Parametric *vs* nonparametric dichotomous choice contingent valuation models: testing the kernel estimator and its revealed performance

Carlos Ordas Criado* Marcella Veronesi†

January 2013

(work-in-progress)

Abstract

Binary choice contingent valuation models have been explored with a variety of flexible estimation techniques to capture the unknown shape of the conditional probability distribution and the nonlinear effects of its determinants. Misspecification tests generally reject the parametric specifications in favor of more flexible counterparts, but they do not guarantee improved out-of-sample performance for the preferred model. This paper applies a simple distribution-free and fully-flexible conditional kernel density estimator, ignored in this literature, to uncover the willingness to pay of citizens for a program that reduces the risk of early death from hazardous waste exposure. We compare its performance with two classic parametric and semiparametric logit specifications. In-sample and out-of-sample testing shows that the nonparametric estimator overwhelmingly outperforms its linear and semiparametric counterparts in terms of their predictive performance. We further provide surface plots of conditional willingness to pay and compare the willingness to pay measures across specifications.

JEL Classification numbers: C14, C35, Q51, Q53

Key Words: binary choice models, contingent valuation, nonparametric estimation, revealed performance, willingness to pay.

*Corresponding author, carlos.ordas@ecn.ulaval.ca, Centre for Research in Economics of the Environment, Food, Transport and Energy (CREATE), Department of Economics, Laval University, Canada.

†marcella.veronesi@env.ethz.ch, Institute for Environmental Decisions (PEPE), ETH Zurich, Switzerland.

1 Introduction

Contingent valuation is a widely used method for placing a value on changes in the provision of goods that are not traded in markets. A frequently used approach to eliciting information about households' willingness to pay for these changes is the so-called dichotomous choice contingent valuation (DCCV) format (Bishop and Heberlein (1979); Hanemann (1984))[1]. Respondents are asked whether or not they are willing to pay $X for the change, or whether they would vote in favor or against a proposed public program in a referendum on a ballot if implementing it costs $X to the household, for example in the form of higher taxes or higher prices of products. The dollar amount $X is varied across respondents, and is usually termed the bid value. When dichotomous choice questions are used, the researcher does not observe the respondent's exact willingness to pay (WTP) amount, and at best the researcher can infer if the respondent's WTP amount is greater than the bid amount (if the respondent is in favor of the program) or less than the bid amount (if the respondent votes against the plan), and form broad intervals around the respondent's WTP.

To estimate the usual welfare measures such as the mean and the median WTP, researchers usually apply parametric models such as binary data models. For example, the simplest such models assumes that an individual's response to the WTP question is motivated by an underlying, and unobserved, WTP amount, which is normally (or logistically) distributed. However, over the last decades researchers have started to argue that parametric methods may lead to biased estimates of the welfare measures because they rely on assumptions of the functional form and of the distribution of the WTP that cannot be known a priory (Kristrom (1990); Li (1996b), Chen and Randall (1997); Creel and Loomis (1997), Cooper (2002), Crooker and Herriges (2004), Huang et al. (2008), Watanabe and Asano (2009)).

In response to this argument, nonparametric and semiparametric methods started to receive more attention. A popular nonparametric estimator used to estimate the WTP is the Turnbull distribution-free estimator (Turnbull (1976); Cosslett (1983)), originally applied in contingent valuation by Carson et al. (1994) and Haab and McConnell (1997). Equivalent nonparametric models have been estimated by Kristrom (1990) and McFadden (1994). However, these nonparametric estimators provide welfare measure estimates unconditional on important socioeconomics and demographics characteristics of the respondents, or characteristics of the good to be valued[2]. For instance, if the respondents have been asked to value

[1] In a contingent valuation study the good is specified in a hypothetical scenario, and no actual transaction takes place. The dichotomous choice approach mimics behavior in regular markets, where people usually purchase, or decline to purchase, a good at the posted price. It also closely resembles people's experience with political markets and propositions on a ballot (see for example Mitchell and Carson (1989) for details on the contingent valuation method).

[2] For example, the Turnbull estimator would not allow testing for scope effects, one of the recommended tests by the NOAA (National Oceanic and Atmospheric Administration) Blue

commodities of different size and quality, the scope test implies checking that WTP increases with the size and the quality of the commodity being valued.

A possible alternative consists in adopting distribution-free semiparametric models[3]. These estimators relax the functional forms and distributional assumptions of the parametric models and allow the estimation of conditional welfare measures. However, they keep some restrictive assumptions Another class of flexible estimators for binary response models are the kernel density estimators, and more particularly the one proposed by Li and Racine (2003) for datasets which involve discrete and continuous variables. The main advantage of the latter approach is that it allows all kind of nonlinearities and interactions between the continuous and discrete factors and therefore account for all kind of effects affecting the conditional density without requiring any (preliminary) parametric formulation. Simple numeric integration further permits to recover any of the DCCV pertinent welfare measures.

This paper contributes to the existing contingent valuation literature in at least two ways. First, we apply for the first time in the DCCV context a simple (distribution-free) kernel estimator, based on generalized product kernels and straightforward bandwidths. Second, we compare its performance with two classic parametric and semiparametric logit specifications by using the "revealed performance" test recently proposed by Racine and Parmeter (2009). The latter methodology has the advantage of relying on the 'true error' of the models as compared to their 'apparent error' which captures models' performance based on a single random draw from the true underlying data generation process and in-sample measures of fit. We show that the kernel estimator overwhelmingly outperforms its linear and semiparametric counterparts in terms of predictive performance. We further discuss the willingness to pay estimates obtained with the parametric and semi/nonparametric approaches for a program that reduces the risk of early death from hazardous waste exposure. Finally, we provide a series of simple plots which allow to explore (conditional) WTP patterns which are of particular relevance for the applied researchers in DCCV.

The reminder of the paper is organized as follows. Section 2 outlines the econometric methodology, starting with the parametric and nonparametric specifications and proceeding with the revealed performance testing procedure. The empirical

Ribbon Panel's contingent valuation guidelines (Arrow et al. (1993)). These conditioning factor are fundamental for policy making, in benefit transfer exercises or for testing the validity of WTP estimates. For example, in a survey eliciting WTP to reduce hazardous waste risks, the scope test requires checking that the mean WTP is increasing in the size of the risk reduction. See also Haab and McConnell (2002, p. 80-83) for other limitations related to the assessment of covariates effects in the Turnbull model. In addition, note that recently Day (2007) found that the Turnbull's estimator may not necessarily converge to the nonparametric maximum likelihood estimates.

[3]See Li (1996a), Chen and Randall (1997), Creel and Loomis (1997), An (2000), Cooper (2002), Belluzzo (2004)

section presents the data in section 3.1, the in-sample and the revealed performance results in section 3.2 and 3.3, respectively, and welfare estimates in section 3.4. We conclude in section 4.

