

**Truncated maximum likelihood regression
with censored responses
and application to the estimation
of the mean hospital cost of stay**

Joint work with

Isabella Locatelli

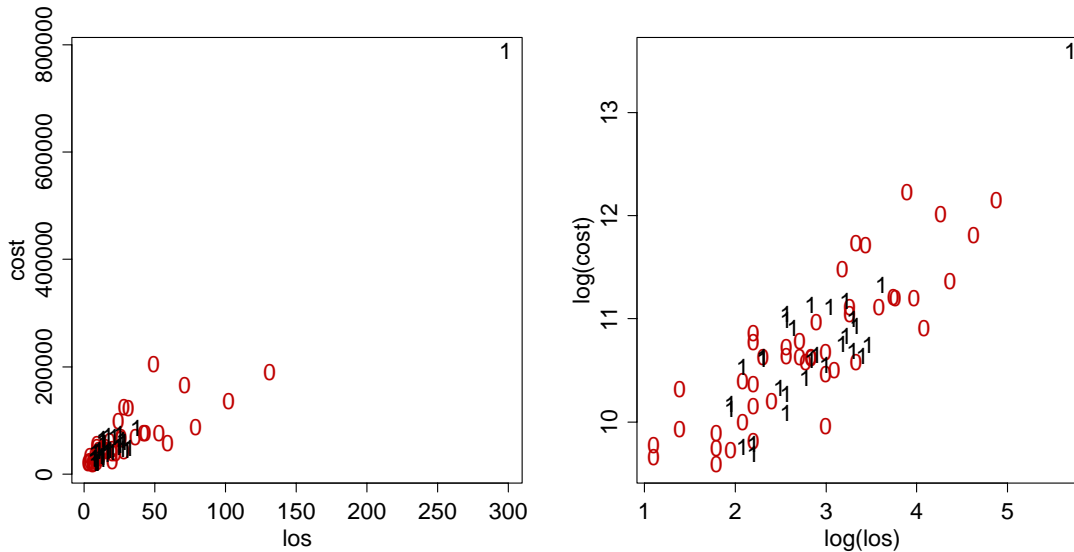
Victor Yohai

Practical problem

Analysis of hospital **cost** and **length of stay (los)**

Goals: estimate $E(\text{los}|\text{covariate values})$ and $E(\text{cost}|\text{covariate values})$

75 stays, major cardiovascular procedures with major cc
CHUV Lausanne 2000



- 45 censored (0: transfer to another hospital); 30 noncensored (1)
- Covariates: age, sex, type of admission (emergency, planned), etc.
- Cost and los distributions are asymmetric
- There are **outliers**

Censored times and costs

Let T be the unobserved survival time (complete los in days) for a patient

C be the censoring time (e.g. time of transfer) of the patient

$T^* = \min(T, C)$ be the observed time

Usual assumption *non-informative censoring* :

T and C are independent

Let U be the cost of a random patient per unit of time (e.g. day: “unit cost”)

Assume:

U does not change over time

U is independent of C

Then

$Y = UT$ is the total unobserved cost of the patient at the survival time

$K = UC$ is the censored total cost of the patient

$Y^* = UT^*$ is the observed total cost of the patient

Can we use standard survival techniques on the cost scale ?

(a) Suppose that U independent from T (and C). Then,

$$\text{Cov}(Y, K) = \text{Var}(U)E(T)E(C) > 0,$$

cost censoring is informative

Example

Estimates of $E(Y) = E(U)E(T) = 50$ with T, C, U lognormal

$E(T) = 5, \sigma(T) = 7; E(C) = 12, \sigma(C) = 7; E(U) = 10, \sigma(U) = \sigma_0$

$\text{Var}(U)^{1/2}$	$\text{Cor}(Y, K)$	KMY	AFTY	KMT	AFTT
2	0.023	50.52	50.39	49.71	49.75
8	0.189	59.78	57.63	49.87	50.16
15	0.306	72.53	67.74	50.07	49.84

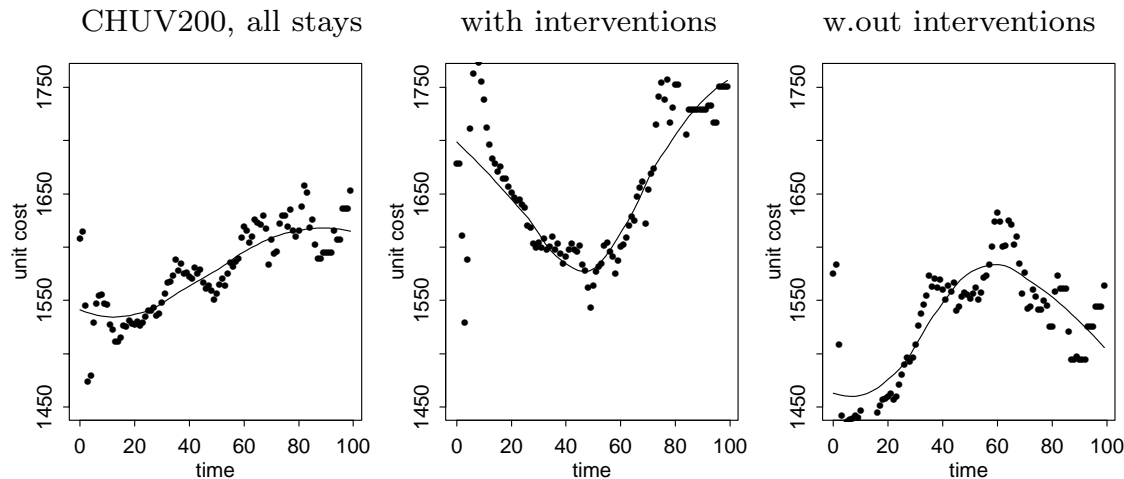
KMY : direct Kaplan-Meier estimate of $E(K)$

AFTY : direct AFT survival estimate of $E(K)$

KMT : Kaplan-Meier estimate of $E(T)$ · estimate of $E(U)$

AFTT : AFT survival estimate of $E(T)$ · estimate of $E(U)$

Mean unit cost of patients which are still at the hospital after a certain time



In general, U and T are not independent.

