Regression analysis of censored data using pseudo-observations

Erik Parner
Section for Biostatistics
Aarhus University

Time to event data

Survival analysis

Competing risks

Pseudo observations

Reference: Parner ET, Andersen PK (2010). Regression analysis of censored data using pseudo-observations. Stata Journal, 10(3): 408-422.

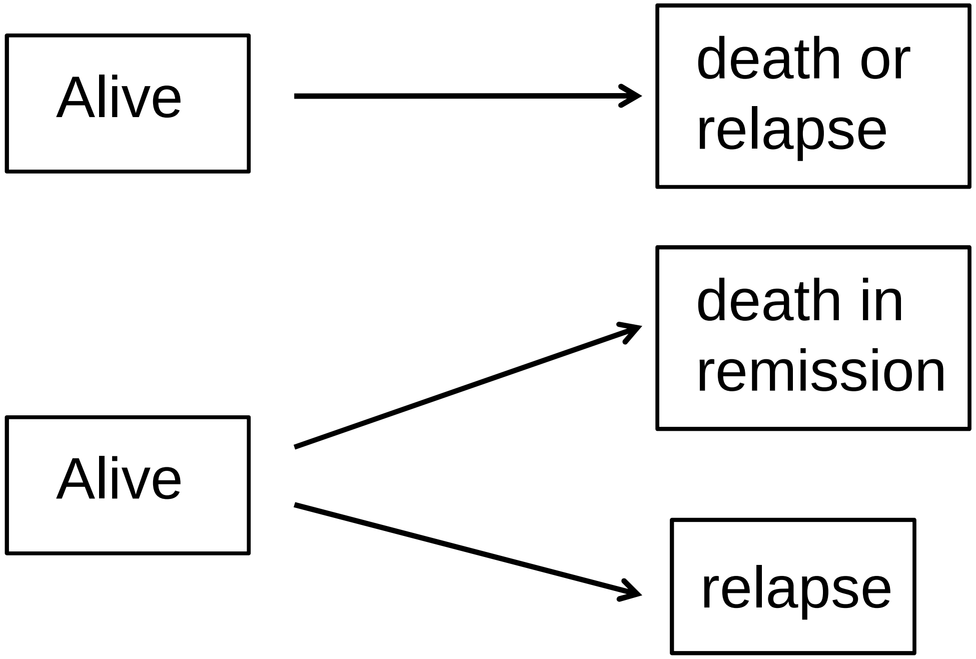
Coming extensions

Example 1: **Bone marrow transplantation for leukemia**

Sibling donor bone marrow transplants matched on human leukocyte antigen.

The data includes information on 137 transplant patients on

- time to death, relapse or lost to follow-up (`tdfs`),
- indicators of relapse and death (`relapse`, `trm`),
- indicator of treatment failure (`dfs`=`relapse|trm`).



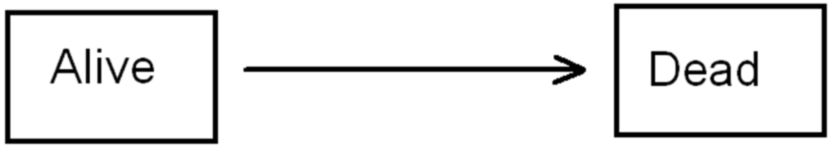
Three factors that may be related to the prognosis:

- **`disease`**;
  1-Acute Lymphocytic Leukemia (ALL),
  2-Low risk Acute Myeloid Leukemia (AML) and
  3-High risk AML,
- the French-American-British Disease grade for AML (**`fab`** = 1 if AML and Grade 4 or 5, 0 otherwise), and
- recipient age at transplant (**`age`**).

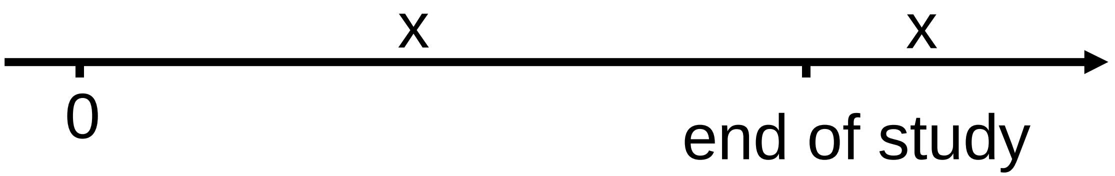
Time-to-event data

Survival analysis: *At most one event per individual.*

Examples: Mortality, disease incidence.



