

Truncated maximum likelihood regression with censored responses

Isabella Locatelli

Alfio Marazzi

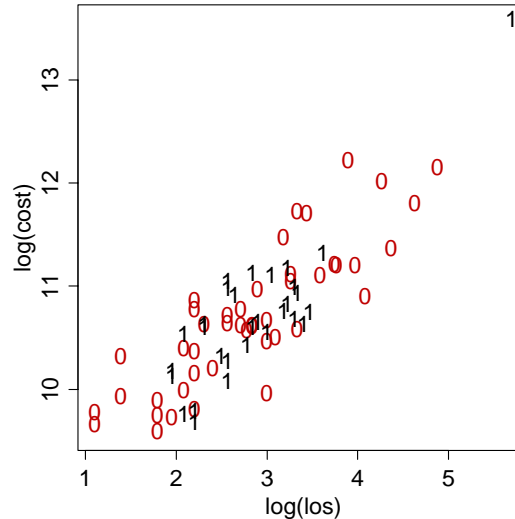
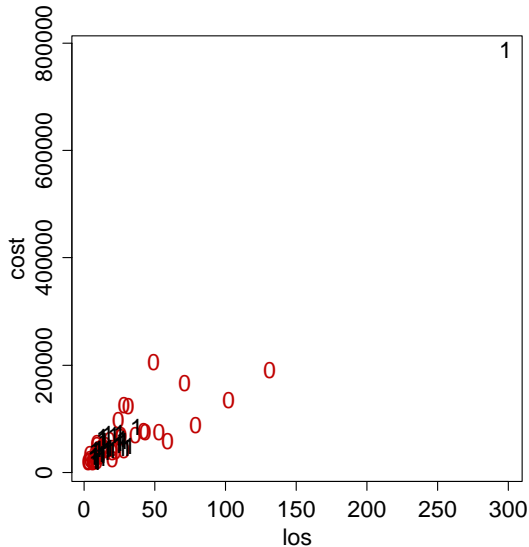
Victor Yohai

Motivation

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

Goal: estimate $E(\text{los}|\text{covariate values})$ and $E(\text{cost}|\text{covariate values})$,
covariates: age, sex, type of admission (emergency, planned), etc

75 stays, major cardiovascular procedures with major cc
Lausanne main hospital, year 2000



- 30 complete stays (1); 45 **censored** (0: transferred)
- Cost and los distributions are **asymmetric**
- There are **outliers**

- In general, cost censoring is *informative*
⇒ standard survival methods cannot be directly used for modeling cost.
- Estimates of the expected total cost must be derived from the survival distribution of los and the distribution of cost per unit of time.

In this talk, we only consider los censoring

KM-type estimates of the survival distribution

Let $y_1, \dots, y_n \sim F$ be a sample of survival times. We observe

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

where c_i are censoring times, independent of y_i .

Problem: define a consistent estimate of F .

Consider the empirical cdf of y :

$$\begin{aligned} F_n(t) &= \frac{1}{n} \left[\sum_{d_i=1} I(y_i < t) + \sum_{d_i=0, y_i^* > t} I(y_i < t) + \sum_{d_i=0, y_i^* \leq t} I(y_i < t) \right] \\ &= \frac{1}{n} \left[\sum_{d_i=1} I(y_i^* < t) + \quad \quad \quad 0 \quad \quad + \quad \quad \quad U \right] \end{aligned}$$

U is unknown !

For $d_i = 0$, $y_i^* \leq t$, replace

$$I(y_i^* < t) \quad \text{by} \quad P(y_i < t | y_i > y_i^*).$$

Then,

$$F_n^*(t) \approx \frac{1}{n} \sum_{d_i=1} I(y_i < t) + \frac{1}{n} \sum_{d_i=0, y_i^* \leq t} P(y_i < t | y_i > y_i^*)$$

(a) Estimate $P(y_i < t | y_i > y_i^*)$ by $(F_n^*(t) - F_n^*(y_i^*)) / (1 - F_n^*(y_i^*))$

$$F_n^*(t) \approx \frac{1}{n} \sum_{d_i=1} I(y_i < t) + \frac{1}{n} \sum_{d_i=0, y_i^* \leq t} \frac{F_n^*(t) - F_n^*(y_i^*)}{1 - F_n^*(y_i^*)}$$

$\Rightarrow F_n^*(y)$ is the Kaplan-Meier **non-parametric** estimate (Efron, 1967).

