Econometrics Journal (2006), volume **9**, pp. 511–540.
doi: 10.1111/j.1368-423X.2006.00196.x

Non-parametric regression for binary dependent variables

MARKUS FRÖLICH

Department of Economics, University of St.Gallen, Bodanstrasse 8, SIAW, 9000 St. Gallen, Switzerland;
E-mail: markus.froelich@unisg.ch

Received: August 2002

Summary Finite-sample properties of non-parametric regression for binary dependent variables are analyzed. Non parametric regression is generally considered as highly variable in small samples when the number of regressors is large. In binary choice models, however, it may be more reliable since its variance is bounded. The precision in estimating conditional means as well as marginal effects is investigated in settings with many explanatory variables (14 regressors) and small sample sizes (250 or 500 observations). The Klein–Spady estimator, Nadaraya–Watson regression and local linear regression often perform poorly in the simulations. Local likelihood logit regression, on the other hand, is 25 to 55% more precise than parametric regression in the Monte Carlo simulations. In an application to female labour supply, local logit finds heterogeneity in the effects of children on employment that is not detected by parametric or semiparametric estimation. (The semiparametric estimator actually leads to rather similar results as the parametric estimator.)

Keywords: *Binary choice, Local parametric regression, Local model, Heterogeneous response, Heterogeneous treatment effect.*

1. INTRODUCTION

In this paper, non-parametric regression for binary dependent variables in finite-samples is analyzed. Binary choice models are of great importance in many economic applications, but non-parametric regression has received relatively little attention so far. Let $Y \in \{0, 1\}$ be a binary outcome variable and X a vector of covariates. Often we are interested in estimating the conditional mean $E[Y|X = x]$ and/or the marginal effects $E[Y|X = x + \Delta x] - E[Y|X = x]$.

Usually, parametric regression models such as maximum likelihood probit or logit are used, which however entail restrictive functional form assumptions. Semiparametric binary choice estimators, such as the Klein and Spady (1993) estimator, relax these restrictions, but still imply assumptions that can be restrictive in empirical applications. The single index restriction, in particular, effectively reduces the heterogeneity in the X characteristics to a single dimension. With the recent emergence of the treatment evaluation literature, however, heterogeneity of treatment effects often has become of interest in itself (see e.g. Heckman *et al.* 1997). For example, in the analysis of the effects of children on female labour supply, the marginal effect of an additional child on the employment probability is likely to depend also on other characteristics. Whereas the effect is usually negative for most women, it might also be positive for some because of increased financial needs (e.g. housing) due to a larger family (particularly if the children are older). If women react differently on the number of children, policy instruments such as subsidized child care, all-day schooling or tax incentives should be targeted more precisely; particularly if the subpopulation of women who increase their labour supply in response to an additional child can be identified and distinguished from those women who reduce their labour supply. Such heterogeneity in the effects on the employment probability can be of substantial interest in many applications, and the estimation model should be sufficiently flexible to not restrict such kind of effect heterogeneity from the outset.[1]

Fully non-parametric regression allows for this flexibility, but is rarely used for the estimation of binary choice applications. A reason might be that the prototypical application of non-parametric regression, which is local linear regression on a low dimensional vector of covariates, is not so well suited for binary choice models. On the one hand, linear probability models often perform poorly in binary choice settings compared to non-linear models such as probit or logit (see e.g. Hyslop 1999). Local non-linear estimation, such as local likelihood logit, might therefore be better suited for binary dependent variables than local linear regression. In addition, local likelihood logit encompasses the parametric logit model for a bandwidth value of infinity.

On the other hand, in many empirical applications one often wants to include a rather large number of covariates.[2] Nonparametric regression in higher dimensions, however, is regarded as highly unreliable due to not only the curse of dimensionality but also small sample variance problems. For example, Seifert and Gasser (1996) show that local linear regression has a very high variance when the data are sparse or clustered.[3] Although the curse of dimensionality does not disappear with binary dependent variables, the finite-sample variance problems are ameliorated because of the boundedness of Y.

The purpose of this paper is to examine the finite-sample performance of local likelihood logit for binary dependent variables with many regressors (relative to the number of observations), of which some are continuous and some are discrete. Local likelihood logit is compared to parametric logit regression, the Klein and Spady (1993) estimator and to local linear and Nadaraya–Watson regression. Whereas Klein–Spady, Nadaraya–Watson and local linear regression perform rather poorly in the Monte Carlo simulations, local logit is often more precise than parametric logit. Even when the logit model is globally true, local likelihood logit does not perform much worse than parametric logit, because in these cases larger bandwidth values are chosen by the cross-validation bandwidth selector. Precision gains are largest for the estimation of the conditional expectations $P(Y = 1|X)$, and somewhat smaller for the estimation of marginal effects.

Local likelihood logit is then applied to analyze the dependence of Portuguese women's labour supply on the number of children. While the parametric logit and the semiparametric Klein–Spady estimator lead to rather similar estimates, local likelihood logit unveils heterogeneity in the marginal effects of children that is unnoticed by the parametric logit estimator. Although the Klein–Spady estimator also detects some heterogeneity in the effects, it seems to be an artefact

[1] Further examples where effect heterogeneity is of interest include the returns to schooling or the effects of training programmes, which can be used for designing optimal treatment rules, (see Manski 2000, 2004).

[2] For identifying causal effects, usually all covariates that affect the outcome variable *and* the treatment variable have to be included, which are often rather many, (see Rubin 1974; Holland 1986 Pearl 2000).

[3] Because the denominator of the estimator can be arbitrarily close to zero. This happens often even at bandwidth values that are only slightly below the optimal value. Due to this, the unconditional finite-sample variance of local linear regression is infinite and the conditional variance is unbounded.

of its larger variability.[4] In Section 2, the local logit estimator is introduced. Section 3 provides the simulation study. Section 4 analyzes female labour supply, and Section 5 concludes.

2. NONPARAMETRIC REGRESSION FOR BINARY DEPENDENT VARIABLES

Let $Y \in \{0, 1\}$ be a binary outcome variable and $X \in \Re^{Q+1}$ a vector of covariates, where for convenience of notation it is supposed that the last element of X is a *constant*. We are interested in estimating the conditional mean $E[Y|X = x]$ and the marginal effects $E[Y|X = x + \Delta x] - E[Y|X = x]$ for particular changes Δx in the covariates. For continuous regressors the marginal effect is often defined as $\partial E[Y|X = x]/\partial x_q$.[5] The standard approach proceeds by specifying a parametric model, for example, a probit or logit model, estimating the coefficients by maximum likelihood and computing the conditional means and marginal effects. The disadvantage of parametric estimation is its reliance on functional form assumptions, which lead to inconsistent estimates if the model is not correctly specified.

Several semiparametric estimators have been suggested to relax these assumptions. A *single index restriction* is often invoked which assumes that the conditional mean can be specified as $E[Y|X = x] = m(x'\theta)$ with m an *unknown* function and θ an unknown coefficient vector. A number of $\sqrt{n}$ consistent estimators of θ have been developed, including iterative methods such as Han (1987), Ichimura (1993), Klein and Spady (1993) and Sherman (1993) and non-iterative methods such as the average derivative estimators of Härdle and Stoker (1989), Powell *et al.* (1989), Stoker (1991) and Horowitz and Härdle (1996). Although θ can often be estimated at $\sqrt{n}$-rate, the function $m(x'\theta)$ is non-parametric, and estimation of $E[Y|X]$ and marginal effects will be at a lower rate. The single index specification permits estimation at the univariate rate and thereby avoids the curse of dimensionality.

Although less restrictive than the parametric models, the single index specification still implies that all individuals can be aligned on a single dimension and restricts interactions between the regressors, which is less appealing for many econometric applications where heterogeneity in responses (for example, treatment effects) is often considered as being important. For example, if Y is female labour supply and one covariate in X represents the number of children, the single index restriction imposes that the labour supply response of, for example, one versus zero children is identical for all women for whom the linear combination $x'\theta$ has the same value, even if they have very different characteristics. For these women, also the effect of five versus two children is supposed to be the same. In particular, the single-index restriction implies that any cross-effects are independent of other characteristics. Hence, although the marginal effects $\partial E[Y|X = x]/\partial x_1$ can vary with x, the ratio of two marginal effects $\frac{\partial E[Y|X=x]/\partial x_1}{\partial E[Y|X=x]/\partial x_2} = \frac{\theta_1}{\theta_2}$ is not permitted to depend on x. This is like a treatment–effect–homogeneity assumption and supposes that, for example, the labour supply response to the number of children is identical to the labour supply response to a change in marginal tax rates (multiplied by the constant θ_1/θ_2) for *every* woman.

[4] Gozalo and Linton (2000) apply least-squares local probit estimation to transport mode choice and find that parametric probit regression misses some important regressor interaction effects, which are detected by local probit. This is also found in the analysis of Portuguese female labour supply, where the estimated effects are examined instead of the coefficients, since the latter are of no direct interest. I also examine the performance of the Klein–Spady estimator, which also misses the structure.