(b) Suppose that $U = bT^a$ (e.g.). Then,

$$\begin{aligned} \text{Cov}(Y, K) = E(C)\text{Cov}(UT, U) &> 0 && \text{if } a < -1, \\ &< 0 && \text{if } a \in (-1, 0), \\ &= 0 && \text{if } a = -1 \text{ or } a = 0. \end{aligned}$$

- In general, cost censoring is informative
 \Rightarrow standard survival techniques cannot be applied to cost data.
- Use a survival technique to estimate the distribution of T .
- Estimate the “mean unit cost”; combine the two steps: e.g., $E(\widehat{Y}) = E(\widehat{U})E(\widehat{T})$.

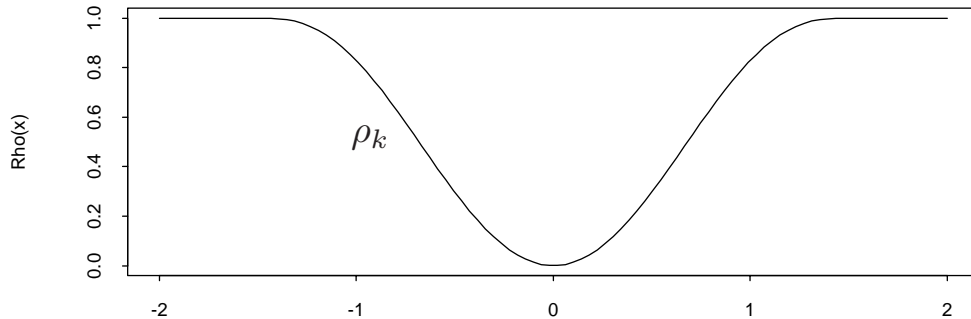
S-estimates : location and scale

Let ρ be a function such that:

(i) $\rho(0) = 0$; (ii) ρ even; (iii) if $|u| < |v|$ then $\rho(u) \leq \rho(v)$; (iv) ρ **bounded**.

Example: Tukey's biweight family:

$$\rho_k(z) = \begin{cases} 3(z/k)^2 - 3(z/k)^4 + (z/k)^6 & \text{if } |z| \leq k, \\ 1 & \text{if } |z| > k, \end{cases}$$



Suppose

$$y_i = \alpha_0 + \sigma_0 e_i, \quad \text{iid.}$$

$e_i \sim F_0$ e.g. a parametric model

For any α , let $F_{n,\alpha}$ be the empirical cdf of the residuals $y_i - \alpha$.

S-estimate of location (Rousseeuw & Yohai, 1984):

$$\tilde{\alpha} = \arg \min_{\alpha} S(\alpha),$$

where $S(\alpha)$ is an “M-scale” defined by

$$\frac{1}{n} \sum \rho \left(\frac{y_i - \alpha}{S(\alpha)} \right) = 0.5 \quad \text{i.e.,} \quad E_{F_{n,\alpha}} \left[\rho \left(\frac{e}{S(\alpha)} \right) \right] = 0.5.$$

S-estimate of scale: $\tilde{\sigma} = S(\tilde{\alpha})$.

S-location of F_0 : $m_0 = \arg \min S_0(\alpha)$,

where $S_0(\alpha)$ solves

$$E_{F_0} \left[\rho \left(\frac{e - \alpha}{S_0(\alpha)} \right) \right] = 0.5.$$

S-scale of F_0 : $s_0 = S_0(\tilde{\alpha}_0)$.

Then, $\tilde{\alpha} - (\tilde{\sigma}/s_0)m_0$ and $\tilde{\sigma}/s_0$

are robust consistent estimates of α_0 and σ_0 .

◦ S-estimates can attain 50% BDP, but inefficient wrt ML when model correct.

Truncated maximum likelihood: location-scale (Marazzi & Yohai, 2004)

Assume $y_i = \alpha_0 + \sigma_0 e_i$, $e_i \sim F_0$, $y_i \sim F_{\alpha_0, \sigma_0}$

Let

$$s_1(z) = [\partial \ln f_{\alpha, \sigma} / \partial \alpha]_{\alpha=0, \sigma=1}, \quad s_2(z) = [\partial \ln f_{\alpha, \sigma} / \partial \sigma]_{\alpha=0, \sigma=1}$$

1. Compute **initial high BDP** $\tilde{\alpha}$, $\tilde{\sigma}$, e.g. S-estimates

2. $r_i = (y_i - \tilde{\alpha}) / \tilde{\sigma}$

$w_i = 0$ if likelihood of y_i is small i.e. : $f_0(r_i) < \eta$, (e.g. 0.01)

$w_i = 1$ if likelihood of y_i is large i.e. : $f_0(r_i) \geq \eta$.

3. Compute **final ML estimates** $\hat{\alpha}$, $\hat{\sigma}$ on retained observations: solve

$$\sum w_i s_1((y_i - \hat{\alpha}) / \hat{\sigma}) = 0,$$

$$\sum w_i s_2((y_i - \hat{\alpha}) / \hat{\sigma}) = 0.$$

Correct $(\hat{\alpha}, \hat{\sigma})$ for consistency if necessary

o $(\hat{\alpha}, \hat{\sigma})$ maintain the BDP of $(\tilde{\alpha}, \tilde{\sigma})$; efficiency wrt ML close to 100%.

Regression with censored data: LS case

$$t_i = x_i^T \beta_0 + \sigma_0 e_i, \quad \beta_0 \in \mathbb{R}^p, \quad x_i \in \mathbb{R}^p,$$

$$(e_i, x_i) \sim H \text{ iid}; \quad e_i \sim F \text{ indep. of } x_i.$$

We observe

$$t_i^* = \min(t_i, c_i), \quad \text{and} \quad d_i = I(t_i < c_i),$$

where the c_i 's are iid censoring times independent of the e_i .

When no censoring: **Least Squares** normal equations

$$\sum r_i(\beta) x_i = 0,$$

where

$$r_i(\beta) = t_i - x_i^T \beta$$

When censoring: **Bukley & James, 1979** modification of normal equations:

$$\frac{1}{n} \left[\sum_{d_i=1} r_i(\beta) x_i + \sum_{d_i=0} \bar{r}_i(\beta) x_i \right] = 0 \quad (\text{NE})$$

where

$$\bar{r}_i(\beta) = E_{F_{n,\beta}^*} [e | e > r_i^*(\beta)]$$

and $F_{n,\beta}^*$ is the KM cdf based on

$$r_i^*(\beta) = t_i^* - x_i^T \beta$$

Note :

$$(\text{NE}) \Leftrightarrow E_{H_{n,\beta}^*} [ex] = 0,$$

where $H_{n,\beta}^*$ is a nonparametric estimate of H based on KM and $r_i^*(\beta)$ (Salibian-Barrera and Yohai, 200?).