Data in standard setting:



Right censoring: We observe either

- the event before the end-of-study.
- or the individ is event free at the end-of-study.

Independent right censoring:

A basic requirement,

those still at risk at time t in our study should be representative for the population at time t.

or, equivalently,

those being censored at time t should be representative for the population at risk at time t.

Note,

- Some denote the condition **non-informative censoring**.
- In a regression analysis the independent censoring should be though of as for given covariates.

Describing the prognosis

Because of incomplete follow-up cause by the censoring, we rarely use basic descriptive and analytic methods such as simple averages for time-to-event data.

3 basic methods are often used to quantify the prognosis

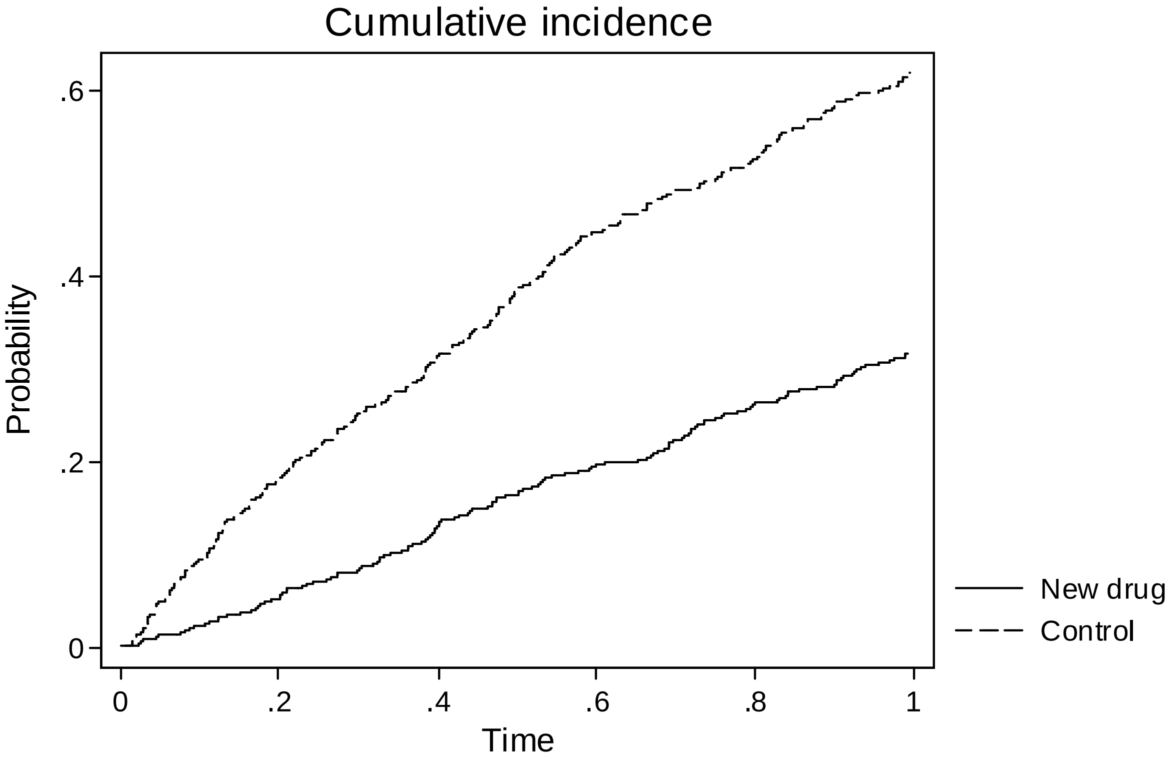
- **The survival function**; $S(t) = P(T>t)$, the probability of being event-free (alive) at time t, or equivalently $F(t)=CIP(t)=1-S(t) = P(T\leq t)$, the risk of event before time t.
- **The hazard function**; $h(t)=P(T \leq t+d \mid T\geq t)/d$, the probability of event (death) before t+d given alive at t, for a small time unit d.
- **The restricted mean**; $E[\min(T, t_0)]$ for at fixed time t_0, or the **expected number of years lost** before time t_0

$$t_0-E[\min(T, t_0)]$$

Example 2 – Evaluating a new drug

An analytic strategy often seen in the epidemiological and clinical literature.

