Computational treatment of the error distribution in nonparametric regression with right-censored and selection-biased data

Géraldine LAURENT
Jointly with Cédric HEUCHENNE

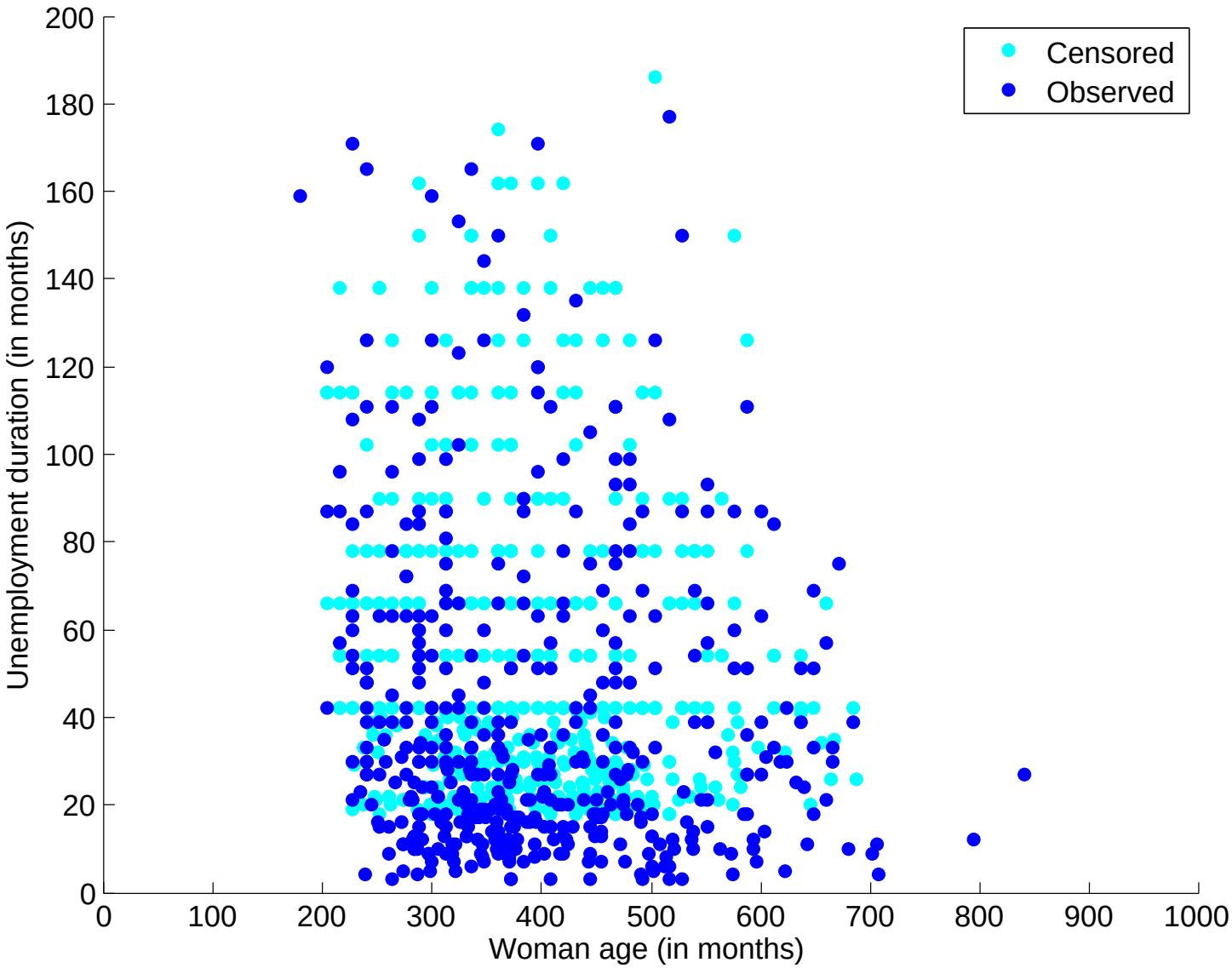
QuantOM, HEC-ULg Management School-University of Liege

Tuesday, 24 August 2010

The Spanish Institute for Statistics studied between 1987 and 1997 the unemployment of active people, and more especially the married women.

For these data, we note that

- the time of unemployment will not be completely observed,
- the age of the woman acts on the future job.


Censored
Observed
Unemployment duration (in months)
Woman age (in months)
0
20
40
60
80
100
120
140
160
180
200
0
100
200
300
400
500
600
700
800
900
1000

Estimation

We consider the nonparametric regression model

$$Y = m(X) + \sigma(X)\varepsilon$$

where

- Y is the response variable
- X is the covariate
- $m(\cdot) = E[Y|\cdot]$ and $\sigma^2(\cdot) = Var[Y|\cdot]$ are unknown smooth functions
- ε is independent of X, with $E[\varepsilon] = 0$ and $Var[\varepsilon] = 1$

Particularity of (X, Y)

- (X, Y) is obtained from cross-sectional sampling
- Y is subject to right censoring.