- **Consistency of Bukley & James:** James & Smith, 1984; Lai & Ying, 1991
- **Consistency of $H_{n,\beta}^*(x, e)$ for $\beta = \beta_0$:** Salibian-Barrera & Yohai, 200?

High BDP regression with symmetric errors and censored data

Salibian-Barrera & Yohai, 200?

Suppose that $\hat{\sigma}$ is a known robust scale of the residuals.

For any β , suppose that γ is a correction of β and

$$Q(\beta, \gamma) = E_{H_{n,\beta}^*} [\rho((e - x_i^T \gamma) / \hat{\sigma})],$$

is the “loss of using $\beta + \gamma$ in place of β ”. Let

$$\hat{\gamma}(\beta) = \arg \min_{\gamma} Q(\beta, \gamma).$$

and note that, if $\beta = \beta_0$, then $\hat{\gamma}(\beta_0) = 0$.

Therefore, define an estimate $\hat{\beta}$ by the equation

$$\hat{\gamma}(\hat{\beta}) = 0.$$

In particular, consider the “loss” $S(\beta, \gamma)$ defined by

$$E_{H_{n,\beta}^*} \left[\rho \left(\frac{e - x_i^T \gamma}{S(\beta, \gamma)} \right) \right] = 0.5$$

and let

$$\tilde{\gamma}(\beta) = \arg \min_{\gamma} S(\beta, \gamma).$$

Then, define the **S-estimate** $\tilde{\beta}$ by the equation

$$\tilde{\gamma}(\tilde{\beta}) = 0.$$

- The **BDP** of $\tilde{\beta}$ is $> 0.5 \times (n - p + 1)/n - m/n$, where m is the number of censored observations in the sample.
- If the error distribution is symmetric and has a unimodal density then, the **S-estimate** is **Fisher consistent** (i.e., $\tilde{\gamma}(\beta_0) \rightarrow 0$ a.s.)
- The efficiency of S-estimates is low; can be improved using MM or TML
- ...

TML regression with asymmetric errors and right-censoring

Model:

$$t_i = x_i^T \beta_0 + \sigma_0 e_i,$$
$$e_i \sim F_0$$

F_0 is the standard version of a parametric asymmetric or symmetric model.

Examples

- $t_i \sim \mathcal{N}(x_i^T \beta_0, \sigma_0^2)$ $F_0 = \Phi$;
- t_i is the log of a Lognormal variable, $F_0 = \Phi$;
- t_i is the log of a Weibull variable, F_0 is a stand. Gumbel min. cdf.
- t_i is a Gumbel max. variable F_0 is a stand. Gumbel max. cdf.

Let m_0 and s_0 be the S-location and the S-scale of F_0 .

Initial estimates

1. For any (σ, β) :
$$v_i^*(\sigma, \beta) = t_i^* - x_i^T \beta - \sigma m_0$$

2. $S(\sigma, \beta, \gamma)$:

$$E_{\hat{H}_{n,\sigma,\beta}} \left[\rho \left(\frac{e - x_i^T \gamma}{S(\sigma, \beta, \gamma)} \right) \right] = 0.5,$$

$\hat{H}_{n,\sigma,\beta}$ is a “parametric” estimate of $H(e, x)$ based on F_0 and $v_i^*(\sigma, \beta)$

$$\tilde{\gamma}(\sigma, \beta) = \arg \min_{\gamma} S(\sigma, \beta, \gamma)$$

3. Define $\tilde{\beta}(\sigma)$:
$$\tilde{\gamma}(\sigma, \beta) = 0$$

We need an estimate of σ_0 for given β :

(a)
$$\hat{\sigma} = \text{MAD}(F_{n,\beta}^*) / \text{MAD}(F_0), \quad F_{n,\beta}^* = \text{KM of residuals}$$

(b)
$$\tilde{\sigma} = S(\tilde{\sigma}, \tilde{\beta}(\tilde{\sigma}), 0) / s_0$$

Truncated maximum likelihood estimates

1. Compute **initial high BDP** S-estimates $\tilde{\beta}, \tilde{\sigma}$
2. Compute standardized residuals $r_i^* = (t_i^* - x_i^T \tilde{\beta}) / \tilde{\sigma}$
3. Compute **final ML estimates**
 - rejecting unlikely observations (unlikely residuals under initial model),
 - replacing censored residuals with expected residuals under the condition that they are larger than the observed residuals.

Step 2 formally:

Let η be a small number (e.g. 0.01),

$$w(e) = I(f_0(e) > 0),$$

$$s_1(z) = [\partial \ln f_{\alpha, \sigma} / \partial \alpha]_{\alpha=0, \sigma=1}, \quad s_2(z) = [\partial \ln f_{\alpha, \sigma} / \partial \sigma]_{\alpha=0, \sigma=1}.$$

let $\tilde{t}_i = x_i^T \tilde{\beta}$ and solve

$$\sum E_{F_{n,i}} w(e) s_1 \left(\frac{\tilde{y}_i + e - x_i^T \hat{\beta}}{\hat{\sigma}} \right) x_i^T = 0,$$

$$\sum E_{F_{n,i}} w(e) s_2 \left(\frac{\tilde{y}_i + e - x_i^T \hat{\beta}}{\hat{\sigma}} \right) = 0,$$

where

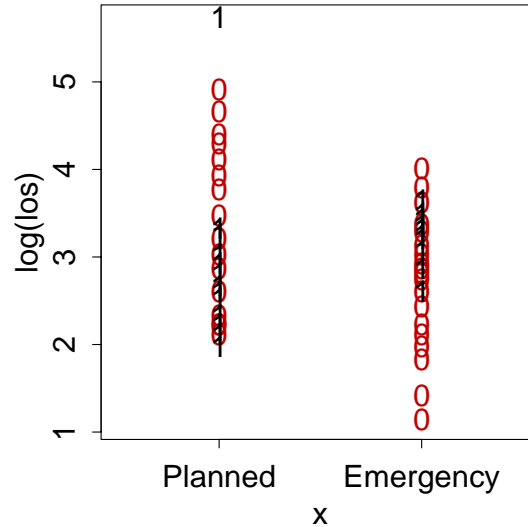
if i non censored $F_{n,i}(e) = I(v_i^* \leq e)$

if i censored

$$F_{n,i}(e) = (F_0(e) - F_0(v_i^*)) / (1 - F_0(v_i^*)) \quad \text{if } e \geq v_i^*, \\ = 0 \quad \text{otherwise.}$$

