i=1}^n \frac{\Delta_i M_i T_i}{\widehat{K}(T_i)} - \frac{1}{n^3}\sum_{i=1}^n \frac{\Delta_i M_i}{\widehat{K}(T_i)} \sum_{i=1}^n \frac{\Delta_i T_i}{\widehat{K}(T_i)} \\ &+ \frac{1}{n^2}\sum_{i=1}^n \frac{1-\Delta_i}{\widehat{K}(C_i)^2}\{G_i(TM) - G_i(M)G_i(T)\} \\ &- \frac{1}{n^2}\sum_{i=1}^n \frac{1-\Delta_i}{\widehat{K}(C_i)^2}[G_i\{TM(C_i)\} - G_i\{M(C_i)\}G_i(T)]\end{aligned}$$

3 Discounting costs and survival time

In cost-effectiveness analysis, it is customary to discount the costs and survival time for a future time because a dollar and a year of life at present may be more valuable than a dollar and a year of life in the future. A 3% annual discount rate is used frequently in practice, but other discount rates may also be considered (Gold et al. 1996). Assume the annual discount rate is β. Consider a small time unit u, where the discount rate over the interval is βu. Then the present value for x_u dollars accumulated at some future time u is $x_0 = x_u(1-\beta u)$ dollars, and the present value for x_{2u} dollars at future time $2u$ is $x_0 = x_{2u}(1-\beta u)^2$ dollars, and so on. Therefore, the present value for x_{nu} dollars accumulated at future time nu is $x_0 = x_{nu}(1-\beta u)^n$. If $t = nu$, then the discounted value of x_t dollars at future time t is

$$x_0 = x_t(1-\beta u)^n = x_t(1-\beta u)^{\frac{t}{u}} \to x_t e^{-\beta t} \quad \text{as} \quad u \to 0$$

If $v(t)$ denotes costs to be incurred at future time t (note that the accumulated cost $M(t)$ is $M(t) = \int_0^t v(u)du$), its discounted cost at the beginning of the study is

$$v_0 = v^d(t) = e^{-\beta t} v(t)$$

In the input data for our `hcost` program, each cost entry is associated with a start date and a stop date. If the option for discounting cost (`drcost()`) is chosen, the program will discount the accumulated costs using its start date. When the time interval between start and stop dates for this cost entry is shorter than the user-specified `k()` (days), the accumulated costs are discounted using the start date. However, if the time interval is longer than `k()`, then the interval is divided into `k()`-day subintervals, and the costs at each subinterval are discounted at the start date of this subinterval. The default value for `k()` is `k(90)`.

A similar strategy can be used to discount years of life in the future at an annual discount rate of β, as if "life" values were measured in dollars. For one year of life in the future time t, its present value is $e^{-\beta t}$. Therefore, the current value (discounted value) of survival time T is

$$T^d = \int_0^T e^{-\beta s} ds = \frac{1}{\beta}(1 - e^{-\beta T})$$

4 Incremental cost-effectiveness ratio

If a new treatment has a greater health benefit but lower costs when compared with its competitor, the new strategy is undoubtedly preferred. However, if a program has higher costs but a greater health benefit than its competitor, a decision must be made about which of the two programs to adopt. The incremental cost-effectiveness ratio (ICER) is a useful measure under these circumstances.

The ICER is defined as the additional costs one must pay for saving one additional year of life. Mathematically, it can be expressed as $\gamma = (\mu_2^M - \mu_1^M)/(\mu_2^T - \mu_1^T)$. It can be estimated by plugging in the ZT estimator (2) or BT estimator (1) for the mean cost $\widehat{\mu}_k^M$ as well as the estimator for mean survival time (3), $\widehat{\mu}_k^T$, for each group k, $k = 1, 2$. Bootstrap methods (Efron and Tibshirani 1986, 1993; Hwang 1995; Mushlin et al. 1998; Jiang, Wu, and Williams 2000; O'Brien and Briggs 2002; Jiang and Zhou 2004) are commonly used to construct confidence intervals (CIs) for the ICER. Although many researchers believe that the bootstrap method provides better coverage, Hwang (1995) and Jiang, Wu, and Williams (2000) show that the bootstrap method is only first-order accurate, similar to the Fieller method (Fieller 1954). Therefore, the Fieller method, if used correctly, can be a reliable and computationally efficient way to obtain CIs for the ICER.

