The Stata Journal (2002)
2, Number 2, pp. 164–182

G-estimation of causal effects, allowing for time-varying confounding

Jonathan A. C. Sterne
University of Bristol, UK
jonathan.sterne@bristol.ac.uk

Kate Tilling
King's College London
kate.tilling@bristol.ac.uk

Abstract. This article describes the `stgest` command, which implements G-estimation (as proposed by Robins) to estimate the effect of a time-varying exposure on survival time, allowing for time-varying confounders.

Keywords: st0014, G-estimation, time-varying confounding, survival analysis

1 Introduction

In this article, we describe the use of G-estimation to estimate causal effects. This method is used in studies where subjects are studied over a period of time, and the subject characteristics are measured at the start of the study (the *baseline* measurements) and on a number of subsequent occasions. Such studies are known as *cohort studies* by epidemiologists and as *panel studies* by social scientists. Subjects are *followed-up* until the occurrence of the outcome event, or until they are *censored* (e.g., because they reach the scheduled end of follow-up or because they withdraw from the study). The outcome event could be death from a particular cause, or the occurrence of a particular disease or other life event (e.g., the first successful job application for a panel of job seekers). The time between the start of follow-up and the occurrence of the outcome event is called the *failure time*.

Our aim is to identify factors associated with the occurrence of the outcome event. We will call such factors *exposures*; these could be risk factors for disease (such as alcohol consumption) or treatment interventions (e.g., antiretroviral therapy for HIV-infected patients). We will deal only with binary exposure variables, for which subjects can always be classified as exposed or unexposed to the risk factor or treatment. The control of *confounding* is a fundamental problem in the analysis and interpretation of such studies. A confounding variable (*confounder*) is one that is associated with both the occurrence of the outcome and with the exposure of interest. For example, smoking will usually confound the association between alcohol consumption and the occurrence of cancer. Variables on the causal pathway between exposure and the outcome event should not be treated as confounders. For example, when estimating the effect of an antihypertensive (blood pressure-lowering) drug on the occurrence of heart disease, we should not control for blood pressure after the start of treatment. Controlling for a covariate that is intermediate on the pathway between exposure and outcome will estimate only the direct effect of the exposure (ignoring the effect mediated through the covariate).

Exposure effects controlled for confounding may be estimated via stratification (e.g., using Mantel–Haenszel methods) or by using regression models that include both the exposure and the confounder(s) as covariates. We will focus on Cox and Weibull regression models for the analysis of cohort studies. When exposures and confounders are measured repeatedly, we may estimate their association with the outcome by splitting follow-up time into the periods between measurements, and assuming that the values measured at the start of the period remain constant until the next measurement occasion. We will refer to such estimates as *time updated* effects.

The problem addressed here is that *standard methods for the analysis of cohort studies can lead to biased estimates of time-updated exposure effects*. This is because of *time-varying confounding*. As defined by Mark and Robins (1993), a covariate is a *time-varying confounder* for the effect of exposure on outcome if

1. past covariate values predict current exposure, and
2. current covariate value predicts outcome.

If, in addition, past exposure predicts current covariate value, then standard survival analyses with time-updated exposure effects will give biased exposure estimates, whether or not the covariate is included in the model.

For example, consider a study to estimate the effect of antiretroviral therapy (ART) on AIDS-free survival in patients infected with HIV. Markers of disease progression (e.g., CD4 counts) are used to decide when to administer ART, but are also affected by ART. CD4 count is a *time-varying confounder* for the effect of ART on survival times because

1. past values of CD4 count predict whether an individual is treated (condition 1), and
2. CD4 count predicts survival time (condition 2).

In addition, ART affects subsequent CD4 count, and so standard approaches to the analysis of time-updated exposure effects will give biased estimates of the effect of ART. For example, analyses of the effect of ART on survival times could employ three possible strategies:

1. The crude estimate (not controlled for confounding) of the effect of ART will be biased, because ART tends to be given to individuals who are more immunosuppressed (their CD4 count is low) and who therefore tend to experience higher rates of AIDS and death.
2. Controlling for the baseline values of confounders such as CD4 count will still give biased estimates of the effect of ART, because this ignores the fact that individuals who started treatment after the start of the study will tend to be those who became immunosuppressed.

3. Controlling for time-updated measurements of confounders such as CD4 count will still give biased estimates of the effect of ART, because ART acts at least partly by raising CD4 counts. Such models would therefore ignore the effect of ART, which acts through raising CD4 count.

2 Methods

The method of G-estimation of causal effects in the presence of time-varying confounding was introduced by Robins; see, for example, Robins et al. (1992), Witteman et al. (1998), or Tilling et al. (2002). We briefly outline the method here.

The concept of the *counterfactual* failure time is fundamental to G-estimation. For subject i, the counterfactual failure time U_i is defined as the failure time that would have occurred if the subject had been unexposed throughout follow-up. U_i is called the counterfactual failure time because it is unobservable for subjects who were exposed at any time. For subjects who were unexposed throughout follow-up, U_i is equal to their observed failure time. We assume that exposure accelerates failure time by a factor $\exp(-\psi)$, which we will call the *causal survival time ratio*. The purpose of the G-estimation procedure is to estimate the unknown parameter ψ.

