The copula-graphic estimator in censored nonparametric location-scale regression models

Aleksandar Sujica, Ingrid Van Keilegom

Université catholique de Louvain

YRD 2013, February 1, LLN

Outline

Outline

- Survival analysis
- Informative censoring
- Copulas
- Location-scale regression

Survival analysis

Example

Y = time to first relapse of a cancer (survival time)

X = age of patient

How to estimate $F(y|x)$?

Survival analysis

Example

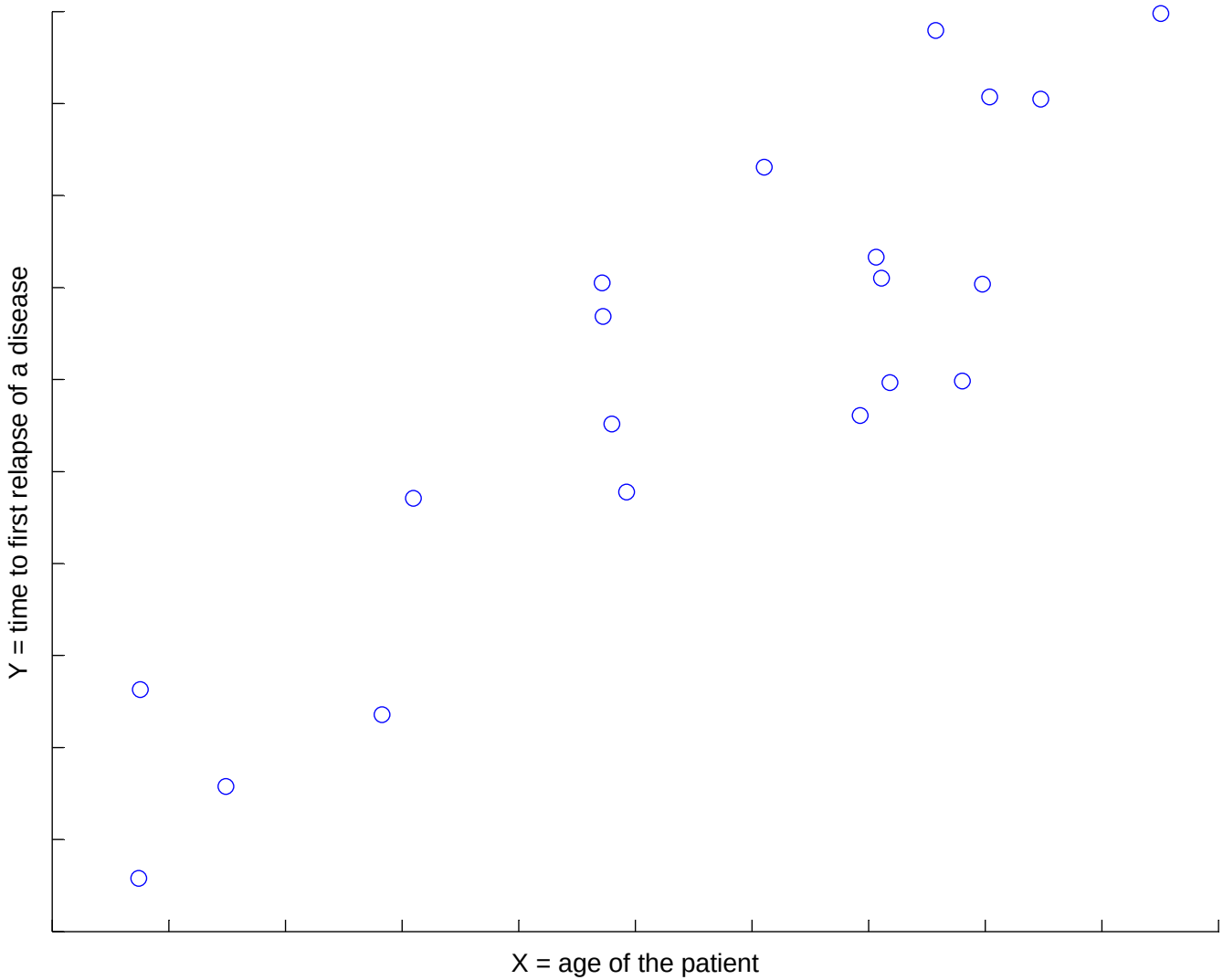
Y = time to first relapse of a cancer (survival time)

X = age of patient

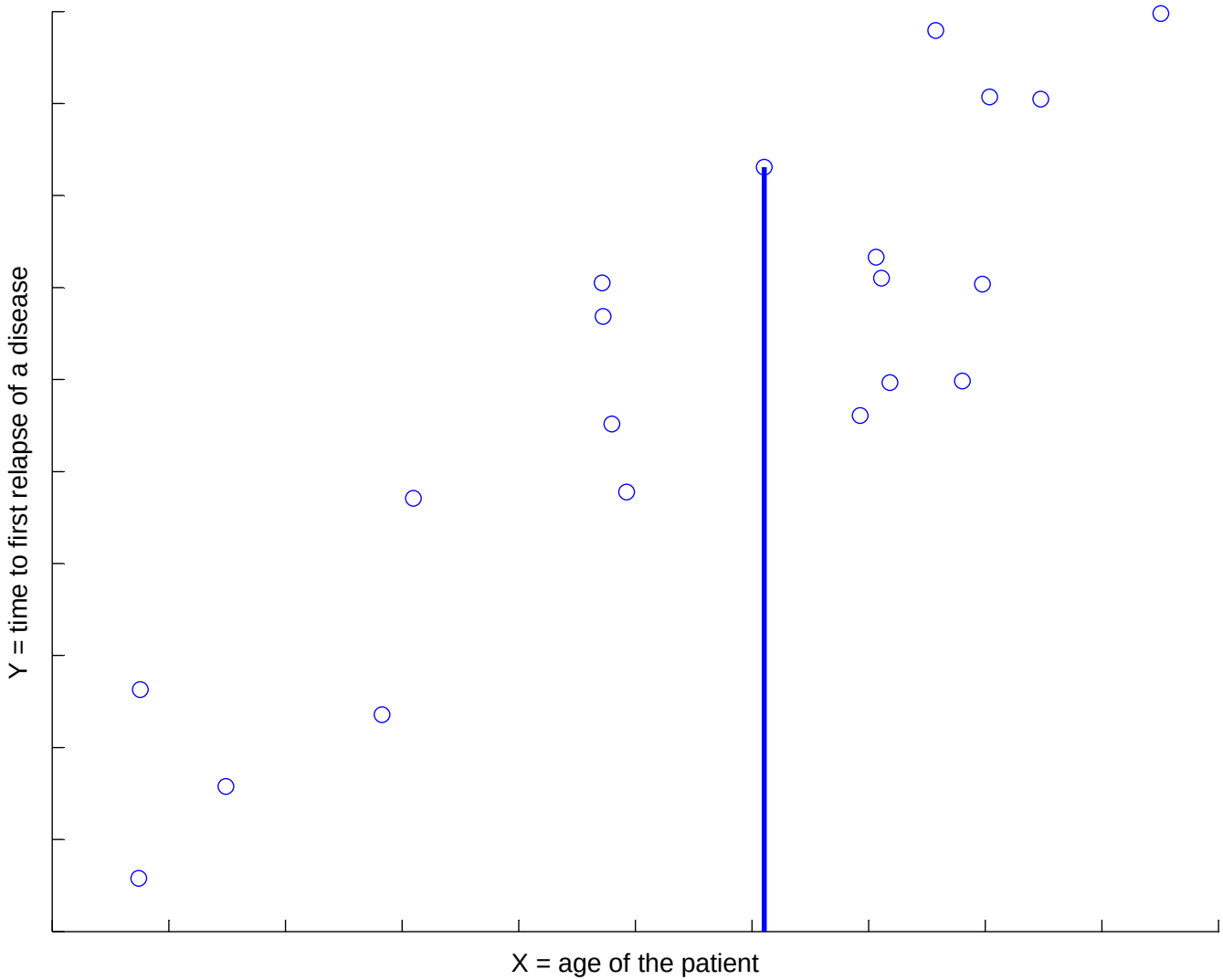
How to estimate $F(y|x)$?

