Testing for one-sided alternatives in nonparametric censored regression

Cédric Heuchenne

QuantOM[*], *HEC-Management School of University of Liège*

Université de Liège

Juan Carlos Pardo-Fernández

Departamento de Estatística e IO

Universidade de Vigo

Abstract

Assume that we have two populations (X_1, Y_1) and (X_2, Y_2) satisfying two general nonparametric regression models $Y_j = m_j(X_j) + \varepsilon_j$, $j = 1, 2$, where $m(\cdot)$ is a smooth location function, ε_j has zero location and the response Y_j is possibly right-censored. In this paper, we propose to test the null hypothesis $H_0 : m_1 = m_2$ versus the one-sided alternative $H_1 : m_1 < m_2$. We introduce two test statistics for which we obtain the asymptotic normality under the null and the alternative hypotheses. Although the tests are based on nonparametric techniques, they can detect any local alternative converging to the null hypothesis at the parametric rate $n^{-1/2}$. The practical performance of a bootstrap version of the tests is investigated in a simulation study. An application to a data set about unemployment duration times is also included.

Key Words: Bootstrap; Comparison of regression curves; Kernel estimation; Nonparametric regression; Nonparametric regression residuals; Right censoring; Survival analysis.

[*]Centre for Quantitative Methods and Operations Management

1 Introduction and motivation of the test

Comparing two populations is an important problem in statistics. When several variables enter the study in each population, regression models are commonly used to describe the relationship between two or more of them. In that case it is interesting to compare the corresponding regression models.

In this article we work with two fully nonparametric regression models, which describe the relation between a response variable and a covariate. We also assume that the responses may be right censored.

More precisely, let (X_1, Y_1) and (X_2, Y_2) be two independent pairs of variables, where, for $j = 1, 2$, the relation between the response variable, Y_j, and the covariate, X_j, is established in terms of a general nonparametric regression model of the type

$$Y_j = m_j(X_j) + \varepsilon_j, \tag{1.1}$$

where m_j is an unknown conditional location function and ε_j is a regression error with zero conditional location (see (1.3) hereunder). Note that no structure is imposed between the error and the covariate, so the model includes heteroscedasticity and covariate-dependent errors.

In each population, the response variable may be right censored. This means that there exists a censoring variable, C_j, such that only the minimum of Y_j and C_j is observable. Hence, the observations will not come from the pair (X_j, Y_j), but from the vector (X_j, Z_j, Δ_j), where $Z_j = \min\{Y_j, C_j\}$ and $\Delta_j = I(Y_j \leq C_j)$, $j = 1, 2$. ($I(\cdot)$ denotes the indicator function.)

Several regression models for censored data have been considered in the literature. For our purposes, we will assume that the response variable, Y_j, and the censoring variable, C_j, are independent, given a value of the covariate, $X_j = x$. Under this censoring model, the conditional location function of model (1.1) is defined as

$$m_j(x) = \int_0^1 F_j^{-1}(s|x) J(s) ds, \tag{1.2}$$

where $F_j(\cdot|x)$ is the conditional distribution of Y_j given $X_j = x$, $F_j^{-1}(s|x) = \inf\{t; F_j(t|x) \geq s\}$ is the corresponding quantile function, and $J(s)$ is a given score function satisfying $\int_0^1 J(s)ds = 1$. This definition implies that the conditional location of the errors in (1.1) is

$$\int_0^1 F_j^{\varepsilon^{-1}}(s|x) J(s) ds = 0, \tag{1.3}$$

where $F_j^{\varepsilon}(y|x) = P(Y_j - m_j(x) \leq y | X_j = x)$, $j = 1, 2$. Different choices of the function J lead to different conditional location functions. In particular, if $J(s) = I(0 \leq s \leq 1)$,

then $m_j(x) = E(Y_j|X_j = x)$ is the conditional mean function. However, it may happen that this choice for the function J is not appropriate because of the inconsistency of the estimator of the conditional distribution $F_j(\cdot|x)$ in the right tail due to the censoring. A useful choice is $J(s) = (q-p)^{-1}I(p \leq s \leq q)$, with $0 \leq p < q \leq 1$, which leads to trimmed means. The conditional median or other conditional quantiles can be seen as limits of trimmed means.

The development of analytical tools is especially motivated when censoring is present in the data. Indeed, as pointed out in Fan and Gijbels (1994), visual tools for regression (scatter plots, residuals plots, etc.) are not directly applicable to check the shape of the regression curves due to censoring. For instance, when comparing models which are clearly different (one above the other one), the censoring mechanism may transform the data from the first model in small values, as if those data were generated by the second model, making impossible any visual decision about equality of curves.

In this article we will introduce a test for the equality of the regression curves given in model (1.1) against one-sided alternatives. In order to make this comparison, we assume that the covariates X_1 and X_2 have common support, R_X, and that $m_1(x) \leq m_2(x)$ for all $x \in R_X$. In the test, the null hypothesis states that the regression curves given in model (1.1) are equal

$$H_0 : m_1(x) = m_2(x) \text{ for all } x \in R_X, \tag{1.4}$$

and the alternative hypothesis states that one of the curves is above the other one

$$H_1 : m_1(x) < m_2(x) \text{ on some open interval of } R_X. \tag{1.5}$$

The practical motivation of this kind of tests comes from the fact that, in certain situations, we may have some additional information about the alternative hypothesis. For example, in an economical context, that would be the case when one of the groups to be compared receives an incentive and it is assumed that this incentive will never lead to undesired effects, but at worst will not have any impact on the considered response variables. For more details on this motivation, see the example in Section 5.

The problem of testing for the equality of regression curves in nonparametric setups has been widely treated in the literature. We can basically distinguish two classes of articles: the ones dealing with general alternatives (that is, testing for the equality versus the inequality) and papers considering one-sided alternatives. Among the first group, see for example the articles by Neumeyer and Dette (2003) Pardo-Fernández *et al.* (2007), and, more recently, Sriheda and Stute (2010). The testing procedures proposed in the first two cited papers are somehow related to the ones we will introduce in the present paper because they are based on estimation of the residuals of the regression models. In the group of articles dealing with one-sided alternatives, we can cite Hall et al. (1997), Koul

and Schick (1997, 2003), Neumeyer and Dette (2005) and Neumeyer and Pardo-Fernández (2009).

All the previous references are concerned with testing procedures for completely observed data. Up to now, very little literature has been devoted to the comparison of curves when censored data are present. In this context, Pardo-Fernández and Van Keilegom (2006) developed testing procedures for general alternatives (equality versus inequality). These authors worked under models where the response is right-censored and the error term is multiplicative (of the form $\sigma(X)\varepsilon$, where $\sigma(X)$ is an unknown scale function and ε is independent of X), and their test statistics are based on the comparison of estimators of the error distribution. In the present paper we will extend the ideas introduced in Neumeyer and Pardo-Fernández (2009) to a censored data context when one-sided comparison of curves tests are required.

The paper is organized as follows. In the next section, two testing procedures are described in detail. Section 3 summarizes the main asymptotic results, including the asymptotic normality of the proposed statistics under the null as well as under the alternative hypotheses. In Section 4, the practical performance of bootstrap versions of the proposed testing procedures are investigated by means of some simulations. Section 5 contains an application to a data set related with unemployment duration times. Finally, the Appendix contains the assumptions, functions and proofs needed to obtain the main results of Section 3.

2 Description of the testing procedures

As mentioned in Section 1, we now develop two methods that enable to test H_0 versus H_1 when right censoring in the response variables is present. The first statistic tries to compute the differences between the responses of both samples and some "reference" curve while the second one directly estimates the difference between both curves. That can lead to different practical behaviors that will be displayed in Section 4.

The first procedure is based on an extension of Neumeyer and Pardo-Fernández (2009) to the censored data case. First, we will explain the basic ideas of the test before introducing the estimators needed to construct the final test statistic. Let us define $m_R(\cdot)$ as some "reference" curve satisfying $m_1(x) \leq m_R(x) \leq m_2(x)$ for all $x \in R_X$ and

$$\varepsilon_j^R = Y_j - m_R(X_j),$$

which can also be expressed as

$$\varepsilon_j^R = \varepsilon_j + (m_j(X_j) - m_R(X_j)).$$

Under the null hypothesis, we clearly have $\varepsilon_j^R = \varepsilon_j$ since $m_j(X_j) = m_R(X_j)$, $j = 1, 2$. This implies

$$E[(Y_j - m_R(X_j))J(F_j(Y_j|X_j))] = E[m_j(X_j) - m_R(X_j)] = 0,\ j = 1, 2,$$

where (1.2) and the definition of $J(\cdot)$ are used (see Section 1). However, under the alternative hypothesis, we have either

$$E[\varepsilon_2^R J(F_2(Y_2|X_2))] > 0 \ \text{ and } \ E[\varepsilon_1^R J(F_1(Y_1|X_1))] \leq 0,$$

or

$$E[\varepsilon_1^R J(F_1(Y_1|X_1))] < 0 \ \text{ and } \ E[\varepsilon_2^R J(F_2(Y_2|X_2))] \geq 0,$$

since under H_1, $m_2(x) > m_1(x)$ on some open interval of $x \in R_X$. That suggests to construct a test statistic based on the difference

$$E[w_2(X_2)\varepsilon_2^R J(F_2(Y_2|X_2))] - E[w_1(X_1)\varepsilon_1^R J(F_1(Y_1|X_1))], \tag{2.1}$$

where $w_j(\cdot)$, $j = 1, 2$, are given positive weight functions. Indeed, the above quantity becomes greater and greater when H_1 goes away from H_0.