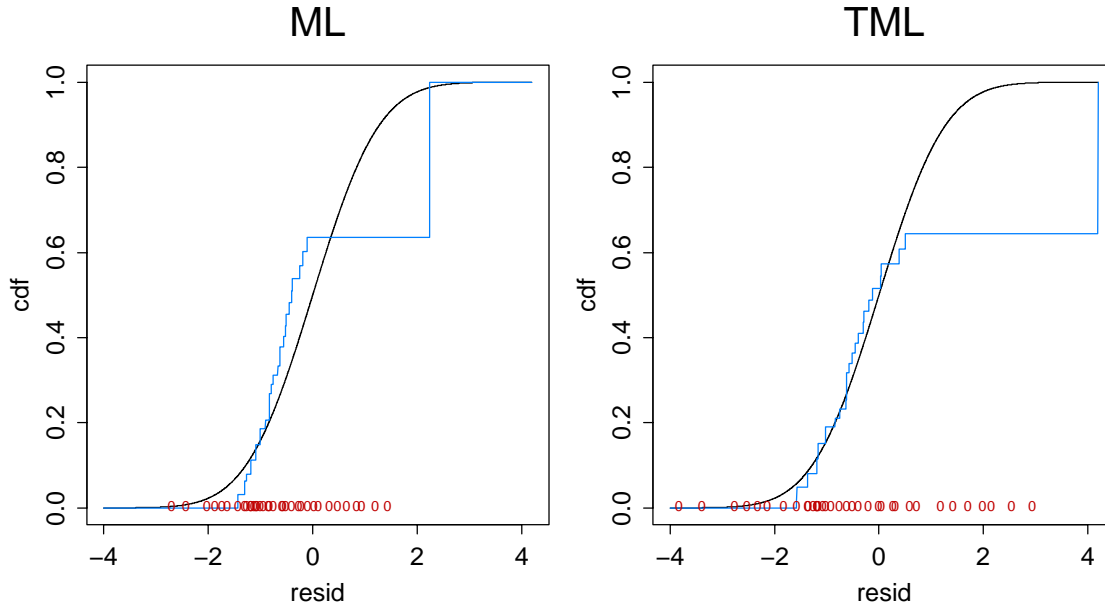
Example 1

75 stays, major cardiovascular procedures with major cc
45 are censored (0)



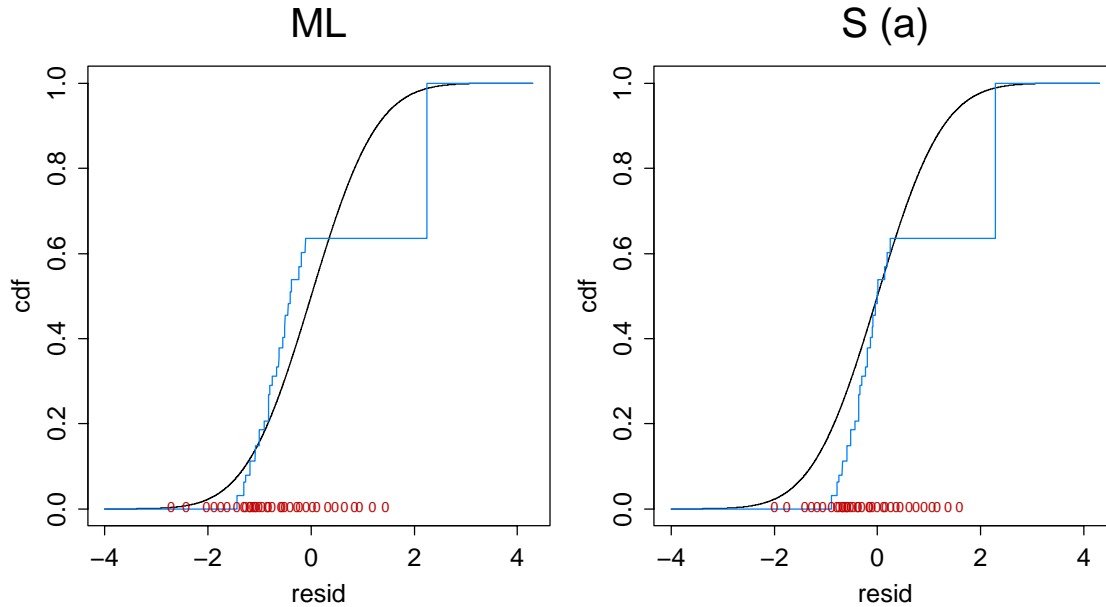
Model: $\log(\text{los}) = \alpha + \beta x + \sigma \cdot e, \quad e \sim N(0, 1),$
 $x = 0$: Planned admission,
 $x = 1$: Emergency admission.

	ML	$S^{(a)}$	TML
$\hat{\alpha}$	3.41	3.00	2.97
$\hat{\beta}$	0.45	0.45	0.62
$\hat{\sigma}$	1.02	1.18	0.65



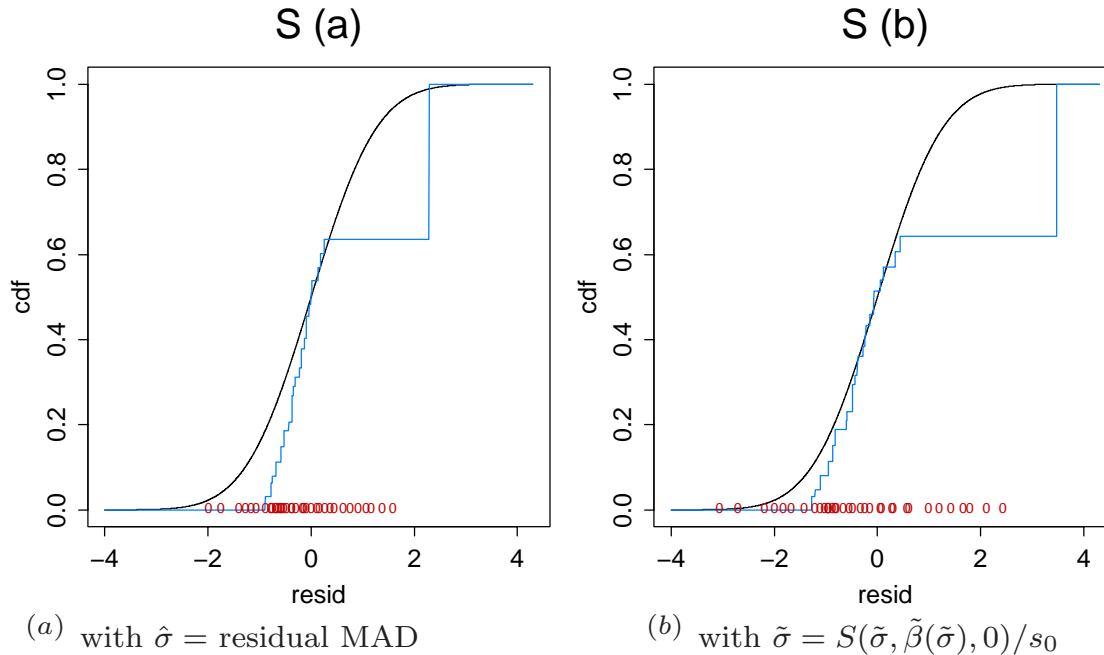
(a) with $\hat{\sigma} = \text{residual MAD}$