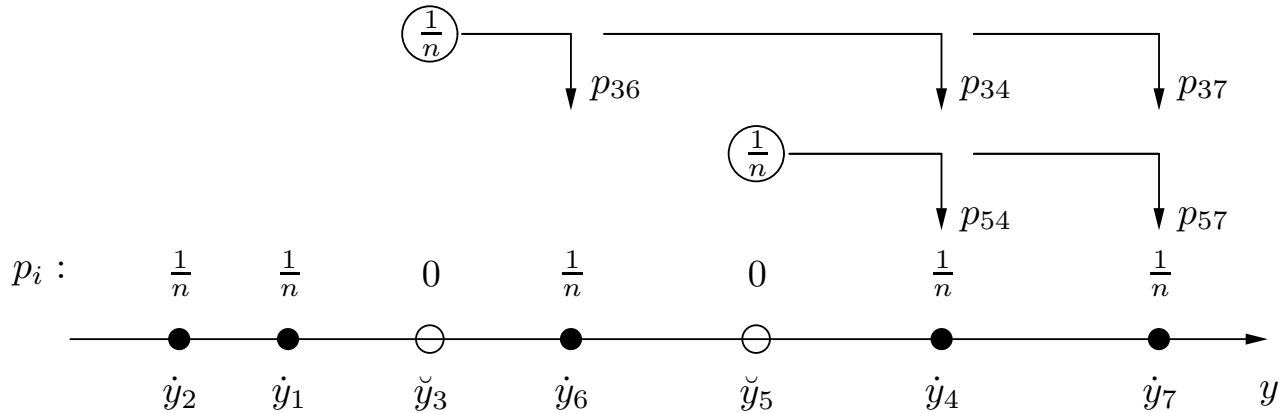
(b) Estimate $P(y_i < t | y_i > y_i^*)$ by $(F_0(t) - F_0(y_i^*)) / (1 - F_0(y_i^*))$
where F_0 is a **parametric** model of F .

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^* .

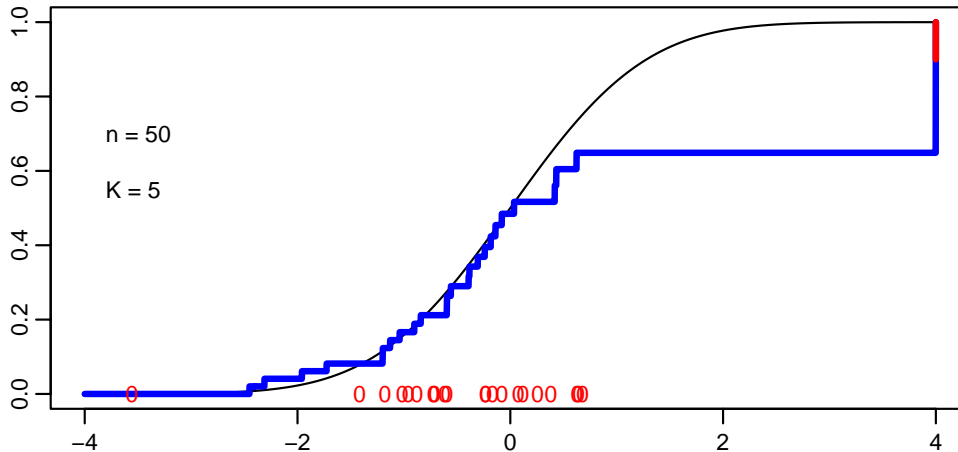


- $p_i = 0$ if $y_i = \check{y}_i$
- the mass $1/n$ of \dot{y}_i is assigned to \dot{y}_i
- the mass $1/n$ of \check{y}_j is distributed among the \dot{y}_i with $\dot{y}_i > \check{y}_j$ with probabilities $p_{ji} = P(y = \dot{y}_i | y > \check{y}_j)$, i.e.,

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

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.

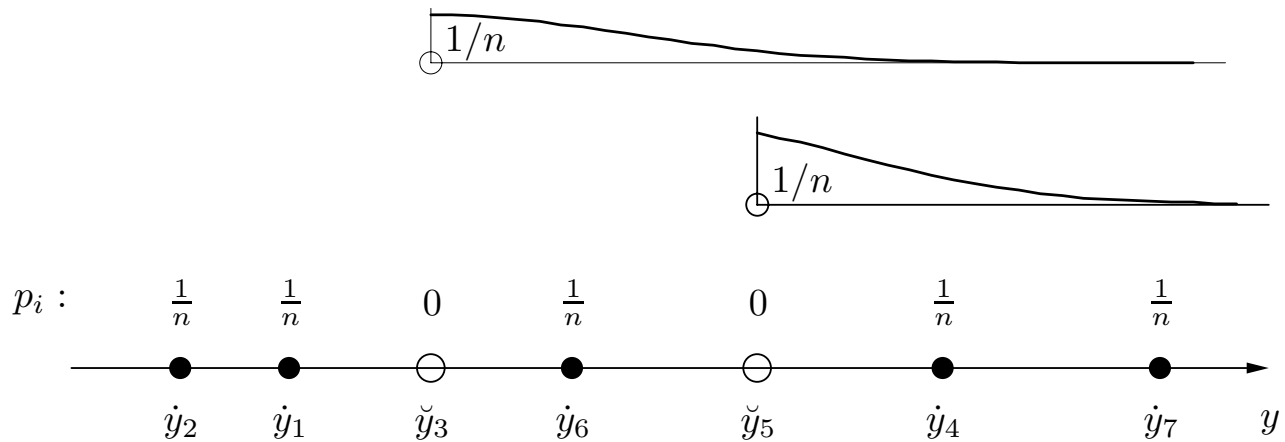


The worst case is when all (J) censored observations are located between the non-censored good observations and K non-censored outliers. Then, the mass assigned to the outliers is $(K + J)/n$.

\Rightarrow 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 $\sum p_i = 1 - m/n$.

