

Research Group: *Food, Farms, and Firms*

December 16, 2009

Nonparametric vs Parametric Binary Choice Models: An Empirical Investigation

CHRISTOPHE BONTEMPS, JEFFREY S. RACINE
AND MICHEL SIMIONI

Nonparametric vs parametric binary choice models: An empirical investigation

Christophe Bontemps*, Jeffrey S. Racine[†] & Michel Simioni[‡]

December 16, 2009

Abstract

The estimation of conditional probability distribution functions (PDFs) in a kernel nonparametric framework has recently received attention. As emphasized by Hall, Racine & Li (2004), these conditional PDFs are extremely useful for a range of tasks including modelling and predicting consumer choice. The aim of this paper is threefold. First, we implement nonparametric kernel estimation of PDF with a binary choice variable and both continuous and discrete explanatory variables. Second, we address the issue of the performances of this nonparametric estimator when compared to a classic on-the-shelf parametric estimator, namely a probit. We propose to evaluate these estimators in terms of their predictive performances, in the line of the recent "revealed performance" test proposed by Racine & Parmeter (2009). Third, we provide a detailed discussion of the results focusing on environmental insights provided by the two estimators, revealing some patterns that can only be detected using the nonparametric estimator.

Keywords: Binary choice models, Nonparametric estimation, specification test, tap water demand

*Toulouse School of Economics, Gremaq-INRA, 21 alle de Brienne, 31000 Toulouse France.

[†]McMaster University, Canada.

[‡]Toulouse School of Economics, Gremaq-INRA & IDEL.

⁰We are grateful to Céline Nauges for her support and helpful comments. We also thank Yves Surry for giving us the opportunity to attend the lecture on nonparametric econometrics given at the SLU (Uppsala, Sweden) in June 2008, which has been the starting point of this work. We also thank the participants of the AAEA & ACCI 2009 Joint Annual Meeting (Milwaukee, USA) for comments.

1 Introduction

Recent developments in nonparametric estimation techniques have challenged conventional (parametric) practices when dealing with empirical questions. Most attention has been devoted to cases where the response variable is continuous (see Pagan (1999)). The existing literature addressing the issue of discrete response variables is more scarce. In this framework, efforts have been mostly done in the field of semiparametric tools. These tools have softened parametric specification assumptions involved in the computation of the probability that a given outcome of the response variable occurs as a function of a set of explanatory variables. For example, single index models allow for an unknown link function while maintaining the assumption that the explanatory variables enter the probability through a parametric index. These models have been surveyed by Horowitz (1998). In other models, explanatory variables enter through a general nonparametric additive structure, the link function being specified as the usual logistic or normal cumulative density function. Few attempts to introduce purely nonparametric estimation methods can be found in the literature on discrete choice (see Briesch, Chintagunta & Matzkin (2002)). While these methods produce results that are consistent with the economic foundations of choice theory, they are difficult to implement in practice.

The estimation of conditional PDFs in a kernel nonparametric framework has recently received attention. As emphasized by Hall et al. (2004), these conditional PDFs are extremely useful for a range of tasks including modelling and predicting consumer choice. The aim of this paper is threefold. First, we implement nonparametric kernel estimation of PDF with a binary choice variable and both continuous and discrete explanatory variables. Second, we address the issue of the performances of this nonparametric estimator when compared to a classic on-the-shelf parametric estimator, namely a probit. We propose to evaluate these estimators in terms of their predictive performances, in the line of the recent "revealed performance" test proposed by Racine & Parmeter (2009). Third, we provide a detailed discussion of the results focusing on environmental insights provided by the two estimators, emphasizing some patterns that can only be detected using the nonparametric estimator.

The empirical application reported here concerns an environmental question that was treated in Bontemps & Nauges (2009). In France, despite an access to safe public drinking water, and in spite of its excessively high price compared to tap water, 42% of the population still regularly drink bottled water. Using scanner data on French consumption combined with raw water quality and other environmental data, and using a Probit model, Bontemps & Nauges (2009) show that raw water bad quality seems to be the most important factor driving the decision not to drink tap water. The estimated effect is found to be stronger for low-income households. Significant direct impact of socioeconomic and demographic households' characteristics, as well as the role of cultural/regional factors are revealed. Overall, this study shows that pollution of raw water implies indirect costs for households who instead of drinking water from the tap spend up to 100 times more for bottled water. We revisit this question using the same dataset and using a kernel nonparametric estimator of the PDF in order to recover estimates of the probability to drink tap water. We address the three issues mentioned above, and show that the nonparametric estimator outperforms the parametric one.

In the paper, we provide an estimation of the probability of drinking tap water using a non-

parametric kernel conditional probability specification based on the conditional PDF of the decision variable. This specification uses the same information set than the probit specification and we use recent developments on generalized kernel estimation to deal with both continuous and discrete variables (see Li & Racine (2007)). Moreover, this setup enables to assess the relevance of the explanatory variables through the data-driven bandwidth selection method. We find a similar set of significant variables with e.g. the raw water bad quality index. We find similar in-sample predictive performances, that is the overall correct classification ratio is 70.04% when considering the NP specification while it is of 69.17% for the probit specification. We then address the issue of selecting a preferred model specification between the fully NP and the probit. We take the view that fitted statistical models are approximations, a perspective that differs from that of consistent models selection which posits a finite dimensional ‘true model’. Recently, p Racine & Parmeter (2009) propose a test based on this idea that we implement here. The test results indicate that the NP specification possesses an expected ‘true error’ that is statistically significantly lower than that for the parametric specification and is therefore preferred. Finally, we contrast the parametric specification with the nonparametric specification on some interesting cases.

The rest of the paper proceeds as follows. Section 2 outlines the nonparametric estimator and the test procedure. Section 3 presents the empirical application of these methods to the environmental question raised in Bontemps & Nauges (2009). Section 4 presents some concluding remarks.