	ML	$S^{(a)}$
$\hat{\alpha}$	3.41	3.00
$\hat{\beta}$	0.45	0.45
$\hat{\sigma}$	1.02	1.18



(a) with $\hat{\sigma} = \text{residual MAD}$

	$S^{(a)}$	$S^{(b)}$
$\hat{\alpha}$	3.00	2.95
$\hat{\beta}$	0.45	0.57
$\hat{\sigma}$	1.18	0.79



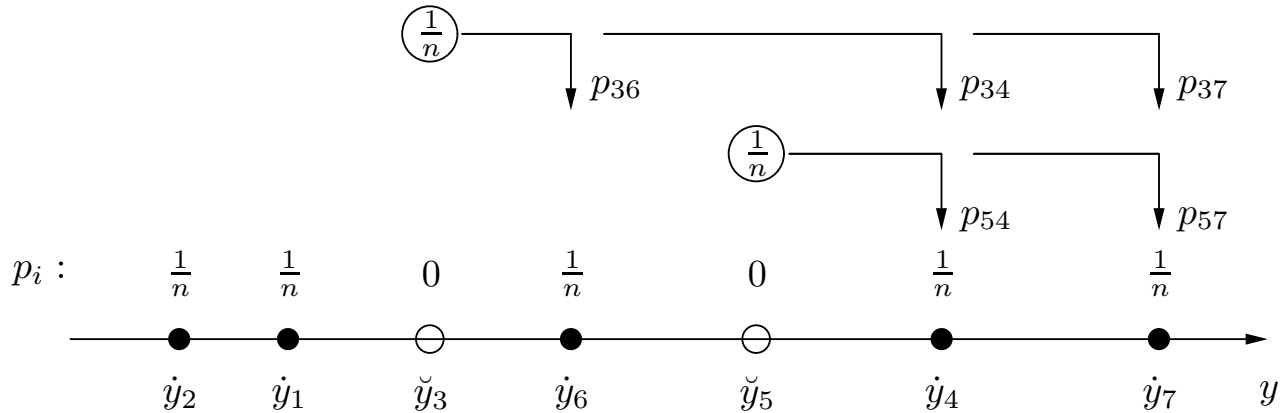
Thus : MAD based on KM is not sufficiently robust;
 a full parametric approach is required if proportion of censoring is high

Interpretation of KM

Let \check{y}_i the censored y 's,

\dot{y}_i the non-censored y 's,

p_i the mass assigned to y_i^* .

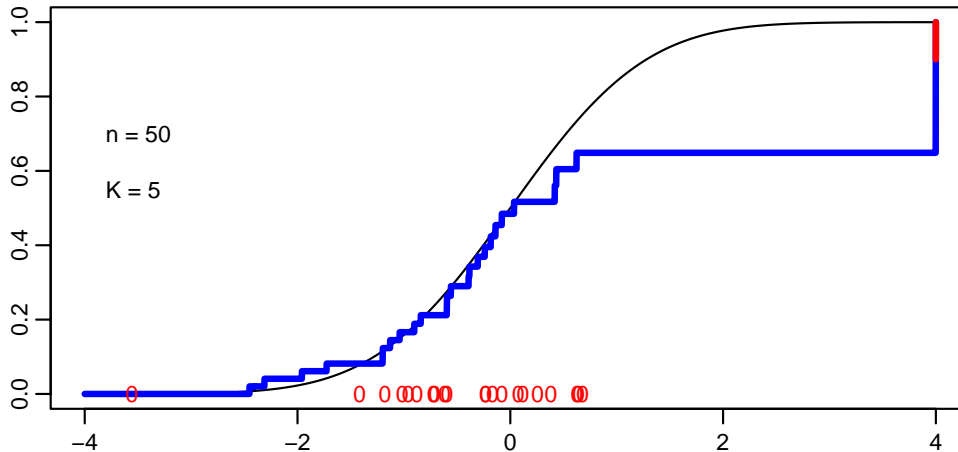


- $p_i = 0$ if $y_i = \check{y}_i$
- the mass $1/n$ of \dot{y}_i is assigned to \dot{y}_i
- the mass $1/n$ of \check{y}_j is distributed among the \dot{y}_i with $\dot{y}_i > \check{y}_j$ with probabilities $p_{ji} = P(y = \dot{y}_i | y > \check{y}_j)$, i.e.,

$$p_i = \frac{1}{n} + \sum_{\check{y}_j < \dot{y}_i} p_{ji}$$

Consequences

- The mass assigned by KM to K large non-censored outliers is $> K/n$ because they receive part of the mass of the censored observations.



The worst case is when all (J) censored observations are located between the non-censored good observations and K non-censored outliers. Then, the mass assigned to the outliers is $(K + J)/n$.

⇒ The BDP of a robust estimate based on KM is reduced by J/n .

- Let m be the number of censored observations larger than $\max(\hat{y}_i)$. Then $\sum p_i = 1 - m/n$.

Estimating a mean cost under informative cost censoring

Consider a time period $[0, \tau]$ (e.g. one month).

Divide $[0, \tau]$ in intervals $[t_k, t_{k+1})$ of length 1 (e.g., days).

