Survival analysis : from basic concepts to open research questions

Ecole d'été, Villars-sur-Ollon, 2-5 September 2018

Ingrid Van Keilegom

ORSTAT – KU Leuven

KU LEUVEN

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Table of Contents

1. Basic concepts
2. Cure models
   - Introduction
   - Ongoing research
3. Dependent censoring
   - Introduction
   - Ongoing research
4. Measurement errors
   - Introduction
   - Ongoing research

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Part I : Basic concepts

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Basic concepts

What is ‘Survival analysis’ ?

- Survival analysis (or duration analysis) is an area of statistics that models and studies the time until an event of interest takes place.
- In practice, for some subjects the event of interest cannot be observed for various reasons, e.g.
  - the event is not yet observed at the end of the study
  - another event takes place before the event of interest
  - ...
- In survival analysis the aim is
  - to model ‘time-to-event data’ in an appropriate way
  - to do correct inference taking these special features of the data into account.

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Examples

- Medicine :
  - time to death for patients having a certain disease
  - time to getting cured from a certain disease
  - time to relapse of a certain disease
- Agriculture :
  - time until a farm experiences its first case of a certain disease
- Sociology (‘duration analysis’) :
  - time to find a new job after a period of unemployment
  - time until re-arrest after release from prison
- Engineering (‘reliability analysis’) :
  - time to the failure of a machine

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Common functions in survival analysis

- Let T be a non-negative continuous random variable, representing the time until the event of interest
- Denote

$F(t) = P(T \leq t)$	distribution function
$f(t)$	probability density function

- For survival data, we consider rather

$S(t)$	survival function
$H(t)$	cumulative hazard function
$h(t)$	hazard function

- Knowing one of these functions suffices to determine the other functions

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Survival function :

$$S(t) = P(T > t) = 1 - F(t)$$

- Probability that a randomly selected individual will survive beyond time t
- Decreasing function, taking values in $[0, 1]$
- Equals 1 at $t = 0$ and 0 at $t = \infty$

Cumulative hazard function :

$$H(t) = -\log S(t)$$

- Increasing function, taking values in $[0, +\infty]$
- $S(t) = \exp(-H(t))$

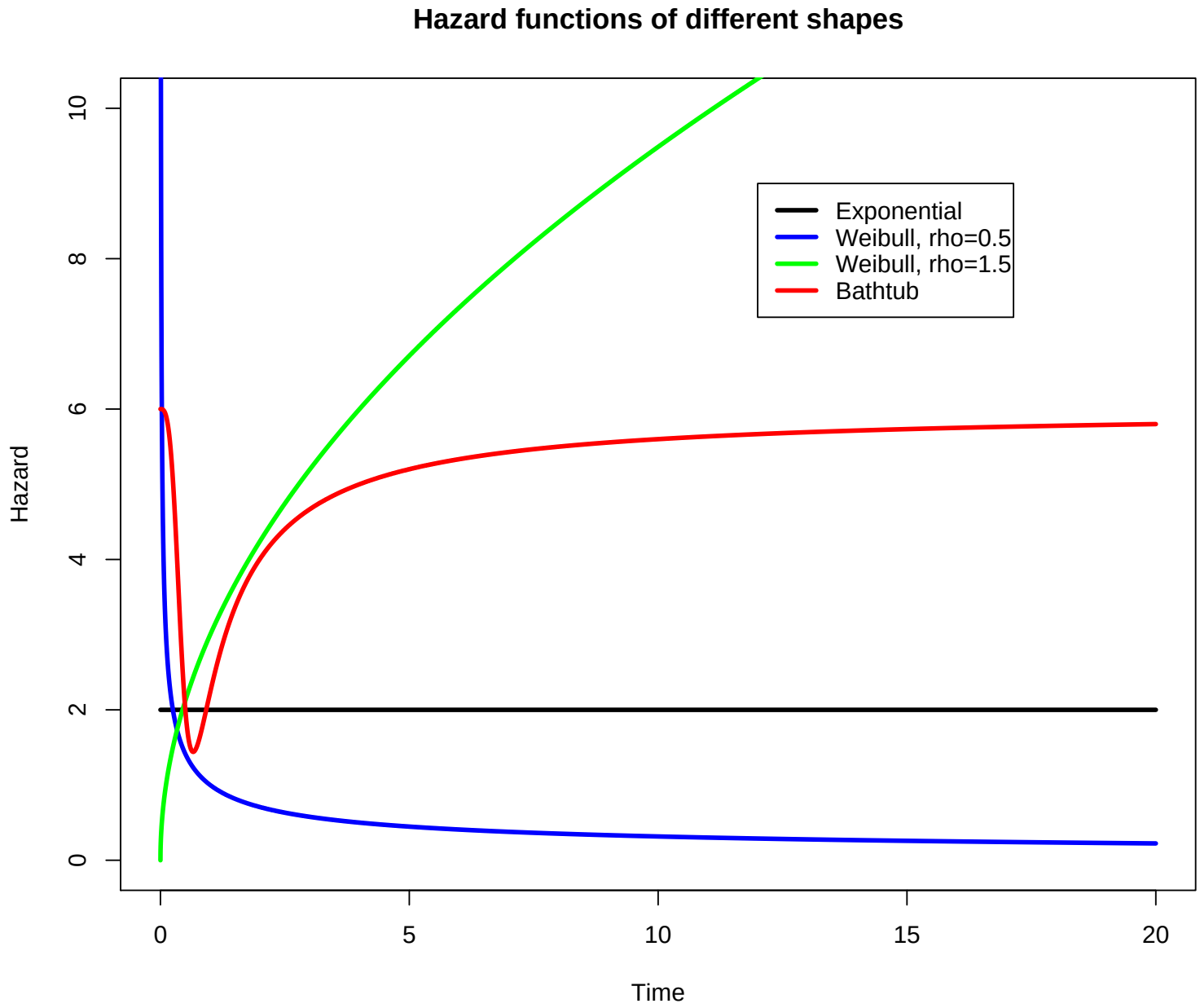
Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Hazard function (or hazard rate) :

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} \\ &= \frac{1}{P(T \geq t)} \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t} \\ &= \frac{f(t)}{S(t)} = \frac{-d}{dt} \log S(t) = \frac{d}{dt} H(t) \end{aligned}$$

- $h(t)$ measures the instantaneous risk of dying right after time t given the individual is alive at time t
- Positive function (not necessarily increasing or decreasing)
- The hazard function $h(t)$ can have many different shapes and is therefore a useful tool to summarize survival data

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research



Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Random right censoring :

- For certain individuals under study, only a lower bound for the true survival time is observed
- Ex : In a clinical trial, some patients have not yet died at the time of the analysis of the data
- Two latent variables :

$$T = \text{survival time}$$

$$C = \text{censoring time}$$

$\Rightarrow$ Data : (Y, Δ) with

$$Y = \min(T, C)$$

$$\Delta = I(T \leq C)$$

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

- Censoring can occur for various reasons :
  - end of study
  - lost to follow up
  - competing event (e.g. death due to some cause other than the cause of interest)
  - patient withdrawing from the study, change of treatment, ...
- We assume that T and C are independent (called independent censoring)

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

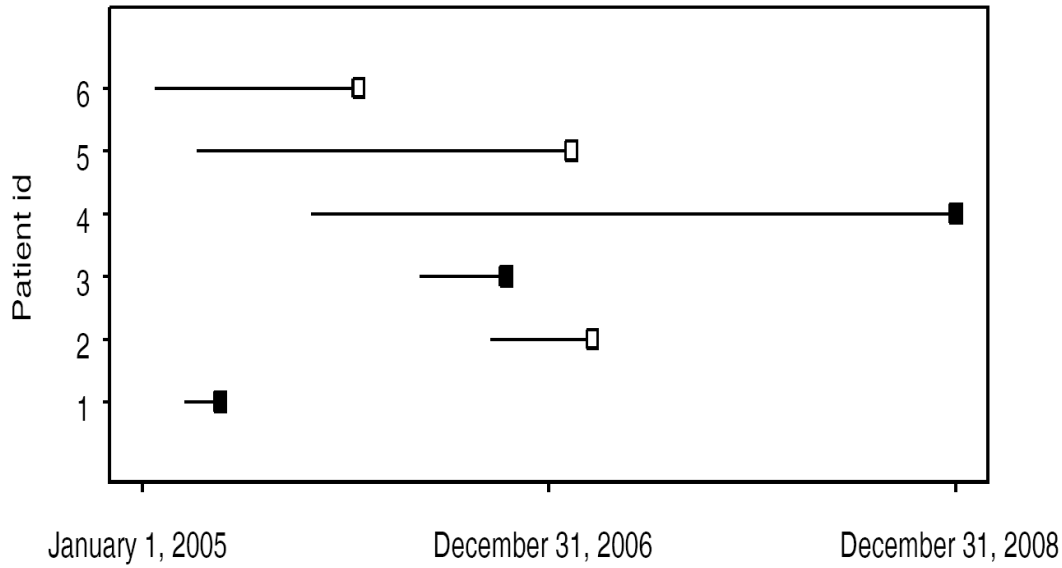
Example : Random right censoring in HIV study

- Study enrolment: January 2005 - December 2006
- Study end: December 2008
- Objective: HIV patients followed up to death due to AIDS or AIDS related complication (time in month from confirmed diagnosis)
- Possible causes of censoring :
  - death due to other cause
  - lost to follow up / dropped out
  - still alive at the end of study

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Table: Data of first 6 patients in HIV study

Patient id	Entry Date	Date last seen	Status	Time	Censoring
1	18 March 2005	20 June 2005	Dropped out	3	0
2	19 Sept 2006	20 March 2007	Dead due to AIDS	6	1
3	15 May 2006	16 Oct 2006	Dead due to accident	5	0
4	01 Dec 2005	31 Dec 2008	Alive	37	0
5	9 Apr 2005	10 Feb 2007	Dead due to AIDS	22	1
6	25 Jan 2005	24 Jan 2006	Dead due to AIDS	12	1



Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Nonparametric estimation

Likelihood for randomly right censored data

- Random sample of size n : (Y_i, Δ_i) $(i = 1, \ldots, n)$ with

$$\begin{aligned} Y_i &= \min(T_i, C_i) \\ \Delta_i &= I(T_i \leq C_i) \end{aligned}$$

and where

$T_1, \ldots, T_n$ (latent) survival times
$C_1, \ldots, C_n$ (latent) censoring times

- Denote

$f(\cdot)$ and $F(\cdot)$ for the density and distribution of T
$g(\cdot)$ and $G(\cdot)$ for the density and distribution of C

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

It can be shown that the likelihood for random right censored data equals :

$$\prod_{i=1}^{n} \left[(1 - G(Y_i))f(Y_i)\right]^{\Delta_i} \left[(1 - F(Y_i))g(Y_i)\right]^{1-\Delta_i}$$

We assume that censoring is uninformative, i.e. the distribution of the censoring times does not depend on the parameters of interest related to the survival function.

$\Rightarrow$ The factors $(1 - G(Y_i))^{\Delta_i}$ and $g(Y_i)^{1-\Delta_i}$ are non-informative for inference on the survival function

$\Rightarrow$ They can be removed from the likelihood, leading to

$$\prod_{i=1}^{n} f(Y_i)^{\Delta_i} S(Y_i)^{1-\Delta_i} = \prod_{i=1}^{n} h(Y_i)^{\Delta_i} S(Y_i)$$

where $S(\cdot) = 1 - F(\cdot)$ (survival function)
$h(\cdot) = f(\cdot)/S(\cdot)$ (hazard function)

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Kaplan-Meier (KM) estimator of the survival function

- Kaplan and Meier (*JASA*, 1958)
- Nonparametric estimation of the survival function for right censored data
- Based on the order in which events and censored observations occur

Notations :

- n observations $Y_1, \ldots, Y_n$ with censoring indicators $\Delta_1, \ldots, \Delta_n$
- r distinct event times ($r \leq n$)
- ordered event times : $Y_{(1)}, \ldots, Y_{(r)}$ and corresponding number of events: $d_{(1)}, \ldots, d_{(r)}$
- $R_{(j)}$ is the size of the risk set at event time $Y_{(j)}$

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

- Log-likelihood for right censored data :

$$\sum_{i=1}^{n} \left[\Delta_i \log f(Y_i) + (1 - \Delta_i) \log S(Y_i) \right]$$

- Replacing the density function $f(Y_i)$ by $S(Y_{i-}) - S(Y_i)$, yields the nonparametric log-likelihood :

$$\log L = \sum_{i=1}^{n} \left[\Delta_i \log(S(Y_{i-}) - S(Y_i)) + (1 - \Delta_i) \log S(Y_i) \right]$$

- Aim : finding an estimator $\hat{S}(\cdot)$ which maximizes $\log L$
- It can be shown that the maximizer of $\log L$ takes the following form :

$$\hat{S}(t) = \prod_{j: Y_{(j)} \leq t} (1 - h_{(j)}),$$

for some $h_{(1)}, \ldots, h_{(r)}$

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

- Plugging-in $\hat{S}(\cdot)$ into the log-likelihood, gives after some algebra :

$$\log L = \sum_{j=1}^{r} \left[d_{(j)} \log h_{(j)} + \left(R_{(j)} - d_{(j)} \right) \log(1 - h_{(j)}) \right]$$

- Using this expression to solve

$$\frac{d}{dh_{(j)}} \log L = 0$$

leads to

$$\hat{h}_{(j)} = \frac{d_{(j)}}{R_{(j)}}$$

- Plugging in this estimate $\hat{h}_{(j)}$ in $\hat{S}(t) = \prod_{j: Y_{(j)} \leq t} (1 - h_{(j)})$ we obtain :

$$\hat{S}(t) = \prod_{j: Y_{(j)} \leq t} \frac{R_{(j)} - d_{(j)}}{R_{(j)}} = \text{ Kaplan-Meier estimator}$$

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

- Step function with jumps at the event times
- If the largest observation, say Y_n, is censored :
  - $\hat{S}(t)$ does not attain 0
  - Impossible to estimate $S(t)$ consistently beyond Y_n
  - Various solutions :
    - Set $\hat{S}(t) = 0$ for $t \geq Y_n$
    - Set $\hat{S}(t) = \hat{S}(Y_n)$ for $t \geq Y_n$
    - Let $\hat{S}(t)$ be undefined for $t \geq Y_n$
- When all data are uncensored, the Kaplan-Meier estimator reduces to the empirical distribution function

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Asymptotic normality of the KM estimator

The variance can be consistently estimated by (Greenwood formula)

$$\widehat{Var}(\hat{S}(t)) = \hat{S}^2(t) \sum_{j:Y_{(j)} \leq t} \frac{d_{(j)}}{R_{(j)}(R_{(j)} - d_{(j)})}$$

Asymptotic normality of $\hat{S}(t)$:

$$\frac{\hat{S}(t) - S(t)}{\sqrt{\widehat{Var}(\hat{S}(t))}} \xrightarrow{d} N(0,1)$$

Nelson-Aalen estimator of the cumulative hazard function

Proposed by Nelson (1972) and Aalen (1978) :

$$\hat{H}(t) = \sum_{j:Y_{(j)} \leq t} \frac{d_{(j)}}{R_{(j)}} \qquad \text{for } t \leq Y_{(r)}$$

The estimator is also asymptotically normal

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Point estimate of the mean survival time

- Nonparametric estimator can be obtained using the Kaplan-Meier estimator, since

$$\mu = E(T) = \int_0^\infty t f(t) dt = \int_0^\infty S(t) dt$$

$\Rightarrow$ We can estimate μ by replacing $S(t)$ by the KM estimator $\hat{S}(t)$

- But, $\hat{S}(t)$ is inconsistent in the right tail if the largest observation (say Y_n) is censored
  - Proposal 1 : assume Y_n experiences the event immediately after the censoring time :

$$\hat{\mu}_{Y_n} = \int_0^{Y_n} \hat{S}(t) dt$$

  - Proposal 2 : restrict integration to a predetermined interval $[0, t_{max}]$ and consider $\hat{S}(t) = \hat{S}(Y_n)$ for $Y_n \leq t \leq t_{max}$:

$$\hat{\mu}_{t_{max}} = \int_0^{t_{max}} \hat{S}(t) dt$$

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Point estimate of the median survival time

- Advantages of the median over the mean :
  - As survival function is often skewed to the right, the mean is often influenced by outliers, whereas the median is not
  - Median can be estimated in a consistent way (if censoring is not too heavy)
- An estimator of the p^{th} quantile x_p is given by :

$$\hat{x}_p = \inf\left\{t \mid \hat{S}(t) \leq 1 - p\right\}$$

$\Rightarrow$ An estimate of the median is given by $\hat{x}_{p=0.5}$

- The variance of $\hat{x}_p$ can be estimated by :

$$\widehat{Var}(\hat{x}_p) = \frac{\widehat{Var}(\hat{S}(x_p))}{\hat{f}^2(x_p)},$$

where $\hat{f}$ is an estimator of the density f

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

- Estimation of f involves smoothing techniques and the choice of a bandwidth sequence
  $\Rightarrow$ We prefer not to use this variance estimator in the construction of a CI
- Thanks to the asymptotic normality of $\hat{S}(x_p)$:

$$P\left(-z_{\alpha/2} \leq \frac{\hat{S}(x_p) - S(x_p)}{\sqrt{\widehat{Var}(\hat{S}(x_p))}} \leq z_{\alpha/2} \right) \approx 1 - \alpha,$$

with obviously $S(x_p) = 1 - p$.

$\Rightarrow$ A $100(1-\alpha)\%$ CI for x_p is given by

$$\left\{ t : -z_{\alpha/2} \leq \frac{\hat{S}(t) - (1-p)}{\sqrt{\widehat{Var}(\hat{S}(t))}} \leq z_{\alpha/2} \right\}$$

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Example : Schizophrenia patients

- Schizophrenia is one of the major mental illnesses encountered in Ethiopia

  $\rightarrow$ disorganized and abnormal thinking, behavior and language + emotionally unresponsive

  $\rightarrow$ higher mortality rates due to natural and unnatural causes

- Project on schizophrenia in Butajira, Ethiopia

  $\rightarrow$ survey of the entire population (68491 individuals) in the age group 15-49 years

$\Rightarrow$ 280 cases of schizophrenia identified and followed for 5 years (1997-2001)

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Table: Data on schizophrenia patients

Patid	Time	Censor	Education	Onset	Marital	Gender	Age
1	1	1	1	37	3	1	44
2	3	1	3	15	2	2	23
3	4	1	6	26	1	1	33
4	5	1	12	25	1	1	31
5	5	0	5	29	3	1	33
...							
278	1787	0	2	16	2	1	18
279	1792	0	2	23	1	1	25
280	1794	1	2	28	1	1	35

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

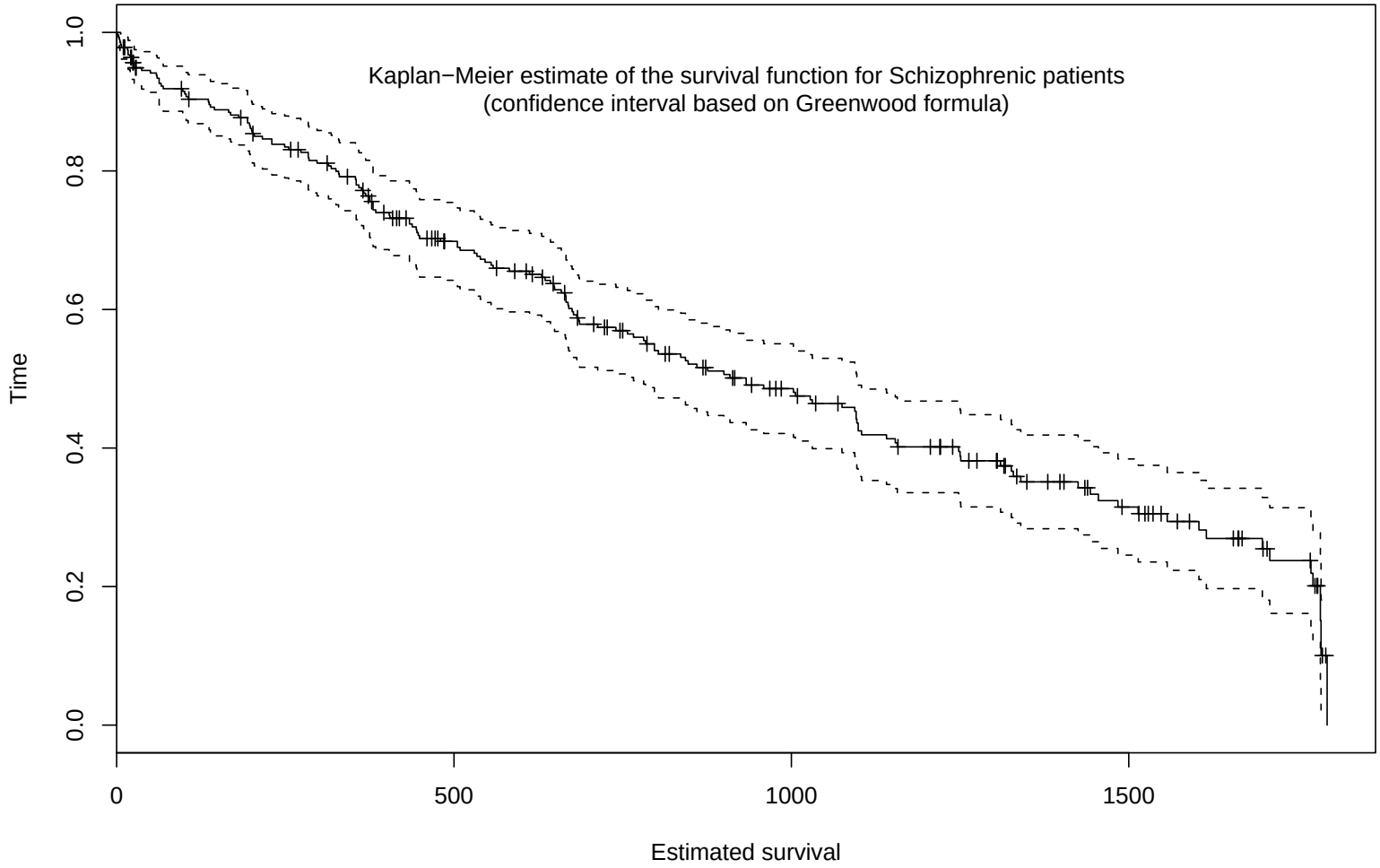
In R : survfit

```
schizo <- read.table("c://...//Schizophrenia.csv", header=T,sep=";")
KM_schizo_g <- survfit(Surv(Time,Censor)~1,data=schizo,
   type="kaplan-meier", conf.type="plain")
plot(KM_schizo_g, conf.int=T, xlab="Estimated survival", ylab="Time",
   yscale=1)
mtext("Kaplan-Meier estimate of the survival function for Schizophrenic
   patients", 3,-3)
mtext("(confidence interval based on Greenwood formula)", 3,-4)
```

