

The World's Largest Open Access Agricultural & Applied Economics Digital Library

This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.

Help ensure our sustainability.

Give to AgEcon Search

AgEcon Search
http://ageconsearch.umn.edu
aesearch@umn.edu

Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.

THE STATA JOURNAL

Editor

H. Joseph Newton Department of Statistics Texas A & M University College Station, Texas 77843 979-845-3142; FAX 979-845-3144 jnewton@stata-journal.com

Associate Editors

Christopher Baum Boston College

Rino Bellocco Karolinska Institutet

David Clayton Cambridge Inst. for Medical Research

Mario A. Cleves

Univ. of Arkansas for Medical Sciences

William D. Dupont Vanderbilt University

Charles Franklin

University of Wisconsin, Madison

Joanne M. Garrett

University of North Carolina

Allan Gregory Queen's University

James Hardin

University of South Carolina

Ben Jann ETH Zurich, Switzerland

Stephen Jenkins University of Essex

Ulrich Kohler WZB, Berlin

Jens Lauritsen

Odense University Hospital

Stata Press Production Manager Stata Press Copy Editors

Editor

Nicholas J. Cox Geography Department **Durham University** South Road Durham City DH1 3LE UK n.j.cox@stata-journal.com

Stanley Lemeshow Ohio State University

J. Scott Long Indiana University

Thomas Lumley University of Washington, Seattle

Roger Newson

King's College, London

Marcello Pagano Harvard School of Public Health

Sophia Rabe-Hesketh

University of California, Berkeley

J. Patrick Royston MRC Clinical Trials Unit, London

Philip Ryan University of Adelaide

Mark E. Schaffer

Heriot-Watt University, Edinburgh

Jeroen Weesie Utrecht University

Nicholas J. G. Winter Cornell University

Jeffrey Wooldridge Michigan State University

Lisa Gilmore Gabe Waggoner, John Williams

Copyright Statement: The Stata Journal and the contents of the supporting files (programs, datasets, and help files) are copyright © by StataCorp LP. The contents of the supporting files (programs, datasets, and help files) may be copied or reproduced by any means whatsoever, in whole or in part, as long as any copy or reproduction includes attribution to both (1) the author and (2) the Stata Journal.

The articles appearing in the Stata Journal may be copied or reproduced as printed copies, in whole or in part, as long as any copy or reproduction includes attribution to both (1) the author and (2) the Stata Journal.

Written permission must be obtained from StataCorp if you wish to make electronic copies of the insertions. This precludes placing electronic copies of the Stata Journal, in whole or in part, on publicly accessible web sites, fileservers, or other locations where the copy may be accessed by anyone other than the subscriber.

Users of any of the software, ideas, data, or other materials published in the Stata Journal or the supporting files understand that such use is made without warranty of any kind, by either the Stata Journal, the author, or StataCorp. In particular, there is no warranty of fitness of purpose or merchantability, nor for special, incidental, or consequential damages such as loss of profits. The purpose of the Stata Journal is to promote free communication among Stata users.

The Stata Journal, electronic version (ISSN 1536-8734) is a publication of Stata Press, and Stata is a registered trademark of StataCorp LP.

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 Heinzl (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 Δ_i denote the treatment status of the ith 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, ..., n and β 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 ε_i . The distribution can be estimated nonparametrically from the Kaplan–Meier product-limit estimator (Kaplan and Meier 1958) as

$$\widehat{F}(\varepsilon) = 1 - \prod_{\varepsilon_i < \varepsilon} \left(1 - \frac{d_i}{n_i} \right) \tag{1}$$

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

$$Y_i^* = Y_i(1 - \Delta_i) + E(Y_i | Y_i > Z_i)\Delta_i$$
 (2)

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

$$E(Y_i | Y_i > Z_i) = \alpha + \beta X_i + E\left\{\varepsilon_i | \varepsilon_i > Z_i - (\alpha + \beta X_i)\right\}$$

where the expectation

$$E\left\{\varepsilon_{i} \mid \varepsilon_{i} > Z_{i} - (\alpha + \beta X_{i})\right\} = \int_{Z_{i} - (\alpha + \beta X_{i})}^{\infty} \frac{\varepsilon dF}{1 - F\left\{Z_{i} - (\alpha + \beta X_{i})\right\}}$$

After substituting F with \hat{F} in (1), the realization of Y_i^* becomes

$$y_i^* = y_i(1 - \delta_i) + \alpha + \beta x_i + \left\{ \frac{\sum_{\varepsilon_j > \varepsilon_i} w_j \varepsilon_j}{1 - \widehat{F}(\varepsilon_i)} \right\} \delta_i$$
 (3)

where w_j are the steps of \widehat{F} at ε_j . Based on y_i^* and x_i , the Buckley–James estimator of α and β can be obtained by the least-squares principle

$$\widehat{\beta} = \frac{(x - \bar{x})y_i^*}{(x - \bar{x})'(x - \bar{x})}, \qquad \widehat{\alpha} = \bar{y}^* - \widehat{\beta}\bar{x}$$
(4)

where x is the design matrix and \bar{x} is a vector of means of the covariates. Given an initial value of α and β , 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 α and β 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][n], \underline{i} terate(#) \underline{t} olerance(#) \underline{d} ispnum(#)

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 $\mathtt{untreated} = 1 - \mathtt{treatment}$ was generated, indicating whether an individual's outcome is *not* censored. The indicator variable $\mathtt{untreated}$ 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

vars: 7 27 Mar 2005 10:00 size: 39,884 (96.2% of memory free) (_dta has notes)

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

Sorted by: familyid id

. list in 1/6, sep(0) ab(9) noobs

familyid	id	sbp	treatment	age	gender	bmi
10189 10189	1018919 1018920	106.5 132	0	54 60	0 1	27.28 28.09
20021	2002119	130.5	0	66	0	20.13
20021	2002120	142.25	0	68	1	30.8
20022	2002219	139	1	62	0	33.65
20022	2002220	126	0	68	1	29.06

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.