If ψ were known, then for a subject who experienced the outcome event and who was exposed throughout follow-up, U_i would be equal to their observed failure time multiplied by $\exp(\psi)$, since

$$\text{Failure time if continuously exposed} = \text{Failure time if unexposed} \times \exp(-\psi)$$

and so

$$\begin{aligned} U_i &= \text{Failure time if continuously unexposed} \\ &= \text{Failure time if continuously exposed} \times \exp(\psi) \end{aligned}$$

Similarly, for any subject who experienced the outcome event at time T_i, the counterfactual failure time $U_{i,\psi}$ could be derived from the observed failure time by

$$U_{i,\psi} = \int_0^{T_i} \exp(\psi e_i(t))dt \tag{1}$$

where $e_i(t)$ is 1 if subject i is exposed at time t and 0 if subject i is unexposed. As explained earlier, we assume that exposure is constant between measurement occasions. For example, if subject i experienced the outcome event at 5 years and was exposed for three of these, then

$$U_{i,\psi} = 3\exp(\psi) + 2$$

However, for the reasons given in the introduction, in the presence of time-varying confounding, ψ cannot be estimated from the data using standard methods (e.g., using a Weibull or other accelerated failure time model).

G-estimation provides estimates of ψ, allowing for time-varying confounding. The main assumption underlying the procedure is that there is *no unmeasured confounding*, which means that we have measured all variables that contribute to the process that determines whether a subject is exposed at each measurement occasion. If this assumption holds, then providing that we can account for the time varying confounding, associations between exposure and the outcome can be attributed unambiguously to the effect of the exposure. Conditional on this assumption (which cannot be tested using the data), individuals' exposure status at each measurement occasion will be independent of their counterfactual failure time U_i. An example of this assumption is that, conditional on past weight, smoking status, blood pressure and cholesterol measurements (the confounders), the decision of an individual to quit smoking (the exposure) is independent of what his/her survival time would have been had he/she never smoked. Exposure does not have to be independent of subjects' actual life expectancy (smokers may choose to quit precisely because they recognize that smoking has already affected their health, and thus reduced their life expectancy).

The assumption of no unmeasured confounders implies that exposure at each measurement occasion is independent of U_i. The G-estimation procedure therefore searches for the value ψ_0 for which exposure at each measurement occasion is independent of U_{i,ψ_0}. This is done by fitting a logistic regression model relating measured exposure e_{it} at each measurement occasion t to $U_{i,\psi}$, controlling for all confounders $\{x_{ijt}\}$:

$$\text{logit}(e_{it}) = \alpha U_{i,\psi} + \sum_j \beta_j x_{ijt} \tag{2}$$

The confounders in this regression model will typically include the other covariates at the current time point t, the values of the exposure and the other covariates at previous time-points and the values of the exposure and the other covariates at baseline. Subjects contribute an observation for each occasion at which the exposure and confounders were measured.

A series of logistic regression models defined by equation 2 are fitted for a range of different values of ψ. The G-estimate ψ_0 is the value of ψ for which the Wald statistic for α is zero; that is, the p-value is 1, meaning that there is no association between current exposure and U_{i,ψ_0}. The upper and lower limits of the 95 percent confidence interval for ψ_0 are the two values for which the two-sided p-values for the Wald statistic of α are 0.05.

The G-estimate ψ_0 is minus the log of the "causal survival time ratio". Thus, $\exp(-\psi_0)$ estimates the ratio of the survival time of a continuously exposed person to that of an otherwise identical person who was never exposed. This ratio is the amount by which continuous exposure multiplies time to the outcome event. If $\exp(-\psi_0) > 1$, then exposure is beneficial (i.e., exposure increases time to the outcome event). The causal interpretation is justified because (i) changes in exposure precede the occurrence of the outcome, and (ii) providing the assumption of no unmeasured confounders is valid, the estimated association between exposure and outcome can be attributed to the effect of exposure rather than to any confounding factor. Similar causal interpretations

can be made from randomized controlled trials, in which criterion (i) is justified by the experimental design, and criterion (ii) by the randomized allocation of exposure (treatment).

Censoring

The counterfactual survival time, $U_{i,\psi}$, can only be derived from the observed data for a subject who experiences the event. If the study has a planned end of follow-up (at time C_i for individual i) that occurs before all subjects have experienced the outcome event, then some subjects' counterfactual failure times will not be estimable. If C_i is independent of the counterfactual survival time, then this problem can be overcome by replacing $U_{i,\psi}$ with an indicator variable $\Delta_{i,\psi}$ that takes the value 1 if the event would have been observed both if individual i had been exposed throughout follow-up and if they had been unexposed throughout follow-up, and the value 0 otherwise; see Witteman et al. (1998),

$$\Delta_{i,\psi} = \text{ind}(U_{i,\psi} < C_{i,\psi}) \tag{3}$$

where $C_{i,\psi} = C_i$ if $\psi \geq 0$ and $C_{i,\psi} = C_i \exp(\psi)$ if $\psi < 0$. Thus, $\Delta_{i,\psi}$ is zero for all subjects who do not experience an event during follow-up, and may also be zero for some of those who did experience an event. Unlike $U_{i,\psi}$, $\Delta_{i,\psi}$ is estimable for all subjects.

Competing risks

Subjects may also be censored by competing risks. For example, in the study of the effect of ART on AIDS-free survival, subjects could withdraw from the study because they felt too ill to participate in further follow-ups, or be withdrawn from the study because they were prescribed an alternative treatment. In each of these cases, censoring is not independent of the underlying counterfactual survival time. Thus, the above method for dealing with censoring by planned end of study cannot be used to deal with censoring by competing risks.

As outlined by Witteman et al. (1998), censoring due to competing risks is dealt with by modeling the censoring mechanism, and using each individual's estimated probability of being censored to adjust the analysis. Multinomial logistic regression (using all available data) is used to relate the probability of being censored at each measurement occasion to the exposure and covariate history, and hence to estimate the probability of being uncensored to the end of the study for each individual. The inverse of this probability is used to weight the contributions of individuals to the logistic regression models used in the G-estimation process. This approach means that observations within the same individual are no longer independent, so the logistic regression models use robust standard errors allowing for clustering within individuals. This is equivalent to the procedure suggested by Witteman et al., to use a robust Wald test from a generalized estimating equation with an independence working correlation matrix. The confidence intervals obtained using this procedure are conservative.