Now, for $j = 1, 2$, suppose that we have n_j i.i.d. replications $\{(X_{ij}, Z_{ij}, \Delta_{ij}), i = 1, \ldots, n_j\}$ of (X_j, Z_j, Δ_j). Let $n = n_1 + n_2$. Since Y_j, $j = 1, 2$, is possibly right-censored, we cannot estimate the two expectations of (2.1) with simple averages. We therefore use the idea of artificial data points already developed, for example, in Heuchenne and Van Keilegom (2007). In the present case, it consists of the following steps:

1. replacing each unknown censored $\varepsilon_{ij}^R J(F_j(Y_{ij}|X_{ij}))$ (where $\varepsilon_{ij}^R = Y_{ij} - m_R(X_{ij})$) by a quantity that takes censoring into account,

2. introducing the obtained data points into a classical estimator for the mean (here the simple average).

To achieve step 1 above, define

$$\epsilon(x, z, \delta, F) = \left\{ \delta z J(F(z|x)) + (1 - \delta)\frac{\int_z^{+\infty} y J(F(y|x)) dF(y|x)}{1 - F(z|x)} \right\},$$

and note that

$$
\begin{aligned}
&E[\epsilon(X_j, Z_j - m_R(X_j), \Delta_j, F_j^{\varepsilon_R})|X_j] \\
&= \int_{-\infty}^{+\infty} \int_{m_R(X_j)+e}^{+\infty} eJ(F_j^{\varepsilon_R}(e|X_j))dG_j(u|X_j)dF_j^{\varepsilon_R}(e|X_j) \\
&\quad + \int_{-\infty}^{+\infty} \int_{u-m_R(X_j)}^{+\infty} \frac{\int_{u-m_R(X_j)}^{+\infty} yJ(F_j^{\varepsilon_R}(y|X_j))dF_j^{\varepsilon_R}(y|X_j)}{1 - F_j^{\varepsilon_R}(u - m_R(X_j)|X_j)} dF_j^{\varepsilon_R}(e|X_j)dG_j(u|X_j) \\
&= \int_{-\infty}^{+\infty} eJ(F_j^{\varepsilon_R}(e|X_j))[1 - G_j(m_R(X_j) + e|X_j)]dF_j^{\varepsilon_R}(e|X_j) \\
&\quad + \int_{-\infty}^{+\infty} \int_{-\infty}^{m_R(X_j)+e} eJ(F_j^{\varepsilon_R}(e|X_j))dG_j(u|X_j)dF_j^{\varepsilon_R}(e|X_j) \\
&= E[\varepsilon_j^R J(F_j(Y_j|X_j))|X_j] \qquad (2.2)
\end{aligned}
$$

for $F_j^{\varepsilon_R}(y|X_j) = P(\varepsilon_j^R \leq y|X_j)$ and $G_j(u|X_j) = P(C_j \leq u|X_j)$, $j = 1, 2$. As a consequence, replacing $\varepsilon_j^R J(F_j(Y_j|X_j))$ for $j = 1, 2$, by $\epsilon(X_j, Z_j - m_R(X_j), \Delta_j, F_j^{\varepsilon_R})$ in (2.1) will not change its value.

Next, we estimate the quantities in (2.2). The distribution $F_j(y|x)$, which is needed to estimate $F_j^{\varepsilon_R}(y|X_j) = F_j(y + m_R(X_j)|X_j)$, is replaced by the Beran (1981) estimator, defined by (in the case of no ties):

$$\hat{F}_j(y|x) = 1 - \prod_{Z_{ij} \leq y, \Delta_{ij}=1} \Big\{1 - \frac{W_{ij}(x, a_n)}{\sum_{k=1}^{n_j} I(Z_{kj} \geq Z_{ij}) W_{kj}(x, a_n)}\Big\}, \qquad (2.3)$$

where

$$W_{ij}(x, a_n) = \frac{K\left(\frac{x - X_{ij}}{a_n}\right)}{\sum_{k=1}^{n_j} K\left(\frac{x - X_{kj}}{a_n}\right)},$$

K is a kernel function and $\{a_n\}$ is a bandwidth sequence. For the sake of simplicity we present the method and the theory with a single bandwidth a_n, but in practice two bandwidth sequences are required, one for each population (say a_{1n} and a_{2n}).

Given a deterministic function p such that $0 \leq p(x) \leq 1$ for all $x \in R_X$, we propose to estimate $m_R(x)$ by

$$\hat{m}_R(x) = p(x)\hat{m}_1(x) + (1 - p(x))\hat{m}_2(x),$$

where

$$\hat{m}_j(x) = \int_0^1 \hat{F}_j^{-1}(s|x)J(s)\, ds \qquad (2.4)$$

estimates $m_j(x)$ and $\hat{F}_j^{-1}(s|x) = \inf\{t; \hat{F}_j(t|x) \geq s\}$, $j = 1, 2$. That leads to the test statistic

$$T_{n1} = \Big(\frac{n_1 n_2}{n}\Big)^{1/2} \sum_{j=1,2} (-1)^j n_j^{-1} \sum_{i=1}^{n_j} w_j(X_{ij}) \epsilon(X_{ij}, Z_{ij} - \hat{m}_R(X_{ij}), \Delta_{ij}, \hat{F}_j^{\varepsilon_R}), \tag{2.5}$$

where $\hat{F}_j^{\varepsilon_R}(y - \hat{m}_R(X_{ij})|X_{ij}) = \hat{F}_j(y|X_{ij})$, for all y.

Finally, the second procedure can be simply seen as a weighted sum of differences between both curves at all the data points

$$T_{n2} = \Big(\frac{n_1 n_2}{n}\Big)^{1/2} \sum_{j=1,2} n_j^{-1} \sum_{i=1}^{n_j} w_j(X_{ij})(\hat{m}_2(X_{ij}) - \hat{m}_1(X_{ij})). \tag{2.6}$$

As it turns out, both statistics estimate

$$D = \Big(\frac{n_1 n_2}{n}\Big)^{1/2} \int_{R_X} (m_2(x) - m_1(x)) f(x) dx, \tag{2.7}$$

where either

$$f(x) = f_{X_1}(x)(1 - p(x))w_1(x) + f_{X_2}(x)p(x)w_2(x)$$

for T_{n1}, or

$$f(x) = f_{X_1}(x)w_1(x) + f_{X_2}(x)w_2(x)$$

for T_{n2}, and $f_{X_j}(\cdot)$ denotes the density of X_j, $j = 1, 2$. Obviously, both $f(x)$ can be made equal by appropriate choices of w_j and p, $j = 1, 2$.

3 Asymptotic theory

In the next results, we obtain the asymptotic distributions of the test statistics under the alternative hypothesis H_1 and local alternatives of the type

$$H_{1n} : m_2(x) = m_1(x) + n^{-1/2} r(x), \text{ where } r(x) \geq 0 \text{ for all } x \in R_X, \tag{3.1}$$

which includes the null hypothesis H_0 when $r = 0$.

The assumptions mentioned in the results below, as well as their proofs, are given in the Appendix.

