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

Géraldine Laurent[1] and Cédric Heuchenne[2]

[1] QuantOM, HEC-Management School of University of Liège
boulevard du Rectorat, 7 Bât.31, B-4000 Liège, Belgium,
G.Laurent@student.ulg.ac.be
[2] QuantOM, HEC-Management School of University of Liège
boulevard du Rectorat, 7 Bât.31, B-4000 Liège, Belgium,
C.Heuchenne@ulg.ac.be

Abstract. Consider the regression model $Y = m(X) + \sigma(X)\varepsilon$, where $m = E[Y|X]$ and $\sigma^2(X) = Var[Y|X]$ are unknown smooth functions and the error ε (with unknown distribution) is independent of X. The pair (X, Y) is subject to parametric selection bias and the response to right censoring. We construct a new estimator for the cumulative distribution function of the error ε, and develop a bootstrap technique to select the smoothing parameter involved in the procedure. The estimator is studied via extended simulations and applied to real unemployment data.

Keywords: Nonparametric regression, selection bias, right censoring, bootstrap, bandwidth selection

1 Introduction and model

Let (X, Y) be a bivariate random vector, where Y is the unemployment duration of an individual and X is, for example, his age when he lost his job. The objective is to study the relation between Y and X. In Figure 1, an example of a scatter plot with these two variables is displayed. It comes from the Spanish Institute for Statistics and is completely described in Section 4. Unfortunately, this kind of data set suffers from some 'incompleteness' (due to sampling), as explained hereunder.

Indeed, (X, Y) is supposed to be obtained from cross-sectional sampling meaning that only individuals whose unemployment duration is in progress at a fixed sampling time are observed and followed. As a result, a bias appears due to the length of Y: conditionally on X, longer durations have a larger probability to be observed. Moreover, we assume that durations of the followed individuals are possibly right-censored; for example, this may happen if an individual stops the follow-up or if the follow-up itself comes to an end.

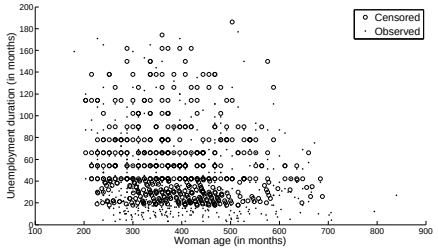

Fig. 1. Scatter plot of unemployment data

In this context, the following general nonparametric regression model can be assumed in most applications:

$$Y = m(X) + \sigma(X)\varepsilon, \tag{1}$$

where $m(X) = E[Y|X]$ and $\sigma^2(X) = Var[Y|X]$ are unknown smooth functions and ε (with zero mean, unit variance and distribution F_ε) is independent of X. This enables to define the error ε and estimate its distribution $F_\varepsilon(\cdot)$. Indeed, such an estimator can be very useful in the sense that it is naturally related to the commonly used graphical procedures based on visual examination of the residuals (see Atkinson 1985). Furthermore, a complete set of testing procedures can be based on this estimated distribution (e.g. tests for the model (1), goodness-of-fit tests for $F_\varepsilon(\cdot)$, $m(\cdot)$ and $\sigma(\cdot)\ldots$).

As explained above, the incompleteness of the data is characterized by two phenomena: cross-sectional sampling and right censoring. We can therefore model them by using the following variables.

1. T, the truncation variable (duration between the time point when the individual loses his job and the sampling time) assumed to be here independent of Y conditionally on X (usual assumption when truncated data are present): T is observed if $Y \geq T$.
2. C, the censoring variable making Y (larger or equal to T) observable only if $Y \leq C$. (Y, T) is assumed to be independent of $C - T$, conditionally on $T \leq Y$ and X (assumption needed to construct conditional distribution estimators with censored data).

Here, $F_{T|X}(y|x) = P(T \leq y|x)$ is assumed to be a parametric function. This assumption is satisfied by classical length-biased data but also by other types of selection biases where the process that counts individuals who lose their job can be considered as parametric. By construction, we also impose that the support of $F_{Y|X}(y|x) = P(Y \leq y|x)$ is included into the support of $F_{T|X}(y|x)$ and that the lower bound of the support of $F_{T|X}(y|x)$ is zero. Defining $Z = \min(C - T, Y - T)$ and $\Delta = I(Y \leq C)$, we therefore obtain

a sample $\{(X_1, T_1, Z_1, \Delta_1), \ldots, (X_n, T_n, Z_n, \Delta_n)\}$ of independent copies of (X, T, Z, Δ) with the same distribution as (X, T, Z, Δ) conditionally on $Y \geq T$. Special cases of these data have been widely studied in the literature (see, e.g., de Uña-Alvarez and Iglesias-Perez (2008) for a literature overview).