In SAS : proc lifetest

```
title1 'Kaplan-Meier estimate of the survival function for Schizophrenic
   patients';
proc lifetest method=km width=0.5 data=schizo;
time Time*Censor(0);
run;
```

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research



Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

```
> KM_schizo_g
Call: survfit(formula = Surv(Time, Censor) ~ 1, data = schizo, type =
"kaplan-meier", conf.type = "plain")

      n  events  median 0.95LCL 0.95UCL
    280     163     933     766    1099

> summary(KM_schizo_g)
Call: survfit(formula = Surv(Time, Censor) ~ 1, data = schizo, type =
"kaplan-meier", conf.type = "plain")

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1    280       1    0.996 0.00357       0.9894        1.000
    3    279       1    0.993 0.00503       0.9830        1.000
    4    277       1    0.989 0.00616       0.9772        1.000
 …
 1770     13       1    0.219 0.03998       0.1409        0.298
 1773     12       1    0.201 0.04061       0.1214        0.281
 1784      8       2    0.151 0.04329       0.0659        0.236
 1785      6       2    0.100 0.04092       0.0203        0.181
 1794      1       1    0.000      NA           NA           NA
```

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Proportional hazards models

The semiparametric proportional hazards (PH) model

- Cox, 1972
- Popular regression model in survival analysis
- We will work with semiparametric proportional hazards models, but there also exist parametric variations

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Simplest expression of the model

- Case of two treatment groups (Treated vs. Control) :

  $$h_T(t) = \psi h_C(t),$$

  with $h_T(t)$ and $h_C(t)$ the hazard function of the treated and control group
- Proportional hazards model :
  - Ratio $\psi = h_T(t)/h_C(t)$ is constant over time
  - $\psi < 1$ ($\psi > 1$): hazard of the treated group is smaller (larger) than the hazard of the control group at any time
  - Survival curves of the 2 treatment groups can never cross each other

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

More generalizable expression of the model

- Consider a treatment covariate x_i (0 = control, 1 = treatment) and an exponential relationship between the hazard and the covariate x_i :

$$h_i(t) = \exp(\beta x_i) h_0(t),$$

with

  - $h_i(t)$: hazard function for subject i
  - $h_0(t)$: hazard function of the control group
  - $\exp(\beta) = \psi$: hazard ratio (HR) or relative risk

- Other functional relationships can be used between the hazard and the covariate

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

More complex model

- Consider a set of covariates $x_i = (x_{i1}, \ldots, x_{ip})^T$ for subject i :

$$h_i(t) = h_0(t) \exp(\beta^T x_i),$$

with

  - β : the $p \times 1$ parameter vector
  - $h_0(t)$: the baseline hazard function (i.e. hazard for a subject with $x_{ij} = 0$, $j = 1, \ldots, p$)

- Proportional hazards (PH) assumption : ratio of the hazards of two subjects with covariates x_i and x_j is constant over time :

$$\frac{h_i(t)}{h_j(t)} = \frac{\exp(\beta^T x_i)}{\exp(\beta^T x_j)}$$

- Semiparametric PH model : leave the form of $h_0(t)$ completely unspecified and estimate the model in a semiparametric way

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Fitting the semiparametric PH model

- Based on likelihood maximization
- As $h_0(t)$ is left unspecified, we maximize a so-called partial likelihood instead of the full likelihood :

$$L(\beta) = \prod_{j=1}^{r} \frac{\exp\left(x_{(j)}^T \beta\right)}{\sum_{k \in R(Y_{(j)})} \exp\left(x_k^T \beta\right)}$$

where

- r observed event times
- $Y_{(1)}, \ldots, Y_{(r)}$ ordered event times
- $x_{(1)}, \ldots, x_{(r)}$ corresponding covariate vectors
- $R(Y_{(j)})$ risk set at time $Y_{(j)}$

- It can be shown that the partial likelihood is actually a profile likelihood, in which the baseline hazard is profiled out.
- This expression is used to estimate β through numerical maximization

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Inference under the Cox model

- Variance-covariance matrix of $\hat{\beta}$ can be approximated by the inverse of the information matrix evaluated at $\hat{\beta}$ $\rightarrow$ $Var(\hat{\beta}_h)$ can be approximated by $[I(\hat{\beta})]^{-1}_{hh}$
- Properties (consistency, asymptotic normality) of $\hat{\beta}$ are well established (Gill, 1984)
- A 100(1-α)% confidence interval for β_h is given by

$$\hat{\beta}_h \pm z_{\alpha/2}\sqrt{Var(\hat{\beta}_h)}$$

- Testing hypotheses of the form

$$H_0 : \beta_1 = \beta_{10}$$

$$H_1 : \beta_1 \neq \beta_{10}$$

regarding a subvector β_1 of β, can be done using the Wald, score or likelihood-ratio test, exactly as in parametric regression models.

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Example : Active antiretroviral treatment cohort study

- CD4 cells protect the body from infections and other types of disease
  $\rightarrow$ if count decreases beyond a certain threshold the patients will die
- As HIV infection progresses, most people experience a gradual decrease in CD4 count
- Highly Active AntiRetroviral Therapy (HAART)
  - AntiRetroviral Therapy (ART) + 3 or more drugs
  - Not a cure for AIDS but greatly improves the health of HIV/AIDS patients
- Data from a study conducted in Ethiopia :
  - 100 individuals older than 18 years and placed under HAART for the last 4 years
  - only use data collected for the first 2 years

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Table: Data of HAART Study

Pat ID	Time	Censo-ring	Gen-der	Age	Weight	Func. Status	Clin. Status	CD4	ART
1	699	0	1	42	37	2	4	3	1
2	455	1	2	30	50	1	3	111	1
3	705	0	1	32	57	0	3	165	1
4	694	0	2	50	40	1	3	95	1
5	86	0	2	35	37	0	4	34	1
...									
97	101	0	1	39	37	2	.	.	1
98	709	0	2	35	66	2	3	103	1
99	464	0	1	27	37	.	.	.	2
100	537	1	2	30	76	1	4	1	1

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

How is survival influenced by gender and age ?

- Define agecat $= 1$ if age < 40 years
  $= 2$ if age ≥ 40 years
- Define gender $= 1$ if male
  $= 2$ if female
- Fit a semiparametric PH model including gender and agecat as covariates :
  - $\hat{\beta}_{\text{agecat}} = 0.226$ (HR=1.25)
  - $\hat{\beta}_{\text{gender}} = 1.120$ (HR=3.06)
  - Inverse of the observed information matrix :

$$I^{-1}(\hat{\beta}) = \begin{bmatrix} 0.4645 & 0.1476 \\ 0.1476 & 0.4638 \end{bmatrix}$$

  - 95% CI for $\hat{\beta}_{\text{agecat}}$: [-1.11, 1.56]
    95% CI for HR of old vs. young : [0.33, 4.77]
  - 95% CI for $\hat{\beta}_{\text{gender}}$: [-0.21, 2.45]
    95% CI for HR of female vs. male : [0.81, 11.64]

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Survival function estimation in the semiparametric model

- Survival function for subject with covariate x_i :

$$\begin{aligned} S_i(t) &= \exp(-H_i(t)) \\ &= \exp(-H_0(t)\exp(\beta^t x_i)) \\ &= (S_0(t))^{\exp(\beta^t x_i)} \end{aligned}$$

  with $S_0(t) = \exp(-H_0(t))$ and $H_0(t) = \int_0^t h_0(s)ds$

- Estimate the baseline cumulative hazard $H_0(t)$ by

$$\hat{H}_0(t) = \sum_{j:Y_{(j)} \le t} \frac{d_{(j)}}{\sum_{k \in R(Y_{(j)})} \exp\left(x_k^t \hat{\beta}\right)},$$

- Define

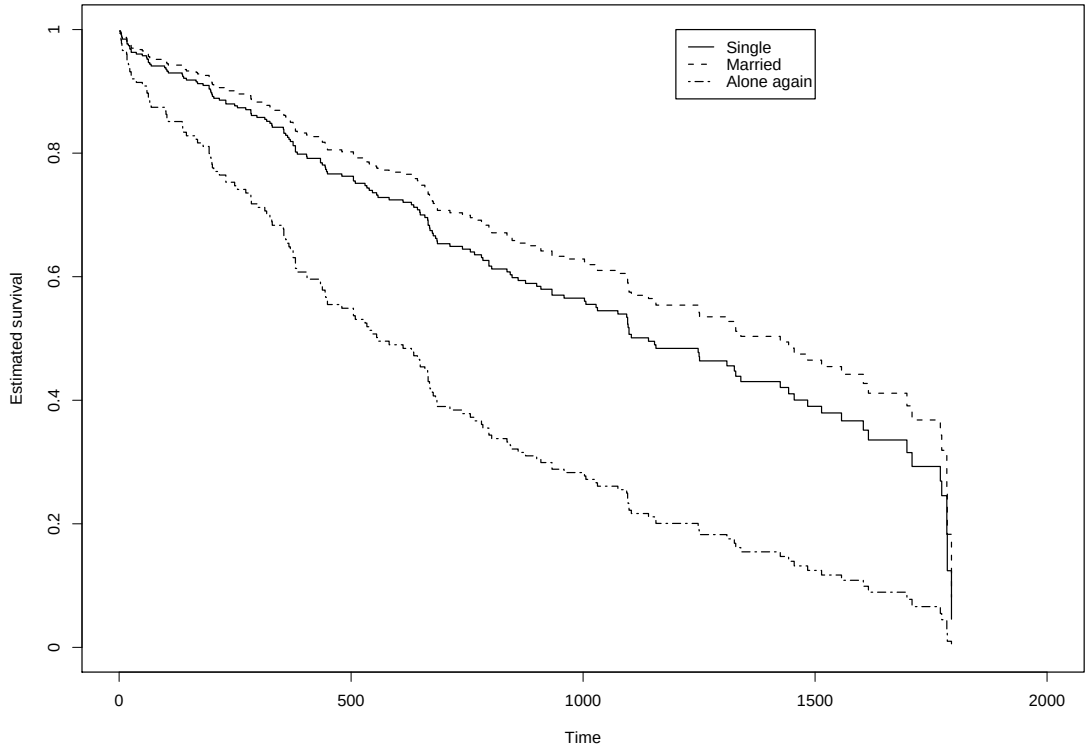
$$\hat{S}_i(t) = \left(\hat{S}_0(t)\right)^{\exp(\hat{\beta}^t x_i)},$$

  with $\hat{S}_0(t) = \exp(-\hat{H}_0(t))$

- It can be shown that the estimator is asymptotically normal

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Example : Survival function estimates for marital status groups in the schizophrenic patients data



Consider e.g. survival at 505 days :

Single group :	0.755	95% CI : [0.690, 0.827]
Married group :	0.796	95% CI : [0.730, 0.867]
Alone again group :	0.537	95% CI : [0.453, 0.636]

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

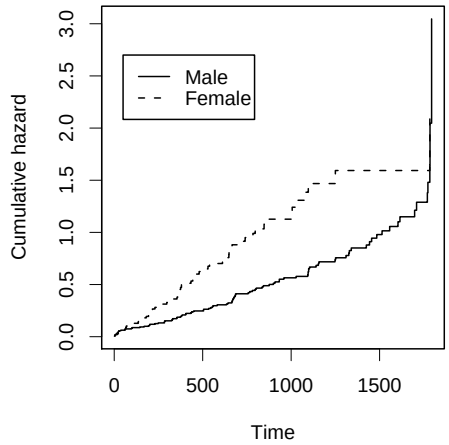
Checking the proportional hazards assumption

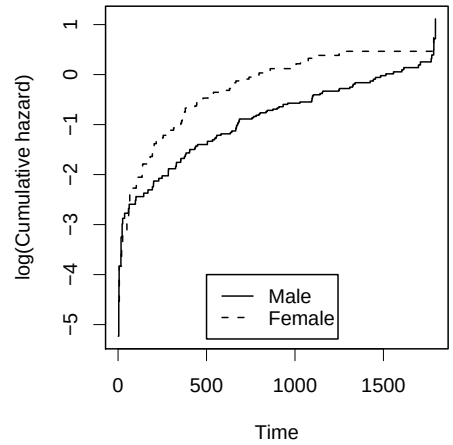
- PH assumption : hazard ratio between two subjects with different covariates is constant over time
- Diagnostic plots :
  - Consider for simplicity the case of a covariate with r levels
  - Estimate the cumulative hazard function for each level of the covariate by means of the Nelson-Aalen estimator $\Rightarrow \hat{H}_1(t), \hat{H}_2(t), \ldots, \hat{H}_r(t)$ should be constant multiples of each other :

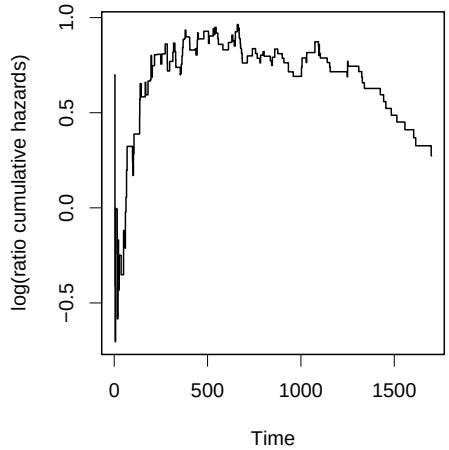
Plot	PH assumption holds if
$\log(\hat{H}_1(t)), ..., \log(\hat{H}_r(t))$ vs t	parallel curves
$\log(\hat{H}_j(t)) - \log(\hat{H}_1(t))$ vs t	constant lines
$\hat{H}_j(t)$ vs $\hat{H}_1(t)$	straight lines through origin

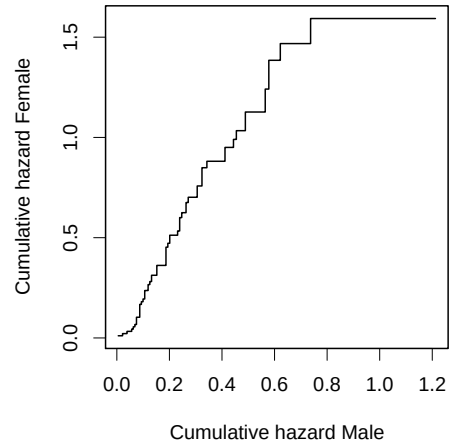
Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Example :









Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Parametric survival models

Some common parametric distributions

- Exponential distribution : $S_0(t) = \exp(-\lambda t)$
- Weibull distribution : $S_0(t) = \exp(-\lambda t^{\rho})$
- Log-logistic distribution : $S_0(t) = \dfrac{1}{1 + (t\lambda)^{\kappa}}$
- Log-normal distribution : $S_0(t) = 1 - F_N\left(\dfrac{\log(t) - \mu}{\sqrt{\gamma}}\right)$

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Parametric survival models

The parametric models considered here have two representations :

◇ Accelerated failure time model (AFT) :

$$S_i(t) = S_0(\exp(\theta^T x_i)t),$$

where

- $\theta = (\theta_1, \ldots, \theta_p)^T =$ vector of regression coefficients
- $\exp(\theta^T x_i) =$ acceleration factor
- S_0 belongs to a parametric family of distributions

Hence,

$$h_i(t) = \exp\left(\theta^T x_i\right) h_0\left(\exp(\theta^T x_i)t\right)$$

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

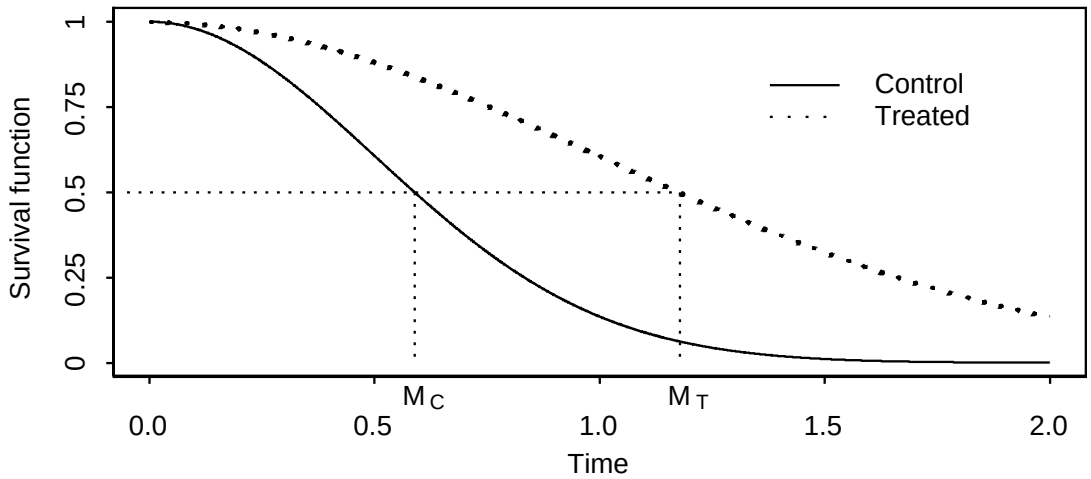
and

$$M_i = \exp(-\theta^T x_i) M_0$$

where M_i = median of S_i, since

$$S_0(M_0) = \frac{1}{2} = S_i(M_i) = S_0\big(\exp(\theta^T x_i) M_i\big)$$

Ex : For one binary variable (say treatment (T) and control (C)), we have $M_T = \exp(-\theta) M_C$:



Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

◇ Linear model :

$$\log T = \mu + \gamma^T x + \sigma W,$$

where

- $\mu =$ intercept
- $\gamma = (\gamma_1, \ldots, \gamma_p)^T =$ vector of regression coefficients
- $\sigma =$ scale parameter
- W has known distribution, that is
  - independent of x (random design)
  - the same for all x (fixed design)

  and the mean and variance of W are fixed to identify the model

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

- These two models are equivalent, if we choose
  - $S_0 =$ survival function of $\exp(\mu + \sigma W)$
  - $\theta = -\gamma$

  Indeed,

$$\begin{aligned} S_i(t) &= P(T_i > t) \\ &= P(\log T_i > \log t) \\ &= P(\mu + \sigma W_i > \log t - \gamma^t x_i) \\ &= S_0\big(\exp(\log t - \gamma^t x_i)\big) \\ &= S_0\big(t \exp(\theta^t x_i)\big) \end{aligned}$$

  $\Rightarrow$ The two models are equivalent

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Special case : the Weibull distribution

- Consider the accelerated failure time model

$$S_i(t) = S_0\big(\exp(\theta^t x_i)t\big),$$

where $S_0(t) = \exp(-\lambda t^\alpha)$ is Weibull

$$\Rightarrow S_i(t) = \exp\big(-\lambda \exp(\beta^t x_i) t^\alpha\big) \text{ with } \beta = \alpha\theta$$

$$\Rightarrow f_i(t) = \lambda\alpha t^{\alpha-1} \exp(\beta^t x_i) \exp\big(-\lambda \exp(\beta^t x_i) t^\alpha\big)$$

$$\Rightarrow h_i(t) = \alpha\lambda t^{\alpha-1} \exp(\beta^t x_i) = h_0(t) \exp(\beta^t x_i),$$

with $h_0(t) = \alpha\lambda t^{\alpha-1}$ the hazard of a Weibull

$\Rightarrow$ We also have a Cox PH model

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

- The above model is also equivalent to the following linear model :

$$\log T = \mu + \gamma^t x + \sigma W,$$

where W has a standard extreme value distribution, i.e. $S_W(w) = \exp(-e^w)$. Indeed,

$$\begin{aligned} P(W > w) &= P\big(\exp(\mu + \sigma W) > \exp(\mu + \sigma w)\big) \\ &= S_0\big(\exp(\mu + \sigma w)\big) \\ &= \exp\big(-\lambda \exp(\alpha\mu + \alpha\sigma w)\big) \end{aligned}$$

Since W has a known distribution, we fix $\lambda \exp(\alpha\mu) = 1$ and $\alpha\sigma = 1$ (identifiability constraint), and hence

$$P(W > w) = \exp(-e^w)$$

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

- It follows that

  Weibull accelerated failure time model

  = Cox PH model with Weibull baseline hazard

  = Linear model with standard extreme value error distribution

  and

  - $\theta = -\gamma = \beta/\alpha$
  - $\alpha = 1/\sigma$
  - $\lambda = \exp(-\mu/\sigma)$

- Note that the Weibull distribution is the only continuous distribution that can be written as an AFT model and as a PH model

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Estimation

- It suffices to estimate the model parameters in one of the equivalent model representations. Consider e.g. the linear model :

$$\log T = \mu + \gamma^T x + \sigma W$$

- The likelihood function for right censored data equals

$$\begin{aligned} L(\mu, \gamma, \sigma) &= \prod_{i=1}^{n} f_i(Y_i)^{\Delta_i} S_i(Y_i)^{1-\Delta_i} \\ &= \prod_{i=1}^{n} \left[\frac{1}{\sigma Y_i} f_W\left(\frac{\log Y_i - \mu - \gamma^T x_i}{\sigma} \right) \right]^{\Delta_i} \\ &\quad \times \left[S_W\left(\frac{\log Y_i - \mu - \gamma^T x_i}{\sigma} \right) \right]^{1-\Delta_i} \end{aligned}$$

Since W has a known distribution, this likelihood can be maximized w.r.t. its parameters μ, γ, σ

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

- Let

$$(\hat{\mu}, \hat{\gamma}, \hat{\sigma}) = \text{argmax}_{\mu,\gamma,\sigma} L(\mu, \gamma, \sigma)$$

- It can be shown that
  - $(\hat{\mu}, \hat{\gamma}, \hat{\sigma})$ is asymptotically unbiased and normal
  - The estimators of the accelerated failure time model (or any other equivalent model) and their asymptotic distribution can be obtained from the Delta-method

Basic concepts

Cure models

Introduction

Ongoing research

Dependent censoring

Introduction

Ongoing research

Measurement errors

Introduction

Ongoing research

Part II : Cure models

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Introduction to cure models

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Introduction

- In classical survival models, we assume that all individuals will experience the event of interest, so

$$\lim_{t \to \infty} S(t) = 0$$

where

$$S(t) = P(T > t)$$

and T is the time until the event of interest occurs.

