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.

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

http://www.stata-journal.com

The *Stata Journal* is indexed and abstracted in the following:

- CompuMath Citation Index®
- Current Contents/Social and Behavioral Sciences®
- RePEc: Research Papers in Economics
- Science Citation Index Expanded (also known as SciSearch®)
- Scopus™
- Social Sciences Citation Index®

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

The Stata Journal (2010)
10, Number 3, pp. 408–422

Regression analysis of censored data using pseudo-observations

Erik T. Parner
University of Aarhus
Aarhus, Denmark
parner@biostat.au.dk

Per K. Andersen
University of Copenhagen
Copenhagen, Denmark
P.K.Andersen@biostat.ku.dk

Abstract. We draw upon a series of articles in which a method based on pseudovalues is proposed for direct regression modeling of the survival function, the restricted mean, and the cumulative incidence function in competing risks with right-censored data. The models, once the pseudovalues have been computed, can be fit using standard generalized estimating equation software. Here we present Stata procedures for computing these pseudo-observations. An example from a bone marrow transplantation study is used to illustrate the method.

Keywords: st0202, stpsurv, stpci, stpmean, pseudovalues, time-to-event, survival analysis

1 Introduction

Statistical methods in survival analysis need to deal with data that are incomplete because of right-censoring; a host of such methods are available, including the Kaplan–Meier estimator, the log-rank test, and the Cox regression model. If one had complete data, standard methods for quantitative data could be applied directly for the observed survival time X, or methods for binary outcomes could be applied by dichotomizing X as $I(X > \tau)$ for a suitably chosen τ. With complete data, one could furthermore set up regression models for any function $f(X)$ and check such models using standard graphical methods such as scatterplots or residuals for quantitative or binary outcomes.

One way of achieving these goals with censored survival data and with more-general event history data (for example, competing-risks data) is to use a technique based on pseudo-observations, as recently described in a series of articles. Thus the technique has been studied in modeling of the survival function (Klein et al. 2007), the restricted mean (Andersen, Hansen, and Klein 2004), and the cumulative incidence function in competing risks (Andersen, Klein, and Rosthøj 2003; Klein and Andersen 2005; Klein 2006; Andersen and Klein 2007).

The basic idea is simple. Suppose a well-behaved estimator $\widehat{\theta}$, for the expectation $\theta = E\{f(X)\}$, is available—for example, the Kaplan–Meier estimator for $S(t) = E\{I(X > t)\}$—based on a sample of size n. The ith pseudo-observation ($i = 1, \dots, n$) for $f(X)$ is then defined as $\widehat{\theta}_i = n \times \widehat{\theta} - (n-1) \times \widehat{\theta}_{-i}$ where $\widehat{\theta}_{-i}$ is the estimator applied to the sample of size $n-1$, which is obtained by eliminating the ith observation from the dataset. The pseudovalues are generated once, and the idea is to replace the incompletely observed

$f(X_i)$ by $\widehat{\theta}_i$. That is, $\widehat{\theta}_i$ may be used as an outcome variable in a regression model or it may be used to compute residuals. $\widehat{\theta}_i$ also may be used in a scatterplot when assessing model assumptions (Perme and Andersen 2008; Andersen and Perme 2010). The intuition is that, in the absence of censoring, $\theta = E\{f(X)\}$ could, obviously, be estimated as $(1/n)\sum_i f(X_i)$, in which case the ith pseudo-observation is simply the observed value $f(X_i)$. The pseudovalues are related to the jackknife residuals used in regression diagnostics.

We present three new Stata commands—`stpsurv`, `stpci`, and `stpmean`—that provide a new possibility in Stata for analyzing regression models and that generate pseudovalues (respectively) for the survival function (or the cumulative distribution function, "the cumulative incidence") under right-censoring, for the cumulative incidence in competing risks, and for the restricted mean under right-censoring. Cox regression models can be fit using the pseudovalue function for survival probabilities in several time points. Thereby, the pseudovalue method provides an alternative to Cox regression, for example, in situations where rates are not proportional. As discussed by Perme and Andersen (2008), residuals for model checking may also be obtained from the pseudovalues. An example based on bone marrow transplantation data is presented to illustrate the methodology.

In section 2, we briefly present the general pseudovalue approach to censored data regression. In section 3, we present the new Stata commands; and in section 4, we show examples of the use of the commands. Section 5 concludes with some remarks.