2 Econometric methodology and revealed performance

2.1 Binary response models

Let's consider two competing logistic models, one with a set of linear explanatory variables (model (1)) defined Generalized Linear Model (GLM), and a semiparametric model (model (2)) defined Generalized Additive Model (GAM) with an additive (semiparametric) explanatory structure

$$\text{Prob}(Y|X) = \{1 + \exp(X'^c\alpha_{glm} + X'^d\beta_{glm} + \tilde{X}'^d\tilde{\beta}_{glm})\}^{-1} \tag{1}$$

$$\text{Prob}(Y|X) = \{1 + \exp(\sum_{j=1}^{p} f_j(X_p^c) + X'^d\beta_{gam} + \tilde{X}'^d\tilde{\beta}_{gam})\}^{-1} \tag{2}$$

Y is the dependent dichotomous unordered factor $Y = 0$ or 1, $X = (X^c, X^d, \tilde{X}^d)$ denotes a vector of p continuous, q discrete unordered and r discrete ordered explanatory components, and the α's and β's represent parameters associated to the related regressors. The only difference between models (1) and (2) stems from continuous components which are estimated in a flexible (but additive) way in the semiparametric specification. Both logistic models possess well-known estimation procedures such as the maximum likelihood estimator (MLE henceforth) for equation (1) and a penalized MLE for equation (2). An obvious drawback of the latter two specifications is the arbitrary assumption made on the link function (logit) and the absence of interaction between the regressors, and more particularly between the discrete factors and the continuous variables. A simple way of relaxing all parametric and additivity restrictions is to consider a fully nonparametric formulation. Let $f(.)$ and $m(.)$ be the joint and marginal densities of (X, Y) and X respectively. The density of Y conditional on X for a given realization of (X, Y), denoted by $(x, y) = (x^c, x^d, \tilde{x}^d)$, can be written as

$$g(y|x) = \frac{f(x, y)}{m(x)} \tag{3}$$

Equation 3 can be naturally estimated with the help of product kernels. Given the discrete nature of Y and $(X^d, \tilde{X}^d) \in X$, Li and Racine (2003) have proposed the use of 'generalized product kernels' which account for mixed types of variables. The unknown joint and marginal density $f(x, y)$ and $m(x)$ in equation (3) can be replaced by their kernel fits $\hat{m} = \hat{m}(x^c, x^d, \tilde{x}^d)$ and $\hat{f}(x, y) = \hat{f}(x^c, x^d, \tilde{x}^d, y^d)$ given by

$$\hat{m}(x) = n^{-1} \sum_{i=1}^{n} \prod_{j=1}^{p} W(X_{ij}^c, x_j^c) \prod_{j=1}^{q} l(X_{ij}^d, x_j^d) \prod_{j=1}^{P} \tilde{l}(\tilde{X}_{ij}^d, \tilde{x}_j^d) \quad (4)$$

$$\hat{f}(x,y) = n^{-1} \sum_{i=1}^{n} \prod_{j=1}^{p} W(X_{ij}^c, x_j^c) \prod_{j=1}^{q} l(X_{ij}^d, x_j^d) \prod_{j=1}^{P} \tilde{l}(\tilde{X}_{ij}^d, \tilde{x}_j^d) \times l(Y_i^d, y) \quad (5)$$

where

$$W(X_{ij}^c, x_j^c) = \frac{1}{b_j} K(\frac{X_{ij}^c - x_j^c}{b_j}) \quad \text{for the continuous regressor}$$

$$l(X_{ij}^d, \tilde{x}_j^d) = \begin{cases} 1-\lambda_j & \text{if } X_{ij} = x_j^d \\ \frac{\lambda_j}{c_j - 1} & \text{otherwise} \end{cases} \quad \text{for the discrete unordered factor}$$

$$\tilde{l}(\tilde{X}_{ij}^d, \tilde{x}_j^d) = \begin{cases} 1 & \text{if } \tilde{X}_{ij} = \tilde{x}_j^d \\ \gamma^{|\tilde{X}_{ij} - \tilde{x}_j^d|} & \text{otherwise} \end{cases} \text{for the discrete unordered factor}$$

The functions $W()$, $l()$ and $\tilde{l}$ deserve some comments. The kernel function $K()$ in the $W()$ term is a traditional symmetric and univariate probability density and b_j is the bandwidth. The $l()$ term is a kernel for unordered discrete variables X_{ij}^d with c_j categories proposed by Aitchison and Aitken (1976), where $\lambda_j \in [0, (c_j - 1)/c_j]$ represents the bandwidth. We notice that when λ_j is set to 0, $l()$ becomes and indicator function that drives to the "frequency" approach of discrete variables in nonparametric estimation. When $\lambda_j = (c_j - 1)/c_j$, $l() = 1/c_j$ for all (X_{ij}, x^d) pairs. Finally, the term $\tilde{l}()$ is a kernel for ordered discrete variables proposed by Wang and Van Ryzin (1981), where $\gamma_j \in [0, 1]$ stands for the bandwidth. Again we can see that $\gamma_j = 0$ transforms $\tilde{l}()$ in an indicator function while $\gamma_j = 1$ assigns uniform weights.

Estimating $\hat{g}(y|x)$ with the estimators (4) and (5) involves choosing a value for each bandwidth b_j, λ_j, and γ_j according to some optimality criterion. Hall et al. (2004) show that least squares cross-validation produces asymptotically optimal smoothing for the relevant components while eliminating irrelevant ones by oversmoothing. The main drawbacks of this approach is that it is computationally lengthy and that the multidimensional optimization process can result in undersmoothed patterns over some dimensions due to discontinuities over the support. Alternatively, one can simply employ the robust normal-reference rule-of-thumb of Silverman (1986, p. 47) for the continuous regressors and use the 'frequency' approach. We adopt the latter methodology and control for the pertinence of the marginal effects with the help of bootstrapped confidence intervals and partial regression plots.

2.2 Revealed global performance

In the presence of theoretically equivalent specifications such as those of section 2.1, the adequacy of a particular model can be checked with misspecification tests. However, these procedures do not ensure that the 'correct' specification performs better than their 'inadequate' counterparts in terms of expected performance on independent data drawn from the same underlying data generation process (DGP). By the same token, a parametric model that successfully passes a misspecification test may perform worse than alternative specifications. As contingent valuation is often carried out for assessing the pertinence of implementing some public policy measures outside the training dataset, testing the predictive performance of the competing models is of prime relevance for policy makers. Racine and Parmeter (2009) show how the 'forecast ability' of cross-sectional models can help the practitioners in discriminating between approximate models. This section provides a non-technical description of this approach for binary choice models. We redirect the reader to Racine and Parmeter (2009) for a detailed presentation in a more general context.