- ~~Survival analysis~~
- Informative censoring
- Copulas
- Location-scale regression

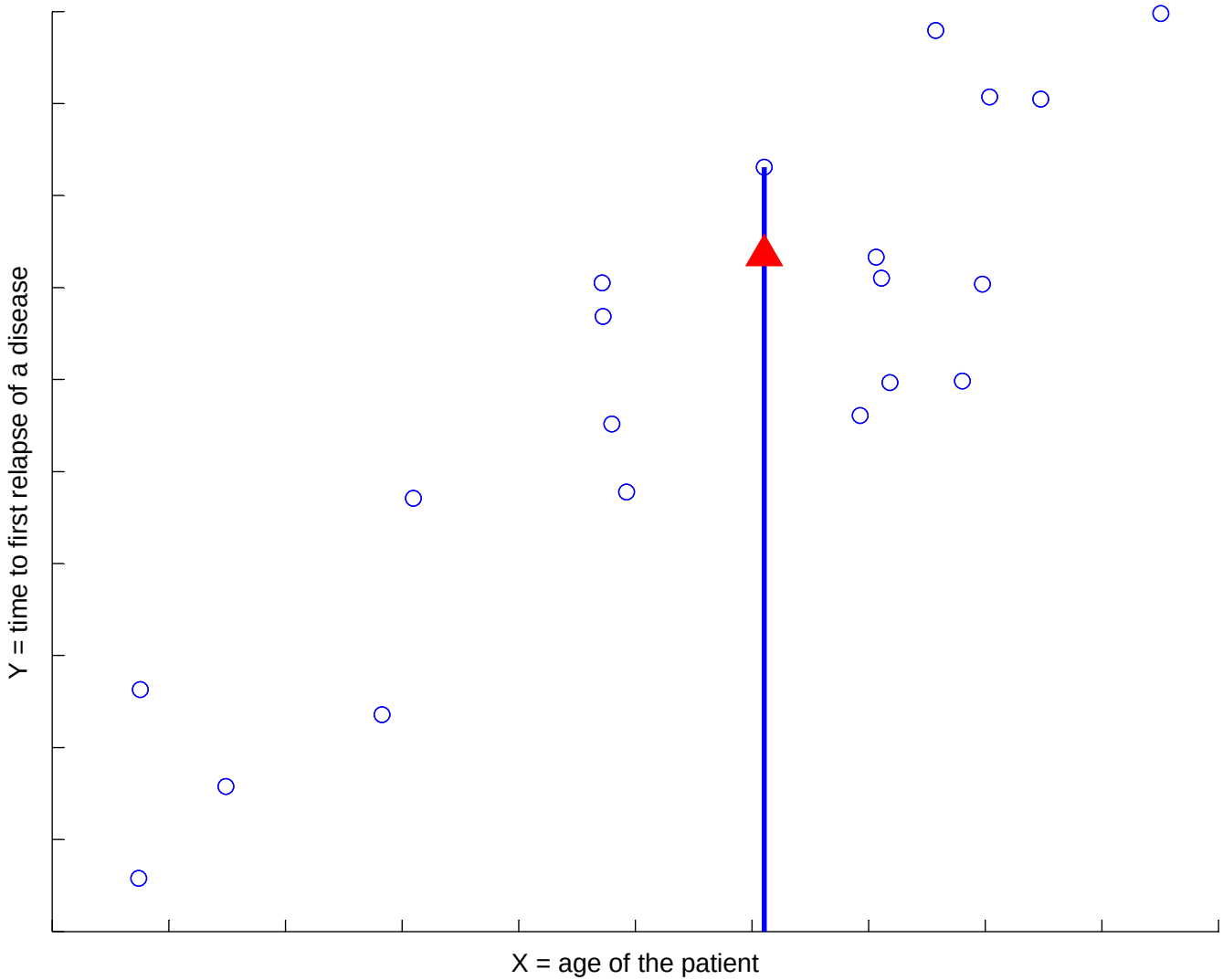
Censoring



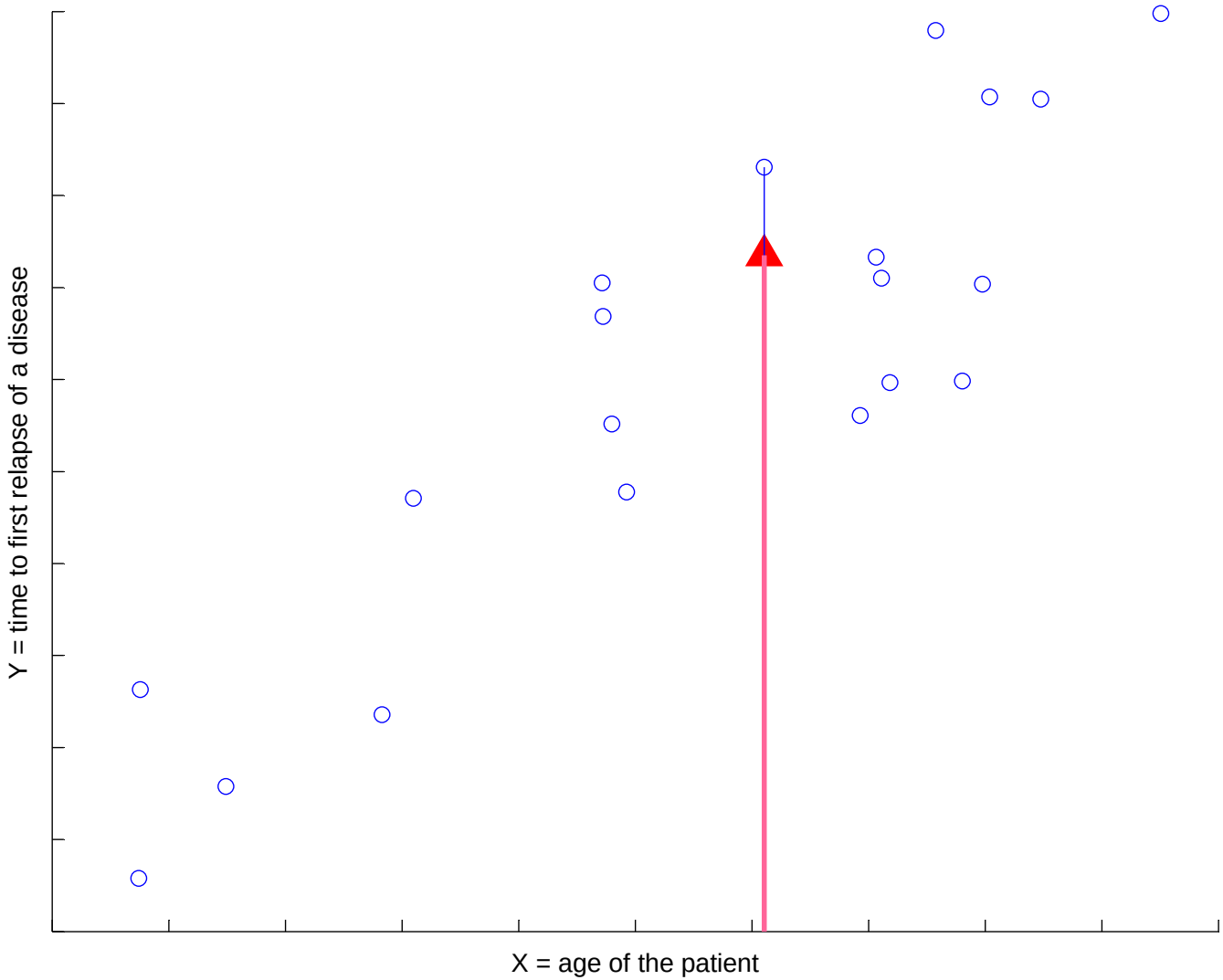
Censoring



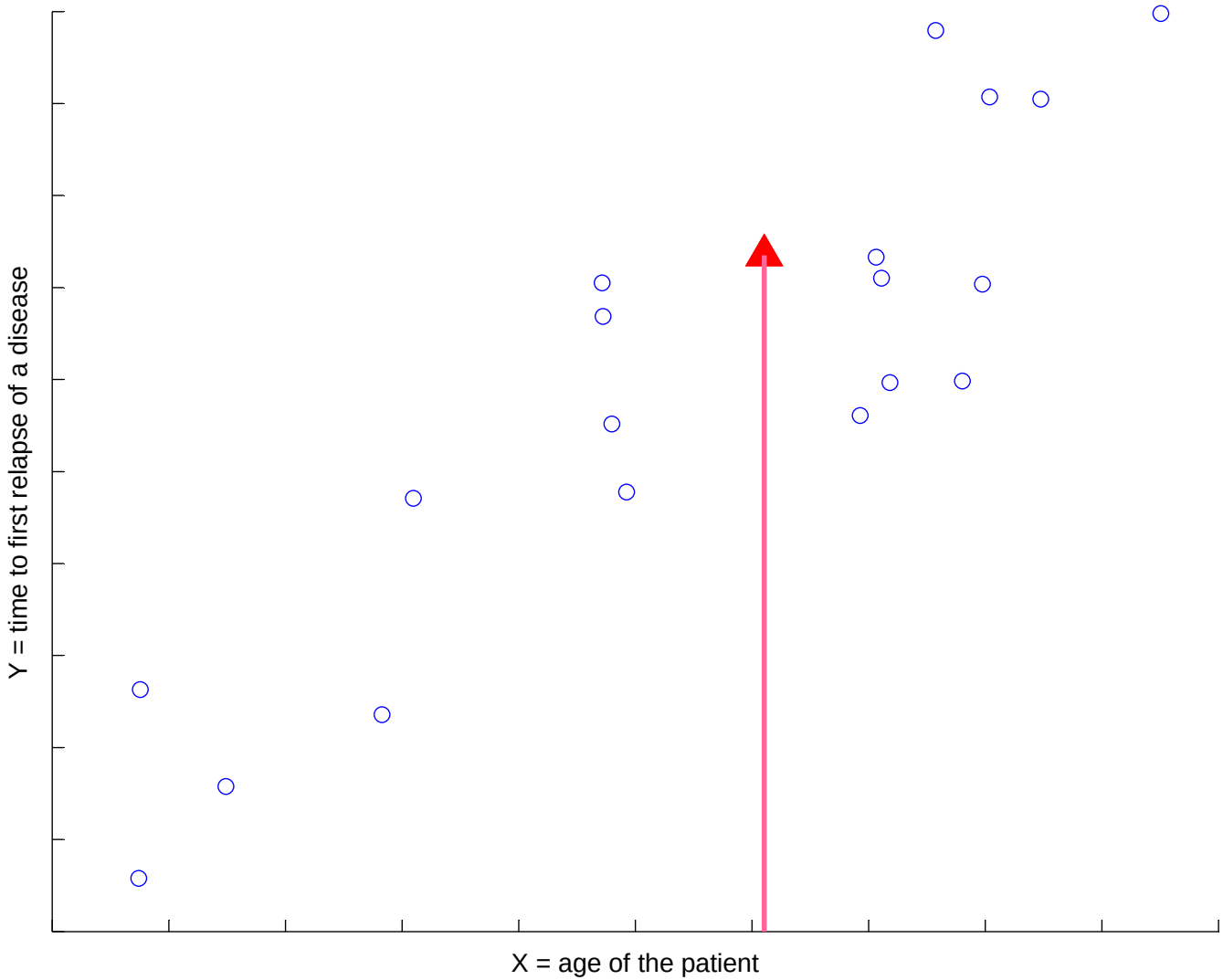
Censoring



Censoring



Censoring



Censoring

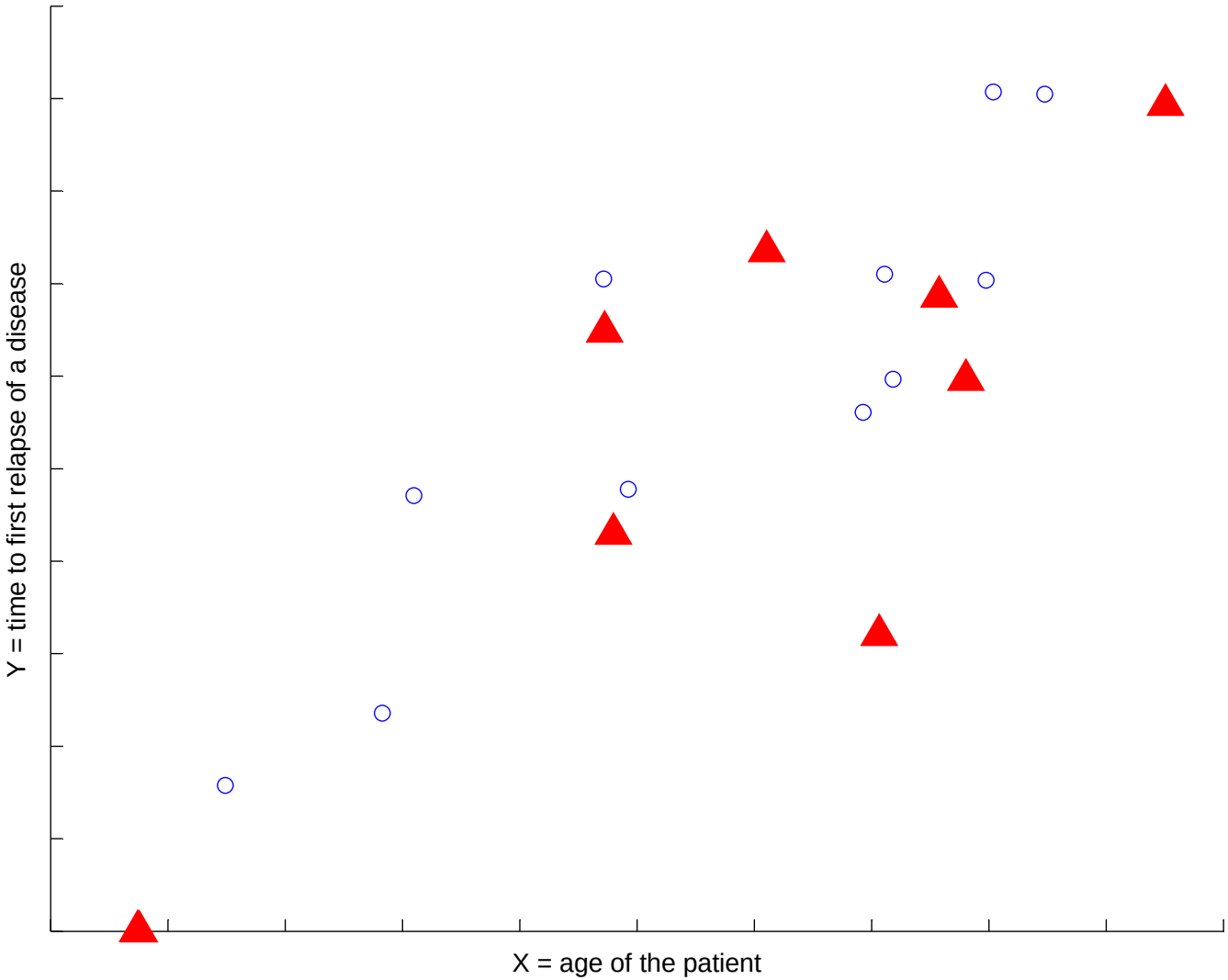
What is Censoring ?

$\Rightarrow$ Instead of observing Y, we observe (T, Δ), where

- $T = \min(Y, C)$
- $\Delta = I(Y \leq C)$,

and C is the censoring time.

Informative censoring



Copula

What is a copula ?

A function containing ALL INFORMATION about DEPENDENCE between two random variables

Sklar's theorem

If F and G are continuous, there exists a unique copula $\mathscr{C}$ such that

$$P(Y > y, C > c) = \mathscr{C}\left(\overline{F}(y), \overline{G}(c)\right),$$

where $\bar{F}(y) = 1 - F(y) = P(Y > y)$

$\bar{G}(c) = 1 - G(c) = P(C > c)$

Copula

What is a copula ?

A function containing ALL INFORMATION about DEPENDENCE between two random variables

Sklar's theorem

If F and G are continuous, there exists a unique copula $\mathscr{C}$ such that

$$P(Y > y, C > c) = \mathscr{C}\left(\overline{F}(y), \overline{G}(c)\right),$$

