

# ROC curves in nonparametric location-scale regression models

WENCESLAO GONZÁLEZ-MANTEIGA

*Departamento de Estatística e IO,  
Universidade de Santiago de Compostela*

JUAN CARLOS PARDO-FERNÁNDEZ

*Departamento de Estatística e IO, Universidade de Vigo*

INGRID VAN KEILEGOM

*Institute de Statistique, Université catholique de Louvain*

## Abstract

The receiver operating characteristic curve (ROC curve) is a tool of extensive use to analyse the discrimination capability of a diagnostic variable in medical studies. In certain situations, the presence of a covariate related to the diagnostic variable can increase the discriminating power of the ROC curve. In this article we model the effect of the covariate over the diagnostic variable by means of nonparametric location-scale regression models. We propose a new nonparametric estimator of the conditional ROC curve and study its asymptotic properties. We also present some simulations and an illustration to a data set concerning diagnosis of diabetes.

*Key Words:* area under the curve, conditional ROC curve, location-scale regression models, nonparametric regression, relative distribution.

*Running headline:* ROC curves in nonparametric regression

# 1 Introduction

## 1.1 ROC curves

In medical studies, or in general in health studies, the diagnosis of an individual or a patient is very often based on a characteristic of interest, which may lead to some classification errors. These classification errors are calibrated on the basis of two indicators: *sensitivity* (probability of diagnosing a diseased person as diseased) and *specificity* (probability of diagnosing a healthy person as healthy).

When the diagnostic characteristic, or diagnostic variable, is of a continuous type, here denoted by  $Y$ , the classification will necessarily be based on a *cutoff* value,  $c$ : if  $Y \geq c$  then the individual is classified as diseased, and if  $Y < c$  then the individual is classified as healthy. Let  $F_1$  denote the distribution of  $Y$  in the diseased population, and let  $F_0$  denote the distribution of  $Y$  in the healthy population. In that case, the following geometrical locus is of special interest:

$$\{ (1 - F_0(c), 1 - F_1(c)), c \in \mathbb{R} \}, \quad (1)$$

which is obtained by varying the cutoff values in the complement of the specificity versus the sensitivity. The geometrical locus (1) is called the *receiver operating characteristic curve* (ROC curve), and it is a very extensively used tool to analyse the discrimination power of the diagnostic variable. In practice, the ROC curve is usually reparametrized in the interval  $(0, 1)$ , as follows:

$$\{ (p, 1 - F_1(F_0^{-1}(1 - p))), p \in (0, 1) \}.$$

The estimation of the ROC curve has been intensively treated in the literature, specially during the last ten years, both from parametric and non-parametric points of view. The books of Pepe (2004) and Krzanowski Hand (2009) are general and good references on this topic.

Several estimators have been proposed when the ROC curve is identified as

$$\text{ROC}(p) = 1 - F_1(F_0^{-1}(1 - p)), \quad 0 < p < 1.$$

For that, assume that two samples,  $\{Y_{01}, \dots, Y_{0n_0}\}$  and  $\{Y_{11}, \dots, Y_{1n_1}\}$ , are available from the populations  $F_0$  and  $F_1$ , respectively. Those estimates are of the form

$$\widehat{\text{ROC}}(p) = 1 - \hat{F}_1(\hat{F}_0^{-1}(1 - p)),$$

where  $\hat{F}_0$  and  $\hat{F}_1$  are either empirical estimates  $\hat{F}_j(t) = F_{jn_j}(t) = n_j^{-1} \sum_{i=1}^{n_j} I(Y_{ji} \leq t)$ , or smooth estimates  $\hat{F}_j(t) = (F_{jn_j} * K_h)(t)$  (here  $K_h(u) = \int_{-\infty}^u h^{-1}k(h^{-1}u)du$  is the cumulative distribution function of the rescaled version of the kernel  $k$ ,  $h$  is a bandwidth or smoothing parameter, and  $*$  denotes convolution). See, among others, the aforementioned book of Pepe (2004) and the papers by Lloyd (1998), Lloyd Yong (1999), Zou, Hall Shapiro (1997), Zhou Harezlak (2002) and Hall Hyndman (2003). Other smoothing procedures are treated in the papers by Qiu Le (2001) and by Peng Zhou (2004), while Wan Zhang (2007) present a semiparametric approach. Besides, the ROC curve can also be interpreted in terms of the relative distribution or relative density, see e.g. Hancock Morris (1999) and Molanes-López (2007).

Related to the ROC curve, several markers, such as the area under the curve (AUC) or the index of Youden, are considered as summaries of the discrimination capability of the ROC curve. The AUC is the most commonly used one and it is given by

$$\text{AUC} = \int_0^1 \text{ROC}(p) dp.$$

Clearly, under the assumption of independence between populations,  $\text{AUC} = P(Y_1 > Y_0)$ , where  $Y_0$  and  $Y_1$  are random variables with distributions  $F_0$  and  $F_1$ , respectively. The AUC takes values between 0.5 (low discrimination power) and 1 (high discrimination power).

A widely used family of ROC curves is obtained when the distributions  $F_0$  and  $F_1$  only differ from their location parameters,  $\mu_0$  and  $\mu_1$ , and scale parameters,  $\sigma_0$  and  $\sigma_1$ . More specifically, when the distributions  $F_0$  and  $F_1$  are Gaussian, the obtained ROC curve is called a *binormal ROC curve*:

$$\text{ROC}(p) = \Phi(a + b \Phi^{-1}(p)),$$

where  $\Phi$  is the cumulative distribution function of a standard normal,  $\Phi^{-1}$  is the corresponding quantile function,  $a = (\mu_1 - \mu_0)/\sigma_1$  and  $b = \sigma_0/\sigma_1$ . In that case, the area under the curve is simply  $\text{AUC} = \Phi(a/\sqrt{1+b^2})$ ; see e.g. Pepe (2004), page 83.

## 1.2 ROC curves with covariates

In many studies, a covariate (or vector of covariates),  $X$ , is available along with the diagnostic variable,  $Y$ . The information contained in  $X$  may increase the discrimination

capability of the ROC curve. A general framework to incorporate the information in the covariate is given by location-scale regression models:

$$Y_0 = \mu_0(X_0) + \sigma_0(X_0)\varepsilon_0, \quad (2)$$

$$Y_1 = \mu_1(X_1) + \sigma_1(X_1)\varepsilon_1, \quad (3)$$

where, for  $j = 0, 1$ ,  $\mu_j(\cdot) = E(Y_j|X_j = \cdot)$  and  $\sigma_j^2(\cdot) = Var(Y_j|X_j = \cdot)$  are the conditional mean and conditional variance of the response  $Y_j$  given the covariate  $X_j$  in each population, respectively, and the error  $\varepsilon_j$  is independent of  $X_j$ .

The parametric case, with  $\mu_j(x) = \alpha_j + \beta_j x$  ( $j = 0, 1$ ) and constant variances, has been studied and applied in the recent literature. See, for instance, Pepe (1997, 1998, 2004) or Faraggi (2003). In the latter paper by Faraggi, a data set concerning fingerstick glucose measurements as a marker for diabetes is analysed and the age of the patients is considered as the covariate. This data set was previously discussed in Smith Thompson (1996), and we will reconsider it in our illustration in section 5.

More recently, Zheng Heagerty (2004) in a context where the diagnostic marker changes over time, estimated the ROC curve induced from model (2)-(3) on the basis of pilot spline estimators for the mean functions and variance functions.

In other contributions in nonparametric setups, the ROC curve is directly modelled through a generalized linear model of a semiparametric type where the ROC curve is considered as the response variable (see, for instance, Cai Pepe, 2002).

In this paper, we present a new nonparametric estimator of the conditional ROC curve under the general model (2)-(3). The estimating process, which makes use of the estimation of the distribution of the regression errors, is described in section 2. In section 3 we state several theoretical results concerning the asymptotic behaviour of the proposed estimator. Some simulations are presented in section 4, and section 5 contains an illustration to the abovementioned data set. Finally, the appendix contains the proofs of the theoretical results.