Theorem 1 *Assume (A1)–(A5), $n_j/n \to \kappa_j$, $j = 1, 2$, and the function $p(\cdot)$ is twice continuously differentiable. Under H_{1n},*

$$T_{n1} \xrightarrow{L} N((\kappa_1 \kappa_2)^{1/2} d, \gamma_0^2),$$

where

$$d = \int r(x)f(x)dx < \infty,$$

$$\begin{aligned}\gamma_0^2 \;=\; & E[\kappa_1(w_2(X_2)\epsilon(X_2, Z_2 - m_2(X_2), \Delta_2, F_2^{\varepsilon}) - \phi_2(X_2, Z_2, \Delta_2))^2 \\ & +\kappa_2(w_1(X_1)\epsilon(X_1, Z_1 - m_1(X_1), \Delta_1, F_1^{\varepsilon}) - \phi_1(X_1, Z_1, \Delta_1))^2]\end{aligned}$$

and $\phi_j(x, z, \delta)$, $j = 1, 2$, *is defined in the Appendix. Under* H_1,

$$T_{n1} - D \xrightarrow{L} N(0, \gamma_1^2),$$

where

$$\begin{aligned}\gamma_1^2 \;=\; & \kappa_1 Var[w_2(X_2)\epsilon(X_2, Z_2 - m_R(X_2), \Delta_2, F_2^{\varepsilon_R}) - \phi_2(X_2, Z_2, \Delta_2)] \\ & +\kappa_2 Var[w_1(X_1)\epsilon(X_1, Z_1 - m_R(X_1), \Delta_1, F_1^{\varepsilon_R}) - \phi_1(X_1, Z_1, \Delta_1)].\end{aligned}$$

Theorem 2 *Assume (A1), (A2) (i), the function* $J(\cdot)$ *is continuously differentiable, (A3)–(A5) and* $n_j/n \to \kappa_j$, $j = 1, 2$. *Under* H_{1n},

$$T_{n2} \xrightarrow{L} N((\kappa_1\kappa_2)^{1/2}d, \tau_0^2),$$

where

$$\tau_0^2 = E[f^2(X_2)\kappa_1\eta_2^2(Z_2, \Delta_2|X_2) + f^2(X_1)\kappa_2\eta_1^2(Z_1, \Delta_1|X_1)]$$

and $\eta_j(z, \delta|x)$, $j = 1, 2$, *is defined in the Appendix. Under* H_1,

$$T_{n2} - D \xrightarrow{L} N(0, \tau_0^2 + \tau_1^2),$$

where $\tau_1^2 = \kappa_1 Var[w_2(X_2)(m_2(X_2) - m_1(X_2))] + \kappa_2 Var[w_1(X_1)(m_2(X_1) - m_1(X_1))]$.

Remark 1. The previous theorems give the asymptotic distributions of the test statistics T_{n1} and T_{n2} under both null and alternative hypotheses. The null hypothesis corresponds to $r(x) = 0$, for all $x \in R_X$, which implies $d = 0$, and hence the asymptotic distribution of T_{n1} is $N(0, \gamma_0^2)$ and the asymptotic distribution of T_{n2} is $N(0, \tau_0^2)$. Therefore, the test that rejects the null hypothesis H_0 versus the alternative H_1 when the observed value of T_{n1} (respectively, T_{n2}) is larger than $z_{1-\alpha}\gamma_0$ (respectively, $z_{1-\alpha}\tau_0$), has asymptotic significance level α, where $z_{1-\alpha}$ denotes the $(1-\alpha)$-quantile of the standard normal distribution. On the other hand, the above theorems show that our tests can detect any local alternative, H_{1n}, converging to the null hypotheses at the parametric rate $n^{-1/2}$, since d is positive in that case. Finally, under the alternative hypotheses H_1, the theorems show that both test

statistics are shifted and both asymptotic variances are modified due to the introduction of the differences between the curves. Also note that the testing procedures are only reasonable when $\kappa_1 > 0$ and $\kappa_2 > 0$.

Remark 2. The proposed testing procedures can be obviously extended to the case where the order of the curves under the alternative is unspecified. In this context, we need to assume either $m_2(x) \geq m_1(x)$ for all $x \in R_X$, or $m_2(x) \leq m_1(x)$ for all $x \in R_X$. Under the null hypothesis, $H_0 : m_1 = m_2$ and under the alternative

$$H_1 : m_2 > m_1 \text{ or } m_1 > m_2$$

on some open interval of R_X. The test statistics given in (2.5) and (2.6) may still be used but at the significance level α, the null hypothesis is rejected when their values are either larger than $z_{1-\alpha/2}\gamma_0$ (or $z_{1-\alpha/2}\tau_0$) or smaller than $z_{\alpha/2}\gamma_0$ (or $z_{\alpha/2}\tau_0$).

Remark 3. In some situations, as explained in Section 1, it may happen that the conditional location function (1.2) for a given function $J(\cdot)$ cannot be consistently estimated due to the presence of censoring. It is typically the case if two conditional means have to be compared, with $J(s) = I(0 \leq s \leq 1)$. However, this problem can be avoided in many situations if the models (1.1) satisfy stronger assumptions. For example, in the classical homoscedastic case, where the error distribution is the same in both models and independent of the covariate, the null hypothesis is equivalent to the equality of two regression curves with a function $\tilde{J}(\cdot)$ (having the same properties as $J(\cdot)$) chosen in an appropriate way. Indeed, for $j = 1, 2$, consider the models $Y_j = m_j(X_j) + \varepsilon_j$, where $m_j(x) = \int_0^1 F_j^{-1}(s|x)J(s)ds$ and ε_j is independent of X_j, and suppose that ε_1 and ε_2 have the same distribution F_ε. In that case, it easy to prove that $F_j^{-1}(s|x) = m_j(x) + F_\varepsilon^{-1}(s)$. Now consider a score function $\tilde{J}(\cdot)$. The conditional location functions defined with respect to $\tilde{J}(\cdot)$ are, for $j = 1, 2$,

$$\tilde{m}_j(x) = \int_0^1 F_j^{-1}(s|x)\tilde{J}(s)ds = \int_0^1 (m_j(x) + F_\varepsilon^{-1}(s))\tilde{J}(s)ds = m_j(x) + c$$

where $c = \int_0^1 F_\varepsilon^{-1}(s)\tilde{J}(s)ds$. This means that if the null hypothesis holds for a particular choice of the score function $J(\cdot)$, it will necessarily hold for any other choice of $J(\cdot)$. Therefore this function can be chosen in a convenient way in order to have consistent estimators of the location functions.

Remark 4. An interesting extension is the problem of testing the equality of k regression curves versus an ordered alternative. We have

$$H_0 : m_1 = m_2 = m_3 = \cdots = m_k,$$

versus

$$H_1 : m_1 \leq m_2 \leq m_3 \leq \cdots \leq m_k,$$

where some of the above inequalities have to be strict on some open interval. A direct way to generalize our procedures would be to define a "reference" curve between two consecutive curves:

$$m_R(x) = p(x)m_j(x) + (1 - p(x))m_{j+1}(x),$$

for some particular j, $1 \leq j \leq k-1$. This choice guarantees that

$$E[\varepsilon_1^R J(F_1(Y_1|X_1))] \leq E[\varepsilon_2^R J(F_2(Y_2|X_2))] \leq \ldots \leq E[\varepsilon_k^R J(F_k(Y_k|X_k))],$$

with some of the inequalities being strict under the alternative hypothesis (and the same is obviously true for $E[m_s(X_s) - m_R(X_s)]$, $s = 1, \ldots, k$). Therefore, based on these relations, estimators of the above quantities (computed in the same way as in Section 2) can be constructed and compared to test the equality of the k curves versus an ordered alternative.

Remark 5. In order to use the testing procedures described above, the variances of the statistics T_{n1} and T_{n2} under the null hypothesis are needed. Unfortunately, those variances are too complicated and contain too many unknown quantities. Alternatively, we use a bootstrap procedure to estimate the critical values of the tests in practical situations. The method is based on a smoothed version of the naive bootstrap described in Efron (1981) adapted to the regression context.

The main idea of the smooth bootstrap consists of replacing the regular bootstrap observations resampled from an empirical estimator of a distribution function by observations resampled from an estimator of the corresponding density. Suppose that bootstrap resamples are needed from a general distribution estimator $\hat{H}$. Under the naive bootstrap, the bootstrap observations, say ν_b^*, are simulated according to the discrete probability mass function given by $\hat{H}$. Instead, under the smooth bootstrap, the bootstrap observations are of the form $\nu_b^* + \omega S_b$, where ν_b^* are drawn from $\hat{H}$, S_b are random variables with mean zero and variance one simulated independently from ν_b^*, and ω is a small constant. The term ωS_b adds a small perturbation to the original bootstrap observation, and makes the procedure equivalent to resampling from a density estimator.

In our case, the general idea described above will be applied to the residuals of the regression models. Basically, we will need to reconstruct the regression models on the basis of bootstrap samples. For that purpose, in each population, we need to draw resamples of residuals from the estimator of the conditional distribution of the error given the covariate.

The idea of the smooth bootstrap appears when those bootstrap residuals are drawn from a smoothed version of the empirical estimators of the conditional distribution by using the method described above.