2 Some methodological details

2.1 The general approach

In this section, we briefly introduce censored data regression based on pseudo-observations; see, for example, Andersen, Klein, and Rosthøj (2003) or Andersen and Perme (2010) for more details. Let $X_1, \ldots, X_n$ be independent and identically distributed survival times, and suppose we are interested in a parameter of the form

$$\theta = E\{f(X)\}$$

for some function $f(\cdot)$. This function could be multivariate, for example,

$$f(X) = \{f_1(X), \ldots, f_M(X)\} = \{I(X > \tau_1), \ldots, I(X > \tau_M)\}$$

for a series of time points $\tau_1, \ldots, \tau_M$, in which case,

$$\theta = (\theta_1, \ldots, \theta_M) = \{S(\tau_1), \ldots, S(\tau_M)\}$$

where $S(\cdot)$ is the survival function for X. More examples are provided below. Furthermore, let $Z_1, \ldots, Z_n$ be independent and identically distributed covariates. Also suppose we are interested in a regression model of $\theta = E\{f(X_i)\}$ on Z_i—for example, a generalized linear model of the form

$$g[E\{f(X_i) \mid Z_i\}] = \beta^T Z_i$$

where $g(\cdot)$ is the link function. If right-censoring prevents us from observing all the X_is, then it is not simple to analyze this regression model. However, suppose $\widehat{\theta}$ is an approximately unbiased estimator of the marginal mean $\theta = E\{f(X)\}$ that may be computed from the sample of right-censored observations. If $f(X) = I(X > \tau)$, then $\theta = S(\tau)$ may be estimated using the Kaplan–Meier estimator. The ith pseudo-observation is now defined, as suggested in section 1, as

$$\widehat{\theta}_i = n \times \widehat{\theta} - (n-1) \times \widehat{\theta}_{-i}$$

Here $\widehat{\theta}_{-i}$ is the "leave-one-out" estimator for θ based on all observations but the ith: X_j, $j \neq i$. The idea is to replace the possibly incompletely observed $f(X_i)$ by $\widehat{\theta}_i$ and to obtain estimates of the βs based on the estimating equation:

$$\sum_i \left\{ \frac{\partial}{\partial \beta} g^{-1}(\beta^T Z_i) \right\}^T V_i^{-1}(\beta) \left\{ \widehat{\theta}_i - g^{-1}(\beta^T Z_i) \right\} = \sum_i U_i(\beta) = U(\beta) = 0 \qquad (1)$$

In (1), V_i is a working covariance matrix. Graw, Gerds, and Schumacher (2009) showed that for the examples studied in this article, $E\{f(X_i) \,|\, Z_i\} = E(\widehat{\theta}_i \,|\, Z_i)$, and thereby (1) is unbiased, provided that censoring is independent of covariates; see also Andersen and Perme (2010). A sandwich estimator is used to estimate the variance of $\widehat{\beta}$. Let

$$I(\beta) = \sum_i \left\{ \frac{\partial}{\partial \beta} g^{-1}(\beta^T Z_i) \right\}^T V_i^{-1}(\beta) \left\{ \frac{\partial g^{-1}(\beta^T Z_i)}{\partial \beta} \right\}$$

and

$$\widehat{\mathrm{Var}} \left\{ U\left(\widehat{\beta}\right) \right\} = \sum_i U_i\left(\widehat{\beta}\right)^T U_i\left(\widehat{\beta}\right)$$

then

$$\widehat{\mathrm{Var}}\left(\widehat{\beta}\right) = I\left(\widehat{\beta}\right)^{-1} \widehat{\mathrm{Var}} \left\{ U\left(\widehat{\beta}\right) \right\} I\left(\widehat{\beta}\right)^{-1}$$

The estimator of β can be shown to be asymptotically normal (Graw, Gerds, and Schumacher 2009; Liang and Zeger 1986), and the sandwich estimator converges in probability to the true variance. Once the pseudo-observations have been computed, the estimators of β can be obtained by using standard software for generalized estimating equations.

The pseudo-observations may also be used to define residuals after fitting some standard model (for example, a Cox regression model) for survival data; see Perme and Andersen (2008) or Andersen and Perme (2010).