Setting: Comparing a new drug to a control drug.
Data: time to event.



The analytic strategy involves, all available in Stata:

Step 1. Estimating the cumulative incidence proportions:

```
stset time, failure(failure==1)
sts list, at(0 1) by(group) failure
```

CIP(new): 32% (95% CI: 28% - 36%)
CIP(control): 62% (95% CI: 58% - 66%).

Step 2. Testing for a difference between the two groups.

```
sts test group
```

Log-rank test: p<0.001.

Step 3. Quantifying the difference.

```
stcox i.group
```

HR: 2.9 (95% CI: 2.4 - 3.6).

The analysis in step 1 focus on CIP, whereas the analyses in step 2-3 focus on rates. The analyses in step 2-3 are based on the assumption of proportional rates.
Why this analytic strategy?

Let’s take a closer look of the `sts list` command.

```
. sts list, at(0 1) by(group) failure

             Beg.                Failure        Std.
    Time    Total     Fail       Function       Error    [95% Conf. Int.]
---------------------------------------------------------------
group=1
       0        0        0        0.0000           .          .          .
       1      361      139        0.2780      0.0200   0.2409     0.3195
group=2
       0        0        0        0.0000           .          .          .
       1      182      318        0.6360      0.0215   0.5939     0.6781
---------------------------------------------------------------
```

The standard error (se) is Greenwood’s formula for the CIP, although the confidence intervals are based on the log-log transformation. So

$$se_{KM}(CIP_2 - CIP_1) = \sqrt{se(CIP_2)^2 + se(CIP_1)^2}$$

$$= \sqrt{0.0200^2 + 0.0215^2} = 0.0294$$

The confidence based on the Kaplan-Meier estimate

$$\begin{aligned}\text{CI}_{KM}(CIP_2 - CIP_1) &= 0.6360 - 0.2780 \pm 1.96 \cdot 0.0294 \\ &= 0.3580,\ \text{CI} = (0.3004, 0.4156)\end{aligned}$$

Comments

- The confidence interval for the relative risk is not easily estimated using the information from `sts list` command. The user have to apply the δ-method.
- Difficult to adjust for effect of covariates.

We may compute a confidence interval for the risk difference using the **pseudo-observation method**.

The pseudo value method creates a transformation of the time-to-event data given by the change in the Kaplan-Meier estimate when leaving out a single observation from the data set.

We then use these pseudo observations in a regression analysis – Generalized estimation equations (GEE) or Generalized linear models (GLM) - as if we had time-to-event data with no censoring.