Using a continuous parametric model



- the mass $1/n$ of \dot{y}_i is assigned to \dot{y}_i
- the mass $1/n$ of \check{y}_j is distributed over $y > \check{y}_j$ with density $\sim f_0(y) \mid y > \check{y}_j$

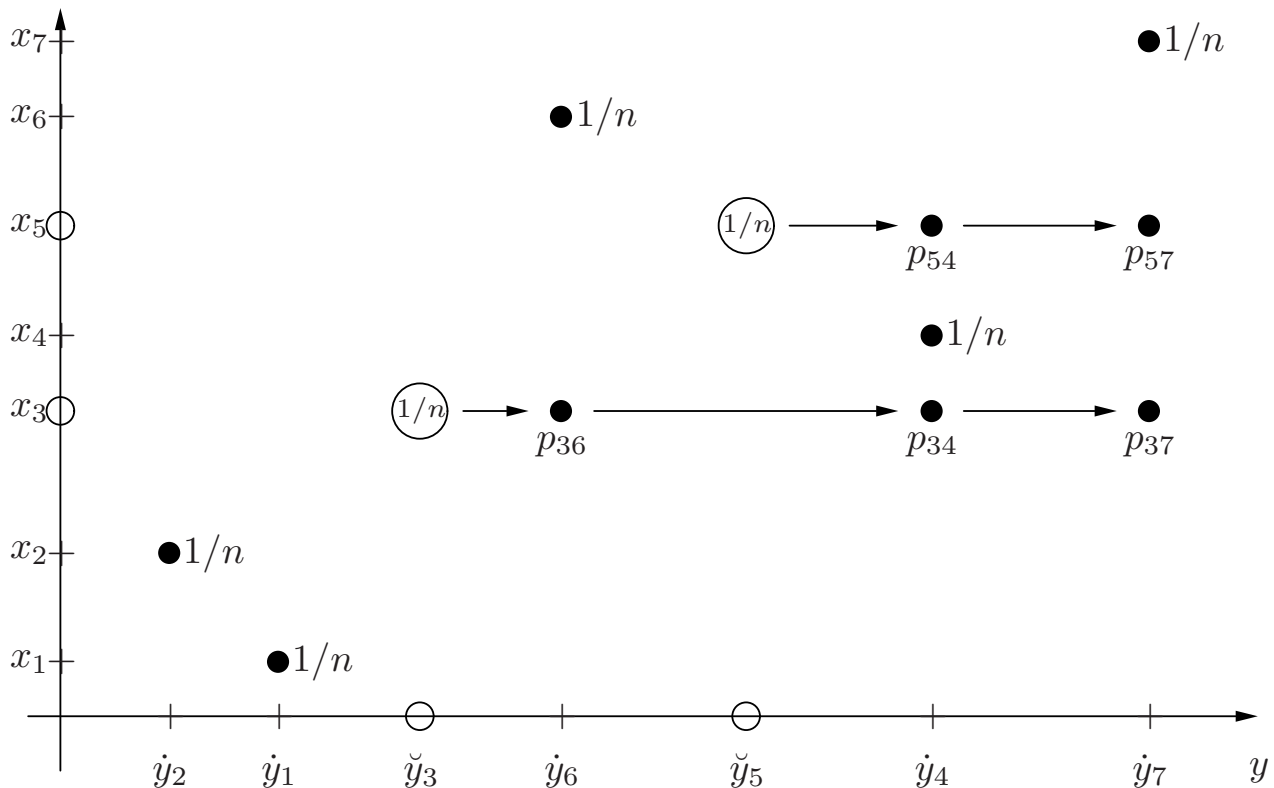
Consequence

- The mass assigned to K outliers is K/n .