- This assumption is realistic when studying e.g.
  - Time to death (all causes confounded)
  - Time to failure of a machine
  - Time to retirement
  - ...

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

- However, in many situations, a fraction of the population will never experience the event of interest :
  - Medicine : time until recurrence of a certain disease
  - Economics : time to find a new job after a period of unemployment
  - Demography : time to a second child after a first one
  - Finance : time until a bank goes bankrupt
  - Marketing : time until someone buys a new product
  - Sociology : time until a re-arrest for released prisoners
  - Education : time taken to solve a problem
  - ...

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

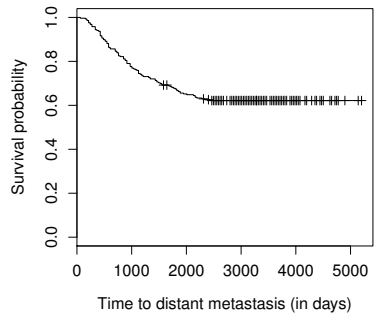
- Two groups of individuals :
  - Cured individuals
  - Susceptible individuals
- The survival function is not proper :

$$\lim_{t \to \infty} S(t) > 0$$

- Cure rate = probability of being cured :

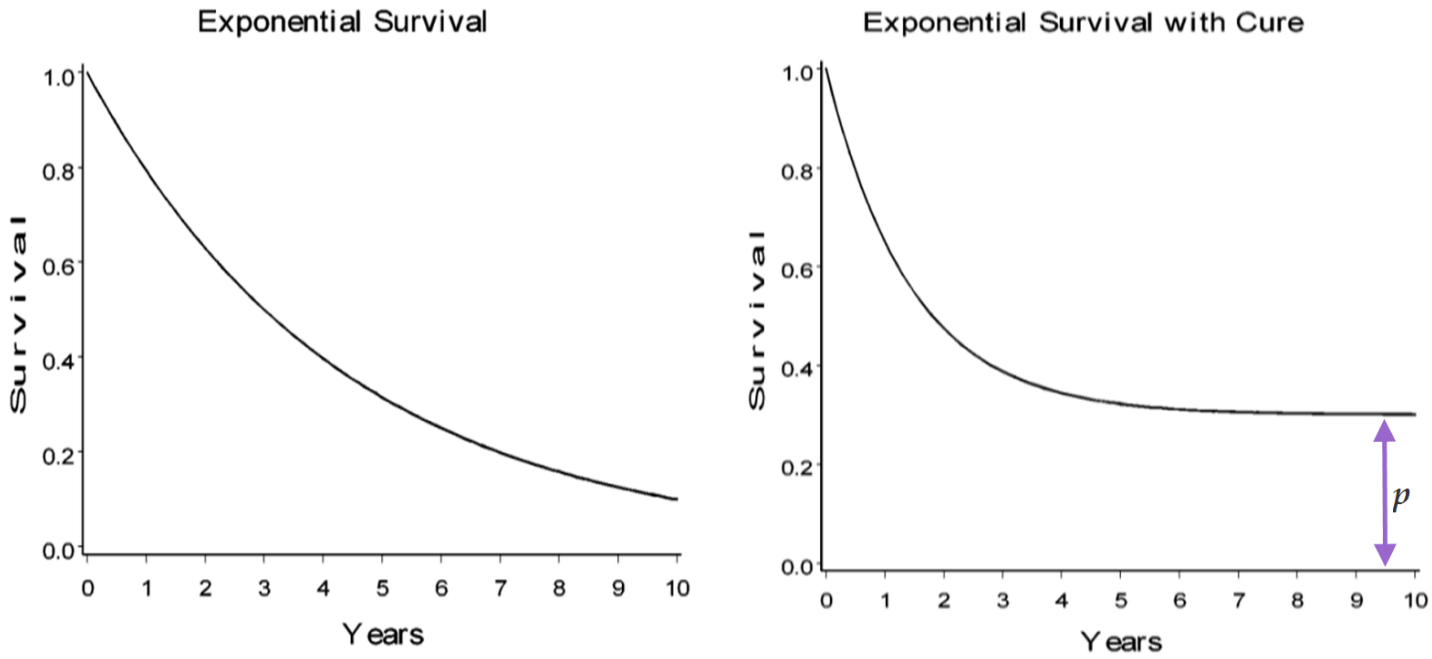
$$1 - p = \lim_{t \to \infty} S(t)$$

- Example : Kaplan-Meier plot of time to distant metastasis for breast cancer patients :



$\Rightarrow$ Height of the plateau corresponds to $1 - p$

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research



Example of exponential model with cure
(where height of the plateau = cure rate = $1 - p$)

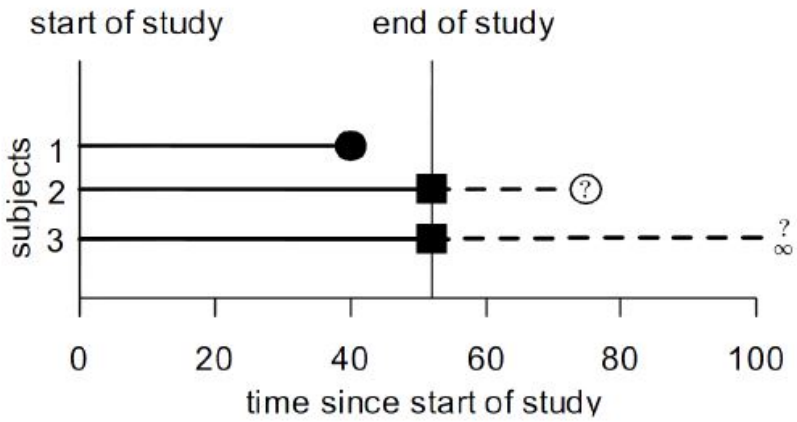
Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

- The binary variable

$$B = I(T < \infty)$$

indicating if someone is cured or not, is latent

- The observable variables are still Y and Δ as before, but
  - when $\Delta = 1$, the individual is susceptible
  - when $\Delta = 0$, we don't know whether he is susceptible or cured



Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

- Cure models are also called
  - 'split population models' in economics
  - 'limited-failure population life models' in engineering
- How can we know that we need to use a cure model if we cannot distinguish cured observations from censored uncured observations ?
  - Informal: 'if we have a long plateau that contains a large number of data points, we can be confident that (almost) all observations in the plateau correspond to cured observations'
  - Context of the study

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Is a cure model identified ?

Or : how can we know whether a censored observation in the right tail is cured or not cured ?
Let

$$\begin{aligned} S(t) &= P(T > t|B = 0)P(B = 0) + P(T > t|B = 1)P(B = 1) \\ &= 1 - p + pS_u(t), \end{aligned}$$

where $S_u(t) = P(T > t|B = 1)$ is the (proper) survival function of the susceptibles.
Let $F_u = 1 - S_u$
$G =$ the censoring distribution
τ_F is the right endpoint of the support of F (for any F)
If

$$\tau_{F_u} \leq \tau_G,$$

then the model is identified !

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Cure regression models

Two main families exist :

- Mixture cure models :

$$S(t \mid x, z) = p(z) S_u(t \mid x) + 1 - p(z),$$

where

- X and Z are two vectors of covariates
- $p(z) = P(B = 1 \mid Z = z)$ is the probability of being susceptible (incidence part)
- $S_u(t \mid x) = P(T > t \mid X = x, B = 1)$ is the (proper) conditional survival function of the susceptibles (latency part)

$\rightarrow$ the cure rate is $1 - p(z)$

The model has been proposed by Boag (1949), Berkson and Gage (1952), Farewell (1982)

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

- ◇ Promotion time cure models (also called bounded cumulative hazard models or PH cure models) :

$$S(t \mid x) = \exp\{-\theta(x)F(t)\},$$

where

- X is the complete vector of covariates
- $\theta(x)$ captures the effect of the covariates x on the survival function $S(t \mid x)$

$\rightarrow$ proportional hazards structure

$\rightarrow$ the cure rate is $P(B = 0 \mid X = x) = \exp\{-\theta(x)\}$

The model has been proposed by Yakovlev et al (1996)

There also exist models that unify the mixture and the promotion time cure model into one over-arching model

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Is it important to account for cure?

Simulate data from a mixture cure model

$$S(t \mid x, z) = p(z)S_u(t \mid x) + 1 - p(z)$$

with

- Incidence : logistic regression model with $Z = (1, Z_1, Z_2)^T$, average cure proportion of 32%
- Latency : exponential model with covariate $X = Z$
- Censoring times follow an exponential distribution, average censoring rate of 34%
- $n = 300$
- For each dataset, we fit
  - a Cox PH model
  - a mixture cure model

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

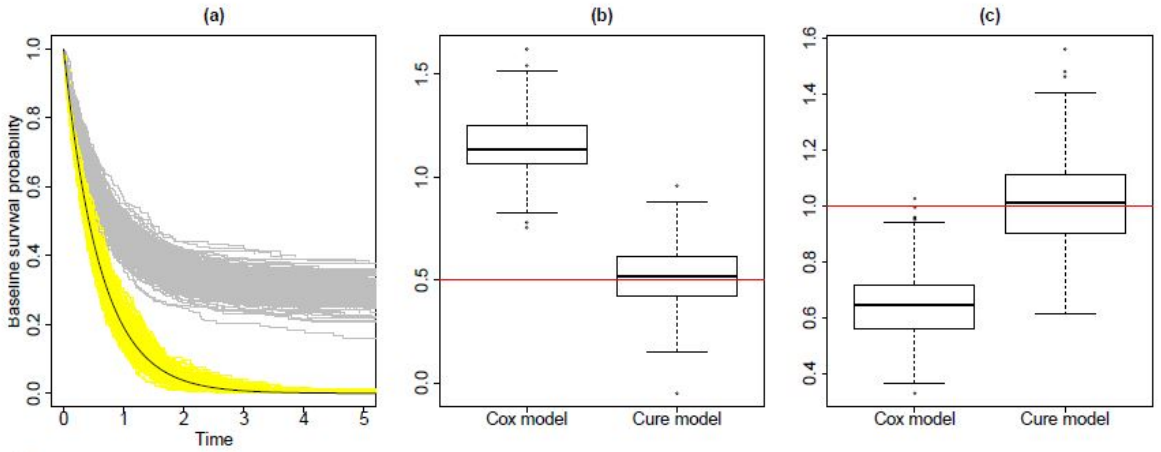


Figure 2

(a) True baseline survival function (black curve) and estimated baseline survival function over 250 datasets from the classical Cox PH model (grey curves) and the mixture cure model (yellow curves) - (b) Boxplots of $\hat{\beta}_1$ (red line : β_1) - (c) Boxplots of $\hat{\beta}_2$ (red line : β_2)

$\Rightarrow$ Not taking into account the presence of a cure fraction in survival data has important consequences that may lead to wrong conclusions

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Examples

Example 1 : Breast cancer data

- Time to distant metastasis (in days)
- 286 patients with a lymph-node-negative breast cancer
- Covariates :
  - Age : range = [26-83], median = 52
  - Estrogen receptor status : 0 = ER- (77 pts), 1 = ER+ (209 pts)
  - Size of the tumor : range = [1-4], median = 1
  - Menopausal status : 0 = premenopausal (129 pts), 1 = postmenopausal (157 pts)

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

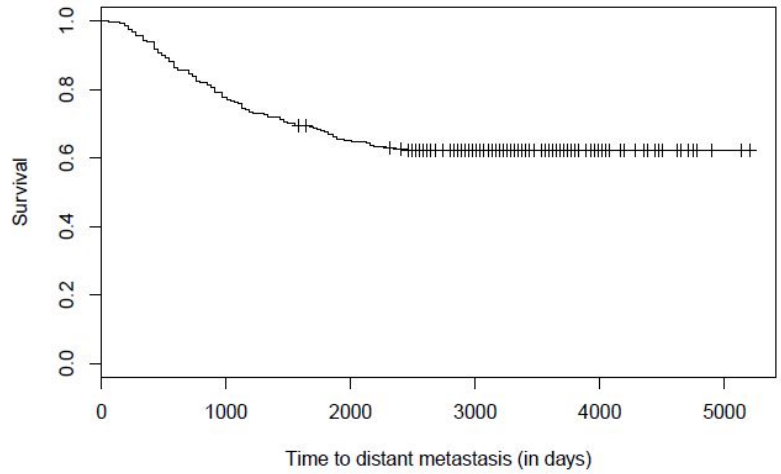


Figure 3

Kaplan & Meier (1958) estimator for the data from Wang et al. (2005) (+ : censored observations)

$\rightarrow$ 179 patients are right-censored, among which 88.3% are censored after the last observed event time

$\rightarrow$ strong medical evidence for a fraction of cure in breast cancer relapse

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Example 2 : Personal loan data

- Data from a U.K. financial institution
- Data used in Stepanova and Thomas (2002), Tong et al. (2012)
- Application information for 7521 loans
- Default observed for 376 out of 7521 observations (5%)

Var number	Description	Type
v1	The gender of the customer (1=M, 0=F)	categorical
v2	Amount of the loan	continuous
v3	Number of years at current address	continuous
v4	Number of years at current employer	continuous
v5	Amount of insurance premium	continuous
v6	Homephone or not (1=N, 0=Y)	categorical
v7	Own house or not (1=N, 0=Y)	categorical
v8	Frequency of payment (1=low/unknown, 0=high)	categorical

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Note that

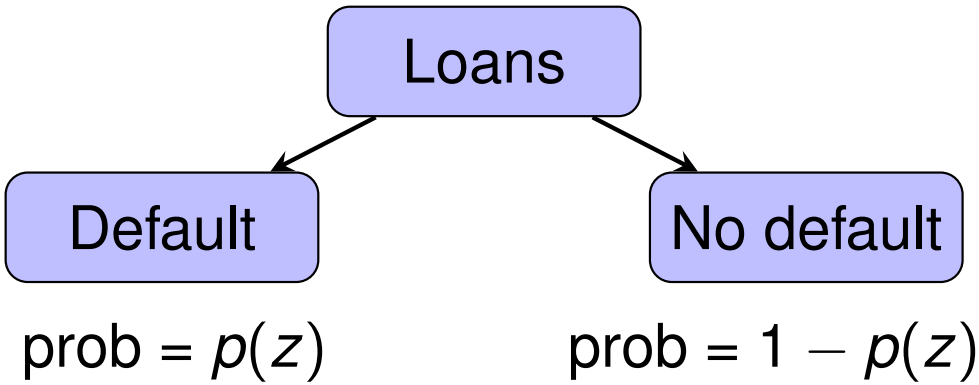
- heavy right censoring
- default will not/never take place for a large part of the population

$\Rightarrow \lim_{t\to\infty} S(t) \neq 0$

$\Rightarrow$ we use a mixture cure model

$\Delta_i = 1 \Rightarrow$ the individual is susceptible

$\Delta_i = 0 \Rightarrow$ we do not know whether default will ever take place or not



Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Mixture cure models

Recall the model :

$$S(t \mid x, z) = p(z)S_u(t \mid x) + 1 - p(z)$$

Incidence :

- models the probability of being susceptible $p(z) = P(B = 1 \mid Z = z)$
- Most often logistic regression model :

$$p(z) = \frac{\exp(z^T\alpha)}{1 + \exp(z^T\alpha)}$$

Latency :

- models the conditional survival function of the susceptibles $S_u(t \mid x) = P(T > t \mid X = x, B = 1)$
  - parametric model
  - Cox PH model
  - AFT model, ...

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Fully parametric model

Ex: Logistic/Weibull model (Farewell, 1982)

- Conditional survival function of the uncured :

$$S_u(t \mid x) = \exp(-(\lambda e^{\beta^T x})t^{\rho})$$

  with $\lambda > 0$ the shape parameter and $\rho > 0$ the scale parameter.

- Maximum likelihood estimation :
  - Numerical optimization, e.g., Newton-Raphson
  - Variance of the estimators via the inverse of the observed information matrix

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Logistic / Cox PH model

- Conditional survival function of the uncured:

$$S_u(t \mid x) = S_u(t)^{\exp(x^T \beta)}$$

  with the baseline survival function $S_u(t)$ left unspecified.
- The PH assumption remains valid for the susceptibles but is not valid anymore at the level of the population
  $\Rightarrow$ Partial likelihood approach developed for the Cox PH model can not be used
- Several approaches have been proposed :
  - Approaches based on the marginal likelihood
  - Approaches based on the EM algorithm

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Other mixture cure models

- Logistic / semi parametric AFT models
- Other link functions in the incidence: probit, complementary log-log link function
- Flexible semiparametric models, e.g.
  - Cox model for latency and single-index structure in the incidence : $p(z) = g(\gamma^T z)$ where $g(\cdot)$ is unspecified
  - Logistic regression for incidence and non-parametric model in the latency
- Non-parametric mixture cure models

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Promotion time cure models

- Also called bounded cumulative hazard model or PH cure model
- Introduced by Yakovlev et al (1996) and formally proposed by Tsodikov (1998)
- Idea : since, in the presence of cure, the survival function is improper, the idea is to 'bound' the cumulative hazard function

$$H(t) = \theta F(t)$$

with $F(\cdot)$ a proper distribution function and $\theta > 0$
In this way

$$\lim_{t \to \infty} H(t) = \theta$$

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

- If θ depends on covariates, the (improper) survival function is then given by

$$S(t \mid x) = \exp\{-\theta(x)F(t)\}$$

where

  - X is the complete vector of covariates (with an intercept)
  - $\theta(x)$ captures the effect of the covariates x on the survival function $S(t \mid x)$

- This formulation has a proportional hazards structure
- This model has a specific biological interpretation (leading to the name 'promotion time model')
- Usually, $\theta(x) = \exp(\beta^T x)$, and F is unspecified
- The cure rate is $1 - \exp\{-\theta(x)\}$

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

References

- Book :
  - Maller and Zhou (1996)
- Review papers :
  - Peng and Taylor (2014)
  - Amico and VK (2018)

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

The focused information criterion for a mixture cure model

(joint with Gerda Claeskens)

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Proportional hazards mixture cure model

We consider the model

$$S(t \mid x, z) = p(z) S_u(t \mid x) + 1 - p(z)$$

where

- survival function : proportional hazards model, i.e.

$$\begin{aligned} S_u(t|x) &= S_u(t)^{\exp(x^T\beta)} \\ &= \exp\Big(- \exp(x^T\beta) H_u(t) \Big) \end{aligned}$$

  where $S_u(\cdot)$ and $H_u(\cdot)$ are the baseline survival and baseline cumulative hazard function of the susceptibles

- cure rate : logistic model, i.e.

$$p(z) = \frac{\exp(z^T\alpha)}{1 + \exp(z^T\alpha)} \quad \text{or} \quad \log\left(\frac{p(z)}{1 - p(z)}\right) = z^T\alpha$$

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

The data consist of iid vectors $(X_i, Z_i, Y_i, \Delta_i)$, $i = 1, \ldots, n$, with

$$Y_i = \min(T_i, C_i), \quad \Delta_i = I(T_i \leq C_i),$$

and C_i is independent of T_i given (X_i, Z_i).

