Buckley–James method for analyzing censored data, with an application to a cardiovascular disease and an HIV/AIDS study

James Cui
Monash University, Melbourne, Australia
james.cui@med.monash.edu.au

Abstract. The Buckley–James method and the Cox proportional hazards model were proposed in the 1970s. Both methods can be used to analyze survival-type data, although the former focuses on calculation of the expected value of the survival time and the latter on the relative risk of explanatory variables on the failure event. In cardiovascular disease epidemiological studies, it is essential to correct the effect of taking antihypertensive medicine, which means we need to calculate the expected blood pressure for people who take the medicine. I developed a Stata program to calculate the Buckley–James estimate. I will describe how to use this program to calculate the expected value of a censored outcome and illustrate the method through an example from a cardiovascular disease and an HIV/AIDS study.

Keywords: st0093, Buckley–James method, censoring, expectation, survival

1 Introduction

The Buckley–James method (Buckley and James 1979) was proposed in 1979, a few years after Sir David Cox proposed the Cox proportional hazards model (Cox 1972). Both methods can be used to analyze survival-type data, although the former focuses on calculation of the expected value of the survival time and the latter on the relative risk of explanatory variables on the failure event. Today, the Cox model dominates the analysis of survival data. I examine whether the Buckley–James method can be used to calculate the expected survival time. Stare, Harrell, Jr., and Heinzel (2001) suggested that the lack of use of the Buckley–James method in the past 20 years is due to lack of appropriate software. I developed a Stata program to perform the Buckley–James calculation.

In cardiovascular disease epidemiological studies, when a person takes antihypertensive medicine, the blood pressure (BP) measured on this individual is usually lower than the inherited blood pressure that the person should have. For finding the right loci of the genes that determine the BP, the inherited BP (not the measured BP) should be used in relevant analyses (Terwillinger 2002; Olson 2002). We need to use appropriate statistical methods to adjust the measured BP to a higher level (the so-called restored BP) to be close to the inherited BP. This is also called the correction of the effect of antihypertensive medicine, which is the same as calculation of the expected blood pressure (Cui, Hopper, and Harrap 2002, 2003).
Statistically speaking, the inherited BP is right-censored by the measured BP for people who take antihypertensive medicine (Turnbull 1976). We do not know exactly the value of the inherited BP, but we do know that it is at least the value of the measured BP. The problem that we need to address here is the estimation of the expected value of the censored BP when relevant covariates are taken into account. The Buckley–James method fits this purpose. This method models the censored outcome as a linear function of the covariates, while the Kaplan–Meier estimator (Kaplan and Meier 1958) estimates the distribution of the random error.

I will describe how to use the Buckley–James method to calculate the expected value of a censored outcome and illustrate the method using an example from a cardiovascular disease and an HIV/AIDS study. The analysis was performed using Stata 9.

2 The Buckley–James method

The following notations are used. For a sample of $n$ individuals, let $\Delta_i$ denote the treatment status of the $i$th individual; i.e., $\Delta_i = 1$ if the person is treated and $\Delta_i = 0$ otherwise, $i = 1, 2, \ldots, n$. Let $Y_i$ represent the value of the inherited BP, which is observed when $\Delta_i = 0$ and right-censored when $\Delta_i = 1$. In the latter case, although we do not know exactly what $Y_i$ is, we observe another quantity, $Z_i$, and know that $Z_i < Y_i$.

Buckley and James (1979) proposed a least-squares method to estimate the parameters in the model

$$Y_i = \alpha + X_i\beta + \varepsilon_i$$

where $i = 1, 2, \ldots, n$ and $\beta$ is a $p \times 1$ vector of regression coefficients corresponding to the $1 \times p$ covariate vector $X_i$, and $p$ is the number of covariates. No specific parametric form was assumed for the distribution function of the error term $\varepsilon_i$. The distribution can be estimated nonparametrically from the Kaplan–Meier product-limit estimator (Kaplan and Meier 1958) as

$$\hat{F}(\varepsilon) = 1 - \prod_{\varepsilon_i \leq \varepsilon} \left(1 - \frac{d_i}{n_i}\right)$$

where $d_i = \sum_{j=1}^{n} I(\varepsilon_j = \varepsilon_i \text{ and } \Delta_j = 0)$ and $n_i = \sum_{j=1}^{n} I(\varepsilon_j > \varepsilon_i)$.