The paper is organized as follows. In the next section, we describe the estimation procedure in detail. $\sigma(x)$ are obtained Section 3 presents the results of a simulation study while Section 4 is devoted the analysis of the unemployment data introduced hereabove.

2 Description of the method

To address the problem introduced in Section 1, we first propose to write

$$H_{X,Y}(x, y) = \mathbb{P}(X \leq x, Y \leq y | T \leq Y \leq C),$$

the observed joint distribution of (X, Y), as

$$H_{X,Y}(x, y) = \frac{\mathbb{P}(X \leq x, Y \leq y, T \leq Y \leq C)}{\mathbb{P}(T \leq Y \leq C)}.$$

We can show that

$$\mathbb{P}(X \leq x, Y \leq y, T \leq Y \leq C) = \int_{r \leq x} \int_{s \leq y} \int_{u \leq s} (1 - \mathcal{G}(s - u|r))\, dF_{T|X}(u|r) dF_{Y|X}(s|r) dF_X(r)$$

and that

$$\mathbb{P}(T \leq Y \leq C) = \iint \int_{u \leq y} (1 - \mathcal{G}(y - u|x))\, dF_{T|X}(u|x) dF_{Y|X}(y|x) dF_X(x),$$

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

$$H_{X,Y}(x, y) = (E[w(X, Y)])^{-1} \int_{r \leq x} \int_{s \leq y} w(r, s) dF_{X,Y}(r, s), \tag{2}$$

where the weight function is defined by

$$w(x, y) = \int_{t \leq y} (1 - \mathcal{G}(y - t|x))\, dF_{T|X}(t|x). \tag{3}$$

In particular, a similar expression can be obtained for a constant follow-up τ, i.e. $C = T + \tau$ where τ is a positive constant. By applying the same reasoning, it's easy to check that the weight $w(x, y)$ can be written as

$$w(x, y) = \int_{0 \vee y - \tau}^{y} dF_{T|X}(t|x). \tag{4}$$

Thanks to (2), we have

$$dF_{X,Y}(x,y) = \frac{E[w(X,Y)]}{w(x,y)} dH_{X,Y}(x,y),$$

leading to

$$F_\varepsilon(e) = \mathbb{P}\left(\frac{Y-m(X)}{\sigma(X)} \leq e\right) = \iint_{\{(x,y):\frac{y-m(x)}{\sigma(x)} \leq e\}} \frac{E[w(X,Y)]}{w(x,y)} dH_{X,Y}(x,y). \tag{5}$$

Next, we estimate the unknown quantities in (5). For $\mathcal{G}(y-t|x)$, we use the Beran (1981) estimator, defined by (in the case of no ties):

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

where

$$W_i(x,h_n) = \frac{K\left(\frac{x-X_i}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)},$$

K is a kernel function and h_n is a bandwidth sequence tending to 0 when $n \to \infty$. We thus obtain for $w(x,y)$

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

For $m(\cdot)$ and $\sigma(\cdot)$, we use

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

obtained by extending the conditional estimation methods introduced in de Uña-Alvarez and Iglesias-Perez (2008). Consequently, the estimator of the error distribution is

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

where

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

and $\hat{H}_{X,Y}(x,y)$ is the bivariate empirical distribution based on pairs (X_i,Y_i) verifying $T_i \leq Y_i \leq C_i$, $i = 1,\ldots,n$.

Remark 2.1 Under some assumptions, weak convergence of $\hat{F}_\varepsilon$ can be obtained by extensions of the proofs of de Uña-Alvarez and Iglesias-Perez (2008) and Ojeda Cabrera and Van Keilegom (2008). For more information about that, details can be given on request to the authors.