Maximum likelihood estimation :

The likelihood under the PH mixture cure model is given by

$$L_n(\alpha, \beta, H) = \prod_{i=1}^{n} \left[\left\{ \pi(Z_i^T \alpha) H\{Y_i\} e^{X_i^T \beta} e^{-H(Y_i) \exp(X_i^T \beta)} \right\}^{\Delta_i} \times \left\{ 1 - \pi(Z_i^T \alpha) + \pi(Z_i^T \alpha) e^{-H(Y_i) \exp(X_i^T \beta)} \right\}^{1-\Delta_i} \right],$$

where $\pi(t) = \exp(t)/[1 + \exp(t)]$.

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Define

$$(\widehat{\alpha}, \widehat{\beta}, \widehat{H}_u) = \text{argmax}_{\alpha,\beta,H}\, L_n(\alpha, \beta, H).$$

Asymptotic properties of $(\widehat{\alpha}, \widehat{\beta}, \widehat{H}_u)$ have been established by Fang, Li and Sun (2005) and Lu (2007) :

$$n^{1/2}\big(\widehat{H}_u(\cdot) - H_u(\cdot)\big) \Rightarrow \text{ Gaussian process}$$

and

$$n^{1/2}\big(\widehat{\alpha} - \alpha, \widehat{\beta} - \beta\big) \xrightarrow{d} \text{Multivariate normal}$$

(for the case where the model is correctly specified)

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Variable selection in a mixture cure model

The parameters in the model are

- α : for logistic model on cure rate $\pi(\cdot)$
- $\beta, H_u(\cdot)$: for Cox PH model on survival function $S_u(\cdot|\cdot)$

Suppose we are interested in a certain quantity

$$\mu = \mu(\alpha, \beta, H_u(\cdot)),$$

which we call the focus.

Of interest : Variable selection in order to estimate as well as possible (in MSE sense) the focus μ.

Literature on variable selection for mixture cure models :

- Scolas et al (2016) (using Lasso)
- Dirick et al (2015) (using AIC)

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Examples :

- Personalized prediction of the (unconditional) survival of a given patient (or for given values of x and z) :

$$S(t|x,z) = p(z)S_u(t|x) + 1 - p(z)$$

- Personalized prediction of the (unconditional) risk :

$$h(t|x,z) = \frac{p(z)f_u(t|x)}{p(z)S_u(t|x) + 1 - p(z)}$$

- Mean or median survival time for given values of x and z (conditional or unconditional)
- Probability of being cured for given z : $p(z)$

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

How to do variable selection ?

Note that

- Incorporating the full vectors x and z will lead to a full model with a large variance but a smaller bias as compared to a narrow model that leaves out all components of x and z, resulting in a large bias but a smaller variance.
- One could construct intermediate model selection scenarios where some of the components of x and z are protected (i.e. forced to be present in all models). The unprotected variables take part in the model selection step.

For simplicity, we ignore this division and assume that all components of x and z are unprotected.

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Focused Information Criterion (FIC)

General idea : 'best' model depends on the focus and is selected by minimizing the *MSE* of the estimator of the focus.

References : Claeskens and Hjort (2008), Cambridge.

Some notation : In each submodel we estimate the focus μ by maximixing the semiparametric likelihood introduced before, and we define

$$\widehat{\mu}_{S_1,S_2} = \mu(\widehat{\alpha}_{S_1,S_2}, \widehat{\beta}_{S_1,S_2}, \widehat{H}_{uS_1,S_2}(\cdot)),$$

where S_1 is the subset of $\{1,\ldots,p\}$ (logistic) and S_2 is the subset of $\{1,\ldots,q\}$ (Cox PH) that indicates which components of x and z are present in the considered model.

Define

$$(\widehat{S}_1, \widehat{S}_2) = \text{argmin}_{S_1,S_2}\, FIC(S_1, S_2) = \text{argmin}_{S_1,S_2}\, \widehat{MSE}(\widehat{\mu}_{S_1,S_2})$$

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

In order to be able to calculate the MSE of each submodel, we need to make an assumption regarding the true model.

We work with local misspecification :

- The true hazard rate is

$$H_{u,\text{true}}(t|x) = H_{0u}(t) \exp\left(x^T(\beta_0 + b/\sqrt{n})\right),$$

- The true logistic model is

$$\text{logit}\{p_{\text{true}}(z)\} = z^T\left(\alpha_0 + a/\sqrt{n}\right),$$

where α_0 and β_0 are known, and a and b do not depend on the sample size n.

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Asymptotic theory

Define

$$\langle \widehat{H}_u - H_{0u},\ \widehat{\alpha} - \alpha_0,\ \widehat{\beta} - \beta_0 \rangle (g)$$
$$= \int_0^\tau g_1(t)\, d(\widehat{H}_u - H_{0u})(t) \ +\ g_2^T(\widehat{\alpha} - \alpha_0,\ \widehat{\beta} - \beta_0),$$

where $g = (g_1(\cdot), g_2)$.

Note that

- If $g = (0, e_k)$, then
  $\langle \widehat{H}_u - H_{0u},\ \widehat{\alpha} - \alpha_0,\ \widehat{\beta} - \beta_0 \rangle (g)$
  = k-th component of $(\widehat{\alpha} - \alpha_0, \widehat{\beta} - \beta_0)$
- If $g = (I(\cdot \le t), 0)$, then
  $\langle \widehat{H}_u - H_{0u},\ \widehat{\alpha} - \alpha_0,\ \widehat{\beta} - \beta_0 \rangle (g) = \widehat{H}_u(t) - H_{0u}(t)$

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Note that

$$U\Big(H_{0u},\ \alpha_0 + \frac{a}{\sqrt{n}},\ \beta_0 + \frac{b}{\sqrt{n}}\Big) = 0 \quad \text{and} \quad U_n(\widehat{H}_u, \widehat{\alpha}, \widehat{\beta}) = 0,$$

where

$$\begin{aligned} U_n(\Gamma)(g) &= U_n(H_u, \alpha, \beta)(g) \\ &= U_{n1}(\Gamma)(g_1) + U_{n2}(\Gamma)(g_2) \\ &= \text{score operator} \end{aligned}$$

and

$$U(\Gamma) = EU_n(\Gamma),$$

where the expected value is with respect to the true model.

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

For any submodel (S_1, S_2),

$$n^{1/2}\langle \widehat{H}_{uS_1,S_2} - H_{0u},\ \widehat{\alpha}_{S_1,S_2} - \alpha_0,\ \widehat{\beta}_{S_1,S_2} - \beta_0 \rangle$$

converges weakly to a Gaussian process G with covariance function

$$\begin{aligned}\mathrm{Cov}\big(G(g), G(\tilde{g})\big) = \int_0^\tau & \sigma_1\big(\sigma^{-1}_{S_1,S_2}(g), 0\big)(t)\, \sigma^{-1}_{S_1,S_2(1)}(\tilde{g})(t)\, dH_0(t) \\ & + \big(\sigma^{-1}_{S_1,S_2(2)}(\tilde{g}), 0\big)^T \sigma_2\big(\sigma^{-1}_{S_1,S_2}(g), 0\big),\end{aligned}$$

and with mean function

$$E\big(G(g)\big) = B_1\big(\sigma^{-1}_{S_1,S_2(1)}(g)\big) + B_2\big(\sigma^{-1}_{S_1,S_2(2)}(g), 0\big).$$

Note that

- If $g = (0, e_k)$, we get the asymptotic normality of the k-th component of $n^{1/2}(\widehat{\alpha}_{S_1,S_2} - \alpha_0,\ \widehat{\beta}_{S_1,S_2} - \beta_0)$
- If $g = (I(\cdot \le t), 0)$, we get the asymptotic normality of $n^{1/2}(\widehat{H}_{uS_1,S_2}(t) - H_{0u}(t))$

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Hence,

$$n^{1/2}\left(\widehat{\mu}_{S_1,S_2} - \mu_0\right) \xrightarrow{d} N\left(Bias(\mu, S_1, S_2, a, b), Var(\mu, S_1, S_2)\right).$$

Estimation of $Bias(\mu, S_1, S_2, a, b)$ and $Var(\mu, S_1, S_2)$:

- ◇ Variance : plug-in estimation of the asymptotic variance
- ◇ Bias : based on $\widehat{a} = n^{1/2}\widehat{\alpha}_{Full}$ and $\widehat{b} = n^{1/2}\widehat{\beta}_{Full}$

Hence,

$$FIC(S_1, S_2) = \widehat{MSE}(\widehat{\mu}_{S_1,S_2}).$$

This result can now be used to select the best model for μ by minimizing $FIC(S_1, S_2)$ over all possible submodels.

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Simulations

Only preliminary simulation results ...

Consider the following Cox/logistic cure model :

$$S(t|x,z) = p(z)S_u(t|x) + 1 - p(z)$$

where

- $X, Z \sim \text{Unif}[-1,1]$
- $p(z) = \frac{\exp(\alpha_0+\alpha_1 z)}{1+\exp(\alpha_0+\alpha_1 z)}$, with $\alpha_0 = \alpha_1 = 2$
- $S_u(t|x) = [\exp(-1.65t)]^{\exp(\beta_1 x)}$, with $\beta_1 = 2$
- $C \sim \text{Exp}(mean = 1.7)$

Then,

$$\% \text{ cure} = 0.2 \quad \text{and} \quad \% \text{ censoring} = 0.4$$

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Focus parameters :

$$\mu_j = H_{0u}(t)$$

for $t = 1^{st}, 2^{nd}$ or 3^{rd} quartile of baseline cumulative survival function ($j = 1, 2, 3$)

9 candidate models :

	logistic		Cox		Estimated MSE ($\times 10^3$)			True MSE ($\times 10^3$)		
	X	Z	X	Z	μ_1	μ_2	μ_3	μ_1	μ_2	μ_3
1	1	1	1	1	1.62	7.30	29.9	1.56	6.45	36.0
2	1	1	1	0	1.57	6.73	26.1	1.57	6.13	33.6
3	1	1	0	1	22.2	20.2	37.4	25.0	27.5	56.7
4	1	0	1	1	1.72	8.31	36.8	1.78	8.16	50.2
5	1	0	1	0	1.55	6.66	25.9	1.63	6.59	35.9
6	1	0	0	1	15.5	9.50	68.8	17.6	14.2	83.5
7	0	1	1	1	1.50	6.62	26.9	1.42	5.80	34.7
8	0	1	1	0	1.47	6.26	24.3	1.43	5.54	32.4
9	0	1	0	1	12.4	5.46	100.1	14.1	10.9	124.2

The true model is model 8.

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

	logistic		Cox		FIC model selection prob.		
	X	Z	X	Z	μ_1	μ_2	μ_3
1	1	1	1	1	0.08	0.01	0.02
2	1	1	1	0	0.07	0.02	0.10
3	1	1	0	1	0.00	0.05	0.11
4	1	0	1	1	0.01	0.00	0.00
5	1	0	1	0	0.18	0.09	0.20
6	1	0	0	1	0.00	0.19	0.12
7	0	1	1	1	0.20	0.05	0.07
8	0	1	1	0	0.46	0.20	0.33
9	0	1	0	1	0.00	0.39	0.05

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Data analysis

Personal loan data :

- Data from a U.K. financial institution
- Data used in Stepanova and Thomas (2002), Tong et al. (2012)
- Application information for 7521 loans
- Default observed for 376 out of 7521 observations (5%)

Var number	Description	Type
v1	The gender of the customer (1=M, 0=F)	categorical
v2	Amount of the loan	continuous
v3	Number of years at current address	continuous
v4	Number of years at current employer	continuous
v5	Amount of insurance premium	continuous
v6	Homephone or not (1=N, 0=Y)	categorical
v7	Own house or not (1=N, 0=Y)	categorical
v8	Frequency of payment (1=low/unknown, 0=high)	categorical

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Note that

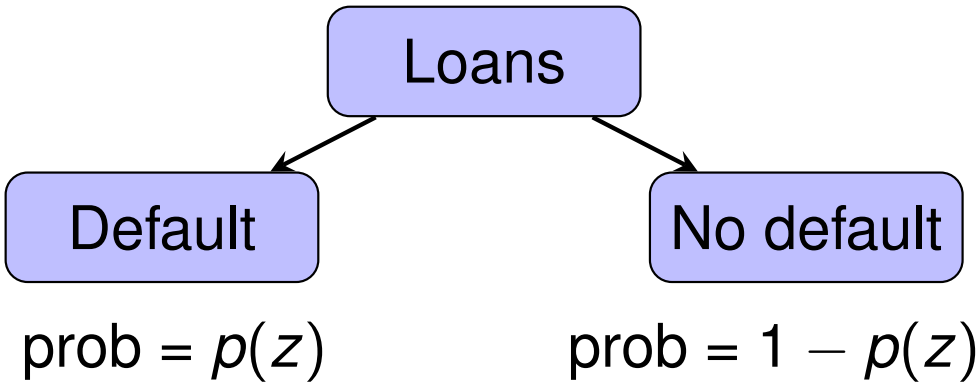
- heavy right censoring
- default will not/never take place for a large part of the population

$\Rightarrow \lim_{t\to\infty} S(t) \neq 0$

$\Rightarrow$ we use a mixture cure model

$\Delta_i = 1 \Rightarrow$ the individual is susceptible

$\Delta_i = 0 \Rightarrow$ we do not know whether default will ever take place or not



Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

- 7521 observations and 8 variables
- Default observed for 376 out of 7521 observations
- 2 covariate vectors (α and β), empty models excluded : $(2^8 - 1) \times (2^8 - 1) = 65025$ FICs to calculate !
- Focus : probability of cure $1 - p(z)$ at $z = \text{median}(Z)$

Part	v1	v2	v3	v4	v5	v6	v7	v8
Cure rate	1	1	1	1	1	0	0	1
Survival of uncured	1	0	1	0	1	1	1	1

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Conclusions

- We considered a proportional hazards mixture cure model, and developed the asymptotic distribution of the estimators of the model components under local misspecification of the model.
- This asymptotic distribution can then be used to select the best variables to estimate a certain quantity (focus) in the model via FIC minimization.

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Part III : Dependent censoring

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Introduction to dependent censoring

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Introduction

- Random right censoring assumes that the survival time (T) and the censoring time (C) are independent
- We observe

$$Y = \min(T, C) \text{ and } \Delta = I(T \leq C),$$

so we observe either T or C, but not both

$\Rightarrow$ Relation between T and C not identifiable in general

$\Rightarrow$ Relation between T and C needs to be specified in order to identify the model

$\Rightarrow$ Independence assumption is most natural assumption, and holds true in many contexts

(See Tsiatis, 1975)

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Independence of T and C is satisfied if

- Administrative censoring : individuals alive at the end of the study are censored
  $\Rightarrow$ Censoring is unrelated to survival time
  $\Rightarrow$ Independence assumption makes sense
- Censoring happens for other reasons that are completely unrelated to the event of interest
  Eg. In medical studies, patients might move, die because of car accident, etc.
- Many other contexts

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Independence of T and C might be doubtful if

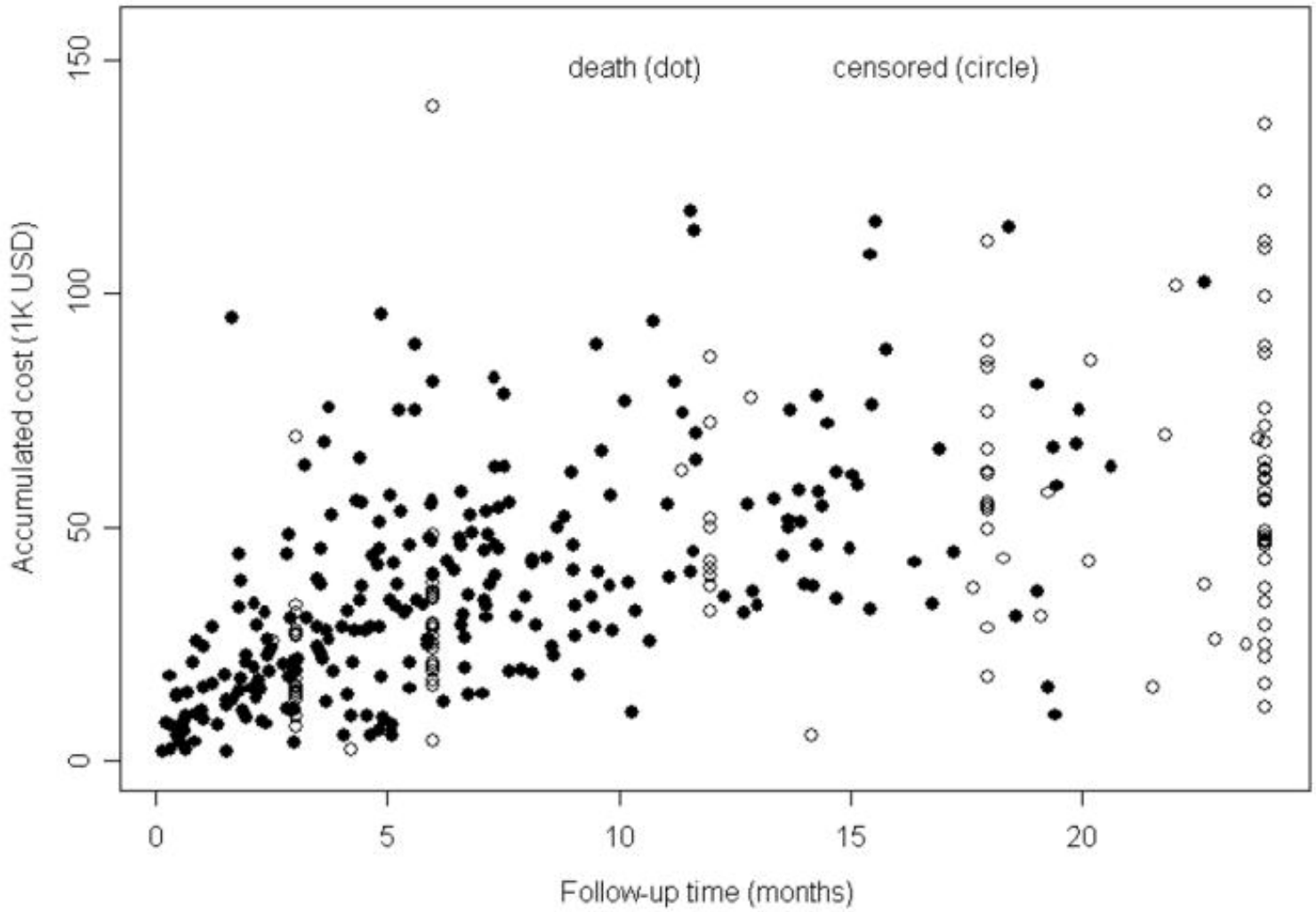
- ◇ Medical studies : Patients may withdraw from the study
  - because their condition is deteriorating or because they are showing side effects which need alternative treatments (positive relation between T and C)
  - because their health condition has improved and so they no longer follow the treatment (negative relation between T and C)
- ◇ Unemployment studies : Unemployed people with low chances on the job market could decide to go abroad to improve their chances, leading to censoring times that depend on the duration of unemployment

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

- ◇ Transplantation studies : Often the length of time a patient has to wait before he gets transplanted (C) depends on his/her medical condition, so on his time to death (T)
- ◇ Health economics :
  - Let U be the medical cost, then
    $$U = A(T)$$
    for some increasing function A
  - Suppose that the cost accumulation rate is constant over time, but the rate may vary from individual to individual :
    $$A(T) = RT,$$
    where R is the cost accumulation rate
  - If T is censored by C, then U is censored by $A(C) = RC$, and so we observe $\min(RT, RC)$
  - Clearly, RT and RC are dependent

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Example of accumulated medical cost data :



Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Note that

- ◇ The independence between T and C can not be tested in practice !
- ◇ It needs to be motivated based on the context of the study
- ◇ Standard methods may lead to wrong or biased inference

$\Rightarrow$ It is important to propose a model under which the dependence between T and C can be identified, and which is flexible enough to cover a wide range of situations

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

What happens if independence is assumed when T and C are in reality correlated ?