[5] However, whereas $E[Y|X = x + \Delta x] - E[Y|X = x]$ is bounded, $\partial E[Y|X = x]/\partial x_q$ may not be.

Many of the other well-known semiparametric approaches also restrict interactions and response heterogeneity in that $\frac{\partial E[Y|X=x]/\partial x_1}{\partial E[Y|X=x]/\partial x_2}$ is not permitted to depend on any other covariates than x_1 and x_2. This includes the *generalized additive model* (GAM), the *partial linear model* and the *generalized partial linear model*.[6,7]

Nonparametric regression is more flexible. Although it is subject to the curse of dimensionality and does not achieve $\sqrt{n}$ convergence, it may still perform well in finite samples.[8] Local polynomial regression is the most popular class of estimators, (see e.g. Fan and Gijbels 1996). Apart from Nadaraya–Watson (=local constant) regression, however, local polynomial regression is not particularly suited for binary choice models as it does not incorporate the restriction that $E[Y|X] \in [0, 1]$. An immediate solution is to cap the estimates at 0 and at 1, which however makes the objective function non-differentiable and also implies that estimated marginal effects may be exactly zero at many x values.

Instead of local polynomials, other local models may be more appropriate. Let $g(x, \theta_x)$ be a *known* function with unknown coefficient vector θ_x. The conditional mean function can be modelled locally as

$$E[Y|X = x] = g(x, \theta_x). \tag{1}$$

In contrast to the parametric and semiparametric models, the coefficient vector θ_x is allowed to vary arbitrarily with x. Local parametric modelling includes Nadaraya–Watson (local constant) kernel regression with $g(x, \theta_x) = \theta_x$ and local linear regression with $g(x, \theta_x) = x'\theta_x$. For binary choice models, the logit specification

$$E[Y|X = x] \doteq \frac{1}{1 + \mathrm{e}^{-x'\theta_x}} \tag{2}$$

[6] The *generalized additive model* (GAM) assumes that $m(x)$ with $\dim(x) = K$ can be written as a sum of unknown non-parametric functions of each component of x with a known link function G, e.g. the logit link,

$$G\,(E[Y|X = x]) = \alpha + m_1(x_1) + m_2(x_2) + \cdots + m_K(x_K),$$

where the functions m_k are unknown, (see Hastie and Tibshirani 1990). The *partial linear model* separates the components of x into two sets x_{S_1} and x_{S_2} and assumes that x_{S_1} enters linearly whereas x_{S_2} can enter non-parametrically through an unknown function m:

$$E[Y|X = x] = x_{S_1}\beta + m(x_{S_2})$$

(see e.g. Robinson 1998). The *generalized partial linear model* extends this model with a *known* link function

$$G\,(E[Y|X = x]) = x_{S_1}\beta + m(x_{S_2}).$$

In the last two models, usually the set x_{S_2} contains only one regressor, which thus eliminates response heterogeneity. If x_{S_2} contains several regressors often a *generalized partial linear index model* is used:

$$G\,(E[Y|X = x]) = x_{S_1}\beta + m(x'_{S_2}\theta).$$

(see e.g. Pagan and Ullah 1999; Härdle *et al.* 2004, Fan and Gijbels 1996).

[7] More flexible are semiparametric models with multiple indices: $E[Y|X = x] = \alpha + \sum_{l=1}^{q} m_l(x'\theta_l)$, which includes projection pursuit (Friedman and Stuetzle 1981) and neural networks (White 1989; Kuan and White 1994).

[8] In addition, the non-parametric regression estimates are often used as plug in estimates in some semiparametric estimation problem of the type discussed in Newey (1994) and Chen *et al.* (2003). Examples are partial means, average treatment effects or average derivative estimators, (see e.g. Blundell and Powell 2004; Frölich 2004, 2005, 2006; Hahn 1998; Heckman *et al.* 1998; Hirano *et al.* 2003; Imbens 2004). Here, $\sqrt{n}$-consistency can be achieved despite the low precision in the non-parametric plug-in estimator, and the approaches with semiparametric structure mentioned above become less attractive.

is convenient, since it imposes the range restriction and is differentiable.[9] If the logit form is closer to the true regression curve than a constant or linear specification, a non-parametric estimator incorporating this local model will be less biased than kernel or local linear regression. Local logit encompasses the global logit model (where θ_x does not vary with x) and if the global logit model were indeed correct, local logit estimation would be unbiased.

Several approaches to estimate $\hat{\theta}_x$ have been suggested. Local non-linear least-squares regression Gozalo and Linton (2000) estimates $\hat{\theta}_x$ from a sample of n i.i.d. observations $\{(Y_i, X_i)\}_{i=1}^n$ as

$$\hat{\theta}_x = \arg\min_{\theta_x} \sum_{i=1}^{n} (Y_i - g(X_i, \theta_x))^2 \, K_H(X_i - x), \tag{3}$$

where $K_H(X_i - x)$ is a kernel function and H a vector of bandwidth values.

More general is local likelihood estimation which estimates $\hat{\theta}_x$ as

$$\hat{\theta}_x = \arg\max_{\theta_x} \sum_{i=1}^{n} \ln L\,(Y_i, g(X_i, \theta_x))\, K_H(X_i - x), \tag{4}$$

where $\ln L(Y_i,\ g(X_i,\ \theta_x))$ is the log-likelihood contribution of observation $(Y_i,\ X_i)$. For H converging to infinity, the local neighbourhood widens and the local estimator would converge to the global parametric estimator.

Local likelihood estimation includes local least squares as a special case when the likelihood function for normal errors is used. Local likelihood estimation has been introduced by Tibshirani and Hastie (1987), and its properties have been analyzed in Fan *et al.* (1995), Fan and Gijbels (1996), Fan *et al.* (1998) and Eguchi *et al.* (2003), among others.[10] Local likelihood estimation has been used for density and hazard estimation,[11] but only very rarely has it been applied to estimating regression functions for binary dependent variables. Tibshirani and Hastie (1987) and Fan *et al.* (1995) consider exemplary biometric applications of local likelihood logit, but only for one-dimensional X. Fan *et al.* (1998) present a brief simulation study and observed good performance of the local likelihood logit estimator, again only for one-dimensional X. For higher-dimensional X, Tibshirani and Hastie (1987) suggest non-parametric additive modelling for the regressors X, which, however, restricts interactions of the regressors.[12] Only in the context of local least squares, an example for non-parametric regression with a binary dependent variable Y and unrestricted interactions among the regressors X has been illustrated in Gozalo and Linton (2000).

[9] Fan *et al.* (1995) and Fan *et al.* (1998) also consider higher, order expansions in the logit domain, e.g. with a quadratic term: $E[Y|X = x] \doteq \frac{1}{1+e^{-\alpha_x - x'\beta_x - x'\gamma_x x}}$, where x does not contain a constant. A rather general result for local modelling with local polynomials entering through a link function is that odd order polynomials lead to simpler asymptotic bias expressions and with a bias of same order in the interior as in boundary regions, (see e.g. Fan *et al.* 1995; Carroll *et al.* 1998). This argument thus favours a linear or cubic expansion over a quadratic expansion. However, if x is higher dimensional, estimation with quadratic or even cubic terms could be difficult, since the number of coefficients proliferate quickly with $\dim(x)$ and local multicollinearity may occur more frequently.

[10] See also Carroll *et al.* (1998) who proposed local estimating equations.

[11] See e.g. Copas (1995), Hjort and Jones (1996), Loader (1996), Staniswalis (1989), Fan and Gijbels (1996), Eguchi and Copas (1998), Bebchuk and Betensky (2001) and Park *et al.* (2002).

[12] Blundell and Powell (2004) mentioned local likelihood probit estimation as one possible plug-in estimator for the average structural function.