2 Nonparametric estimation and test procedures

2.1 Nonparametric PDF estimator

Let $f(\cdot)$ and $m(\cdot)$ denote the joint and marginal density of (X, Y) and X , respectively, where Y is a binary discrete variable and where we allow X to include both continuous, unordered, and ordered variables. For what follows, we shall refer to Y as a dependent variable (i.e., Y is explained), and to X as covariates (i.e., X is the vector of the explanatory variables). The density of Y conditional on X is then defined as

$$g(y|x) = \frac{f(x, y)}{m(x)} \quad (1)$$

Consider the kernel estimators of the previous joint and marginal densities we denote by \hat{f} and \hat{m} . We then estimate the conditional density by replacing the unknown densities in (1) by their estimators, i.e.

$$\hat{g}(y|x) = \frac{\hat{f}(x, y)}{\hat{m}(x)} \quad (2)$$

As we are facing a mix of discrete (unordered and ordered) and continuous variables when estimating the two unconditional densities, we use the “generalized product kernel” estimator

proposed by Li and Racine (2003). Let $X = (X^c, X^d, \tilde{X}^d)$ represent the division of X into its p continuous, q discrete unordered, and r discrete ordered components, then the estimator of the marginal density $m(\cdot)$ for a given realization of X denoted by $x = (x^c, x^d, \tilde{x}^d)$ is given by

$$\begin{aligned}\hat{m}(x) &= \hat{m}(x^c, x^d, \tilde{x}^d) \\ &= n^{-1} \sum_{i=1}^n \prod_{j=1}^p W(X_{ij}^c, x_j^c) \prod_{j=1}^q l(X_{ij}^d, x_j^d) \prod_{j=1}^r \tilde{l}(\tilde{X}_{ij}^d, \tilde{x}_j^d)\end{aligned}\quad (3)$$

where we use different kernels depending on the nature of the variable under consideration. That is:

- For a continuous variable x_j^c , we use the function $W(\cdot)$ defined as

$$W(X_{ij}^c, x_j^c) = \frac{1}{h_j} K\left(\frac{X_{ij}^c - x_j^c}{h_j}\right)$$

where $K(\cdot)$ is a traditional kernel function, i.e. a symmetric, univariate probability density, and h_j is the bandwidth.

- For an unordered discrete variable x_j^d with c_j categories, we use Aitchison & Aitken (1976) kernel given by:

$$l(X_{ij}^d, x_j^d) = \begin{cases} 1 - \lambda_j & \text{if } X_{ij}^d = x_j^d \\ \frac{\lambda_j}{c_j - 1} & \text{otherwise.} \end{cases}$$

where the bandwidth λ_j belongs to the interval $[0, (c_j - 1)/c_j]$. Note that when $\lambda_j = 0$ the kernel l becomes an indicator function, i.e., the function which is usually chosen as the kernel in the "frequency" approach of discrete variables in nonparametric estimation. When $\lambda_j = (c_j - 1)/c_j$, the kernel $l(X_{ij}^d, x_j^d) = 1/c_j$ for all values of X_{ij}^d and x_j^d .

- For an ordered discrete variable \tilde{x}_j^d , we use Wang & Van Ryzin (1981) kernel given by:

$$\tilde{l}(\tilde{X}_{ij}^d, \tilde{x}_j^d) = \begin{cases} 1 & \text{if } \tilde{X}_{ij}^d = \tilde{x}_j^d \\ \frac{1}{\gamma_j} & \text{otherwise.} \end{cases}$$

where the bandwidth γ_j belongs to the interval $[0, 1]$. Again, when $\gamma_j = 0$, the kernel \tilde{l} becomes an indicator function, and, when $\gamma_j = 1$, this kernel is a uniform weight function.

Similarly, the estimator of the joint density $f(\cdot)$ for a given realization of (X, Y) denoted by $(x, y) = (x^c, x^d, \tilde{x}^d, y)$ is given by

$$\begin{aligned} \hat{f}(x, y) &= \hat{f}(x^c, x^d, \tilde{x}^d, y) \\ &= n^{-1} \sum_{i=1}^n \prod_{j=1}^p W(X_{ij}^c, x_j^c) \prod_{j=1}^q l(X_{ij}^d, x_j^d) \prod_{j=1}^r \tilde{l}(\tilde{X}_{ij}^d, \tilde{x}_j^d) \times l(Y_i^d, y) \end{aligned}$$

The computation of the two previous estimators involves the choice of the bandwidths h_j , $j = 1, \dots, p$, λ_j , $j = 1, \dots, q$, and γ_j , $j = 1, \dots, r$. Recently, Hall et al. (2004) show that we could pursue a least-squares cross validation to selecting these bandwidths. They consider the following criterion which is based upon a weighted integrated square error:

$$\sum_{x^d} \int \{\hat{g}(y|x) - g(y|x)\}^2 m(x) M(x^c) dx dy$$

where $M(\cdot)$ is a weight function. Moreover the choice of this criterion allows to determine which components of the vector X are relevant when conducting conditional inference. Indeed, the data-driven chosen bandwidths will exhibit a markedly dichotomous behavior. On one hand, the minimization of the least-squares cross-validation criterion will assign large smoothing parameters to the irrelevant components (i.e., $h_j \rightarrow \infty$ or $\lambda_j \rightarrow (c_j - 1)/c_j$ or $\gamma_j \rightarrow 1$ depending on the nature of the variable), and, consequently, will shrink these components toward the uniform distribution on the respective marginals. On the other hand, the minimization will assign smoothing parameters of conventional size to the relevant components. For example, $h_j = O_p(n^{-1/(p+5)})$. To sum up, cross-validation produces asymptotically optimal smoothing for relevant components, while eliminating irrelevant ones by oversmoothing.

2.2 Preferred model selection

We consider two non-nested model specifications, a parametric probit specification and a nonparametric kernel conditional probability specification. Both models use identical information sets and deliver estimates of the probability that $Y = 1$ conditional on the covariates X .

We approach the issue of selecting a parametric versus nonparametric specification from the perspective that fitted statistical models are approximations. Clearly our perspective is distinct from that of consistent model selection which posits a finite-dimensional ‘true model’. We therefore consider selection of a parametric versus nonparametric specification not from the perspective of a test that posits that one model is ‘true’ hence set up a null as such. Rather, both are at best decent approximations, therefore we select that model that has lowest expected ‘true error’.

Our approach is therefore firmly embedded in the statistics literature dealing with ‘apparent’ versus ‘true’ error estimation; for a detailed overview of expected apparent and excess error, we direct the reader to Efron (1982, Chapter 7). In effect, in-sample measures of fit such as the standard

error of the regression or R^2 and so forth measure ‘apparent error’ which will be smaller than ‘true error’ which is the expected error when the model is used to predict new draws from the underlying data generating process. For example, for a continuous regression model $Y_i = g(X_i) + \varepsilon_i$, one might compute the Average Square Prediction Error or ASPE given by $ASPE = n_1^{-1} \sum_{i=1}^{n_1} (Y_i - \hat{g}(X_i))^2$ which is a measure of apparent error. But all such in-sample measures are fallible which is why they cannot be recommended as guides for model selection. Our procedure can be thought of as a means of *estimating* a model’s *true error* and *testing* whether the true error is statistically smaller for one model than another. The statistics literature on cross-validated estimation of excess error is a well-studied field. However, this literature deals with model specification within a class of models (i.e., which predictor variables should be used, whether or not to conduct logarithmic transformations on the dependent variable and so forth) and proceeds by minimizing excess error. Our purpose here is substantively different. Here we pose the question of whether the true error associated with one model differs *significantly* from that for another model. We adopt the ‘revealed performance’, or RP, test proposed by Racine & Parmeter (2009).