where $\bar{F}(y) = 1 - F(y) = P(Y > y)$

$\bar{G}(c) = 1 - G(c) = P(C > c)$

Copula

We will use Archimedean copula i.e.

$$\mathscr{C}(u,v) = \varphi^{-1}\big(\varphi(u)+\varphi(v)\big),\ 0 \le u,v \le 1,$$

$\Rightarrow$ what gives

$$P(Y > y, C > c) = \varphi^{-1}\big(\varphi(\overline{F}(y)) + \varphi(\overline{G}(c))\big).$$

For each patient we also observe a covariate X
$\Rightarrow$ We will have for every x a generator φ_x such that

$$P(Y > y, C > c|X = x) = \varphi_x^{-1}\big(\varphi_x(\overline{F}(y|x)) + \varphi_x(\overline{G}(c|x))\big),$$

where $\overline{F}(y|x) = P(Y > y|X = x)$

$\overline{G}(c|x) = P(C > c|X = x)$

Copula

We will use Archimedean copula i.e.

$$\mathscr{C}(u,v) = \varphi^{-1}\big(\varphi(u)+\varphi(v)\big),\ 0 \le u,v \le 1,$$

$\Rightarrow$ what gives

$$P(Y > y, C > c) = \varphi^{-1}\big(\varphi(\overline{F}(y)) + \varphi(\overline{G}(c))\big).$$

For each patient we also observe a covariate X
$\Rightarrow$ We will have for every x a generator φ_x such that

$$P(Y > y, C > c|X = x) = \varphi_x^{-1}\big(\varphi_x(\overline{F}(y|x)) + \varphi_x(\overline{G}(c|x))\big),$$