Compared to local least-squares regression, the *local likelihood logit* estimator has the advantage of nesting the efficient estimator when the logit specification is true. In this case, the parametric maximum likelihood logit estimator is efficient, and for bandwidth values converging to infinity the local likelihood logit estimator converges to the parametric logit estimator. For normal errors, i.e. when assuming that ε_i is normally distributed in $Y_i = \mu(X_i) + \varepsilon_i$, local non-linear least-squares regression and local likelihood estimation are identical. But the assumption of normal errors is often not appropriate for binary dependent variables and is likely to lead to efficiency losses compared to a more sensible specification such as $Y_i = 1(\mu(X_i) + \varepsilon_i > 0)$, which leads to the local likelihood logit estimator for logistic ε_i. In addition, the local likelihood logit estimator is globally concave and usually converged much faster to the solution than local non-linear least squares.[13]

2.1. *Local likelihood logit estimation*

The *local likelihood logit* estimator is $\hat{E}[Y|X = x] = \frac{1}{1+e^{-x'\hat{\theta}_x}}$,[14] where

$$\hat{\theta}_x = \arg\max_{\theta_x} \sum_{i=1}^{n} \left(Y_i \ln\left(\frac{1}{1+e^{-X_i'\theta_x}}\right) + (1 - Y_i)\ln\left(\frac{1}{1+e^{X_i'\theta_x}}\right)\right) K_H(X_i - x). \quad (5)$$

In many empirical applications, X may contain continuous as well as discrete variables. In principle, discrete variables could be accommodated by forming separate cells for each combination of the values of the discrete regressors and conducting separate regressions within each cell. However, more precise estimates can be obtained by smoothing also over the discrete regressors. Discrete regressors can easily be incorporated in the local model $g(\cdot)$. For including discrete regressors also in the distance metric of the kernel function $K(X_i - x)$, Racine and Li (2004) suggested a hybrid product kernel that coalesces continuous and discrete regressors. They distinguish three types of regressors: continuous, discrete with natural ordering (number of children) and discrete without natural ordering (bus, train, car). Suppose that the variables in X are arranged such that the first q_1 regressors are continuous, the regressors $q_1 + 1, \ldots, q_2$ are discrete with natural ordering and the remaining $Q - q_2$ regressors are discrete without natural ordering. Then the kernel weights $K(X_i - x)$ are computed

[13] As an alternative to local models, it is often suggested to include a sufficient number of interaction terms in a global parametric model. Although this might be a convenient approach in practice, some problems should be noted. If the number of covariates is large, the number of interaction terms can quickly exceed the number of observations. Even if all Q covariates are binary, 2^Q different interaction terms can be formed. The estimates of such 'saturated models' can be very imprecise because no smoothing over the covariates takes place, see e.g. Racine and Li (2004). In binary choice models estimated by maximum likelihood, several of the interaction terms might predict the outcome perfectly, thus leading to numerical problems and undefined estimates. Although fully interacted models are often problematic in practice, a careful data-driven procedure to select from the many possible interaction terms might lead to similar results as a non-parametric approach. This, however, is beyond the scope of this paper.

[14] With $g(x, \theta_x)$ as the local model, the conditional mean is estimated as $\hat{E}[Y|X = x] = g(x, \hat{\theta}_x)$. Marginal effects can be estimated either by estimating two conditional means $\hat{E}[Y|X = x + \Delta x] - \hat{E}[Y|X = x] = g(x + \Delta x, \hat{\theta}_{x+\Delta x}) - g(x, \hat{\theta}_x)$ or from within the model as $g(x + \Delta x, \hat{\theta}_x) - g(x, \hat{\theta}_x)$.

as

$$K_{h,\delta,\lambda}(X_i - x) = \prod_{q=1}^{q_1} \kappa\left(\frac{X_{q,i} - x_q}{h}\right) \prod_{q=q_1+1}^{q_2} \delta^{|X_{q,i}-x_q|} \prod_{q=q_2+1}^{Q} \lambda^{1(X_{q,i} \neq x_q)}, \tag{6}$$

where $X_{q,i}$ and x_q denote the qth element of X_i and x, respectively. $1(\cdot)$ is the indicator function. κ is a symmetric *univariate* kernel function. h, δ and λ are bandwidth parameters with $0 \leq \delta, \lambda \leq 1$. This kernel function $K_{h,\delta,\lambda}(X_i - x)$ measures the distance between X_i and x through three components: The first term is the standard product kernel for continuous regressors. The second term measures the distance between the ordered discrete regressors and assigns geometrically declining weights. The third term measures the mismatch between the unordered discrete regressors. δ controls the amount of smoothing for the ordered and λ for the unordered discrete regressors. The larger δ and/or λ the more smoothing takes place with respect to the discrete regressors. If δ and λ are both 1, the discrete regressors would not affect the kernel weights and the non-parametric estimator would 'smooth globally' over the discrete regressors. On the other hand, if δ and λ are both zero, smoothing would proceed only within each of the cells defined by the discrete regressors but not between them.

Principally, instead of using only three bandwidth values h, δ, λ for all regressors, a different bandwidth could be employed for each regressor. This would substantially increase the computational burden for bandwidth selection and might lead to additional noise due to estimating these bandwidth parameters. Alternatively, groups of similar regressors could be formed, with each group assigned a separate bandwidth parameter. Particularly if the ranges assumed by the ordered discrete variables vary considerably, those variables that take on many different values should be separated from those with only few values. Moreover, the continuous regressors should be scaled to same mean and same standard deviation to adjust for different scopes and measurement scales and to improve numerical stability.

2.2. Bandwidth choice

A large number of alternative bandwidth selection methods for non-parametric regression have been developed. For many plug-in methods, experience with their small sample behaviour for higher-dimensional x is limited, though. In addition, with local parametric regression, the appropriate choice of the bandwidth parameters h, δ and λ depends also on how well the specified function g resembles the true conditional mean function. If the parametric hyperplane encompasses the true conditional mean function, the optimal bandwidth values would be $(h, \delta, \lambda) = (\infty, 1, 1)$, corresponding to (global) parametric regression. Otherwise the bandwidths should converge to zero with increasing sample size.[15]

[15] Eguchi *et al.* (2003) analyze optimal bandwidth choice for local likelihood estimation under small h asymptotics and large h asymptotics. If the local model is in an α-neighbourhood of the true conditional expectation function, the optimal bandwidth tends to infinity as n increases to infinity. If the local model is far from the true conditional expectation function, the optimal bandwidth tends to zero as n increases. For some intermediate cases, the optimal bandwidth may be constant. They argue that standard plug-in methods for bandwidth choice, which are based on small h asymptotics, are not appropriate if it is a priori unknown how far the local model is from the true expectation function. On the other hand, they conjecture that cross-validation and bootstrap methods should be consistent in both situations, i.e. when the optimal bandwidth tends to infinity and when it tends to zero.

Bandwidth choice by *cross-validation* permits this ambiguity, i.e. if it is not a priori known how well the local parametric model fits the data. Cross-validation selects the bandwidths to minimize out-of-sample prediction error. For minimizing squared prediction error, the bandwidths are chosen to minimize the least-squares criterion CV_{LS}

$$CV_{\mathrm{LS}} = \sum_{i=1}^{n} (Y_i - g(X_i, \hat{\theta}_{-X_i|h,\delta,\lambda}))^2, \tag{7}$$

where $\hat{\theta}_{-X_i|h,\delta,\lambda}$ is the leave-one-out coefficients estimate for the estimation of $E[Y|X = X_i]$ that is obtained from the data sample without observation i. The sum of squared errors indicates how well the estimator is able to predict $E[Y|X]$ for the sample distribution of X.

In the context of local likelihood estimation, Staniswalis (1989) suggested a different cross-validation criterion based on maximizing the leave-one-out fitted likelihood function

$$CV_{ML}(h, \lambda) = \sum_{i=1}^{n} Y_i \ln g(X_i, \hat{\theta}_{-X_i|h,\delta,\lambda}) + (1 - Y_i) \ln(1 - g(X_i, \hat{\theta}_{-X_i|h,\delta,\lambda})). \tag{8}$$

2.3. *Inference and confidence intervals*

Fan and Gijbels (1996), Fan *et al.* (1998) and Carroll *et al.* (1998)[16] discuss estimation of bias and variance of the local likelihood estimator. A consistent estimator of the local variance of $\hat{\theta}_x$ can be obtained as

$$\left(\sum \Lambda_i \bar{\Lambda}_i X_i X_i' K_i\right)^{-1} \left(\sum (Y_i - \Lambda_i)^2 X_i X_i' K_i^2\right) \left(\sum \Lambda_i \bar{\Lambda}_i X_i X_i' K_i\right)^{-1}$$

where $\Lambda(u) = \frac{1}{1+e^{-u}}$ is the logit function and $\Lambda_i = \Lambda(X_i'\hat{\theta}_x)$ and $\bar{\Lambda}_i = 1 - \Lambda_i$ and $\Lambda_x = \Lambda(x'\hat{\theta}_x)$ and $\bar{\Lambda}_x = 1 - \Lambda_x$.[17] For a small sample size, Carroll *et al.* (1998) and Galindo *et al.* (2001) propose a degrees of freedom correction.

The variance approximation for $\hat{E}[Y|X = x]$ can then be obtained by the delta method as

$$\mathrm{Var}\left(\hat{E}\,[Y|X = x]\right) \approx \Lambda_x^2 \bar{\Lambda}_x^2 \cdot x'\widehat{\mathrm{Var}}(\hat{\theta}_x)x.$$

Similarly, the variance of the marginal effects could be obtained by applying the delta method to

$$\frac{\partial E\,[Y|X = x]}{\partial x} = \Lambda(x'\hat{\theta}_x)\bar{\Lambda}(x'\hat{\theta}_x)\,\hat{\theta}_x$$

[16] Carroll *et al.* (1998) analyzed local estimating equations (which contains local likelihood as a special case where the score of the likelihood function gives the moment conditions) and derived consistent estimators for the variance, which correspond to the following expressions.