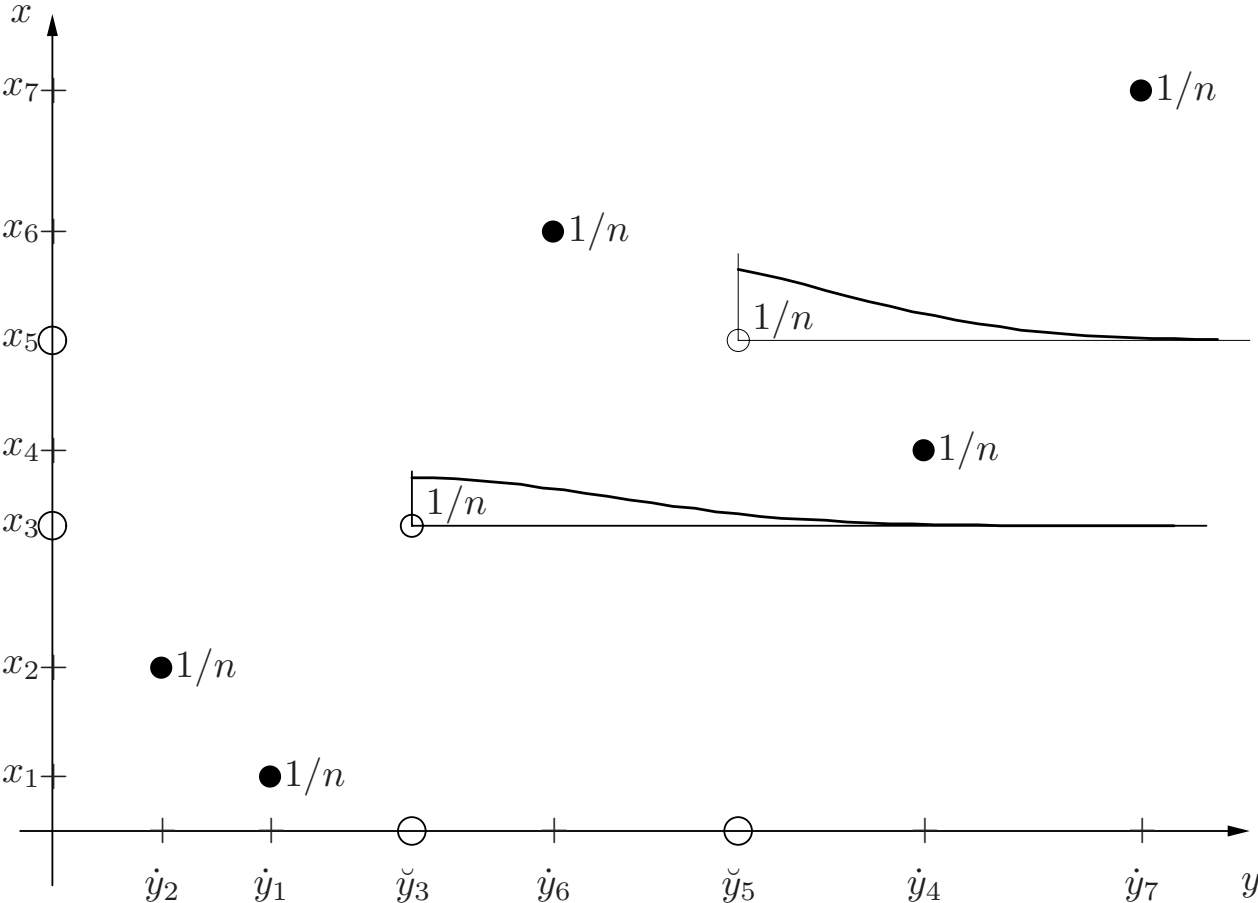
Estimating the joint distribution of y and a regressor x

Let (x_i, y_i) be a sample of $H(x, y) = F(y)G(x)$. We observe (x_i, y_i^*, d_i) .

- (a) Nonparametric consistent estimate $H_n^*(x, y)$ of $H(x, y)$ based on KM (Salibian-Barrera & Yohai, 200?)



(b) Semi-parametric estimate $\hat{H}_n(x, y)$ of $H(x, y)$
 based on conditional densities $\sim f_0(y) \mid y > \check{y}_i$



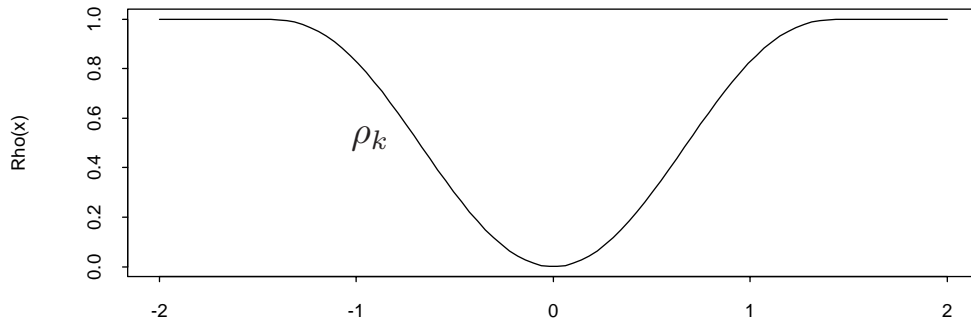
S-estimates : location and scale

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

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

$$u_i \sim F_0 \quad \text{e.g. a parametric model}$$

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

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

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

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

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

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

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

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

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

We observe

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

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

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

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

where

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

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

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

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

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

Note :

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

where $H_{n,\beta}^*$ is the nonparametric estimate of H based on KM and $r_i^*(\beta)$.

- **Consistency of Bukley & James:** James & Smith, 1984; Lai & Ying, 1991
- **Consistency of $H_{n,\beta}^*(x, u)$ 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

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

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

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

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:

$$y_i = x_i^T \beta_0 + \sigma_0 u_i,$$
$$u_i \sim F_0$$

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

Examples

- $y_i \sim \mathcal{N}(x_i^T \beta_0, \sigma_0^2)$ $F_0 = \Phi$;
- y_i is the log of a Lognormal variable, $F_0 = \Phi$;
- y_i is the log of a Weibull variable, F_0 is a stand. Gumbel min. cdf.
- y_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) : $v_i^*(\sigma, \beta) = y_i^* - x_i^T \beta - \sigma m_0$

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

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

$\hat{H}_{n,\sigma,\beta}$ is a semiparametric estimate of $H(u, x)$ based on $v_i^*(\sigma, \beta)$

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

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

We need an estimate of σ_0 for given β :

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

(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 $r_i^* = (y_i^* - x_i^T \tilde{\beta}) / \tilde{\sigma}$
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),