As mentioned in section 2.1, the data for the binary choice model consists of pairs $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$, where X_i is a $1 \times (q + p + r)$ vector of predictors and Y is a dichotomous response. Let $Z_i = (X_i, Y_i)$ represent random draws from the unknown distribution function F, so that $Z_1, Z_2, \ldots, Z_{n_1} \sim F$. We observe the *training* dataset $Z_1 = z_1, Z_1 = z_2, \ldots, Z_{n_1} = z_{n_1}$ and denote $Z^{n_1} = (Z_1, Z_2, \ldots, Z_{n_1})$ and $z^{n_1} = z_1, z_2, \ldots, z_{n_1}$. Having observed $Z^{n_1} = z^{n_1}$, we estimate a binary choice model and we use it to predict new values of the response variable

$$\hat{g}_{z^{n_1}}(x^{n_2}) \tag{6}$$

where the superscript n_2 indicates a new set of observations, $z^{n_2} = z_{n_1+1}, z_{n_1+2}, \ldots, z_n$, distinct from z^{n_1} and with $n_2 = n - n_1$. Instead of comparing some fit measure of the 'apparent' error (McFaddens' or the correctly classified ratio) of two competing models, say $\hat{g}^{gam}_{z_{n_1}}(x^{n_1})$ vs $\hat{g}^{np}_{z_{n_1}}(x^{n_1})$, we are interested in contrasting their 'true error', i.e., the error obtained when the fitted model is used to predict new draws from the underlying DGP as in equation (6). The 'revealed performance' test (RP) proposed by Racine and Parmeter (2009) aims at (i) estimating a model's true error and (ii) testing whether the true error is significantly lower for one model compared to another.

The description of the procedure requires the introduction of some notations and definitions borrowed in part from Efron (1982, p.51). The 'true error' of model $g(.)$ is given by

$$E_{n2,F}[L(Y^{n_2} - \hat{g}_{Z^{n_1}}(X^{n_2}))] \tag{7}$$

where $L(.)$ designates an arbitrary loss function, i.e., a function which measures the degree of wrongness between the estimated value and the true or desired value. Note that the expectation in 7 is taken over the new points $Z_{n_1+1}, Z_{n_1+2}, \ldots, Z_n \sim F$, independent of the training dataset $(Z_1, Z_2, \ldots, Z_{n_1})$ of $g(.)$. A realization of the 'true error' based on the observed $z^{n_2} = z_{n_1+1}, z_{n_1+2}, \ldots, z_n$ can be written as

$$\frac{1}{n_2} \sum_{i=n_1+1}^{n} L(y_i - \hat{g}_{z^{n_1}}(x_i^{n_2})). \tag{8}$$

Equation 8 represents an average prediction error which could correspond for instance to the percentage of wrongly classified predictions of the binary choice model $\hat{g}(.)$ returned by a loss function for the z^{n2} dataset. Next we define the 'expected true error' to be

$$E(E_{n_2,F}[L(.)]), \tag{9}$$

where the $E(.)$ stands for the expectation operator over all potential model surfaces $\hat{g}_{Z^{n_1}}(.)$ for the selected loss function $L()$. When comparing two models, the model with the lowest 'expected true error' will be the closest to the true DGP and therefore will be preferred on the basis of its revealed performance. If the researcher were given S splits of the data, the empirical distribution of (8) could be constructed for each model, and a simple pairwise test of means' equality could be applied to determine significant differences. Racine and Parmeter (2009) provide an algorithm to build such distributions and carry out these tests.

For the rest of this section, we follow Bontemps et al. (2009), and implement the procedure advocated by Racine and Parmeter (2009) based on two predictive indicator for binary response models: the Correctly Classified Ratio (CCR) and the Area Under the "Receiver Operating Characteristic" curve (AUROC, see Egan (1975)). Consider the loss function

$$Q(Y_i, \eta_i, \alpha) = \begin{cases} 0 & \text{if } Y_i = 1 \text{ and } \eta_i > \alpha \text{ or if } Y_i = 0 \text{ and } \eta_i \leq \alpha \\ 1 & \text{otherwise} \end{cases} \tag{10}$$

where η_i designates the probability assigned by the model to the ith observation, α is the cut-off value used that observation i into the category 0 or 1. For a given cut-off (usually $\alpha = 0.5$), the $CCR(\alpha)$ can be expressed as

$$CCR(\alpha) = 1 - \frac{\sum_{i=1}^{n} Q(Y_i, \eta_i, \alpha)}{n}. \tag{11}$$

An obvious drawback of the CCR index is its dependence on the arbitrary parameter α. The latter index is linked to the confusion matrix presented in Table

1 where ON and OP are the total numbers of observed 0 and 1 respectively, TN (TP) stands for 'true negative' ('true positive) occurring when the observed value y and its predicted outcome $\hat{y}$ coincide at the level 0 (1), *i.e.,* when $\eta_i \leq \alpha$ ($\eta_i > \alpha$), and FN (TN) for 'false negative' ('true negative') when the observed value y is 0 (1) and the predicted value $\hat{y}$ is 1 (0), *i.e.,* when $\eta_i \leq \alpha$ ($\eta_i > \alpha$). The "Receiver Operating Characteristic" curve (ROC) plots the percentage of predicted true positive (TP/OP called *specificity*) versus 1 - the percentage of predicted false negative (1- the *specificity* index TN/ON), letting the cut-off value varying between its extremes. The Area Under the ROC curve (AUROC) can be used as summary classification measure independent of any cut-off-value α. The AUROC indicator lies between 0.5 (worthless classification) and 1 (perfect classification).

Table 1: Notations for the model preference measures

Confusion matrix

		Predicted		
		0	1	Total
Obs.	0	TN	FP	ON
	1	FN	TP	OP
Total				n

Indices

Accuracy (CCR) $= \frac{TN+TP}{n}$

Sensitivity (TPR) $= \frac{TP}{OP}$

Specificity (SPC) $= \frac{TN}{ON}$

We can now describe the algorithmic procedure which allows to construct the empirical distributions of the 'true errors' based on the CCR and AUROC measures:

1. Resample *without* replacement pairwise from $Z = \{X_i, Y_i\}_{i=1}^n$ and call these $Z_\star = \{X_i^\star, Y_i^\star\}_{i=1}^n$;

2. Let the first n_1 resampled observations form the training sample $Z_\star^{n_1} = \{X_i^\star, Y_i^\star\}_{i=1}^{n_1}$ and the remaining $n_2 = n - n_1$ observations form an evaluation sample $Z_\star^{n_2} = \{X_i^\star, Y_i^\star\}_{i=n_1+1}^{n_2}$;

3. 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 logit fixed, fit each model on the training observations $Z_\star^{n_1}$, and then obtain predicted values from the evaluation points $Z_\star^{n_1}$ that were not used to fit the model.

4. Compute the CCR(0.5)/AUROC of each model.

5. Repeat this a large number of times, S = 1'000 in our case, yielding S draws of CCR/AUROC for the two models.

With these empirical distributions at hand for the competing models, two simple tests of differences in means (t-test and Mann-Whitney-Wilcoxon) allow to find the best model in terms of revealed performance.