More precisely, the bootstrap algorithm for our testing procedure consists of the following steps. For B fixed and $b = 1, \ldots, B$,

1. For $j = 1, 2$ and $i = 1, \ldots, n_j$:

   - Let $Y^*_{ijb} = \hat{m}_R(X_{ij}) + \hat{\varepsilon}^*_{ijb}$, where $\hat{\varepsilon}^*_{ijb}$ is drawn from a smoothed version of the empirical distribution $\hat{F}^{\varepsilon}_j(\cdot|X_{ij})$, as described above (that is, $\hat{\varepsilon}^*_{ijb} = \hat{\nu}^*_{ijb} + \omega_j S_{ijb}$, with $\hat{\nu}^*_{ijb}$ simulated from $\hat{F}^{\varepsilon}_j(\cdot|X_{ij})$, S_{ijb} is a random variable with mean zero and variance one, and ω_j is a small constant).
   - Similarly, select C^*_{ijb} from a smoothed version of $\hat{G}_j(\cdot|X_{ij})$, the Beran (1981) estimator of the distribution $G_j(\cdot|X_{ij}) = P(C_j \leq \cdot|X_{ij})$ obtained by replacing Δ_{ij} by $1 - \Delta_{ij}$ in the expression of $\hat{F}_j(\cdot|X_{ij})$.
   - Let $Z^*_{ijb} = \min(Y^*_{ijb}, C^*_{ijb})$ and $\Delta^*_{ijb} = I(Y^*_{ijb} \leq C^*_{ijb})$.

2. The two bootstrap samples are $\{(X_{ij}, Z^*_{ijb}, \Delta^*_{ijb}),\ i = 1, \ldots, n_j\}$ for $j = 1, 2$.

3. Compute T^*_{n1b} and T^*_{n2b}, the test statistics calculated with both above bootstrap samples.

Let $T^*_{n1(b)}$ be the $b-$th order statistic of $T^*_{n11}, \ldots, T^*_{n1B}$, and analogously for $T^*_{n2(b)}$. Then $T^*_{n1(\lfloor(1-\alpha)B\rfloor+1)}$ and $T^*_{n2(\lfloor(1-\alpha)B\rfloor+1)}$ (where $\lfloor\cdot\rfloor$ denotes the integer part) approximate the $(1-\alpha)-$quantiles of the distributions of T_{n1} and T_{n2} under H_0, respectively. Note that the bootstrap samples are constructed under the null hypothesis.

Similar bootstrap procedures were used in other testing problems in nonparametric regression (see, for instance, Pardo-Fernández *et al.*, 2007, or Pardo-Fernández and Van Keilegom, 2006). Neumeyer (2006, 2009) proved the consistency of this type of smooth bootstrap mechanisms in nonparametric regression problems with complete data. A formal proof of the consistency of the described bootstrap in the present context of censored regression is beyond the scope of the present article, although we believe that some of the ideas used in Neumeyer (2009) could be borrowed to the present setting.

4 Simulations

In this section we present the results of a simulation study which illustrates the practical performance of the proposed tests based on the bootstrap mechanism described in Remark 5. For the sake of simplicity, the function p is chosen to be the constant $p(x) = 0.5$, and the weight functions are $w_1(x) = w_2(x) = 1$, for all x.

We choose $J(s) = 0.75^{-1} I(0 \leq s \leq 0.75)$ as the score function in the definition of the conditional location functions, which corresponds to conditional trimmed means. The regression models considered in the simulations are:

$$
\begin{array}{lll}
(i) & m_1(x) = \sin(2\pi x); & m_2(x) = \sin(2\pi x) \\
(ii) & m_1(x) = x; & m_2(x) = x + 0.25 \\
(iii) & m_1(x) = 0; & m_2(x) = 0.5x \\
(iv) & m_1(x) = \sin(2\pi x); & m_2(x) = \sin(2\pi x) + 0.5x \\
(v) & m_1(x) = 0; & m_2(x) = 0.5 \exp\{-100(x-0.5)^2\}
\end{array}
$$

Model (i) corresponds to the null hypothesis, while models (ii), (iii), (iv) and (v) correspond to the alternative hypothesis.

In each population ($j = 1$ or 2), the distribution of the covariates is Uniform in $[0, 1]$. The conditional distribution of the errors ε_j given $X_j = x$ is exponential of parameter $\lambda_j(x)$, recentered in such a way that $\int_0^1 F_j^{\varepsilon^{-1}}(s|x) J(s) ds = 0$, as established in (1.3). In each case we consider a homoscedastic and a heteroscedastic situation. In the homoscedastic case the parameters of the conditional distribution of the errors are

$$\lambda_1(x) = 0.50^{-1} \quad \text{and} \quad \lambda_2(x) = 0.75^{-1}, \tag{4.1}$$

while in the heteroscedastic case we set

$$\lambda_1(x) = (0.25 + 0.50x)^{-1} \quad \text{and} \quad \lambda_2(x) = (0.50 + 0.50x)^{-1}. \tag{4.2}$$

The censoring variables are given by the model $C_j = m_j(X_j) + \rho_j$, where the conditional survival function of ρ_j given that $X_j = x$ is chosen to be $1 - F_j^{\rho}(y|x) = (1 - F_j^{\varepsilon}(y|x))^{\beta_j}$, with $\beta_j > 0$. The value β_j controls the amount of censored data. In fact, this censoring mechanism allows us to have the same amount of censoring over the whole support of the covariates and it is easy to see that the expected proportion of uncensored data in each population is $(1 + \beta_j)^{-1}$. The case $\beta_j = 1/3$ (25% of censoring) will be considered in the simulations. Note that the choice of the function J is reasonable under the proposed censoring mechanism.

The tables display the observed proportion of rejections in 1000 trials for sample sizes $(n_1, n_2) = (50, 50)$, $(50, 100)$ and $(100, 100)$. In all cases we worked with $B = 100$ bootstrap replications and significance levels $\alpha = 0.05$ and $\alpha = 0.10$. For the kernel needed to calculate the weights that appear in the Beran estimator, we choose the kernel of Epanechnikov $K(u) = 0.75(1 - u^2) I(|u| < 1)$.

Concerning the choice of the smoothing parameters a_{jn}, in the tables we show the results obtained for two choices of the bandwidths, depending on the sample sizes: $a_j = 1.5 n_j^{-3/10}$ and $a_j = 2 n_j^{-3/10}$. For $n_j = 50$, the bandwidths are 0.46 and 0.62, respectively; for $n_j = 100$, the bandwidths are 0.38 and 0.50, respectively. The selection of the

smoothing parameters is still an open question in most testing problems in nonparametric regression setups, and so far there is no automatic method to choose them, especially in a censored data setup. In practical applications, as we will do in the next section, the test can be performed for a large range of bandwidths and a decision can be made on the basis of the resulting p-values.

The variables S_{ijb} needed in the smooth version of the bootstrap are constructed with a standard normal random variable multiplied by $\omega_j = 0.1a_{jn}$, where a_{jn} is the bandwidth used to construct the Nadaraya-Watson weights in each population. With these choices, a small perturbation is added to the values resampled from the empirical distribution functions, as recommended in the smooth bootstrap.

Table 1 displays the results of the tests for models $(i) - (v)$ under the assumption of homoscedasticity, according to (4.1). For model (i), the level is well approximated for both choices of the bandwidth. On the other hand, the tests reach reasonable power under models $(ii) - (v)$. The power increases as the sample sizes increase. The results for both test statistics are very similar in most cases, although the test based on T_{n2} seems to reach a slightly better power than the test based on T_{n1}. Similar comments can be made for Table 2, which displays the analogous results under the assumption of heteroscedasticity, according to (4.2).

[Tables 1 and 2 (at the end of the manuscript) to be placed around here]

5 Data analysis

In this section, we use the proposed methods to analyze a data set related to unemployment duration times. The survey *Encuesta de Población Activa* (Labour Force Survey) is carried out by the Instituto Nacional de Estadística (Spanish bureau for official statistics) to collect information about employment in Spain. About 60000 homes are surveyed each three months for this purpose. Here, the available information corresponds to unemployed married women in the autonomous region of Galicia (NW of Spain) collected during the period 1987–1997. For each woman, we observe a certain number of variables, including her age at the entrance in the study. In total, 1007 observations are considered, with ages ranging from 19 to 61. Besides, we also know if each woman receives or not any type of public unemployment subsidy during the observation period.

Each unemployed woman is followed up for the next 18 months after her entrance in the survey. If a woman is still unemployed when the follow-up ends, then a censored observation appears because the true unemployment duration time cannot be observed. In this data set, 563 observations are censored (55.9%) out of the 1007.

In this framework, if we assume that subsidy can never increase unemployment duration time (due to characteristics of the different subsidies), the following question can be asked: for each age, does the subsidy have a global significant positive effect on the unemployment duration time? To answer this question, we create two groups of observations according to the fact if the woman receives or not a subsidy, and we consider her age as the covariable. The group of women who receive a subsidy consists of 284 observations (143 censored), with ages ranging from 20 to 59. On the other hand, the group of women who do not receive any subsidy consists of 723 observations (420 censored), with ages ranging from 19 to 61. Figure 1 shows the scatter plot of the response variable 'unemployment duration time' (measured in days) versus the covariate 'age' (measured in years). The estimated conditional location functions for each group with $J(s) = 0.75^{-1} I(0 \leq s \leq 0.75)$ are also included in the graph, obtained with bandwidth equal to 15 in both groups.