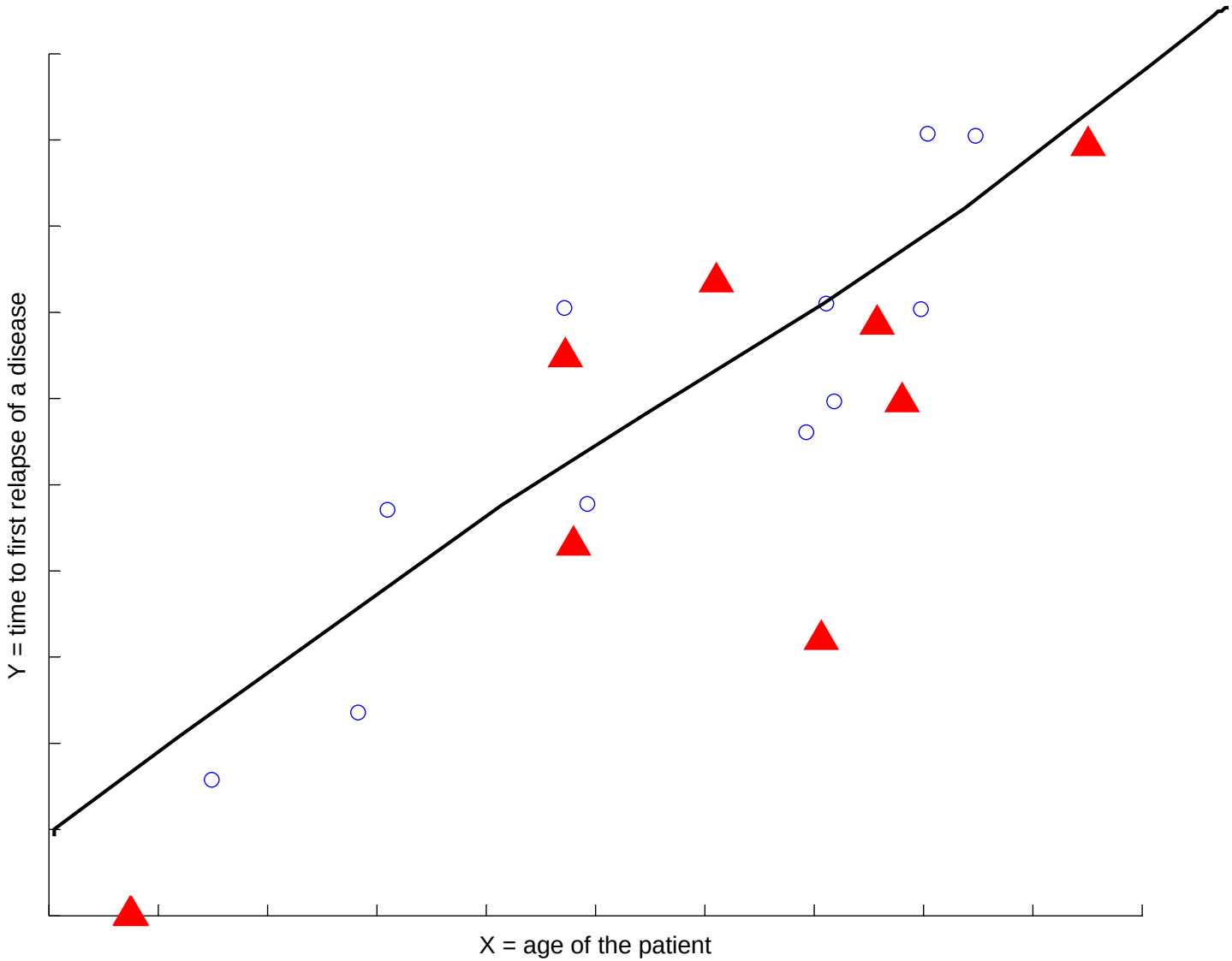
where $\overline{F}(y|x) = P(Y > y|X = x)$

$\overline{G}(c|x) = P(C > c|X = x)$

- ~~Survival analysis~~
- ~~Informative censoring~~
- ~~Copulas~~
- Location-scale regression

Location-scale regression

$$Y = m(X) + \sigma(X)\varepsilon, \quad X \perp\!\!\!\perp \varepsilon$$



- ~~Survival analysis~~
- ~~Informative censoring~~
- ~~Copulas~~
- ~~Location-scale regression~~

Our model

- $(T_i = \min(Y_i, C_i), \Delta_i, X_i) \quad i = 1, ..., n$
- $P(Y < y, C < c | X = x) = \varphi_x^{-1}(\varphi_x(F(y|x)) + \varphi_x(G(c|x)))$
- $Y = m(X) + \sigma(X)\varepsilon, \;\; X \perp\!\!\!\perp \varepsilon$

History and Contribution

Example

Y = time to first relapse of a cancer (survival time)

C = time until leaving the study

X = age of the patient

How to estimate $F(y|x)$?

$\Rightarrow$ Beran (1981)

Y, C given X are independent

$\Rightarrow$ Van Keilegom & Akritas (1999)

Y, C given X are independent

$Y = m(X) + \sigma(X)\varepsilon, \;\; X \perp\!\!\!\perp \varepsilon$

$\Rightarrow$ Braekers & Veraverbeke (2005)

Y, C given X are copula dependent and copula is known

History and Contribution

Example

Y = time to first relapse of a cancer (survival time)

C = time until leaving the study

X = age of the patient

How to estimate $F(y|x)$?

$\Rightarrow$ Beran (1981)

Y, C given X are independent

$\Rightarrow$ Van Keilegom & Akritas (1999)

Y, C given X are independent

$Y = m(X) + \sigma(X)\varepsilon, \;\; X \perp\!\!\!\perp \varepsilon$

$\Rightarrow$ Braekers & Veraverbeke (2005)

Y, C given X are copula dependent and copula is known

History and Contribution

Example

Y = time to first relapse of a cancer (survival time)

C = time until leaving the study

X = age of the patient

How to estimate $F(y|x)$?

$\Rightarrow$ Beran (1981)

Y, C given X are independent

$\Rightarrow$ Van Keilegom & Akritas (1999)

Y, C given X are independent
$Y = m(X) + \sigma(X)\varepsilon, \;\; X \perp\!\!\!\perp \varepsilon$

$\Rightarrow$ Braekers & Veraverbeke (2005)

Y, C given X are copula dependent and copula is known

History and Contribution

Example

Y = time to first relapse of a cancer (survival time)

C = time until leaving the study

X = age of the patient

How to estimate $F(y|x)$?

$\Rightarrow$ Beran (1981)

Y, C given X are independent

$\Rightarrow$ Van Keilegom & Akritas (1999)

Y, C given X are independent
$Y = m(X) + \sigma(X)\varepsilon, \;\; X \perp\!\!\!\perp \varepsilon$

$\Rightarrow$ Braekers & Veraverbeke (2005)

Y, C given X are copula dependent and copula is known

	Fully nonparametric model	Location-scale regression
Independent censoring	Beran (1981)	VK & Akritas (1999)
Informative censoring	Braekers & Veraverbeke (2005)	Our model

Outline

Estimating $\bar{F}(y|x) = P(Y > y|X = x)$

For simplicity reasons assume that $\sigma(x) = 1$

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

We can show that

$$\bar{F}(y|x) = \bar{F}_e(y - m(x)),$$

where $\bar{F}_e(t) = P(\varepsilon > t)$.

We will estimate $\bar{F}(y|x)$ by estimating $\bar{F}_e(\cdot)$ and $m(\cdot)$

$$\hat{\bar{F}}(y|x) = \hat{\bar{F}}_e(y - \hat{m}(x))$$

Estimating $\bar{F}(y|x) = P(Y > y|X = x)$

For simplicity reasons assume that $\sigma(x) = 1$

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

We can show that

$$\bar{F}(y|x) = \bar{F}_e(y - m(x)),$$

where $\bar{F}_e(t) = P(\varepsilon > t)$.

We will estimate $\bar{F}(y|x)$ by estimating $\bar{F}_e(\cdot)$ and $m(\cdot)$

$$\hat{\bar{F}}(y|x) = \hat{\bar{F}}_e(y - \hat{m}(x))$$

Estimating $\bar{F}(y|x) = P(Y > y|X = x)$

For simplicity reasons assume that $\sigma(x) = 1$

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

We can show that

$$\bar{F}(y|x) = \bar{F}_e(y - m(x)),$$

where $\bar{F}_e(t) = P(\varepsilon > t)$.

We will estimate $\bar{F}(y|x)$ by estimating $\bar{F}_e(\cdot)$ and $m(\cdot)$

$$\hat{\bar{F}}(y|x) = \hat{\bar{F}}_e(y - \hat{m}(x))$$

Estimating $\bar{F}(y|x) = P(Y > y|X = x)$

For simplicity reasons assume that $\sigma(x) = 1$

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

We can show that

$$\bar{F}(y|x) = \bar{F}_e(y - m(x)),$$

where $\bar{F}_e(t) = P(\varepsilon > t)$.