## 2 Methodology

Consider that along with the diagnostic variables in the healthy population,  $Y_0$ , and in the diseased population,  $Y_1$ , we have two univariate continuous covariates referring

to the same characteristic,  $X_0$  and  $X_1$ . The relation between the diagnostic variables and the covariates is established in terms of the nonparametric location-scale regression models (2)-(3), where we assume for  $j = 0, 1$  that  $\mu_j(\cdot) = E(Y_j|X_j = \cdot)$  and  $\sigma_j^2(\cdot) = \text{Var}(Y_j|X_j = \cdot)$  are unknown smooth functions, and  $\varepsilon_j$  is independent of  $X_j$ . For  $j = 0, 1$ , let  $G_j(y) = P(\varepsilon_j \leq y)$ ,  $F_j(y|x) = P(Y_j \leq y|X_j = x)$  and  $F_{X_j}(x) = P(X_j \leq x)$ , and denote the support of  $X_j$  by  $R_{X_j}$ . The intersection of  $R_{X_0}$  and  $R_{X_1}$  is denoted by  $R_X$  and is supposed to be non-empty. The probability density functions of the above distributions will be denoted by lower case letters (i.e.,  $g_j(y)$ ,  $f_j(y|x)$  and  $f_{X_j}$ , for  $j = 0, 1$ ).

For a fixed value  $x$  in  $R_X$ , the conditional ROC curve is defined by, for  $0 < p < 1$ ,

$$\begin{aligned} \text{ROC}_x(p) &= 1 - F_1(F_0^{-1}(1-p|x)|x) \\ &= 1 - G_1\left(\sigma_1^{-1}(x)\{G_0^{-1}(1-p)\sigma_0(x) + \mu_0(x) - \mu_1(x)\}\right) \\ &= 1 - G_1\left(G_0^{-1}(1-p)b(x) - a(x)\right), \end{aligned}$$

where

$$a(x) = \frac{\mu_1(x) - \mu_0(x)}{\sigma_1(x)} \quad \text{and} \quad b(x) = \frac{\sigma_0(x)}{\sigma_1(x)},$$

and where for any distribution function  $F$  and any  $0 \leq s \leq 1$ ,  $F^{-1}(s) = \inf\{y : F(y) \geq s\}$ . Suppose we have a sample  $(X_{01}, Y_{01}), \dots, (X_{0n_0}, Y_{0n_0})$  of i.i.d. data generated from model (2) and another sample  $(X_{11}, Y_{11}), \dots, (X_{1n_1}, Y_{1n_1})$  of i.i.d. data generated from model (3), that is independent of the first sample. Let  $N = n_0 + n_1$ . Based on these data, we propose the following estimator of the conditional ROC curve:

$$\widehat{\text{ROC}}_x(p) = 1 - \int \hat{G}_1\left(\hat{G}_0^{-1}(1-p+hu)\hat{b}(x) - \hat{a}(x)\right)k(u)du, \quad (4)$$

where  $k$  is a probability density function (kernel),  $h = h_N$  is a bandwidth sequence, and for  $j = 0, 1$ ,

$$\begin{aligned} \hat{G}_j(y) &= n_j^{-1} \sum_{i=1}^{n_j} I(\hat{\varepsilon}_{ji} \leq y), \\ \hat{\varepsilon}_{ji} &= \frac{Y_{ji} - \hat{\mu}_j(X_{ji})}{\hat{\sigma}_j(X_{ji})} \quad (i = 1, \dots, n_j), \\ \hat{\mu}_j(x) &= \sum_{i=1}^{n_j} W_{ji}(x, g) Y_{ji}, \quad \hat{\sigma}_j^2(x) = \sum_{i=1}^{n_j} W_{ji}(x, g) [Y_{ji} - \hat{\mu}_j(X_{ji})]^2, \end{aligned}$$

and

$$W_{ji}(x, g) = \frac{k_g(x - X_{ji})}{\sum_{l=1}^{n_j} k_g(x - X_{jl})},$$

with  $g = g_N$  a second bandwidth sequence, and  $k_g(\cdot) = k(\cdot/g)/g$ . Finally,  $\hat{a}(x) = [\hat{\mu}_1(x) - \hat{\mu}_0(x)]/\hat{\sigma}_1(x)$  and  $\hat{b}(x) = \hat{\sigma}_0(x)/\hat{\sigma}_1(x)$ . Note that  $\widehat{\text{ROC}}_x(p)$  can also be written as :

$$\widehat{\text{ROC}}_x(p) = 1 - \frac{1}{n_1} \sum_{i=1}^{n_1} K \left( \frac{\hat{G}_0(\{\hat{\varepsilon}_{1i} + \hat{a}(x)\}/\hat{b}(x)) - 1 + p}{h} \right),$$

where  $K$  is the distribution function corresponding to the kernel  $k$ .

The ROC curve is defined in terms of distribution functions of continuous random variables, and hence it is a continuous curve. This motivates the construction of the smooth estimator proposed in (4), which ensures that the estimated ROC curve is also continuous. The bandwidth  $h$  determines the smoothness of the estimated ROC curve.

Also note that the estimator of the conditional ROC curve given in (4) can be considered simply in terms of empirical distributions of the regression residuals, without adding any smoothing to the ROC curve, by taking  $h = 0$ :

$$\widetilde{\text{ROC}}_x(p) = 1 - \hat{G}_1 \left( \hat{G}_0^{-1}(1 - p)\hat{b}(x) - \hat{a}(x) \right). \quad (5)$$

This estimator, which we can call the “empirical” conditional ROC curve estimator, is also a valid estimator of the conditional ROC curve, but it has the drawback of not being continuous.

On the other hand, the bandwidth  $g$  is used to locally estimate the regression and variance functions. In principle, one could use different bandwidths for each of the curves  $\mu_0(x)$ ,  $\mu_1(x)$ ,  $\sigma_0(x)$  and  $\sigma_1(x)$ , but for simplicity of presentation we will restrict here to one bandwidth.

Other estimators of  $\text{ROC}_x(p)$  can be considered, based on smoothing of each of the empirical distributions  $\hat{G}_0(\cdot)$  and  $\hat{G}_1(\cdot)$ . See e.g. Hall Hyndman (2003) and Qiu Le (2001) for the case without covariates. We follow here the approach used, among others, by Peng Zhou (2004) and López-de Ullibarri et al. (2008) and apply smoothing on the ROC curve itself.

### 3 Main result

The following result is an i.i.d. representation for the ROC-process  $\widehat{\text{ROC}}_x(p) - \text{ROC}_x(p)$ . Note that the main term of this representation does not depend on the bandwidth  $h$ , as its contribution is asymptotically negligible. The assumptions under which the results below are valid, are given in the appendix.

**Theorem 1** Assume (A1)-(A3). Then, for  $0 < p < 1$  and for a fixed  $x$  in  $R_X$ ,

$$\begin{aligned} & \widehat{\text{ROC}}_x(p) - \text{ROC}_x(p) \\ &= g_1(G_0^{-1}(1-p)b(x) - a(x)) \left\{ \hat{A}_x + G_0^{-1}(1-p)\hat{B}_x \right\} + g^2\beta_x(p) + \hat{R}_x(p), \end{aligned}$$

where

$$\begin{aligned} \hat{A}_x &= \sigma_1^{-1}(x) \sum_{j=0}^1 (-1)^{j+1} f_{X_j}^{-1}(x) n_j^{-1} \sum_{i=1}^{n_j} k_g(x - X_{ji})(Y_{ji} - \mu_j(X_{ji})), \\ \hat{B}_x &= \frac{1}{2} \sigma_1^{-2}(x) \sum_{j=0}^1 (-1)^{j+1} \left( \frac{\sigma_0(x)}{\sigma_1(x)} \right)^{2j-1} f_{X_j}^{-1}(x) n_j^{-1} \sum_{i=1}^{n_j} k_g(x - X_{ji}) \sigma_j^2(X_{ji}) (\varepsilon_{ji}^2 - 1), \\ \beta_x(p) &= -\frac{1}{2} \mu_2^k \int \frac{\partial^2}{\partial t^2} E[\varphi(t, Y_1, c_x(1-p)) | X_1 = v] |_{t=v} dF_{X_1}(v) \\ &\quad + \frac{1}{2} \mu_2^k g_1(G_0^{-1}(1-p)b(x) - a(x)) \left\{ \sigma_1^{-1}(x) \sum_{j=0}^1 (-1)^{j+1} \left[ \mu_j''(x) + 2\mu_j'(x) \frac{f'_{X_j}(x)}{f_{X_j}(x)} \right] \right. \\ &\quad \left. + G_0^{-1}(1-p) \sigma_1^{-2}(x) \frac{1}{2} \sum_{j=0}^1 (-1)^{j+1} \left( \frac{\sigma_0(x)}{\sigma_1(x)} \right)^{2j-1} \left[ (\sigma_j^2(x))'' + 2(\sigma_j^2(x))' \frac{f'_{X_j}(x)}{f_{X_j}(x)} \right] \right\}, \\ \varphi(x, y, z) &= g_1(z) \sigma_1^{-1}(x) \left[ y - \mu_1(x) + \frac{z}{2\sigma_1(x)} \left\{ (y - \mu_1(x))^2 - \sigma_1^2(x) \right\} \right], \end{aligned}$$

and where  $\mu_2^k = \int u^2 k(u) du$  and  $\sup_{\delta < p < 1-\delta} |\hat{R}_x(p)| = o_P((Ng)^{-1/2})$ , for any small  $\delta > 0$ .