2.2 The survival function

Suppose we are interested in the survival function $S(\tau_j) = \Pr(X > \tau_j)$ at a grid of time points $\tau_1 < \cdots < \tau_M$, for a survival time X. Hence, $\theta = (\theta_1, \ldots, \theta_M)$ where

$\theta_j = S(\tau_j)$. When $M = 1$, we consider the survival function at a single point in time. Under right-censoring, the survival function is estimated by the Kaplan–Meier estimator (Kaplan and Meier 1958),

$$\widehat{S}(t) = \prod_{t_j \le t} \frac{Y_j - d_j}{Y_j}$$

where $t_1 < \cdots < t_D$ are the distinct event times, Y_j is the number at risk, and d_j is the number of events at time t_j. The cumulative distribution function is then estimated by $\widehat{F}(t) = 1 - \widehat{S}(t)$. In this case, the link function of interest could be the cloglog function

$$\text{cloglog}\,\{F(\tau)\} = \log\left[-\log\{1 - F(\tau)\}\right]$$

which is equivalent to a Cox regression model for the survival function evaluated in τ.

2.3 The mean survival time

The mean time-to-event is the area under the survival curve:

$$\mu = \int_0^\infty S(u)du \tag{2}$$

For right-censored data, the estimated survival function (the Kaplan–Meier estimator) does not always converge down to zero. Then the mean cannot be estimated reliably by plugging the Kaplan–Meier estimator into (2). An alternative to the mean is the restricted mean, defined as the area under the survival curve up to a time $\tau < \infty$ (Klein and Moeschberger 2003), which is equal to $\theta = \mu_\tau = E\{\min(X, \tau)\}$. The restricted mean survival time is estimated by the area under the Kaplan–Meier curve up to time τ. That is,

$$\widehat{\mu}_\tau = \int_0^\tau \widehat{S}(u)du$$

An alternative mean is the conditional mean given that the event time is smaller than τ, $\mu_\tau^c = E(X \mid X \le \tau)$, which is similarly estimated by

$$\widehat{\mu}_\tau^c = \int_0^\tau \frac{\widehat{S}(u) - \widehat{S}(\tau)}{1 - \widehat{S}(\tau)} du$$

For the restricted and conditional mean, a link function of interest could be the log or the identity.

2.4 The cumulative incidence

Under competing risks, the cumulative incidence function is estimated in a different way. Suppose the event of interest has hazard function $h_1(t)$ and the competing risk has hazard function $h_2(t)$. The cumulative incidence function for the event of interest is then given as

$$F_1(t) = \int_0^t h_1(u) \exp\left[-\int_0^u \{h_1(v) + h_2(v)\}\, dv\right] du$$

If $t_1 < \cdots < t_D$ are the distinct times of the primary event and the competing risk combined, Y_j is the number at risk, d_{1j} is the number of the primary events at time t_j, and d_{2j} is the number of competing risks at time t_j. Then the cumulative incidence function of the primary event is estimated by

$$\widehat{F}_1(t) = \sum_{t_j \le t} \left(\frac{d_{1j}}{Y_j}\right) \prod_{t_i < t_j} \left\{\frac{Y_i - (d_{1i} + d_{2i})}{Y_i}\right\}$$

Again the link function of interest could be cloglog corresponding to the regression model for the competing risks cumulative incidence studied by Fine and Gray (1999).

3 The stpsurv, stpmean, and stpci commands

3.1 Syntax

Pseudovalues for the survival function, the mean survival time, and the cumulative incidence function for competing risks are generated using the following syntaxes:

`stpsurv` [*if*] [*in*], `at(`*numlist*`)` [`generate(`*string*`)` `failure`]

`stpmean` [*if*] [*in*], `at(`*numlist*`)` [`generate(`*string*`)` `conditional`]

`stpci` *varname* [*if*] [*in*], `at(`*numlist*`)` [`generate(`*string*`)`]

`stpsurv`, `stpmean`, and `stpci` are for use with st data. You must, therefore, `stset` your data before issuing these commands. Frequency weights are allowed in the `stset` command. In the `stpci` command for the cumulative incidence function in competing risks, an indicator variable for the competing risks should always be specified. The pseudovalues are by default stored in the `pseudo` variable when one time point is specified and are stored in variables `pseudo1`, `pseudo2`, ... when several time points are specified. The names of the pseudovariables are changed by the `generate()` option.

3.2 Options

`at(`*numlist*`)` specifies the time points in ascending order of which pseudovalues should be computed. `at()` is required.

`generate(`*string*`)` specifies a variable name for the pseudo-observations. The default is `generate(pseudo)`.

`failure` generates pseudovalues for the cumulative incidence proportion, which is one minus the survival function.

`conditional` specifies that pseudovalues for the conditional mean should be computed instead of those for the restricted mean.