Converting survival time ratios to hazard ratios

The parameter estimated by the G-estimation procedure, the causal survival time ratio, describes the association between exposure and survival using the accelerated failure time parameterization. In epidemiology, the more usual parameterization for survival analysis is that of proportional hazards. It is therefore useful to be able to express the causal survival time ratio in the proportional hazards parameterization. One obvious way to do this is via Weibull models, the only model that can be expressed in either parameterization.

The Weibull hazard function at time t is $h(t) = \phi\gamma t^{\gamma-1}$, where ϕ is referred to as the scale parameter and γ as the shape parameter. If the vector of covariates x_i does not affect γ, then the Weibull regression model can be written as either the usual epidemiological proportional hazards model

$$h(t, x_i) = h_0(t)\exp(\beta^T x_i) \tag{4}$$

or as an accelerated failure time model,

$$T_i = \exp(\theta^T x_i + \epsilon) \tag{5}$$

where T_i is the failure time for individual i, and ϵ has an extreme value distribution with scale parameter $1/\gamma$. The Weibull shape parameter γ can thus be used to express results from the accelerated failure time parameterization as proportional hazards: $\theta = -\beta/\gamma$. If the underlying survival times are assumed to follow a Weibull distribution, the Weibull shape parameter can therefore be used to express the G-estimated survival ratio as a hazard ratio for the exposure.

3 The stgest command

`stgest` estimates the effect of a time-varying exposure variable, *expvar*, on survival, accounting for possible confounding by the list of (time-varying or non time-varying) variables specified in *confvars* and, optionally, the lagged or baseline effects of one or more of these variables, specified using the `lagconf()` and `baseconf()` options.

Use of G-estimation requires a dataset in which exposures have been measured on at least two occasions, and the time until the occurrence of outcome of interest, or of censoring, is also recorded. The data should be in long `st` format, with subject identifier specified using the `id` option of the `stset` command, and each line of the dataset corresponding to an examination. We explain how to deal with censoring because of competing risks later.

3.1 Syntax

`stgest` *expvar confvars*, `visit(`*varname*`)` [`lasttime(`*varname*`) range(`*numlist*`)`
`step(`#`) tol(`#`) lagconf(`*varlist*`) firstvis(`#`) baseconf(`*varlist*`)`
`pnotcens(`*varname*`) idcens(`*varname*`) saveres(`*filename*`) replace detail`
`round(`#`)`]

`makelag` *varlist*, `firstvis(`#`) visit(`*varname*`)`

`makebase` *varlist*, `firstvis(`#`) visit(`*varname*`)`

`gesttowb`

4 Options

`visit(`*varname*`)` specifies the variable identifying the measurement occasion (examination). At least two measurement occasions are needed. If lagged confounders are to be used, then three measurement occasions are needed, since events occurring between the first (baseline) and second examinations are not included in the analyses. The `visit` option must be specified.

`lasttime(`*varname*`)` must be specified unless all subjects experience the outcome event. It contains the time at which follow-up would have been completed for each patient, had they not experienced the outcome event.

`range(`*numlist*`)` provides the lower and upper ends of the range of estimates for the causal parameter to be considered in the estimation procedure. The default is -5 to 5. Unless the `step()` option is specified, the program conducts an interval bisection search for the best estimate of the causal parameter, together with corresponding upper and lower 95% confidence intervals.

`step(`#`)` is the increment to be used in the search for the best estimate of the causal parameter. If this option is used, the program conducts a grid search instead of an interval bisection search.

`tol(`#`)` is an integer (default 3) specifying the tolerance for the interval bisection search. The search ends when successive values of the estimate of the causal parameter differ by less than $10^{-\text{tol}}$.

`lagconf(`*varlist*`)` gives a list of variables whose lagged confounding effect should be controlled for in the analysis. The lagged value is defined as the value at the previous occasion defined by `visit()`. Corresponding variables with names prefixed by L are created in the dataset.

`firstvis(#)` is the number of the first measurement occasion after which outcome events contribute to the analysis. If this is specified, then only follow-up from the examination after the baseline is considered in estimating the causal effect. By default `firstvis()` is the minimum value of `visit()`. If lagged confounders are to be used, then `firstvis()` must be at least one greater than the minimum value of `visit()`.

`baseconf(`*varlist*`)` gives a list of variables whose baseline confounding effect should be controlled for in the analysis. The baseline value is defined as the minimum value of `visit()`. Corresponding variables with names prefixed by B are created in the dataset.

`pnotcens(`*varlist*`)` specifies a variable containing the cumulative probability of remaining uncensored by competing risks to the end of follow-up, for each individual. If this is not specified, it is assumed that there is no censoring by competing risks. This is derived from a logistic regression with censoring at each examination as the outcome.

`idcens(`*varname*`)` must be specified if `pnotcens()` is specified. `idcens()` is an indicator variable that shows whether the individual was censored due to competing risks (that is, for reasons other than the occurrence of the event of interest). Where there are competing risks, robust standard errors are used to take into account the fact that the probability of being censored is the same for all observations on a given individual.

`saveres(`*filename*`)` requests that the z-statistic for each value in `range()` be saved in *filename*. If this is not specified, no results are saved.

`replace` allows results previously saved in *filename* to be overwritten.

`detail` displays output from the regression model fitted at each iteration.

`round(#)` is rarely needed. It is used when there are problems in creating the indicator variable used in the logistic regression of exposure on counterfactual failure time, allowing for censoring.