As a consequence, we get the weak convergence of the ROC-process. The proof can be obtained by applying the central limit theorem for triangular arrays to the random variables  $(Ng)^{1/2} \hat{A}_x$  and  $(Ng)^{1/2} \hat{B}_x$ . Both the case of undersmoothing ( $C = 0$ ) and the optimal bandwidth  $g = C^{1/5} N^{-1/5}$  with  $0 < C < \infty$  are considered.

**Corollary 2** Assume (A1)-(A3). Then, for a fixed  $x$  in  $R_X$  and for a small  $\delta > 0$ , the process  $(Ng)^{1/2}(\widehat{\text{ROC}}_x(p) - \text{ROC}_x(p))$  ( $\delta < p < 1 - \delta$ ) converges weakly to a Gaussian process

$$W_x(p) = g_1(G_0^{-1}(1-p)b(x) - a(x)) \left\{ W_{1x} + G_0^{-1}(1-p)W_{2x} \right\} + C^{1/2}\beta_x(p),$$

where  $C$  is defined in assumption (A1), and where  $W_{1x}$  and  $W_{2x}$  are normal random

variables with zero mean, and

$$\begin{aligned} \text{Var}(W_{1x}) &= \sigma_1^{-2}(x) \|k\|_2^2 \sum_{j=0}^1 f_{X_j}^{-1}(x) \lambda_j^{-1} \sigma_j^2(x) \\ \text{Var}(W_{2x}) &= \frac{1}{4} \frac{\sigma_0^2(x)}{\sigma_1^2(x)} \|k\|_2^2 \sum_{j=0}^1 f_{X_j}^{-1}(x) \lambda_j^{-1} E(\varepsilon_j^4 - 1) \\ \text{Cov}(W_{1x}, W_{2x}) &= \frac{1}{2} \frac{\sigma_0(x)}{\sigma_1^2(x)} \|k\|_2^2 \sum_{j=0}^1 f_{X_j}^{-1}(x) \lambda_j^{-1} \sigma_j(x) E(\varepsilon_j^3), \end{aligned}$$

with  $\lambda_j = \lim_{N \rightarrow \infty} n_j/N$  ( $j = 0, 1$ ), and where  $\|k\|_2^2 = \int k^2(u) du$ .

This result can now be used to obtain the limiting distribution of any continuous functional of the ROC-process. A well known particular case is the conditional version of the so-called *area under the curve* (AUC), which, for a fixed  $x$  in  $R_X$ , we define by

$$\text{AUC}_x = \int_{\delta}^{1-\delta} \text{ROC}_x(p) dp. \quad (6)$$

For technical reasons, we restrict the integration to the interval to  $[\delta, 1 - \delta]$ , which can however be made arbitrarily close to  $[0, 1]$ . The estimator is

$$\widehat{\text{AUC}}_x = \int_{\delta}^{1-\delta} \widehat{\text{ROC}}_x(p) dp.$$

The proof of the following result is an immediate consequence of the continuous mapping theorem.

**Corollary 3** *Assume (A1)-(A3). Then, for a fixed  $x$  in  $R_X$ ,*

$$(Ng)^{1/2} (\widehat{\text{AUC}}_x - \text{AUC}_x) \xrightarrow{d} N(d_x, s_x^2),$$

where  $d_x = C^{1/2} \int_{\delta}^{1-\delta} \beta_x(p) dp$  and

$$\begin{aligned} s_x^2 &= \text{Var} \left( \int_{\delta}^{1-\delta} W_x(p) dp \right) \\ &= \gamma_{1x}^2 \text{Var}(W_{1x}) + \gamma_{2x}^2 \text{Var}(W_{2x}) + 2\gamma_{1x}\gamma_{2x} \text{Cov}(W_{1x}, W_{2x}), \\ \gamma_{1x} &= \int_{\delta}^{1-\delta} g_1(G_0^{-1}(1-p)b(x) - a(x)) dp, \\ \gamma_{2x} &= \int_{\delta}^{1-\delta} g_1(G_0^{-1}(1-p)b(x) - a(x)) G_0^{-1}(1-p) dp. \end{aligned}$$



## 4 Simulations

In this section we present a small simulation study. We are mainly interested in the global performance of the proposed estimator of the conditional ROC curve and in the effect of the smoothing parameter  $h$ . We have simulated data from two scenarios:

- **Scenario 1:**

Regression functions:  $\mu_0(x) = 0$ ;  $\mu_1(x) = x$ .

Conditional variance functions:  $\sigma_0^2(x) = \sigma_1^2(x) = 0.5^2$ .

- **Scenario 2:**

Regression functions:  $\mu_0(x) = 0.5 \sin(2\pi x)$ ;  $\mu_1(x) = \sin(\pi x)$ .

Conditional variance functions:  $\sigma_0^2(x) = \sigma_1^2(x) = (0.25 + 0.5x)^2$ .

In both scenarios, the covariates  $X_0$  and  $X_1$  are uniformly distributed on  $[0,1]$ , and the regression errors  $\varepsilon_0$  and  $\varepsilon_1$  have standard normal distribution. The true ROC curves, presented here as a surface, and the true conditional AUC, presented as a function of the values of the covariate, are depicted in Figure 1 (scenario 1) and Figure 2 (scenario 2).

[ **Figure 1 to be placed here** ]

[ **Figure 2 to be placed here** ]

The estimator of the conditional ROC curves was calculated on a grid of points of the form  $\{(x_l, p_r) \in (0, 1) \times (0, 1), l = 1, \dots, n_x, r = 1, \dots, n_p\}$ . More precisely, in all cases we take

$$x_l = 0.05 + (l - 1) \frac{0.90}{n_x - 1}, \text{ for } l = 1, \dots, n_x,$$
$$p_r = 0.05 + (r - 1) \frac{0.90}{n_p - 1}, \text{ for } r = 1, \dots, n_p,$$

with  $n_x = 25$  and  $n_p = 25$ . The estimators of the regression curves,  $\mu_0(\cdot)$  and  $\mu_1(\cdot)$ , and variance curves,  $\sigma_0^2(\cdot)$  and  $\sigma_1^2(\cdot)$ , which are needed in the construction of the estimator of the conditional ROC curve, are based on the kernel of Epanechnikov  $k(u) = 0.75(1 - u^2)I(|u| < 1)$  and on cross-validation bandwidths: for  $j = 0, 1$ , a regular cross-validation procedure is used to estimate  $\mu_j$ , and then the same bandwidth is used to estimate  $\sigma_j^2$ .

The discrepancy between the estimator and the true ROC surface is measured in terms of the empirical version of the global mean squared error (MSE):

$$\text{MSE} = \frac{1}{n_x} \sum_{l=1}^{n_x} \frac{1}{n_p} \sum_{r=1}^{n_p} (\widehat{\text{ROC}}_{x_l}(p_r) - \text{ROC}_{x_l}(p_r))^2.$$

Table 1 displays the averages and standard deviations of the MSEs obtained in 1000 data sets simulated from scenario 1. The estimators of the ROC curves were calculated with different values of the smoothing parameter  $h$ , ranging from 0 to 0.25. The case  $h = 0$  corresponds to the empirical estimator given in (5). As expected, the MSE decreases as the sample sizes increase. The effect of the parameter  $h$  is not very important, although introducing a small amount of smoothing in the estimator produces a better behaviour in terms of MSE with respect to the empirical estimator. The required amount of smoothing to improve the MSE decreases as the sample sizes get larger. Figure 3 shows the boxplots of the 1000 estimated MSEs for several sample sizes and several values of the smoothing parameter.