Consider

$$(\log T, \log C) \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right),$$

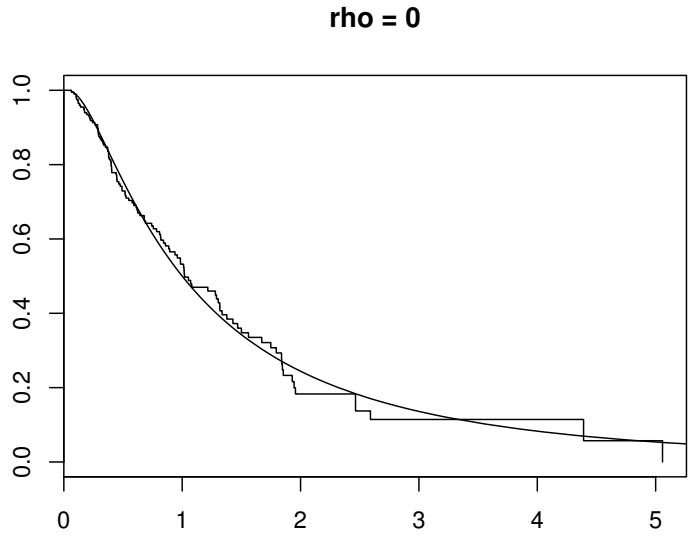
where $\rho = 0, \pm 0.3, \pm 0.6$ or ± 0.9

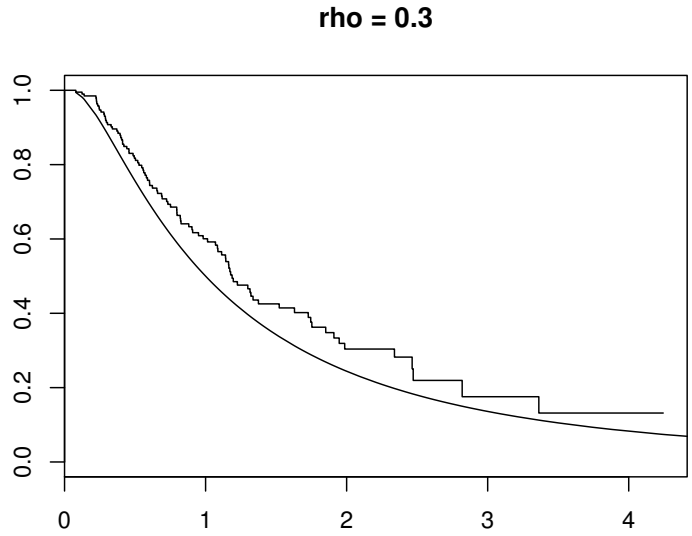
Further, let $Y = \min(T, C)$ and $\Delta = I(T \leq C)$

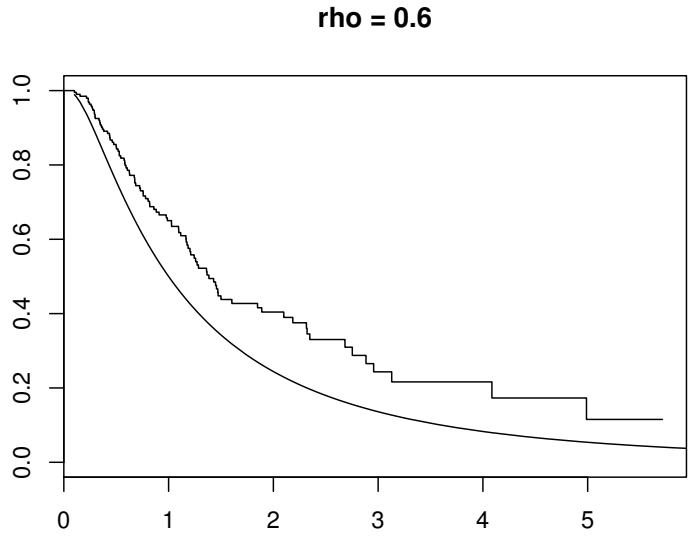
For an arbitrary sample of size $n = 200$, we calculate

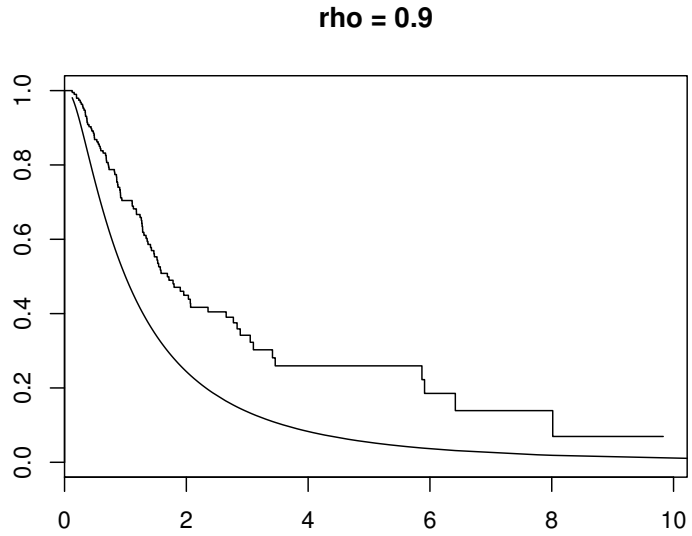
- the true survival function $S(t)$ of $T \sim \exp(N(0, 1))$
- the Kaplan-Meier estimator $\hat{S}(t)$ (which assumes $T \perp\!\!\!\perp C$)

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research



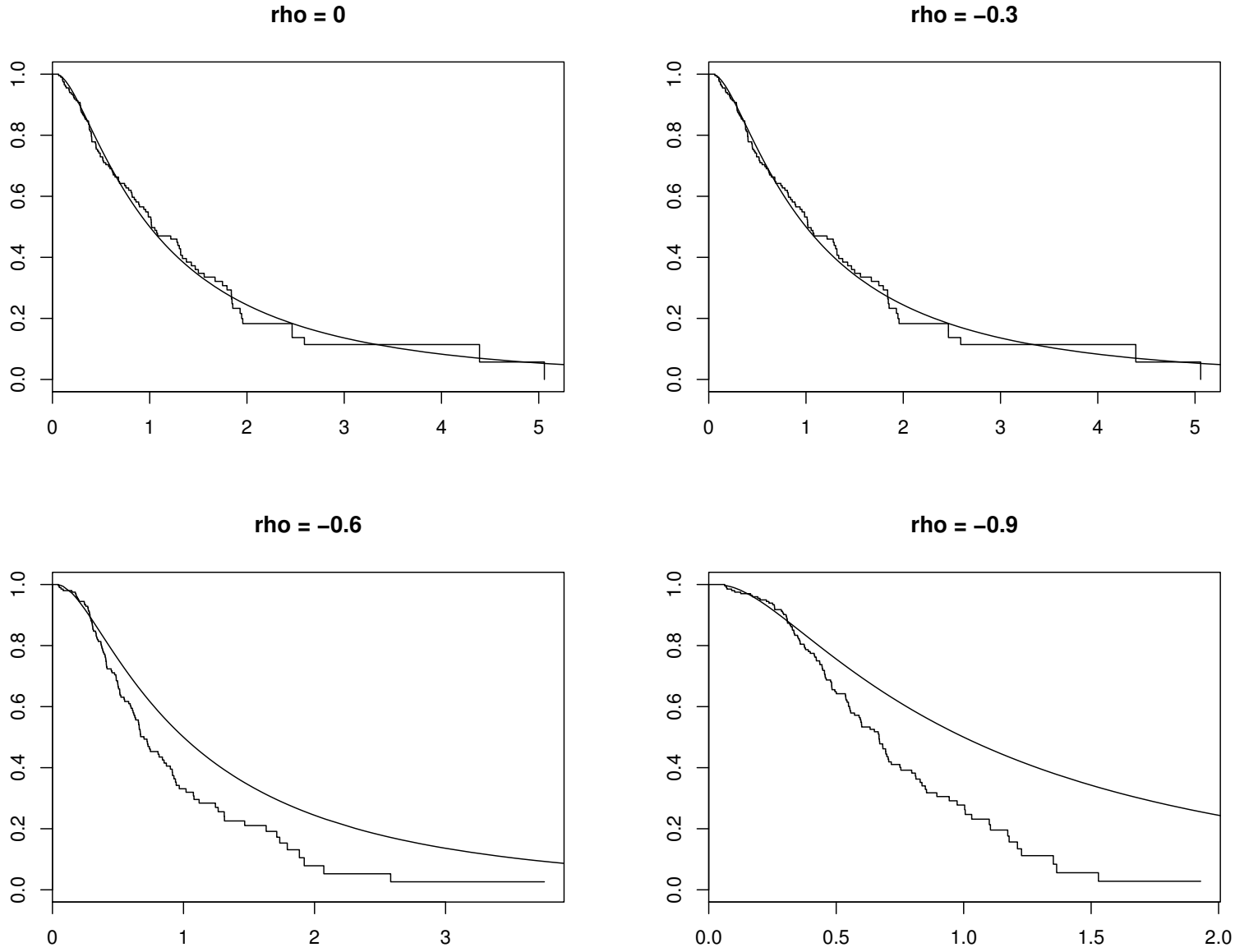






$\Rightarrow$ The larger ρ, the more the Kaplan-Meier estimator lies above the true survival function

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research



$\Rightarrow$ The smaller ρ, the more the Kaplan-Meier estimator lies below the true survival function

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Example : liver transplant data

- See Collett, 2015
- 281 patients were registered for a liver transplant
- 75 patients died while waiting for a transplant
- $T =$ time to death while waiting for a liver transplant
- $C =$ time at which the patient receives a transplant
- Livers were given on the basis of patient's health condition
- Patients who get a transplant tend to be those who are closer to death $\Rightarrow$ dependent censoring

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

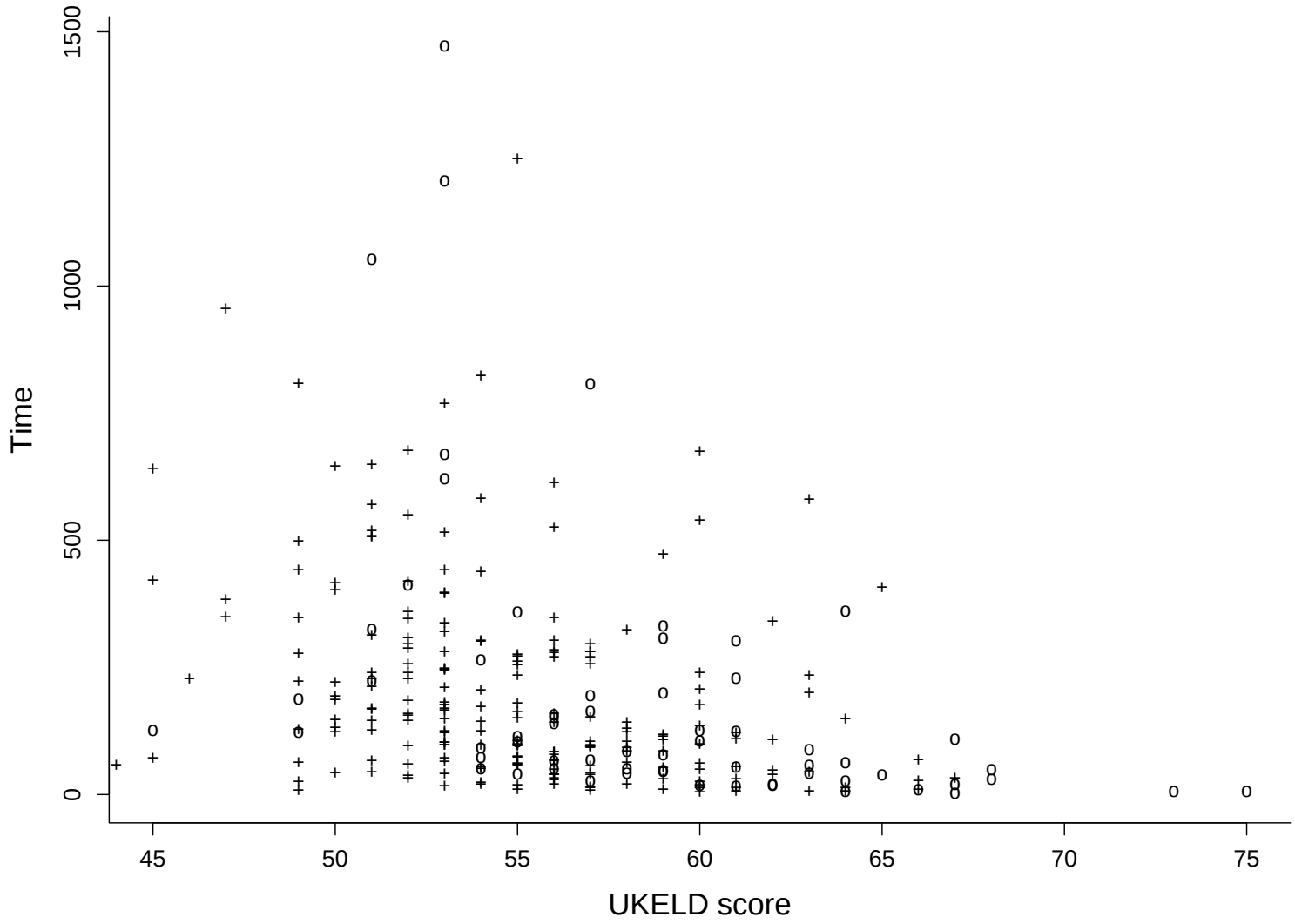
- Covariates :
  - Age of the patients in years (X_1)
  - Gender (1 = male, 0 = female) (X_2)
  - Body mass index (BMI) in kg/m^2 (X_3)
  - UKELD score: UK end-stage liver disease score (X_4)
- We could model these data using eg. an accelerated failure time model

$$\log T = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon,$$

and estimate the β's using one of the classical methods
But : the estimated coefficients will be biased
$\Rightarrow$ We need other methods that take dependent censoring into account

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Scatter plot of the survival time versus the UKELD score :



Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Existing approaches to take dependent censoring into account

Without covariates :

- ◇ **Bounds on marginal distribution** :
  Slud and Rubinstein (1983) studied bounds for the marginal survival function, rather than exact estimators
- ◇ **Copula approach** :
  - Zheng and Klein (1995) : modelling of the bivariate distribution of T and C by means of a known copula function, and estimation of the marginal distribution of T nonparametrically under this copula model
  - Rivest and Wells (2001) : special case of Archimedean copulas

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

With covariates :

- **Copula approach** : extension of the Zheng and Klein's method to the Cox model (Huang and Zhang, 2008)
- **Inverse probability of censoring weighted (IPCW) method** : the weights are derived from a Cox model for the censoring time (Collett, 2015)
- **Multiple imputation method** : the censored failure times are imputed under departures from independent censoring within the Cox model (Jackson et al., 2014)
- **Auxiliary information** : adjust for dependent censoring in the estimation of the marginal survival function (Scharfstein and Robins, 2002, Hsu et al, 2015)
- **Accumulated medical cost data** : specific methods exist for this particular type of dependent censoring (Lin et al, 1997, among others).

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Flexible parametric model for survival data subject to dependent censoring

(joint with Negera Wakgari Deresa)

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Proposed model

Objective : To propose a flexible parametric model that allows for dependence between T and C

Model :

$$\begin{cases} \Lambda_\theta(T) = X^T\beta + \epsilon_T \\ \Lambda_\theta(C) = W^T\eta + \epsilon_C, \end{cases}$$

where

- $T = \log(\text{survival time})$, $C = \log(\text{censoring time})$
- Λ_θ is a parametric family of monotone transformations
- $\begin{pmatrix} \epsilon_T \\ \epsilon_C \end{pmatrix} \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_T^2 & \rho\sigma_T\sigma_C \\ \rho\sigma_T\sigma_C & \sigma_C^2 \end{pmatrix}\right)$
- $X = (1, \tilde{X}^T)^T$ and $W = (1, \tilde{W}^T)^T$
- $(\epsilon_T, \epsilon_C) \perp\!\!\!\perp (X, W)$

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Note that

$$\text{Corr}\big(\Lambda_\theta(T), \Lambda_\theta(C)|X, W\big) = \rho \in [-1, 1]$$

$\Rightarrow$ The model allows T and C to be dependent (given X and W) !

We will show that the ρ-parameter is identified.

We will work with the Yeo and Johnson (2000) family of transformations :

$$\Lambda_\theta(t) = \begin{cases} \{(t+1)^\theta - 1\}/\theta & t \geq 0, \theta \neq 0 \\ \log(t+1) & t \geq 0, \theta = 0 \\ -\{(-t+1)^{2-\theta} - 1\}/(2-\theta) & t < 0, \theta \neq 2 \\ -\log(-t+1) & t < 0, \theta = 2 \end{cases}$$

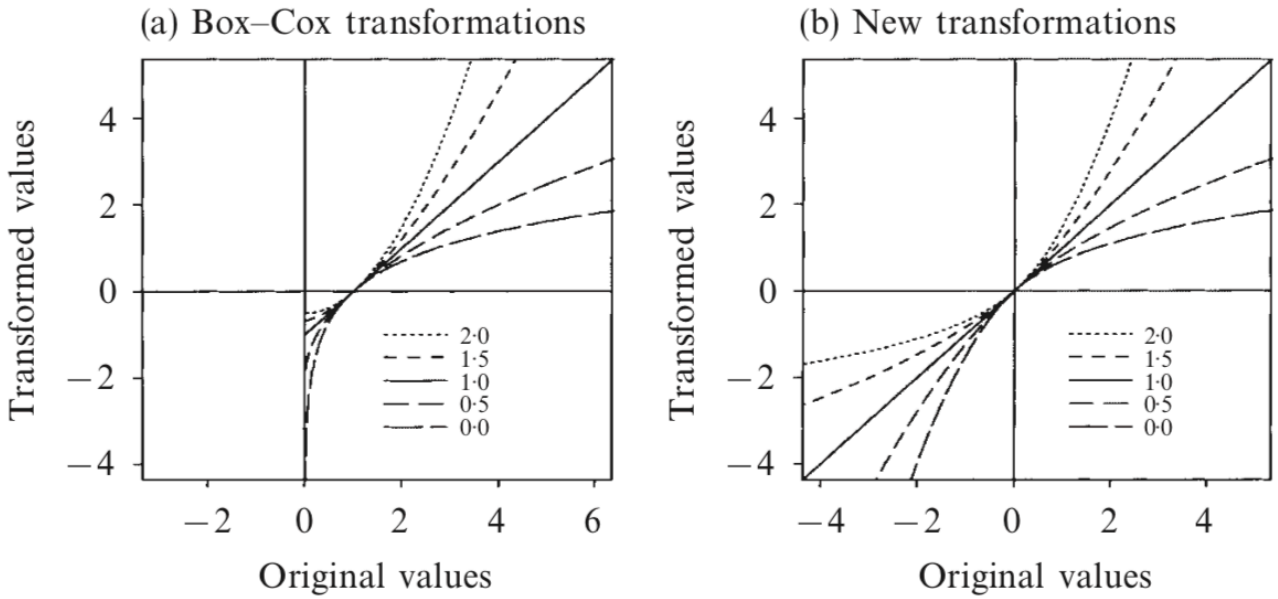
Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Why the Yeo-Johnson transformation ?

- It generalizes the well-known Box-Cox transformation to the whole real line :

  - Box-Cox(θ) maps $\mathbb{R}^+$ to $(-1/\theta, \infty)$
  - Yeo-Johnson(θ) maps $\mathbb{R}$ to $\mathbb{R}$ for $0 \leq \theta \leq 2$

- $\theta = 1 : \Lambda_\theta(T) = T =$ log(survival time)
  $1 < \theta \leq 2 : \Lambda_\theta(T)$ is convex and lies above T
  $0 \leq \theta < 1 : \Lambda_\theta(T)$ is concave and lies below T



Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Identifiability and estimation

Theorem (Identifiability of the model)

Suppose that $Var(\tilde{X})$ and $Var(\tilde{W})$ have full rank.

Then, the proposed model is identifiable.

This means that if for $j = 1, 2$, the pair (T_j, C_j) satisfies the proposed model with parameters

$$\alpha_j = (\theta_j, \beta_j, \eta_j, \sigma_{T_j}, \sigma_{C_j}, \rho_j),$$

and if $Y_j = \min(T_j, C_j)$ and $\Delta_j = I(T_j \leq C_j)$, then

$$f_{Y_1,\Delta_1|X,W}(\cdot,\cdot \mid x, w; \alpha_1) \equiv f_{Y_2,\Delta_2|X,W}(\cdot,\cdot \mid x, w; \alpha_2)$$

for almost every (x, w), implies that $\alpha_1 = \alpha_2$, i.e.

$$\theta_1 = \theta_2, \beta_1 = \beta_2, \eta_1 = \eta_2, \sigma_{T_1} = \sigma_{T_2}, \sigma_{C_1} = \sigma_{C_2}, \rho_1 = \rho_2$$

The proof is based on Basu and Ghosh (1978).