If we weighted the observed and censored outcome as

$$Y_i^* = Y_i(1 - \Delta_i) + E(Y_i | Y_i > Z_i)\Delta_i$$

it can be shown that $E(Y_i^*) = E(Y_i)$; for details see Smith (2002). When $\Delta_i = 1$, the second term in (2) becomes

$$E(Y_i | Y_i > Z_i) = \alpha + \beta X_i + E\{\varepsilon_i | \varepsilon_i > Z_i - (\alpha + \beta X_i)\}$$
where the expectation
\[ E \{ \varepsilon_i | \varepsilon_i > Z_i - (\alpha + \beta X_i) \} = \int_{Z_i - (\alpha + \beta X_i)}^{\infty} \frac{\varepsilon dF}{1 - F \{ Z_i - (\alpha + \beta X_i) \}} \]

After substituting \( F \) with \( \hat{F} \) in (1), the realization of \( Y_i^* \) becomes
\[
y_i^* = y_i(1 - \delta_i) + \alpha + \beta x_i + \left( \sum_{\varepsilon_i > \varepsilon_j} w_j \varepsilon_j \right) \delta_i
\]

where \( w_j \) are the steps of \( \hat{F} \) at \( \varepsilon_j \). Based on \( y_i^* \) and \( x_i \), the Buckley–James estimator of \( \alpha \) and \( \beta \) can be obtained by the least-squares principle
\[
\hat{\beta} = \frac{(x - \bar{x})y_i^*}{(x - \bar{x})(x - \bar{x})}, \quad \hat{\alpha} = \hat{y}^* - \hat{\beta} \bar{x}
\]

where \( x \) is the design matrix and \( \bar{x} \) is a vector of means of the covariates. Given an initial value of \( \alpha \) and \( \beta \), their final estimates can be calculated iteratively from the above estimation (3) through (4). Then the restored outcome \( y^* \) can be obtained accordingly from (3) by replacing \( \alpha \) and \( \beta \) with their final estimates. The standard error of \( y^* \) can be obtained using the bootstrap method.

Some authors have also investigated the effect of adding extra covariates to the linear regression model (Cook and Weisberg 1982). A plot can be drawn to assess whether there is any relationship between the scatterplot of \( y^* \) and a new covariate, and the scatterplot of \( y^* \) adjusted for covariates already in the model and a new covariate. However, these methods are beyond the scope of this paper.

### 3 The buckley program

#### 3.1 Syntax

```
buckley depvar treatvar varlist [ if ] [ in ] [ , iterate(#) tolerance(#) dispnum(#)
```

#### 3.2 Description

`buckley` uses the Buckley–James method (Buckley and James 1979) to estimate the regression coefficients and generate the expected value of the censored outcome. `depvar` is the dependent variable whose value is right-censored when the treatment variable `treatvar = 1`. Otherwise, it is observed exactly when `treatvar = 0`. `varlist` is a list of names of covariate variables. At least one covariate variable must be specified.
3.3 Options

iterate(#) specifies the maximum number of iterations allowed in the optimization, which must be a positive integer. The default is iterate(100).

tolerance(#) specifies the convergence criterion for the change in the sum of squares of the difference of regression coefficients between iterations. It must be a positive number between 0 and 1. The default is tolerance(1e-6).

dispnum(#) specifies how frequently to display the iteration; e.g., dispnum(50) displays every 50th iteration. # must be a positive integer. The default is dispnum(100).

3.4 Remarks

Within the Stata program, a variable called untreated = 1 – treatment was generated, indicating whether an individual’s outcome is not censored. The indicator variable untreated is what we usually used in the survival analysis, specifying whether an individual has a failure event.

The Stata program generates a matrix, called coefbj, to store the regression coefficients and generates a new variable, called varbj, to store the expected value of the censored outcome. If a user wants to rerun the program using the current dataset, it is better to rename varbj to avoid this variable’s being replaced by the newly generated variable. The program can be slow if the dataset is large or the number of covariates in the model is large.

4 Example 1

The bp data come from a cardiovascular disease epidemiological study (Harrap et al. 2000), which consists of 767 pairs of parents and 15.6% (n = 240) of them have taken an antihypertensive medicine. Because only a few offspring take the antihypertensive medicine and hence have censored BP, I dropped offspring data in this analysis.