We will estimate $\bar{F}(y|x)$ by estimating $\bar{F}_e(\cdot)$ and $m(\cdot)$

$$\hat{\bar{F}}(y|x) = \hat{\bar{F}}_e(y - \hat{m}(x))$$

$$\hat{F}(y|x) = \hat{F}_e(y - \hat{m}(x))$$

$\hat{m}(\cdot)$

$\hat{F}_e(\cdot)$

X_i Δ_i T_i

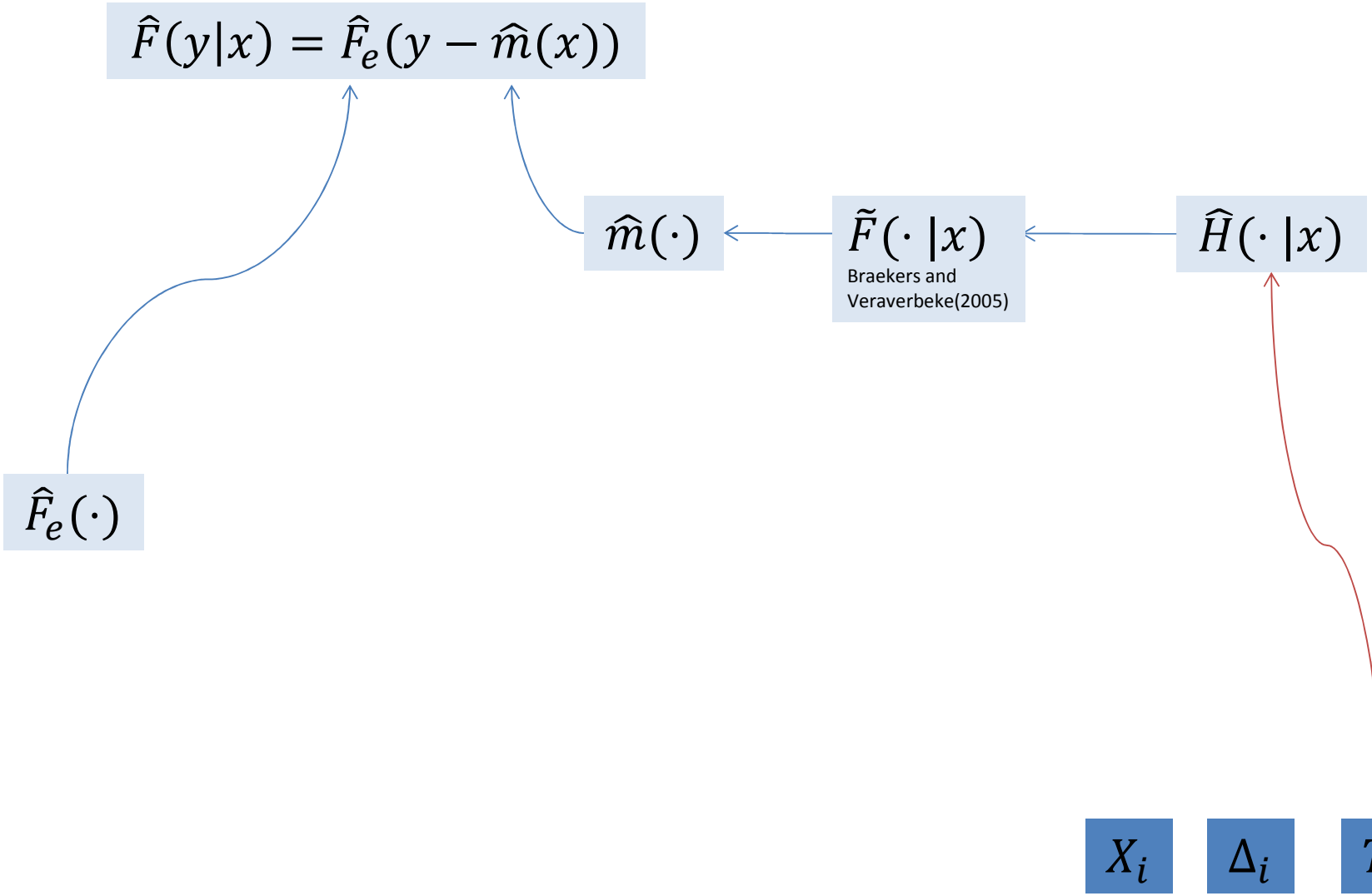
$$\hat{F}(y|x) = \hat{F}_e(y - \hat{m}(x))$$