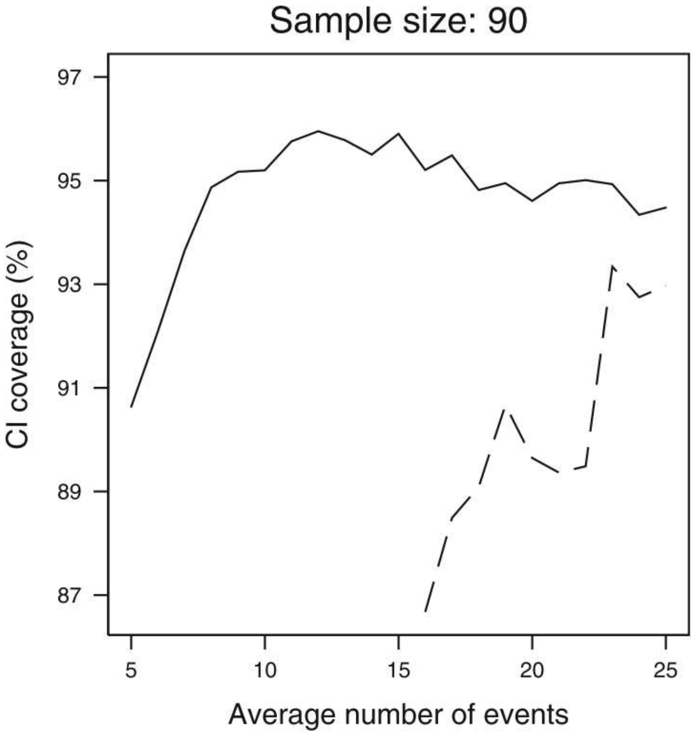
```
. stpsurv , at(1) failure
Computing pseudo observations (progress dots indicate percent
completed)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50
..................................................   100
Generated pseudo variable: pseudo

. glm pseudo i.group , link(id) vce(robust) noheader

Iteration 0:   log pseudolikelihood = -652.95466
---------------------------------------------------------------
        |            Robust
 pseudo | Coef.   Std. Err.      z    P>|z|  [95% Conf.Interval]
--------+------------------------------------------------------
2.group |  .358   .0294161   12.17   0.000  .3003456 .4156544
  _cons |  .278   .0200458   13.87   0.000   .238711  .317289
---------------------------------------------------------------
```

$$\mathrm{CI}_{KM}(CIP_1 - CIP_0) = 0.3580,\ \mathrm{CI} = (0.3004, 0.4156)$$

Almost the same as calculation by hand from the Kaplan-Meier estimate.

Comment.

Based on simulations, the pseudo value method is newer worse in terms of coverage and mean square error than the Kaplan-Meier for comparing two groups[1]. One example;



Pseudo observation

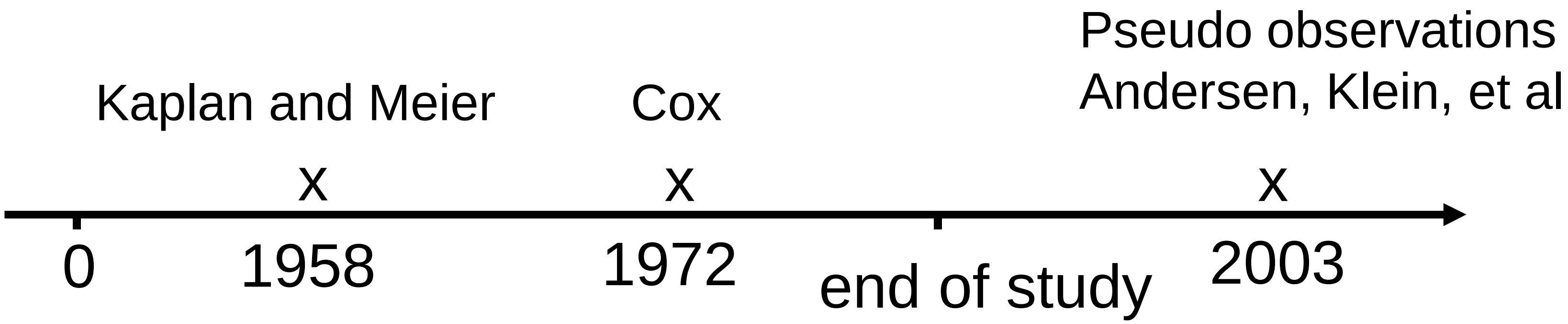
Kaplan-Meier

[1] Reference: Hansen, Andersen, Parner. Events per variable for risk differences and relative risks using pseudo-observations. Lifetime Data Analysis 2014.

So why this analytic strategy in Example 2?

Analyses of CIP have not been available in standard statistical packages.

Most analyst censor at around 2000;



The pseudo value method

Let $T_1, \ldots, T_n$ denote time-to-event outcomes with explanatory variables $X_1, \ldots, X_n$.

We are interested in a parameter of the form

$$\theta = \mathrm{E}[f(T_i)]$$

The function could be multivariate, for example

$$f(T_i) = (f_1(T_i), \ldots, f_M(T_i)) = (1(T_i > t_1), \ldots, 1(T_i > t_M))$$

for a series of time points $t_1, \ldots, t_M$ in which case

$$\theta = (\theta_1, \ldots, \theta_M) = (S(t_1), \ldots, S(t_M))$$

where $S(\cdot)$ is the survival function of T_i.