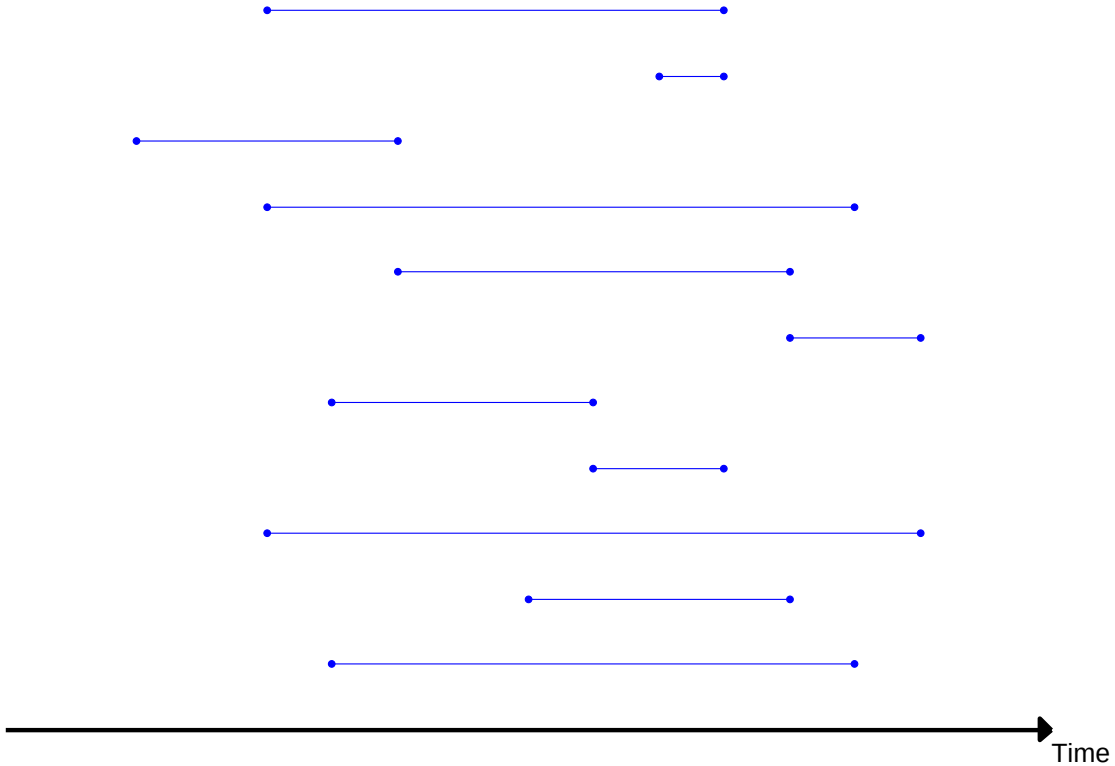
We study the variable Y delimited by

$$T \leq Y \leq C$$

where

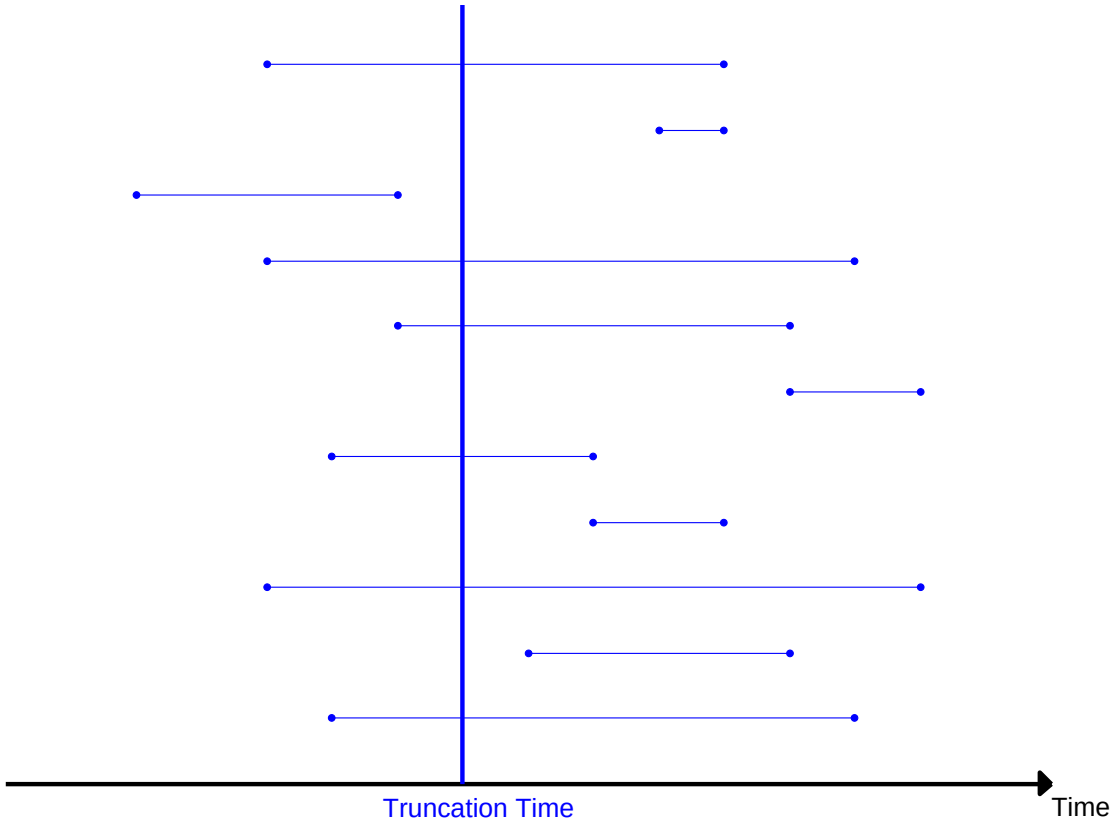
- T is the truncation variable
- C is the censoring variable.