From the above output, the expected outcome is

$$\widehat{\text{sbp}} = 61.377 + 0.669\,\text{age} + 3.056\,\text{gender} + 1.163\,\text{bmi}$$

The following table shows the summary statistics of the restored outcome varbj and the original measured outcome sbp. For untreated individuals, the values of varbj are the same as those of sbp. However, for treated individuals, the mean of varbj is higher than that of sbp (152.2 vs. 135.8). But the standard deviation of varbj is lower than that of sbp (14.3 vs. 17.0). Although varbj has the same maximum value as sbp does, the minimum value of varbj is higher than that of sbp (125.6 vs. 91.5).

. 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

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 addition and draw the scatterplot of addition against 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.

```
. 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)")
```

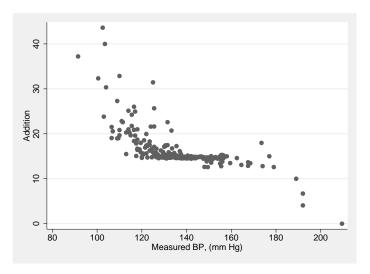


Figure 1: Scatterplot of the amount of addition against the measured BP.

tabstat addi	ition,	stats(mean	p50	sd	min	max)	for	mat(%7	7.2f)	
variable		mean		p50			sd		min		max
addition		16.34	14	1.92		4.	48		0.00	4	3.61

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.

The expected time from HIV infection to AIDS diagnosis for those who have not developed AIDS is

$$\widehat{\text{time}} = 83.1084 - 2.3148(\text{age} - 30)$$

From the following table, we see that the expected incubation period (varbj) is longer than the censored incubation period (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.

	tabstat	time	varbi.	by(aids)	stats(n mean	sd min max)	col(stats)	f(%7.2f)	long
•	cabbcac	CIMC	varb,	Dy (arab)	boacb (II mcan	bu min max/	COT (BUTES)	1 (/01 • 21/	TOHE

aids	variable	N	mean	sd	min	max
0	time	26.00	81.92	22.74	24.00	97.00
	varbj	26.00	104.74	12.42	68.69	121.77
1	time	25.00	41.36	25.77	1.00	84.00
	varbj	25.00	41.36	25.77	1.00	84.00
Total	time	51.00	62.04	31.57	1.00	97.00
	varbj	51.00	73.67	37.68	1.00	121.77

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 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) r(iter)	number of observations number of iterations	r(k) r(sumsq)	number of covariates sum of squares of difference in coefficients between iterations
Macros r(depvar) r(covar)	name of dependent variable names of covariates	r(treatvar) r(varbj)	name of treatment variable name of variable for expected outcome
Matrices coefbj	coefficient vector		

9 References

- Buckley, J. and I. James. 1979. Linear regression with censored data. *Biometrika* 66: 429–436.
- Cook, R. D. and S. Weisberg. 1982. Residuals and Influence in Regression. New York: Chapman & Hall.
- Cox, D. R. 1972. Regression models and life-tables (with discussion). Journal of the Royal Statistical Society, Series B 34: 187–220.
- Cui, J., J. L. Hopper, and S. B. Harrap. 2002. Genes and family environment explain correlations between blood pressure and body mass index. *Hypertension* 40: 7–12.
- ———. 2003. Antihypertensive treatments obscure familial contributions to blood pressure variation. *Hypertension* 47: 207–210.
- Harrap, S. B., M. Stebbing, J. L. Hopper, H. N. Hoang, and G. G. Giles. 2000. Familial patterns of covariation for cardiovascular risk factors in adults: The Victorian Family Heart Study. American Journal of Epidemiology 152: 704–715.
- Kaplan, E. L. and P. Meier. 1958. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 53: 457–481.
- Neaton, J. D., R. H. Grimm, Jr., R. J. Prineas, J. Stamler, G. A. Grandits, P. J. Elmer, J. A. Cutler, J. M. Flack, J. A. Schoenberger, and R. McDonald. 1993. Treatment of Mild Hypertension Study. Final results. Treatment of Mild Hypertension Study Research Group. Journal of the American Medical Association 270: 713–724.
- Olson, J. M. 2002. Linkage analysis, model-free. In *Biostatistical Genetics and Genetic Epidemiology*, ed. R. C. Elston, J. Olson, and L. Palmer, 460–472. Chichester, UK: Wiley.
- Selvin, S. 1995. Practical Biostatistical Methods. Pacific Grove, CA: Duxbury.
- Smith, P. J. 2002. Analysis of Failure and Survival Data. Boca Raton, FL: Chapman & Hall/CRC.

- Stare, J., F. E. Harrell, Jr., and H. Heinzl. 2001. BJ: an S-plus program to fit linear regression models to censored data using the Buckley-James method. *Computer Methods and Programs in Biomedicine* 64: 45–52.
- Terwillinger, J. D. 2002. Linkage analysis, model-based. In *Biostatistical Genetics and Genetic Epidemiology*, ed. R. C. Elston, J. Olson, and L. Palmer, 448–460. Chichester, UK: Wiley.
- Turnbull, B. W. 1976. The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society, Series B* 38: 290–295.

About the Author

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