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

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

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

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

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

where

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

if i censored

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

Monte Carlo results

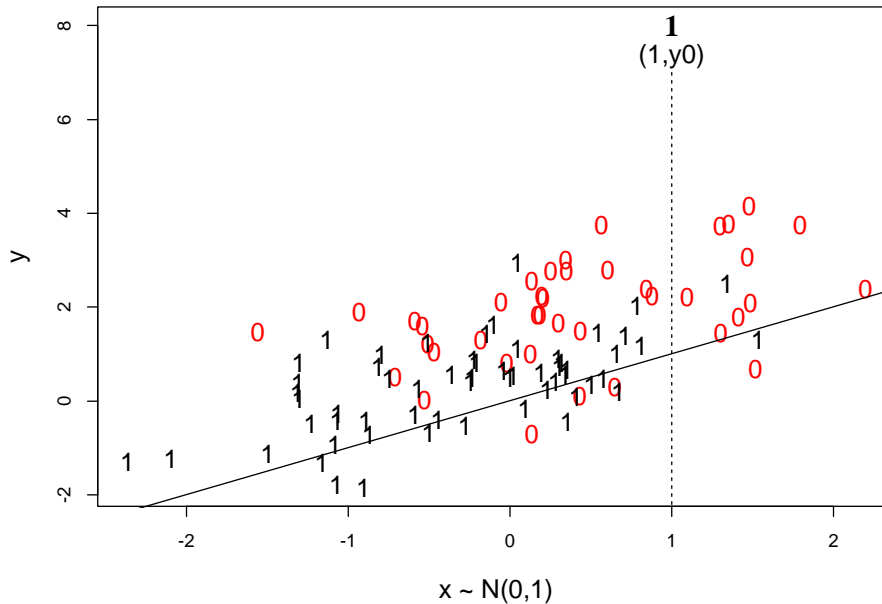
1000 samples from model:

$$y_i = 0 + 1 \cdot x_i + 1 \cdot u_i, \quad i = 1, \dots, 100$$

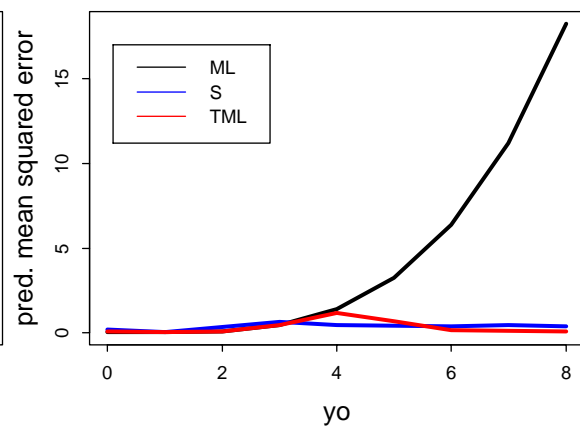
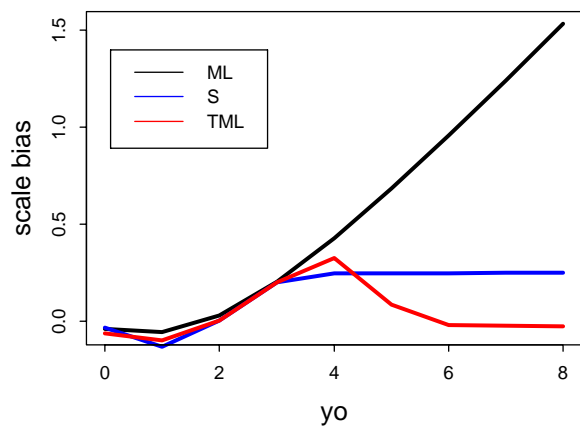
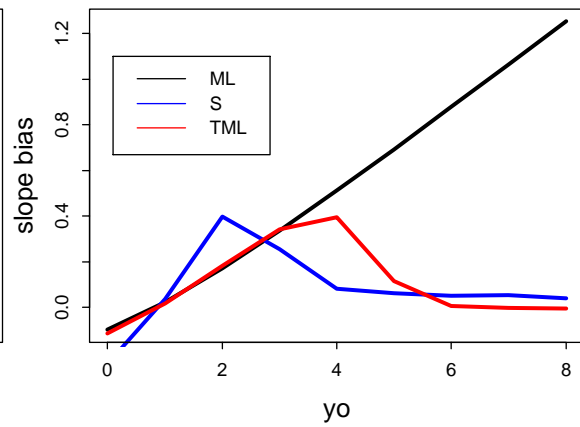
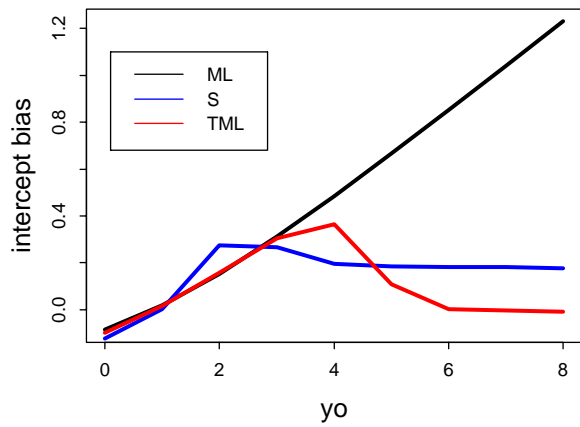
$$x_i \sim N(0, 1), \quad y_i^* = \min(y_i, c_i),$$

$$u_i \sim N(0, 1) \quad \text{or} \quad u_i \sim \text{log-Weibull}, \quad c_i \sim N(1, 1), .$$