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Estimation

- The data consist of i.i.d. replications $(Y_i, \Delta_i, X_i, W_i)$, $i = 1, \ldots, n$ of (Y, Δ, X, W)
- The model parameters are estimated by maximizing the likelihood function
- The likelihood function is given by

$$
\begin{aligned}
&L(\alpha) \\
&= \prod_{i=1}^{n} \Bigg[\frac{1}{\sigma_T} \Big\{ 1 - \Phi\Big(\frac{\Lambda_\theta(Y_i) - W_i^T \eta - \rho \frac{\sigma_C}{\sigma_T} (\Lambda_\theta(Y_i) - X_i^T \beta)}{\sigma_C (1-\rho^2)^{1/2}} \Big) \Big\} \\
&\qquad \times \phi\Big(\frac{\Lambda_\theta(Y_i) - X_i^T \beta}{\sigma_T} \Big) \Lambda'_\theta(Y_i) \Bigg]^{\Delta_i} \\
&\quad \times \Bigg[\frac{1}{\sigma_C} \Big\{ 1 - \Phi\Big(\frac{\Lambda_\theta(Y_i) - X_i^T \beta - \rho \frac{\sigma_T}{\sigma_C} (\Lambda_\theta(Y_i) - W_i^T \eta)}{\sigma_T (1-\rho^2)^{1/2}} \Big) \Big\} \\
&\qquad \times \phi\Big(\frac{\Lambda_\theta(Y_i) - W_i^T \eta}{\sigma_C} \Big) \Lambda'_\theta(Y_i) \Bigg]^{1-\Delta_i}
\end{aligned}
$$

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

- Note that the likelihood can not be factorized in a part only depending on the parameters of T and another part only depending on the parameters of C
- The only exception is when $\rho = 0$, in which case the likelihood reduces to the usual normal likelihood under independent censoring
- Define

$$\hat{\alpha} = (\hat{\theta}, \hat{\beta}, \hat{\eta}, \hat{\sigma}_T, \hat{\sigma}_C, \hat{\rho}) = \text{argmax}_{\alpha \in A} L(\alpha)$$

where

$$A = \{(\theta, \beta, \eta, \sigma_T, \sigma_C, \rho) : 0 \leq \theta \leq 2, \beta \in \mathbb{R}^p, \eta \in \mathbb{R}^q, \sigma_T > 0, \sigma_C > 0, -1 < \rho < 1\}$$

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Asymptotic theory

We assume that our model is potentially misspecified. Let $\alpha^* = (\theta^*, \beta^*, \eta^*, \sigma_T^*, \sigma_C^*, \rho^*)$ be the parameter vector that minimizes the Kullback-Leibler Information Criterion (KLIC), given by

$$E\left[\log\left\{\frac{f_{Y,\Delta|X,W}(Y,\Delta \mid X,W)}{f_{Y,\Delta|X,W}(Y,\Delta \mid X,W;\alpha)}\right\}\right],$$

where the expectation is taken with respect to the true density $f_{Y,\Delta,X,W}$.

Theorem (Consistency)

Under regularity conditions $(A1)$ to $(A3)$ in White (1982),

$$(\hat{\theta}, \hat{\beta}, \hat{\eta}, \hat{\sigma}_T, \hat{\sigma}_C, \hat{\rho}) \xrightarrow{P} (\theta^*, \beta^*, \eta^*, \sigma_T^*, \sigma_C^*, \rho^*) \quad \text{as } n \to \infty$$

If the model is correctly specified the KLIC attains its unique minimum at $\alpha^* = \alpha$

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Theorem (Asymptotic normality)

Under regularity conditions $(A1)$ to $(A6)$ in White (1982),

$$n^{1/2}\Big((\hat{\theta},\hat{\beta},\hat{\eta},\hat{\sigma}_T,\hat{\sigma}_C,\hat{\rho}) - (\theta^*,\beta^*,\eta^*,\sigma_T^*,\sigma_C^*,\rho^*)\Big) \xrightarrow{d} N(0,V),$$

where $V = A(\alpha^*)^{-1}B(\alpha^*)A(\alpha^*)^{-1}$, with

$$A(\alpha) = \left(E\left\{\frac{\partial^2}{\partial\alpha_i\partial\alpha_j}\log f_{Y,\Delta|X,W}(Y,\Delta \mid X,W;\alpha)\right\}\right)_{i,j=1}^{p+q+4},$$

$$B(\alpha) = \left(E\left\{\frac{\partial}{\partial\alpha_i}\log f_{Y,\Delta|X,W}(Y,\Delta \mid X,W;\alpha) \times \frac{\partial}{\partial\alpha_j}\log f_{Y,\Delta|X,W}(Y,\Delta \mid X,W;\alpha)\right\}\right)_{i,j=1}^{p+q+4}$$

If the model is correctly specified, $V = A(\alpha)^{-1}$, the inverse of Fisher's information matrix.

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Simulations

- The model :

$$\begin{cases} \Lambda_\theta(T) = 2 + 1.2X_1 + 1.5X_2 + \epsilon_T \\ \Lambda_\theta(C) = 2.5 + 0.5X_1 + X_2 + \epsilon_C, \end{cases}$$

  where

  - $X_1 \sim \text{Bern}(0.5)$ and $X_2 \sim U[-1, 1]$
  - $\theta = 0$ or 1.5

- Setting 1 : $(\epsilon_T, \epsilon_C) \sim N_2(\mu, \Sigma)$

  with $\mu = (0, 0)$ and $(\sigma_T, \sigma_C, \rho)$ = (1, 1.5, 0.75)

- Setting 2 : $(\epsilon_T, \epsilon_C) \sim t_v(\mu, \Sigma)$, $v = 15$

- The censoring rate is approximately 45% under both settings

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Setting 1 : $(\epsilon_T, \epsilon_C) \sim N_2(\mu, \Sigma)$, $n = 300$

θ	0			1.5		
Par.	Bias	RMSE	CR	Bias	RMSE	CR
	Dependent censoring model					
β_0	-0.001	0.144	0.947	-0.019	0.169	0.948
β_1	-0.008	0.182	0.944	-0.022	0.188	0.940
β_2	-0.004	0.169	0.940	-0.022	0.183	0.931
σ_1	0.004	0.097	0.944	-0.007	0.109	0.940
ρ	-0.028	0.208	0.956	-0.031	0.215	0.954
θ	-0.002	0.031	0.949	-0.019	0.101	0.944
	Independent censoring model					
β_0	0.207	0.249	0.709	0.156	0.234	0.880
β_1	0.201	0.268	0.812	0.169	0.255	0.867
β_2	0.111	0.212	0.904	0.074	0.214	0.924
σ_1	-0.030	0.095	0.906	-0.051	0.115	0.871
θ	-0.021	0.038	0.881	-0.080	0.130	0.858

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Setting 2 : $(\epsilon_T, \epsilon_C) \sim t_{15}(\mu, \Sigma)$, $n = 300$

θ	0			1.5		
Par.	Bias	RMSE	CR	Bias	RMSE	CR
	Dependent censoring model					
β_0	0.032	0.159	0.945	0.056	0.190	0.938
β_1	0.035	0.207	0.928	0.048	0.218	0.929
β_2	0.037	0.187	0.930	0.058	0.206	0.923
σ_1	0.012	0.113	0.928	0.033	0.126	0.922
ρ	-0.045	0.240	0.938	-0.050	0.223	0.942
θ	0.004	0.034	0.917	0.028	0.103	0.902
	Independent censoring model					
β_0	0.242	0.283	0.631	0.247	0.307	0.735
β_1	0.229	0.300	0.781	0.244	0.320	0.770
β_2	0.140	0.235	0.881	0.156	0.260	0.871
σ_1	-0.025	0.106	0.907	-0.015	0.115	0.913
θ	-0.017	0.038	0.861	-0.037	0.107	0.895

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Data Application

- See Collett, 2015
- 281 patients were registered for a liver transplant
- 75 patients died while waiting for a transplant
- $T =$ time to death while waiting for a liver transplant
- $C =$ time at which the patient receives a transplant
- Livers were given on the basis of patient's health condition
- Patients who get a transplant tend to be those who are closer to death $\Rightarrow$ dependent censoring
- Covariates :
  - Age of the patients in years
  - Gender (1 = male, 0 = female)
  - Body mass index (BMI) in kg/m^2
  - UKELD score: UK end-stage liver disease score

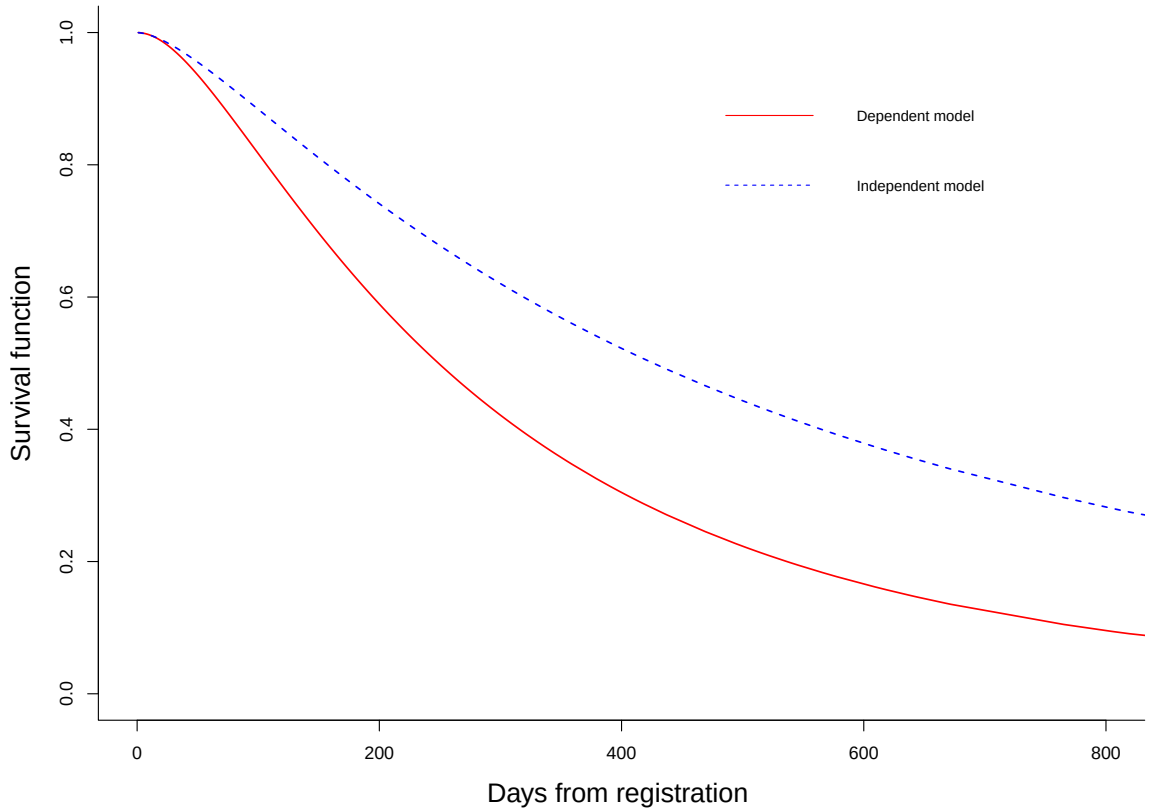
Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Parameter estimates :

	Dependent model				Independent model			
var.	Est.	SE	BSE	p-value	Est.	SE	BSE	p-value
Age	-0.165	0.096	0.108	0.084	-0.267	0.109	0.104	0.014
Gender	0.915	0.895	0.957	0.307	0.988	1.318	1.460	0.456
BMI	-0.086	0.065	0.063	0.181	-0.121	0.085	0.082	0.155
UKELD	-0.610	0.214	0.181	0.005	-0.678	0.237	0.186	0.004
θ	1.764	0.196	0.158	0.000	1.680	0.195	0.156	0.000
ρ	0.730	0.250	0.249	0.004				

- The parameter estimates are somewhat different for the two models
- The UKELD score is negatively related to the survival time
- $\hat{\rho} = 0.73 \Rightarrow$ strong correlation

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research



- Estimated survival at Age = 50, UKELD = 60, BMI = 25 and Gender = 0
- Six months survival rate : 79% under the independent model, 67% under the dependent model
- The 80% survival rate is overestimated by almost two months

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Possible extensions

- Extension to competing risks, and to regimes with independent (administrative) and dependent censoring
- Relaxing parametric assumption on transformation function :

$$\begin{cases} H(T) = X^T\beta + \epsilon_T \\ H(C) = W^T\eta + \epsilon_C, \end{cases}$$

where

  - H is an unknown monotone transformation
  - $(\epsilon_T, \epsilon_C) \sim N_2(0, \Sigma)$

- More flexible regression functions, using kernel or spline methods
- Replace bivariate normality assumption by assumption involving Gaussian or elliptical copulas

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Conclusions

- We proposed a flexible parametric model for survival data subject to dependent censoring
- The proposed model is identifiable
- Our approach allows to estimate the association between T and C
- A simulation study shows the good performance of the proposed model

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Part IV : Measurement errors

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Introduction to measurement errors

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Introduction

What is measurement error ? Some examples :

- inaccurate measurement devices (eg. scale, thermometer)
- imprecise recording (eg. self-reporting in surveys)
- temporal variation (eg. blood pressure)

Distinction should be made between

- measurement errors (continuous variables)
- misclassification (non-continuous variables)

We focus on measurement errors.

Our focus will be on regression models in which covariates are subject to measurement error.

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Taking measurement error into account is

- essential to do valid estimation and inference in these models
- not necessary for prediction

Possible causes of measurement error :

- inaccuracies due to a measuring device
- a biased attitude during data collection
- miscategorization
- high expenses of measuring process
- incomplete information because of missing observations
- ...

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Consequences when measurement error is not taken into account :

- biased model estimators : attenuation towards zero in simple linear models
- features of the data are often less obvious and power of tests is lower

Reviews on measurement error problems :

- Carroll, Ruppert, Stefanski and Crainiceanu (2006)
- Schennach (2016)
- Yan (2014) (in survival analysis)

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Example : simple linear regression

Consider

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

where $\beta_0 = 1$, $\beta_1 = 1$, $X \sim U[0, 1]$ and $\varepsilon \sim N(0, 0.25^2)$.
Instead of observing X, we observe

$$W = X + U,$$

where $U \sim N(0, 0.5^2)$ and $X \perp\!\!\!\perp U$.
For an arbitrary sample of size $n = 50$, we obtain

$$\hat{\beta}_0 = 0.98 \quad \text{and} \quad \hat{\beta}_1 = 1.04$$

when regressing Y on X, and

$$\hat{\beta}_0 = 1.26 \quad \text{and} \quad \hat{\beta}_1 = 0.53$$

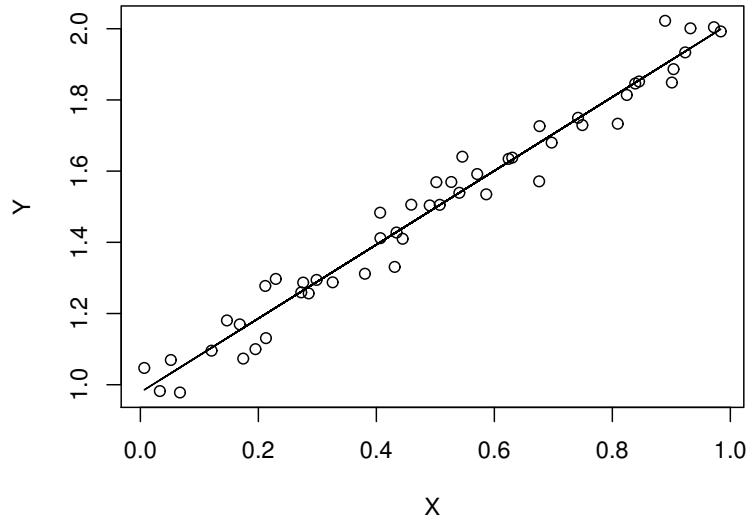
when regressing Y on W !

$\Rightarrow$ Slope is underestimated

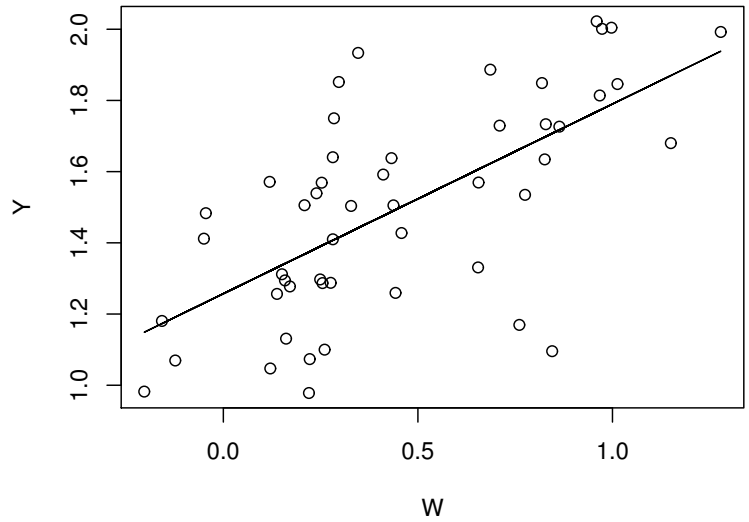
$\Rightarrow$ Effect of X on Y decreases because of measurement error on X

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

X versus Y :



W versus Y :



Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Measurement error models

Let X = true, unobservable covariate
U = error
W = observed covariate

- Classicial measurement error model :

$$W = X + U, \quad X \perp\!\!\!\perp U$$

Hence, $\text{Var}(W) = \text{Var}(X) + \text{Var}(U) > \text{Var}(X)$
Example : blood pressure : X = long-term value, W = observed value

- Berkson model :

$$X = W + U, \quad W \perp\!\!\!\perp U$$

Hence, $\text{Var}(X) > \text{Var}(W)$
Example : rounding errors

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

We will work with the classical measurement error model.

Assuming that $X \perp\!\!\!\perp U$ does not suffice to uniquely identify the distribution of X.

Example :

- Suppose $W = W_1 + W_2 + W_3$ and that $(W_1, W_2, W_3) \perp\!\!\!\perp$
- If we only know that $W = X + U$ and that $X \perp\!\!\!\perp U$, then many possibilities exist :
  - $X = W_1, U = W_2 + W_3$
  - $X = W_1 + W_2, U = W_3$
  - $X = W_2, U = W_1 + W_3$
  - many many more

$\Rightarrow$ We need to impose more assumptions to distinguish X from U

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Three options to identify the model :

- ◇ Option 1 : Additional data are available, which can be of the following form :
  - validation data (containing X for some of the observations)
  - repeated measurements of W
  - panel data (longitudinal data)
  - instrumental variables
- ◇ Option 2 : U has a completely known distribution
  Eg. $U \sim N(0, \sigma^2)$ with σ^2 known
- ◇ Option 3 : U has a partially known distribution, and some other aspects of the model are known

Note that theoretical identifiability in a measurement error context does not always result in practical identifiability.

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Let us look in more detail at option 3 :

- Make heavy independence assumptions (Reiersøl 1950 for linear regression, and Schennach and Hu 2013 for certain nonlinear models)
- Nonparametric deconvolution :
  - Matias (2002)
  - Butucea and Matias (2005), Butucea et al (2008)
  - Meister (2006, 2007)
  - Schwarz and Van Bellegem (2010)
  - Delaigle and Hall (2016)

  However these papers suffer from one or more of the following problems :
  - focus on estimation of distribution of X
  - many tuning parameters, for which no clear guidelines are given
  - no practical implementation, focus on minimax rates
  - identifiability issues (theoretical and practical)

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Existing approaches to take measurement error into account

- Method of moments (Fuller, 1987)
- Regression calibration (Carroll & Stefanski, 1990)
- Score function based approaches (Nakamura, 1990)
- Bayesian methodologies (Gustafson, 2004)
- Simulation-extrapolation (SIMEX) (Cook & Stefanski, 1994)
- Multiple imputation (Cole et al, 2006)
- ...

See Carroll et al (2006), Buonacorsi (2010) and Schennach (2016) for more details on these correction approaches

But, all methods require the error distribution to be known, unless validation or auxiliary data are available !

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Let us now focus on Simex.

Simex method :

- Simulation-Extrapolation (Simex) algorithm
- Simulation-based method for correcting the bias due to measurement error
- Consistent estimators when the true extrapolation function is used

References :

- Cook and Stefanski, 1994
- Stefanski and Cook, 1995
- Carroll, Küchenhoff, Lombard and Stefanski, 1996
- Carroll, Ruppert, Stefanski and Crainiceanu, 2006
- Many more

The distribution of U needs to be known to apply Simex, and is assumed to be $N(0, \sigma^2)$ with σ^2 known.