Option (a) : Let U_k be the cost incurred over day k :

$$U_k = U \cdot I(T \geq t_k).$$

Then

$$Y = \sum U_k.$$

The distribution of U_k among those who “survive” t_k is the same as the distribution of U_k among those who do not survive t_k and

$$E(Y) = \sum E(U_k | T \geq t_k) P(T \geq t_k) = \sum E(U_k | T^* \geq t_k) P(T \geq t_k)$$

- $P(T \geq t_k)$ can be estimated using a survival model.
- $E(U_k | T^* \geq t_k)$ can be (consistently) estimated by :
 - the mean of the U_k 's among patients under observation at t_k or
 - the TML of the U_k 's among patients under observation at t_k .

Option (b)

$$E(Y) = \sum E(Y|t_k \leq T < t_{k+1})P(t_k \leq T < t_{k+1}) + E(Y|T > \tau)P(T > \tau)$$

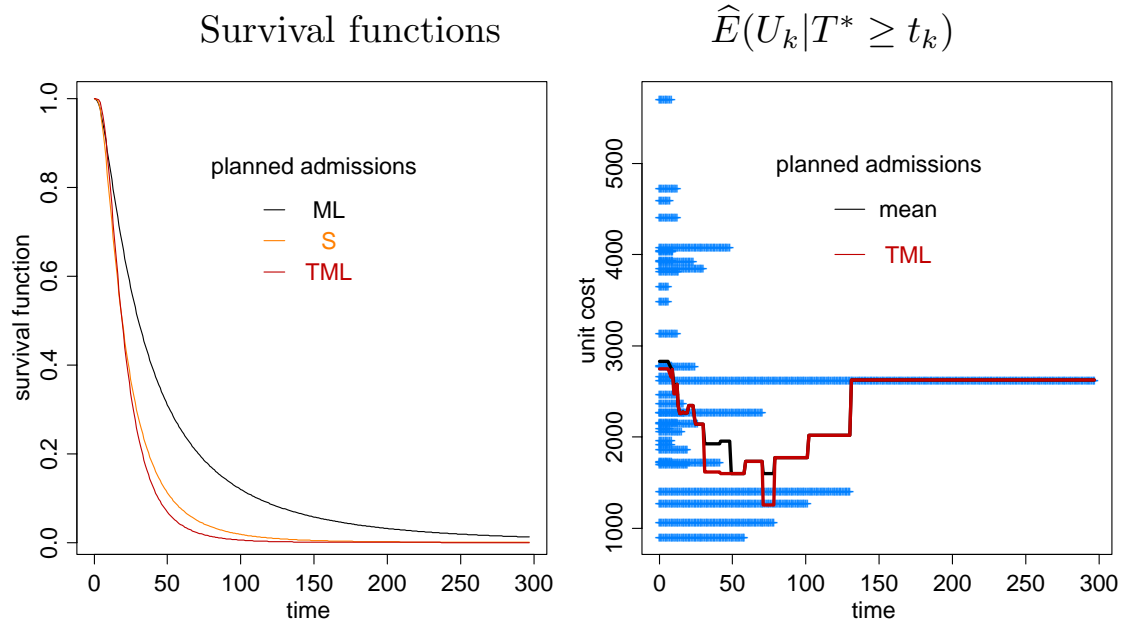
and

≈ 0

$$E(Y|t_k \leq T < t_{k+1}) = E(Y|t_k \leq T^*, T < t_{k+1})$$

- $P(t_k \leq T < t_{k+1})$ can be estimated using a survival model.
- $E(Y|t_k \leq T^*, T < t_{k+1})$ can be (consistently) estimated by :
 - the mean of the observed Y 's among those who “go home” in (t_k, t_{k+1}) or
 - the TML of the observed Y 's among those who “go home” in (t_k, t_{k+1}) .
- Unfortunately, we usually have very few observations in each single day.
 \Rightarrow use a functional model for $E(Y|t \leq T^*, T < t + \delta)$?

Example 1: continuation

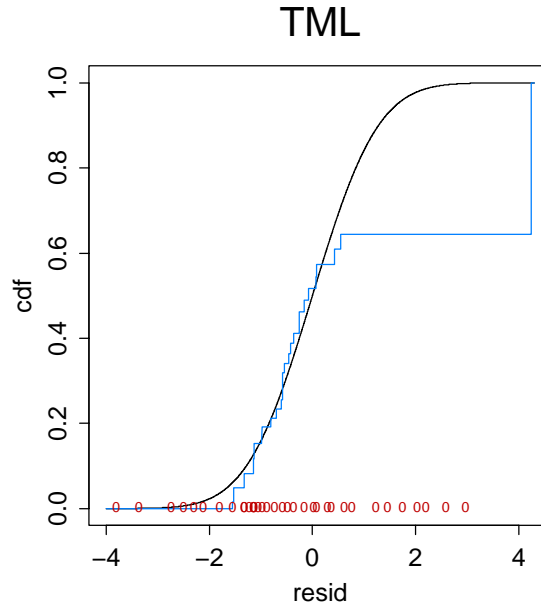
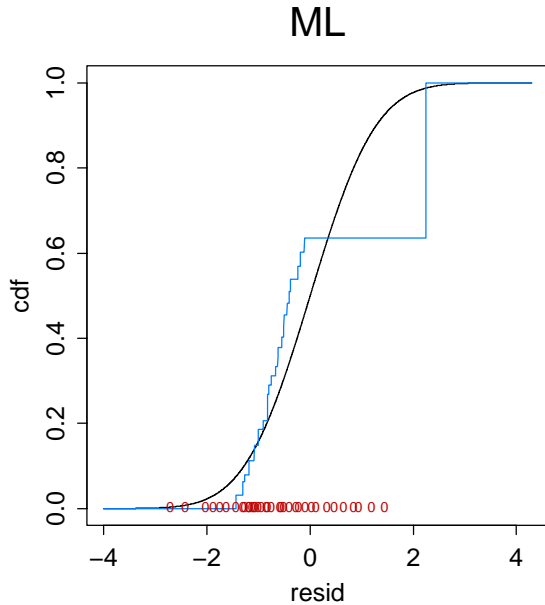