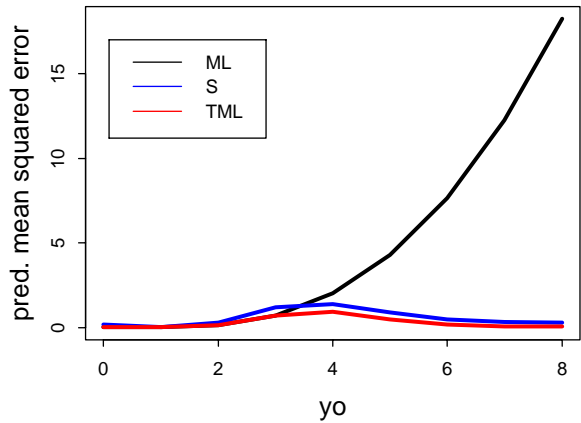
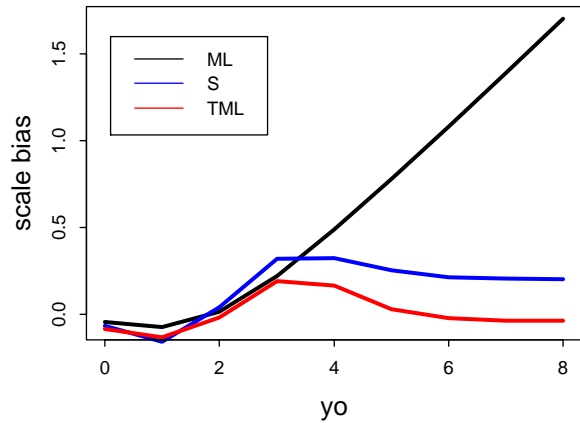
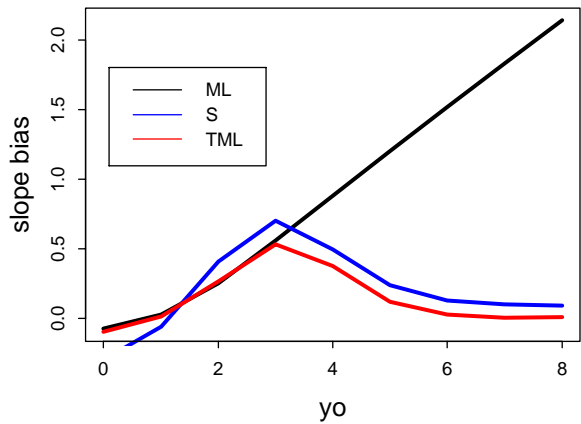
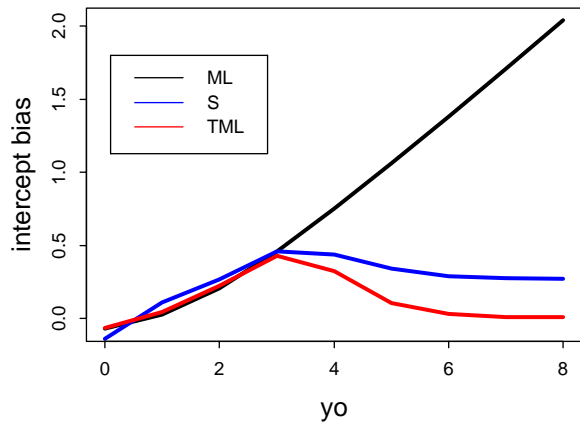
10 outliers in $(1, y_0)$ for $y_0 = 0, 1, 2, \dots, 8$