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Consider a regression model with regression coefficients β.

Simulation step

For $\lambda = \lambda_1, \ldots, \lambda_K$ (= increasing amounts of error, eg. $0, 0.5, 1, 1.5, 2$) :

- For $b = 1, \ldots, B$ = number of datasets to be generated:
  - Add error to the mismeasured covariates :

    $$W_{b,i}(\lambda) = W_i + \sqrt{\lambda}\,\sigma Z_{b,i}$$

    where $Z_{b,i} \sim_{iid} N(0,1)$, for $b = 1, \ldots, B$ and $\lambda = \lambda_1, \ldots, \lambda_K$.
    The variance of these contaminated data is

    $$Var(W_{b,i}(\lambda)|X_i) = (1+\lambda)\sigma^2$$

  - Estimate the parameters β of the regression model under consideration using a naive estimation procedure, i.e. a method that does not take into account the measurement error $\Rightarrow \widehat{\beta}_b(\lambda)$.
- Compute $\widehat{\beta}(\lambda) = \frac{1}{B}\sum_{b=1}^{B} \widehat{\beta}_b(\lambda)$.

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Extrapolation step

- Model the $\widehat{\beta}(\lambda)$ as a function of λ, using for example a linear, quadratic, or cubic regression.
- Extrapolate back to -1, since at this point,

$$\text{Var}(W_{b,i}(-1)) = 0$$

This leads to the following Simex estimator of β :

$$\widehat{\beta}_{Simex} = \widehat{\beta}(-1).$$

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

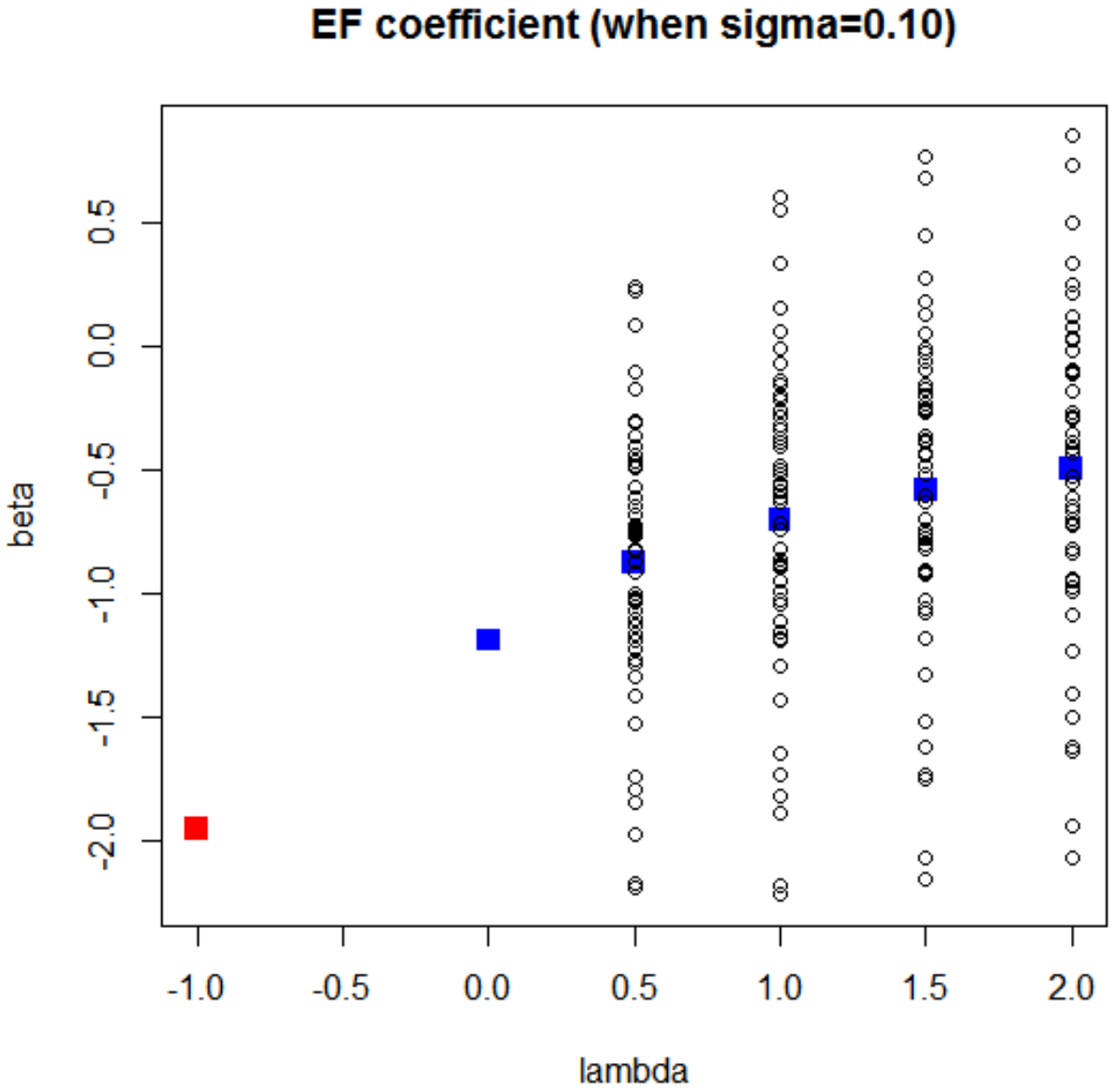


Figure: Visual representation of the SIMEX approach.

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Cox model with measurement error using Simex

Consider a Cox proportional hazards model with measurement error in the covariates :

$$h(t|x) = h_0(t)\exp(x^T\beta), \quad t \geq 0,$$

where x is a vector of covariates (without intercept)
β is a vector or regression coefficients
$h(t|x)$ is the hazard rate at time t given x
$h_0(t)$ is an unspecified baseline hazard

Prentice (1982) showed that not correcting for measurement error in the Cox model leads to biased estimators of β.

Yan (2014) gives a review of correction methods for the Cox model, including the Simex method.

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Consider the following Cox model with two covariates :

- 3 models for X_1 :
  $X_1 \sim$ 2 Beta(1,1)-1
  2 Beta(0.7,0.5)-1
  N(-0,1) truncated at $[-2, 2]$ (not. $N(0, 1, -2, 2)$)
- $U_1 \sim N(0, \sigma^2)$
- $X_2 \sim$ Bernoulli(0.5), $U_2 = 0$ (no measurement error)
- $\beta_1 = 1$, $\beta_2 = -0.5$

Instead of observing T we observe

$$Y = \min(T, C) \quad \text{and} \quad \Delta = I(T \leq C),$$

where C is the random censoring time, assumed to be independent of T given $X = (X_1, X_2)$.

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Assume the following model for T and C :

- $T|X \sim \text{Exp}\big(\mu = 0.5 \exp(-X^T\beta)\big)$ (so $h_0(t) = 2$)
- $C \perp\!\!\!\perp X$ and $C \sim \text{Exp}(\mu = 3)$

The censoring rate is $P(C < T) = 0.43$.

We compare different estimation methods :

- the 'naive' method, that does not take measurement error into account
- Simex using the true (but unknown) σ
- Simex using two misspecified values of σ (0.75σ and 1.25σ)

A quadratic extrapolation function is used in the second step of the Simex algorithm.

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Table: Simulation results for $n = 300$

			Naive (no correction)		Simex (true σ)		Simex (0.75σ)		Simex (1.25σ)	
f_X	σ		β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
2 Beta(1,1)-1	.144	Bias	-.054	-.007	.012	-.012	-.017	-.010	.050	-.015
		SD	.136	.164	.150	.165	.144	.165	.156	.166
		MSE	.021	.027	.023	.027	.021	.027	.027	.028
	.289	Bias	-.217	.001	-.036	-.012	-.114	-.007	.059	-.020
		SD	.139	.168	.182	.172	.163	.170	.199	.175
		MSE	.066	.028	.034	.030	.040	.029	.043	.031
	.433	Bias	-.387	.026	-.147	.010	-.246	.016	-.038	.003
		SD	.112	.162	.168	.173	.147	.169	.190	.179
		MSE	.162	.027	.050	.030	.082	.029	.038	.032
2 Beta(.7,.5)-1	.166	Bias	-.057	-.004	.017	-.010	-.017	-.007	.059	-.014
		SD	.131	.164	.147	.167	.140	.165	.155	.169
		MSE	.021	.027	.022	.028	.020	.027	.028	.029
	.332	Bias	-.234	.008	-.045	-.008	-.128	-.001	.050	-.017
		SD	.124	.161	.165	.168	.148	.164	.185	.172
		MSE	.070	.026	.029	.028	.038	.027	.037	.030
	.499	Bias	-.399	.028	-.156	.004	-.258	.014	-.050	-.006
		SD	.101	.157	.155	.169	.133	.163	.176	.175
		MSE	.169	.025	.048	.029	.084	.027	.033	.031
N(0,1,-2,2)	.440	Bias	-.238	.023	-.040	-.001	-.127	.009	.060	-.013
		SD	.088	.173	.123	.187	.108	.180	.139	.195
		MSE	.064	.030	.017	.035	.028	.032	.023	.038
	.880	Bias	-.557	.056	-.320	.025	-.413	.037	-.227	.012
		SD	.071	.160	.1209	.184	.103	.173	.137	.194
		MSE	.316	.029	.117	.034	.181	.031	.071	.038
	1.32	Bias	-.739	.081	-.564	.061	-.629	.068	-.505	.054
		SD	.054	.175	.097	.194	.082	.185	.110	.202
		MSE	.549	.037	.327	.041	.402	.039	.267	.044

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

The table shows that

- β_2 is not affected a lot by the measurement error in X_1 nor by the correction
- The situation is very different for β_1:
  - When the error is not taken into account, $\hat{\beta}_1$ is biased, and this bias increases with the value of σ
  - Simex always decreases this bias, but does not make it disappear, because
    - Simex assumes that $U \sim N(0, \sigma^2)$
    - Simex with a quadratic extrapolant tends to yield conservative corrections (see Carroll et al, 2006)
  - This decrease in bias comes at the cost of a higher variance, but the MSE is usually smaller with Simex than with the Naive method
  - In general :

$$\text{Bias}(\hat{\beta}_1|1.25\sigma) < \text{Bias}(\hat{\beta}_1|\sigma) < \text{Bias}(\hat{\beta}_1|0.75\sigma)$$

(due to conservative behavior of Simex method)

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Flexible parametric approach to classical measurement error variance estimation without auxiliary data

(joint with Aurélie Bertrand and Catherine Legrand)

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Assume that

- $W = X + U$ with $X \perp\!\!\!\perp U$
- U is normal and $E(U) = 0$, but $\mathrm{Var}(U)$ is unknown
- no validation or auxiliary data are available

Our goal :

- Identification and estimation of $\mathrm{Var}(U)$ (and of f_X)
- Estimation of regression models with measurement error in some of the covariates

We like to develop a stable and feasible practical method, that makes minimal model assumptions

Note that we do not assume that $\mathrm{Var}(U)$ is known, which is an important relaxation of what is commonly assumed in the literature.

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Methodology for estimating error variance

We assume that X has compact support, so it can be written as

$$X = aS + b,$$

with $a > 0$ and $b \in \mathbb{R}$ unknown
S having density f_S defined on $[0, 1]$.

Recall that we observe $W = X + U$, with $X \perp\!\!\!\perp U$ and $U \sim N(0, \sigma^2)$.

Hence, the unknown model parameters are a, b, σ^2 and f_S.

Note that the density of W is

$$f_W(w) = \frac{1}{a\sigma} \int f_S\Big(\frac{x-b}{a}\Big)\phi\Big(\frac{w-x}{\sigma}\Big)\, dx \tag{1}$$

Theorem

There exist unique a, b, σ^2, and a unique density f_S such that (1) holds true. Hence, the model is identifiable.

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

- The proof follows from Schwarz and Van Bellegem (2010), who prove the identifiability for any P_X belonging to

$$\{P \in \mathcal{P} \,|\, \exists A \in \mathcal{B}(\mathbb{R}) : |A| > 0 \text{ and } P(A) = 0\},$$

where $\mathcal{B}(\mathbb{R})$ = set of Borel sets in $\mathbb{R}$
$\mathcal{P}$ = set of all probability distributions on $\mathbb{R}$
$|A|$ = Lebesgue measure of A.

- Other error densities that allow to identify the model :
  - Cauchy
  - stable, ...

(see Schwarz and Van Bellegem, 2010).

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

To approximate the density f_S of S, we will make use of Bernstein polynomials. Why ?

- leads to rich and flexible parametric family of densities
- any continuous density can be approximated arbitrarily well by Bernstein polynomials (see below)
- requires only one regularization parameter
- compared to nonparametric deconvolution methods, it converges faster and is less sensitive to tuning parameters

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

A Bernstein polynomial of degree m is

$$B_m(s) = \sum_{k=0}^{m} \alpha_{k,m} b_{k,m}(s), \qquad s \in [0,1],$$

where $\alpha_{k,m} \in \mathbb{R}$, and

$$b_{k,m}(s) = \binom{m}{k} s^k (1-s)^{m-k}, \qquad k = 0, \ldots, m,$$

are Bernstein basis polynomials.
Bernstein (1912) shows that any continuous function $f(s)$ defined on $[0,1]$ can be uniformly approximated by such a polynomial :

$$\lim_{m\to\infty} \sup_{0\le s\le 1} \left| \sum_{k=0}^{m} f\left(\frac{k}{m}\right) b_{k,m}(s) - f(s) \right| = 0.$$

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Note that if we take $f \equiv f_S$,

$$\sum_{k=0}^{m} f_S\left(\frac{k}{m}\right) b_{k,m}(s)$$

$$= \sum_{k=0}^{m} \theta_{k,m} \underbrace{\frac{\Gamma(m+2)}{\Gamma(k+1)\Gamma(m-k+1)} s^k (1-s)^{m-k}}_{= \text{Beta}_{k+1,m-k+1}(s)},$$

where

$$\theta_{k,m} = f_S\left(\frac{k}{m}\right) \binom{m}{k} \frac{\Gamma(k+1)\Gamma(m-k+1)}{\Gamma(m+2)}.$$

This representation shows that f_S is approximated by a mixture of Beta$(k+1, m-k+1)$ densities $(k = 0, \ldots, m)$.

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

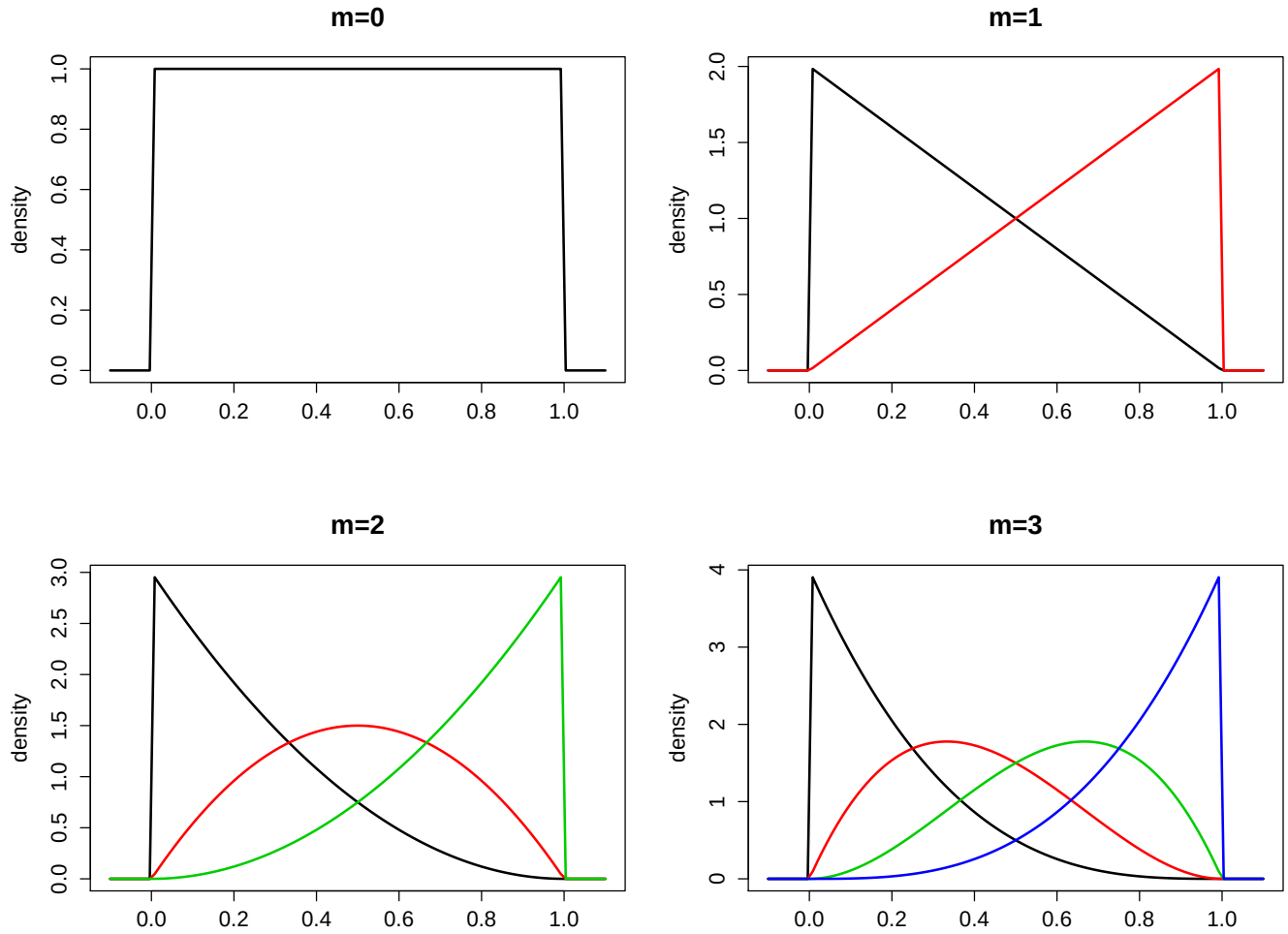


Figure: Representation of the Beta densities appearing in the Bernstein polynomials of degree $m = 0$, 1, 2 and 3.

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Recall that

$$f_W(w) = \frac{1}{a\sigma} \int f_S\Big(\frac{x-b}{a}\Big)\phi\Big(\frac{w-x}{\sigma}\Big)\, dx$$

which can now be approximated by

$$\begin{aligned}&\widetilde{f}_{W,m}(w;\sigma,a,b,\theta_m)\\ &= \frac{1}{a\sigma}\sum_{k=0}^{m}\theta_{k,m}\int \text{Beta}_{k+1,m-k+1}\Big(\frac{x-b}{a}\Big)\phi\Big(\frac{w-x}{\sigma}\Big)\, dx,\end{aligned}$$

which is a flexible $m+3$-dimensional parametric family of densities.

Note that

$$\lim_{m\to\infty}\sup_w \left|\widetilde{f}_{W,m}(w;\sigma,a,b,\theta_m) - f_W(w)\right| = 0,$$

as long as f_S is continuous.

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

When we have a sample $W_1, \ldots, W_n \overset{iid}{\sim} W$, the log-likelihood function of the set of parameters (σ, a, b, θ_m) is then

$$\mathcal{L}(\sigma, a, b, \theta_m) = \sum_{i=1}^{n} \log \widetilde{f}_{W,m}(W_i; \sigma, a, b, \theta_m).$$

This function can be maximized numerically with respect to (σ, a, b, θ_m) in order to obtain

$$(\widehat{\sigma}_m, \widehat{a}_m, \widehat{b}_m, \widehat{\theta}_m),$$

for a given value of m, the degree of the Bernstein polynomial.

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Note that a model with $m = m_1$ is nested in a model with $m = m_2$ for $m_1 < m_2$ (see Wang and Ghosh, 2012).
Hence, the quality of the approximation improves when m increases.

On the other hand, a large value of m implies a large number of parameters $\theta_{k,m}$ to be estimated, which could impair the quality of the estimated model.

Hence, we suggest choosing m using a model selection criterion.
Simulations suggest that BIC performs better than AIC thanks to the choice of more parsimonious models :

$$BIC(m) = (m+3)\log(n) - 2\mathcal{L}(\widehat{\sigma}_m, \widehat{a}_m, \widehat{b}_m, \widehat{\theta}_m), \qquad m \geq 0.$$

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Note that for a given value of m, $(\widehat{\sigma}_m, \widehat{a}_m, \widehat{b}_m, \widehat{\theta}_m)$ maximizes the likelihood of a potentially misspecified model

$\Rightarrow$ Its asymptotic properties can be derived based on the results in White (1982) on misspecified parametric models.