Real World



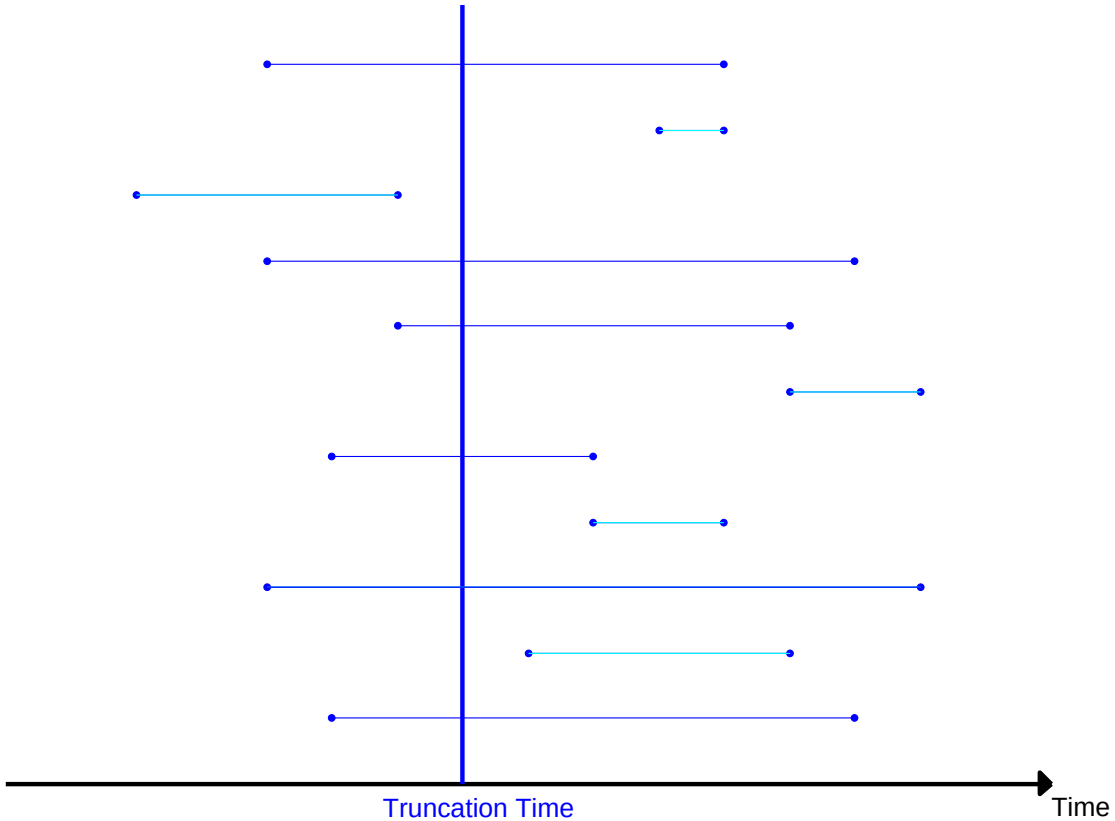
We use as notation F for cdf

Real World



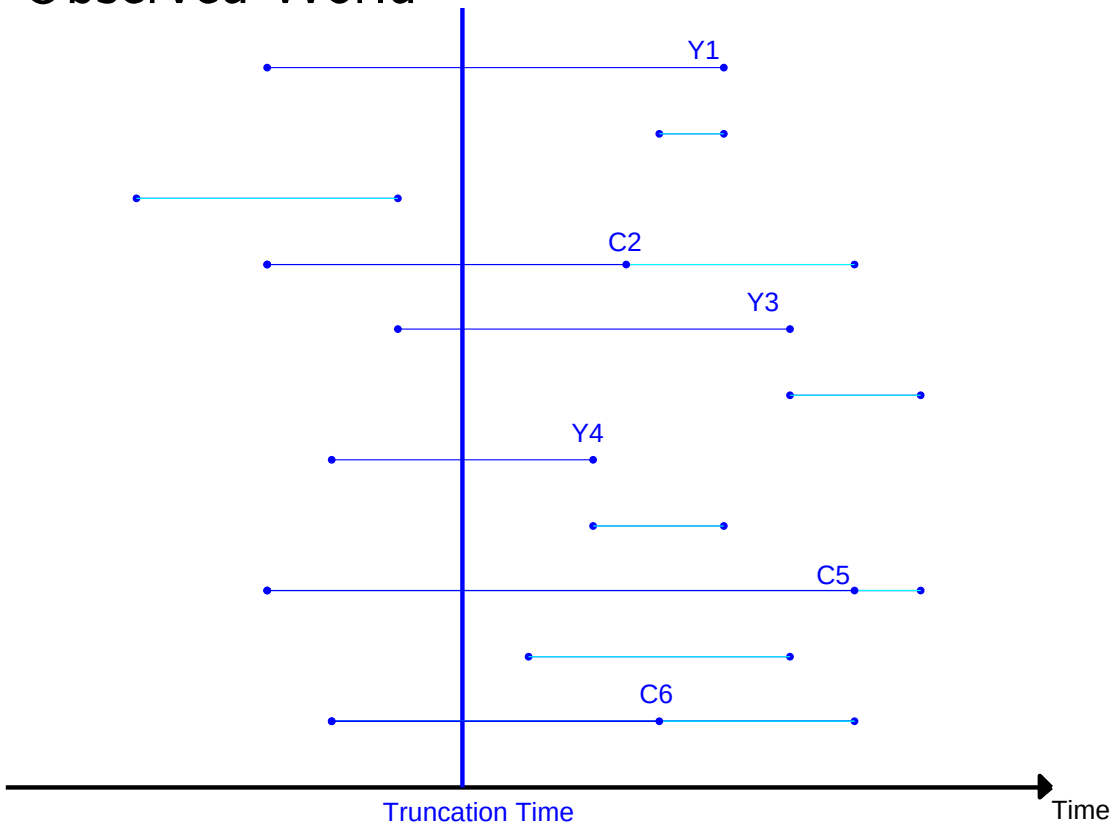
We use as notation F for cdf

Real World



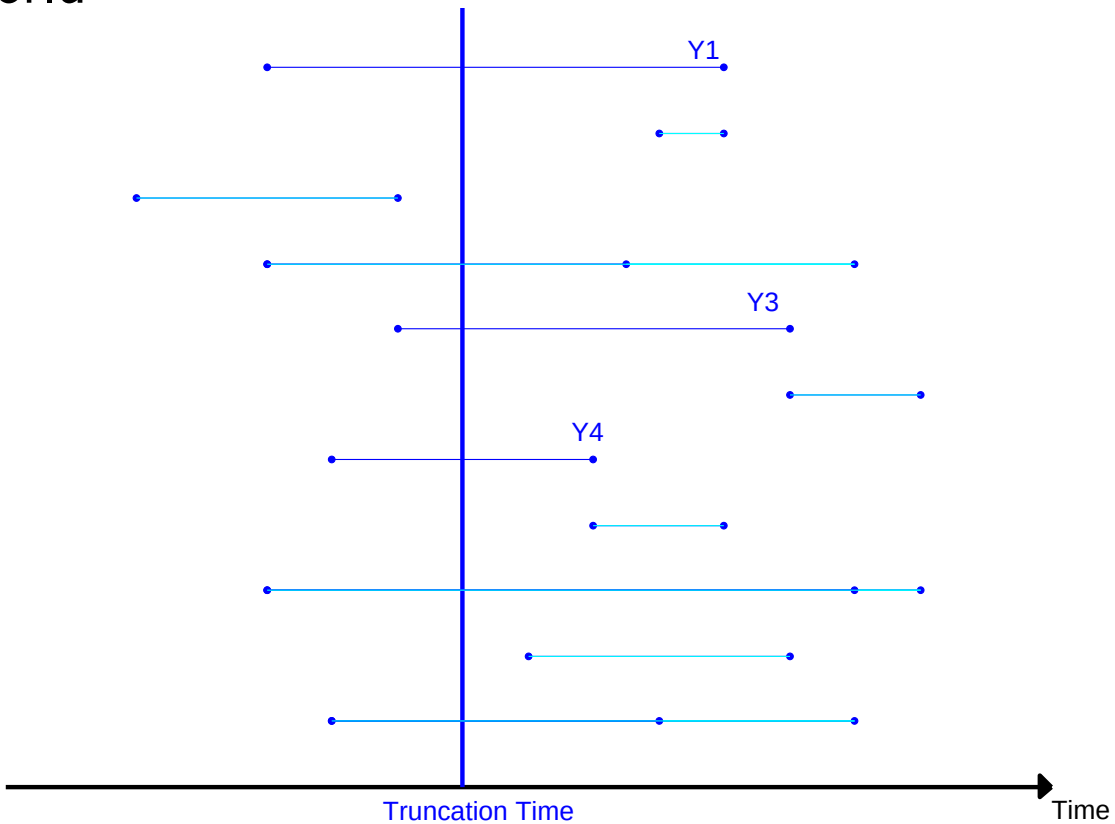
We use as notation F for cdf

Intermediate Observed World



We use as notation $\mathcal{H}$ for cdf, n the sample size

Observed World



We use as notation H for cdf

Aim : Estimation of the error distribution

$$F_\varepsilon(e) = \mathbf{P}(\varepsilon \leq e)$$

with

$$(X, Y) \text{ where } T \leq Y \leq C$$

where

- the distribution $F_{T|X}$ is a parametric distribution
- the distribution $F_{C-T|X}$ is completely unknown

Assumptions:

- the variables Y and T are independent, conditionally on X
- for each value x, the support of $F_{Y|X}(\cdot|x)$ is included into the support of $F_{T|X}(\cdot|x)$
- the lower bound of the T support is zero
- the variables (T, Y) and $C - T$ are independent, conditionally on $T \leq Y$, X

We have

$$\begin{aligned} H_{X,Y}(x,y) &= \mathbf{P}(X \leq x, Y \leq y | T \leq Y \leq C) \\ &= (E[w(X,Y)])^{-1} \int_{r \leq x} \int_{s \leq y} w(r,s) dF_{X,Y}(r,s), \end{aligned}$$

the weight function $w(x,y)$ is defined by

$$w(x,y) = \int_{t \leq y} \{1 - \mathcal{G}(y-t|x)\} \, dF_{T|X}(t|x)$$

where $\mathcal{G}(z|x) = \mathbf{P}(C - T \leq z | X = x, T \leq Y)$.

In particular, if $C = T + \tau$ where τ is a positive constant, the weight function is

$$w(x, y) = \int_{0 \vee y - \tau}^{y} dF_{T|X}(t|x)$$

by applying the same procedure.

We obtain

$$F_{X,Y}(x,y) = \int_{r\leq x}\int_{s\leq y} \frac{E[w(X,Y)]}{w(r,s)} dH_{X,Y}(r,s)$$

Therefore,

$$\begin{aligned} F_{\varepsilon}(e) &= \mathbf{P}\left(\frac{Y-m(X)}{\sigma(X)} \leq e\right) \\ &= \iint_{\left\{(x,y):\frac{y-m(x)}{\sigma(x)}\leq e\right\}} dF_{X,Y}(x,y) \\ &= \iint_{\left\{(x,y):\frac{y-m(x)}{\sigma(x)}\leq e\right\}} \frac{E[w(X,Y)]}{w(x,y)} dH_{X,Y}(x,y) \end{aligned}$$

Thus, the estimator is

$$\hat{F}_{\varepsilon}(e) = \frac{1}{M} \sum_{i=1}^{n} \frac{\hat{E}[w(X,Y)]}{\hat{w}(X_i, Y_i)} I\{\hat{\varepsilon}_i \leq e, \Delta_i = 1\}$$

with

$$\hat{\varepsilon}_i = \frac{Y_i - \hat{m}(X_i)}{\hat{\sigma}(X_i)}, \qquad M = \sum_{i=1}^{n} \Delta_i,$$

$$\hat{E}[w(X,Y)] = \left(\frac{1}{M} \sum_{i=1}^{n} \frac{\Delta_i}{\hat{w}(X_i, Y_i)} \right)^{-1}$$

where the functions $\hat{m}(\cdot)$, $\hat{\sigma}(\cdot)$ and $\hat{w}(\cdot,\cdot)$ are nonparametric estimators.

For $\mathcal{G}(t|x)$, we use the Beran (1981) estimator defined by

$$\hat{\mathcal{G}}(t|x) = 1 - \prod_{Z_i \leq t, \Delta_i = 0} \left(1 - \frac{W_i(x, h_n)}{\sum_{j=1}^{n} W_j(x, h_n) I\{Z_j \geq Z_i\}} \right)$$

where

- $Z_i = \min(C_i - T_i, Y_i - T_i)$ and $\Delta_i = I\{Y_i \leq C_i\}$
- $W_i(x, h_n) = \frac{K\left(\frac{x - X_i}{h_n}\right)}{\sum_{j=1}^{n} K\left(\frac{x - X_j}{h_n}\right)}$ are the Nadaraya-Watson weights
- K is a kernel function
- h_n is a bandwidth sequence tending to 0 when $n \to \infty$

$$=> \hat{w}(x, y) = \int_{t \leq y} \left\{ 1 - \hat{\mathcal{G}}(y - t|x) \right\} dF_{T|X}(t|x)$$

The estimators of $m(\cdot)$ and $\sigma(\cdot)$ are given by

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} \frac{W_i(x,h_n) Y_i \Delta_i}{\hat{w}(x,Y_i)}}{\sum_{i=1}^{n} \frac{W_i(x,h_n) \Delta_i}{\hat{w}(x,Y_i)}},$$

$$\hat{\sigma}^2(x) = \frac{\sum_{i=1}^{n} \frac{W_i(x,h_n) \Delta_i (Y_i - \hat{m}(x))^2}{\hat{w}(x,Y_i)}}{\sum_{i=1}^{n} \frac{W_i(x,h_n) \Delta_i}{\hat{w}(x,Y_i)}},$$

extension of the estimators in de Uña-Alvarez and Iglesias-Pérez (2008).

Asymptotic results

Under some assumptions,

$$\hat{F}_\varepsilon(e) - F_\varepsilon(e) = \sum_{i=1}^{n} V(X_i, Y_i, Z_i, \Delta_i, e) + o_p(n^{\frac{-1}{2}})$$

uniformly in e.
$=>$ Weak convergence of the process

$$\sqrt{n}(\hat{F}_\varepsilon(e) - F_\varepsilon(e)) \rightarrow \Omega(e)$$

where Ω is a Gaussian process with zero mean and complex covariance.