Before introducing the RP test, let us recall how predictive performance is measured in binary choice models. Different indices can be used to measure this predictive performance. Efron (1978) gives a detailed discussion of such indices. The most common index used is the Correctly Classified Ratio (*CCR*) or *accuracy*. This index measures the exact proportion of right predictions over the considered sample. That is, for each observation i , we compute the value of the loss function

$$Q(Y_i, \eta_i, \alpha) = \begin{cases} 0 & \text{if } Y_i = 1 \text{ and } \eta_i > \alpha \text{ or if } Y_i = 0 \text{ and } \eta_i \leq \alpha \\ 1 & \text{otherwise} \end{cases}$$

where η_i is the probability assigned by the binary choice model to the i th observation, and α is the cut-off-value used to map the classifier, namely the probabilities η_i , to classes of predicted 0 or 1. For a given cut-off-value (usually, $\alpha = 0.5$), the *CCR* is then computed as

$$CCR(\alpha) = 1 - \frac{\sum_{i=1}^n Q(Y_i, \eta_i, \alpha)}{n}. \quad (4)$$

This index can be also linked to the so-called confusion matrix we define in Table 1 where *ON* and *OP* are the total numbers of observed 0 and 1 respectively, *TN* stands for ‘true negative’, occurring when both the observed value and the prediction outcome ($\eta_i \leq \alpha$) are 0, and *FN* for ‘false negative’, when the observed value is 0 and the prediction outcome is 1 ($\eta_i > \alpha$), while *FP* and *TP* are the ‘false’ and ‘true positive’ respectively.

The *CCR* index can be shown to be very sensitive to the choice of the cut-off-value. An other well-known classification performance metric that circumvents this drawback, is the ‘Receiver Operating Characteristic’ curve (*ROC*), described in Egan (1975). This curve is a graphical plot of the *sensitivity* (percentage of predicted true positive) versus $1 - \textit{specificity}$ (percentage of predicted false positive) (see Table 1 for more precise definitions), letting the classification cut-off-value vary between its extremes. The *AUROC*, i.e., the ‘Area Under the Receiver Operating Characteristic’ curve, can then be computed as a summary measure of the classification performance. These second measure whose values are between 0.5 (worthless classification) and 1 (perfect classification), provides a better evaluation ratio than the *CCR* index, since it is independent of any cut-off-value.

Table 1: Notations

		Predicted			$Accuracy$ (CCR) = $\frac{TN+TP}{n}$
		0	1	Total	
Obs.	0	TN	FP	ON	$Sensitivity$ (TPR) = $\frac{TP}{TP+FN}$
	1	FN	TP	OP	
Total				n	$Specificity$ (SPC) = $\frac{TN}{TN+FP}$
Confusion matrix					Index

Given the two previous measures (CCR and $AUROC$), we can define two different ways of measuring the 'apparent error' when estimating a binary choice model. For instance, the second term in the definition of the CCR measure (see equation 4) can be viewed as the empirical realization of $E_{n_1, \hat{F}}[Q(Y^{n_1}, \eta_{n_1}^{n_1}), 0.5]$ where $E_{n_1, F}$ denotes the expectation over the n_1 observed points $Z^{n_1} = \{Y_i, X_i\}_{i=1}^{n_1}$ which are independently and identically distributed with empirical cumulative distribution function \hat{F} (we refer Z^{n_1} as the training sample, terminology borrowed from the literature on statistical discriminant analysis), $Y^{n_1} = \{Y_i\}_{i=1}^{n_1}$, and $\eta_{n_1}^{n_1} = \{\eta_i\}_{i=1}^{n_1}$, i.e. the vector of the assigned probabilities to the observations calibrated using the observed sample. In order to implement the RP test, we are interested in estimating a quantity known as 'expected true error'. Following Efron (1982), we can define the 'true error' to be

$$E_{n_2, F}[Q(Y^{n_2}, \eta_{n_1}^{n_2}, 0.5)]$$

where $E_{n_2, F}$ denotes the expectation over the n_2 new points $Z^{n_2} = \{Y_i, X_i\}_{i=n_1+1}^n$ which are independently and identically distributed with cumulative distribution function F , and are independent of the training sample Z^{n_1} (we refer Z^{n_2} as the evaluation sample), $Y^{n_2} = \{Y_i\}_{i=n_1+1}^n$, and $\eta_{n_1}^{n_2} = \{\eta_i\}_{i=n_1+1}^n$, i.e. the vector of the probabilities assigned to the new points calibrated using the training sample. Next, we define the 'expected true error' as

$$E(E_{n_2, F}[Q(\cdot)])$$

where the expectation is taken over all potential classifiers $\eta_{n_1}^{n_2}$, for the selected loss function $Q(\cdot)$. When comparing two approximate models, the model possessing the lower 'expected true error' will be preferred in applied settings.

A realization of the 'true error' based upon the observed $z^{n_2} = \{y_i, x_i\}_{i=n_1+1}^n$ is given by

$$\frac{1}{n_2} \sum_{i=n_1+1}^n Q(y_i, \eta_{n_1, i}^{n_2}, 0.5) \tag{5}$$

In the following, we consider the corresponding *CCR* for ease of interpretation. If we consider S splits of the data into training and evaluation samples, we can then construct the empirical distribution function of loss, or equivalently, of *CCR*. The algorithm used to construct the empirical distributions of *CCR* for the two competing models proceeds as follows:

- (i) Resample without replacement pairwise from $Z = \{X_i, Y_i\}_{i=1}^n$ and call these $Z_* = \{X_i^*, Y_i^*\}_{i=1}^n$.
- (ii) Let the first n_1 of the resampled observations form a training sample $Z_*^{n_1} = \{X_i^*, Y_i^*\}_{i=1}^{n_1}$ and the remaining $n_2 = n - n_1$ observations form an evaluation sample, *i.e.*, $Z_*^{n_2} = \{X_i^*, Y_i^*\}_{i=n_1+1}^{n_2}$.
- (iii) Holding the degree of smoothing at that for the full sample (*i.e.*, the bandwidths scaling factors) of the nonparametric model and the functional form of the Probit fixed, fit each model on the training observations $Z_*^{n_1}$, and then obtain predicted values from the evaluation $Z_*^{n_2}$ that were not used to fit the model.
- (iv) Compute the *CCR*(0.5) of each model.
- (v) Repeat this a large number of times, say, $S = 10,000$, yielding S draws of *CCR* for the two models.

We clone this procedure with *AUROC* replacing *CCR*.