4 Example data

To illustrate the pseudovalue approach, we use data on sibling-donor bone marrow transplants matched on human leukocyte antigen (Copelan et al. 1991). The data are available in Klein and Moeschberger (2003). The data include information on 137 transplant patients on time to death, relapse, or lost to follow-up (`tdfs`); the indicators of relapse and death (`relapse`, `trm`); the indicator of treatment failure (`dfs` = `relapse`|`trm`); and three factors that may be related to outcome: `disease` [acute lymphocytic leukemia (ALL), low-risk acute myeloid leukemia (AML), and high-risk AML], the French–American–British (FAB) disease grade for AML (`fab` = 1 if AML and grade 4 or 5; 0 otherwise), and recipient age at transplant (`age`).

4.1 The survival function at a single time point

We will first examine regression models for disease free survival at 530 days based on the Kaplan–Meier estimator. Disease free survival probabilities for the single prognostic factor FAB at 530 days (figure 1) can be compared using information obtained using the Stata `sts list` command, which evaluates the Kaplan–Meier estimator.

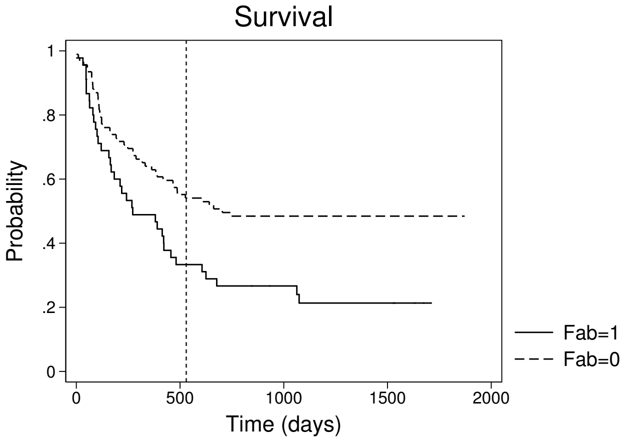

Figure 1. Disease free survival

Based on the `sts list` output below, the risk difference (RD) for FAB is computed as RD = 0.333 − 0.541 = −0.207 [95% confidence interval: −0.379, −0.039] and the relative risk (RR) for FAB is RR = 0.333/0.541 = 0.616, where FAB = 0 is chosen as the reference group. The confidence interval of the RD is based on computing the standard error of the RD as $(0.0522^2 + 0.0703^2)^{1/2}$. The confidence interval for the RR is not easily estimated using the information from the `sts list` command.

```
. use bmt

. stset tdfs, failure(dfs==1)

     failure event:  dfs == 1
obs. time interval:  (0, tdfs]
 exit on or before:  failure
──────────────────────────────────────────────────────────────────────────────
      137  total obs.
        0  exclusions
──────────────────────────────────────────────────────────────────────────────
      137  obs. remaining, representing
       83  failures in single record/single failure data
   107138  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =      2640

. sts list, at(0 530) by(fab)

         failure _d:  dfs == 1
   analysis time _t:  tdfs

             Beg.                          Survivor      Std.
    Time     Total     Fail                Function     Error     [95% Conf. Int.]
───────────────────────────────────────────────────────────────────────────────
fab=0
       0         0        0                  1.0000         .          .         .
     530        49       42                  0.5408    0.0522     0.4334    0.6364
fab=1
       0         0        0                  1.0000         .          .         .
     530        16       30                  0.3333    0.0703     0.2018    0.4704
───────────────────────────────────────────────────────────────────────────────
Note:  survivor function is calculated over full data and evaluated at
       indicated times; it is not calculated from aggregates shown at left.
```

Now we turn to the pseudovalues approach. We start by computing the pseudovalues at 530 days using the `stpsurv` command. The pseudovalues are stored in the `pseudo` variable.

```
. stpsurv, at(530)
Computing pseudo observations (progress dots indicate percent completed)
────┼─── 1 ───┼─── 2 ───┼─── 3 ───┼─── 4 ───┼─── 5
..................................................    50
..................................................   100
Generated pseudo variable: pseudo
```