[17] Fan *et al.* (1998) and Carroll *et al.* (1998) propose modified versions of the variance estimator. Fan *et al.* (1998) replaces $(Y_i - \Lambda_i)^2$ in the above expression by $\Lambda_x \bar{\Lambda}_x$, whereas Carroll *et al.* (1998) replace $(Y_i - \Lambda_i)^2$ by $\Lambda_i \bar{\Lambda}_i$. Both modifications should lead to similar results if the smoothing window is small.

to obtain

$$\operatorname{Var}\left(\frac{\partial \hat{E}\,[Y|X=x]}{\partial x}\right) \approx \Lambda_x^2 \bar{\Lambda}_x^2 (I + (\bar{\Lambda}_x - \Lambda_x)\hat{\theta}_x x')\widehat{\operatorname{Var}}(\hat{\theta}_x)(I + (\bar{\Lambda}_x - \Lambda_x) x \hat{\theta}_x').$$

For obtaining confidence intervals also local bias has to be taken into account. Various methods to estimate the bias have been proposed, e.g. plug-in procedures using the asymptotic formulae for the bias and empirical procedures as suggested in Ruppert (1997) and Fan *et al.* (1998). The empirical approaches are usually based on higher order polynomials in the logit link function, and there seems to be only limited experience with these methods for higher dimensional X. With an estimate of the local bias, pointwise confidence intervals can be constructed as described in Fan *et al.* (1998), exploiting the asymptotic normality of the estimate.

To avoid estimation of the local bias, undersmoothing through the choice of a smaller bandwidth value would reduce the magnitude of the bias at the expense of a larger variance. With sufficient undersmoothing, the bias might become negligible compared to the variance such that confidence intervals can be based on estimated variance only. However, this undersmoothing clearly sacrifices precision. A more satisfying solution would be based on bootstrap confidence intervals as derived in Galindo *et al.* (2001).

2.4. Local multicollinearity

Compared to conventional Nadaraya–Watson regression with unbounded kernel, a well-defined maximizer of the local likelihood may not always exist or may be susceptible to numerical problems when the smoothing window is small. First, it may happen that *all* observations in the local smoothing window are zeros or ones. Second, perfect prediction in the local smoothing window may result in coefficient estimates converging to infinity. Third, collinearity of the regressors in the local smoothing window can render the estimates undefined. These problems have not been addressed in the literature so far, but could be quite important in finite samples. Obviously such problems should be of less concern when using a kernel with unbounded support, such as the Gaussian, compared to a compact kernel, such as the Epanechnikov. Nevertheless, even with the Gaussian kernel near-multicollinearity and perfect prediction can lead to very imprecise or undefined estimates.

The first problem could easily be solved by simply defining the estimate as one or zero if all observations in the smoothing window are one or zero.[18] In this case, marginal effects can no longer be computed directly from the local parametric model but require estimation of the conditional expectation function at x and $x + \Delta x$.

To overcome the other two problems, there are essentially two approaches: locally increasing bandwidths or dropping regressors. Locally dropping regressors that cause collinearity reduces the

[18] Fan *et al.* (1998) suggested to use only bandwidth values that are sufficiently large such that all local smoothing windows contain zeros as well as ones. This, however, would not be the best solution if the true conditional expectation function $E[Y|X=x]$ were indeed zero or one for some values of x, i.e. where the local logit model would be misspecified. Defining the estimate as zero or one is more in line with linear smoothers, which estimate the conditional expectation function as some weighted average of the observed Y in the smoothing window.

complexity of the local model,[19] whereas increasing the bandwidth values reduces the localization of the model. Different versions of the local likelihood logit estimator have been examined in the simulations and led to largely similar results. Overall local likelihood logit with Gaussian kernel and increasing bandwidths in case of collinearity gave the best results in the Monte Carlo. (More details are discussed in the supplementary appendix.)

3. FINITE SAMPLE PROPERTIES

In this section, the finite sample behaviour of local likelihood logit, local constant, local linear and Klein–Spady regression is analyzed for various simulation designs with 14 covariates (4 continuous, 10 binary) and samples of size 250 and 500, respectively. Hence, relative to the number of observations, the estimation problem can be considered as rather high-dimensional, since even the binary variables alone generate 1024 different cells. The out-of-sample prediction performance is examined for the conditional mean $E[Y|X]$ and for the marginal effects. Samples $\{(Y_i, X_i)\}_{i=1}^n$ of size n are drawn as well as validation samples $\{X_j\}_{j=1}^n$. From the sample $\{(Y_i, X_i)\}_{i=1}^n$ the conditional mean $E[Y|X = X_j]$ is predicted at all locations X_j and compared to the true conditional mean $E[Y|X = X_j]$.[20] The marginal effects are estimated for all 14 variables separately. For a binary variable, the effect of a change from 0 to 1 is estimated. For a continuous variable, the effect of an increase by 1 is estimated.[21]

The four continuous variables X_1^c, X_2^c, X_3^c, X_4^c are drawn from different χ^2 distributions, and the 10 binary variables $X_1^b, \ldots, X_{10}^b$ are Bernoulli distributed. Four different designs are considered, which differ in the dependence structure among the covariates. In designs 1 and 2, the continuous variables are uncorrelated, while they are correlated in designs 3 and 4. The binary variables are uncorrelated in designs 1 and 3 but correlated in designs 2 and 4.

X-design 1: The continuous variables $X_1^c, X_2^c, X_3^c, X_4^c$ are independent and are distributed χ^2 with 1, 2, 3 and 4 degrees of freedom, respectively. The binary variables $X_1^b, \ldots, X_{10}^b$ are distributed *Bernoulli*$(p = 0.5)$. All variables are independent of each other.

X-design 2: The continuous variables are distributed independently as in X-design 1. The binary variables are dependent: $X_1^b \sim \mathit{Bernoulli}(0.5)$ and $X_k^b \sim \mathit{Bernoulli}(p = 0.3 + 0.4\bar{X}_{k-1}^b)$, where $\bar{X}_{k-1}^b = \frac{1}{k-1}\sum_{l=1}^{k-1} X_l^b$ is the mean of the realized values of the 'preceding' binary variables. Thus,

[19] If all regressors except the constant are dropped, the estimator reduces to the Nadaraya–Watson estimator.

[20] The data sample $\{(Y_i, X_i)\}_{i=1}^n$ and the validation sample $\{X_j\}_{j=1}^n$ are drawn from the same population. The main interest is often not in estimating the conditional mean and marginal effects at the locations of the data but at some other location x_0, provided x_0 is in the support of X_i. In the analysis of discrimination, we might be interested in predicting the wages of women if they had the distribution of the human capital characteristics X that is observed among men. In a treatment evaluation context, we might be interested in predicting the earnings after participation in a training programme for individuals who did not participate, by controlling for differences in the characteristics X. For such analyses we need the support of X to be identical in both subpopulations, or more precisely, that the support of X in the validation sample is a subset of the support of X in the data sample.

[21] More precisely, the marginal effect is estimated as $\hat{E}[Y|X = \ddot{X}_j] - \hat{E}[Y|X = \dot{X}_j]$, where $\ddot{X}_j$ and $\dot{X}_j$ differ from X_j only in the component corresponding to the variable for which the effect will be estimated. For the effect of a binary variable, the corresponding element is set to 1 in $\ddot{X}_j$ and to 0 in $\dot{X}_j$. For a continuous variable, $\dot{X}_j$ equals X_j and the corresponding element in $\ddot{X}_j$ is increased by one.

if all preceding variables are one, the probability that the next variable also takes the value one is 0.7. The correlation among the binary variables lies between 0.1 and 0.4.

X-design 3: The binary variables are independent *Bernoulli*(0.5) variables as in X-design 1. The continuous variables are positively correlated. X_1^c is distributed $\chi^2_{(1)}$, X_2^c is generated as X_1^c plus an independent $\chi^2_{(1)}$, X_3^c is generated as X_2^c plus an independent $\chi^2_{(1)}$, and X_4^c is generated as X_3^c plus an independent $\chi^2_{(1)}$. The implied correlation among the continuous variables lies between 0.5 and 0.9.

X-design 4: The continuous variables and the binary variables are dependent and generated as in X-design 3 and X-design 2, respectively.