3 Practical implementation and simulations

3.1 Bandwidth selection procedure

We want to determine the smoothing parameter h_n which minimizes

$$MISE = E[\int \{\hat{F}_{\varepsilon,h_n}(e) - F_\varepsilon(e)\}^2 de]. \tag{6}$$

Considering asymptotic expansions for (6) will lead to complicated expressions with too many unknown quantities. As a consequence, we develop a bootstrap procedure. This is an extension of the method of Li and Datta (2001) to the truncation case.

The bootstrap procedure is explained below.
For $b = 1, \ldots, B$,

1. for $i = 1, \ldots, n$,
   Step 1. Generate $X^*_{i,b}$ from the distribution

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

   where $\hat{E}[w(X,Y)|X = \cdot] = \left(\sum_{j=1}^{n} \frac{W_j(\cdot, g_n)\Delta_j}{\hat{w}(\cdot, Y_j)} \right)^{-1}$, and where g_n is a pilot bandwidth asymptotically larger than the original h_n.
   Step 2. Select at random a $Y^*_{i,b}$ from the distribution

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

   Step 3. Draw $T^*_{i,b}$ from the distribution $F_{T|X}(.|X^*_{i,b})$. If $T^*_{i,b} > Y^*_{i,b}$, then reject the datum $(X^*_{i,b}, Y^*_{i,b}, T^*_{i,b})$ and go to Step 1. Otherwise, go to Step 4.
   Step 4. Select at random a $(C-T)^*_{i,b}$ from $\hat{\mathcal{G}}(.|X^*_{i,b})$ calculated with g_n and compute $C^*_{i,b} = T^*_{i,b} + (C-T)^*_{i,b}$.
   Step 5. Define $Z^*_{i,b} = \min(Y^*_{i,b} - T^*_{i,b}, C^*_{i,b} - T^*_{i,b})$ and $\Delta^*_{i,b} = I\{Y^*_{i,b} \leq C^*_{i,b}\}$.

2. Compute $\hat{F}^*_{\varepsilon,h_n,b}$ with the obtained resample $\{(X^*_{i,b}, T^*_{i,b}, Z^*_{i,b}, \Delta^*_{i,b}) : i = 1, ..., n\}$

From this, the mean squared of the error distribution can be approximated by

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

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

Table 1. Results for the MISE for the regression model (7)

3.2 Simulations

We study the MISE (obtained from (6) where $E[\cdot]$ is etimated by the average over all the samples and h_n is defined by the above bootstrap procedure) of the error distribution for two homoscedastic and two heteroscedastic models. In the homoscedastic cases, we computed error distributions based on $Y_i - \hat{m}(X_i)$, avoiding the estimation of $\sigma(X_i)$, $i = 1, \ldots, n$. For each model, we consider both finite and infinite supports. We chose to work with the Epanechnikov kernel. The simulations are carried out for samples of size $n = 100$, and $B = 250$, and the results are obtained by using 250 simulations.

In the first setting, we generate i.i.d. samples from the homoscedastic regression model

$$Y = X + \varepsilon, \tag{7}$$

where X has a uniform distribution on $[1,7321;2]$ and ε has a uniform distribution on $[-\sqrt{3};\sqrt{3}]$.

Table 1 summarizes the simulation results for different distributions of T and $C-T$. Clearly, the MISE decreases when the censoring percentage (third column) decreases whatever the distributions of T and $C-T$. Notice that, for the same censoring percentage, the MISE is weaker when T has a beta distribution instead of a uniform distribution. It's explained by the shape of the beta distribution.

In the second setting, we consider a heteroscedastic regression model

$$Y = X^2 + X * \varepsilon, \tag{8}$$

where X has a uniform distribution on $[2;2\sqrt{3}]$ and ε has a uniform distribution on $[-\sqrt{3};\sqrt{3}]$.

In Table 2, when looking at a heteroscedastic instead of a homoscedastic model, introduced variability seems to increase the MISE in a reasonable way. The MISE increasing is not surprising because we don't estimate $\sigma(x)$ in the homoscedastic model. If the distributions of T or $C-T$ depend on X, the MISE doesn't seem to vary significantly whatever the support of ε.