We are interested in a regression analysis of θ=E[$f(T_i)$] on X_i of the form

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

where g is a link function.

Right-censoring prevents us from observing all the T_i.

Suppose that $\hat{\theta}$ is an approximately unbiased estimator of the marginal mean θ = E[$f(T_i)$] which may be computed from the sample of right censored observations.

If $f(T_i)$ = I($T_i > t$) then $\theta = S(t)$ may be estimated using the Kaplan-Meier estimator.

The ith pseudo-observation is now defined as

$$\hat{\theta}_i = n \cdot \hat{\theta} - (n-1) \cdot \hat{\theta}_{-i}$$

Where $\hat{\theta}_{-i}$ is the "leave-one-out" estimator for θ based on all observation but the ith.

Estimates of the β's are obtained using the estimating equation

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

V_i is a working covariance matrix.

A sandwich estimator is used to estimate the variance

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

$$V\hat{a}r(U(\hat{\beta})) = \sum_i U_i(\hat{\beta})^T U_i(\hat{\beta})$$

$$V\hat{a}r(\hat{\beta}) = I(\hat{\beta})^{-1} V\hat{a}r(U(\hat{\beta})) I(\hat{\beta})^{-1}$$

The presented variance estimate is slightly conservative.

Comments:

- When V_i is a independence working covariance matrix the estimation procedure corresponds to fitting a **generalized linear model** with robust variance estimation.

- Events per variable rule of thumb[1];

  RD: event per variable=10,

  RR: event per variable=15.

  [1]Hansen, Andersen, Parner. (2014).

- We need not understand **why** it work.

  Its rather complicated. As for the Cox regression.

  We need to understand **how** it works.

  A good reference is

  Klein JP, Logan B, Harhoff M, Andersen PK. Analyzing survival curves at a fixed point in time. Stat Med. 2007;26(24):4505-19.

- When $f(T_i) = \mathrm{I}(T_i > t)$

$$\begin{aligned} g(\mathrm{E}[f(T_i) \mid X_i]) &= g(P(T_i > t \mid X_i)) \\ &= g(p_i) \\ &= \beta_0 + \beta_1 \cdot X_{i1} + \ldots + \beta_d \cdot X_{id} \end{aligned}$$

and in practice we often consider the two link functions $g(p)=p$ and g(p)=log(p).

Stata syntax

Pseudo-values for the survival function, the mean survival time and the cumulative incidence function for competing risks.

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

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

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

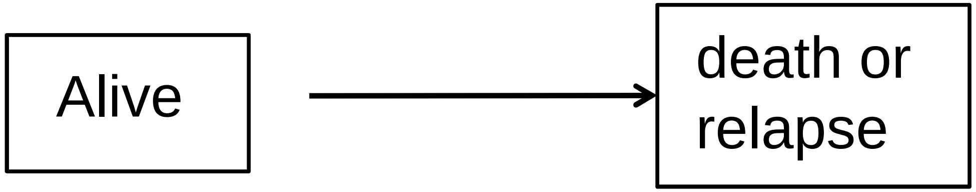
You must `stset` your data first.

Frequency weights are allowed in `stset` command.

In `stpci` an indicator variable for the competing risks should always be specified.