Bandwidth selection

We want to determine the smoothing parameter h_n which minimizes

$$MISE = E\left[\int \left\{\hat{F}_{\varepsilon,h_n}(e) - F_\varepsilon(e)\right\}^2 de\right]$$

We consider bootstrap procedure which is an extension of Li and Datta (2001).

For $b = 1, \ldots, B$,
For $i = 1, \ldots, n$

Step 1 Generate $X^*_{i,b}$ from

$$\hat{F}_X(\cdot) = \sum_{j=1}^{n} \frac{\hat{E}[w(X,Y)]}{\hat{E}[w(X,Y)|X=\cdot]} I\{X_j \leq \cdot, \Delta_j = 1\},$$

where $\hat{E}[w(X,Y)|X=\cdot] = \sum_{j=1}^{n} W_j(\cdot, g_n)\Delta_j / \sum_{j=1}^{n} \frac{W_j(\cdot, g_n)\Delta_j}{\hat{w}(\cdot, Y_j)}$

and g_n is a pilot bandwidth

Step 2 Generate $Y^*_{i,b}$ from

$$\hat{F}_{Y|X}(\cdot|X^*_{i,b}) = \sum_{j=1}^{n} \frac{\hat{E}[w(X,Y)|X = X^*_{i,b}]W_j(X^*_{i,b}, g_n)}{\hat{w}(X^*_{i,b}, Y_j)(\sum_{k=1}^{n} W_k(X^*_{i,b}, g_n)\Delta_k)} I\{Y_j \leq \cdot, \Delta_j = 1\}$$

Step 3 Draw $T^*_{i,b}$ from the distribution $F_{T|X}(\cdot|X^*_{i,b})$.

- If $T^*_{i,b} > Y^*_{i,b}$, then reject $(X^*_{i,b}, Y^*_{i,b}, T^*_{i,b})$ and go to Step 1.
- Otherwise, go to Step 4.

Step 4 Select at random $V^*_{i,b}$ from $\hat{\mathcal{G}}(\cdot|X^*_{i,b})$ calculated with g_n

Step 5 Define

- $Z^*_{i,b} = \min(Y^*_{i,b} - T^*_{i,b}, V^*_{i,b})$
- $\Delta^*_{i,b} = I\{Y^*_{i,b} - T^*_{i,b} \leq V^*_{i,b}\}$.

Compute $\hat{F}^*_{\varepsilon,h_n,b}$, the error distribution based on

- bandwidth h_n
- resample $\{(X^*_{i,b}, T^*_{i,b}, Z^*_{i,b}, \Delta^*_{i,b}) : i = 1, \ldots, n\}$.

The expression of the MISE can be approximated by

$$\text{argmin}_{h_n} B^{-1} \sum_{b=1}^{B} \int \{\hat{F}^*_{\varepsilon,h_n,b}(e) - \hat{F}_{\varepsilon,g_n}(e)\}^2 de.$$

Simulations

We consider

- model $Y = X + \varepsilon$ where
  - $X \sim U([1.7321; 2])$
  - $\varepsilon \sim U\left(\left[-\sqrt{3}; \sqrt{3}\right]\right)$
- model $\log Y = X + \varepsilon$ where
  - $X \sim U([0; 1])$
  - $\varepsilon \sim N(0; 1)$
- model $Y = X^2 + X * \varepsilon$ where
  - $X \sim U\left(\left[2; 2 * \sqrt{3}\right]\right)$
  - $\varepsilon \sim U\left(\left[-\sqrt{3}; \sqrt{3}\right]\right)$
- model $\log Y = X^2 + X * \varepsilon$ where
  - $X \sim U([0; 1])$
  - $\varepsilon \sim N(0; 1)$

where X and ε are independent in each model

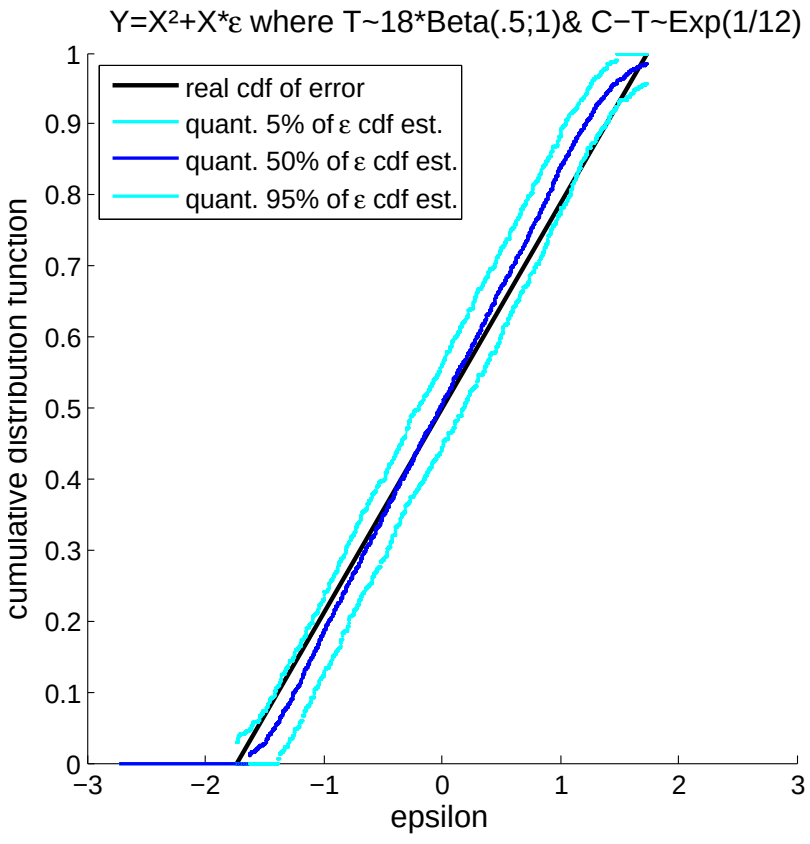
Homoscedastic model : $Y = X + \varepsilon$