5 Example

We will illustrate the use of the `stgest` command to estimate the effect of smoking on rates of heart disease, using data from the Caerphilly study, a longitudinal study of cardiovascular risk factors. Results will be compared with those from standard survival analyses. Participants (all of whom are men) were recruited between 1979 and 1983 (examination 1), when they were aged 44 to 60. Further examinations took place during the periods 1984 to 1988 (examination 2), 1989 to 1993 (examination 3), and 1993 to 1997 (examination 4). All subjects were followed until the end of 1998.

The dataset analyzed here is based on a total of 1756 subjects who had complete data at examinations 1 and 2. The outcome variable (`mi`) is the occurrence of either a myocardial infarction or death from coronary heart disease. Variable `miexitdt` gives

the exit date for each subject, defined as the first (minimum) of (a) date of occurrence of the outcome, (b) date of death, (c) date of emigration, and (d) end of scheduled follow-up (31 December 1998). In the following displays, we will list data for ids 1021, 1022, and 1023. Id 1021 died from coronary heart disease on 18 June 1996, while ids 1022 and 1023 survived until the scheduled end of follow-up.

```
. gen byte touse=0
. replace touse=1 if id==1021|id==1022|id==1023
(11 real changes made)
. list id visit examdat mi miexitdt onsdod if touse
          id  visit    examdat     mi   miexitdt     onsdod
  16.   1021      1  10sep1979      0  18jun1996  18jun1996
  17.   1021      2  31jul1984      0  18jun1996  18jun1996
  18.   1021      3  17mar1992      1  18jun1996  18jun1996
  19.   1022      1  10sep1979      0  31dec1998  14dec1999
  20.   1022      2  19sep1984      0  31dec1998  14dec1999
  21.   1022      3  20nov1989      0  31dec1998  14dec1999
  22.   1022      4  28oct1993      0  31dec1998  14dec1999
  23.   1023      1  10sep1979      0  31dec1998          .
  24.   1023      2  03oct1984      0  31dec1998          .
  25.   1023      3  20nov1989      0  31dec1998          .
  26.   1023      4  08nov1993      0  31dec1998          .
```

Variable `exitdate` is the date at the end of each time interval, defined as `miexitdt` for the subject's last examination, and the date of the subsequent examination otherwise.

```
. by id, sort: gen exitdate=examdat[_n+1]
. by id, sort: replace exitdate=miexitdt if _n==_N
. format exitdate %d
```

Because we wish to control for the baseline effect of smoking and the other covariates, both standard survival analyses and G-estimation will begin at the date of examination 2. We therefore define a variable `examdat2` containing this date for each id, and use this to create variable `agebase` (age at examination 2).

```
. gen edat2=examdat if phase==2
. egen examdat2=max(edat2), by(id)
. format examdat2 %d
. label var examdat2 "Date of 2nd exam (start of follow up for G estimation)"
. drop edat2
. gen agebase=(examdat2-dob)/365.25
. replace agebase=agebase/10
. label var agebase "Age at baseline (10 year units)"
```

We can now `stset` the data and are ready for survival analyses and G-estimation.

```
. stset exitdate, id(id) failure(mi) origin(time examdat2) scale(365.25)
                id:  id
     failure event:  mi ~= 0 & mi ~= .
obs. time interval:  (exitdate[_n-1], exitdate]
 exit on or before:  failure
    t for analysis:  (time-origin)/365.25
            origin:  time examdat2
```

```
     6377  total obs.
     1756  obs. end on or before enter()
-----------------------------------------------------------------------------
     4621  obs. remaining, representing
     1756  subjects
      244  failures in single failure-per-subject data
 18547.87  total analysis time at risk, at risk from t =         0
                            earliest observed entry t =         0
                                 last observed exit t =  14.47502

. list id visit examdat exitdate mi _t0 _t _d _st if touse, noobs nodisp

    id  visit    examdat   exitdate     mi    _t0     _t    _d   _st
  1021      1  10sep1979  31jul1984      0      .      .     .     0
  1021      2  31jul1984  17mar1992      0   0.00   7.63     0     1
  1021      3  17mar1992  18jun1996      1   7.63  11.88     1     1
  1022      1  10sep1979  19sep1984      0      .      .     .     0
  1022      2  19sep1984  20nov1989      0   0.00   5.17     0     1
  1022      3  20nov1989  28oct1993      0   5.17   9.11     0     1
  1022      4  28oct1993  31dec1998      0   9.11  14.28     0     1
  1023      1  10sep1979  03oct1984      0      .      .     .     0
  1023      2  03oct1984  20nov1989      0   0.00   5.13     0     1
  1023      3  20nov1989  08nov1993      0   5.13   9.10     0     1
  1023      4  08nov1993  31dec1998      0   9.10  14.24     0     1
```

Variable `cursmoke` is an indicator variable that records whether the subject was a smoker at each examination.

```
. list id visit examdat cursmok if touse
          id  visit    examdat  cursmok
  16.   1021      1  10sep1979      0
  17.   1021      2  31jul1984      0
  18.   1021      3  17mar1992      0
  19.   1022      1  10sep1979      1
  20.   1022      2  19sep1984      1
  21.   1022      3  20nov1989      1
  22.   1022      4  28oct1993      0
  23.   1023      1  10sep1979      1
  24.   1023      2  03oct1984      1
  25.   1023      3  20nov1989      1
  26.   1023      4  08nov1993      1
```