We have performed the test based on the statistics T_{n1} and T_{n2} for ten values of the bandwidths covering a reasonable range: $11, 12, \ldots, 20$. The p-values were calculated with 500 bootstrap replications. In all cases, the p-values were less than 0.002. Since several tests are performed, one should apply some correction to prevent troubles from multiple testing. In our case, even if the very conservative Bonferroni's correction is applied to take into account for multiple testing, the null hypothesis is rejected at a level 0.05. This fact confirms the difference between the curves in this example, that is, for each age, the general behavior of the unemployment duration time for the group of women not receiving any subsidy is significantly globally greater than the corresponding time in the group of women who receive the subsidy.

[Figure 1 (at the end of the manuscript) to be placed around here]

Appendix

The following notations are needed in the statement of the asymptotic results given in Section 3.

$$\xi_j(z,\delta,y|x) = (1-F_j(y|x))\left\{-\int_{-\infty}^{y\wedge z}\frac{dH_{1,j}(s|x)}{(1-H_j(s|x))^2}+\frac{I(z\le y,\delta=1)}{1-H_j(z|x)}\right\},$$

$$\eta_j(z,\delta|x) = -f_{X_j}^{-1}(x)\int_{-\infty}^{+\infty}\xi_j(z,\delta,v|x)J(F_j(v|x))dv,$$

$$\rho_{1,j}(v^1,v^2) = -w_j(x^1)\left[\delta^1 J(F_j(z^1|x^1))+(1-\delta^1)\frac{\int_{z^1}^{+\infty}J(F_j(y|x^1))dF_j(y|x^1)}{1-F_j(z^1|x^1)}\right]$$
$$\times p(x^1)^{j-1}(1-p(x^1))^{2-j}\eta_{3-j}(z^2,\delta^2|x^1),$$

$$\rho_{2,j}(v^1,v^2) = w_j(x^1)\left\{\left[-\delta^1 J(F_j(z^1|x^1))-(1-\delta^1)\frac{\int_{z^1}^{+\infty}J(F_j(y|x^1))dF_j(y|x^1)}{1-F_j(z^1|x^1)}\right]\right.$$
$$\times p(x^1)^{2-j}(1-p(x^1))^{j-1}\eta_j(z^2,\delta^2|x^1)$$
$$+\left[\delta^1(z^1-m_R(x^1))J'(F_j(z^1|x^1))+(1-\delta^1)\right.$$
$$\left.\times\frac{\int_{z^1}^{+\infty}(y-m_R(x^1))J(F_j(y|x^1))dF_j(y|x^1)}{(1-F_j(z^1|x^1))^2}\right]\xi_j(z^2,\delta^2,z^1|x^1)$$
$$+(1-\delta^1)\left[\frac{\int_{z^1}^{+\infty}(y-m_R(x^1))J(F_j(y|x^1))d\xi_j(z^2,\delta^2,y|x^1)}{1-F_j(z^1|x^1)}\right.$$
$$\left.\left.+\frac{\int_{z^1}^{+\infty}(y-m_R(x^1))J'(F_j(y|x^1))\xi_j(z^2,\delta^2,y|x^1)dF_j(y|x^1)}{1-F_j(z^1|x^1)}\right]\right\},$$

$$\phi_j(x,z,\delta) = \sum_{\lambda=0,1}\int\rho_{1,3-j}((x,y,\lambda),(z,\delta))f_{X_{3-j}}(x)dH_{\lambda,3-j}(y|x)$$
$$-\sum_{\lambda=0,1}\int\rho_{2,j}((x,y,\lambda),(z,\delta))f_{X_j}(x)dH_{\lambda,j}(y|x),$$

where $H_j(\cdot|x) = P(Z_j\le\cdot|x)$, $H_{\delta,j}(\cdot|x) = P(Z_j\le\cdot,\Delta_j=\delta|x)$, $j=1,2$, $\delta,\lambda=0,1$, $v^1=(x^1,z^1,\delta^1)$, $v^2=(z^2,\delta^2)$, $x^1,x^2\in R_X$, $z^1,z^2\in I\!R$, $\delta^1,\delta^2=0,1$, and $J'(s)$ denotes the first derivative of $J(s)$ with respect to s. For a (sub)distribution $L(y|x)$, we will use the notations $l(y|x)=L'(y|x)=\frac{\partial}{\partial y}L(y|x)$, $\dot{L}(y|x)=\frac{\partial}{\partial x}L(y|x)$ and similar notations will be used for higher order derivatives.

The assumptions needed for the results of Section 3 are listed below.

$(A1)(i)$ $na_n^3(\log n)^{-3}\to\infty$ and $na_n^4\to 0$.
(ii) R_X is a compact interval.

(*iii*) K is a density with compact support, $\int uK(u)du = 0$ and K is twice continuously differentiable.

$(A2)(i)$ Let T_x^j be any value less than the upper bound of the support of a random variable with distribution function $H_j(\cdot|x)$ such that $\inf_{x\in R_X}(1 - H_j(T_x^j|x)) > 0$, $j = 1, 2$. There exist $0 \leq s_{0j} \leq s_{1j} \leq 1$ such that $s_{1j} \leq \inf_x F_j(T_x^j|x)$, $s_{0j} \leq \inf\{s \in [0,1]; J(s) \neq 0\}$, $s_{1j} \geq \sup\{s \in [0,1]; J(s) \neq 0\}$ and $\inf_{x\in R_X} \inf_{s_{0j}\leq s\leq s_{1j}} f_j(F_j^{-1}(s|x)|x) > 0$, for $j = 1, 2$.
(*ii*) J is three times continuously differentiable, $\int_0^1 J(s)ds = 1$ and $J(s) \geq 0$ for all $0 \leq s \leq 1$.

$(A3)(i)$ $F_{X_j}(x)$ (the cumulative distribution function of X_j, $j = 1, 2$) is three times continuously differentiable and $\inf_{x\in R_X} f_{X_j}(x) > 0$.
(*ii*) $w_j(x)$, $j = 1, 2$, is twice continuously differentiable.

$(A4)(i)$ $L(y|x)$ is continuous,
(*ii*) $L'(y|x) = l(y|x)$ exists, is continuous in (x, y) and $\sup_{x,y} |yL'(y|x)| < \infty$,
(*iii*) $L''(y|x)$ exists, is continuous in (x, y) and $\sup_{x,y} |y^2L''(y|x)| < \infty$,
(*iv*) $\dot{L}(y|x)$ exists, is continuous in (x, y) and $\sup_{x,y} |y\dot{L}(y|x)| < \infty$,
(*v*) $\ddot{L}(y|x)$ exists, is continuous in (x, y) and $\sup_{x,y} |y^2\ddot{L}(y|x)| < \infty$,
(*vi*) $\ddot{L}'(y|x)$ exists, is continuous in (x, y) and $\sup_{x,y} |y\ddot{L}'(y|x)| < \infty$,
for $L(y|x) = H_j(y|x)$ and $H_{1,j}(y|x)$, $j = 1, 2$.

$(A5)$ For the density $f_{X_j|Z_j,\Delta_j}(x|z,\delta)$ of X_j given (Z_j, Δ_j),
(*i*) $\sup_{x,z} |f_{X_j|Z_j,\Delta_j}(x|z,\delta)| < \infty$,
(*ii*) $\sup_{x,z} |\dot{f}_{X_j|Z_j,\Delta_j}(x|z,\delta)| < \infty$,
(*iii*) $\sup_{x,z} |\ddot{f}_{X_j|Z_j,\Delta_j}(x|z,\delta)| < \infty$,
for $\delta = 0, 1$, and where $\dot{f}_{X_j|Z_j,\Delta_j}(x|z,\delta)$ and $\ddot{f}_{X_j|Z_j,\Delta_j}(x|z,\delta)$ denote respectively the first and second derivatives of $f_{X_j|Z_j,\Delta_j}(x|z,\delta)$ with respect to x.