3 Empirical analysis

3.1 Data

The dataset used in this paper comes from a mail survey carried out by Loomis and duVair (1993) over 1000 households living in California. The respondents were asked to vote on a series of public programs that would reduce the chances of early death from exposure to heavy metals. The question was "Suppose the State of California put program A (A = 1, 2 or 3) to increase funding for waste minimization on the next ballot. If it costs your household $X, would you vote for the program (yes or no)?". The survey also collects information about the relative importance that the respondent placed on hazardous waste exposure of the household as compared to other community problems, as well as other socio-demographic information (further details can be found in Loomis and duVair (1993)). The sample used in this paper has been previously explored with different binary choice specifications by Loomis and duVair (1993), Creel and Loomis (1997), and Cooper (2002).[4] The first study estimated simple linear specifications with a logit MLE while the remaining two studies looked into misspecification issues for a variety of flexible vs less flexible models. In this paper we concentrate on the specification adopted in Creel and Loomis (1997).

In addition to the acceptation or rejection of the proposed programs, we observe the following explanatory variables for the respondents:

- OTHER: an unordered discrete variable which ranks the level of risk exposure of the respondent to hazardous waste with respect to other problems in his/her community. 1 = high, 2 = medium, 3 = low, 4 = don't know;
- RISKRED: an ordered discrete variable which measures the percentage of risk reduction with respect to the existing risk of early death. Three possible levels are considered 25%; 50%, and 75%.
- EDUC: the number of years of education of the respondent (varies between 5 and 20 years in the sample);
- AGE: age of the respondent in years, treated as a continuous variable;
- INC: median of the income category in which income's respondent lies (among 12 classes), treated as a continuous variable;

[4]We thank Michale Creel and John Loomis for providing the data.

- BID: cost of the program in U.S. dollars (between 0 and 1100); treated as a continuous variable.

The final sample includes 791 observations with complete records for the variables defined above.

3.2 In-sample performance

In this section we present the estimates obtained with the three specifications, the parametric (Generalized Linear Model, GLM) and semiparametric (Generalized Additive Model, GAM) logit and the kernel fit, over the full dataset. Table 2 shows the coefficients of the linear portions for the logit models with MLE-based confidence tests. The confidence levels of the nonparametric components of the semiparametric model are also reported but their impact on the conditional probability of accepting a program of waste reduction must be checked on the partial regression plots of Figure 1, along with the nonparametric fits. Note also that Table 2 provides the bandwidths for the continuous regressors obtained with the robust version of Silverman's rule-of-thumb. Confidence intervals for the nonparametric fits can be found on the partial plots.

Note, first, that in Table 2 the coefficients for the dummies OTHER_1 to OTHER_3 in the logit specifications are decreasing, that is the respondents who feel the least concerned with heavy metal exposure are the less likely to accept a program. The magnitude of OTHER_4 lies in between of OTHER_2 and OTHER_3, which makes sense. However, only OTHER_3 appears as being significant at the 5% level. RISKRED also appears as being highly significant and positive as expected. Regarding the continuous regressors, AGE is the only determinant that does not have a significant effect on respondents' choices, even when it is entered in the logit specification as a flexible term. The shape of the nonparametric components can be better judged in Figure 1. We notice that, except AGE, the continuous components of the GAM fit display nonlinearities which are very similar to those displayed by the nonparametric fits. Moreover, most of the logit fits fall within the bootstrapped confidence intervals drawn around the nonparametric fit. The only exception comes from the variable BID for which the fully nonparametric model indicates a larger drop in the probability of voting yes at large bid levels. A striking feature of the nonparametric fits is that it is fully coherent with basic principles of economic theory: the larger the income levels, the larger the propensity for the respondent to accept the program; the larger the cost (bid) of the program, the lower the likelihood to vote yes in a potential ballot.

We now turn to the confusion matrices for the three specifications in Tables 3 and 4. As it is standard in the literature, we employ a cut-off of 0.5 to map the predicted probabilities for each observation to one of the 0/1 classes. The results clearly indicate that the linear logit underperforms in terms of CCR compared to

Table 2: Generalized linear (GLM), additive (GAM) logit and nonparametric (NP) estimates.

	GLM fit	GAM fit	NP fit bw
other	-	-	0.000
other_1	-0.791	0.085	
other_2	-0.982*	-0.200	
other_3	-2.270***	-1.475***	
other_4	-1.193	-0.458	
riskred	0.748**	1.036**	0.000
education	0.008***	s.**	1.263
age	-0.003	n.s.	7.507
income	7.e-06**	s.***	11933
bid	-0.003***	s.***	78.422

the semiparametric logit, which is in turn significantly dominated by the nonparametric model. Therefore, accounting for nonlinearities and interactions between the regressors seems to be of importance in terms of within-sample classification performance.

Table 3: Confusion matrix for $\alpha = 0.5$ with full sample ($n = 791$)

Linear logit

		Predicted 0	Predicted 1	
Obs.	0	165	184	349
	1	74	368	442
		239	552	791

Additive logit

		Predicted 0	Predicted 1	
Obs.	0	204	145	349
	1	79	363	442
		283	508	791

Nonparametric

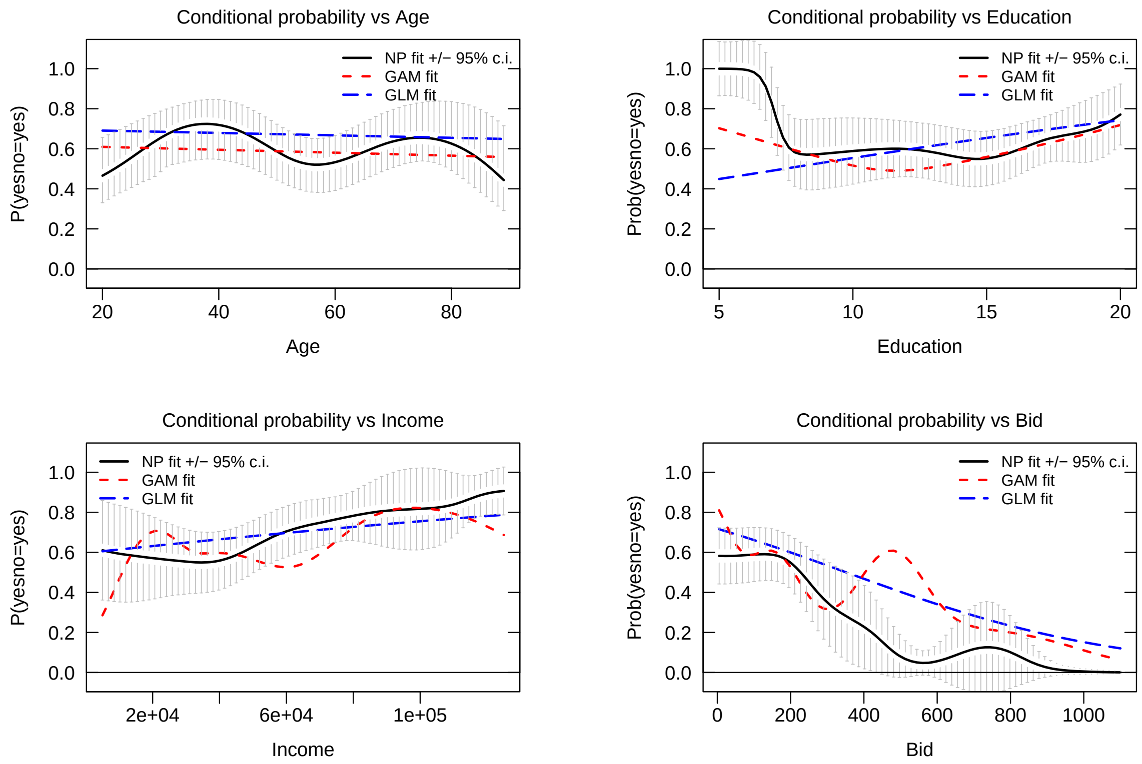
		Predicted 0	Predicted 1	
Obs.	0	270	79	349
	1	18	424	442
		288	503	791