The Y observations are generated according to one of the five Y-designs:

Y-design 1: Linear index model without interaction or higher-order terms

$$Y = 1\left(-8 - X_1^c + 2X_2^c - 3X_3^c + 4X_4^c + 2\sum_{k=1}^{10} X_k^b(-1)^k + noise > 0\right).$$

Y-design 2: Linear index model with squared and interaction terms

$$\begin{aligned} Y &= 1 \quad \text{if } Y^* \geq 0, \text{ where} \\ Y^* &= 8 - X_1^{c2} + X_2^{c2} - X_3^2 + X_4^{c2} + 3X_1^c - 5X_2^c + 7X_3^c - 9X_4^c \\ &\quad + 2\sum_{j=1}^{10} X_k^b(-1)^k - X_1^cX_1^b + X_2^cX_2^b - X_3^cX_3^b + X_4^cX_4^b + noise \end{aligned}$$

Y-design 3: Linear index model with interaction terms

$$\begin{aligned} Y &= 1 \quad \text{if } Y^* \geq 0, \text{ where} \\ Y^* &= -8 - X_1^c + 2X_2^c - 3X_3^c + 4X_4^c + 2\sum_{k=1}^{10} X_k^b(-1)^k \\ &\quad - 3X_1^cX_1^bX_2^b + 3X_2^cX_3^bX_4^b - 3X_3^cX_6^bX_7^b + 3X_4^cX_8^bX_9^b + noise. \end{aligned}$$

Y-design 4: Nonlinear model with lower and upper threshold

$$\begin{aligned} Y &= 1 \quad \text{if } 8 \leq Y^* < 15, \text{ where} \\ Y^* &= 2\sqrt{\left|10 - X_1^c - X_2^c + X_3^c + X_4^c\right|} - 0.3\left(X_1^c + X_2^c\right)\sum_{k=1}^{4} X_k^b + 0.2\left(X_3^c + X_4^c\right)\sum_{k=5}^{10} X_k^b + noise \end{aligned}$$

Y-design 5: Index model with two-regimes

$$\begin{aligned} Y &= 1\left(Y_1^* \geq 0\right) \text{ if } \sum_{k=1}^{10} kX_k^b \text{ is below its mean, and} \\ Y &= 1\left(Y_2^* \geq 0\right) \text{ otherwise, where} \\ Y_1^* &= -4 - X_1^c + X_2^c - X_3^c + X_4^c - X_1^cX_1^b + X_2^cX_2^b - X_3^cX_3^b + X_4^cX_4^b + noise \\ Y_2^* &= -4 + \left(-X_1^c + X_2^c - X_3^c + X_4^c\right)^2 + 4\left(-X_1^cX_1^b + X_2^cX_2^b - X_3^cX_3^b + X_4^cX_4^b\right) + noise. \end{aligned}$$

The first three Y-designs correspond to the latent index threshold passing model familiar from utility maximization theory: An individual chooses a certain option (purchasing a good, participating in the labour force) if her idiosyncratic latent utility exceeds a certain threshold (opportunity cost, reservation wage). In Y-design 1, the latent index is a linear combination of the regressors, as it is, for instance, modelled in a logit, probit or single-index model. In Y-design 2, square and interaction terms enter the latent index. Interaction terms with the binary regressors are also included in Y-design 3. Y-design 4 represents a situation where a certain option is only chosen if a latent index is neither too large nor too small. As an example, consider the relationship between wages and the decision to work overtime. Overtime work will be attractive neither at very low wages nor at very high wages due to the income (wealth) effect. Y-design 5 models different behavioural rules for two different subpopulations. According to their binary regressors, each individual belongs either to subpopulation one or to subpopulation two, and each subpopulation faces a different outcome relationship. Such segregation might for instance be generated by administrative regimes which induce different incentives among eligible and non-eligible groups, for example, affirmative action programmes, preferential tax treatments, exemptions from social security or pension contributions, etc.

Four different designs for the noise are considered: In the first variant, noise is drawn from a *logistic* distribution, a symmetric distribution with relatively little mass in the tails. This implies that for Y-design 1 and logistic noise, the global logit model is correctly specified and the parametric logit estimator, which is used as the benchmark estimator, is consistent and efficient. In the second variant, referred to as *heteroskedastic* noise, the noise is drawn from a t_2 distribution and multiplied by $0.14\sqrt{\sum X_k^c \sum X_k^b}$. This leads to a heteroskedastic error with substantial probability mass in the tails.[22] In the third variant, called *bimodal*, the noise is drawn from a mixture of two normals $N(2, 1)$ and $N(-2, 1)$ leading to a symmetric bimodal distribution. In the fourth variant, called *asymmetric*, the noise is drawn from the asymmetric $\chi^2_{(3)}$ distribution. These four variants thus capture various features of the noise distribution.[23] Except for the logistic noise in Y-design 1, the global parametric logit estimator is always misspecified.

3.1. Implementation of the estimators

All estimators use as regressors $X_1^c, \ldots, X_4^c, X_1^b, \ldots, X_{10}^b$ and a constant, but no interaction or higher-order terms. The benchmark parametric *logit* is estimated by maximum likelihood.[24]

The semiparametric *Klein–Spady* estimator is implemented as in Gerfin (1996) with $m(\cdot)$ estimated by one-dimensional kernel regression and the bandwidth selected by generalized cross-validation. To be more specific, for a given bandwidth value h, the Klein–Spady estimator maximizes the quasi-log-Likelihood function:

$$\hat{\beta}_h = \arg\max_{\beta} Q(\beta, h) = \sum_i Y_i \ln \hat{m}_i + (1 - Y_i) \ln(1 - \hat{m}_i)$$

[22] The variance of the t_2 distribution is infinite.

[23] The mean of Y depends on the Y-design, X-design and the noise and varies between 0.43 and 0.56.

[24] If the parametric logit did not converge, e.g. collinearity or perfect prediction, a new sample is drawn.

where

$$\hat{m}_i = \frac{\sum_l Y_l \phi\left(\frac{X_l'\beta - X_i'\beta}{h \cdot std(X'\beta)}\right)}{\sum_l \phi\left(\frac{X_l'\beta - X_i'\beta}{h \cdot std(X'\beta)}\right)}$$

where $std(X'\beta)$ is the standard deviation of $X'\beta$ in the sample. For a given bandwidth value h, the estimates $\hat{\beta}_h$ are obtained by Newton–Raphson iteration (with step size 0.2). Since the coefficients β are identified only up to scale, the first coefficient is normalized as one. The bandwidth is chosen that minimizes the generalized cross-validation criterion

$$GCV(h) = \frac{\sum_i (Y_i - \hat{m}_i)^2}{\left(\sum_i (1 - \xi_i)\right)^2}$$

where

$$\xi_i = \frac{\phi(0)}{\sum_l \phi\left(\frac{X_l'\hat{\beta}_h - X_i'\hat{\beta}_h}{h \cdot std(X'\hat{\beta}_h)}\right)}$$

is the kernel weight accorded to observation i in the estimation of $\hat{m}_i$. The $GCV(h)$ is computed for 30 different values of $h \in \{0.02, 0.04, \ldots, 0.60\}$ and the bandwidth corresponding to the minimum is chosen. With this estimated bandwidth $\hat{h}$, the coefficients $\hat{\beta}_{\hat{h}}$ are estimated and estimates of the conditional mean at a value x are obtained as

$$\hat{E}[Y|X = x] = \frac{\sum_l Y_l \phi\left(\frac{X_l'\hat{\beta}_{\hat{h}} - x'\hat{\beta}_{\hat{h}}}{\hat{h} \cdot std(X'\hat{\beta}_{\hat{h}})}\right)}{\sum_l \phi\left(\frac{X_l'\hat{\beta}_{\hat{h}} - x'\hat{\beta}_{\hat{h}}}{\hat{h} \cdot std(X'\hat{\beta}_{\hat{h}})}\right)}.$$

For the *non-parametric* Nadaraya–Watson, local linear and local logit estimators, the bandwidths are chosen either according to the least-squares criterion CV_{LS} (7) or according to the likelihood criterion CV_{ML} (8).[25] The kernel weights $K(X_i - x)$ for given bandwidth values h, λ are computed in all estimators as

$$K_{h,\lambda}(X_i - x) = \prod_{k=1}^{10} \lambda^{1(X_{k,i}^b \neq x_k^b)} \prod_{k=1}^{4} \kappa\left(\frac{X_{k,i}^c - x_k^c}{h}\right), \tag{9}$$

where κ is either the Epanechnikov kernel $\kappa(u) = \frac{3}{4}(1 - u^2)1_{[-1,1]}(u)$ or the Gaussian kernel.

For local linear and local logit regression, the 14 regressors (plus a constant) enter not only in the kernel function $K(X_i - x)$ but also in the local specification $g(x, \theta_x)$. To improve numerical accuracy and to ensure that all continuous regressors are of similar magnitude, the continuous variables are scaled to the same standard deviation prior to estimation. The estimated marginal effects refer to the unscaled variables, though.