Proof of Theorem 1. First, easy calculations show that

$$\begin{aligned}
\left(\frac{n}{n_1 n_2}\right)^{1/2} T_{n1} \;=\; & \frac{1}{n_1 n_2 a_n} \sum_{i=1}^{n_2} \sum_{j=1}^{n_1} K(\frac{X_{i2} - X_{j1}}{a_n}) \rho_{1,2}((X_{i,2}, Z_{i2}, \Delta_{i2}), (Z_{j1}, \Delta_{j1})) \\
& + \frac{1}{n_2^2 a_n} \sum_{i=1}^{n_2} \sum_{j=1}^{n_2} K(\frac{X_{i2} - X_{j2}}{a_n}) \rho_{2,2}((X_{i2}, Z_{i2}, \Delta_{i2}), (Z_{j2}, \Delta_{j2})) \\
& - \frac{1}{n_1 n_2 a_n} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} K(\frac{X_{i1} - X_{j2}}{a_n}) \rho_{1,1}((X_{i1}, Z_{i1}, \Delta_{i1}), (Z_{j2}, \Delta_{j2})) \\
& - \frac{1}{n_1^2 a_n} \sum_{i=1}^{n_1} \sum_{j=1}^{n_1} K(\frac{X_{i1} - X_{j1}}{a_n}) \rho_{2,1}((X_{i1}, Z_{i1}, \Delta_{i1}), (Z_{j1}, \Delta_{j1})) \\
& + \sum_{j=1,2} \frac{1}{n_j} \sum_{i=1}^{n_j} (-1)^j w_j(X_{ij}) \left[\Delta_{ij}(Z_{ij} - m_R(X_{ij})) J(F_j(Z_{ij}|X_{ij})) \right. \\
& \qquad \left. + (1 - \Delta_{ij}) \frac{\int_{Z_{ij}}^{+\infty} (y - m_R(X_{ij})) J(F_j(y|X_{ij})) dF_j(y|X_{ij})}{1 - F_j(Z_{ij}|X_{ij})} \right] \\
& + o_P(n^{-1/2}) \\
=\; & E_{1,2}^n + E_{2,2}^n - E_{1,1}^n - E_{2,1}^n + E_3^n + o_P(n^{-1/2}).
\end{aligned}$$

$E_{1,2}^n + E_{2,2}^n - E_{1,1}^n - E_{2,1}^n$ consists of the random terms caused by the estimators $\hat{F}_j(y|x)$, $j = 1, 2$, and $\hat{m}_R(x)$ in (2.5). They are therefore obtained by introducing $F_j(y|x)$, $j = 1, 2$, and $m_R(x)$ in (2.5) and using uniform consistency and asymptotic representations for the estimators of those quantities. Those properties can be found (with small adaptations), for example, in Propositions 4.3, 4.5 and 4.8 of Van Keilegom and Akritas (1999) and

Theorem 2.3 of González-Manteiga and Cadarso-Suárez (1994). Next, it is clear that

$$
\begin{aligned}
E_3^n \;=\; & \sum_{j=1,2} \frac{1}{n_j} \sum_{i=1}^{n_j} (-1)^j \{ w_j(X_{ij}) \, [\Delta_{ij} (m_j(X_{ij}) - m_R(X_{ij})) J(F_j(Z_{ij}|X_{ij})) \\
& \qquad + (1 - \Delta_{ij})(m_j(X_{ij}) - m_R(X_{ij})) \frac{\int_{Z_{ij}}^{+\infty} J(F_j(y|X_{ij})) dF_j(y|X_{ij})}{1 - F_j(Z_{ij}|X_{ij})} \Big] \\
& \qquad - E[w_j(X_j)(m_j(X_j) - m_R(X_j))] \} \\
& + \sum_{j=1,2} \frac{1}{n_j} \sum_{i=1}^{n_j} (-1)^j w_j(X_{ij}) \, [\Delta_{ij} \varepsilon_{ij} J(F_j(Z_{ij}|X_{ij})) + (1 - \Delta_{ij}) \\
& \qquad \times \frac{\int_{Z_{ij} - m_j(X_{ij})}^{+\infty} e J(F_j^{\varepsilon}(e|X_{ij})) dF_j^{\varepsilon}(e|X_{ij})}{1 - F_j(Z_{ij}|X_{ij})} \Big] \\
& + \int_{R_X} (m_2(x) - m_1(x)) f(x) dx, \\
= \; & E_{3,1}^n + E_{3,2}^n + E_{3,3},
\end{aligned}
$$

where $f(x) = w_2(x) f_{X_2}(x) p(x) + w_1(x) f_{X_1}(x)(1 - p(x))$. Note that all the terms of the above expression are sums of zero mean i.i.d. random variables except the last one.

Next, we treat $E_{1,2}^n$, $E_{2,2}^n$, $E_{1,1}^n$ and $E_{2,1}^n$. Since $E_{2,2}^n$ and $E_{2,1}^n$ are double sums only using one sample, we first remove terms for which $i = j$. That leads to

$$
E_{2,k}^n = \frac{1}{n_k^2 a_n} \sum_{i \neq j}^{n_k} K(\frac{X_{ik} - X_{jk}}{a_n}) \rho_{2,k}((X_{ik}, Z_{ik}, \Delta_{ik}), (Z_{jk}, \Delta_{jk})) + o_P(n_k^{-1/2}), \;\; k = 1, 2.
$$

We go on developing $E_{2,k}^n$, $k = 1, 2$, with

$$
\begin{aligned}
E_{2,k}^n \;=\; & (n_k^2 a_n)^{-1} \sum_{i \neq j} \{ A_k^*(V_i, V_j) + E[A_k(V_i, V_j)|V_i] + E[A_k(V_i, V_j)|V_j] - E[A_k(V_i, V_j)] \} \\
& + o_P(n_k^{-1/2}) \\
= \; & T_{1,k}^n + T_{2,k}^n + T_{3,k}^n + T_{4,k}^n + o_P(n_k^{-1/2}),
\end{aligned}
$$

where

$$
A_k(V_i, V_j) = K(\frac{X_{ik} - X_{jk}}{a_n}) \rho_{2,k}((X_{ik}, Z_{ik}, \Delta_{ik}), (Z_{jk}, \Delta_{jk})),
$$

$A_k^*(V_i, V_j) \;=\; A_k(V_i, V_j) \;-\; E[A_k(V_i, V_j)|V_i] \;-\; E[A_k(V_i, V_j)|V_j] \;+\; E[A_k(V_i, V_j)]$ and $V_i \;=$

$(X_{ik}, Z_{ik}, \Delta_{ik})$, for $k = 1, 2$. Consider

$$
\begin{aligned}
&E[A_k(V_i, V_j)|V_i] \\
&= \sum_{\delta=0,1} \int\int \rho_{2,k}((X_{ik}, Z_{ik}, \Delta_{ik}), (z, \delta)) K(\frac{X_{ik} - x}{a_n}) h_{\delta,k}(z|x) f_{X_k}(x)\, dz\, dx \\
&= a_n \sum_{\delta=0,1} \int\int \rho_{2,k}((X_{ik}, Z_{ik}, \Delta_{ik}), (z, \delta)) K(u) (h_{\delta,k}(z|X_{ik}) - a_n u \dot{h}_{\delta,k}(z|X_{ik})) \\
&\qquad \times (f_{X_k}(X_{ik}) - a_n u f'_{X_k}(X_{ik}))\, dz\, du + O(a_n^3) \\
&= a_n f_{X_k}(X_{ik}) \sum_{\delta=0,1} \int \rho_{2,k}((X_{ik}, Z_{ik}, \Delta_{ik}), (z, \delta)) h_{\delta,k}(z|X_{ik})\, dz + O(a_n^3) = O(a_n^3) \qquad \text{(A.1)}
\end{aligned}
$$

$i = 1, \ldots, n_k$ and $k = 1, 2$, since

$$
\sum_{\delta=0,1} \int \eta_k(z, \delta|x) h_{\delta,k}(z|x) dz = \sum_{\delta=0,1} \int \xi_k(z, \delta, y|x) h_{\delta,k}(z|x) dz = 0
$$

for all $x \in R_X$ and for all $y \leq T_x^k$. Note that the term $O(a_n^3)$ in the second equality of (A.1) is obtained from the third terms of the Taylor expansions of order two for $f_{X_k}(X_{ik} - u a_n)$ and $h_{\delta,k}(z|X_{ik} - u a_n)$ which lead after integration to uniformly bounded terms of order $O(a_n^3)$ under (A1)-A(5). Hence, we also have $E[A_k(V_i, V_j)] = O(a_n^3)$. In a similar way, using three Taylor expansions of order 2, we get

$$
\begin{aligned}
E[A_k(V_i, V_j)|V_j] \;=\; & a_n \sum_{\delta=0,1} \int\int K(u) \rho_{2,k}((X_{jk} + a_n u, z, \delta), (Z_{jk}, \Delta_{jk})) \\
& \qquad \times h_{\delta,k}(z|X_{jk} + a_n u) f_{X_k}(X_{jk} + a_n u) dz du \\
=\; & a_n \sum_{\delta=0,1} \int\int K(u) \left[\rho_{2,k}((X_{jk}, z, \delta), (Z_{jk}, \Delta_{jk})) \right. \\
& \qquad \left. + a_n u \dot{\rho}_{2,k}((X_{jk}, z, \delta), (Z_{jk}, \Delta_{jk}))\right] \\
& \qquad \times [h_{\delta,k}(z|X_{jk}) + a_n u \dot{h}_{\delta,k}(z|X_{jk})] \\
& \qquad \times [f_{X_k}(X_{jk}) + a_n u f'_{X_k}(X_{jk})] dz du \\
& + O(a_n^3) \\
=\; & a_n f_{X_k}(X_{jk}) \sum_{\delta=0,1} \int \rho_{2,k}((X_{jk}, z, \delta), (Z_{jk}, \Delta_{jk}))\, dH_{\delta,k}(z|X_{jk}) \\
& + O(a_n^3), \qquad \text{(A.2)}
\end{aligned}
$$

where $\dot{\rho}_{2,k}((x^1, z^1, \delta^1), (z^2, \delta^2))$ denotes the derivative of $\rho_{2,k}((x^1, z^1, \delta^1), (z^2, \delta^2))$ with respect to x^1.