Normal errors



log-Weibull errors

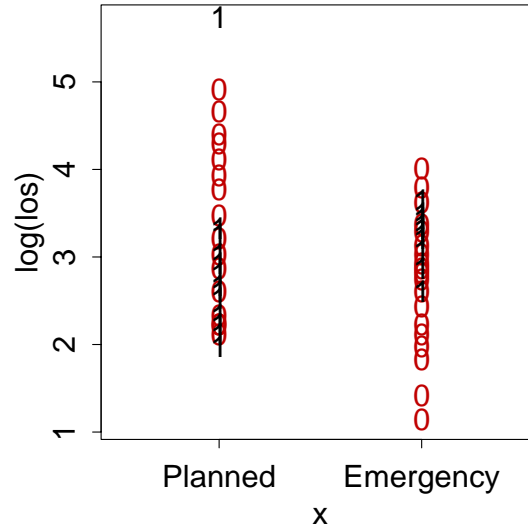


n×variance at nominal model: $u_i \sim N(0, 1)$

| n | Intercept | | | Slope | | | Scale | | |
|------|-----------|-------|-------|-------|-------|-------|-------|-------|-------|
| | ML | S | TML | ML | S | TML | ML | S | TML |
| 20 | 1.388 | 2.495 | 1.899 | 1.706 | 3.794 | 2.258 | 0.816 | 2.501 | 1.901 |
| 50 | 1.356 | 2.085 | 1.460 | 1.507 | 3.756 | 1.651 | 0.811 | 2.164 | 1.208 |
| 100 | 1.226 | 2.056 | 1.296 | 1.399 | 4.178 | 1.454 | 0.730 | 1.917 | 0.966 |
| 200 | 1.247 | 1.916 | 1.258 | 1.481 | 4.524 | 1.584 | 0.721 | 1.891 | 0.995 |
| 500 | 1.358 | 2.004 | 1.392 | 1.301 | 4.302 | 1.383 | 0.727 | 1.844 | 1.132 |
| 1000 | 1.218 | 1.939 | 1.293 | 1.277 | 4.002 | 1.355 | 0.728 | 1.856 | 0.979 |

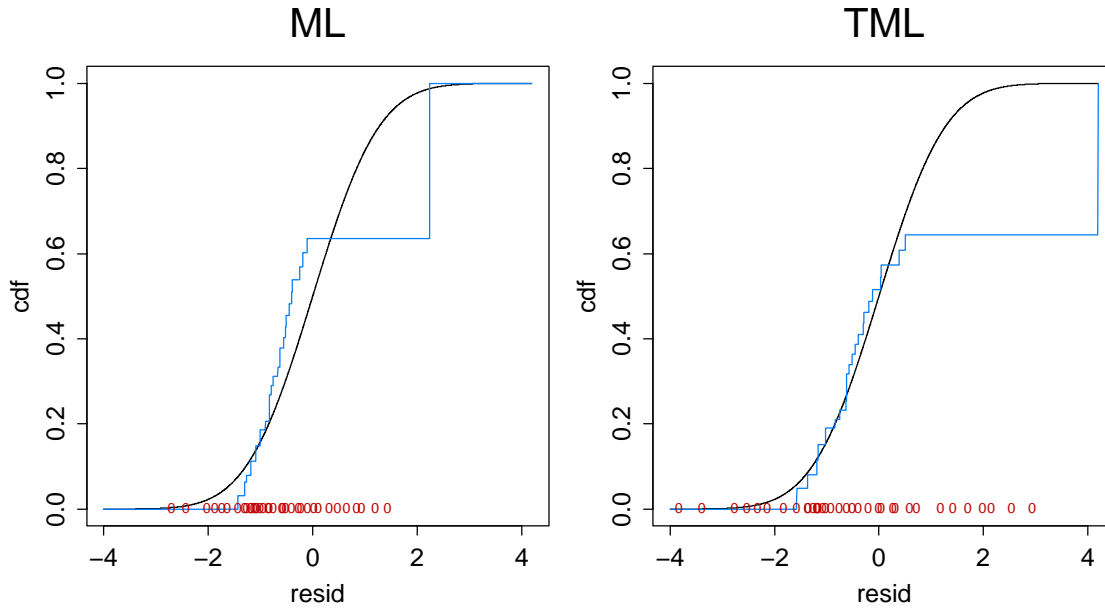
Example

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



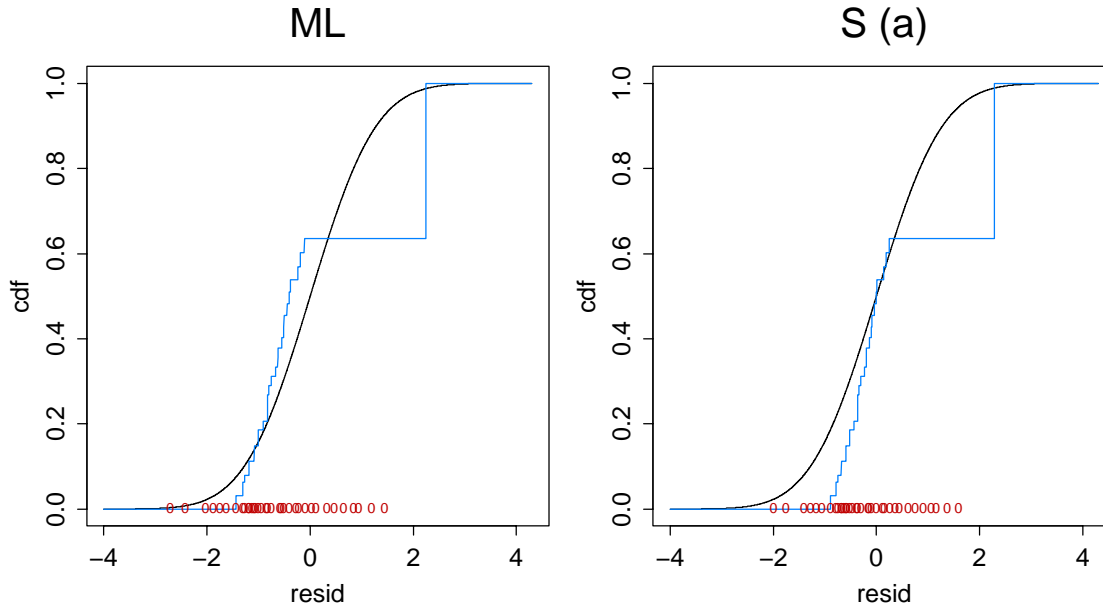
Model: $\log(\text{los}) = \alpha + \beta x + \sigma u, \quad u \sim N(0, 1),$
 $x = 0$: Planned admission,
 $x = 1$: Emergency admission.