3 Empirical application

3.1 Data

The data set used is based on two main sources. First, data on French households' purchases are provided by TNS Worlpanel, for the year 2001. This database contains information on French households' purchases of food items as well as households' socioeconomics and demographics.¹ We define each household as a "tap water drinker" or "tap water non-drinker" from observed purchases of non-alcoholic drinks.² Second we use various sources of environmental information at the township level to compute the index of *raw water bad quality* (RWBQ hereafter) as in Bontemps & Nauges (2009). Data on price of tap water (IFEN-SCEES-Agences de l'eau, 2001), data on raw water (Ministry of Health, 2001), and data on manure spreading by the Ministry of Agriculture (2000) at the township level were collected. An information on the management of water supply chosen at the township level (public versus private) is also included via the dummy variable *deleg*. Finally, we compute the index of raw water bad quality and merged it to the households' panel through the residential address information of each household.

In addition to the *raw water bad quality* index, we observe the following socioeconomics and demographics at the household level:

¹Purchases of 10,000 surveyed households are recorded all over the year 2001.

²We follow the definition of Bontemps & Nauges (2009) by defining the household as "tap water drinker" (resp. "tap water non-drinker") if the average consumption of non-alcoholic drinks by person by day is lower (resp. greater) than 0.5 liters. The set of non-alcoholic drinks includes: bottled water, tea drinks, sodas, tonics, fruits and vegetables juices, etc.

- Head of household’s education level (**diploma**). We distinguish four education levels: head without diploma (reference in the probit model), head with diploma less than the *baccalaurat* (diplo.L), head with the *baccalaurat* or a higher diploma (diplo.Q), head for whom information is missing (diplo.C).
- Household’s monthly income (before income taxes) (**Income**).
- Rural or urban location (**rural**): we build an indicator variable which takes the value of 1 if the household lives in a "commune" of less than 2,000 inhabitants, and 0 otherwise.
- Retirement status (**iret**): we build an indicator variable which takes the value of 1 if household’s head is retired and 0 otherwise.³
- Household’s geographic location (**region**). We follow here the regional division chosen by TNS. France is divided in 8 main zones: Paris, East, North, West, Middle-West, Middle-East, South-East, and South-West

The sample gathers 4,623 households from 1,282 distinct townships ("communes") all over France. A complete description of the sample is given in Bontemps & Nauges (2009). We recall the main feature of this sample. On this sample 68% of households in the sample are classified as "tap water drinkers". This percentage is close to what is usually found in polls at the nation level. Regional differences are also observed in our sample, the highest percentages of tap water drinkers are in the Middle-East (82%), South-West (80%), and South-East (77%). The lowest proportion of tap water drinkers is observed in the North (56%). The average raw water bad quality index is of 0.93, varying from 0.87 in the North of France up to 0.97 in the Paris Region and in the west of France, these two regions being particularly affected by nitrogen pollution. Household monthly income varies from an average of 1,820 euros in Middle-West to an average of 2,490 euros in Paris and surroundings. The proportion of households living in rural area varies from 1% (Paris and surroundings) to 15% in the Western region where farming activity dominates. The share of retired households is quite homogeneous across regions, except in the South-East which attracts a lot of retired people with its warmer climate.

3.2 In-sample global performances

As mentioned in 2.2, we estimate two non-nested model specifications, a parametric probit specification and a nonparametric kernel conditional probability specification. Both models use identical information sets and deliver estimates of the probability of drinking tap water conditional on the covariates $X = \{\text{RWBQ}, \text{Income}, \text{diploma}, \text{region}, \text{deleg}, \text{rural}, \text{iret}\}$. For the probit specification, we add the interactions $\text{RWBQ} \times \text{Income}$ and $\text{RWBQ} \times \text{iret}$ in order to use the "preferred" parametric specification of Bontemps & Nauges (2009).

Table 2 reports the estimates of the parameters of the probit specification and the bandwidths selected by least-squares cross-validation for the nonparametric specification.⁴ Empirical significance levels allow to test the significance of the associated variables in the parametric specification

³This variable was proven more significant than age of household’s head in Bontemps & Nauges (2009).

⁴All the computations are made using the `np` package of R software (see Hayfield & Racine (2008)).

Table 2: Probit coefficient estimates and significance vs nonparametric bandwidth estimates and associated scale factors:

	Estimate	$Pr(Z > z)$	Bandwidth	upper bound
(Intercept)	2.3296	0.0000	-	-
RWBQ	-1.8113	0.0021	0.1801905	∞
Income	-0.5492	0.0155	1.294752	∞
diploma	-	-	0.8634835	1
diplo.L	-0.1328	0.0464	-	-
diplo.Q	0.0435	0.4433	-	-
diplo.C	-0.0229	0.5703	-	-
Region	-	-	0.1208747	0.875
Region2	-0.0284	0.7376	-	-
Region3	-0.5879	0.0000	-	-
Region4	-0.0590	0.3836	-	-
Region5	-0.0468	0.5887	-	-
Region6	0.3706	0.0000	-	-
Region7	0.1486	0.0576	-	-
Region8	0.2974	0.0005	-	-
deleg	-0.0178	0.6966	0.5	0.5
rural	0.2397	0.0095	0.0721212	0.5
iret	-1.3491	0.0089	3.253532e-13	0.5
RWBQ×Income	0.5789	0.0166	-	-
RWBQ× iret	0.9461	0.0871	-	-
irob	-	-	9.802058e-15	0.5

The bandwidths are chosen by minimizing a least-square cross-validation criterion.

The upper bound for a bandwidth, is equal to $(c_j - 1)/c_j$ in the case of an unordered discrete variable with c_j categories, and 1 in the case of an ordered one.

while the magnitudes of the selected bandwidths reveal the relevance of the associated variables in the nonparametric specification. We observe that significance tests and relevance assessments provide coherent results. All the continuous variables have significant effects on the probability to drink tap water in the parametric specification, and are relevant in the nonparametric specification as they are far from being "smoothed-out". For the discrete variables, **rural** and **iret** are significant and their bandwidths are far from their upper bounds, indicating clearly their relevance. This is not the case for the variable **deleg**. For the ordered discrete variables, the two specifications show that some regional effects exist while the diploma effects seem to be less clear.

Table 3 reports the confusion matrices for the two specifications. As usually done when evaluating binary choice models, we use a cut-off value $\alpha = 0.5$ to map the classifiers, namely the estimated probabilities, to classes of predicted 0 ("the household does not drink tap water") or 1 ("the household does"). The comparison of these confusion matrices is not very conclusive, the two specifications tending to over-predict the fact of drinking tap water. Moreover, the two

specifications produce similar values for the usual global performance measures as shown in Table 4.