Zhao and Tian (2001) used Fieller's Theorem to obtain CIs for the ICER. Because asymptotically $x = \widehat{\mu}_2^M - \widehat{\mu}_1^M$ and $y = \widehat{\mu}_2^T - \widehat{\mu}_1^T$ are bivariate normally distributed, the $100(1 - 2\alpha)\%$ confidence limits for the ICER γ are

$$\frac{xy - z_\alpha^2 s_{xy} \pm \{(xy - z_\alpha^2 s_{xy})^2 - (x^2 - z_\alpha^2 s_{xx})(y^2 - z_\alpha^2 s_{yy})\}^{\frac{1}{2}}}{y^2 - z_\alpha^2 s_{yy}} \tag{4}$$

where s_{xx}, s_{yy}, and s_{xy} are respectively the variances of x and y and the covariance of x and y, and z_α is the cutoff point with tail area α of the standard normal distribution. Because the two treatment groups are independent, s_{xx}, s_{yy}, and s_{xy} can be estimated by the summation of $\widehat{\mathrm{Var}}(\widehat{\mu}_k^M)$, $\widehat{\mathrm{Var}}(\widehat{\mu}_k^T)$, and $\widehat{\mathrm{Cov}}(\widehat{\mu}_k^M, \widehat{\mu}_k^T)$ over the two groups k=1, 2, respectively.

Note that the cost-effectiveness analysis is normally performed when one treatment group is simultaneously more effective and more costly than the other group. The formula (4) can be used to obtain a finite CI only if there is a statistically significant difference between the mean survival times of the two treatment groups—that is, if the denominator of (4) is positive. If the denominator of (4) is negative, indicating that the difference between the effects of two treatments is not statistically significant, the CI for the ICER is exclusive and thus infinite; that is, the CI is formed by $(-\infty, CL_L) \cup (CL_U, +\infty)$, where CL_L and CL_U are the lower and upper limits obtained from (4). Although this infinite CI is correct and reflects our knowledge about the true ICER, its interpretation can be difficult (Stinnett and Mullahy 1998; Briggs, Mooney, and Wonderling 1999; Wang and Zhao 2008). Alternative methods, such as the cost-effectiveness acceptability curve, may be considered in this situation (Fenwick et al. 2006; Löthgren and Zethraeus 2000). For this reason, we do not rou-

tinely perform a cost-effectiveness analysis, but we do provide an option for calculating the ICER and its CI.

5 The hcost command

5.1 Syntax

`hcost` *idvar costvar* [*if*] [*in*] `, start(`*varname*`) stop(`*varname*`) l(#)`
[`group(`*varname*`) drsurv(#) drcost(#) icer(#) k(#) level(#)`
`method(#)`]

The `hcost` command estimates mean costs within a time horizon L with possibly censored data. The data must be processed by the `stset` command (see [ST] **stset**) before using `hcost`.

5.2 Options

`start(`*varname*`)` specifies the variable that contains the start time of accumulated costs (in days). `start()` is required unless `method()` is `0` and `drcost()` is `0`.

`stop(`*varname*`)` specifies the variable that contains the end time of accumulated costs (in days). `stop()` is required unless `method()` is `0` and `drcost()` is `0`.

`l(#)` specifies the time limit L (in days) for calculating mean costs and mean survival time. `l()` is required.

`group(`*varname*`)` specifies the variable that contains the treatment group information if there is more than one group. If there is only one group, this option need not be specified. There cannot be more than two groups.

`drsurv(#)` specifies the annual discount rate for survival time. The value must be between 0 and 1. The default is `drsurv(0)`; that is, survival time is already discounted in the data.

`drcost(#)` specifies the annual discount rate for costs. The value must be between 0 and 1. The default is `drcost(0)`; that is, cost is already discounted in the data. Each cost entry is discounted according to its start date. If the time interval for the cost entry is too big, it will be divided into `k()`-day subintervals first.

`icer(#)` specifies the option for performing cost-effectiveness analysis: `1` for calculating the ICER and its CI and `0` otherwise. The default is `icer(0)`.

`k(#)` specifies the number of days in a time interval for discounting costs. The default is `k(90)`. When the time interval between start and stop dates is smaller than `k()`, the costs accumulated are discounted using the start date; if the time interval is bigger than `k()`, then the interval is divided into `k()`-day subintervals first, and the costs at each subinterval are discounted at the start date of the subinterval.

`level(#)` specifies the level for CIs. The value must be between 10 and 99. The default is `level(95)`.

`method(#)` specifies the method for mean costs estimation: `1` for using cost history (ZT estimator), `0` for using only total costs (BT estimator). The default is `method(1)`. If `method()` is `0` and `drcost()` is `0`, `start()` and `stop()` are not required. When `start()` and `stop()` are not provided, the costs are assumed to be collected within the time limit L.