In these analyses, we will control for the following variables, which may confound the association between smoking and heart disease.

```
              storage  display
variable name   type   format    variable label
-----------------------------------------------------------------------------
hearta          byte   %5.0g     Previous heart attack reported by subject
gout            byte   %5.0g     Previous gout reported by subject
highbp          byte   %5.0g     Previous high blood pressure reported by subject
diabet          byte   %5.0g     Previous diabetes reported by subject
fib75           byte   %8.0g     Fibrinogen above 75th centile
chol75          byte   %8.0g     Cholesterol above 75th centile
hbpsyst         byte   %5.0g     Measured high systolic blood pressure
hbpdias         byte   %5.0g     Measured high diastolic blood pressure
obese           byte   %5.0g     Obese at current visit
thin            byte   %5.0g     Underweight at current visit
```

To examine the effects of baseline smoking controlling for the baseline effects of other variables, we use the utility program `makebase` (supplied with the `stgest` package) to create variables (with names prefixed with B) containing the baseline value of all covariates:

```
. makebase cursmok hearta gout highbp diabet fib75 chol75 hbpsyst hbpdias /*
 */  obese thin, firstvis(1) visit(visit)
Baseline confounders
              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------------------
Bcursmok        byte   %9.0g
Bhearta         byte   %9.0g
Bgout           byte   %9.0g
Bhighbp         byte   %9.0g
Bdiabet         byte   %9.0g
Bfib75          byte   %9.0g
Bchol75         byte   %9.0g
Bhbpsyst        byte   %9.0g
Bhbpdias        byte   %9.0g
Bobese          byte   %9.0g
Bthin           byte   %9.0g
```

We now use Cox regression to examine the effect of smoking at baseline, controlling for the baseline values of the covariates. This shows that subjects who were smokers had a substantially increased hazard of subsequent heart attacks.

```
. stcox B* agebase
  (output omitted)
No. of subjects =         1756                     Number of obs   =      4621
No. of failures =          244
Time at risk    =  18547.87132
                                                   LR chi2(12)     =    111.72
Log likelihood  =   -1695.9464                     Prob > chi2     =    0.0000
------------------------------------------------------------------------------
          _t |
          _d |  Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    Bcursmok |   1.605305   .2165014     3.51   0.000     1.232421    2.091009
     Bhearta |   2.025315   .4094604     3.49   0.000     1.362713    3.010101
       Bgout |    1.67308   .3970742     2.17   0.030     1.050751    2.663996
     Bhighbp |   1.210737   .1807883     1.28   0.200     .9035401    1.622377
     Bdiabet |   1.611559   .6771107     1.14   0.256     .7073046    3.671859
      Bfib75 |   2.139609   .3182323     5.11   0.000     1.598571    2.863762
     Bchol75 |   1.308254   .1816931     1.93   0.053     .9964956    1.717547
    Bhbpsyst |   1.021101   .1562174     0.14   0.891     .7565613     1.37814
    Bhbpdias |   1.764604   .2802593     3.58   0.000     1.292579    2.409003
      Bobese |   .8838818   .1801146    -0.61   0.545     .5928422    1.317799
       Bthin |   .3796157   .2220934    -1.66   0.098     .1206009    1.194917
    agebase |   1.638064   .2384586     3.39   0.001     1.231455     2.17893
------------------------------------------------------------------------------
```

A second utility program `makelag` creates variables (with names prefixed with L) containing the lagged value of covariates (i.e., the value at the previous examination).

```
. makelag cursmok hearta gout highbp diabet fib75 chol75 hbpsyst hbpdias /*
 */  obese thin, firstvis(1) visit(visit)
Lagged confounders
```

variable name	storage type	display format	value label	variable label
Lcursmok	byte	%9.0g		
Lhearta	byte	%9.0g		
Lgout	byte	%9.0g		
Lhighbp	byte	%9.0g		
Ldiabet	byte	%9.0g		
Lfib75	byte	%9.0g		
Lchol75	byte	%9.0g		
Lhbpsyst	byte	%9.0g		
Lhbpdias	byte	%9.0g		
Lobese	byte	%9.0g		
Lthin	byte	%9.0g		

To examine the effect of current smoking, controlling for other confounders and for chronic damage caused by previous smoking, we would fit a model including the current, lagged, and baseline values of all covariates. Such a model appears to show that there is no effect of smoking. However, this analysis is not valid because of time-dependent confounding.

```
. stcox cursmok agebase hearta gout highbp diabet fib75 chol75 hbpsyst hbpdias
> obese thin B* L*
```

(*output omitted*)

_t _d	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
cursmok	1.057524	.2165462	0.27	0.785	.7079332	1.57975
agebase	1.623111	.2409387	3.26	0.001	1.213372	2.171215
hearta	1.858263	.4145168	2.78	0.005	1.200141	2.877279
gout	1.132199	.2803371	0.50	0.616	.6968859	1.839432
highbp	1.046366	.1853165	0.26	0.798	.7394895	1.480593
diabet	1.525949	.5353171	1.20	0.228	.7672392	3.034936
fib75	1.540378	.2135747	3.12	0.002	1.173836	2.021376
chol75	1.460741	.2412332	2.29	0.022	1.056822	2.019037
hbpsyst	1.021339	.1658041	0.13	0.897	.7429951	1.403957
hbpdias	1.178434	.1849555	1.05	0.296	.8663812	1.602882
obese	.7288805	.184787	-1.25	0.212	.4434639	1.197993
thin	.7028389	.484119	-0.51	0.609	.1821981	2.711239
Bcursmok	1.115229	.2502954	0.49	0.627	.7183324	1.731421
Bhearta	1.359836	.3645049	1.15	0.252	.8041209	2.299596
Bgout	1.490102	.5008544	1.19	0.235	.7710977	2.879537
Bhighbp	.9011563	.1792485	-0.52	0.601	.6102223	1.330798
Bdiabet	1.101382	.6544871	0.16	0.871	.343652	3.52986
Bfib75	1.844116	.3171563	3.56	0.000	1.316425	2.583333
Bchol75	1.170021	.2230618	0.82	0.410	.8052191	1.700095
Bhbpsyst	.7803709	.1417425	-1.37	0.172	.5466298	1.114061
Bhbpdias	1.424024	.2592141	1.94	0.052	.9967215	2.034515
Bobese	.7065833	.2168224	-1.13	0.258	.387225	1.289328
Bthin	.3312229	.2712597	-1.35	0.177	.0665299	1.649012
Lcursmok	1.406188	.3691871	1.30	0.194	.8405529	2.352458
Lhearta	1.354716	.3841175	1.07	0.284	.7771375	2.361559