Table 3: Confusion matrices on the sample ($n = 4623$ and $\alpha = 0.5$)

Probit specification				Nonparametric specification					
		Predicted				Predicted			
		0	1			0	1		
Obs.	0	143	1324	1467	Obs.	0	134	1333	1467
	1	101	3055	3156		1	52	3104	3156
		244	4379	4623			186	4437	4623

Table 4: Global performances ($n = 4623$ and $\alpha = 0.5$)

Model	Sensitivity	Specificity	CCR
Probit	96.79 %	9.74 %	69.17 %
Nonparametric	98.35 %	9.13 %	70.04 %

In Figure 1 we report two graphs representing the global performances of the two models when the cut-off-value (α) is varying. The nonparametric specification provides better accuracy for a given range of the cut-off-values, roughly $\alpha \in [0.45, 0.70]$, than the parametric specification, the difference between the CCR values being very small outside this interval. In the same way, the ROC curve for the nonparametric specification always dominates the ROC curve for the parametric specification but the difference between the areas under these two curves is quite small (0.023).

To sum up, the nonparametric specification weakly outperforms the Probit specification when considering in-sample performances measures. But, let us recall first that the Probit specification with the chosen interactions between explanatory variables was selected to fit the data best using these criteria in Bontemps & Nauges (2009). Second, as emphasized by (Racine & Parmeter 2009), there is no guarantee that the nonparametric specification will perform any better than the Probit specification, even though the former may indeed exhibit an apparent marked improvement in (in-sample) fit according to the chosen performance measures. We will see in the following that focusing on out-of-sample predictive ability provides a useful tool for discriminating among the two specifications.

3.3 Test results

We consider the RP test using $S = 10,000$, where S is the number of splits of the data into two independent samples of size $n_1 = n - n_2$ and n_2 . In our example we have $n = 4,623$ and we report results for $n_2 = 250$.⁵ For each split into two independent samples, we fit each model to the n_1

⁵Results are qualitatively unchanged for other choices of n_2 and are available from the authors upon request.

Figure 1: Global performances with varying cut-off-value α

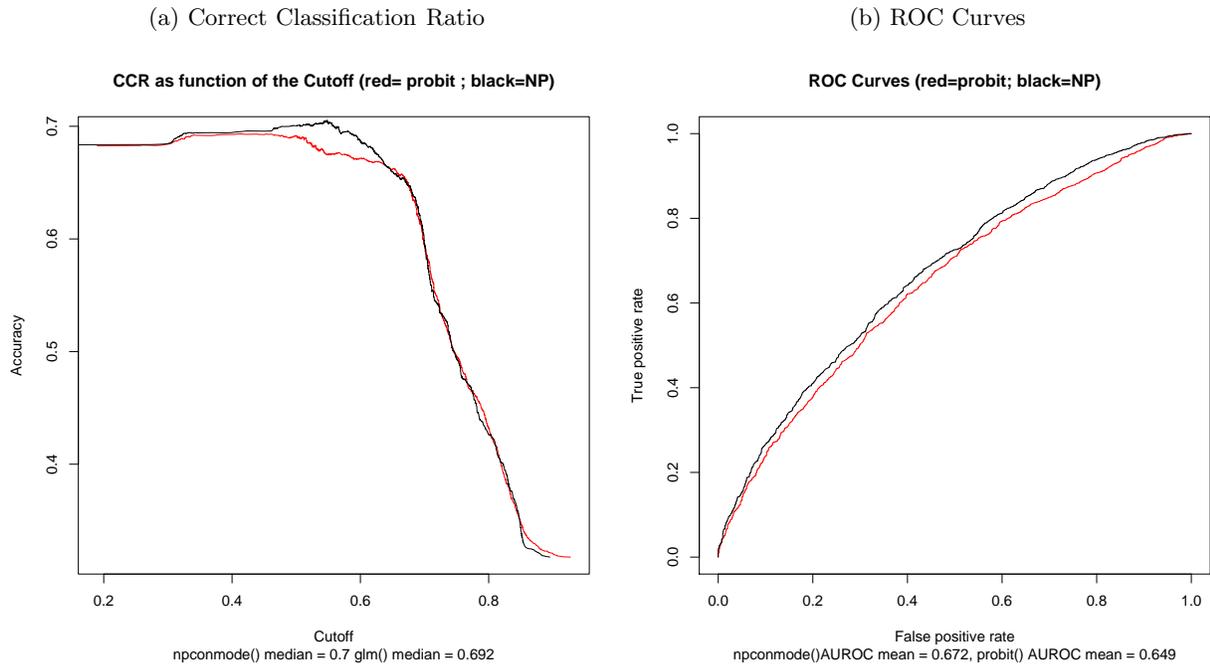
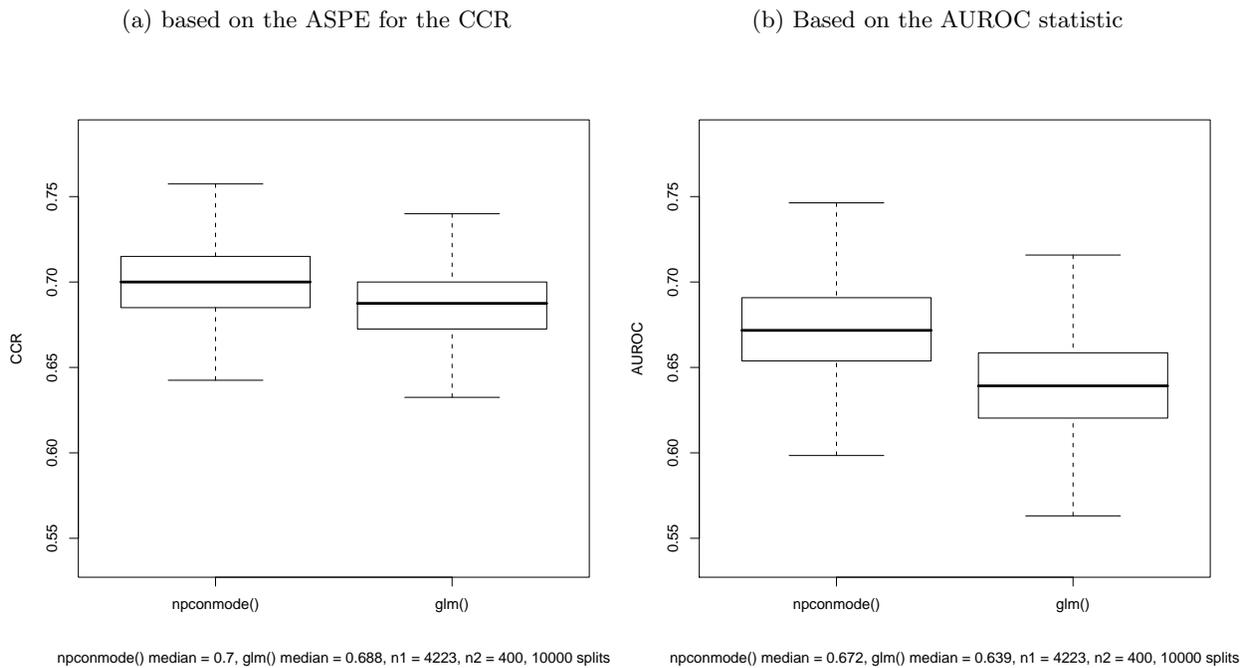