The pseudovalues are analyzed in generalized linear models with an identity link function and a log link function, respectively.

```
. glm pseudo i.fab, link(id) vce(robust) noheader
Iteration 0:   log pseudolikelihood = -96.989802

------------------------------------------------------------------------------
             |               Robust
      pseudo |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       1.fab |  -.2080377   .0881073    -2.36   0.018    -.3807248   -.0353506
       _cons |   .5406774   .0522411    10.35   0.000     .4382867    .6430681
------------------------------------------------------------------------------

. glm pseudo i.fab, link(log) vce(robust) eform noheader
Iteration 0:   log pseudolikelihood = -123.14846
Iteration 1:   log pseudolikelihood = -101.53512
Iteration 2:   log pseudolikelihood = -96.991808
Iteration 3:   log pseudolikelihood = -96.989802
Iteration 4:   log pseudolikelihood = -96.989802

------------------------------------------------------------------------------
             |               Robust
      pseudo |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       1.fab |   .6152278   .1440588    -2.07   0.038     .3887968    .9735298
------------------------------------------------------------------------------
```

The generalized linear models with an identity link function and a log link function fit the relations

$$\begin{aligned} p_i &= E\left(X_i\right) = \beta_0 + \beta_1 \times \text{FAB}_i \\ \log(p_i) &= \log\{E\left(X_i\right)\} = \widetilde{\beta}_0 + \widetilde{\beta}_1 \times \text{FAB}_i \end{aligned}$$

respectively, where $p_i = S_i(530)$ is disease free survival probability at 530 days for individual i. Hence, based on the pseudovalues approach, we estimate the RD for FAB by RD $= -0.208$ [95% confidence interval: -0.381, -0.035] and the RR for FAB by RR $= 0.615$ [95% confidence interval: 0.389, 0.974]. The results are very similar to the direct computation from the Kaplan–Meier using the `sts list` command. We now obtain the confidence interval for the RR.

Suppose we wish to compute the RR for FAB, adjusting for `disease` as a categorical variable and `age` as a continuous variable. Using the same pseudovalues, we fit the generalized linear model.

(*Continued on next page*)

```
. glm pseudo i.fab i.disease age, link(log) vce(robust) eform noheader
Iteration 0:   log pseudolikelihood = -114.83229
Iteration 1:   log pseudolikelihood = -93.440112
Iteration 2:   log pseudolikelihood = -88.620704
Iteration 3:   log pseudolikelihood = -88.601028
Iteration 4:   log pseudolikelihood = -88.601013
Iteration 5:   log pseudolikelihood = -88.601013

------------------------------------------------------------------------------
             |               Robust
      pseudo |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       1.fab |   .6322634   .1665066    -1.74   0.082     .3773412    1.059405
             |
     disease |
          2  |   1.951343    .412121     3.17   0.002     1.289914    2.951931
          3  |   1.005533   .3586364     0.02   0.988     .4998088    2.022965
             |
         age |   .9856265   .0080274    -1.78   0.075      .970018    1.001486
------------------------------------------------------------------------------
```

Patients with AML and grade 4 or 5 (FAB = 1) have a 27% reduced disease free survival probability at 530 days, when adjusting for disease and age.

4.2 The survival function at several time points

In this example, we compute pseudovalues at five data points roughly equally spaced on the event scale: 50, 105, 170, 280, and 530 days. To fit the model $\log[-\log\{S(t\,|\,Z)\}] = \log\{\Lambda_0(t)\} + \beta Z$, we can use the cloglog link on the pseudovalues on failure probabilities; that is, we fit a Cox regression model for the five time points simultaneously.

```
. stpsurv, at(50 105 170 280 530) failure
Computing pseudo observations (progress dots indicate percent completed)
────┼─── 1 ───┼─── 2 ───┼─── 3 ───┼─── 4 ───┼─── 5
..................................................    50
..................................................   100
Generated pseudo variables: pseudo1-pseudo5

. generate id=_n

. reshape long pseudo, i(id) j(times)
(note: j = 1 2 3 4 5)

Data                               wide   ->   long
-----------------------------------------------------------------------------

Number of obs.                      137   ->     685
Number of variables                  32   ->      29
j variable (5 values)                     ->   times
xij variables:
            pseudo1 pseudo2 ... pseudo5   ->   pseudo
-----------------------------------------------------------------------------
```

```
. glm pseudo i.times i.fab i.disease age, link(cloglog) vce(cluster id) noheader
Iteration 0:   log pseudolikelihood = -468.74476
Iteration 1:   log pseudolikelihood = -457.41878  (not concave)
Iteration 2:   log pseudolikelihood = -406.98781
Iteration 3:   log pseudolikelihood = -365.23278
Iteration 4:   log pseudolikelihood =  -350.7435
Iteration 5:   log pseudolikelihood = -349.97156
Iteration 6:   log pseudolikelihood = -349.96409
Iteration 7:   log pseudolikelihood = -349.96409
                                    (Std. Err. adjusted for 137 clusters in id)
------------------------------------------------------------------------------
             |               Robust
      pseudo |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       times |
          2  |   1.114256   .3269323     3.41   0.001     .4734805    1.755032
          3  |   1.626173   .3567925     4.56   0.000     .9268721    2.325473
          4  |   2.004267   .3707305     5.41   0.000     1.277649    2.730885
          5  |   2.495327   .3824645     6.52   0.000     1.745711    3.244944
             |
       1.fab |   .7619547    .354821     2.15   0.032     .0665183    1.457391
             |
     disease |
          2  |  -1.195542   .4601852    -2.60   0.009    -2.097489   -.2935959
          3  |   .0036343   .3791488     0.01   0.992    -.7394838    .7467524
             |
         age |   .0130686   .0146629     0.89   0.373    -.0156702    .0418074
       _cons |  -2.981582   .6066311    -4.91   0.000    -4.170557   -1.792607
------------------------------------------------------------------------------
```