The following tables list the relevant variable names used in the analysis and their first six observations. The variable familyid is a unique number shared by all family members (here the pair of parents), and id is a unique number for each individual in the study. Variable sbp is the measured systolic blood pressure (unit mm Hg), and treatment indicates whether an individual takes an antihypertensive medicine. Variables age, gender, and bmi are the age, gender, and body-mass index of an individual, respectively.
. use bp
. describe
Contains data from bp.dta
obs: 1,534
vars: 7 27 Mar 2005 10:00
size: 39,884 (96.2% of memory free) (_dta has notes)

<table>
<thead>
<tr>
<th>storage</th>
<th>display</th>
<th>value</th>
<th>variable name</th>
<th>type</th>
<th>format</th>
<th>label</th>
<th>variable label</th>
</tr>
</thead>
<tbody>
<tr>
<td>familyid</td>
<td>float</td>
<td>%9.0g</td>
<td>Family number</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>id</td>
<td>long</td>
<td>%12.0g</td>
<td>Individual number</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sbp</td>
<td>float</td>
<td>%9.0g</td>
<td>Systolic blood pressure</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>treatment</td>
<td>byte</td>
<td>%8.0g</td>
<td>Anti-hypertensive medicine (1/0)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>age</td>
<td>float</td>
<td>%9.0g</td>
<td>Age (year)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gender</td>
<td>byte</td>
<td>%8.0g</td>
<td>Gender (Male 1, Female 0)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bmi</td>
<td>float</td>
<td>%9.0g</td>
<td>Body mass index</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Sorted by: familyid id
. list in 1/6, sep(0) ab(9) noobs

<table>
<thead>
<tr>
<th>familyid</th>
<th>id</th>
<th>sbp</th>
<th>treatment</th>
<th>age</th>
<th>gender</th>
<th>bmi</th>
</tr>
</thead>
<tbody>
<tr>
<td>10189</td>
<td>1018919</td>
<td>106.5</td>
<td>0</td>
<td>54</td>
<td>0</td>
<td>27.28</td>
</tr>
<tr>
<td>10189</td>
<td>1018920</td>
<td>132</td>
<td>0</td>
<td>60</td>
<td>1</td>
<td>28.09</td>
</tr>
<tr>
<td>20021</td>
<td>2002119</td>
<td>130.5</td>
<td>0</td>
<td>66</td>
<td>0</td>
<td>20.13</td>
</tr>
<tr>
<td>20021</td>
<td>2002120</td>
<td>142.25</td>
<td>0</td>
<td>68</td>
<td>1</td>
<td>30.8</td>
</tr>
<tr>
<td>20022</td>
<td>2002219</td>
<td>139</td>
<td>1</td>
<td>62</td>
<td>0</td>
<td>33.65</td>
</tr>
<tr>
<td>20022</td>
<td>2002220</td>
<td>126</td>
<td>0</td>
<td>68</td>
<td>1</td>
<td>29.06</td>
</tr>
</tbody>
</table>

### 4.1 Fitting the Buckley–James method to the data

We used the following command to estimate the Buckley–James regression coefficients, with the option of a maximum of 50 iterations and the convergence criterion set to 1e–10. Every 25th iteration will be displayed, so we can see the process of the iteration procedure. The regression coefficients will be listed when the iteration procedure finishes. They can also be obtained by listing the matrix `coefbj` separately.

. buckley sbp treatment age gender bmi, i(50) t(1e-10) d(25)
(output omitted)

Regression coefficients:

<table>
<thead>
<tr>
<th>coefbj[1,4]</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
</tr>
<tr>
<td>varbj</td>
</tr>
</tbody>
</table>

From the above output, the expected outcome is

\[
\text{sbp} = 61.3777 + 0.669 \text{age} + 3.056 \text{gender} + 1.163 \text{bmi}
\]
The following table shows the summary statistics of the restored outcome $\text{varbj}$ and the original measured outcome $\text{sbp}$. For untreated individuals, the values of $\text{varbj}$ are the same as those of $\text{sbp}$. However, for treated individuals, the mean of $\text{varbj}$ is higher than that of $\text{sbp}$ (152.2 vs. 135.8). But the standard deviation of $\text{varbj}$ is lower than that of $\text{sbp}$ (14.3 vs. 17.0). Although $\text{varbj}$ has the same maximum value as $\text{sbp}$ does, the minimum value of $\text{varbj}$ is higher than that of $\text{sbp}$ (125.6 vs. 91.5).