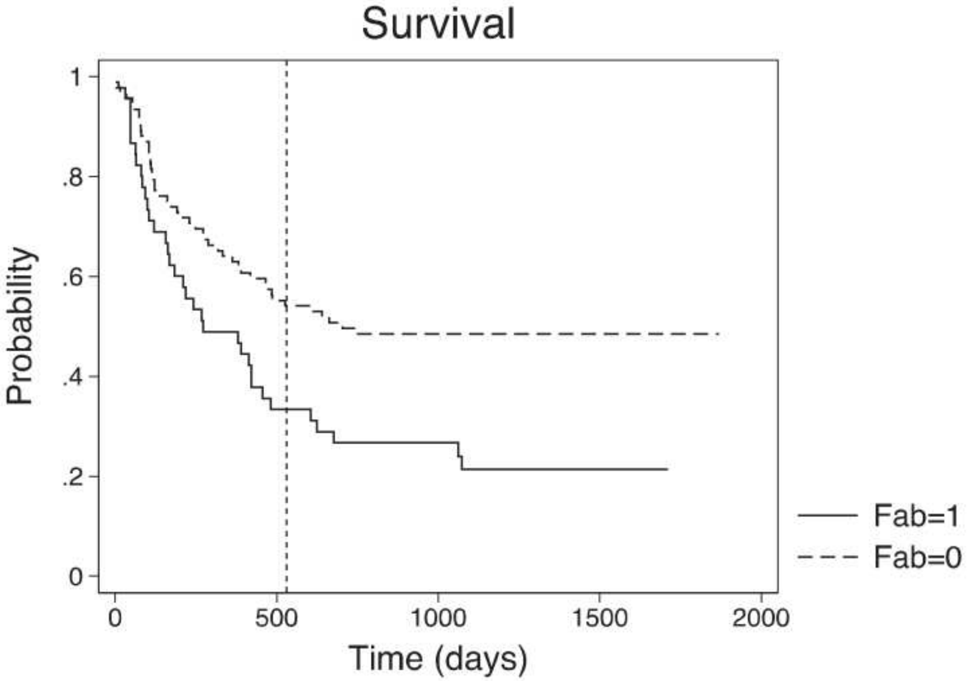
Example 1: Bone marrow transplantation for leukemia

Disease free survival probabilities for the single prognostic factor **FAB** at 530 days.



Alive

death or relapse



```
. use bmt, clear
. stset tdfs, failure(dfs==1)
  *** output omitted ***
. stpsurv , at(530)
  *** output omitted ***
. * GLM analysis of the pseudo values at 530 days.
. glm pseudo i.fab , fam(gauss) 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
---------------------------------------------------------------
```

Model: $p_i = S_i(530) = \beta_0 + \beta_1 \cdot FAB_i$

We estimate the risk difference for FAB by RD=-0.208 (95% CI:-0.381,-0.035).

```
. glm pseudo i.fab , fam(gauss) link(log) vce(robust) eform

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
 _cons | .5406774  .0522411  -6.36   0.000   .4473978  .6534053
----------------------------------------------------------------
```

Model: $\log(p_i) = \tilde{\beta}_0 + \tilde{\beta}_1 \cdot \mathrm{FAB}_i$

We estimate the the relative risk for FAB by RR=0.615 (95% CI: 0.389,0.974).

Suppose we wish to compute the relative risk for **FAB**, adjusting for disease as a categorical variable and **age** as a continuous variable.

Using the same pseudo-values we fit the generalized linear model.

```
. glm pseudo i.fab i.disease age, fam(gauss) link(log) ///
                                  vce(robust) eform
  *** output omitted ***
---------------------------------------------------------------
        |             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
  _cons | .5873784  .1602365  -1.95   0.051  .3441207  1.002594
---------------------------------------------------------------
```

The survival function at several time points

We compute pseudo-values at 5 data points roughly equally spaced on the event scale: 50, 105, 170, 280 and 530 days.

Often the interest would be to see if risk difference or relative risks change over time.

Here, for illustration, we will use another link cloglog, f(p)=log[-log[1-p]], to fit the model

$$\log[-\log\{S(t|X_i)\}] = \log(\Lambda_0(t))+\beta X_i$$

i.e. a Cox regression model for the 5 time points simultaneously.