Note that for $T^n_{1,k}$, $E[T^n_{1,k}] = 0$, resulting, by Chebyshev's inequality, in

$$\begin{aligned} P(|T^n_{1,k}| > R(n_k a_n)^{-1}) &\leq R^{-2}(n_k a_n)^2 E[(T^n_{1,k})^2] \\ &= R^{-2} n_k^{-2} \sum_{j \neq i} \sum_{m \neq l} E[A^*_k(V_i, V_j) A^*_k(V_l, V_m)], \end{aligned}$$

for any $R > 0$. Since $E[A^*_k(V_i, V_j)] = 0$, the terms for which $i, j \neq l, m$ are zero. The terms for which either i or j equals l or m and the other differs from l and m, are also zero, because, for example when $i = l$ and $j \neq m$,

$$E[A^*_k(V_i, V_j) E[A^*_k(V_i, V_m)|V_i, V_j]] = 0.$$

Thus, only the $2n_k(n_k - 1)$ terms for which (i, j) equals (l, m) or (m, l) remain. Since $A^*_k(V_i, V_j)$ is bounded by $CK(\frac{X_{ik} - X_{jk}}{a_n}) + O(a_n)$ for some constant $C > 0$, we have (in the case (i, j) equals (l, m)) that

$$E[A^*_k(V_i, V_j)^2] \leq C^2 a_n \int f^2_{X_k}(x)\, dx \int K^2(u)\, du + O(a_n^2) = O(a_n).$$

The case (i, j) equals (m, l) is treated similarly. It now follows that

$$T^n_{1,k} = o_P(n_k^{-1} a_n^{-1}), \tag{A.3}$$

which is $o_P(n_k^{-1/2})$. By (A.1), (A.2), (A.3), we finally obtain

$$\begin{aligned} E^n_{2,k} \;=\; & \frac{1}{n_k} \sum_{i=1}^{n_k} \sum_{\delta=0,1} \int \rho_{2,k}((X_{ik}, z, \delta), (Z_{ik}, \Delta_{ik})) f_{X_k}(X_{ik}) dH_{\delta,k}(z|X_{ik}) \\ & + o_P(n^{-1/2}),\ k = 1, 2. \end{aligned}$$

The terms $E^n_{1,2}$ and $E^n_{1,1}$ are treated in a very similar way such that

$$\begin{aligned} E^n_{1,3-k} \;=\; & \frac{1}{n_k} \sum_{i=1}^{n_k} \sum_{\delta=0,1} \int \rho_{1,3-k}((X_{ik}, z, \delta), (Z_{ik}, \Delta_{ik})) f_{X_{3-k}}(X_{ik}) dH_{\delta,3-k}(z|X_{ik}) \\ & + o_P(n^{-1/2}),\ k = 1, 2, \end{aligned}$$

and both first terms of $E^n_{2,k}$ and $E^n_{1,3-k}$ are sums of i.i.d random variables with zero means. Since $E^n_{1,3-k}$, $E^n_{2,k}$ for $k = 1, 2$, and $E^n_{3,2}$ don't depend on the difference between $m_2(\cdot)$ and $m_1(\cdot)$, they are involved in the asymptotic representation for any hypothesis (H_0, H_{1n} or H_1). $E^n_{3,1}$ is null under H_0, negligible under H_{1n} and modifies the variability of the asymptotic representation under H_1. Applications of the central limit theorem finish the proof.□

Proof of Theorem 2. First, write

$$\begin{aligned}\sum_{j=1,2}\frac{1}{n_j}\sum_{i=1}^{n_j} w_j(X_{ij})(\hat{m}_2(X_{ij})-\hat{m}_1(X_{ij})) &= \sum_{j=1,2}\frac{1}{n_j}\sum_{i=1}^{n_j} w_j(X_{ij})(m_1(X_{ij})-\hat{m}_1(X_{ij}))\\ &\quad-\sum_{j=1,2}\frac{1}{n_j}\sum_{i=1}^{n_j} w_j(X_{ij})(m_2(X_{ij})-\hat{m}_2(X_{ij}))\\ &\quad+\sum_{j=1,2}\frac{1}{n_j}\sum_{i=1}^{n_j} w_j(X_{ij})(m_2(X_{ij})-m_1(X_{ij}))\\ &= E_1^n - E_2^n + E_3^n\end{aligned}$$

Clearly,

$$\begin{aligned}E_3^n = &\sum_{j=1,2}\frac{1}{n_j}\sum_{i=1}^{n_j}\left\{w_j(X_{ij})(m_2(X_{ij})-m_1(X_{ij})) - \int_{R_X} w_j(x)(m_2(x)-m_1(x))f_{X_j}(x)dx\right\}\\ &+\int_{R_X}(m_2(x)-m_1(x))f(x)dx,\end{aligned}$$

for $f(x)=f_{X_1}(x)w_1(x)+f_{X_2}(x)w_2(x)$. We now focus on E_1^n. Using the asymptotic development of $\hat{m}_1(\cdot)-m_1(\cdot)$, it can be written as

$$\begin{aligned}&\frac{-1}{n_1^2 a_n}\sum_{i=1}^{n_1}\sum_{k=1}^{n_1} w_1(X_{i1})K(\frac{X_{i1}-X_{k1}}{a_n})\eta_1(Z_{k1},\Delta_{k1}|X_{i1})\\ &-\frac{1}{n_1 n_2 a_n}\sum_{i=1}^{n_2}\sum_{k=1}^{n_1} w_2(X_{i2})K(\frac{X_{i2}-X_{k1}}{a_n})\eta_1(Z_{k1},\Delta_{k1}|X_{i2})+o_P(n^{-1/2}).\end{aligned}$$

Both terms above are therefore double sums and E_2^n can be treated similarly. Now, following the lines of Theorem 1 (for the treatment of $E_{k,l}^n$, $k,l=1,2$),

$$E_j^n = -\frac{1}{n_j}\sum_{i=1}^{n_j}\eta_j(Z_{ij},\Delta_{ij}|X_{ij})f(X_{ij})+o_P(n^{-1/2}),\ j=1,2,$$

where $w_j(\cdot)$ is two times differentiable. Therefore, the first term of E_j^n, $j=1,2$, is a sum of i.i.d. random variables with zero mean and arguments similar to the end of Theorem 1 finish the proof.

Acknowledgements

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

author is supported by the Spanish Ministerio de Ciencia e Innovación (project MTM2008-03129), Xunta de Galicia (projects PGIDIT07PXIB300191PR and INBIOMED DXPCT-SUG 2009/063) and Universidade de Vigo. The authors would also thank two anonymous reviewers for their helpful comments about this article.

References

Beran, R. (1981). Nonparametric regression with randomly censored survival data. Technical Report, University of California, Berkeley.

Efron, B. (1981). Censored data and the bootstrap. *Journal of the American Statistical Association*, **76**, 312-319.

Fan, J. and Gijbels, I. (1994). Censored regression: local linear approximations and their applications. *Journal of the American Statistical Association*, **89**, 560–570.

González-Manteiga, W. and Cadarso-Suárez, C. (1994). Asymptotic properties of a generalized Kaplan-Meier estimator with some applications. *Nonparametric Statistics*, **4**, 65–78.

Hall, P., Huber, C. and Speckman, P.L. (1997). Covariate-matched one-sided tests for the difference between functional means. *Journal of the American Statistical Association*, **92**, 1074–1083.

Heuchenne, C. and Van Keilegom, I. (2007). Nonlinear regression with censored data. *Technometrics*, **49**, 34–44.

Koul, H.L. and Schick, A. (1997). Testing for the equality of two nonparametric regression curves. *Journal of Statistical Planning and Inference*, **65**, 293–314.