Dist. of T	Dist. of $C-T$	% Censor.	MISE ($*10^{-3}$)
$T \sim \text{Unif}([0;18])$	$C-T \sim \text{Exp}(0.1)$	0.34	6.9
$T \sim \text{Unif}([0;18])$	$C-T \sim \text{Exp}(0.05)$	0.19	6.2
$T \sim 18 * \text{Beta}(0.5;1)$	$C-T \sim \text{Exp}(1/12)$	0.35	6.3
$T \sim 18 * \text{Beta}(0.5;1)$	$C-T \sim \text{Exp}(1/15)$	0.3	6.2
$T \sim \text{Unif}([0;X+16])$	$C-T \sim \text{Exp}(1/12)$	0.3	6.2
$T \sim 18 * \text{Beta}(0.5;1)$	$C-T \sim \text{Exp}(1/(2X^2-1))$	0.3	6.6

Table 2. Results for the MISE for the regression model (8)

Dist. of T	Dist. of $C-T$	% Censor.	MISE ($*10^{-3}$)
$T \sim \text{Exp}(2)$	$C-T \sim \text{Exp}(0.25)$	0.27	6.6
$T \sim \text{Exp}(2)$	$C-T \sim \text{Exp}(2/9)$	0.25	6.5
$T \sim \text{Exp}(2)$	$C-T \sim \text{Exp}(0.2)$	0.23	6.3

Table 3. Results for the MISE for the heteroscedastic regression model (9)

In the third setting, we study both the homoscedastic and heteroscedastic models

$$\log(Y) = X + \varepsilon \qquad \text{and} \qquad \log(Y) = X^2 + X * \varepsilon, \tag{9}$$

where X has a uniform distribution on $[0;1]$ and ε has a standard normal distribution. In this case, Y is submitted to selection bias and right censoring while the error distribution to estimate is here $\mathbb{P}\left(\frac{\log(Y)-m(X)}{\sigma(X)} \leq e\right)$. This is achieved by a straightforward transformation of expression (5). Results are similar to finite supports but generally less good. To illustrate this, the Table 3 displays some results for the heteroscedastic model.

When looking at the shape of the estimations of the error distributions, we observe that the estimations are quite good for ε-values included between minus 1 and 1 for the homoscedastic models, whatever the support of F_ε. The loss of ε-values in the tails of the distribution is caused by the combined selection bias and right censoring processes (this loss is slightly harder for infinite supports). Concerning the heteroscedastic models, this phenomenon is increased due to local variance estimation. Finally, similar results are obtained for other simulations, in particular when the other weight (fixed censoring $C-T$) is used.

4 Data analysis

The proposed method is illustrated on the unemployment data set introduced in Section 1. These data result from the survey, Encuesta de Población Activa (Labour Force Survey), of the Spanish Institute for Statistics between 1987

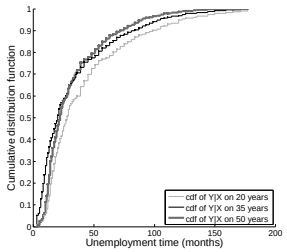

Fig. 2. Representation of $\hat{F}_{Y|X}$ for different values of x

and 1997. The available information consists of 1009 unemployment spells of married women being unemployed at the time of inquiry. Sampled women were asked to provide the date they started searching a job and their age (in years) at this date. After, they were followed for 18 months. If they did not find any job at the end of this period, their unemployment durations were considered as censored. This results in a constant $C-T=\tau$ leading to weights (4). We consider a uniform distribution for the truncation variable. This assumption was informally checked through a graphical comparison between the empirical truncation distribution function and the uniform model (Wang, 1991), showing a good fit.

The bootstrap approximation gives an optimal smoothing parameter of seventy months. The estimator $\hat{F}_{Y|X}(\cdot|x)=\hat{F}_{\varepsilon}\left(\frac{\cdot-\hat{m}(x)}{\hat{\sigma}(x)}\right)$ is displayed in Figure 2 for $x=20,35$ and 50. The 35 years old unemployed women seem to find a job earlier in the short run and later in the long run than the 50 years old unemployed women.

Acknowledgements. Thanks to G. Alvarez-Llorente, M. S. Otero-Giráldez, and J. de Uña-Alvarez (University of Vigo, Spain) for providing the Galician unemployment data.

References

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.*

ATKINSON, A. C. (1985): *Plots, transformations and regression : an introduction to graphical methods of diagnostic.* Clarendon Press, Oxford.

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.*

WANG, M.-C. (1991): Nonparametric estimation from cross-sectional survival data. *Journal of the American Statistical Association 86, 130-143*