Note: For large times the robust $P(T \geq t_k)$ is small

the term $\hat{E}(U_k | T^* \geq t_k)P(T \geq t_k)$ is small

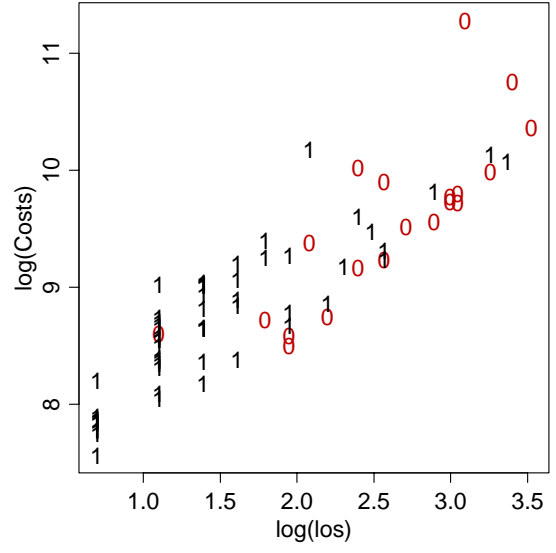
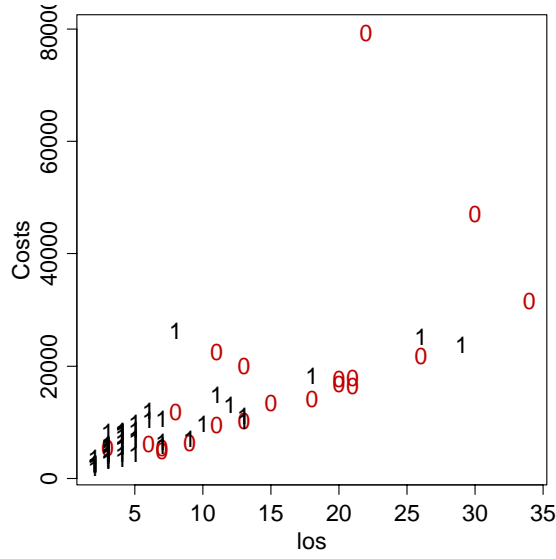
\Rightarrow the estimate of $\hat{E}(U_k | T^* \geq t_k)$ must be precise for small times

	ML	S ^(b)	TML
$\hat{\alpha}$	3.41	2.95	2.94
$\hat{\beta}$	0.45	0.57	0.63
$\hat{\sigma}$	1.02	0.79	0.65
E(Cost Planned)	109 435		57 580
E(Cost Emergency)	77 679		73 413



Example 2

77 malign affections of hepatobiliary system or the pancreas; 20 censored



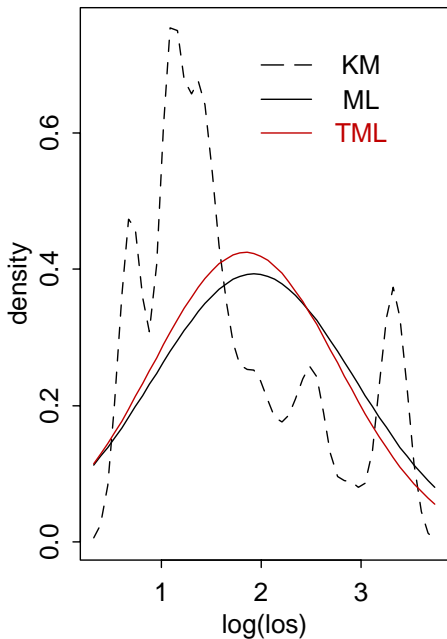
Explore three AFT-models for los: $\log(\text{los}) = \alpha + \sigma \cdot e$,