This model can also fitted by the Stata procedure **`stcox`**.

```
. drop pseudo
. stpsurv , at(50 105 170 280 530) failure
 *** output omitted ***
Generated pseudo variables: pseudo1-pseudo5

. gen 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                  35   ->      32
j variable (5 values)                     ->   times
xij variables:
            pseudo1 pseudo2 ... pseudo5   ->   pseudo
---------------------------------------------------------------
```

```
. glm pseudo ibn.times i.fab i.disease age, fam(gauss)
             link(cloglog) vce(cluster id) nohead noconst eform
  *** output omitted ***
                    (Std. Err. adjusted for 137 clusters in id)
----------------------------------------------------------------
        |              Robust
 pseudo |   exp(b)  Std. Err.    z    P>|z|  [95%Conf.Interval]
--------+-------------------------------------------------------
  times |
    1   | .0507125  .0307638  -4.91   0.000  .0154436  .1665255
    2   | .1545363  .0662937  -4.35   0.000  .066662   .3582473
    3   | .2578417  .1078469  -3.24   0.001  .1135855  .5853067
    4   | .3763201  .1493346  -2.46   0.014  .1728925  .8191031
    5   |  .614925  .2347745  -1.27   0.203  .2909637  1.299587
  1.fab |  2.14246  .7601898   2.15   0.032  1.068781   4.29474
disease |
    2   | .3025399  .1392244  -2.60   0.009  .1227644  .7455777
    3   | 1.003641  .3805293   0.01   0.992  .4773603  2.110136
    age | 1.013154  .0148558   0.89   0.373   .984452  1.042694
----------------------------------------------------------------
```

The rate of treatment failure for FAB patients are 2-fold that of non-FAB patients when adjusting for disease and age.

Without re-computing the pseudo-values we can examine the effect of FAB over time.

```
. gen fab50=(fab==1 & times==1)
. gen fab105=(fab==1 & times==2)
. gen fab170=(fab==1 & times==3)
. gen fab280=(fab==1 & times==4)
. gen fab530=(fab==1 & times==5)
. glm pseudo i.times fab50-fab530 i.disease age, fa(gauss) ///
              lin(cloglog) vce(cluster id) eform
  *** output omitted ***
---------------------------------------------------------------
        |            Robust
 pseudo |   exp(b)  Std. Err.    z  P>|z|   [95% Conf.Interval]
--------+------------------------------------------------------
 *** output omitted ***
  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
 *** output omitted ***
---------------------------------------------------------------
```

```
. 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
```

Similar analysis could be made for the relative risk using the log-link function.

The restricted mean

$$E[\min(T,t_0)] = \int_0^{t_0} S(u)du$$

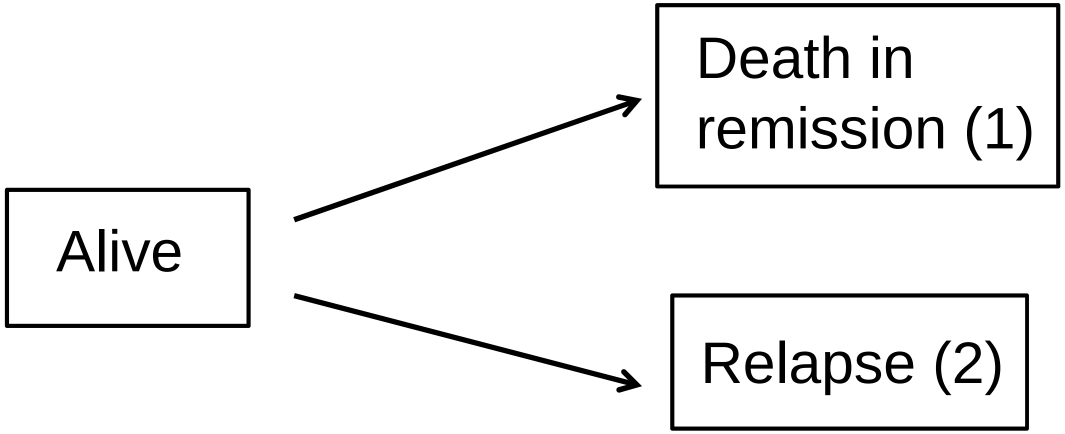
To illustrate we look at a regression model for the mean time to treatment failure restricted to 1500 days. Here we use the identity link function.

```
. use bmt, clear
. stset tdfs, failure(dfs==1)
  *** output omitted ***
. stpmean , at(1500)
  *** output omitted ***
  Generated pseudo variable: pseudo
```

```
. glm pseudo i.fab i.disease age, fam(gauss) li(id) vce(robust)
  *** output omitted ***
----------------------------------------------------------------
        |               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 AML low risk patients have the longest restricted mean life, namely 461.1 days longer than ALL patients within 1500 days.

Competing risk



We need to generalize the 3 ways to quantify the prognosis.

- **The CIP function**; $F_1(t)=CIP_1(t)=P(T\leq t, \text{cause } 1)$, the risk of event of cause 1 before time t.
- **The cause specific hazard function**; $h_1(t)=P(T \leq t+d \text{ , cause } 1 | T\geq t)/d$, the probability of event of cause 1 before t+d given alive at t, for a small time unit d.
- **Life years lost according to causes of death**;

$$\int_0^{t_0} F_1(u)du$$

Cause-specific cumulative risk

Often the interest would be to analyse risk difference or relative risks for the cause specific cumulative risk.

First, for illustration, we will however use the cloglog link function to cumulative incidence of death in remission evaluated at 50, 105, 170, 280 and 530 days.