Finally, we have also considered the estimation of the conditional AUC, as defined in (6), where we take  $\delta = 0.05$ . Figure 4 shows the average of the estimated AUC for several sample sizes, with  $h = 0.10$ . As a reference, we have also included in the graph two bands obtained by means of the logit transformation. The AUC is bounded in the interval  $[0.5, 1]$ , hence the normal asymptotic distribution for  $\widehat{\text{AUC}}_x$  given in corollary 3 might not work very well. Corollary 3 (for the sake of simplicity, we only consider here the case  $C = 0$ ) and the delta method (see, for instance, page 118 in Serfling, 1980) ensure that

$$(Ng)^{1/2}(\text{logit}(\widehat{\text{AUC}}_x) - \text{logit}(\text{AUC}_x)) \xrightarrow{d} N(0, s_x^2/[\text{AUC}_x(1 - \text{AUC}_x)]^2),$$

where  $\text{logit}(p) = \log(p/(1 - p))$ . Note that  $\text{logit}(\widehat{\text{AUC}}_x)$  is not bounded. In Figure 4 we represent the AUC in its original scale, and we have added two bands obtained from the expression

$$\text{logit}^{-1}(\text{logit}(\widehat{\text{AUC}}_x) \pm 2 \text{SE}_x / [\text{AUC}_x(1 - \text{AUC}_x)])$$

where  $\text{logit}^{-1}(u) = \exp(u)/(1 + \exp(u))$  and  $\text{SE}_x$  is the standard deviation of the estimator of the  $\text{AUC}_x$  in the 1000 simulated data sets. The general performance of the estimator of the conditional AUC is good.

[ Table 1 to be placed here ]

[ **Figure 3 to be placed here** ]

[ **Figure 4 to be placed here** ]

Table 2, Figure 5 and Figure 6 show the corresponding results when the data sets are simulated from scenario 2. Similar conclusions can be stated in this case. The lowest values of the MSE are achieved with values of the smoothing parameter  $h$  smaller than the corresponding ones in scenario 1.

[ **Table 2 to be placed here** ]

[ **Figure 5 to be placed here** ]

[ **Figure 6 to be placed here** ]

## 5 Data analysis

As an illustration of the proposed methodology, we present an application to a data set concerning diagnosis of diabetes. This data set has also been analysed in Faraggi (2003) and Smith Thompson (1996).

The data come from a population-based pilot survey of diabetes mellitus in Cairo (Egypt), and consist of post-prandial blood glucose measurements of 286 subjects obtained from a fingerstick. According to the gold standard criteria of the World Health Organization for diagnosing diabetes, 88 subjects were classified as diseased and 198 subjects were classified as healthy. The age of the subject was considered as a relevant covariate in this example, because due to medical reasons (see Smith Thompson (1996) for the details) glucose levels are expected to be higher for older persons who do not suffer from diabetes.

Figure 7 shows the scatter plot of the data for both the healthy and diseased population. The glucose concentration is considered as the diagnostic variable, and the age of the subject as a covariate. We have estimated the conditional ROC curves with the methodology proposed in section 2 in the values of the covariate  $x = 20, 21, \dots, 90$ . The analysis has been performed with several values for the smoothing parameter  $h$ , and very similar results were obtained. Figure 8-(a) shows the complete ROC surface estimated with  $h = 0.10$ . We will keep this value of the smoothing parameter in the rest of the figures. To check visually the effect of the age on the ROC curves, the conditional ROC

curves for ages 30, 50 and 70 are depicted in Figure 8-(b). Clearly, the aging process reduces the capability of the ROC curve to discriminate between diseased and healthy subjects.

[ **Figure 7 to be placed here** ]

[ **Figure 8 to be placed here** ]

The effect of the age on the discrimination power of the ROC curve can be summarized by means of the AUC. Figure 9 shows the AUC as a function of the values of the covariate. As in the simulation study, we use definition (6) with  $\delta = 0.05$ . We have also included in the graph confidence intervals for the AUC obtained by bootstrap. Asymptotic confidence intervals for  $AUC_x$  could be obtained from corollary 3, but the asymptotic variance of the estimator depends on certain unknown quantities that are difficult to estimate. Alternatively, we use a bootstrap of residuals to resample the regression models, and then the percentile method to obtain pointwise bootstrap confidence intervals for the AUC.

More precisely, the bootstrap confidence interval for  $AUC_x$  is obtained with the following algorithm. For fixed  $x$ , and for  $b = 1, \dots, B$  ( $B$  being a large integer),

1. For  $j = 0, 1$ , let  $\{\varepsilon_{ji,b}^*, i = 1, \dots, n_j\}$  be an i.i.d. sample from  $\hat{G}_j$ .
2. Reconstruct bootstrap samples  $\{(X_{ji}, Y_{ji,b}^*), i = 1, \dots, n_j\}$ , for  $j = 0, 1$ , where  $Y_{ji,b}^* = \hat{\mu}_j(X_{ji}) + \hat{\sigma}_j(X_{ji})\varepsilon_{ji,b}^*$ .
3. Repeat the estimation process with the  $b$ th bootstrap sample: estimate the auxiliary functions  $\hat{\mu}_{0,b}^*$ ,  $\hat{\mu}_{1,b}^*$ ,  $\hat{\sigma}_{0,b}^*$ ,  $\hat{\sigma}_{1,b}^*$ ,  $\hat{G}_{0,b}^*$ ,  $\hat{G}_{1,b}^*$  and  $\widehat{ROC}_{x,b}^*$ , and then obtain  $\widehat{AUC}_{x,b}^*$ .

Let  $\widehat{AUC}_{x,(b)}^*$  be the order statistics of the values  $\widehat{AUC}_{x,1}^*, \dots, \widehat{AUC}_{x,B}^*$  obtained in step 3. According to the percentile method,  $(\widehat{AUC}_{x,(\lfloor B\alpha/2 \rfloor)}^*, \widehat{AUC}_{x,(\lfloor B(1-\alpha/2) \rfloor)}^*)$  is a bootstrap confidence interval for  $AUC_x$  of confidence level  $1 - \alpha$  ( $\lfloor \cdot \rfloor$  denotes the integer part).

We have studied the coverage properties of the confidence intervals obtained by the proposed bootstrap procedure in a small simulation study. We use the scenarios given in section 4, from which 1000 data sets were simulated with different sample sizes. Confidence intervals for  $AUC_x$  were obtained for three values of the covariate ( $x = 0.25, 0.50, 0.75$ ). In the simulations we use  $B = 200$ . Table 3 shows the empirical coverages for nominal confidence levels  $1 - \alpha = 0.90$  and  $0.95$ . We display results for  $h = 0$  and  $h = 0.10$ . In

the case of scenario 1, the approximation of the level is very good in almost all cases. In the case of scenario 2, which is more complicated due to the heteroscedasticity of the regression models and the shape of the ROC surface, the empirical coverage is good for  $x = 0.50$ , but somehow poor for  $x = 0.25$  and  $x = 0.75$ .

[ **Table 3 to be placed here** ]

Coming back to the application to real data, in Figure 9 we have depicted the bootstrap confidence intervals of levels 90% and 95% obtained with  $B = 1000$  bootstrap replications for the AUC with respect to the values of the covariate. As seen before, the age of the subject clearly has an important impact on the discrimination power of the glucose measurements as an indicator of diabetes. Similar conclusions can be found in Faraggi (2003), although this author works under a much more restrictive model (linear regression models with homoscedastic normal errors). The advantage of our method is the flexibility incorporated by the nonparametric and heteroscedastic regression models.

[ **Figure 9 to be placed here** ]

## 6 Discussion and concluding remarks

In this article, we have presented and studied a new estimator for the ROC curve under the presence of covariates. This problem has not been treated in the literature with much detail. Our method relies on the assumption of the existence of location-scale regression models explaining the relation between the variable of interest and the covariate in the healthy population and in the diseased population. Location-scale regression models of the type (2)-(3) are flexible enough to model many practical situations.