Lognormal $e \sim N(0, 1)$

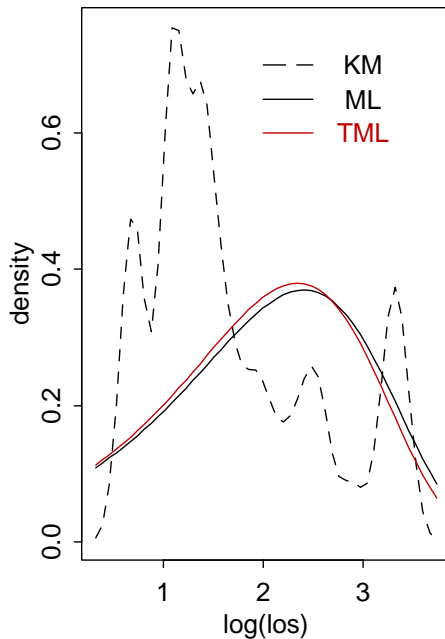
Weibull $e \sim \exp(z \exp(-\exp(z)))$

Gumbel max. $e \sim \exp(-z) \exp(-\exp(-z))$

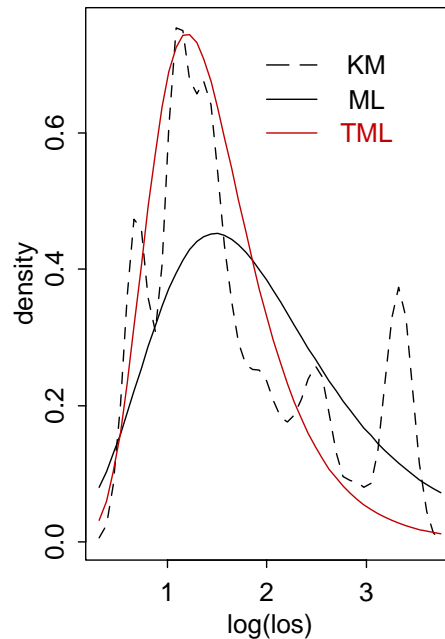
Lognormal

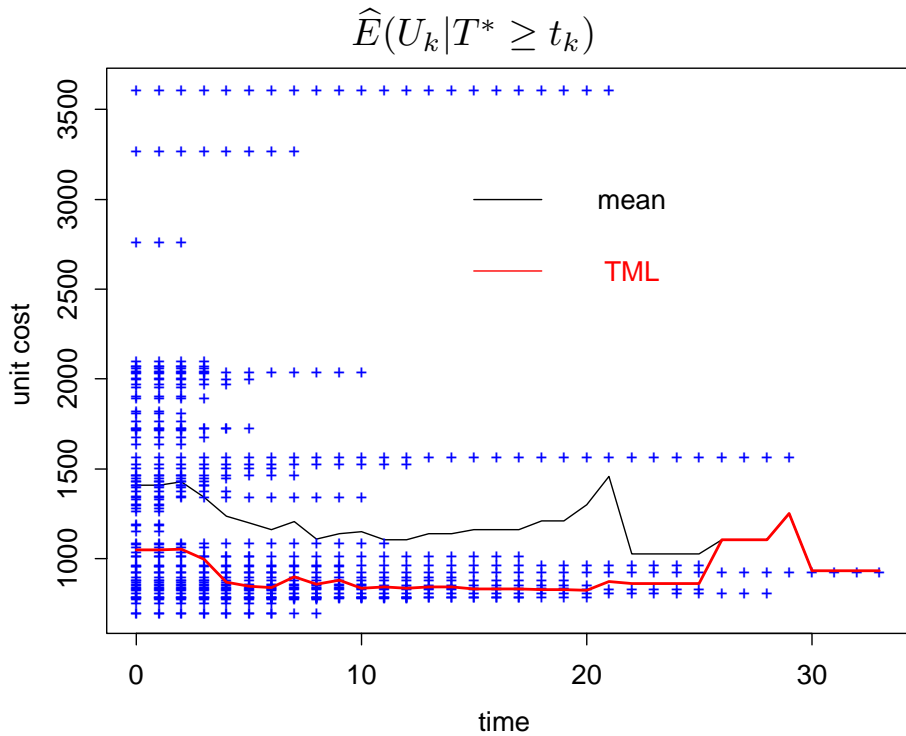


Weibull



Gumbel max.



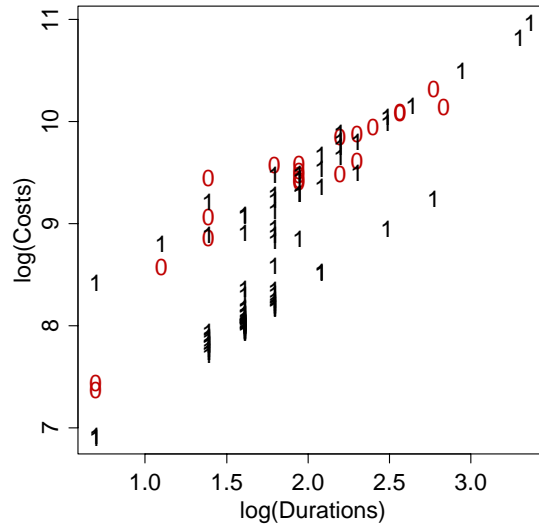
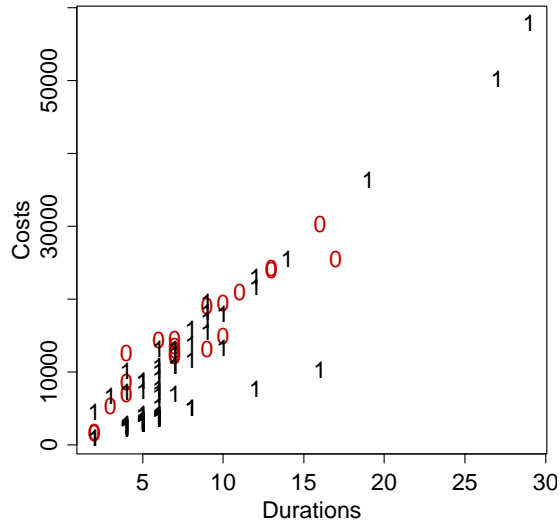


Estimates of the mean total cost

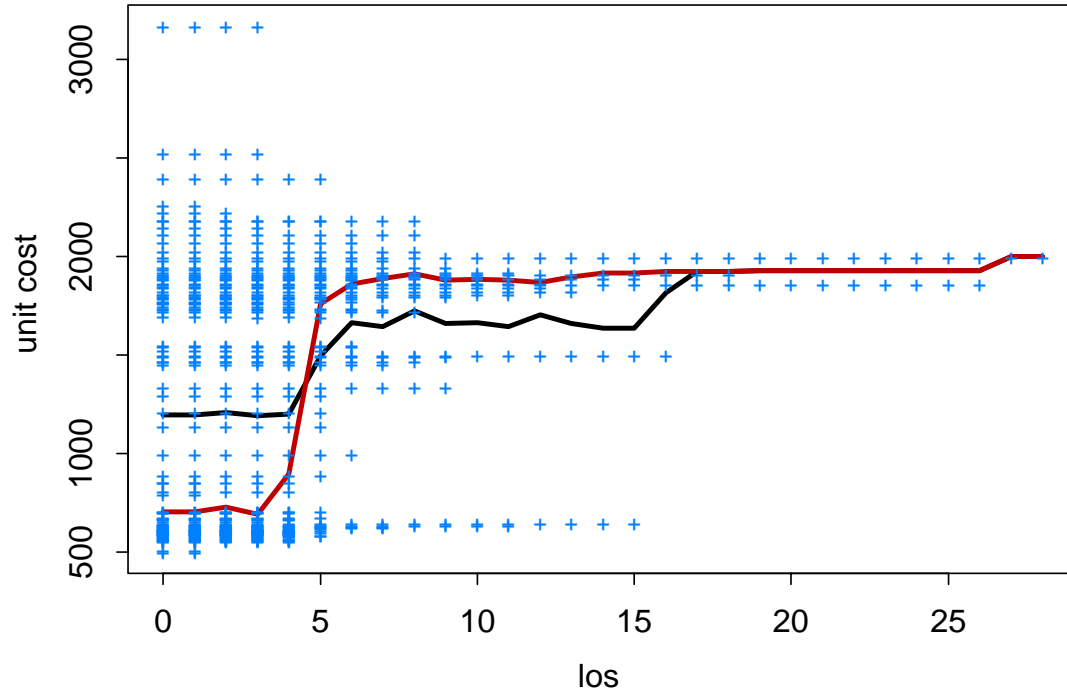
		Time	
		ML	TML
Unit cost	ML	12 745	7 709
	TML	10 084	6 154

Example 3

112 new-born weight > 2.49 Kg, w.out interventions, w. major problems
21 censored



$\hat{E}(U_k | T^* \geq t_k)$ vs los



Tentative conclusions and open problems

- Cost censoring is informative
- Estimation of mean cost with censored data has two components:
 - estimation of the survival (los) distribution
 - estimation of the mean unit cost | survival $> t$
- When the proportion of censoring is high, a full parametric approach is necessary to obtain HBP estimates of AFT-models; KM based estimates are not sufficiently robust.
- Better estimates of the mean total cost (| survival t) should be studied, e.g., using a model for the relationship between Y and T .
- Better estimates of the mean unit cost (| survival $> t$) should be studied, e.g., using a model for the relationship between U and T .