Table 4: Global performances $\alpha = 0.5$ with full sample ($n = 791$)

Model	Sensitivity	Specificity	CCR
Linear logit (GLM)	47.28	83.26	67.38
Additive logit (GAM)	58.45	82.13	71.68
Nonparametric (NP)	77.36	95.93	87.74

Figures 2.a and 2.b report the classification performance for varying cut-off values. We notice in the former plot that the CCR index dominates the logit specifications for any level of the cut-off, and that the semiparametric logit performs

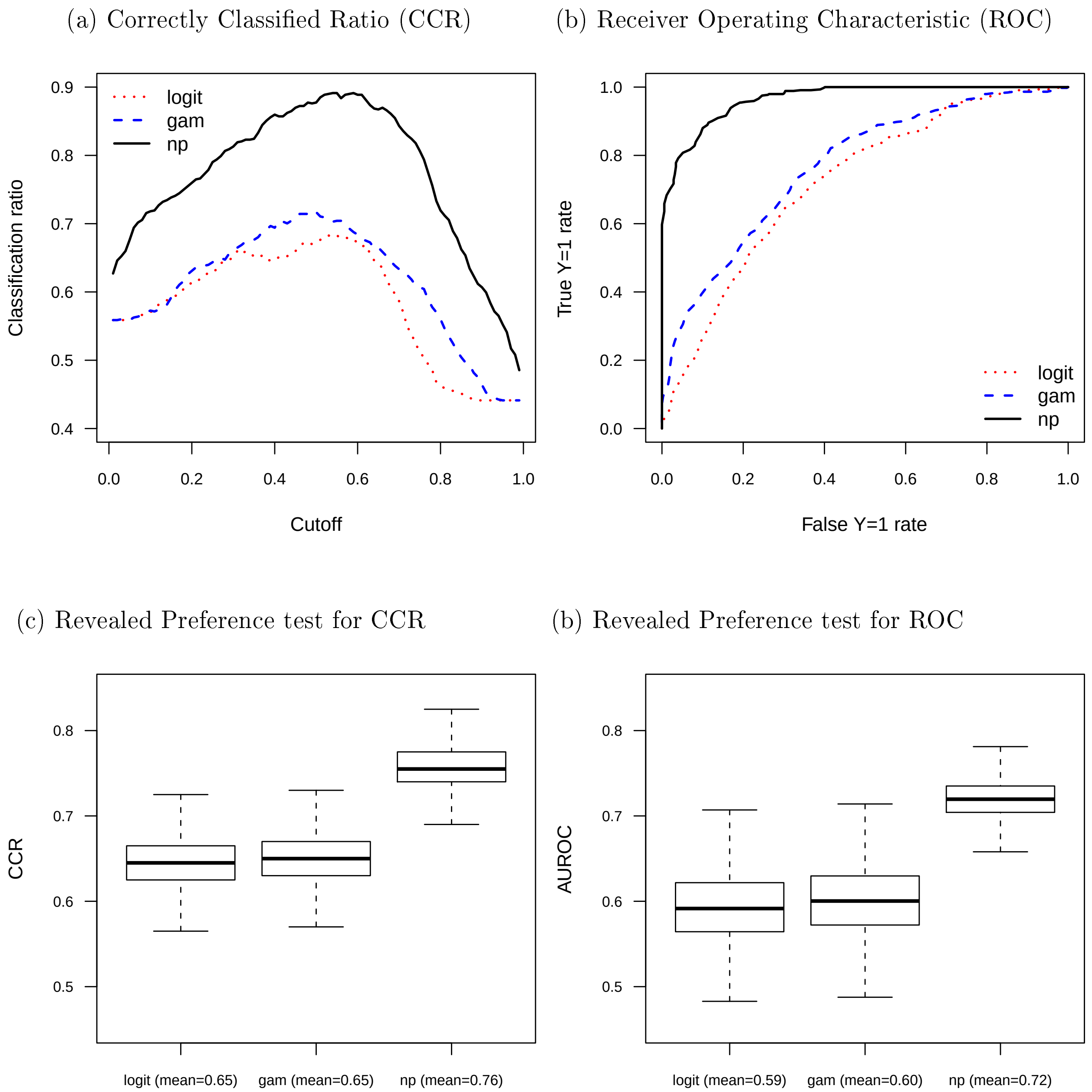
Figure 1: Nonparametric and parametric fits.



Notes: Nonparametric fits with bootstrapped confidence intervals (399 draws conducted on rows of the (y, x) data). All the computations are made using the library *np, v.0.30-3* of R.2.9.2 software, with functions *npcdens* set to *bwmethod="normal-reference"*, *uykertype="aitchisonaitken"*, *uxkertype="aitchisonaitken"*, *oxkertype="wangvanryzin"* and *npplot* set to *plot.errors.boot.method="inid"*, *plot.errors.boot.num=399*. Semiparametric fits carried out with the function *gam* from the library *mgcv, v.1.6-1*.

slightly better compared to its linear counterpart. These results are confirmed by the ROC curves plotted in Figure 2.b. The area under the ROC curve of the nonparametric model is clearly larger than that for the logit models, while we find a positive difference of 0.0253 between the semiparametric model and its linear counterpart.

Figure 2: Global performance with varying cut-off value α.



3.3 Revealed performance test

This section presents the results of the RP test described in section 2.2. We performed the test with 1000 splits of the data ($S = 1000$) into two independent samples of size $n_1 = n - n_2$, where n_2 is arbitrarily[5] set to 200. Recall that the algorithmic procedures requires estimating S times our three specifications over a random selection of n_1 observations (training sample), fitting each model on the remaining n_2 observations, computing the respective CCR and AUROC values and building their empirical distribution. Then, we test whether the expected true error is significantly *lower* for the nonparametric vs the logit specifications, as well as for the semiparametric vs the linear model. This is done with two tests: a t-test and a Mann-Whitney-Wicoxon test. The boxplot of the empirical distributions of the CCR and AUROC are shown in Figures 2.c and 2.d, respectively.