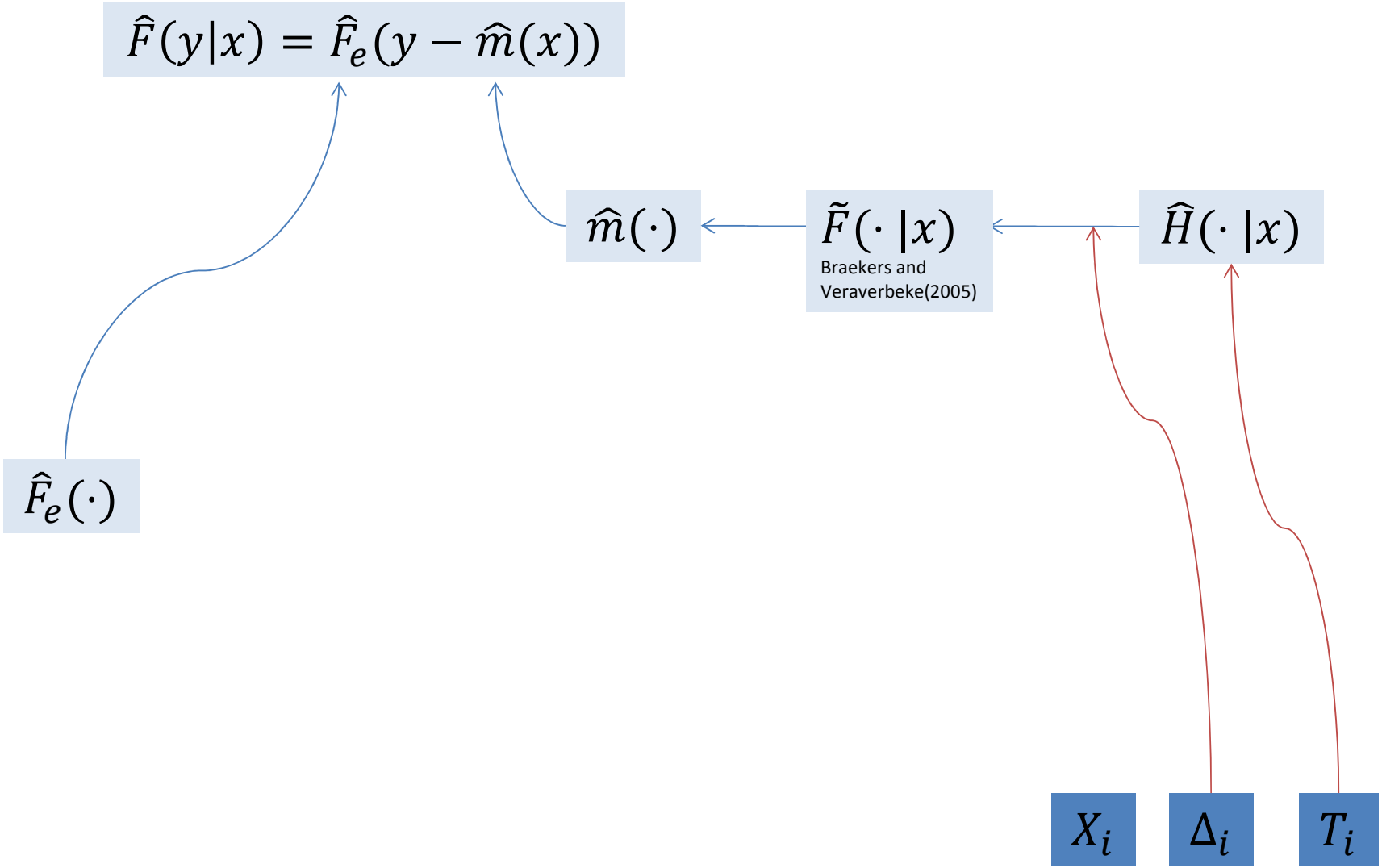
$\hat{m}(\cdot)$

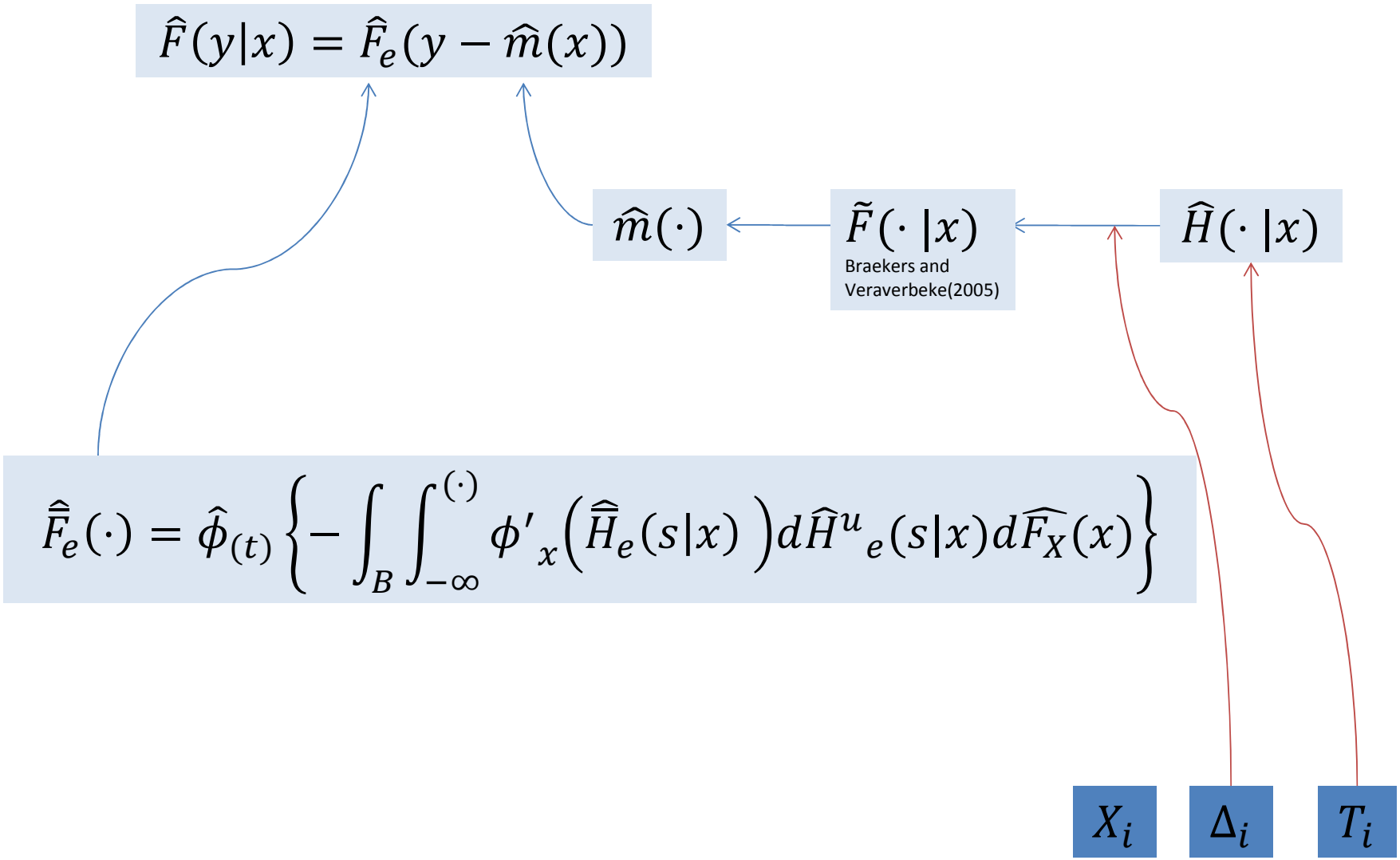
$\tilde{F}(\cdot\,|x)$
Braekers and Veraverbeke(2005)

$\hat{F}_e(\cdot)$

X_i Δ_i T_i


$\hat{F}(y|x) = \hat{F}_e(y - \hat{m}(x))$
$\hat{m}(\cdot)$
$\tilde{F}(\cdot\,|x)$
Braekers and Veraverbeke(2005)
$\hat{H}(\cdot\,|x)$
$\hat{F}_e(\cdot)$
X_i
Δ_i
T_i


$\hat{F}(y|x) = \hat{F}_e(y - \hat{m}(x))$
$\hat{m}(\cdot)$
$\tilde{F}(\cdot\,|x)$
Braekers and Veraverbeke(2005)
$\hat{H}(\cdot\,|x)$
$\hat{F}_e(\cdot)$
X_i
Δ_i
T_i


$\hat{F}(y|x) = \hat{F}_e(y - \hat{m}(x))$
$\hat{m}(\cdot)$
$\tilde{F}(\cdot\,|x)$
Braekers and Veraverbeke(2005)
$\hat{H}(\cdot\,|x)$
$\hat{\bar{F}}_e(\cdot) = \hat{\phi}_{(t)}\left\{-\int_B \int_{-\infty}^{(\cdot)} \phi'_x\left(\hat{\bar{H}}_e(s|x)\right) d\hat{H}^u{}_e(s|x) d\widehat{F_X}(x)\right\}$
X_i
Δ_i
T_i

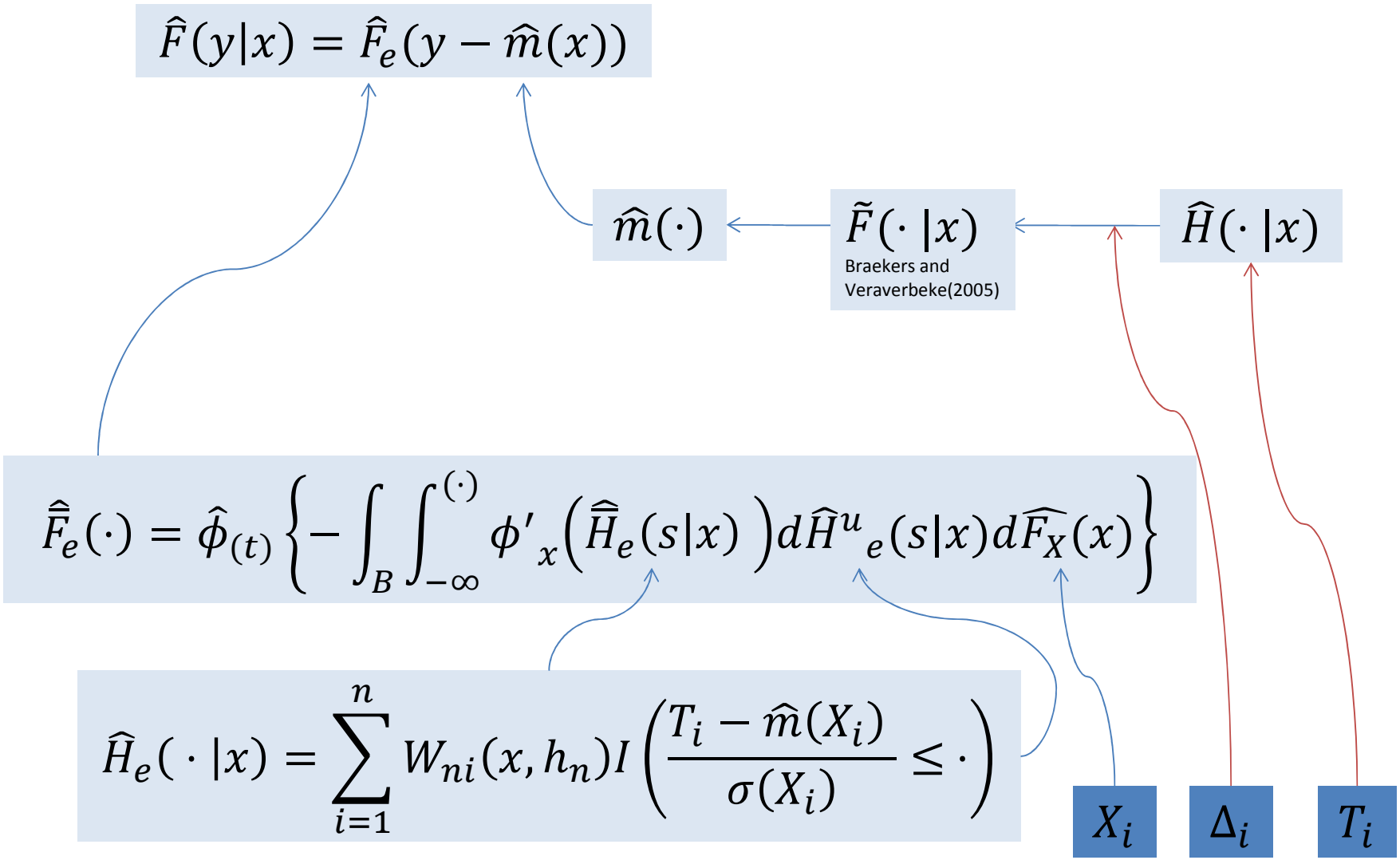
$$\hat{F}(y|x) = \hat{F}_e(y - \hat{m}(x))$$

$\hat{m}(\cdot)$ ← $\tilde{F}(\cdot\,|x)$ Braekers and Veraverbeke(2005) ← $\hat{H}(\cdot\,|x)$

$$\hat{\tilde{F}}_e(\cdot) = \hat{\phi}_{(t)}\left\{-\int_B \int_{-\infty}^{(\cdot)} \phi'_x\left(\hat{\tilde{H}}_e(s|x)\right) d\hat{H}^u{}_e(s|x)\, d\widehat{F_X}(x)\right\}$$

$$\hat{H}_e(\cdot\,|x) = \sum_{i=1}^{n} W_{ni}(x, h_n) I\left(\frac{T_i - \hat{m}(X_i)}{\sigma(X_i)} \leq \cdot\right)$$