Lgout	.9615045	.328462	-0.11	0.909	.4922324	1.87816
Lhighbp	1.304913	.274506	1.27	0.206	.864012	1.970803
Ldiabet	.9766689	.5128373	-0.04	0.964	.3489727	2.7334
Lfib75	1.01087	.1664976	0.07	0.948	.731975	1.396029
Lchol75	.8578317	.1696318	-0.78	0.438	.5822123	1.263929
Lhbpsyst	1.451487	.2812038	1.92	0.054	.9929002	2.121879
Lhbpdias	1.220062	.2124107	1.14	0.253	.8673395	1.716226
Lobese	1.412592	.4112167	1.19	0.235	.7984086	2.499241
Lthin	1.593853	1.348765	0.55	0.582	.3034844	8.370665

To allow for time-dependent confounding, we use `stgest`. This first example ignores censoring due to competing risks, which is dealt with later. This analysis allows for the effect of current, lagged (option `lagconf`), and baseline (option `baseconf`) values of the covariates. We also supply the names of the variable indexing examination number (option `visit`), the first visit from which survival time is counted (option `firstvis`), the scheduled end of follow-up for each individual had they not been censored (option `lasttime`), the range over which we will search for values of ψ (option `range`), and the file in which our results will be saved (option `saveres`). The program automatically creates lagged and baseline values of the covariates with names prefixed by L and B, respectively, and lists these. It then outputs each value of ψ for which it fits the logistic regression (equation 2) and, finally, lists the values of ψ with their corresponding p-values and z statistics.

```
. stgest cursmok agebase fib75 hearta gout highbp diabet chol75 hbpsyst /*
 */ hbpdias obese thin, lagconf(fib75 hearta gout highbp diabet /*
 */ cursmok chol75 hbpsyst hbpdias obese thin) baseconf(fib75 hearta gout /*
 */ highbp cursmok chol75 diabet hbpsyst hbpdias obese thin) /*
 */ visit(visit) firstvis(2) (lasttime(mienddat) range(-1 1) /*
 */ saveres(caergestsmoknocens) replace

confounders: agebase fib75 hearta gout highbp diabet chol75 hbpsyst hbpdias obese
> thin
causvar: cursmok
visit: visit
Range: -1 1, rnum: 2
Search method: interval bisection

Baseline confounders

              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------------------
Bfib75          float  %9.0g
Bhearta         float  %9.0g
Bgout           float  %9.0g
Bhighbp         float  %9.0g
Bcursmok        float  %9.0g
Bchol75         float  %9.0g
Bdiabet         float  %9.0g
Bhbpsyst        float  %9.0g
Bhbpdias        float  %9.0g
Bobese          float  %9.0g
Bthin           float  %9.0g
```

```
Lagged confounders
              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------------------
Lfib75          float  %9.0g
Lhearta         float  %9.0g
Lgout           float  %9.0g
Lhighbp         float  %9.0g
Ldiabet         float  %9.0g
Lcursmok        float  %9.0g
Lchol75         float  %9.0g
Lhbpsyst        float  %9.0g
Lhbpdias        float  %9.0g
Lobese          float  %9.0g
Lthin           float  %9.0g
-1.00 1.00 0.00 0.50 0.25 0.38 0.31 0.28 0.27 0.27 0.28 0.28 0.28 0.44 0.41 0.4
> 2 0.41 0.41 0.41 0.41 0.13 0.06 0.03 0.05 0.04 0.04 0.03 0.03
savres: caergestsmoknocens
           psi       pval          z
  1.        -1          0    9.31326
  2.         0   .0144764    2.44522
  3.    .03125   .0466648    1.98933
  4.  .0322266   .0527831   1.936691
  5.  .0332031   .0527831   1.936691
  6.  .0351563   .0527831   1.936691
  7.  .0390625   .0564285   1.907712
  8.   .046875   .0564285   1.907712
  9.     .0625   .0679246   1.825507
 10.      .125   .3005441   1.035267
 11.       .25   .8140365   .2352219
 12.   .265625   .8268306   .2187683
 13.  .2734375   .8740402   .1585287
 14.  .2773438   .9980356   .0024621
 15.  .2783203   .9980356   .0024621
 16.  .2792969   .8909989  -.1370404
 17.    .28125   .8997204  -.1260146
 18.     .3125   .5281636  -.6308119
 19.      .375   .0902674  -1.693989
 20.    .40625   .0514808   -1.94745
 21.  .4072266   .0492024  -1.966833
 22.  .4082031   .0492024  -1.966833
 23.  .4101563   .0492024  -1.966833
 24.  .4140625   .0414209  -2.039292
 25.   .421875   .0392723  -2.061322
 26.     .4375   .0155952  -2.418253
 27.        .5   .0300632  -2.169256
 28.         1   1.13e-08  -5.710349
G estimate of psi for cursmok: 0.278 (95% CI 0.034 to 0.409)
Causal survival time ratio for cursmok: 0.757 (95% CI 0.665 to 0.967)
```

It is possible that observations are dropped from the logistic regressions if a covariate predicts exposure perfectly. If this problem occurs, the above list of values for `psi`, `pval`, and `z` will include a column labeled `error`, with values of 1 corresponding to the logistic regressions in which the problem occurred. The cause of such problems can be investigated by specifying the `detail` option so that the logistic regression output for each iteration is displayed.