We can see in Table 5 that the nonparametric specification dominates both logit specifications in terms of expected true error and that the same occurs for the semiparametric model when compared to the linear model. We conclude that the nonparametric specification is the more likely to provide willingness to pay measures with unobserved data from the DGP under investigation. Therefore the next section concentrate on the WTP estimates of the nonparametric model.

Table 5: RP tests

Model	CCR		AUROC	
	t-test	MWW-test	t-test	MWW-test
NP < GLM	-84.9***	3946***	-79.2***	7949***
NP < GAM	-83.5***	5024***	-76.5***	8875***
GAM < GLM	-2.38***	472239**	-3.90***	449934***

3.4 Welfare estimation

The main purpose of contingent valuation is to provide welfare measures such as the expected WTP. This can be easily obtained in the context of nonparametric estimations by numerically integrating the conditional densities over the bid support, *i.e.*, $\int_0^\infty \hat{g}(y|x)\,\mathrm{d}BID$. In that respect, we follow Creel and Loomis (1997) and carry the integration over the lower/upper bounds $0/\max(BID)$. Before proceeding to evaluate the expected WTP, we explore how the conditional probability varies over the BID dimension at different levels of the discrete factors (Figure 3) as well as over the continuous covariates' ranges (Figure 4) when the rest of the explanatory variables are kept fixed (at their median).

The upper partial plot in Figure 3 shows that the area under the dotted curve OTHER_1 stochastically dominates that for OTHER_2 and OTHER_3 over most of the BID support, which points toward larger willingness to pay for the respondents who express high concern about heavy metal exposure. By contrast, the

[5]Finding the optimal split size of the data is beyond the scope of this paper but the sensitivity of the results will be tested in the final version of the paper.

probability of voting yes to a waste minimization programs for those who consider risk exposure as a minor problem (OTHER_3) becomes null when the cost of the program exceeds ∼\$300. We can also see in the lower plot that RISKRED does not seem to affect significantly the WTP of the respondents as the conditional density curves lie very close to each other whatever the level of risk reduction offered by the program.

The surface plots in Figure 4 show that the negative relationship between the probability of voting for the program and the BID variable may suffer from boundary effects at the extreme range of the continuous regressors, which violates the monotone decreasing relationship expected in that case over the whole support of the continuous regressors. This is particularly true for $AGE > 57$ and for $EDUC < 15$.

Table 6 compares the expected WTP obtained from the nonparametric approach with those derived from the logit models. We remark that the logit specifications display larger magnitudes. If we put into perspective this result and those obtained with the RP tests, we conclude that the logit models may overestimate the willingness to pay of the population under investigation. More specifically, while the expected WTP estimated for OTHER_1 and OTHER_3 across models in columns 2 and 3, the WTP at the median level of each regressor in column 1 is much lower for the nonparametric model. The latter result also holds for $RISKRED = 0.25$ and $RISKRED = 0.75$ in columns 4 and 5.

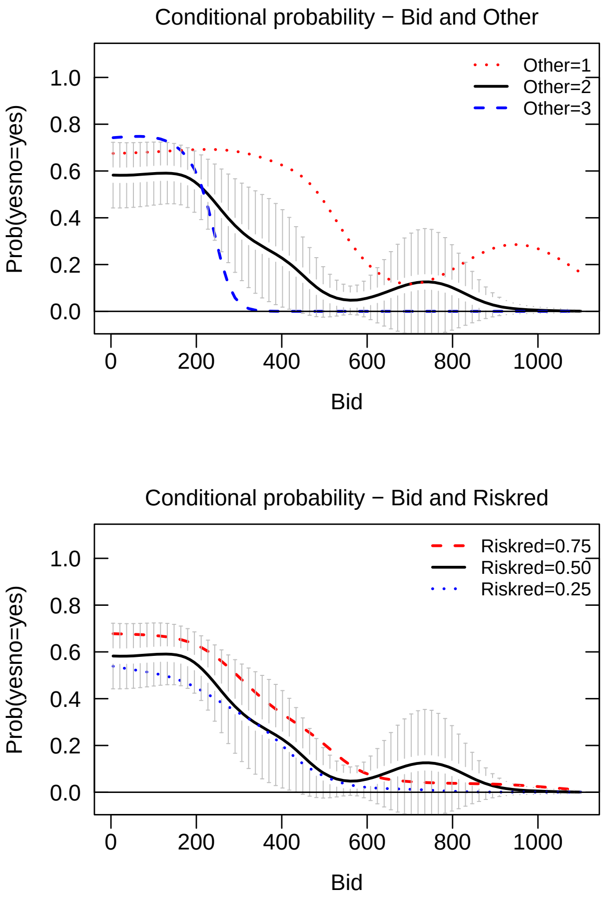
Table 6: Estimated Willingness to Pay

Model	WTP $(x^d_m\ ,\ x^c_m)$	WTP $(o = 1, x^d_m, x^c_m)$	WTP $(o = 3, x^d_m, x^c_m)$	WTP $(r = 0.25, x^d_m, x^c_m)$	WTP $(r = 0.75, x^d_m, x^c_m)$
GLM	428	472	186	387	471
GAM	400	461	175	346	455
NP	239	462	174	187	290

Finally[6], Figure 5 displays expected WTP estimates conditional on income (INC), age (AGE), and education ($EDUC$). We observe that WTP exhibits a decreasing relationship with income at low income levels ($INC <$ \$35000), increasing between \$35000-\$95000, and drops for income > \$85000.