Figure 2: Box-plots of the RP test statistics for the S=10,000 splits of the data



observations, obtain predictions for the values of the covariates in the independent sample of size n_2 , then compute the error associated with the response in the independent sample. We repeat this S times, then test whether the expected error on the independent data is equal nor not. The null is that the expected true error is equal for both models, the alternative that is smaller for the nonparametric model. We use both a simple t -test and also a Mann-Whitney-Wilcoxon test. The P -values for each test are $2.816224e - 12$ and $1.397127e - 11$, respectively, indicating that the nonparametric specification possesses expected true error that is statistically significantly lower than that for the parametric specification and is therefore preferred.

In Figure 2a, results are presented in the form of boxplots of $ASPE$ for each of the two specifications. It can be seen from this figure that a stochastic dominance relationship exists between the nonparametric specification and the Probit one, confirming the previous RP test result and again indicating that the nonparametric model is to be preferred on the basis on independent draws from the data. Does this dominance depend on the choice of a cut-off value $\alpha = 0.5$? We address this question by computing the $AUROC$ value for each of the S replications involved in the RP test. These area do not depend on any chosen cut-off value and thus provide a more robust indicator of the performance of the classification than $ASPE$. In Figure 2b we compare the empirical distributions of the $AUROC$ for the two specifications. The two boxplots are less overlapping than the boxplots in Figure 2a, indicating a clear-cut stochastic dominance of the nonparametric specification.

3.4 Environmental issues

We now focus on the respective insights on tap water consumption raised by the two models. For example, a graphical comparison of the estimated probabilities of drinking tap water expressed as functions of the two continuous variables $RWBQ$ and $Income$ is provided in the 3-Dimensional surface plots (Figure 3 and 4) where we fix the variables $diploma$, $Region$, $deleg$, and $rural$ at chosen values (resp. for household with diploma lower than *Baccalauréat*, in the North of France, no delegation for tap water distribution, not in a rural area) and we let the $iret$ variable change from retired to non-retired. This comparison allows to show how the variables $iret$ and $RWBQ$ interact given the way this interaction is introduced in the two specifications. Both probabilities shift downwards for retired people. We observe that changing from retired to non retired induces a flip in the shape of the probit probability, being increasing with respect to $Income$ and decreasing with respect to $RWBQ$ for non-retired consumers and being increasing with respect to both $Income$ and $RWBQ$ for retired consumers. The nonparametric probability is more flat whatever the value of $Income$ and $RWBQ$ for non retired people, but it exhibits a similar pattern than the Probit one for retired people. This effect for retired people is surprising, a closer look shows that the probability is varying more in the $RWBQ$ dimension than in the $Income$ dimension, capturing an reverse environmental effect. One still observe, that the richer the household, the higher the probability to drink tap water, whatever the environmental quality, but to lower extent. The nonparametric probability is always bigger than the parametric one for retired people, while the two surfaces cross for non retired people. Note that the surfaces for the nonparametric specification never cross the 0.5 cut-off value, being always bigger for retired people and smaller for non retired people. That is, on this example, the $iret$ variable fully discriminates between the tap water drinkers and non-drinkers when using the usual 0.5 cutoff value. This result can be seen on the 2-Dimensional Figure 5, where

Figure 3: Surface plots for the two models (retired=0)

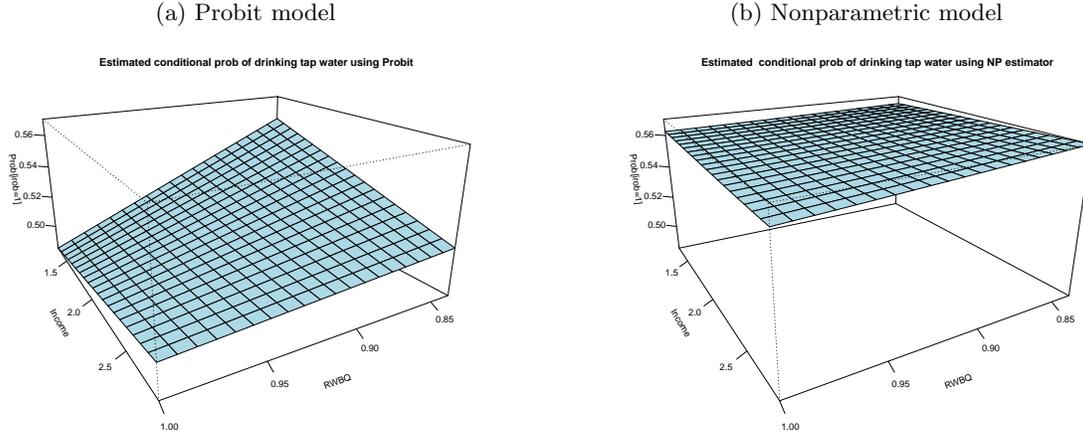
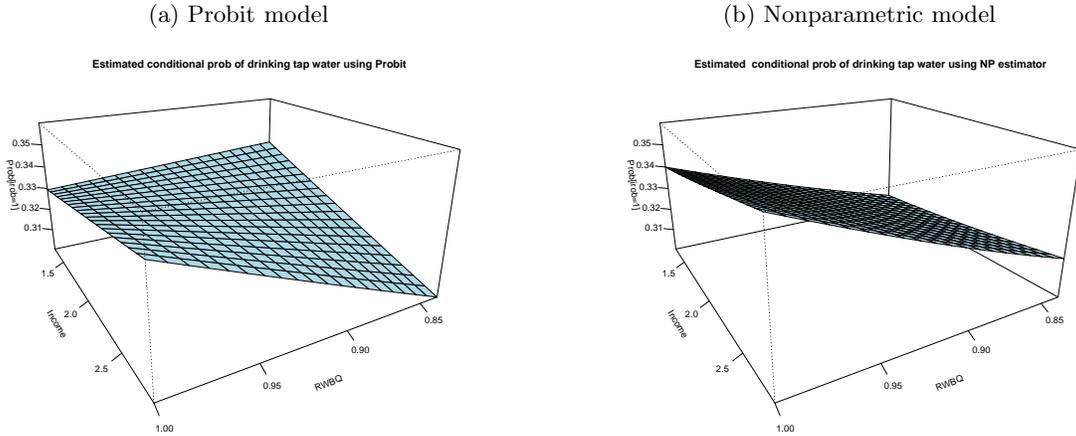