X_i Δ_i T_i

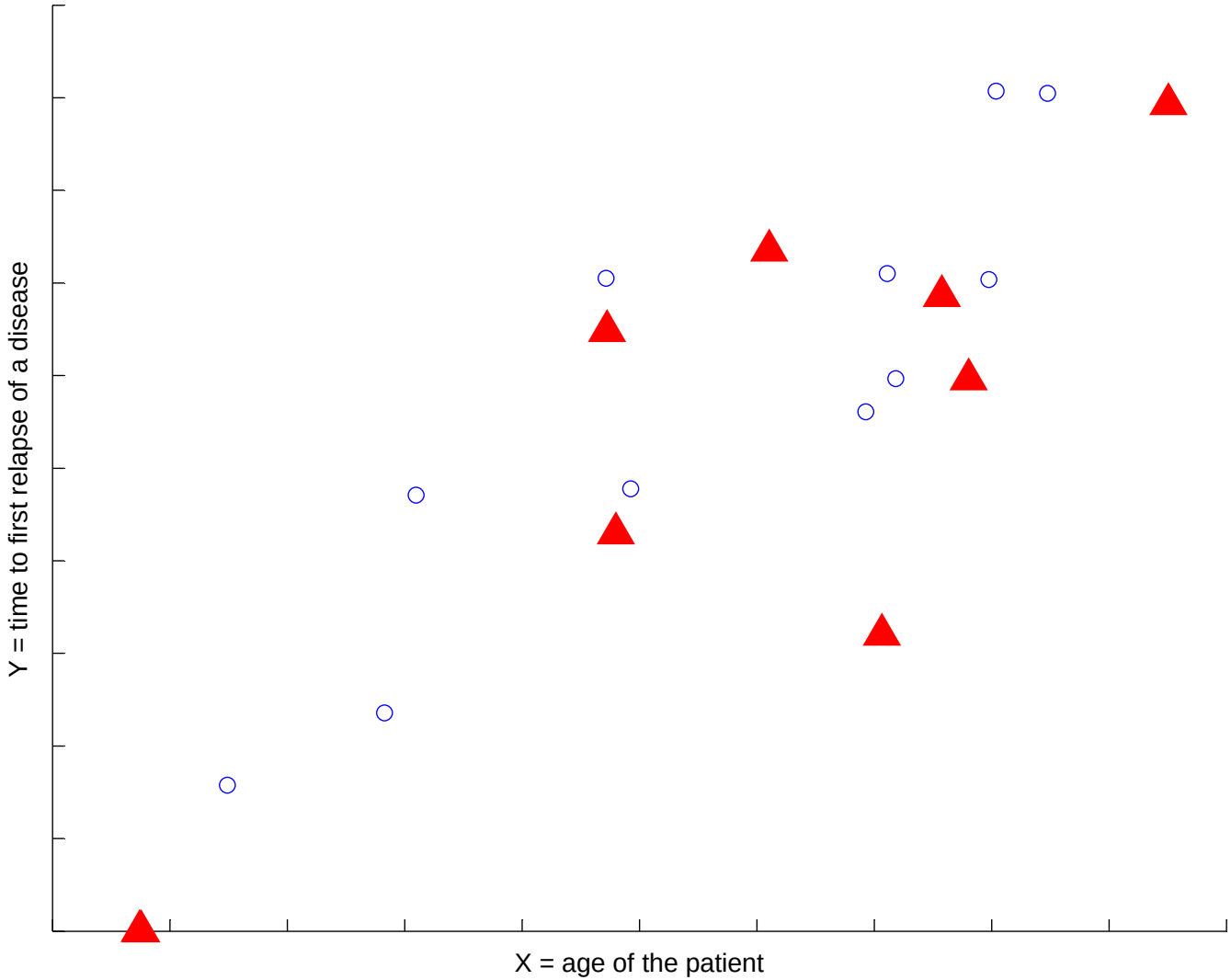

$\hat{F}(y|x) = \hat{F}_e(y - \hat{m}(x))$
$\hat{m}(\cdot)$
$\tilde{F}(\cdot\,|x)$
Braekers and Veraverbeke(2005)
$\hat{H}(\cdot\,|x)$
$\hat{\hat{F}}_e(\cdot) = \hat{\phi}_{(t)}\left\{-\int_B \int_{-\infty}^{(\cdot)} \phi'_x\left(\hat{\hat{H}}_e(s|x)\right) d\hat{H}^u{}_e(s|x) d\widehat{F_X}(x)\right\}$
$\hat{H}_e(\cdot\,|x) = \sum_{i=1}^{n} W_{ni}(x, h_n) I\left(\frac{T_i - \hat{m}(X_i)}{\sigma(X_i)} \leq \cdot\right)$
X_i
Δ_i
T_i

Outline

	Fully nonparametric model	Location-scale regression
Independent censoring	Beran (1981)	VK & Akritas (1999)
Informative censoring	Braekers & Veraverbeke (2005)	New estimator