Local linear regression specifies the conditional expectation function locally as

$$E[Y|X = x] \doteq x'\theta_x, \tag{10}$$

[25] From a grid of 20 × 20 gridpoints: $(h, \lambda) \in \{1, 1.2, \ldots, 1.2^{18}, \infty\} \times \{0.05, 0.10, 0.15, \ldots, 1\}$ for $n = 250$ and $(h, \lambda) \in \{1.2^{-6}, 1.2^{-5}, \ldots, 1.2^{12}, \infty\} \times \{0.05, 0.10, 0.15, \ldots, 1\}$ for $n = 500$.

where a constant is included among the regressors x. For given bandwidth values $E[Y|X = x]$ is estimated by the first element of the weighted least-squares estimate $\hat{\theta}_x$, with

$$\hat{\theta}_x = \left(\mathbf{X}_x' \mathbf{W}_x \mathbf{X}_x\right)^{-1} \mathbf{X}_x' \mathbf{W}_x \mathbf{Y}_x,$$

$$\mathbf{X}_x = \begin{pmatrix} 1 & \left(X_{1,1}^c - x_1^c\right) & \cdots & \left(X_{10,1}^b - x_{10}^b\right) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \left(X_{1,n}^c - x_1^c\right) & \cdots & \left(X_{10,n}^b - x_{10}^b\right) \end{pmatrix} \text{ and } \mathbf{Y}_x = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \text{ and } \mathbf{W}_x = diag\left(K_H(X_i - x)\right) \tag{11}$$

(see Fan and Gijbels 1996; Ruppert and Wand 1994; Seifert and Gasser 1996). Since the local linear estimates may lie outside the interval [0, 1], they are capped at zero and at one.

The *local constant* (Nadaraya Watson) estimator is a special case of local linear regression with all slope coefficients in (10) constrained to zero:

$$\hat{E}[Y|X = x] = \frac{\sum Y_i K_H(X_i - x)}{\sum K_H(X_i - x)}.$$

3.2. Simulation results

The out-of-sample prediction performance for the conditional mean $E[Y\}X]$ and for the marginal effects is assessed by their simulated mean absolute error (MAE), median absolute error (MdAE), mean squared error (MSE) and median squared error (MdSE). For each of the 80 different designs,[26] data samples and validation samples are drawn repeatedly[27] from the same population, and for all observations of the validation sample the conditional means and marginal effects are estimated on the basis of the observations in the data sample, as described above. Since the Monte Carlo study comprehends a large number of different designs and variants of the estimators, only the most salient results are summarized in the following. A supplementary appendix, available from the author's webpage, contains all the detailed simulation results for all 80 designs.

The performance is measured *relative* to the benchmark parametric logit estimator, and all results will be given in percent.[28] Among these four performance measures, generally the relative performance of local logit is worst with respect to MSE and best with respect to MdSE. The relative performance with respect to the other two measures is usually between these two. Therefore, in the following, only the results for the MSE and the MdSE are given. (The results for the mean and median absolute error can be found in the supplementary appendix.)

First, the relative performance of the various estimators in those designs where the parametric logit model is correctly specified is examined in Table 1. These are the four X-designs with Y-design 1 and logistic noise, and Table 1 gives the average over these four designs for sample size 250 and 500, respectively. The first two rows of Table 1 contain the results for estimating the conditional mean $E[Y|X]$. Klein–Spady, Nadaraya–Watson and local linear regression perform

[26] Four X-designs, five Y-designs and four noise variants.

[27] One hundred replications for each design.

[28] Hence, numbers below 100 indicate an improvement over parametric logit regression, whereas numbers above indicate a worse performance.

Table 1. Average prediction performance when global logit is correct.

Sample	Klein–Spady		Nadaraya–Watson		Local linear		Local logit CV_{LS}		Local logit CV_{ML}	
size	MSE	MdSE	MSE	MdSE	MSE	MdSE	MSE	MdSE	MSE	MdSE
	Prediction performance for conditional mean									
250	218.6	295.1	477.6	13198	263.7	13934	115.6	109.4	108.3	106.0
500	250.8	426.7	880.1	23446	478.5	26390	113.6	110.4	110.4	109.1
	Prediction performance for marginal effects of continuous regressors									
250	782.5	560.1	598.4	2548.3	224.0	2249.6	118.4	112.2	108.3	106.1
500	993.7	894.3	1512.0	5416.9	575.1	5077.2	116.2	110.6	111.6	109.7
	Prediction performance for marginal effects of binary regressors									
250	556.5	488.9	230.4	707.0	145.7	1932.1	114.5	111.3	107.1	106.8
500	666.5	780.5	482.9	1355.1	279.9	3845.2	113.5	110.2	109.4	108.5

Note: All figures relative to parametric logit. For Nadaraya Watson and local linear regression, the bandwidth is chosen by CV_{LS}. For local logit, results for bandwidth choice by CV_{LS} and by CV_{ML} are given. Average over the four X-designs with Y-design 1 and logistic noise.

quite poorly with an MSE of about two to nine times the MSE of the parametric logit estimator. The relative performance worsens when the sample size increases, since the parametric logit estimator converges at a faster rate. In contrast, the relative precision of the *local logit* estimator does not change with the sample size and its MSE is only about 10 to 15% larger than for parametric logit. This behaviour can be explained by the very large bandwidths selected by cross-validation, which even increase when the sample size is increased. (Results not shown.) The CV_{ML} cross validation criterion leads here to better results than the conventional least-squares cross validation criterion CV_{LS}.

Table 2 provides the relative performance results averaged over all 80 designs. With the parametric logit now misspecified in almost all of the designs, the relative performance of all estimators improves compared to Table 1, and it also improves with the sample size for Klein–Spady and the local logit estimator. (For NW and local linear, the relative performance worsens with sample size with respect to MAE and MSE, but improves with respect to median absolute error.) Nevertheless, Klein–Spady, NW and local linear are still usually worse than parametric logit,[29] whereas local logit seems to be more precise than parametric logit.

For local logit, the CV_{LS} criterion for bandwidth choice leads to best results with respect to MAE, MdAE and MdSE. The precision gains vis-a-vis parametric logit are about 25% and 55% in terms of MSE and MdSE, respectively, for sample size n = 500. For estimating marginal effects, local logit is still more precise than parametric logit, but the gains in precision are smaller. Choosing the bandwidth by the CV_{ML} criterion, on the other hand, often leads to a smaller MSE and also to a more robust estimator that depends less on the particular design properties, as discussed in the following.

[29] With respect to median absolute error, Klein–Spady is about 5% more precise than parametric logit but only for estimating the conditional mean. Otherwise, Klein–Spady is worse than parametric logit.

Table 2. Average prediction performance over all designs.

Sample	Klein–Spady		Nadaraya Watson		Local linear		Local logit CV_{LS}		Local logit CV_{ML}	
size	MSE	MdSE	MSE	MdSE	MSE	MdSE	MSE	MdSE	MSE	MdSE
					Prediction performance for conditional mean					
250	128.4	104.1	202.8	>10000	132.8	>10000	85.5	53.7	84.7	61.6
500	118.0	104.6	273.4	>10000	175.0	>10000	75.3	45.5	75.7	51.8
					Prediction performance for marginal effects of continuous regressors					
250	254.0	287.4	201.3	>10000	129.0	>10000	93.1	77.5	91.7	81.6
500	204.6	257.5	284.6	>10000	165.7	>10000	86.1	69.8	86.0	77.1
					Prediction performance for marginal effects of binary regressors					
250	227.8	212.9	134.7	>10000	112.4	>10000	103.9	74.0	99.4	80.1
500	186.0	192.8	172.3	2057	132.9	>10000	99.1	63.9	94.4	70.6

Note: All figures relative to parametric logit. For Nadaraya Watson and local linear regression, the bandwidth is chosen by CV_{LS}. For local logit, results for bandwidth choice by CV_{LS} and by CV_{ML} are given. Average over all 80 designs.

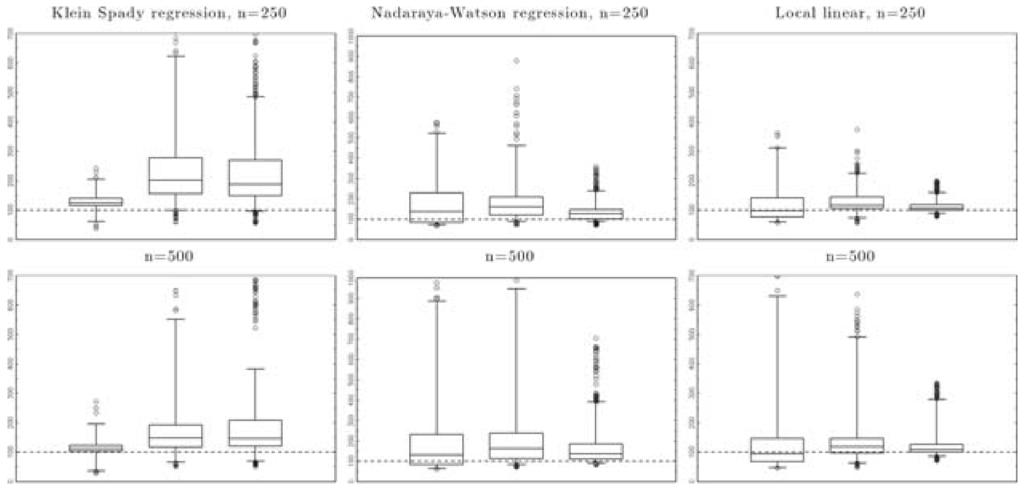

Figure 1. Distribution of relative MSE over the 80 designs. MSE relative to parametric logit. Distribution of the relative MSE over the 80 different designs. Results for Klein–Spady (left-hand side), Nadaraya Watson (middle) and local linear regression (right-hand side). The first box-plot gives the relative MSE for predicting the conditional mean $E[Y|X]$, the second box-plot for the marginal effects of the four continuous regressors and the third box-plot for the marginal effects of the 10 binary regressors. The boxes represent the median and the 25 and 75 percentiles and extend to the 5 and 95 percentiles.