As we have explained in section 1.2, the conditional ROC curve can be expressed in terms of the conditional distribution function and conditional quantile function of the variable of interest,  $Y_0$  or  $Y_1$ , given the covariate,  $X_0$  or  $X_1$ . The main advantage of introducing location-scale regression models is that the conditional ROC curve, in this case, can be rewritten in terms of the distribution function and quantile function of the regression errors. Note that these functions related to the regression errors are not conditional. This means that instead of estimating the conditional distribution of  $Y$

given each value of  $X$ , we only need to estimate the error distribution in each population. Compare with López-de Ullibarri et al. (2008), who proposed a direct estimator of the conditional ROC curve.

We have presented a detailed study of the asymptotic behaviour of the complete ROC-process and the conditional AUC. This theoretical study is also a novelty with respect to other articles on this topic. Simulations and an application to a data set illustrate our methodology. The use of appropriate bootstrap techniques in this context is still an open problem, since the parameters of interest and their estimators are not standard.

## Acknowledgements

The research of W. González-Manteiga and I. Van Keilegom is supported by the Spanish Ministerio de Ciencia e Innovación (project MTM2008-03010). The research of J.C. Pardo-Fernández is supported by the Spanish Ministerio de Ciencia e Innovación (project MTM2008-03129), Xunta de Galicia (project PGIDIT07PXIB300191PR) and Universidade de Vigo. The research of I. Van Keilegom is also supported by IAP research network P6/03 of the Belgian Government (Belgian Science Policy), and by the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC Grant agreement No. 203650. The authors would like to thank two anonymous reviewers for their constructive remarks, and Professor D. Faraggi for providing the data set analysed in this article.

## References

- Akritis, M. Van Keilegom, I. (2001). Nonparametric estimation of the residual distribution. *Scand. J. Statist.*, 28, 549-567.
- Cai, T. Pepe, M. S. (2002). Semiparametric receiver operating characteristic analysis to evaluate biomarkers for disease. *J. Amer. Statist. Assoc.*, 97, 1099-1107.
- Faraggi, D. (2003). Adjusting receiver operating characteristic curves and related indices for covariates. *The Statistician*, 52, 179-192.
- Hall, P. G. Hyndman, R. J. (2003). Improved methods for bandwidth selection when estimating ROC curves. *Statist. Probab. Lett.*, 64, 181-189.

- Handcock, M. S. Morris, M. (1999). *Relative distribution methods in the social sciences*. Springer, New York.
- Krzanowski, W. J. Hand, D. J. (2009). *ROC curves for continuous data*. Chapman and Hall/CRC, Boca Raton.
- Lloyd, C. J. (1998). Using smoothed receiver operating characteristic curves to summarize and compare diagnostic systems. *J. Amer. Statist. Assoc.*, 93, 1356-1364.
- Lloyd, C. J. Yong, Z. (1999). Kernel estimators of the ROC curve are better than empirical. *Statist. Probab. Lett.*, 44, 221-228.
- López-de Ullibarri, I., Cao, R., Cadarso-Suárez, C. Lado, M. J. (2008). Nonparametric estimation of conditional ROC curves: Application to discrimination tasks in computerized detection of early breast cancer. *Comput. Statist. Data Anal.*, 52, 2623-2631.
- Molanes-López, E. M. (2007). *Nonparametric statistical inference for relative curves in two-sample problems*. PhD thesis. University of A Coruña, A Coruña.
- Peng, L. Zhou, X.-H. (2004). Local linear smoothing of receiver operating characteristic (ROC) curves. *J. Statist. Plann. Inference*, 118, 129-143.
- Pepe, M. S. (1997). A regression modelling framework for receiver operating characteristic curves in medical diagnostic testing. *Biometrika*, 84, 595-608.
- Pepe, M. S. (1998). Three approaches to regression analysis of receiver operating characteristic curves for continuous test results. *Biometrics*, 54, 124-135.
- Pepe, M. S. (2004). *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press, Oxford.
- Qiu, P. Le, C. (2001). ROC curve estimation based on local smoothing. *J. Stat. Comput. Simul.*, 70, 55-69.
- Serfling, R. J. (1980). *Approximation theorems of mathematical statistics*. Wiley, New York.
- Smith, P. J. Thompson, T. J. (1996). Correcting for confounding in analyzing receiver operating characteristic curves. *Biom. J.*, 38, 857-863.
- Wan, S. Zhang, B. (2007). Smooth semiparametric receiver operating characteristic curves for continuous diagnostic tests. *Stat. Med.*, 26, 2565-2586.

Zheng, Y. Heagerty, P. J. (2004). Semiparametric estimation of time-dependent ROC curves for longitudinal marker data. *Biostatistics*, 5, 615-632.

Zhou, X.-H. Harezlak, J. (2002). Comparison of bandwidth selection methods for kernel smoothing of ROC curves. *Stat. Med.*, 21, 2045-2055.

Zou, K. H, Hall, W. J and Shapiro, D. E. (1997). Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. *Stat. Med.*, 16, 2143-2156.

Juan Carlos Pardo-Fernández, Departamento de Estatística e Investigación Operativa, Universidade de Vigo. E.U. Enxeñaría Técnica Industrial, Rúa Torrecedeira 86, Vigo (36208), Spain.

E-mail: juancp@uvigo.es

## Appendix: Proofs

### Assumptions

(A1) (i)  $n_j/N \rightarrow \lambda_j$  for some  $0 < \lambda_j < 1$  ( $j = 0, 1$ ). Moreover,  $Ng^5 \rightarrow C$  for some  $0 \leq C < \infty$ ,  $Ng^{3+\alpha}(\log g^{-1})^{-1} \rightarrow \infty$  for some  $\alpha > 0$  and  $Nh^4g \rightarrow 0$ .

(ii)  $R_{X_j}$  is a bounded interval in  $\mathbb{R}$  ( $j = 0, 1$ ).

(iii)  $k$  has compact support,  $\int uk(u)du = 0$  and  $k$  is twice continuously differentiable.

(A2) (i)  $F_{X_j}$  is three times continuously differentiable and  $\inf_{x \in R_{X_j}} f_{X_j}(x) > 0$  ( $j = 0, 1$ ).

(ii)  $\mu_j$  and  $\sigma_j$  are twice continuously differentiable and  $\inf_{x \in R_{X_j}} \sigma_j(x) > 0$  ( $j = 0, 1$ ).

(A3)  $G_j$  is three times continuously differentiable and  $\sup_y |y^2 G_j^{(k)}(y)| < \infty$  for  $k = 1, 2, 3$  and  $j = 0, 1$ . Moreover, for any  $\delta > 0$ ,  $\inf_{\delta < p < 1-\delta} g_0(G_0^{-1}(p)) > 0$ .



**Proof of theorem 1.** For any  $0 < s < 1$ , let  $c_x(s) = G_0^{-1}(s)b(x) - a(x)$  and  $\hat{c}_x(s) = \hat{G}_0^{-1}(s)\hat{b}(x) - \hat{a}(x)$ . Write

$$\begin{aligned}
& \widehat{\text{ROC}}_x(p) - \text{ROC}_x(p) \\
&= - \int \left\{ \hat{G}_1(\hat{c}_x(1-p+hu)) - E[\hat{G}_1(s)]|_{s=\hat{c}_x(1-p+hu)} \right\} k(u) du \\
&\quad - \int \left\{ E[\hat{G}_1(s)]|_{s=\hat{c}_x(1-p+hu)} - G_1(\hat{c}_x(1-p+hu)) \right\} k(u) du \\
&\quad - \int \left\{ G_1(\hat{c}_x(1-p+hu)) - G_1(c_x(1-p+hu)) \right\} k(u) du \\
&\quad - \int \left\{ G_1(c_x(1-p+hu)) - G_1(c_x(1-p)) \right\} k(u) du \\
&= T_{1x}(p) + T_{2x}(p) + T_{3x}(p) + T_{4x}(p).
\end{aligned}$$

We start with  $T_{1x}(p)$ . Using corollary 2 in Akritas Van Keilegom (2001) it follows that  $\sup_y |\hat{G}_1(y) - E[\hat{G}_1(y)]| = O_P(N^{-1/2})$ , and hence,  $\sup_{\delta < p < 1-\delta} |T_{1x}(p)| = o_P((Ng)^{-1/2})$ . On the other hand,

$$\begin{aligned}
T_{2x}(p) &= -\frac{1}{2}g^2\mu_2^k \int \int \frac{\partial^2}{\partial t^2} E[\varphi(t, Y_1, s) | X_1 = v] |_{t=v, s=\hat{c}_x(1-p+hu)} dF_{X_1}(v) k(u) du + o_P(g^2) \\
&= -\frac{1}{2}g^2\mu_2^k \int \frac{\partial^2}{\partial t^2} E[\varphi(t, Y_1, c_x(1-p)) | X_1 = v] |_{t=v} dF_{X_1}(v) + o_P(g^2).
\end{aligned}$$