The estimated survival function in this model for a patient at time t with a set of covariates Z is $S(t) = \exp\{-\Lambda_0(t)e^{\beta Z}\}$, where

$$\begin{aligned}
\Lambda_0(50) &= \exp(-2.9816) = 0.051 \\
\Lambda_0(105) &= \exp(-2.9816 + 1.1143) = 0.155 \\
\Lambda_0(170) &= \exp(-2.9816 + 1.6262) = 0.258 \\
\Lambda_0(280) &= \exp(-2.9816 + 2.0043) = 0.376 \\
\Lambda_0(530) &= \exp(-2.9816 + 2.4953) = 0.615
\end{aligned}$$

The model shows that patients with AML who are at low risk have better disease free survival than ALL patients [RR $= \exp(-1.1955) = 0.30$] and that AML patients with grade 4 or 5 FAB have a lower disease free survival [RR $= \exp(0.7620) = 2.14$].

Without recomputing the pseudovalues, we can examine the effect of FAB over time.

(*Continued on next page*)

```
. generate fab50=(fab==1 & times==1)

. generate fab105=(fab==1 & times==2)

. generate fab170=(fab==1 & times==3)

. generate fab280=(fab==1 & times==4)

. generate fab530=(fab==1 & times==5)

. glm pseudo i.times fab50-fab530 i.disease age, link(cloglog) vce(cluster id)
> noheader eform

Iteration 0:   log pseudolikelihood = -471.86839
Iteration 1:   log pseudolikelihood = -464.24832  (not concave)
Iteration 2:   log pseudolikelihood = -406.31257
Iteration 3:   log pseudolikelihood = -361.28364
Iteration 4:   log pseudolikelihood = -349.90468
Iteration 5:   log pseudolikelihood = -349.44613
Iteration 6:   log pseudolikelihood = -349.43492
Iteration 7:   log pseudolikelihood = -349.43485
Iteration 8:   log pseudolikelihood = -349.43485
                                  (Std. Err. adjusted for 137 clusters in id)
------------------------------------------------------------------------------
             |               Robust
      pseudo |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       times |
          2  |    3.99608   2.023867     2.74   0.006     1.480921    10.78292
          3  |   8.225489   4.601898     3.77   0.000     2.747526    24.62531
          4  |   11.89654   6.835021     4.31   0.000     3.858093    36.68333
          5  |   19.20116   11.25862     5.04   0.000     6.084498    60.59409
             |
       fab50 |   4.047315   3.227324     1.75   0.080     .8480474    19.31586
      fab105 |   2.866106   1.433666     2.11   0.035      1.07525    7.639677
      fab170 |   2.008426    .795497     1.76   0.078     .9240856    4.365155
      fab280 |   2.022028   .7258472     1.96   0.050     1.000533    4.086419
      fab530 |   2.048864   .7838364     1.87   0.061     .9679838     4.33669
             |
     disease |
          2  |   .3024683   .1368087    -2.64   0.008     .1246451    .7339808
          3  |   .9993425   .3815547    -0.00   0.999     .4728471    2.112069
             |
         age |   1.012745   .0148835     0.86   0.389     .9839899     1.04234
------------------------------------------------------------------------------

. test fab50=fab105=fab170=fab280=fab530

 ( 1)  [pseudo]fab50 - [pseudo]fab105 = 0
 ( 2)  [pseudo]fab50 - [pseudo]fab170 = 0
 ( 3)  [pseudo]fab50 - [pseudo]fab280 = 0
 ( 4)  [pseudo]fab50 - [pseudo]fab530 = 0

           chi2(  4) =    1.73
         Prob > chi2 =    0.7855
```

The model shows that there is no statistically significant difference in the FAB effect over time ($p = 0.79$); that is, proportional hazards are not contraindicated for FAB.