The previous tables examined only averages over the 80 different designs. This may be a useful summary statistic, but does not contain any information about the design dependence of the estimator's performance. Some estimators may obtain similar results in all designs, whereas others may be poor in some designs but better in others. Figures 1 and 2 demonstrate how much the relative performance varies across the designs. For each of the 80 different designs, the relative

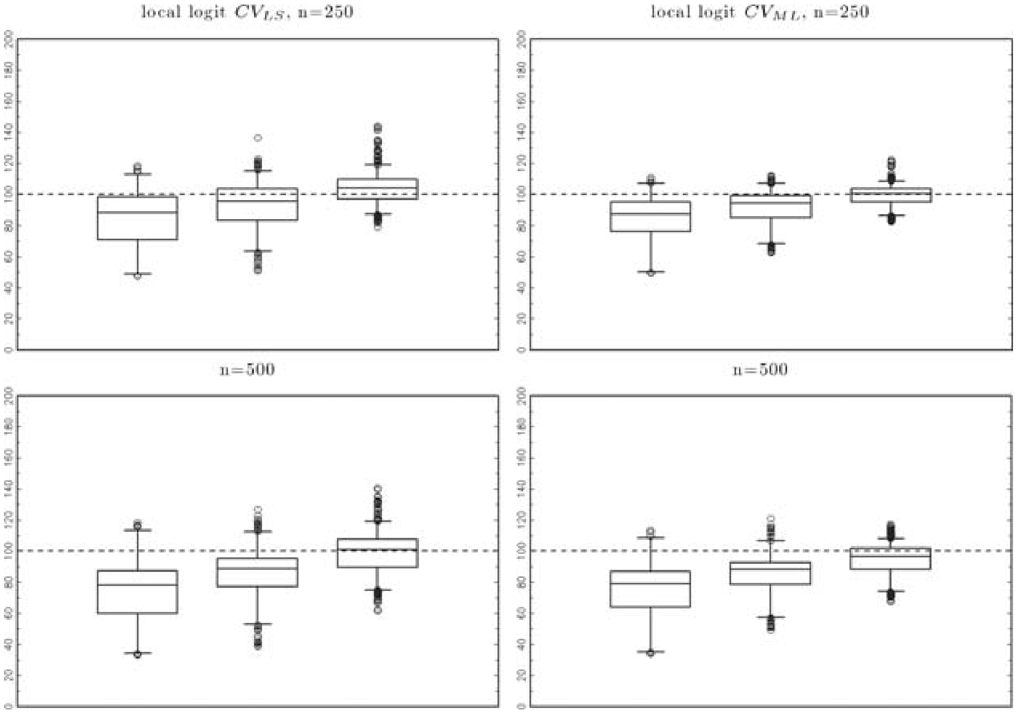

Figure 2. Distribution of relative MSE over the 80 designs. MSE relative to parametric logit. Distribution of the relative MSE over the 80 different designs. Results for local logit with CV_{LS} bandwidth choice (left-hand side) and with CV_{ML} bandwidth choice (right-hand side). The first box-plot gives the relative MSE for predicting the conditional mean $E[Y|X]$, the second box-plot for the marginal effects of the four continuous regressors and the third box-plot for the marginal effects of the 10 binary regressors. The boxes represent the median and the 25 and 75 percentiles and extend to the 5 and 95 percentiles.

MSE in estimating the conditional mean function has been simulated[30] and the first box-plot in each graph shows the distribution of these 80 values. The boxes represent the median and the 25 and 75 percentiles and extend to the 5 and 95 percentiles. Results are capped at 4000. The second box-plot refers to the performance in estimating marginal effects for the continuous covariates, and is based on 320 different values (80 designs × 4 continuous variables). Finally, the third box-plot refers to the performance in estimating marginal effects for the binary covariates, and is based on 800 different values (80 designs x 10 dummy variables). (Figure B2 in the appendix also shows the results for MdSE.)

Figures 1 and 2 document the favourable results for the local logit estimator in this Monte Carlo study. Not only is local logit more precise than parametric logit in the majority of the designs (i.e. the median is below 100), but it also seems to be more robust to design properties than the other estimators, in that its interquartile range and the 5–95 spread are smaller. (Notice that the

[30] These values are shown in Tables B3 to B10 in the supplementary appendix.

scaling of the ordinate varies between the estimators, e.g. the maximum value is close to 700 for the Klein–Spady estimator but much smaller for local logit.)

The *Klein–Spady* estimator performs clearly worse than parametric logit in the majority of the designs, particularly when estimating marginal effects. Its behaviour improves with the sample size, but at n = 500 it is still worse than parametric logit.[31]

For *Nadaraya–Watson* the results are even worse, with a very large relative MSE in a number of designs and a spread increasing with sample size. In terms of MdSE, the Nadaraya–Watson estimator performs very poorly.[32] The results are less extreme for the marginal effects, but still very large. Hence, in terms of median prediction performance, Nadaraya–Watson can be very unreliable.

These findings are similar for *local linear* regression. With respect to MSE, local linear regression behaves better than Nadaraya–Watson, and its MSE varies less with the simulation design. For predicting the conditional mean, it is slightly more precise than parametric logit, and for the marginal effects it performs better than Klein–Spady. In terms of MdSE, however, it can often be very imprecise. The finding that local linear regression performs extremely badly with respect to MdSE but less badly in terms of MSE indicates that local linear regression produces disproportionately many small errors. This could be related to the non-differentiability of the local model, which caps the estimates at 0 and 1. This produces rather many extreme predictions of 0 and 1, whereas in the true data generating processes $E[Y|X] \in \{0, 1\}$ occurs with probability zero.

The results for *local logit* are shown in Figure 2. In the majority of the designs, local logit is more precise than parametric logit, particularly for the conditional mean function. Comparing bandwidth selection by CV_{LS} or by CV_{ML}, it seems that CV_{ML} leads to a smaller spread over the designs. Particularly for n = 250, the interquartile range is smaller with CV_{ML} than CV_{LS}. This holds with respect to MSE as well as MdSE. Although the ranking of CV_{LS} and CV_{ML} with respect to the average performance over the 80 designs is ambiguous (see Table A2, the CV_{ML} criterion led on average to a higher MAE, MdAE and MdSE but a smaller MSE), the CV_{ML} criterion clearly reduced the variance in the performance over the 80 designs: the standard deviation over the 80 designs is smaller for all the four performance measures MAE, MdAE, MSE and MdSE. Although the differences are not very large, this seems to indicate that the CV_{ML} criterion makes the local logit estimator more robust to the particular design properties.[33]

4. HETEROGENEOUS FEMALE LABOUR SUPPLY

The previous section indicated that local likelihood logit can work well even in higher-dimensional settings. In this section, local logit is applied to analyze female labour supply. Determinants of female labour supply have since long been of interest to economists, arguing about the need for subsidized child care or all-day schooling. As confirmed by many studies, female labour force

[31] These findings hold with respect to MSE as well as MdSE, but with a much larger spread for MdSE. In some designs, the relative MdSE is very large whereas it is very small in others.

[32] In more than 20 of the 80 designs the predictions of E[Y}X] are more than 40 times less precise than for parametric logit. (The results are capped at 4000) At the median of the 80 designs, the MdSE is larger than 400. These findings are similar, but somewhat less extreme, with respect to median absolute error.

[33] It should be kept in mind that precision is measured relative to the parametric logit estimator.

participation generally decreases with the number of children and particularly if these children are young. For policy considerations, however, it would be relevant to know whether all women adjust their labour supply in the same way as a reaction on family size or whether some sub-populations react differently. Particularly, for some women, labour supply might be inelastic to family size, whereas for others it might even increase as a reaction on an additional child, e.g. because of increased financial needs. If women's reaction on family size is heterogeneous, the provision of child care subsidies, tax incentives, etc. should be targeted more precisely than if it were largely homogeneous. Therefore, in the analysis of female labour supply, not only should mean effects be estimated but their distribution should also be estimated.

To assess heterogeneity in women's response to family size, the labour force participation of married Portuguese women is analyzed by a reduced form labour supply model. The data are taken from Martins (2001) and consist of 2339 women of whom 60% had been working in 1991. Five explanatory variables are available: age, years of education, husband's monthly wage, number of children below the age of 4 and number of children 4 to 18 years old. Tables A1 and A2 in the Appendix contain descriptive statistics.