The ZT estimator requires the arguments `start()` and `stop()` to be provided. On the other hand, because the BT estimator requires only total costs, the arguments `start()` and `stop()` may be omitted for the BT method. However, if the user wants to discount the costs with the BT method, `start()` and `stop()` are required for discounting purposes.

6 Example

To illustrate the `hcost` command, we use a small simulated dataset as an example. Among the patients in our data, 80 received the conventional therapy (group 0), and 80 received the new treatment (group 1). The first enrolled patient was followed for 69 months and the last for less than 1 month, with an average follow-up of 27 months. Cost data were simulated with start and stop dates for each entry.

Table 1 shows, as an example, several rows of the dataset. Each patient has a unique `id` variable; a survival time variable, `surv`; a death indicator variable, `delta`; a treatment variable, `trt`; and one or more `cost` entries representing different types of costs or costs accumulated at different time intervals defined by the start date variable, `start`, and the stop date variable, `stop`.

Table 1. Some observations from the dataset used as an example

id	start	stop	cost	trt	delta	surv
1	1	1	3694	0	0	575
1	1	9	1	0	0	575
1	1	9	12	0	0	575
1	1	34	106	0	0	575
2	5	11	2	0	0	1166
2	5	26	275	0	0	1166
1001	1	1	0	1	0	1138
1001	1	7	49325	1	0	1138
1001	1	23	112	1	0	1138
1009	1	1	115	1	1	425
1009	1	10	25368	1	1	425
1009	1	24	41	1	1	425

After importing the data, we declare the data as survival data using the command `stset`. This dataset is multiple-record-per-subject survival data; that is, there may be more than one row belonging to one patient. For the BT estimator, the dataset can also be single-record-per-subject survival data as long as the total costs for each subject are provided in one record.

```
. use example
. * Declare the data as survival data by stset before using hcost
. stset surv, failure(delta)
     failure event:  delta != 0 & delta < .
obs. time interval:  (0, surv]
 exit on or before:  failure
------------------------------------------------------------------------------
       9882  total observations
          0  exclusions
------------------------------------------------------------------------------
       9882  observations remaining, representing
       2988  failures in single-record/single-failure data
   10270570  total analysis time at risk and under observation
                                                 at risk from t =         0
                                      earliest observed entry t =         0
                                           last observed exit t =      2082
```

We use the command `hcost` to estimate the mean survival time and the mean costs. The time limit L is chosen to be 1,461 days (that is, 4 years), at which time there are still enough subjects under observation. We estimate the mean survival time and the mean costs for the whole population using the default ZT estimator without discounting either costs or survival time.

```
. hcost id cost, start(start) stop(stop) l(1461)
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
cost	80134.84	4870.968	16.45	0.000	70587.92	89681.76
survival	1164.878	42.04446	27.71	0.000	1082.473	1247.284

```
Method used: ZT (using cost history)
Annual discounting rate for costs: 0
Annual discounting rate for survival time: 0
Time limit L (in days): 1461
```

Next, we show an example where an improperly chosen L may lead to problems. We set L to be 3,000 (days), which is larger than the largest observation time in the dataset (2,082 days).

```
. hcost id cost, start(start) stop(stop) l(3000)
Warning: The time limit l is too large (greater than the last observation time).
r(198);
```

We then calculate the ZT estimator with an annual discounting rate of 3% for both costs and survival time.

```
. hcost id cost, start(start) stop(stop) l(1461) drsurv(0.03) drcost(0.03)
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
cost	77578.64	4602.394	16.86	0.000	68558.12	86599.17
survival	1101.906	39.12913	28.16	0.000	1025.215	1178.598

```
Method used: ZT (using cost history)
Annual discounting rate for costs: .03
Annual discounting rate for survival time: .03
Time limit L (in days): 1461
```

Next, we perform the estimation for the two treatment groups separately using the BT estimator with an annual discounting rate of 3%. This is achieved using the `group()` argument.

```
. hcost id cost, start(start) stop(stop) l(1461) group(trt) method(0) drsurv(0.03)
> drcost(0.03)
Estimates for Group 1 (trt=0)
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
cost	64927.99	7803.73	8.32	0.000	49632.96	80223.02
survival	951.7195	63.09655	15.08	0.000	828.0525	1075.386

```
Estimates for Group 2 (trt=1)
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
cost	107624	9618.589	11.19	0.000	88771.95	126476.1
survival	1252.687	39.27494	31.90	0.000	1175.709	1329.664

```
Estimates for Difference Between Groups (Group 2 - Group 1)
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
cost	42696.05	12386.1	3.45	0.001	18419.74	66972.36
survival	300.9673	74.32157	4.05	0.000	155.2997	446.6349

```
Method used: BT (using total costs)
Annual discounting rate for costs: .03
Annual discounting rate for survival time: .03
Time limit L (in days): 1461
```

We can also obtain the ICER and its CI by using the argument `icer(1)`. In this example, we use the default ZT estimator and an annual discounting rate of 3%.