Next, by condition (A3) we have that  $\sup_{\delta < p < 1-\delta} |T_{4x}(p)| = O(h^2) = o((Ng)^{-1/2})$  if  $Nh^4g \rightarrow 0$ . It remains to consider  $T_{3x}(p)$  :

$$\begin{aligned}
T_{3x}(p) &= - \int g_1(G_0^{-1}(1-p+hu)b(x) - a(x)) \left\{ G_0^{-1}(1-p+hu)[\hat{b}(x) - b(x)] \right. \\
&\quad \left. - [\hat{a}(x) - a(x)] \right\} k(u) du + O_P((Ng)^{-1} \log N) + O_P(n_0^{-1/2}(\log n_0)^{1/2}) \\
&= -g_1(G_0^{-1}(1-p)b(x) - a(x)) \left\{ G_0^{-1}(1-p)[\hat{b}(x) - b(x)] - [\hat{a}(x) - a(x)] \right\} \\
&\quad + O_P(N^{-1/2}(\log N)^{1/2}) + O(h^2), \tag{1}
\end{aligned}$$

which follows from lemma 1 below and since  $\hat{\mu}_j(x) - \mu_j(x) = O_P((Ng)^{-1/2})$  and  $\hat{\sigma}_j(x) - \sigma_j(x) = O_P((Ng)^{-1/2})$  ( $j = 0, 1$ ). Next, note that

$$\begin{aligned}
& G_0^{-1}(1-p)[\hat{b}(x) - b(x)] - [\hat{a}(x) - a(x)] \\
&= G_0^{-1}(1-p)\sigma_1^{-2}(x) \left[ (\hat{\sigma}_0(x) - \sigma_0(x))\sigma_1(x) - (\hat{\sigma}_1(x) - \sigma_1(x))\sigma_0(x) \right] \\
&\quad - \sigma_1^{-1}(x) \left[ \hat{\mu}_1(x) - \mu_1(x) - \hat{\mu}_0(x) + \mu_0(x) \right] + O_P((Ng)^{-1} \log N), \tag{2}
\end{aligned}$$

and that for  $j = 0, 1$ ,

$$\begin{aligned}\hat{\mu}_j(x) - \mu_j(x) &= f_{X_j}^{-1}(x)n_j^{-1} \sum_{i=1}^{n_j} k_g(x - X_{ji})(Y_{ji} - \mu_j(X_{ji})) \\ &\quad + \frac{g^2}{2} \left[ \mu_j''(x) + 2\mu_j'(x) \frac{f'_{X_j}(x)}{f_{X_j}(x)} \right] \mu_2^k + o_P((Ng)^{-1/2}),\end{aligned}\tag{3}$$

$$\begin{aligned}\hat{\sigma}_j(x) - \sigma_j(x) &= \frac{1}{2} \sigma_j^{-1}(x) f_{X_j}^{-1}(x) n_j^{-1} \sum_{i=1}^{n_j} k_g(x - X_{ji}) [(Y_{ji} - \mu_j(X_{ji}))^2 - \sigma_j^2(X_{ji})] \\ &\quad + \frac{g^2}{4\sigma_j(x)} \left[ (\sigma_j^2(x))'' + 2(\sigma_j^2(x))' \frac{f'_{X_j}(x)}{f_{X_j}(x)} \right] \mu_2^k + o_P((Ng)^{-1/2}).\end{aligned}\tag{4}$$

The result now follows, by combining (1), (2), (3) and (4).

**Lemma 1** *Assume (A1)-(A3). Then, for any small  $\delta > 0$ ,*

$$\sup_{\delta < s < 1-\delta} |\hat{G}_0^{-1}(s) - G_0^{-1}(s)| = O_P(n_0^{-1/2}).$$

**Proof.** Let  $I_\delta = [\delta, 1 - \delta]$ , let  $\alpha_n = K_\varepsilon n_0^{-1/2}$  for some  $K_\varepsilon > 0$  and some  $\varepsilon > 0$ . Then,

$$\begin{aligned}&P\left(\sup_{s \in I_\delta} |\hat{G}_0^{-1}(s) - G_0^{-1}(s)| > \alpha_n\right) \\ &\leq P\left(\hat{G}_0^{-1}(s) > G_0^{-1}(s) + \alpha_n \text{ for some } s \in I_\delta\right) \\ &\quad + P\left(\hat{G}_0^{-1}(s) < G_0^{-1}(s) - \alpha_n \text{ for some } s \in I_\delta\right) \\ &= T_1 + T_2.\end{aligned}$$

In what follows, we consider the term  $T_1$ . The term  $T_2$  can be treated in a very similar way.

$$\begin{aligned}T_1 &\leq P\left(\hat{G}_0(G_0^{-1}(s) + \alpha_n) < s \text{ for some } s \in I_\delta\right) \\ &\leq P\left(\sup_y |\hat{G}_0(y) - G_0(y)| > G_0(G_0^{-1}(s) + \alpha_n) - s \text{ for some } s \in I_\delta\right) \\ &= P\left(\sup_y |\hat{G}_0(y) - G_0(y)| > \inf_{s \in I_\delta} \{G_0(G_0^{-1}(s) + \alpha_n) - s\}\right) \\ &\leq P\left(\sup_y |\hat{G}_0(y) - G_0(y)| > K_1 \alpha_n\right),\end{aligned}$$

since  $\inf_{s \in I_\delta} \{G_0(G_0^{-1}(s) + \alpha_n) - s\} > \inf_{\delta/2 < s < 1-\delta/2} g_0(G_0^{-1}(s)) \alpha_n > K_1 \alpha_n$  for some  $K_1 > 0$ .

The latter probability is bounded by  $\varepsilon$  for  $K_\varepsilon$  and  $n_0$  large enough, since  $\sup_y |\hat{G}_0(y) - G_0(y)| = O_P(n_0^{-1/2})$  (see corollary 2 in Akritas Van Keilegom, 2001).

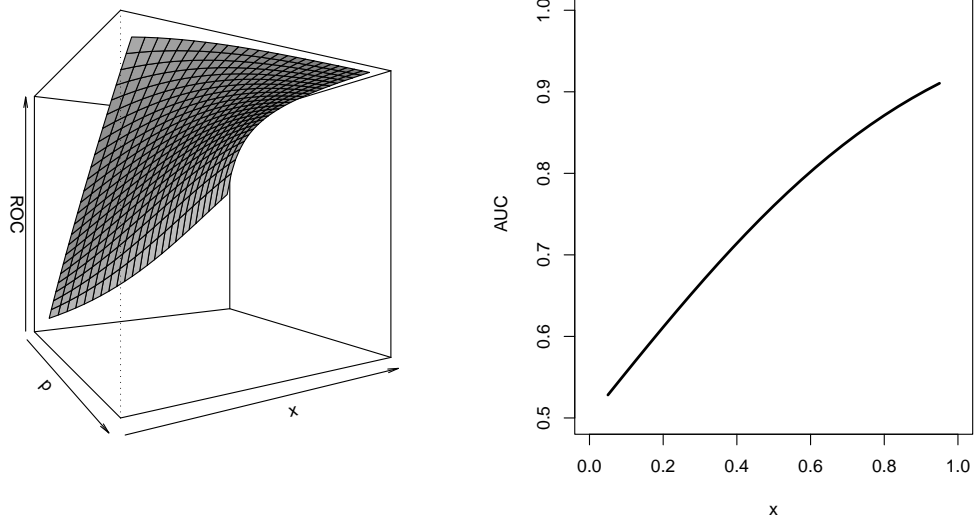


Figure 1: Conditional ROC curves (left) and conditional AUC (right) for scenario 1.