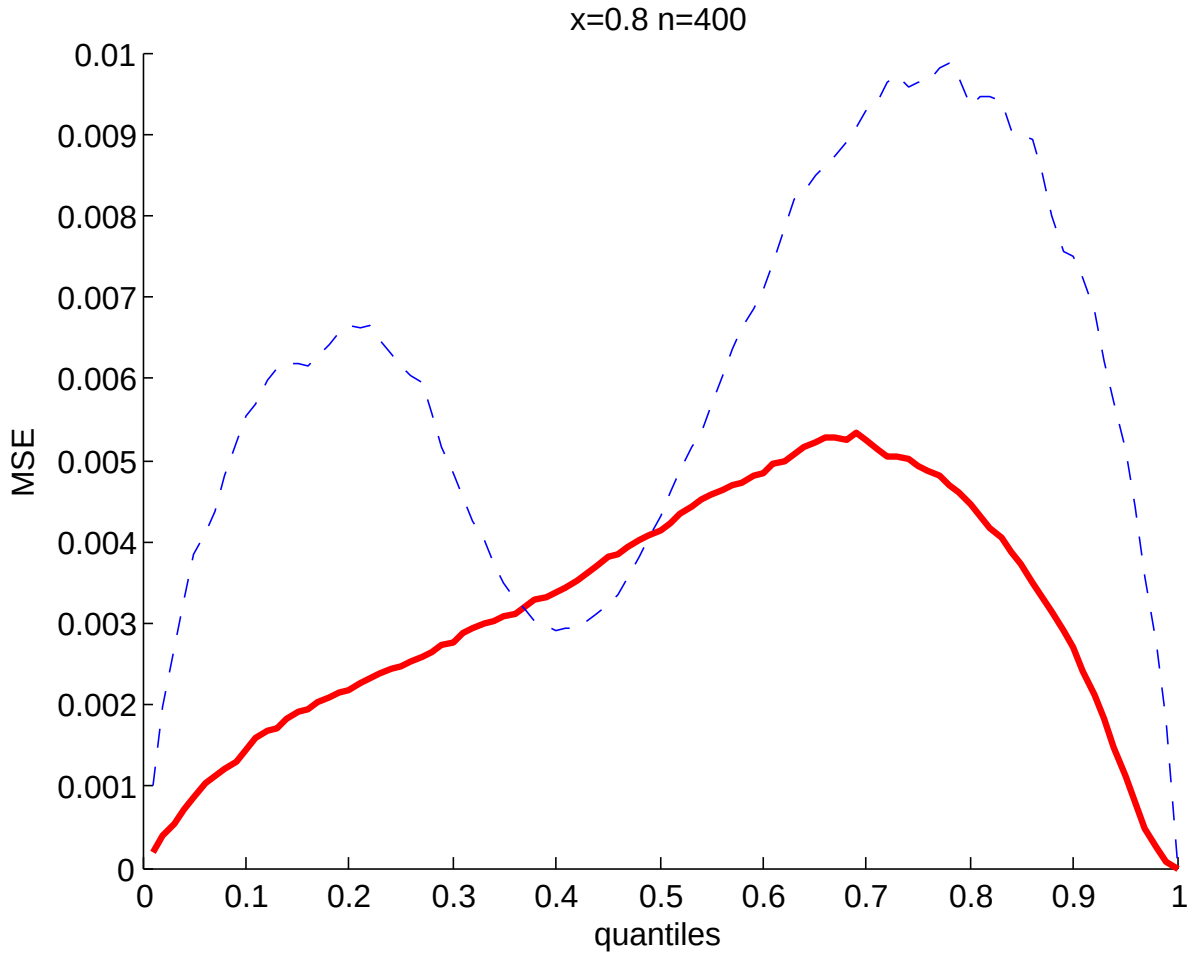
Simulation framework

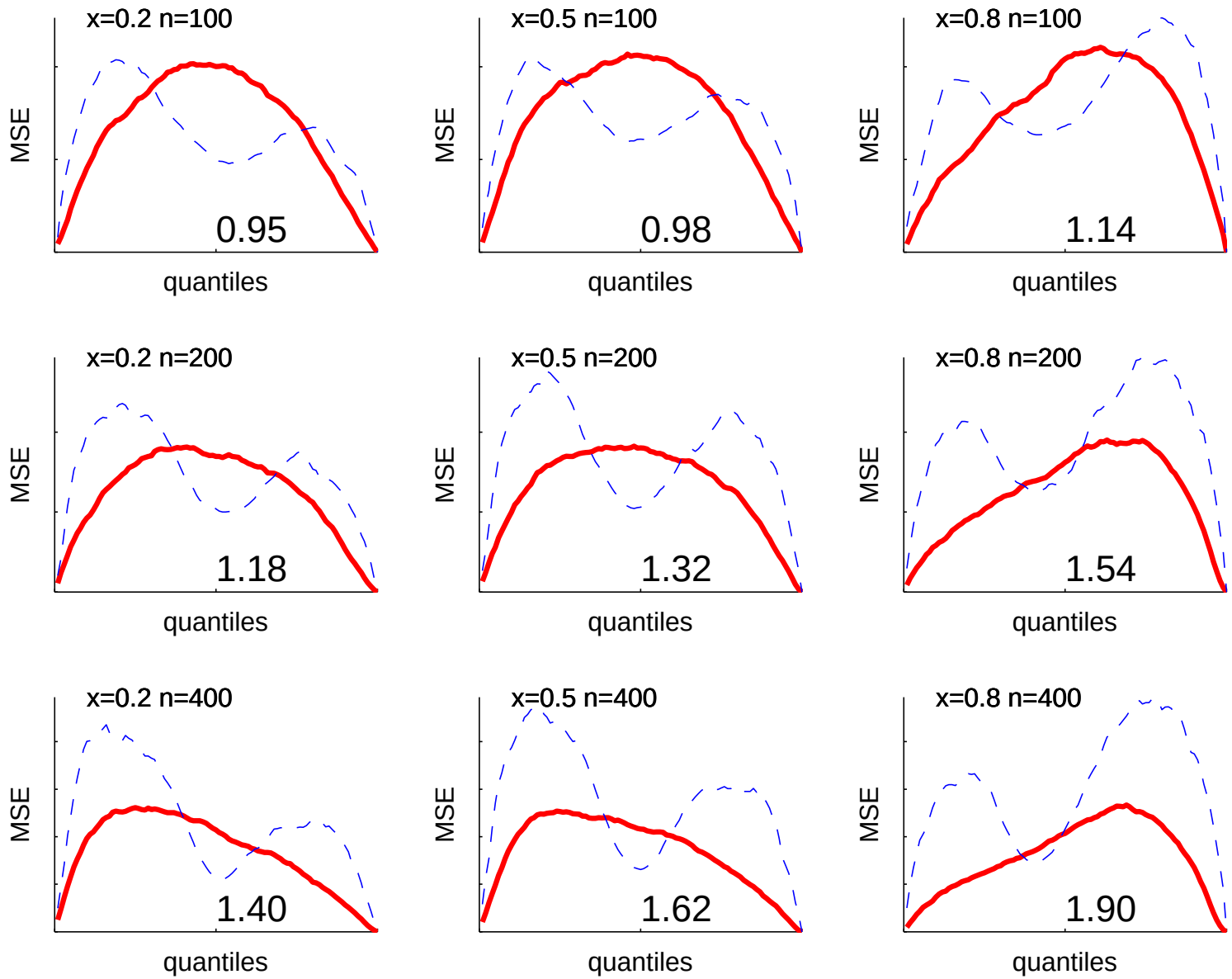
$$Y = 0.1X + 0.015\varepsilon$$



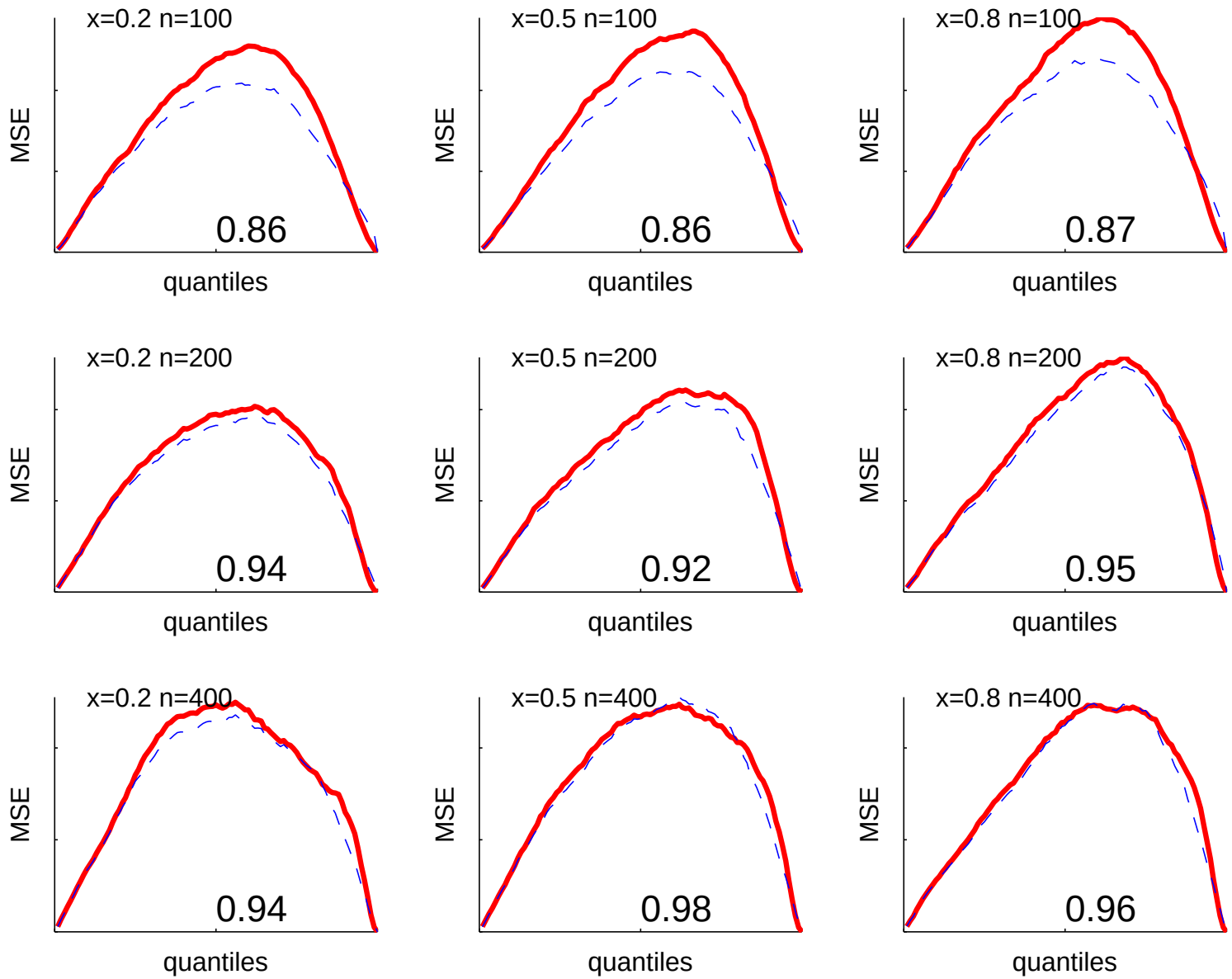
New estimator

Braekers & Veraverbeke (2005)




x=0.2 n=100
MSE
0.95
quantiles
x=0.5 n=100
MSE
0.98
quantiles
x=0.8 n=100
MSE
1.14
quantiles
x=0.2 n=200
MSE
1.18
quantiles
x=0.5 n=200
MSE
1.32
quantiles
x=0.8 n=200
MSE
1.54
quantiles
x=0.2 n=400
MSE
1.40
quantiles
x=0.5 n=400
MSE
1.62
quantiles
x=0.8 n=400
MSE
1.90
quantiles

	Fully nonparametric model	Location-scale regression
Independent censoring	Beran (1981)	VK & Akritas (1999)
Informative censoring	Braekers & Veraverbeke (2005)	New estimator


x=0.2 n=100
0.86
x=0.5 n=100
0.86
x=0.8 n=100
0.87
x=0.2 n=200
0.94
x=0.5 n=200
0.92
x=0.8 n=200
0.95
x=0.2 n=400
0.94
x=0.5 n=400
0.98
x=0.8 n=400
0.96
MSE
quantiles

Outline

We have studied the estimation of $F(y|x) = P(Y \leq y|X = x)$ when

- $Y = m(X) + \sigma(X)\varepsilon$, with ε independent of X
- Y is subject to random right censoring
- Y and C are copula dependent for given X

Future research :

- Tests for the comparison of regression curves
- Tests for the error distribution $F_e(\cdot)$
- Estimation of copula

Thank you for your time :)

Black guy