4.3 The restricted mean

For the restricted mean time to treatment failure, we use the `stpmean` command. To illustrate, we look at a regression model for the mean time to treatment failure restricted to 1,500 days. Here we use the identity link function.

```
. stpmean, at(1500)
Computing pseudo observations (progress dots indicate percent completed)
────┼─── 1 ───┼─── 2 ───┼─── 3 ───┼─── 4 ───┼─── 5
..................................................    50
..................................................   100
Generated pseudo variable: pseudo
. glm pseudo i.fab i.disease age, link(id) vce(robust) noheader
Iteration 0:   log pseudolikelihood = -1065.6767
─────────────┬──────────────────────────────────────────────────────────────
             │               Robust
      pseudo │      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
─────────────┼──────────────────────────────────────────────────────────────
       1.fab │  -352.0442    123.311    -2.85   0.004    -593.7293    -110.359
             │
     disease │
          2  │   461.1214   134.0932     3.44   0.001     198.3036    723.9391
          3  │   78.00616   158.8357     0.49   0.623    -233.3061    389.3184
             │
         age │  -8.169236   5.060915    -1.61   0.106    -18.08845    1.749976
       _cons │    895.118   159.1586     5.62   0.000      583.173    1207.063
─────────────┴──────────────────────────────────────────────────────────────
```

Here we see that low-risk AML patients have the longest restricted mean life, namely, 461.1 days longer than ALL patients within 1,500 days.

4.4 Competing risks

For the cumulative incidence function, we use the `stpci` command to compute the pseudovalues. To illustrate, we use the complementary log–log model to the relapse cumulative incidence evaluated at 50, 105, 170, 280, and 530 days. The event of interest is death in remission. Here relapse is a competing event.

```
. stset tdfs, failure(trm==1)
     failure event:  trm == 1
obs. time interval:  (0, tdfs]
 exit on or before:  failure
──────────────────────────────────────────────────────────────────────────────
      137  total obs.
        0  exclusions
──────────────────────────────────────────────────────────────────────────────
      137  obs. remaining, representing
       42  failures in single record/single failure data
   107138  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =      2640
. generate compet=(trm==0 & relapse==1)
. stpci compet, at(50 105 170 280 530)
Computing pseudo observations (progress dots indicate percent completed)
────┼─── 1 ───┼─── 2 ───┼─── 3 ───┼─── 4 ───┼─── 5
..................................................    50
..................................................   100
Generated pseudo variables: pseudo1-pseudo5

. generate id=_n

. reshape long pseudo, i(id) j(times)
(note: j = 1 2 3 4 5)

Data                               wide   ->   long
─────────────────────────────────────────────────────────────────────────────

Number of obs.                      137   ->    685
Number of variables                  33   ->     30
j variable (5 values)                     ->   times
xij variables:
            pseudo1 pseudo2 ... pseudo5   ->   pseudo
─────────────────────────────────────────────────────────────────────────────

. fvset base none times

. glm pseudo i.times i.fab i.disease age, link(cloglog) vce(cluster id)
> noheader noconst eform

Iteration 0:   log pseudolikelihood = -462.96735  (not concave)
Iteration 1:   log pseudolikelihood = -348.27329
Iteration 2:   log pseudolikelihood = -221.69131
Iteration 3:   log pseudolikelihood = -198.31467
Iteration 4:   log pseudolikelihood = -197.38196
Iteration 5:   log pseudolikelihood = -197.37526
Iteration 6:   log pseudolikelihood = -197.37524
                                    (Std. Err. adjusted for 137 clusters in id)
──────────────────────────────────────────────────────────────────────────────
             │               Robust
      pseudo │     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
─────────────┼────────────────────────────────────────────────────────────────
       times │
          1  │   .0286012   .0292766    -3.47   0.001     .0038467      .21266
          2  │   .0791623   .0547411    -3.67   0.000     .0204131     .306993
          3  │   .1261608   .0823572    -3.17   0.002     .0350965    .4535083
          4  │   .1781601   .1117597    -2.75   0.006     .0521017    .6092124
          5  │   .2383869   .1488814    -2.30   0.022     .0700932    .8107537
             │
       1.fab │   3.104153    1.52811     2.30   0.021     1.182808    8.146518
             │
     disease │
          2  │   .1708985   .1154623    -2.61   0.009     .0454622    .6424309
          3  │   .7829133    .466016    -0.41   0.681     .2438093    2.514068
             │
         age │   1.014382   .0258272     0.56   0.575     .9650037    1.066286
──────────────────────────────────────────────────────────────────────────────
```