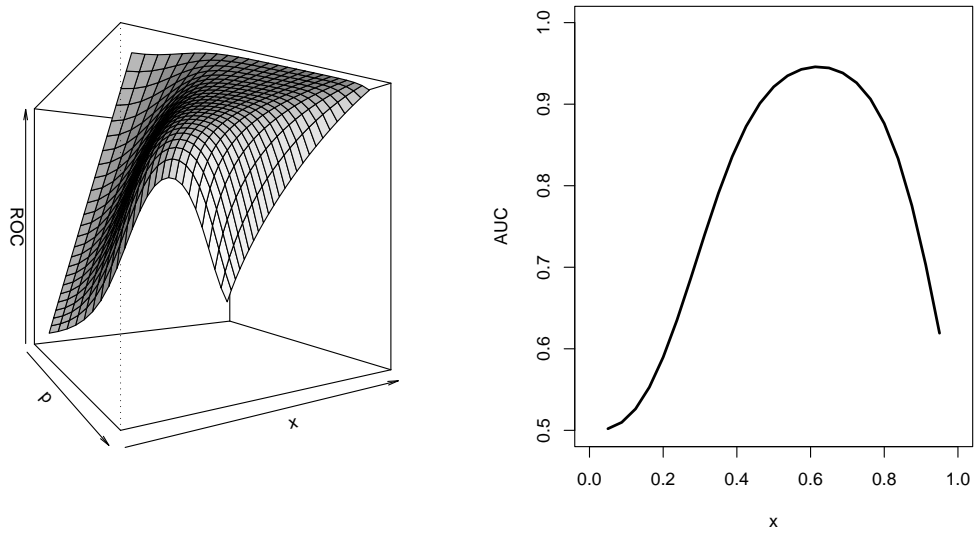
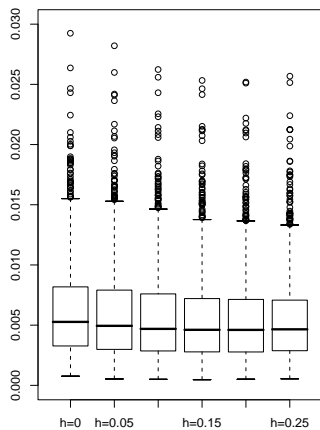


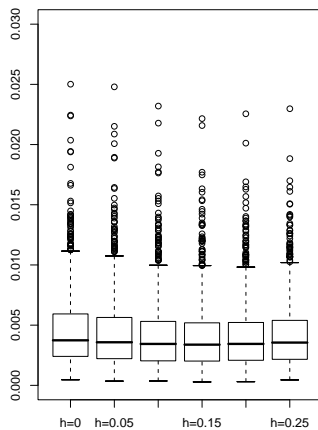
Figure 2: Conditional ROC curves (left) and conditional AUC (right) for scenario 2.

Table 1: Average and standard deviation (sd) of the estimated MSE ( $\times 1000$ ) obtained from 1000 data sets simulated according to scenario 1, for different sample sizes and different values of the smoothing parameter  $h$ .

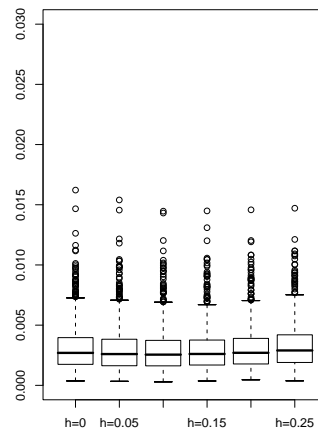
			$h$					
$n_0$	$n_1$		0.00	0.05	0.10	0.15	0.20	0.25
100	100	average	6.411	6.097	5.820	5.644	5.570	5.597
		sd	4.370	4.304	4.178	4.058	3.965	3.903
100	200	average	4.622	4.372	4.170	4.074	4.077	4.178
		sd	3.160	3.111	2.998	2.909	2.856	2.836
200	200	average	3.122	2.994	2.927	2.954	3.067	3.269
		sd	1.905	1.880	1.836	1.818	1.829	1.866



$n_0 = 100, n_1 = 100$



$n_0 = 100, n_1 = 200$



$n_0 = 200, n_1 = 200$

Figure 3: Boxplots of the estimated MSE obtained from 1000 data sets simulated from scenario 1, for different sample sizes and different values of the smoothing parameter  $h$ .

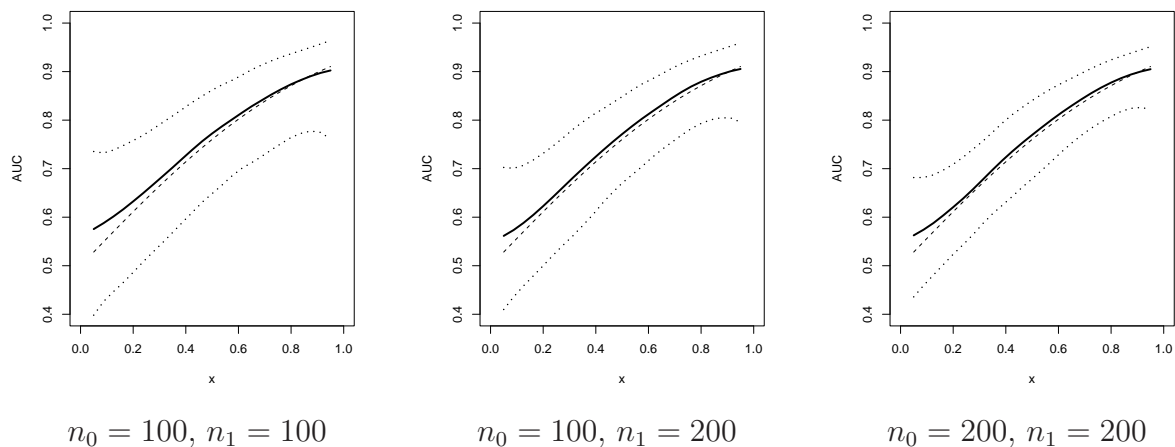


Figure 4: Average of the estimated conditional AUC (solid line) for different sample sizes when data are simulated from scenario 1, and true AUC (dashed line). In all cases  $h = 0.10$ . The bands (dotted lines) are obtained by means of the logit transformation.

Table 2: Average and standard deviation (sd) of the estimated MSE ( $\times 1000$ ) obtained from 1000 data sets simulated according to scenario 2, for different sample sizes and different values of the smoothing parameter  $h$ .

$n_0$	$n_1$		$h$					
			0.00	0.05	0.10	0.15	0.20	0.25
100	100	average	8.849	8.587	8.383	8.360	8.477	8.714
		sd	4.376	4.310	4.224	4.149	4.095	4.062
100	200	average	6.704	6.492	6.349	6.398	6.587	6.896
		sd	3.215	3.155	3.086	3.038	3.014	3.006
200	200	average	4.529	4.424	4.460	4.678	5.021	5.473
		sd	2.092	2.071	2.068	2.089	2.128	2.179

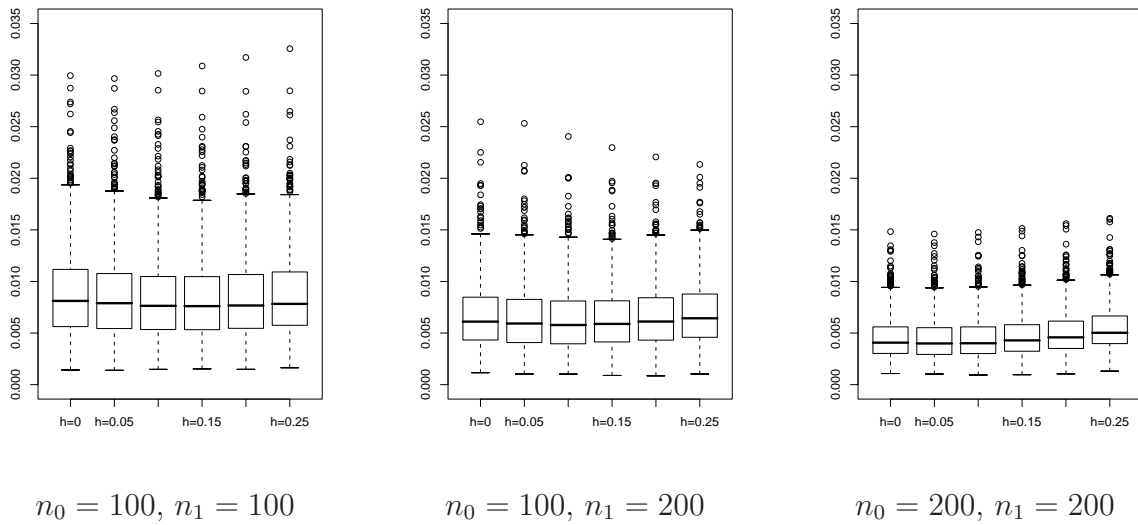


Figure 5: Boxplots of the estimated MSE obtained from 1000 data sets simulated from scenario 2, for different sample sizes and different values of the smoothing parameter  $h$ .