```
. hcost id cost, start(start) stop(stop) l(1461) group(trt) drsurv(0.03)
> drcost(0.03) icer(1)
Estimates for Group 1 (trt=0)
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
cost	64171.55	6553.233	9.79	0.000	51327.45	77015.65
survival	951.7195	63.09655	15.08	0.000	828.0525	1075.386

```
Estimates for Group 2 (trt=1)
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
cost	92181.74	5535.045	16.65	0.000	81333.25	103030.2
survival	1252.687	39.27494	31.90	0.000	1175.709	1329.664

```
Estimates for Difference Between Groups (Group 2 - Group 1)
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
cost	28010.18	8577.971	3.27	0.001	11197.67	44822.7
survival	300.9673	74.32157	4.05	0.000	155.2997	446.6349

```
Method used: ZT (using cost history)
Annual discounting rate for costs: .03
Annual discounting rate for survival time: .03
Time limit L (in days): 1461
Incremental cost-effectiveness ratio (95% CI): 93.067 (38.133, 189.761)
```

From the two examples shown above, we can see that, as expected, the BT estimator produces larger standard errors than the ZT estimator because it does not use the cost history of censored and uncensored subjects.

7 Stored results

`hcost` stores the following in `e()` (when `group()` is not specified):

Scalars

`e(n)`	number of patients	`e(pcensor)`	censoring rate
`e(nobs)`	number of cost observations		

Matrices

`e(b)`	coefficient vector	`e(V)`	variance–covariance matrix of the estimators

If `group()` is specified, similar results are stored in `e(n1)`, `e(n2)`, `e(nobs1)`, `e(nobs2)`, `e(pcensor1)`, `e(pcensor2)`, `e(b1)`, `e(b2)`, `e(V1)`, and `e(V2)` for groups 1 and 2.

The command to list the stored results is shown below.

```
. ereturn list
scalars:
                  e(n1) =  80
               e(nobs1) =  4869
            e(pcensor1) =  .49
                  e(n2) =  80
               e(nobs2) =  5013
            e(pcensor2) =  .75
matrices:
                  e(b1) :  1 x 2
                  e(V1) :  2 x 2
                  e(b2) :  1 x 2
                  e(V2) :  2 x 2
```

The following commands display the estimated mean costs, mean survival time, their variances, and covariance for group 1:

```
.* vector of mean costs and mean survival time for Group 1
. matrix list e(b1)
e(b1)[1,2]
         cost   survival
r1  64171.554   951.7195
. * Variances and covariance between mean costs and mean survival time for Group 1
. matrix list e(V1)
symmetric e(V1)[2,2]
               cost   survival
    cost   42944860
survival  113828.12  3981.1744
```

8 Further comments

A program implementing Lin's estimators for censored costs was developed by Kim and Thompson (2011). However, Lin's estimator is appropriate only when cost data are available as grouped (for example, monthly) data. Our program can easily handle the cost data with any start and stop times using the ZT estimator and therefore is more flexible.

In addition to the survival time, quality-adjusted life-years (QALY) are also widely used to measure effectiveness of treatments. QALY are subject to the same kind of "informative censoring" challenge as costs are (Zhao and Tsiatis 1997, 1999; Willan, Lin, and Manca 2005). Hence, the analysis of QALY should not follow traditional survival techniques, similarly to the analysis of cost data. In our future work, we will extend the program such that QALY also can be used as a measure of effectiveness.

Currently, `hcost` estimates the mean costs and the mean survival time (limited to a time horizon L) without using any covariate information. However, it is often of interest to investigate how covariates affect the costs and effectiveness of different therapies. Therefore, our future work will attempt to incorporate covariate information into our program.

9 References

Bang, H., and A. A. Tsiatis. 2000. Estimating medical costs with censored data. *Biometrika* 87: 329–343.

Briggs, A. H., C. Z. Mooney, and D. E. Wonderling. 1999. Constructing confidence intervals for cost-effectiveness ratios: An evaluation of parametric and non-parametric techniques using Monte Carlo simulation. *Statistics in Medicine* 18: 3245–3262.

Efron, B., and R. J. Tibshirani. 1986. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science* 1: 54–75.

———. 1993. *An Introduction to the Bootstrap*. New York: Chapman & Hall/CRC.