The program displays a graph of z against ψ (Figure 1). This should be approximately linear. The vertical lines show the values of ψ corresponding to the values of z closest to 0, 1.96, and -1.96. If the graph is not linear, then the G-estimation process should be re-run using specified values of the range, because the search algorithm may be unreliable.

The estimated value of ψ_0 corresponds to the value of z closest to 0, with the 95% confidence interval corresponding to the values of z closest to -1.96 and 1.96. The final part of the output shows the causal survival time ratio $\exp(-\psi_0)$, with its 95% CI.

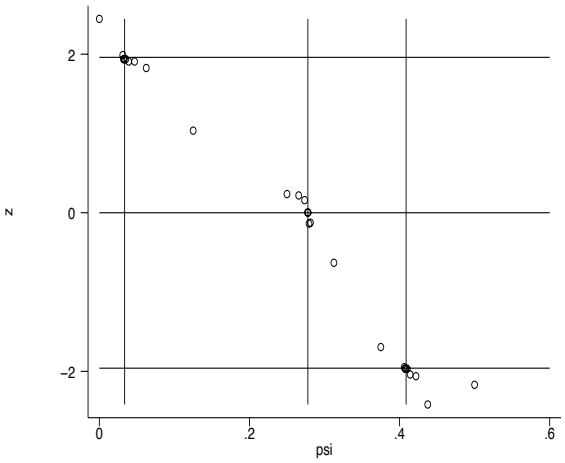

Figure 1: Graph of z against ψ, for a G-estimation that ignores censoring due to competing risks.

To assess the influence of time-dependent confounding, we will compare these results with those from the corresponding Weibull regression:

```
. weibull _t cursmok agebase hearta gout highbp diabet fib75 chol75 hbpsyst /*
  */ hbpdias obese thin B* L* if visit>=2, dead(_d) t0(_t0) hr
```

(*output omitted*)

```
Weibull regression -- entry time _t0
log relative-hazard form                          Number of obs   =       4621
                                                  LR chi2(34)     =     157.01
Log likelihood = -833.45591                       Prob > chi2     =     0.0000
```

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
cursmok	1.055329	.2156308	0.26	0.792	.7070749	1.575107
(*output omitted*)						
p	1.156709	.0784166			1.012788	1.321081
(*output omitted*)						

Note that the hazard ratio for `cursmok` is, as is usually the case, almost identical to that from the Cox regression displayed earlier. The `gesttowb` utility uses the shape parameter γ from the Weibull regression (which is `p` in the Weibull output above) to convert the causal survival time ratio into a corresponding hazard ratio.

```
. gesttowb
g-estimated hazard ratio 1.38 ( 1.04 to  1.60)
```

Because of time-dependent confounding, the standard survival analysis approach to the analysis of time-updated exposures underestimated the effect of smoking (hazard ratio 1.05 compared to the G-estimated hazard ratio of 1.38).

5.1 Allowing for competing risks

To allow for censoring due to competing risks, we first have to model the probability of being censored at each examination. Because we are examining survival from examination 2 onwards, no subject can be censored at examination 1. We model the probability of censoring using one model for examinations 2 and 3, and a separate model for examination 4. For examinations 2 and 3, we use a multinomial logit model with four outcomes: no censoring, MI (our outcome), death from another cause, and lost to follow-up. The model for examination 4 is similar, but there is no loss to follow-up after examination 4 (because it is the last planned examination). We then use the predicted probabilities (for each individual, from the model) of each outcome to calculate pcens, the estimated probability of being censored at a given examination, for each individual.

```
. mlogit cens phase3 agebase hearta gout highbp cursmok hbpsyst hbpdias /*
 */obese thin totchol fibrin if phase~=1&phase~=4
  (output omitted)
Multinomial regression                          Number of obs   =       3285
                                                LR chi2(36)     =     154.79
                                                Prob > chi2     =     0.0000
Log likelihood = -1633.1061                     Pseudo R2       =     0.0452
  (output omitted)
(Outcome cens==No is the comparison group)

. predict pcens2 if e(sample), outcome(2)
(option p assumed; predicted probability)
(3092 missing values generated)

. predict pcens3 if e(sample), outcome(3)
(option p assumed; predicted probability)
(3092 missing values generated)

. gen pcens=pcens2+pcens3
(3092 missing values generated)

. mlogit cens agebase hearta gout highbp diabet cursmok hbpsyst hbpdias /*
 */obese thin totchol fibrin if phase==4
  (output omitted)
Multinomial regression                          Number of obs   =       1336
                                                LR chi2(24)     =      93.39
                                                Prob > chi2     =     0.0000
Log likelihood = -556.46295                     Pseudo R2       =     0.0774
  (output omitted)
(Outcome cens==No is the comparison group)
```

```
. predict pcens42 if e(sample), outcome(2)
(option p assumed; predicted probability)
(5041 missing values generated)
. replace pcens=pcens42 if phase==4
(1336 real changes made)
. replace pcens=0 if phase==1
(1756 real changes made)
```

We use the individual probability of not being censored at each examination to calculate pnotcens, the estimated probability for each individual of being uncensored to the end of examination 4.

```
. gen lpnocens=log(1-pcens)
. egen sumpnoc=sum(lpnocens), by(id)
. gen pnotcens=exp(sumpnoc)
. label var pnotcens "Cumulative probability not censored"
```