```
. use bmt, clear
. stset tdfs, failure(trm==1)
 *** output omitted ***
. gen compet=(trm==0 & relapse==1)
. stpci compet, at(50 105 170 280 530)
 *** output omitted ***
. gen id=_n
. reshape long pseudo, i(id) j(times)
 *** output omitted ***
```

Here we are modeling

$$\mathrm{CIP}_1(t|X) = 1-\exp\{-\Lambda_0(t)\exp(\beta X)\}.$$

Positive values of β for a covariate suggest a larger cumulative incidence for patients with $X = 1$.

This is the Fine-Gray model that is fitted by the Stata procedure **`stcrreg`**.

```
. glm pseudo ibn.times i.fab i.disease age, fam(gauss) ///
             link(cloglog) vce(cluster id) noconst eform
  *** output omitted ***
---------------------------------------------------------------
        |            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
---------------------------------------------------------------
```

The model suggests that the AML low risk patients have the smallest risk of death in remission and the AML FAB 4/5 the highest risk of death in remission.

The interpretation of the β’s are generally difficult. The exp(β) are subhazard ratios relates to the subdistribution hazards

$P(T \leq t+d, \text{cause } 1 \mid T \geq t$ **or {T<t and relapse}**$)$

the instantaneous rate of failure per time unit from cause j among those who are either alive or have had a competing event at time t.

The individuals are then followed after they a competing event, some of these have died from relapse.

Let us model the cause specific cumulative risk directly

$$CIP_1(t|X) = \beta X$$

```
glm pseudo i.fab i.disease age if(times==5), fam(gauss) ///
                          link(id) vce(robust)
  *** output omitted ***
---------------------------------------------------------------
        |             Robust
 pseudo |    Coef.  Std. Err.    z   P>|z|  [95% Conf.
Interval]
--------+------------------------------------------------------
  1.fab | .2600768  .0912255   2.85  0.004   .0812781  .4388754
disease |
     2  |-.2363511  .0888625  -2.66  0.008  -.4105184 -.0621837
     3  | .0079055  .1172756   0.07  0.946  -.2219504  .2377615
    age | .0024241  .0039748   0.61  0.542  -.0053663  .0102145
  _cons | .2075862  .1139077   1.82  0.068  -.0156688  .4308411
---------------------------------------------------------------
```

Life years lost according to causes of death

The overall survival function S(t) and the cause specific CIP functions satisty

$$S(t) + F_1(t) + F_2(t) = 1$$

Hence the **expected number of years lost** before time t_0 can be decomposed

$$t_0 - E[\min(T, t_0)] = \int_0^{t_0} 1 - S(u)du$$

$$= \int_0^{t_0} F_1(u)du + \int_0^{t_0} F_2(u)du$$

A decomposition of number of life years lost according to causes of death.

It cannot be analysed by the Stata functions `stpsurv`, `stpmean` or `stpci`.

Coming extensions

- The number of life years lost according to causes of death.
- Extensions to delayed entry.
- A faster implementation by making the Mata code more vectorized and limiting the number of calculations carried out. Individuals affecting the Kaplan-Meier between the same event times will always have the same Kaplan-Meier based pseudo-observation.