Figure 4: Surface plots for the two models (retired=1)



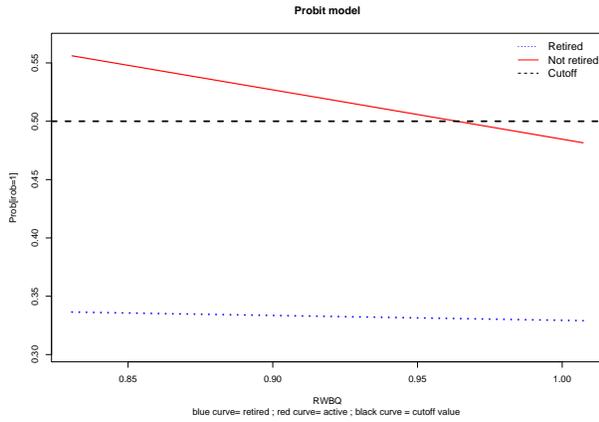
we superpose the estimated probabilities as functions of RWBQ for both retired and non retired, Income being fixed at its second quantile.

In figure 6, we report slices of the previous 3-Dimensional surfaces for a very bad level of environmental quality (the 9th RWBQ quantile). In both models, we observe that the probability of drinking tap water is greater for non retired people, and this probability is slightly increasing with income. Here again, the the ired variable discriminates between tap water drinkers and non drinkers for the nonparametric model when using a cut-off value of 0.5. This is not the case when considering the Probit model.

Let us now consider the specific case of discrete variables that do not interact with RWBQ. In the Probit model, changing the value of such a variable, *ceteris paribus*, induces a translation

Figure 5: Retirement effect and environment

(a) Probit model



(b) Nonparametric model

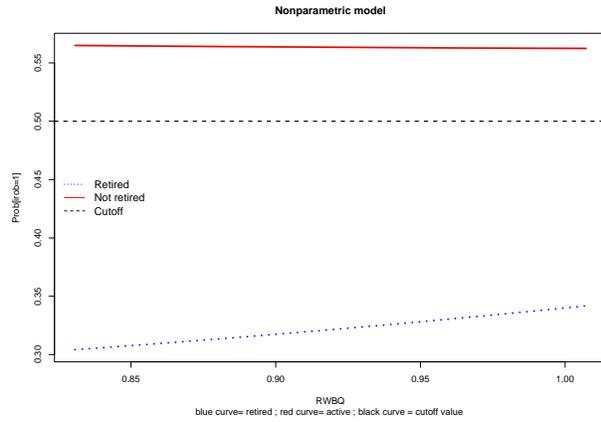
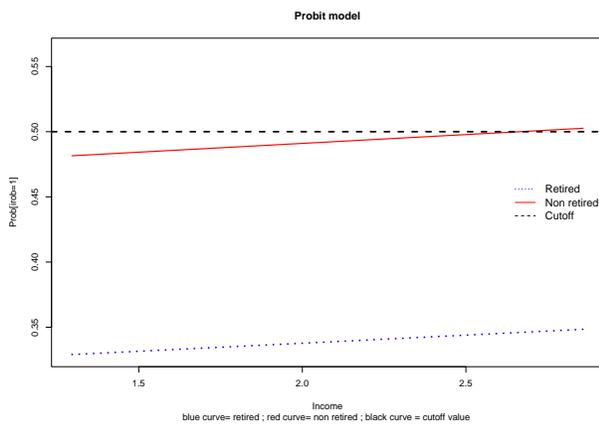


Figure 6: Retirement effect and income

(a) Probit model



(b) Nonparametric model

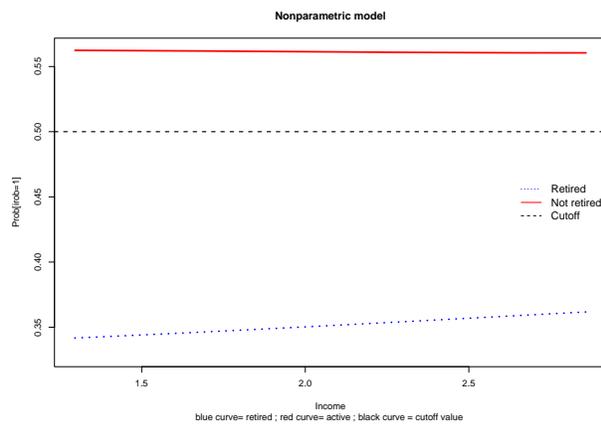
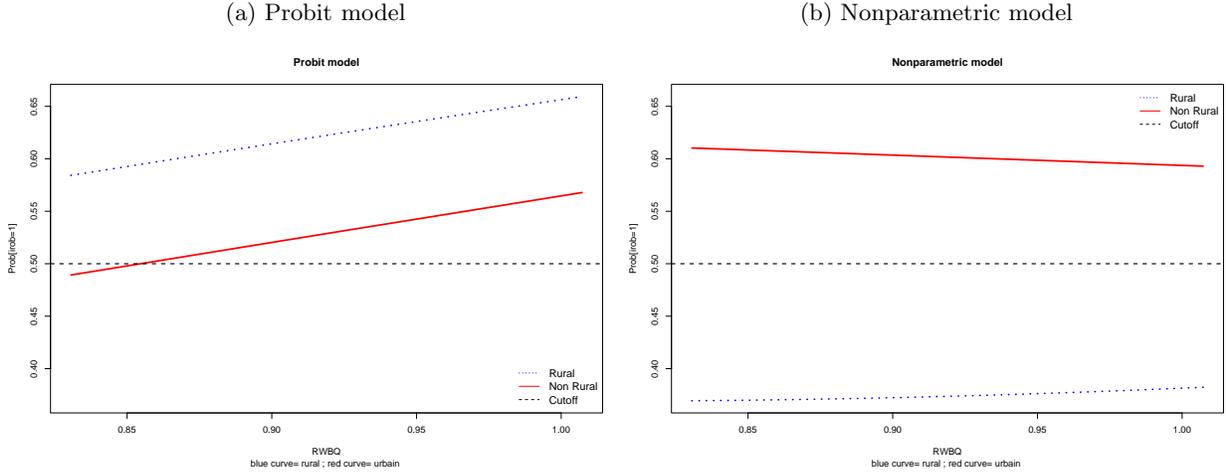


Figure 7: Rural effect



of the estimated probability. The corresponding effect in the nonparametric setting is less clear. Figures 7 and 8 illustrate that feature when changing the respective values of rural and Region.⁶ The rural effect is more pronounced for the nonparametric specification than for the Probit one. Here again, the rural variable clearly discriminates between tap water drinkers and non drinkers for the nonparametric model when using a cut-off value of 0.5. As shown by Bontemps & Nauges (2009), regional effects are important in France due to various cultural habits.