We then use this probability of remaining uncensored to adjust the G-estimation for censoring due to competing risks. This involves specifying two further options: the probability of remaining uncensored to the end of the study (`pnotcens`) and an indicator variable for each id, `idcens`, which takes the value 1 if that subject is censored before the end of the study, and the value 0 otherwise. If these options are specified, the `stgest` command will weight the estimation procedure as described earlier, and use robust standard errors to account for the clustering this induces within individuals.

```
. stgest cursmok agebase fib75 hearta gout highbp diabet chol75 hbpsyst /*
*/hbpdias obese thin, /*
*/visit(visit) firstvis(2) lagconf(fib75 hearta gout highbp diabet /*
*/cursmok chol75 hbpsyst hbpdias obese thin) baseconf(fib75 hearta gout highbp/*
*/cursmok chol75 diabet hbpsyst hbpdias obese thin) lasttime(mienddat)/*
*/idcens(idcrcens) range(-1 1) pnotcens(pnotcens) saveres(caergestsmok) replace
```

(*output omitted*)

```
-1.00 1.00 0.00 0.50 0.25 0.38 0.31 0.28 0.30 0.30 0.31 0.31 0.31 0.75 0.88 0.8
> 1 0.84 0.86 0.85 0.85 0.85 0.84 -0.50 -0.25 -0.13 -0.06 -0.09 -0.08 -0.09 -0.
> 09 -0.09 -0.09
G estimate of psi for cursmok: 0.311 (95% CI -0.077 to 0.844)
Causal survival time ratio for cursmok: 0.733 (95% CI 0.430 to 1.080)
```

(*Continued on next page*)

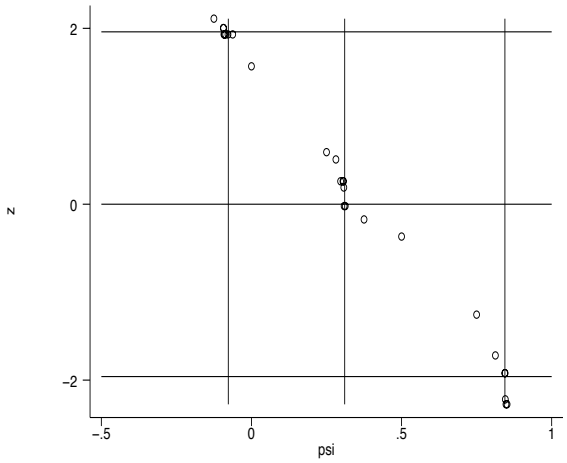

Figure 2: Graph of z against ψ, for G-estimation allowing for censoring due to competing risks.

Again, to assess the influence of time-dependent confounding, we will compare these results with those from the corresponding Weibull regression. To adjust for censoring due to competing risks, we include in the model only those individuals who remained uncensored to the planned end of the study, and weight their contributions by the inverse of the probability of remaining uncensored to the planned end of the study.

```
. gen invpnotc=1/pnotcens
. drop if idcrcens==1
(959 observations deleted)
. weibull _t cursmok agebase hearta gout highbp diabet fib75 chol75 hbpsyst /*
 */ hbpdias obese thin B* L* [pweight=invpnotc] if visit>=2, dead(_d) /*
 */ t0(_t0) hr cluster(id)
  (output omitted)
Weibull regression -- entry time _t0
log relative-hazard form                          Number of obs   =       3999
                                                  Wald chi2(34)   =     178.10
Log likelihood = -946.90352                       Prob > chi2     =     0.0000

                                (standard errors adjusted for clustering on id)
------------------------------------------------------------------------------
             |               Robust
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     cursmok |   1.024165   .2182997     0.11   0.911     .6744302    1.555258
  (output omitted)
-------------+----------------------------------------------------------------
           p |    1.15558   .0947354                      .984051    1.357007
  (output omitted)
------------------------------------------------------------------------------
```

As before, we can use the shape parameter from the Weibull regression to express the G-estimation parameter as a hazard ratio:

```
. gesttowb
g-estimated hazard ratio 1.43 ( 0.91 to  2.65)
```

Again, time-varying confounding has meant that the standard survival analysis substantially underestimates the detrimental effect of smoking on time to MI.

6 Acknowledgments

We thank Ian White and Sarah Walker, who allowed us to use the interval bisection algorithm from their `strbee` command, and Jamie Robins and Miguel Hernan for their help and advice. A previous version of this article was presented at the 2001 UK Stata User Group meeting.

7 References

Mark, S. D. and J. M. Robins. 1993. Estimating the causal effect of smoking cessation in the presence of confounding factors using a rank preserving structural failure time model. *Statistics in Medicine* 12(17): 1605–1628.

Robins, J. M., D. Blevins, G. Ritter, and M. Wulfsohn. 1992. G-estimation of the effect of prophylaxis therapy for Pneumocystis carinii pneumonia on the survival of AIDS patients. *Epidemiology* 3: 319–336.

Tilling, K., J. A. Sterne, and M. Szklo. 2002. Estimating the effect of cardiovascular risk factors on all-cause mortality and incidence of coronary heart disease using g-estimation: the ARIC study. *American Journal of Epidemiology* 155: 710–718.

Witteman, J. C., R. B. D'Agostino, T. Stijnen, W. Kannel, J. C. Cobb, and M. A. de Ridder. 1998. G-estimation of causal effects: isolated systolic hypertension and cardiovascular death in the Framingham Heart Study. *American Journal of Epidemiology* 148(4): 390–401.

About the Authors

Jonathan Sterne is Reader in Medical Statistics in the Department of Social Medicine, University of Bristol. His research interests include statistical methods for life course epidemiology, and bias in meta-analysis and systematic reviews.

Kate Tilling is Lecturer in Medical Statistics in the Department of Public Health Sciences, King's College London. Her research interests include statistical methods for observational studies and longitudinal data analysis.