Here we are modeling $C(t \mid Z) = 1 - \exp\{-\Lambda_0(t)e^{\beta Z}\}$. Positive values of β for a covariate suggest a larger cumulative incidence for patients with $Z = 1$. The model suggests that the low-risk AML patients have the smallest risk of death in remission and the AML FAB 4/5 patients have the highest risk of death in remission.

5 Conclusion

The pseudovalue method is a versatile tool for regression analysis of censored time-to-event data. We have implemented the method for regression analysis of the survival under right-censoring, for the cumulative incidence function under possible competing risks, and for the restricted and conditional mean waiting time. Similar SAS macros and R functions were presented by Klein et al. (2008).

6 References

Andersen, P. K., M. G. Hansen, and J. P. Klein. 2004. Regression analysis of restricted mean survival time based on pseudo-observations. *Lifetime Data Analysis* 10: 335–350.

Andersen, P. K., and J. P. Klein. 2007. Regression analysis for multistate models based on a pseudo-value approach, with applications to bone marrow transplantation studies. *Scandinavian Journal of Statistics* 34: 3–16.

Andersen, P. K., J. P. Klein, and S. Rosthøj. 2003. Generalised linear models for correlated pseudo-observations, with applications to multi-state models. *Biometrika* 90: 15–27.

Andersen, P. K., and M. P. Perme. 2010. Pseudo-observations in survival analysis. *Statistical Methods in Medical Research* 19: 71–99.

Copelan, E. A., J. C. Biggs, J. M. Thompson, P. Crilley, J. Szer, J. P. Klein, N. Kapoor, B. R. Avalos, I. Cunningham, K. Atkinson, K. Downs, G. S. Harmon, M. B. Daly, I. Brodsky, S. I. Bulova, and P. J. Tutschka. 1991. Treatment for acute myelocytic leukemia with allogeneic bone marrow transplantation following preparation with BuCy2. *Blood* 78: 838–843.

Fine, J. P., and R. J. Gray. 1999. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association* 94: 496–509.

Graw, F., T. A. Gerds, and M. Schumacher. 2009. On pseudo-values for regression analysis in competing risks models. *Lifetime Data Analysis* 15: 241–255.

Kaplan, E. L., and P. Meier. 1958. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 53: 457–481.

Klein, J. P. 2006. Modeling competing risks in cancer studies. *Statistics in Medicine* 25: 1015–1034.

Klein, J. P., and P. K. Andersen. 2005. Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function. *Biometrics* 61: 223–229.

Klein, J. P., M. Gerster, P. K. Andersen, S. Tarima, and M. P. Perme. 2008. SAS and R functions to compute pseudo-values for censored data regression. *Computer Methods and Programs in Biomedicine* 89: 289–300.

Klein, J. P., B. Logan, M. Harhoff, and P. K. Andersen. 2007. Analyzing survival curves at a fixed point in time. *Statistics in Medicine* 26: 4505–4519.

Klein, J. P., and M. L. Moeschberger. 2003. *Survival Analysis: Techniques for Censored and Truncated Data*. 2nd ed. New York: Springer.

Liang, K.-Y., and S. L. Zeger. 1986. Longitudinal data analysis using generalized linear models. *Biometrika* 73: 13–22.

Perme, M. P., and P. K. Andersen. 2008. Checking hazard regression models using pseudo-observations. *Statistics in Medicine* 27: 5309–5328.

About the authors

Erik T. Parner has a PhD in statistics from the University of Aarhus. He is an associate professor of biostatistics at the University of Aarhus. His research fields are time-to-event analysis, statistical methods in epidemiology and genetics, and the etiology and changing prevalence of autism.

Per K. Andersen has a PhD in statistics and a DrMedSci degree in biostatistics, both from the University of Copenhagen. He is a professor of biostatistics at the University of Copenhagen. His main research fields are time-to-event analysis and statistical methods in epidemiology.