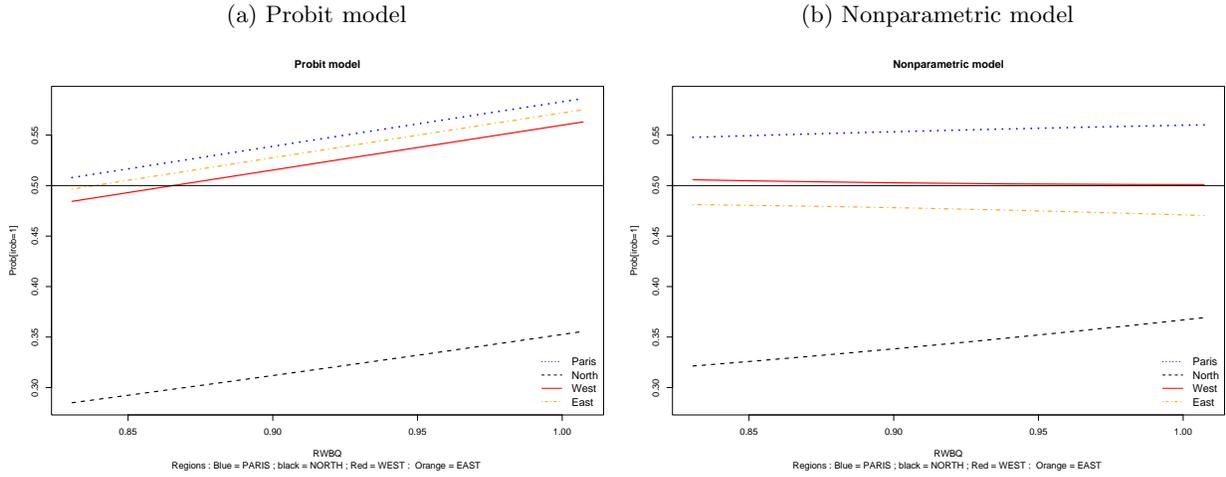
The Probit model shows a less contrasted picture of these regional effects than the nonparametric one. The parametric probabilities are all increasing with the RWBQ, that is when the quality of the environment deteriorates, while it is not the case for all regions for the nonparametric specification. Moreover, two clusters of regions seem to appear in Figure 8a: North *vs* East, West, and Paris, while three appear in Figure 8b: North *vs* East, West *vs* Paris. With a cut-off value of 0.5, people in the North never drink tap water whatever the chosen specification, confirming well-known habits in this region. On the opposite, people in Paris are more likely to drink tap water, this finding may be partly due to water bottle storage problems in flats. The results for the West and East regions are more plausible for the nonparametric specification than for the Probit, people being indifferent between drinking or not tap water in these two regions for the former specification.

4 Concluding remarks

We implement here an estimation of the probability of drinking tap water using a nonparametric kernel conditional probability specification based on the conditional PDF of the decision variable. This specification uses the same information set than a probit specification performed in Bontemps & Nauges (2009). We use recent developments on generalized kernel estimation to deal with both

⁶In the analysis, we fix the variables diploma, income, deleg, and retired at chosen values (resp. for household with diploma lower than *Baccalauréat*, with an income at the 9th quantile, no delegation for tap water distribution, and for retired people). Unless specified, we choose to work with the North region (where the smallest proportion of tap-water drinkers is observed) and in non rural areas.

Figure 8: Regional effect



continuous and discrete variables (see Li & Racine (2007)). The setup chosen here enables to assess the relevance of the explanatory variables through the data-driven bandwidth selection method. We find a similar set of significant or relevant variables, the raw water bad quality index being relevant as in Bontemps & Nauges (2009). We find close in-sample predictive performances, that is the overall correct classification ratio is 70.04% when considering the nonparametric specification while it is of 69.17% for the probit specification. We then address the issue of selecting a preferred model specification between the fully nonparametric and the probit models. We take the view that fitted statistical models are approximations, a perspective that differs from that of consistent models selection which posits a finite dimensional ‘true model’. The RP test recently developed by Racine & Parmeter (2009) and implemented here, shows that the nonparametric specification possesses an expected ‘true error’ that is statistically significantly lower than that for the parametric specification. The nonparametric specification is therefore preferred. The results focusing on environmental insights provided by the two estimators, emphasize some patterns that can only be detected using the nonparametric estimator.

References

- Aitchison, J. & Aitken, C. G. G. (1976), ‘Multivariate binary discrimination by the kernel method’, *Biometrika* **63**, 413.
- Bontemps, C. & Nauges, C. (2009), ‘Carafe ou bouteille ? le rôle de la qualité de l’environnement dans la décision du consommateur’, *Economie et Prévision* **188**, 61–79.
- Briesch, R. A., Chintagunta, P. K. & Matzkin, R. L. (2002), ‘Semiparametric estimation of brand choice behavior’, *Journal of the American Statistical Association* **97**(460), 973–982.
- Efron, B. (1978), ‘Regression and anova with zero-one data: Measures of residual variation’, *Journal of the American Statistical Association* **73**, 113–121.
- Efron, B. (1982), *The Jackknife, the Bootstrap, and Other Resampling Plans*, Society for Industrial Mathematics.
- Egan, J. P. (1975), *Signal Detection Theory and ROC Analysis*, Vol. 195, New York : Academic Press.
- Hall, P., Racine, J. & Li, Q. (2004), ‘Cross-validation and the estimation of conditional probability densities’, *Journal of the American Statistical Association* **99**(468), 1015–1026.
- Hayfield, T. & Racine, J. S. (2008), ‘Nonparametric econometrics: The np package’, *Journal of Statistical Software* **27**(5), 1–32.
- Horowitz, J. (1998), *Semiparametric Methods in Econometrics*, Springer-Verlag, New-York.
- Li, Q. & Racine, J. S. (2007), *Nonparametric Econometrics: Theory and Practice*, Princeton University Press.
- Pagan, A. (1999), *Nonparametric Econometrics*, Cambridge University Press.
- Racine, J. S. & Parmeter, C. F. (2009), *Data-driven model evaluation: a test for revealed performance*. Mac Master University.
- Wang, M.-C. & Van Ryzin, J. (1981), ‘A class of smooth estimators for discrete distributions’, *Biometrika* **68**(1), 301–309.