Let $(\sigma^*_m, a^*_m, b^*_m, \theta^*_m)$ be the parameter vector that minimizes the Kullback-Leibler Information Criterion :

$$E\left[\log\left\{\frac{f_W(W)}{\widetilde{f}_{W,m}(W;\sigma,a,b,\theta_m)}\right\}\right].$$

- ◇ Consistency : Under some regularity conditions,

$$(\widehat{\sigma}_m, \widehat{a}_m, \widehat{b}_m, \widehat{\theta}_m) \xrightarrow{P} (\sigma^*_m, a^*_m, b^*_m, \theta^*_m).$$

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

- Asymptotic normality : Under some regularity conditions,

$$n^{1/2}\Big((\widehat{\sigma}_m, \widehat{a}_m, \widehat{b}_m, \widehat{\theta}_m) - (\sigma_m^*, a_m^*, b_m^*, \theta_m^*)\Big) \xrightarrow{d} N(0, C),$$

where

$$C = A(\gamma^*)^{-1} B(\gamma^*) A(\gamma^*)^{-1},$$

with

$$A(\gamma) = \left(E\left\{\frac{\partial^2}{\partial\gamma_i \partial\gamma_j} \log \widetilde{f}_{W,m}(W;\gamma)\right\}\right)_{i,j},$$

$$B(\gamma) = \left(E\left\{\frac{\partial}{\partial\gamma_i} \log \widetilde{f}_{W,m}(W;\gamma) \cdot \frac{\partial}{\partial\gamma_j} \log \widetilde{f}_{W,m}(W;\gamma)\right\}\right)_{i,j},$$

$\gamma = (\sigma_m, a_m, b_m, \theta_m)$, and $\gamma^* = (\sigma_m^*, a_m^*, b_m^*, \theta_m^*)$.

Note that $C = A(\gamma)^{-1} =$ inverse Fisher matrix if the model is correctly specified.

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Simulations

We are mainly interested in the estimation of σ, and will therefore not report simulation results for the estimation of a, b and f_S.

Consider the following models :

- 8 different densities for X :
  - 2 Beta$(\alpha, \beta) - 1$ with $(\alpha, \beta) = (1, 1), (1, 2), (0.7, 0.5), (3, 2)$
  - Normal$(0, 1)$ truncated at $(t_L, t_U) = (-2, 2), (-1.5, 1.5)$
  - Exponential$(\mu, t_U) - 1$ of mean μ and truncated at t_U, with $(\mu, t_U) = (0.5, 4), (10, 20)$
- NSR = $\frac{\sigma}{\sigma_X} = 0.25, 0.50, 0.75$

For each model, σ was estimated using $m = 0, \ldots, 6$ for each of 500 replicated datasets.

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

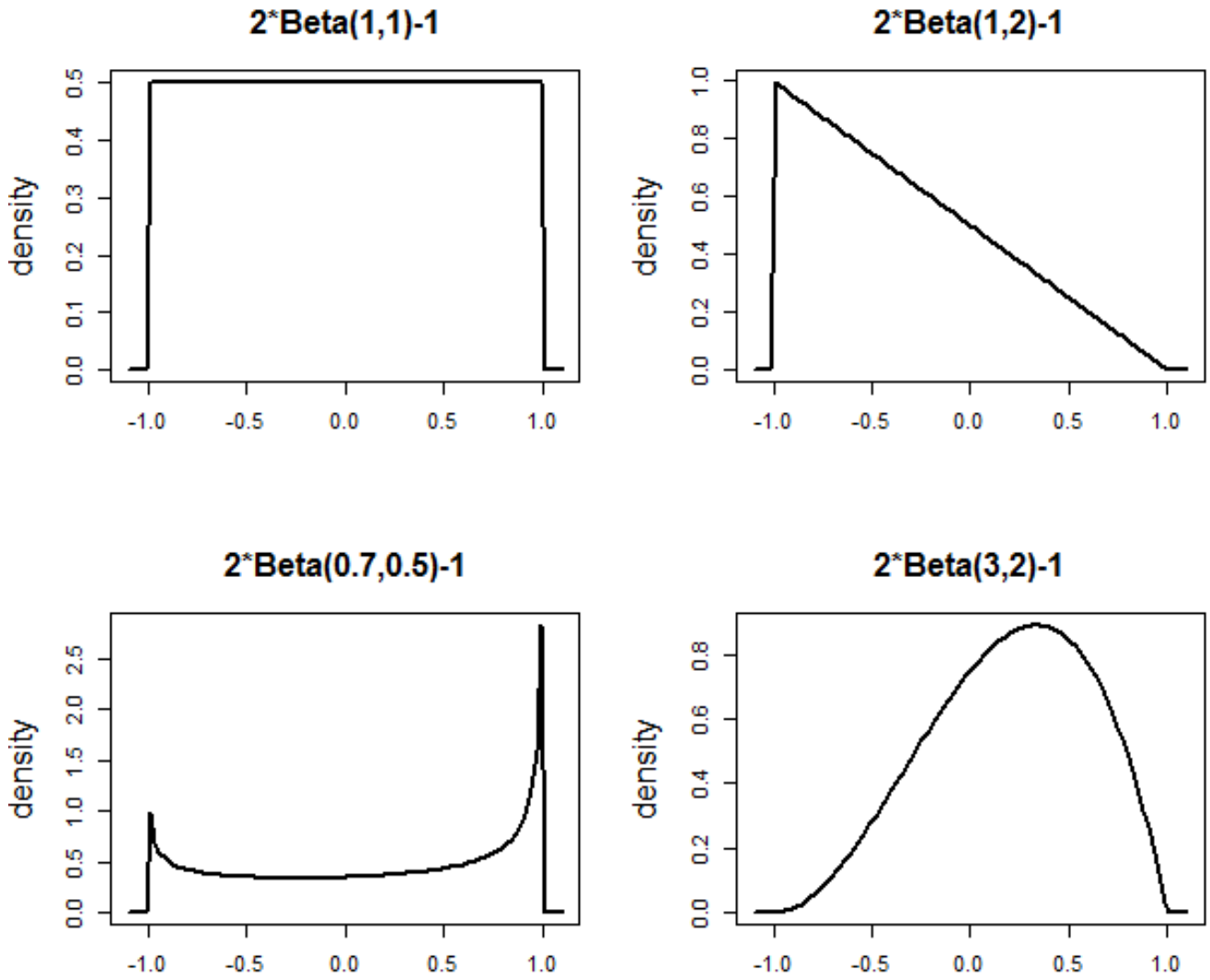


Figure: Representation of the densities f_X considered in the simulation.

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

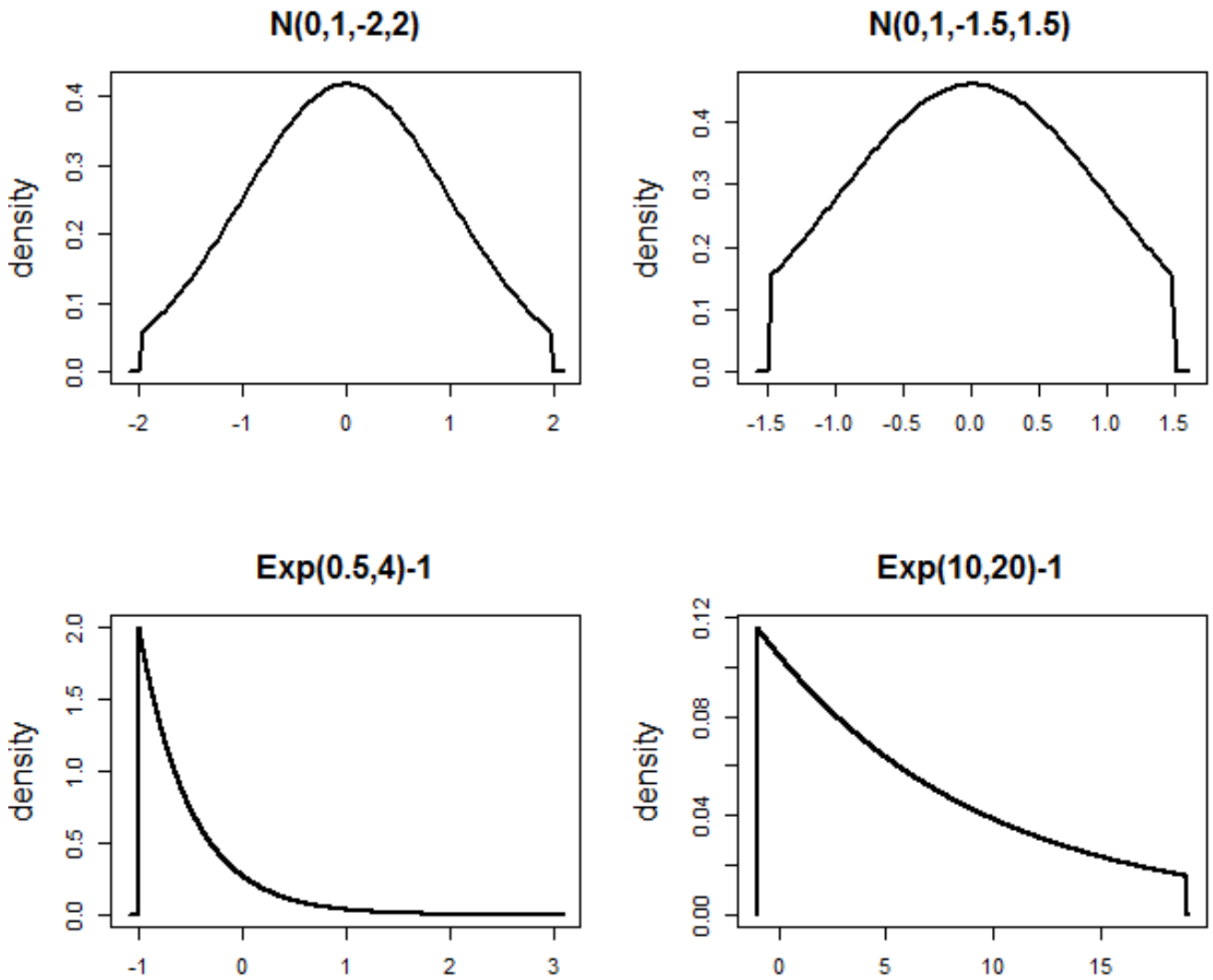


Figure: Representation of the densities f_X considered in the simulation.

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Table: Simulation results for $n = 300$ (RB: relative bias; SD: standard deviation; MSE: mean squared error)

		Estimation of σ				Distribution (in %) of the selected m						
f_X	σ	Bias	RB	SD	MSE	0	1	2	3	4	5	6
2 Beta(1, 1)-1	.144	-.010	-.069	.064	.004	90.0	2.4	6.2	0.8	0.0	0.2	0.4
	.289	-.011	-.037	.076	.006	92.4	3.6	3.0	0.6	0.2	0.0	0.2
	.433	-.003	-.007	.092	.009	93.2	3.0	1.0	1.2	1.2	0.2	0.2
2 Beta(1, 2) -1	.118	-.008	-.069	.053	.003	0.2	90.8	2.0	4.6	1.8	0.6	0.0
	.236	-.006	-.024	.072	.005	7.8	85.2	2.2	3.2	0.4	0.6	0.6
	.354	.014	.040	.101	.010	44.8	50.0	1.6	0.8	1.2	1.0	0.6
2 Beta(.7, .5)-1	.166	-.037	-.221	.063	.005	10.4	28.4	55.6	4.8	0.6	0.0	0.2
	.332	-.067	-.202	.077	.010	44.2	49.2	4.2	2.2	0.0	0.0	0.2
	.499	-.052	-.104	.102	.013	74.2	22.8	1.0	0.8	1.0	0.2	0.0
2 Beta(3, 2)-1	.100	.073	.727	.083	.012	35.4	49.2	1.0	10.8	2.0	1.2	0.4
	.200	.065	.326	.072	.009	65.2	29.2	1.6	1.2	1.4	1.4	0.0
	.300	.059	.196	.078	.010	84.2	12.0	0.8	0.6	1.2	0.4	0.8
N(0,1,-2,2)	.440	.209	.475	.137	.063	93.6	1.6	1.8	0.6	1.0	1.0	0.4
	.880	.116	.132	.177	.045	94.2	3.2	0.4	0.8	0.6	0.0	0.8
	1.32	.032	.024	.229	.053	93.0	2.6	0.4	1.0	1.4	0.6	1.0
N(0,1,-1.5,1.5)	.371	.088	.238	.127	.024	92.4	1.4	2.4	1.6	1.2	0.2	0.8
	.743	.053	.071	.154	.027	96.0	2.6	0.2	0.2	0.4	0.2	0.4
	1.11	.006	.005	.189	.036	97.6	2.0	0.2	0.0	0.2	0.0	0.0
Exp(.5, 4)-1	.124	-.008	-.066	.058	.003	0.4	0.2	1.2	30.8	41.8	19.4	6.2
	.247	-.026	-.104	.037	.002	0.2	0.4	8.6	52.2	27.8	10.0	0.8
	.371	-.029	-.077	.064	.005	3.2	2.0	27.6	46.6	18.4	1.4	0.8
Exp(10, 20)-1	1.31	.001	.001	.947	.897	0.8	70.8	25.0	1.0	1.4	0.6	0.4
	2.63	-.049	-.019	1.02	1.04	3.4	84.4	8.2	2.0	0.6	1.0	0.4
	3.94	.121	.031	1.14	1.32	25.8	64.6	5.2	1.6	1.2	0.4	1.2

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Note that the true value of m is 0 for Beta$(1,1)$
1 for Beta$(1,2)$
3 for Beta$(3,2)$

The other densities are not a mixture of Bernstein polynomials.

The table shows that

- The BIC criterion recovers well the value of m for Beta(1,1) and Beta(1,2), but not for Beta$(3,2)$.
- The selected m tends to decrease with the SNR.
- Smallest relative biases are found for 2 Beta(1,1)-1, 2 Beta(1,2)-1 and both exponential distributions.
- 2 Beta(3,2)-1 and N(0,1,-2,2) yield the worst results, but bias decreases when σ increases.
- Although the model is theoretically identifiable, there appears some practical identifiability problems especially for large values of m, which disappear when a and b are set to their true values.

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Illustration : Cox model estimated with Simex

Our final objective is to be able to estimate regression models, in which covariates are subject to measurement error with unknown variance.

General strategy :

- Estimate σ^2 with the proposed method
- Apply any of the existing methods for estimating regression coefficients when measurement error is present, with σ^2 replaced by $\hat{\sigma}^2$

All methods require the error distribution to be known, unless validation or auxiliary data are available.

We will focus here on Simex.

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

We will illustrate the Simex methodology (with σ^2 replaced by $\hat{\sigma}^2$) on a Cox model with measurement error :

$$h(t|X_1, X_2) = h_0(t) \exp(\beta_1 X_1 + \beta_2 X_2), \quad t \geq 0,$$

where

- $X_1 \sim 2$ Beta(1,1)-1, $X_1 \sim 2$ Beta(0.7,0.5)-1 or $X_1 \sim N(0,1)$ truncated at $[-2, 2]$
- $U_1 \sim N(0, \sigma^2)$
- $X_2 \sim$ Bernoulli(0.5), $U_2 = 0$ (no measurement error)
- $\beta_1 = 1$, $\beta_2 = -0.5$
- $h_0(t) = 2$

T is subject to random right censoring, i.e. instead of observing T we observe

$$Y = \min(T, C) \quad \text{and} \quad \Delta = I(T \leq C),$$

where $C \perp\!\!\!\perp T$ given $X = (X_1, X_2)$.

We take $C \perp\!\!\!\perp X$ and $C \sim \text{Exp}(\mu = 3)$.

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Table: Simulation results for $n = 300$

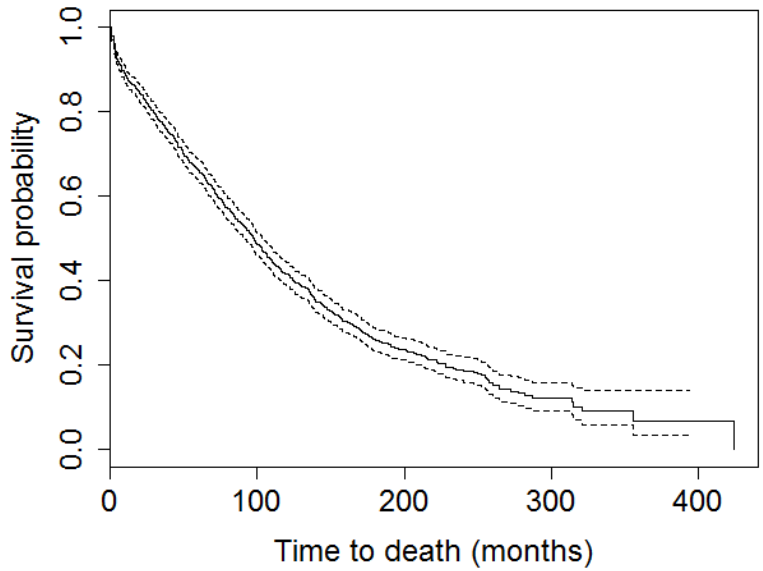
			Naive		Simex (estimated σ)		Simex (true σ)	
f_X	σ		β_1	β_2	β_1	β_2	β_1	β_2
2 Beta(1,1)-1	.144	Bias	-.054	-.007	.013	-.013	.012	-.012
		SD	.136	.164	.163	.165	.150	.165
		MSE	.021	.027	.027	.027	.023	.027
	.289	Bias	-.217	.001	-.038	-.012	-.036	-.012
		SD	.139	.168	.197	.173	.182	.172
		MSE	.066	.028	.040	.030	.034	.030
	.433	Bias	-.387	.026	-.146	.011	-.147	.010
		SD	.112	.162	.186	.173	.168	.173
		MSE	.162	.027	.056	.030	.050	.030
2 Beta(.7,.5)-1	.166	Bias	-.057	-.004	-.007	-.008	.017	-.010
		SD	.131	.164	.160	.166	.147	.167
		MSE	.021	.027	.026	.028	.022	.028
	.332	Bias	-.234	.008	-.109	-.003	-.045	-.008
		SD	.124	.161	.166	.165	.165	.168
		MSE	.070	.026	.040	.027	.029	.028
	.499	Bias	-.399	.028	-.196	.009	-.156	.004
		SD	.101	.157	.160	.167	.155	.169
		MSE	.169	.025	.064	.028	.048	.029
N(0,1,-2,2)	.220	Bias	-.068	-.008	.302	-.060	.007	-.019
		SD	.102	.170	.236	.204	.115	.176
		MSE	.015	.029	.146	.045	.013	.031
	.440	Bias	-.238	.023	.157	-.024	-.040	-.001
		SD	.088	.173	.186	.202	.123	.187
		MSE	.064	.030	.060	.042	.017	.035
	.660	Bias	-.420	.041	-.079	.001	-.174	.011
		SD	.081	.160	.163	.188	.129	.179
		MSE	.183	.027	.033	.035	.047	.032

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Data analysis

We consider data on 1341 patients suffering from 'monoclonal gammapothy of undetermined significance' (MGUS), a precursor lesion for multiple myeloma

Kaplan-Meier curve of the survival time :



Censoring proportion = 30%

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

For each patient the following covariates are recorded :

- hemoglobin
- log(creatinine)
- monoclonal spike
- age
- gender

Hemoglobin, creatinine and monoclonal spike are subject to measurement error, age and gender are supposed to be error free.

We will fit a Cox model to these data using the Simex approach

$\Rightarrow$ first we need to estimate the measurement error variances

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Hemoglobin level	$m = 1$	$m = 2$	**m=3**	$m = 4$	$m = 5$
BIC$\cdot 10^{-3}$	5.7986	5.8053	**5.7770**	5.7834	5.7904
$\widehat{\sigma}$	1.2351	1.4911	**1.3616**	1.2798	1.3385
Creatinine log-level	$m = 9$	$m = 10$	**m=11**	$m = 12$	$m = 13$
BIC$\cdot 10^{-3}$	0.7345	0.7284	**0.7265**	0.7271	0.7294
$\widehat{\sigma}$	0.1836	0.1871	**0.1907**	0.1944	0.1975
Monoclonal spike	$m = 7$	$m = 8$	**m=9**	$m = 10$	$m = 11$
BIC$\cdot 10^{-3}$	2.2785	2.2810	**2.2749**	2.2755	2.2792
$\widehat{\sigma}$	0.1743	0.1731	**0.1780**	0.1685	0.1706

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

		Age	Gender	Hemo-globin	Log-creatinine	Spike
No correction	Estim.	.055	-.389	-.121	.367	.037
	SE	.003	.070	.018	.079	.060
SIMEX	Estim.	.053	-.449	-.183	.349	.030
	SE	.003	.081	.027	.132	.068

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

Conclusions

- New method to estimate the measurement error variance, that is
  - a stable and feasible practical method
  - consistent and asymptotically normal
- Estimation of regression models with measurement error in the covariates with unknown variance
  $\hookrightarrow$ Illustrated with Cox proportional hazards model

Basic concepts
Cure models
Introduction
Ongoing research
Dependent censoring
Introduction
Ongoing research
Measurement errors
Introduction
Ongoing research

The End