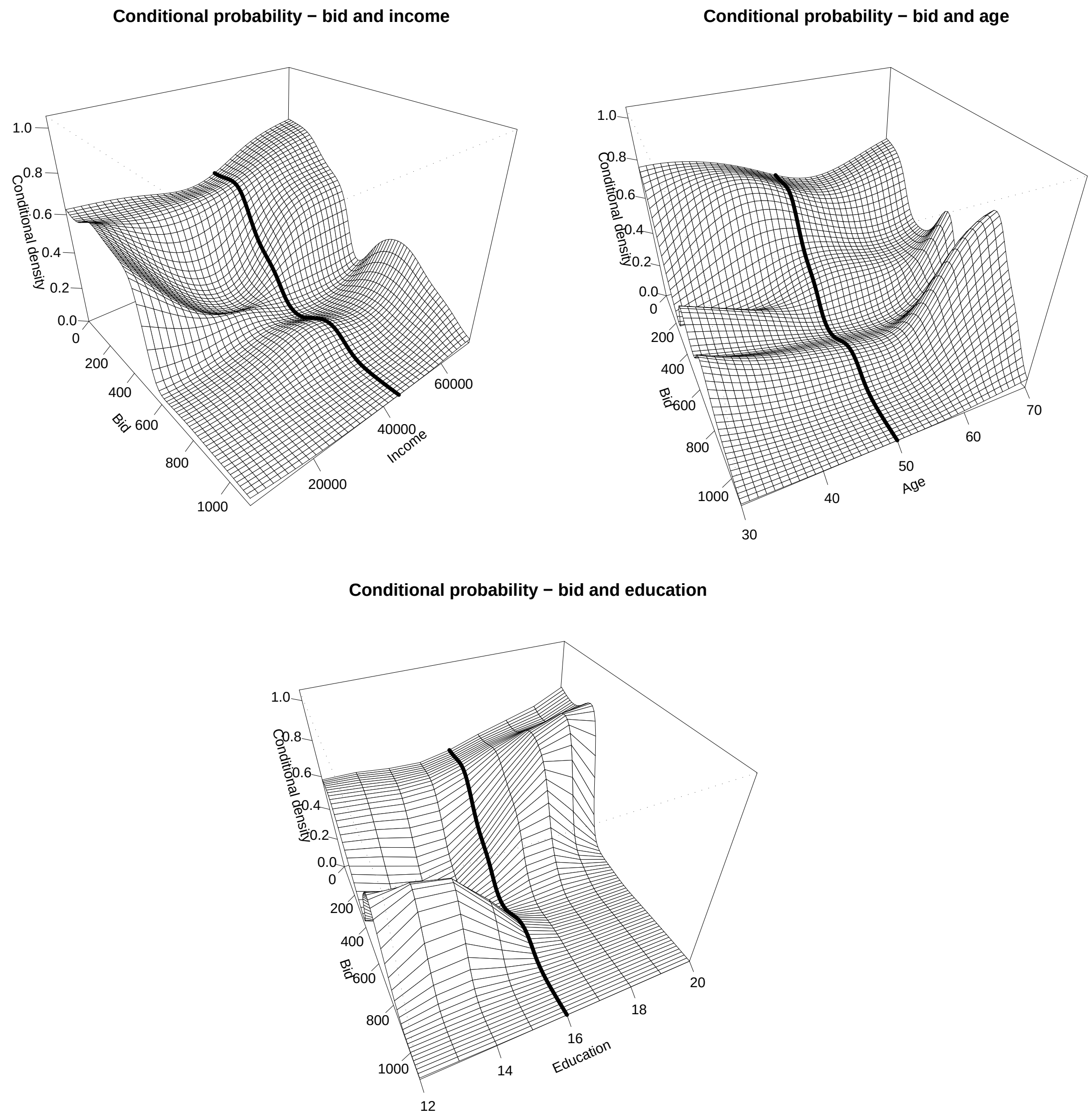
[6]Additional comments for the AGE and EDUC plots in Figure 5 will be added in the final draft of the paper.

Figure 3: Conditional kernel density estimates of accepting a program for reducing waste exposure: bids *vs* discrete factors.



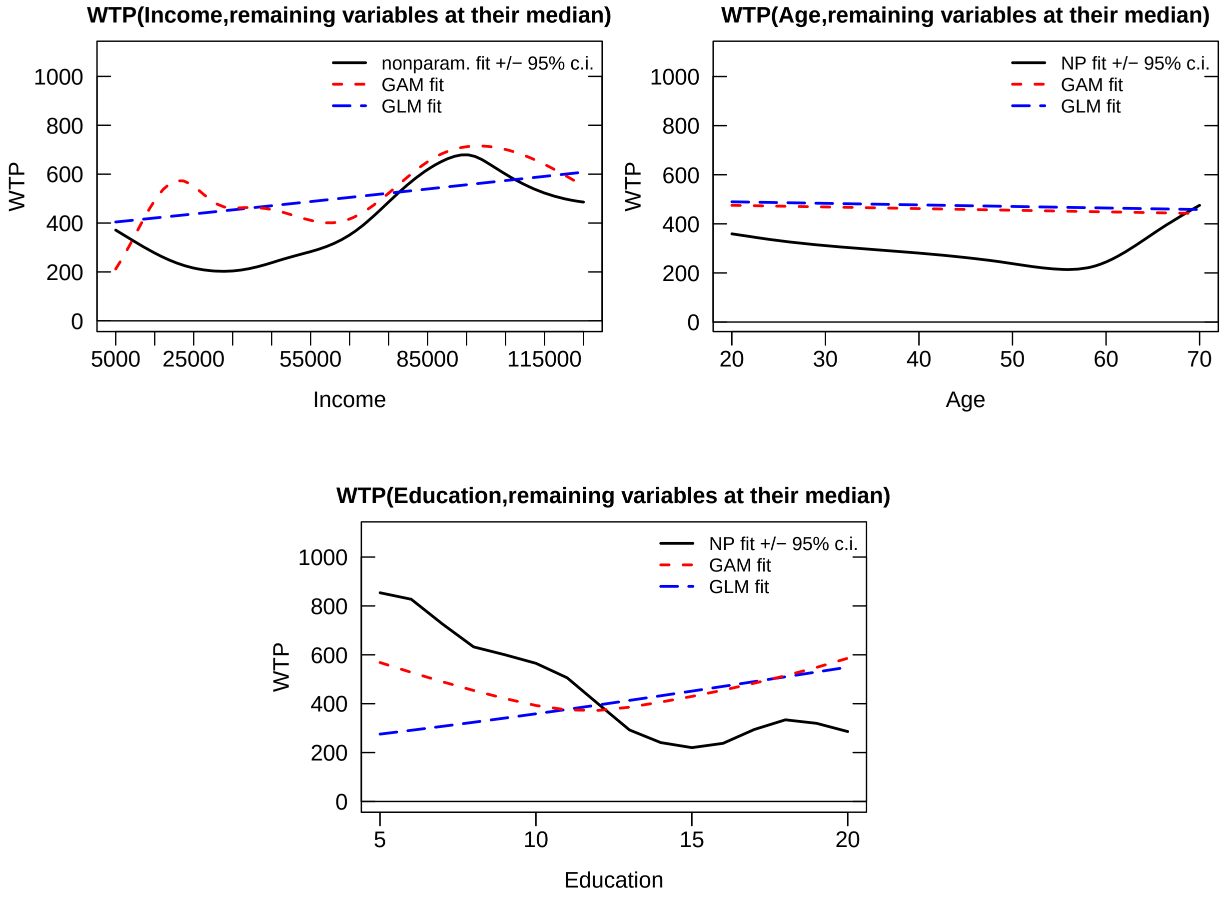
Notes: Nonparametric fits with bootstrapped confidence intervals (399 draws conducted on rows of the (y, x) data). All the computations are made using the library *np*, *v.0.30-3* of R.2.9.2 software, with functions *npcdens* set to *bwmethod="normal-reference"*, *uykertype="aitchisonaitken"*, *uxkertype="aitchisonaitken"*, *oxkertype="wangvanryzin"* and *npplot* set to *plot.errors.boot.method="inid"*, *plot.errors.boot.num=399*.

Figure 4: Conditional kernel density estimates of accepting a program for reducing waste exposure: bids *vs* continuous regressors (income, age and education).