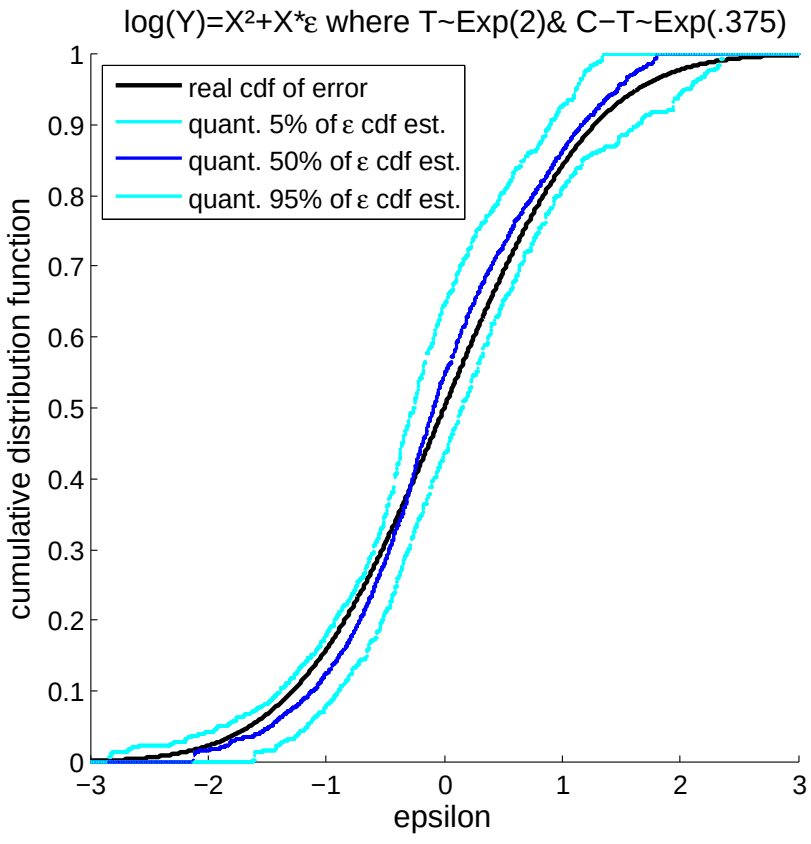
Dist. of T	Dist. of $C - T$	% Censor.	$\widehat{MISE}$ $(*10^{-3})$
Unif$([0; 4])$	Exp$(2/5)$	0.37	5.5
Unif$([0; 4])$	Exp$(2/7)$	0.28	4.9
Unif$([0; X + 2])$	Exp$(2/5)$	0.36	5.2
Unif$([0; X + 2])$	Exp$(2/7)$	0.29	5.0
$4 *$ Beta$(0.5; 1)$	Exp$(2/7)$	0.34	4.2
$4 *$ Beta$(0.5; 1)$	Exp$(2/9)$	0.29	4.0
Unif$([0; 4])$	Exp$(1/(X + 1.5))$	0.28	4.6
$4 *$ Beta$(0.5; 1)$	Exp$(1/(X^2 - 1))$	0.34	4.5

Heteroscedastic model : $Y = X^2 + X * \varepsilon$

Dist. of T	Dist. of $C - T$	% Censor.	$\widehat{MISE}$ $(*10^{-3})$
Unif$([0; 18])$	Exp(0.1)	0.34	6.9
Unif$([0; 18])$	Exp(0.05)	0.19	6.2
$18 *$ Beta$(0.5; 1)$	Exp$(1/12)$	0.35	6.3
$18 *$ Beta$(0.5; 1)$	Exp$(1/15)$	0.29	6.2
Unif$([0; X + 16])$	Exp$(1/12)$	0.29	6.2
$18 *$ Beta$(0.5; 1)$	Exp$(1/(2X^2 - 1))$	0.30	6.6