| | 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 |



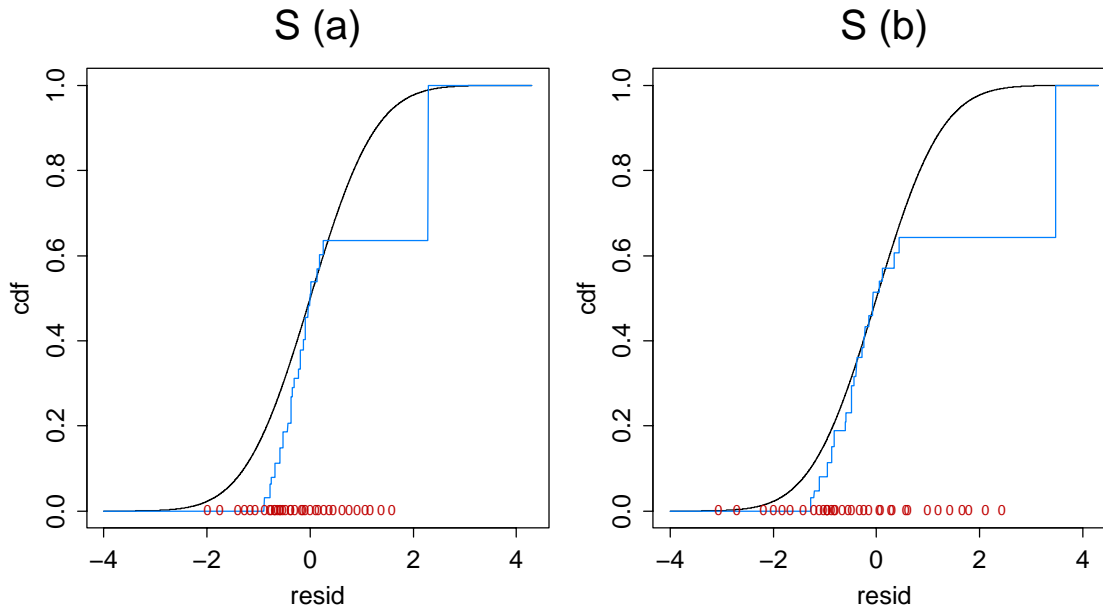
(a) with $\hat{\sigma} = \text{residual MAD}$

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



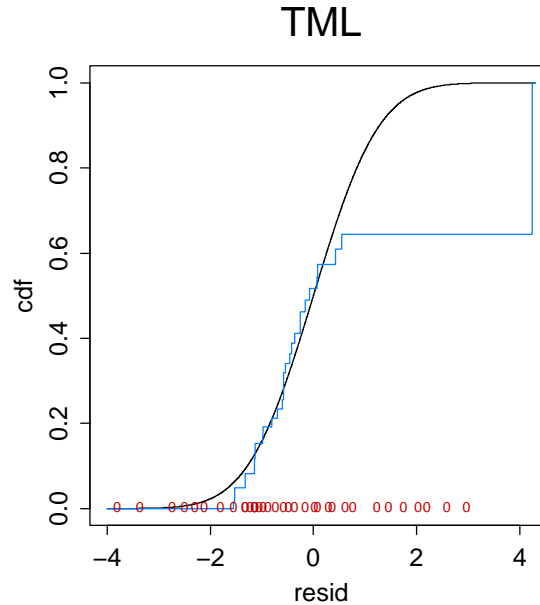
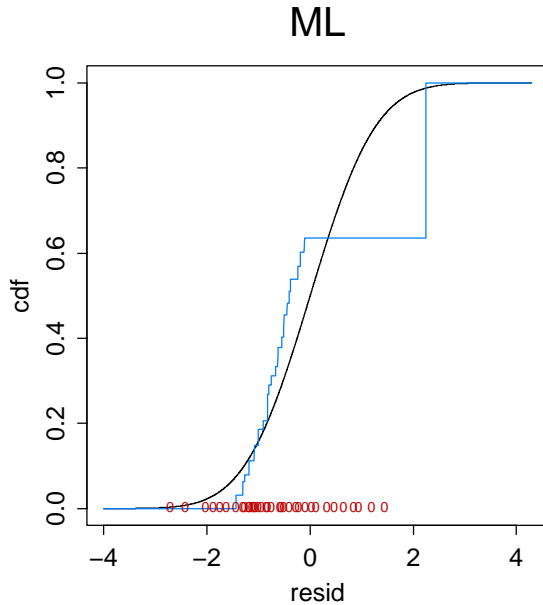
(a) with $\hat{\sigma} = \text{residual MAD}$

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



- (a) with $\hat{\sigma} = \text{residual MAD}$
 (b) with $\tilde{\sigma} = S(\tilde{\sigma}, \tilde{\beta}(\tilde{\sigma}), 0)/s_0$

| | 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 |



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