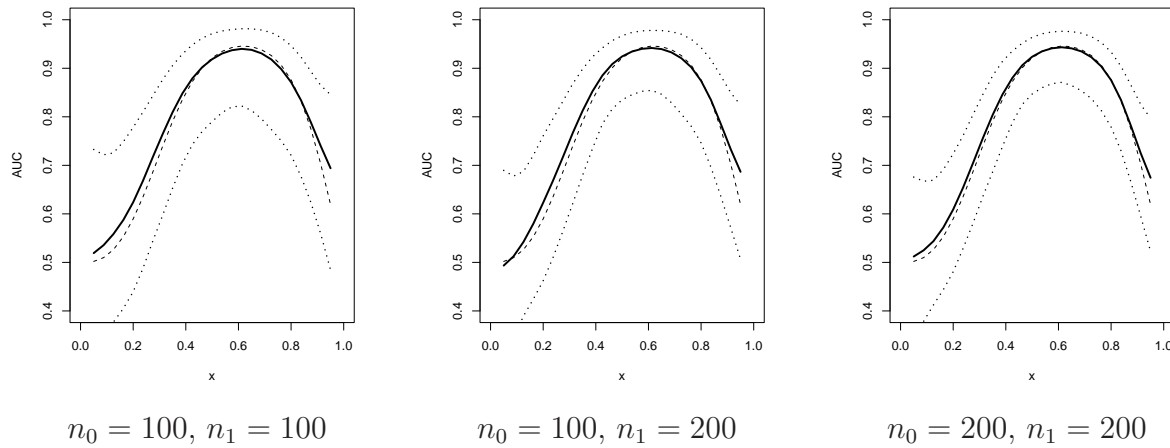


Figure 6: Average of the estimated conditional AUC (solid line) for different sample sizes when data are simulated from scenario 2, and true AUC (dashed line). In all cases  $h = 0.10$ . The bands (dotted lines) are obtained by means of the logit transformation.

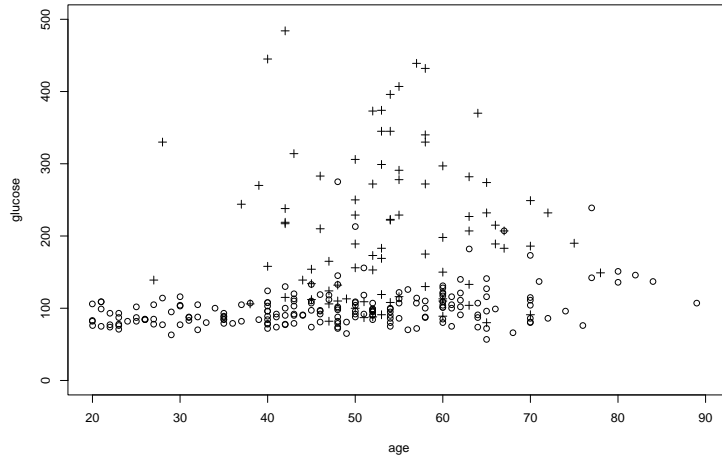


Figure 7: Scatter plot of the diagnostic variable ‘glucose concentration’ with respect to the covariate ‘age of the subject’. The diseased population is represented by crosses and the healthy population is represented by circles.

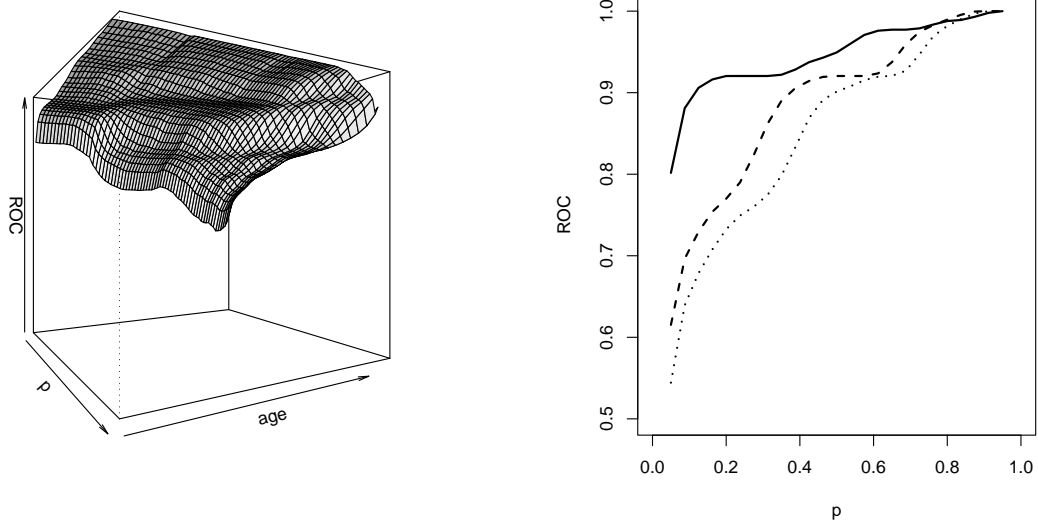


Figure 8: (a) Left: estimated conditional ROC curves. Right: conditional ROC curves for ages 30 (solid line), 50 (dashed line) and 70 (dotted line).

Table 3: Empirical coverages of the bootstrap confidence intervals for  $AUC_x$  obtained from 1000 data sets simulated from scenario 1 and scenario 2.

scenario	$(n_0, n_1)$	$h$	$1 - \alpha :$	$x = 0.25$		$x = 0.50$		$x = 0.75$	
				0.900	0.950	0.900	0.950	0.900	0.950
scenario 1	(100, 100)	0.00		0.851	0.911	0.899	0.949	0.887	0.936
		0.10		0.871	0.925	0.918	0.956	0.884	0.938
	(100, 200)	0.00		0.860	0.933	0.909	0.958	0.906	0.949
		0.10		0.886	0.952	0.927	0.964	0.895	0.946
	(200, 200)	0.00		0.857	0.916	0.895	0.944	0.900	0.948
		0.10		0.869	0.928	0.903	0.953	0.889	0.939
scenario 2	(100, 100)	0.00		0.778	0.839	0.886	0.932	0.865	0.919
		0.10		0.792	0.865	0.870	0.923	0.823	0.887
	(100, 200)	0.00		0.707	0.792	0.903	0.954	0.831	0.899
		0.10		0.752	0.828	0.901	0.943	0.782	0.849
	(200, 200)	0.00		0.790	0.864	0.905	0.956	0.882	0.934
		0.10		0.814	0.877	0.880	0.940	0.832	0.902



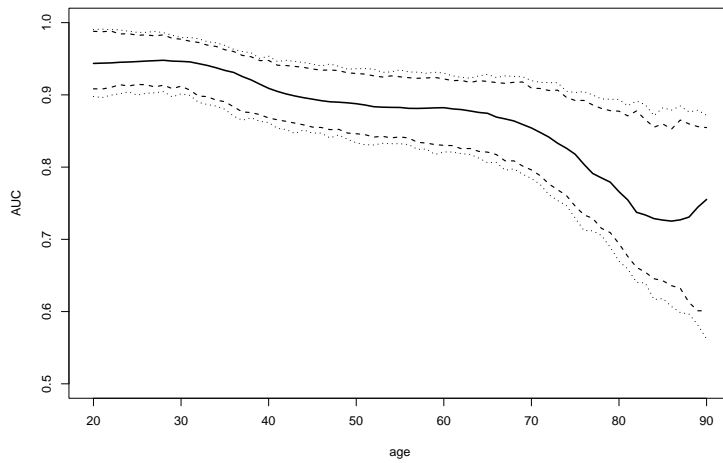


Figure 9: AUC as a function of age (solid line). The dotted and dashed lines represent 90% and 95% pointwise bootstrap confidence intervals, respectively.