Koul, H.L. and Schick, A. (2003). Testing for superiority among two regression curves. *Journal of Statistical Planning and Inference*, **117**, 15–33.

Neumeyer, N. (2006). *Bootstrap procedures for empirical processes of nonparametric residuals*. Habilitationsschrift, Fakultät für Mathematik, Ruhr-Universität Bochum, Bochum.

Neumeyer, N. (2009). Smooth residual bootstrap for empirical processes of nonparametric regression residuals. *Scandinavian Journal of Statististics*, **36**, 204–208.

Neumeyer, N. and Dette, H. (2003). Nonparametric comparison of regression curves: an empirical process approach. *The Annals of Statistics*, **31**, 880–920.

Neumeyer, N. and Dette, H. (2005). A note on one-sided nonparametric analysis of covariance by ranking residuals. *Mathematical Methods of Statistics*, **14**, 80–104.

Neumeyer, N. and Pardo-Fernández, J.C. (2009). A simple test for comparing regression curves versus one-sided alternatives. *Journal of Statistical Planning and Inference*, **139**, 4006–4016.

Pardo-Fernández, J.C. and Van Keilegom, I. (2006). Comparison of regression curves with censored responses. *Scandinavian Journal of Statistics*, **33**, 409–434.

Pardo-Fernández, J.C., Van Keilegom, I. and González-Manteiga, W. (2007). Testing for the equality of k regression curves. *Statistica Sinica*, **17**, 1115–1137.

Srihera, R. and Stute, W. (2010). Nonparametric comparison of regression functions. *Journal of Multivariate Analysis*, **101**, 2039–2059.

Van Keilegom, I. and Akritas, M.G. (1999). Transfer of tail information in censored regression models. *Ann. Statist.*, **27**, 1745–1784.

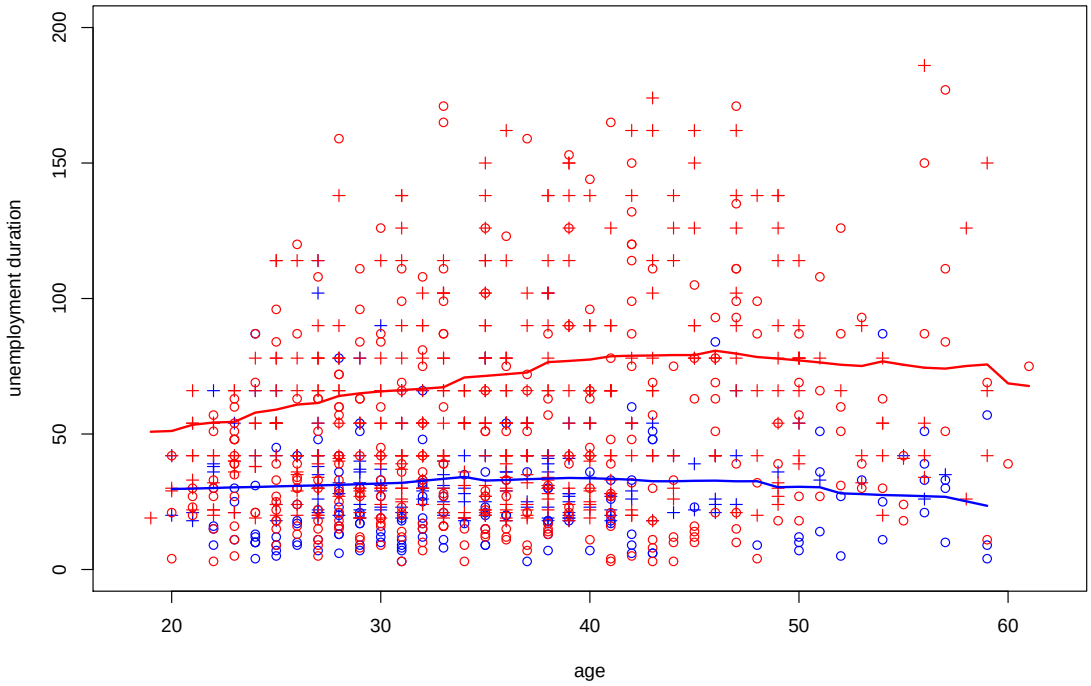

Figure 1: Scatter plot and estimated conditional location curves of the response variable 'unemployment duration' (in days) versus the covariate 'age' (in years) of the group of women receiving any subsidy (in blue) and the group of women not receiving any subsidy (in red). Circles represent uncensored observations, crosses represent censored observations.

Table 1: Empirical proportion of rejections of the test based on the statistic T_{n1} and T_{n2} under models (i)-(vi). The regression models are homoscedastic, as given in (4.1).

			Results for T_{n1}				Results for T_{n2}			
		$a_{jn}:$	$a_{jn}=1.5n_j^{-3/10}$		$a_{jn}=2n_j^{-3/10}$		$a_{jn}=1.5n_j^{-3/10}$		$a_{jn}=2n_j^{-3/10}$	
model	(n_1,n_2)	$\alpha:$	0.100	0.050	0.100	0.050	0.100	0.050	0.100	0.050
(i)	$(50,50)$		0.094	0.040	0.104	0.040	0.096	0.042	0.112	0.040
	$(50,100)$		0.110	0.060	0.114	0.072	0.086	0.046	0.086	0.046
	$(100,100)$		0.087	0.044	0.104	0.049	0.067	0.038	0.079	0.040
(ii)	$(50,50)$		0.882	0.746	0.878	0.746	0.892	0.819	0.886	0.772
	$(50,100)$		0.958	0.893	0.959	0.906	0.962	0.920	0.962	0.911
	$(100,100)$		0.986	0.947	0.982	0.951	0.991	0.968	0.987	0.964
(iii)	$(50,50)$		0.914	0.794	0.904	0.798	0.908	0.799	0.895	0.815
	$(50,100)$		0.972	0.925	0.975	0.927	0.968	0.924	0.966	0.925
	$(100,100)$		0.988	0.958	0.986	0.972	0.990	0.974	0.991	0.965
(iv)	$(50,50)$		0.646	0.481	0.642	0.477	0.709	0.586	0.668	0.533
	$(50,100)$		0.842	0.712	0.790	0.671	0.851	0.755	0.796	0.679
	$(100,100)$		0.881	0.766	0.838	0.742	0.911	0.849	0.883	0.801
(v)	$(50,50)$		0.226	0.115	0.198	0.095	0.243	0.139	0.234	0.110
	$(50,100)$		0.312	0.191	0.267	0.164	0.327	0.202	0.289	0.173
	$(100,100)$		0.383	0.242	0.333	0.165	0.387	0.244	0.371	0.187

Table 2: Empirical proportion of rejections of the test based on the statistic T_{n1} and T_{n2} under models (i)-(vi). The regression models are heteroscedastic, as given in (4.2).

			Results for T_{n1}				Results for T_{n2}			
		$a_{jn}:$	$a_{jn}=1.5n_j^{-3/10}$		$a_{jn}=2n_j^{-3/10}$		$a_{jn}=1.5n_j^{-3/10}$		$a_{jn}=2n_j^{-3/10}$	
model	(n_1,n_2)	$\alpha:$	0.100	0.050	0.100	0.050	0.100	0.050	0.100	0.050
(i)	$(50,50)$		0.093	0.050	0.118	0.058	0.108	0.043	0.110	0.051
	$(50,100)$		0.088	0.056	0.115	0.057	0.070	0.042	0.089	0.054
	$(100,100)$		0.054	0.032	0.087	0.044	0.050	0.028	0.086	0.034
(ii)	$(50,50)$		0.873	0.717	0.882	0.753	0.893	0.799	0.897	0.787
	$(50,100)$		0.958	0.911	0.955	0.911	0.967	0.922	0.973	0.930
	$(100,100)$		0.985	0.961	0.983	0.966	0.993	0.980	0.994	0.977
(iii)	$(50,50)$		0.888	0.775	0.903	0.768	0.905	0.782	0.902	0.775
	$(50,100)$		0.966	0.925	0.975	0.918	0.968	0.926	0.968	0.927
	$(100,100)$		0.986	0.955	0.984	0.962	0.989	0.966	0.995	0.963
(iv)	$(50,50)$		0.646	0.459	0.612	0.439	0.687	0.555	0.621	0.498
	$(50,100)$		0.807	0.688	0.771	0.625	0.791	0.713	0.756	0.626
	$(100,100)$		0.879	0.746	0.812	0.712	0.890	0.805	0.852	0.756
(v)	$(50,50)$		0.239	0.124	0.193	0.107	0.235	0.137	0.225	0.112
	$(50,100)$		0.329	0.195	0.271	0.152	0.334	0.209	0.301	0.171
	$(100,100)$		0.383	0.222	0.329	0.199	0.393	0.230	0.351	0.204