Notes: Nonparametric fits with bootstrapped confidence intervals (399 draws conducted on rows of the (y, x) data). All the computations are made using the library *np, v.0.30-3* of R.2.9.2 software, with functions *npcdens* set to *bwmethod="normal-reference", uykertype="aitchisonaitken", uxkertype="aitchisonaitken", oxkertype="wangvanryzin"* and *npplot* set to *plot.errors.boot.method="inid", plot.errors.boot.num=399.*

Figure 5: Willingness to pay for reducing exposure to hazardous waste for different levels of income, age and education.



4 Conclusion

Dichotomous choice contingent valuation models have been explored with a variety of flexible estimation techniques to obtain consistent estimates of the willingness to pay for non-market resources and avoid arbitrary constraints on the consumers' preferences. Most of them are primarily concerned with misspecification and often conclude that the flexible estimators should be preferred. However, misspecification tests do not ensure improved predictive performance of the retained models, and measures based on the *apparent* error may simply capture overfitting to the particular sample observed.

This papers contributes to the existing literature on the topic in two main aspects. We apply for the first time a nonparametric kernel estimator which allows all kind of nonlinearities and interactions between the continuous and discrete determinants of the binary response and its associated conditional probability. The performance of this estimator is tested against two standard (parametric and semiparametric) logistic models based on the concepts of *true* error and 'revealed' performance recently introduced by Racine and Parmeter (2009). The latter procedure has the advantage of avoiding the overfitting trap which may affect any within-sample performance measure. The approach is illustrated using a contingent evaluation survey of willingness to pay for protection against premature death due to exposure to hazardous waste. We show that the kernel estimator overwhelmingly outperforms the two logit alternatives in terms of revealed performance. We provide a series of simple surface plots which allow to explore (conditional) WTP patterns in a particularly relevant way for the applied researchers in the field.

References

Aitchison, J., and C.G.G Aitken (1976) 'Multivariate binary discrimination by the kernel method.' *Biometrika* 63, 413–417

An, M.Y. (2000) 'A semiparametric ditribution for willingness to pay and statistical inference with dichotomous choice contingent valuation data.' *American Journal of Agricultural Economics* 82, 487–500

Arrow, K., Solow R., R. Radner, and H. Schuman (1993) 'Report of the noaa panel on contingent valuation.' *Federal Register* 58(10), 4601–4615

Belluzzo, W. Jr. (2004) 'Semiparametric approaches to welfare evaluations in binary response models.' *Journal of Business and Economic Statistics* 22, 322–330

Bishop, R.C., and T.A. Heberlein (1979) 'Measuring values of extramarket goods: Are indirect methods biased?' *American Journal of Agricultural Economics* 61, 926–930

Bontemps, C., J Racine, and M. Simioni (2009) 'Nonparametric vs parametric choice models: an empirical investigation.' Ongoing Working Paper, Toulouse School of Economics. Version presented at the AAEA & ACCI Joint Annual Meeting 2009, Milwaukee

Carson, R.T., L. Wilks, and D. Imber (1994) 'Valuing the preservation of australia's kakadu conservation zone.' *Oxford Economics Papers* 46, 727–749

Chen, H.Z., and A. Randall (1997) 'Semi-nonparametric estimation of binary response models with an application to natural resource valuation.' *Journal of Econometrics* 76, 323–340

Cooper, J.C. (2002) 'Flexible functional form estimation of willingness to pay using dichotomous choice data.' *Journal of Environmental Economics and Management* 43, 267–279

Cosslett, S.R. (1983) 'Distribution-free maximum likelihood estimator of the binary choice model.' *Econometrica* 51, 765–782

Creel, M., and J. Loomis (1997) 'Semi-nonparametric-free dichotomous choice contingent valuation.' *Journal of Environmental Economics and Management 32:341-358.* 32, 341–358

Crooker, J.R., and J.A. Herriges (2004) 'Parametric and semi-nonparametric estimation of willingness-to-pay in the dichotomous choice contingent valuation framework.' *Environmental and Resource Economics* 27, 451–480

Day, B. (2007) 'Distribution-free estimation with interval-censored contingent valuation data: Troubles with turnbull?' *Environmental and Resource Economics* 37(4), 777–795

Efron, B. (1982) *The Jacknife, the Bootstrap, and Other Resampling Plans* (Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania 19013)

Egan, J.P. (1975) *Signal Detection Theory and ROC analysis,* vol. 195 (Academic Press, New York)

Haab, T.C., and K.E. McConnell (1997) 'Referendum models and negative willingness to pay: Alternative solutions.' *Journal of Environmental Economics and Management* 32, 251–270

_ (2002) *Valuing Environmental and Natural Resources. The Econometrics of Non-Market Valuation* (Edward Elgar Publishing, Cheltenham, UK)

Hall, P., J.S. Racine, and Q. Li (2004) 'Cross-validation and the estimation of conditional probability densities.' *Journal of the American Statistical Association* 99, 1015–1026

Hanemann, W.M. (1984) 'Welfare evaluation in contingent valuation experiments with discrete responses.' *American Journal of Agricultural Economics* 66, 332–341

Huang, J.-C., D.W. Nychka, and V.K. Smith (2008) 'Semi-parametric discrete choice measures of willingness to pay.' *Economics Letters* 101(1), 91–94

Kristrom, B. (1990) 'A non-parametric approach to the estimation of welfare measures in discrete response valuation studies.' *Land Economics* 66(2), 135–139

Li, C-Z. (1996a) 'Semiparametric estimation of the binary choice model for contingent valuation.' *Land Economics* 72(4), 462–473

Li, Q. (1996b) 'Nonparametric testing of closeness between two unknown distribution functions.' *Econometric Reviews* 15(3), 261–274

Li, Q., and J.S. Racine (2003) 'Nonparametric estimation of distributions with categorical and continuous data.' *Journal of Multivariate Analysis* 86, 266–292

Loomis, J.B., and P.H. duVair (1993) 'Evaluating the effects of alternative risk communication devices on willingness to pay: Results from a dichotomous choice contingent valuation experiment.' *Land Economics* 69, 287–298

McFadden, D. (1994) 'Contingent valuation and social choice.' *American Journal of Agricultural Economics* 76, 689–708

Mitchell, R.C., and R.T. Carson (1989) *Using Surveys to Value Public Goods: The Contingent Valuation Method* (Resource for the Future, The Johns Hopkins University Press.)

Racine, J.S., and C.F. Parmeter (2009) 'Data-driven model evaluation : a test for revealed performance.' Working paper, McMaster University

Silverman, B.W. (1986) *Density estimation for statistics and data analysis* (London: Chapman and Hall)

Turnbull, B. (1976) 'The empirical distribution function with arbitrary grouped, censored, and truncated data.' *Journal of the Royal Statistical Society* 38B, 290–295

Wang, M.-C., and J. Van Ryzin (1981) 'A class of smooth estimators for discrete distributions.' *Biometrika* 68(1), 301–309

Watanabe, M., and K Asano (2009) 'Distribution free consistent estimation of mean wtp in dichotomous choice contingent valuation.' *Environmental and Resource Economics* 44, 1–10