Fenwick, E., D. A. Marshall, A. R. Levy, and G. Nichol. 2006. Using and interpreting cost-effectiveness acceptability curves: An example using data from a trial of management strategies for atrial fibrillation. BMC *Health Services Research* 6: 52.

Fieller, E. C. 1954. Some problems in interval estimation. *Journal of the Royal Statistical Society, Series B* 16: 175–185.

Gold, M. R., J. E. Siegel, L. B. Russell, and M. C. Weinstein. 1996. *Cost-effectiveness in Health and Medicine*. New York: Oxford University Press.

Huang, Y. 2002. Calibration regression of censored lifetime medical cost. *Journal of the American Statistical Association* 97: 318–327.

Hwang, J. T. G. 1995. Fieller's problems and resampling techniques. *Statistica Sinica* 5: 161–171.

Jiang, G., J. Wu, and G. R. Williams. 2000. Fieller's interval and the bootstrap-Fieller interval for the incremental cost-effectiveness ratio. *Health Services and Outcomes Research Methodology* 1: 291–303.

Jiang, H., and X.-H. Zhou. 2004. Bootstrap confidence intervals for medical costs with censored observations. *Statistics in Medicine* 23: 3365–3376.

Kim, L. G., and S. G. Thompson. 2011. Estimation of life-years gained and cost effectiveness based on cause-specific mortality. *Health Economics* 20: 842–852.

Lin, D. Y., E. J. Feuer, R. Etzioni, and Y. Wax. 1997. Estimating medical costs from incomplete follow-up data. *Biometrics* 53: 419–434.

Löthgren, M., and N. Zethraeus. 2000. Definition, interpretation and calculation of cost-effectiveness acceptability curves. *Health Economics* 9: 623–630.

Mushlin, A. I., W. J. Hall, J. Zwanziger, E. Gajary, M. Andrews, R. Marron, K. H. Zou, A. J. Moss, and for the MADIT Investigators. 1998. The cost-effectiveness of automatic implantable cardiac defibrillators: Results from MADIT. *Circulation* 97: 2129–2135.

O'Brien, B. J., and A. H. Briggs. 2002. Analysis of uncertainty in health care cost-effectiveness studies: An introduction to statistical issues and methods. *Statistical Methods in Medical Research* 11: 455–468.

Pfeifer, P. E., and H. Bang. 2005. Non-parametric estimation of mean customer lifetime value. *Journal of Interactive Marketing* 19: 48–66.

Satten, G. A., and S. Datta. 2001. The Kaplan-Meier estimator as an inverse-probability-of-censoring weighted average. *American Statistician* 55: 207–210.

Stinnett, A. A., and J. Mullahy. 1998. Net health benefits: A new framework for the analysis of uncertainty in cost-effectiveness analysis. *Medical Decision Making* 18: S68–S80.

Wang, H., and H. Zhao. 2008. A study on confidence intervals for incremental cost-effectiveness. *Biometrical Journal* 50: 505–514.

Willan, A. R., D. Y. Lin, and A. Manca. 2005. Regression methods for cost-effectiveness analysis with censored data. *Statistics in Medicine* 24: 131–145.

Zhao, H., H. Bang, H. Wang, and P. E. Pfeifer. 2007. On the equivalence of some medical cost estimators with censored data. *Statistics in Medicine* 26: 4520–4530.

Zhao, H., and L. Tian. 2001. On estimating medical cost and incremental cost-effectiveness ratios with censored data. *Biometrics* 57: 1002–1008.

Zhao, H., and A. A. Tsiatis. 1997. A consistent estimator for the distribution of quality adjusted survival time. *Biometrika* 84: 339–348.

———. 1999. Efficient estimation of the distribution of quality-adjusted survival time. *Biometrics* 55: 1101–1107.

About the authors

Shuai Chen is a research associate in the Department of Biostatistics and Medical Informatics at the University of Wisconsin–Madison. She received a PhD in statistics from Texas A&M University and has expertise in statistical methodologies of cost-effectiveness analysis, survival analysis, longitudinal data analysis, and the latent-variable model related to health care.

Jennifer Rolfes received her master's degree in statistics from Texas A&M University. She previously worked as a statistician at StataCorp and is now a business analysis manager at T-Mobile in Seattle, WA.

Dr. Hongwei Zhao is a professor in the Department of Epidemiology and Biostatistics, School of Rural Public Health, Texas A&M Health Science Center. She received her doctoral degree in biostatistics from Harvard in 1997. Dr. Zhao has collaborated extensively with many researchers from different fields such as community medicine, neurology, cancer, and cardiovascular diseases. Her main research interests are survival analysis and longitudinal data analysis. She has made significant contributions in the area of estimating quality-adjusted survival time and medical costs analyses with censored data.