\begin{verbatim}
. tabstat sbp varbj, by(treatment) stats(n mean sd min max) col(stats) f(%7.2f)
> long
  treatment  variable     N    mean   sd   min   max
0       sbp     1294.00   125.61  14.36   87.00  203.50
          varbj   1294.00   125.61  14.36   87.00  203.50
1       sbp     240.00    135.84  16.95   91.50  209.50
          varbj   240.00    152.18  14.25  125.55  209.50
Total  sbp     1534.00    127.21  15.25   87.00  209.50
          varbj   1534.00    129.77  17.28   87.00  209.50
\end{verbatim}

We want to know how much was added to the original measured BP by the Buckley–James method for each treated individual. We generate a new variable called $\text{addition}$ and draw the scatterplot of $\text{addition}$ against $\text{sbp}$, as shown in figure 1. The larger the measured SBP, the smaller the amount of addition. The mean of the addition is 16.3 mm Hg (median 14.9 and SD 4.5), which is close to the values given by Neaton et al. (1993). They showed that the treatment effect for different drugs ranges from 9.1 to 15.9 mm Hg in the reduction of the systolic BP.

\begin{verbatim}
. gen addition = varbj - sbp if treatment==1
(1294 missing values generated)
. scatter addition sbp, xlab(80(20)200) ytitle("Addition")
> xtitle("Measured BP (mm Hg)"
\end{verbatim}
5 Example 2

Here we consider how to calculate the expected survival time from HIV infection to AIDS diagnosis. The data aids comprise 51 HIV-infected individuals (Selvin 1995). The variable id is a unique number for each individual in the study. Variable time specifies the analysis time (months) from HIV infection to AIDS diagnosis (i.e., incubation period) or to the end of the study. Variable aids indicates whether an individual has an AIDS diagnosis (1 for yes and 0 for no). Variable age is the age (years) at the time of HIV infection.

To use the buckley program, we need to generate a new variable, called censor, to indicate whether an individual is censored by the AIDS diagnosis.

```
. use aids, clear
   (AIDS (Selvin 1995:453))
. gen censor = 1 - aids
. gen age30 = age - 30
. buckley time censor age30, i(20) t(1e-6) d(10)
   (output omitted)
Regression coefficients:
   coefbj[1,2]
   age30    _cons
   varbj   -2.3148   83.1084
```
The expected time from HIV infection to AIDS diagnosis for those who have not developed AIDS is

\[ \hat{t} = 83.1084 - 2.3148(\text{age} - 30) \]

From the following table, we see that the expected incubation period (\texttt{varbj}) is longer than the censored incubation period (\texttt{time}) for those who have not developed AIDS, and the mean is 104.7 and 81.9 months, respectively. For those who have developed AIDS, the expected incubation period is the same as the observed incubation period.

<table>
<thead>
<tr>
<th></th>
<th>\texttt{time}</th>
<th>\texttt{varbj}</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>26.00</td>
<td>81.92</td>
</tr>
<tr>
<td></td>
<td>26.00</td>
<td>104.74</td>
</tr>
<tr>
<td>1</td>
<td>25.00</td>
<td>41.36</td>
</tr>
<tr>
<td></td>
<td>25.00</td>
<td>41.36</td>
</tr>
<tr>
<td>Total</td>
<td>51.00</td>
<td>62.04</td>
</tr>
<tr>
<td></td>
<td>51.00</td>
<td>73.67</td>
</tr>
</tbody>
</table>

6 Conclusion

The Buckley–James method can be used to calculate the expected value of a censored outcome as in survival analysis. It can also be used to correct the censored measurements to their underlying values when appropriate covariates are considered in a linear regression model. The Stata program \texttt{buckley} does this.

7 Acknowledgment

I thank Professor Stephen Harrap for providing the blood-pressure data used in this article.
8 Saved results

buckley saves the following results in r():

Scalars
r(N) number of observations
r(k) number of covariates
r(iter) number of iterations
r(sumsq) sum of squares of difference in coefficients between iterations

 Macros
r(depvar) name of dependent variable
r(treatvar) name of treatment variable
r(covar) names of covariates
r(varbj) name of variable for expected outcome

Matrices
coefbj coefficient vector

9 References


Buckley–James method


About the Author

James Cui is a Senior Research Fellow in the Department of Epidemiology and Preventive Medicine, Monash University, Melbourne, Australia.