Interval containing 90% of value of $\hat{F}_\varepsilon$





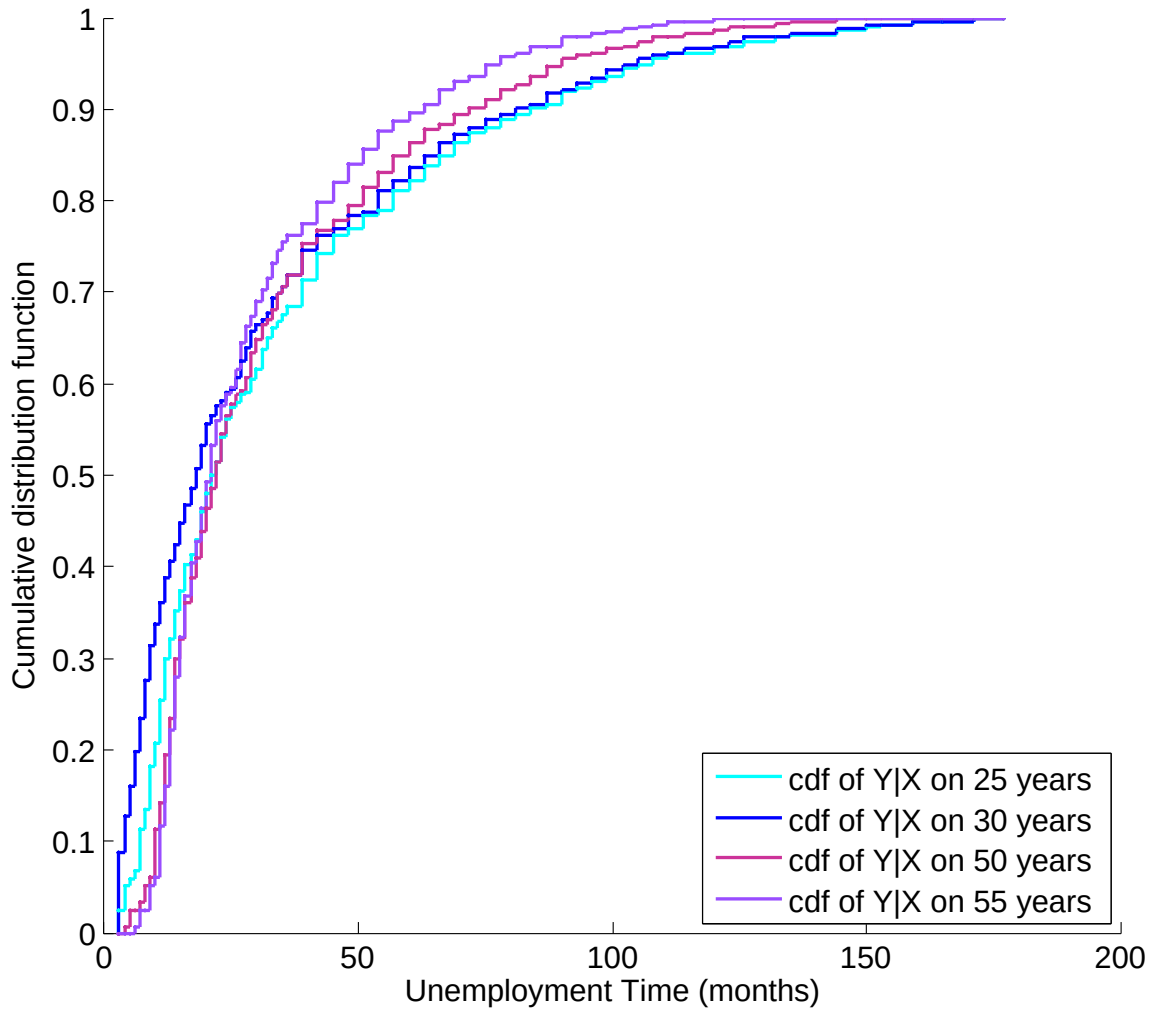
Data analysis

For the real data, we suppose that

- the number of time periods is equal to 1009 but only 446 aren't censored.
- the distribution of T is a uniform one (Wang, 1991);
- the variable C is defined by $C = T + \tau$ where τ is a constant equal to 18 months;

The Bootstrap approximation gives the value of 70 months as the optimal bandwidth.

Representation of $\hat{F}_{Y|X}$ for various ages.



Thank you for your attention

- ASGHARIAN, M., M´LAN, C. E., WOLFSON, D. B. (2002): Length-biased sampling with right-censoring: an unconditional approach. *Journal of the American Statististical Association 97, 201-209*.
- BERAN, R. (1981): *Nonparametric regression with randomly censored survival data*. Technical Report, University of California, Berkeley.
- de UNA-ALVAREZ, J., IGLESIAS-PEREZ, M.C. (2008): Nonparametric estimation of a conditional distribution from length-biased data. *Annals of the Institute of Statistical Mathematics, in press. doi: 10.1007/s10463-008-0178-0*.
- LI, G., DATTA, S. (2001): A bootstrap approach to non-parametric regression for right censored data. *Annals of the Institute of Statistical Mathematics 53, 708-729*.

- OJEDA-CABRERA, J.L., VAN KEILEGOM, I. (2008): Goodness-of-fit tests for parametric regression with selection biased data. *Journal of Statistical Planning and Inference 139 (8), 2836-2850*.
- VAN KEILEGOM, I., AKRITAS, M.G. (1999): Transfer of tail information in censored regression models. *The annals of Statistics 27 (5), 1745-1784*.
- WANG, M.-C. (1991): Nonparametric estimation from cross-sectional survival data. *Journal of the American Statistical Association 86, 130-143*.