For each woman her employment probability $P(Y = 1|X_i)$ given her characteristics X_i is estimated, where Y denotes employment status (1 employed, 0 non-employed). In addition, the marginal effects of the characteristics X_i on employment are estimated, in particular the effects of the number of children. The effect of an additional child on the employment probability depends on all characteristics X_i and thus differs from woman to woman.[34]

The employment probabilities $P(Y = 1|X_i)$ and the marginal effects are estimated by parametric logit, local logit and Klein–Spady. A bandwidth of 0.1 times the standard deviation of the index $x\beta$ was selected for the Klein Spady estimator, and for scale normalization the first coefficient is fixed. For the local logit estimator, all five variables (plus a constant) enter in the local model and in the kernel weighting. The local logit specification is economically appealing as it incorporates monotonicity, decreasing marginal effects and non-saturation. From a simple utility-maximizing labour supply model, the labour supply should usually decrease with the number of children but the effect of an additional child should diminish (e.g. due to returns to scale in child rearing and home production). Nevertheless, the marginal effect should not fall to zero. These implications of the simple model are not incorporated in the local constant or the capped local linear model.

For the kernel weighting, the five regressors are split into two groups: age, education and husband's wage income are treated as continuous variables. The optimal bandwidth for each of these three variables is supposed to be proportional to its standard deviation. By imposing the restrictions that $h_{\text{age}} = h\,Std(age)$, $h_{\text{education}} = h\,Std(\text{education})$ and $h_{\text{wage}} = h\,Std(wage)$, it suffices to estimate a single bandwidth h, while at the same time ensuring that the local neighbourhoods are larger for regressors that display more variation. In the actual implementation of the estimator, this restriction is accommodated by scaling the continuous regressors to mean zero and variance one. The second group of regressors consists of the number of children 0–3 years and 4–18 years old. These two variables are treated as ordered discrete. The same bandwidth value is used for

[34] Note that the estimated effects of children can be interpreted as causal only if the number of children is exogenous given the other characteristics. If some confounding variables that affect both the number of children and the inclination to work are missing, the estimated employment effects are a mixture of the proper causal effect and a selection effect, (see Heckman 1990; Manski 1993). This would change the interpretation of the estimated effects but not the comparison of the different estimators' ability in detecting heterogeneous effects.

Table 3. Logit and Klein–Spady estimated coefficients.

Estimated coefficients	Logit	Klein–Spady
	Dependent variable: Employment	
Constant	1.34	
Age in years	-0.03^{***}	-0.05
Education in years	0.24^{***}	0.28^{***}
Husband's log wage	-0.10	-0.08
Children 0 to 3 years old	-0.39^{***}	-0.36^{***}
Children 4 to 18 years old	-0.13^{***}	-0.17^{***}

Note: Coefficients significant at the 1, 5 or 10% level are marked by ***,** or *, respectively. The Klein–Spady coefficients are divided by -20 to ease comparison with the logit estimates. The first coefficient of the Klein–Spady estimator was normalized to one.

both variables because it is a priori unclear whether a family having an additional older child is more different than a family having an additional younger child. Since both variables enter also in the local model, their coefficients can accommodate the differences in the effects of younger versus older children. Alternatively, two different bandwidth parameters for the children variables could have been included in the cross-validation bandwidth search, but this would have increased computation time considerably. The two bandwidths h, δ were chosen by cross-validation as $(h, \delta) = (3.21, 0.35)$.[35]

The coefficient estimates of the parametric logit and the semiparametric Klein Spady estimator are given in Table 3. The logit coefficients indicate that the probability of employment reduces somewhat with age but increases significantly with educational attainment. Husband's income seems to be of minor relevance, whereas children and particularly small children strongly reduce employment. The Klein–Spady estimates display a similar pattern, but the difference between younger and older children is smaller.

Figure 3 shows the estimated employment probabilities of logit versus Klein Spady (left-side) and logit versus local logit (right-hand side). The comparison of the Klein–Spady to the logit estimates displays a wavelike pattern (that even persists with larger bandwidth values; not shown). On the other hand, the local logit estimates are, apart from very few outliers, similar to the logit estimates, perhaps because of the rather large bandwidth value $h = 3.21$.

To analyze heterogeneity in the response to an additional child, marginal effects are estimated for all 2339 observations. For the continuous variables (age, education, husband's wage), the effect of a 5% increase is estimated.[36] For the estimation of the children effects, different family compositions are considered and compared to the base category zero children. Table 4 shows the marginal effects (in %-points) estimated by parametric logit, Klein–Spady and local logit, respectively. The marginal effects were estimated for all 2339 observations, and the rows labelled 'mean' provide the average over the 2339 observations, while $Q_{0.05}$, $Q_{0.25}$, $Q_{0.75}$ and $Q_{0.95}$ refer to their 5, 25, 75 and 95 percentiles (with respect to the 2339 observations). On average, all three estimators predict similar marginal effects for the continuous variables, e.g. an increase in educational attainment increases the employment probability by 1.6 %-points. For the effects of

[35] From a grid of 20 × 20 gridpoints: $(h, \delta) \in \{0.3, 0.3 \cdot 1.2, \ldots, 0.3 \cdot 1.2^{18}, \infty\} \times \{0.05, 0.10, 0.15, \ldots, 1\}$

[36] More precisely, the effect of a 2.5% increase compared to a 2.5% decrease in the continuous variable.

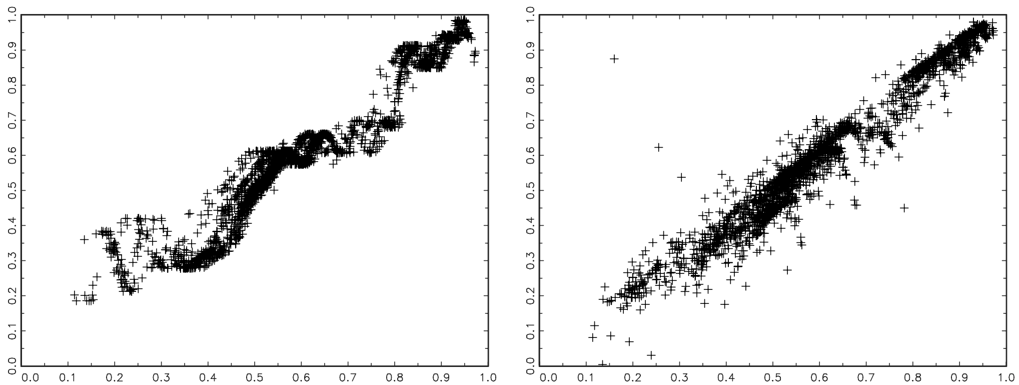

Figure 3. Estimated employment probabilities logit vs. Klein–Spady and logit vs. local logit Abscissa: Estimated employment probability corresponding to logit model. Ordinate: Estimated employment probability corresponding to Klein–Spady (left-hand side) or local logit (right-hand side); 2339 observations.

children, on the other hand, parametric logit produces larger estimates than Klein–Spady and local logit. Compared to no children, having an older child reduces employment probability by 2.6 %-points, whereas a younger child reduces employment by 7.9 %-points. Relative to no children, two children reduce the employment likelihood by 5.3 to 16 %-points and three children by 8 to 24 %-points, depending on their age structure. Local logit and Klein–Spady regression estimate the effect of two children to only about 4 to 11 %-points.

More interesting, however, is the distribution of these marginal effects in the population, given by the quantiles in Table 4. The effects estimated by parametric logit are not spread out very much. For example, the effect of an older child is more negative than −3.2 for a quarter of the population, whereas it is less negative than −2.2 for the quarter of the population with the weakest reaction on a child. It is also apparent that for the parametric logit estimator the estimated effects are always positive or always negative and never change sign in the population, e.g. the employment effects of children are negative for all women. This pattern is due to the globally monotone logit specification and exemplifies how parametric regression may overlook heterogeneity in the effects.

This is different with the semi- and non-parametric estimators. According to local logit, an older child reduces employment by more than 2.5 %-points for a quarter of the population but it *increases* employment by at least 2.7 %-points for an other quarter of the population. For 5% of all women, the increase in the employment probability is even larger than 5.6 %-points. Although for larger family sizes the employment effects become more negative, still at least 5% of the population exhibits positive effects even for two children. (For the Klein–Spady estimator this holds even for three children.) Thus, for a part of the population, having one (or two) children does not reduce labour force participation and might even increase it.

A graphical summary of these marginal effects in form of box-plots, which cover 95% of their distribution mass (2.5 to 97.5 percentile), is provided in Figure 4. The left-hand gives the results according to parametric logit, the middle picture for local logit and the right-hand for Klein–Spady. Besides positive employment effects of children for parts of the population, particularly for one child, it can also be seen from Figure 4 and Table 4 that the Klein–Spady estimates are generally the most variable. Their interquartile ranges and their standard deviations (not shown in

Table 4. Distribution of marginal effects on employment probability (in %-points).

	Age	Education	Husb's wage	One old child	One young child	Two older children	One old & 1 young	Two young children	Three older children	Two old & 1 young	Two young & 1 old	Three young children
							Logit					
Mean	−1.2	1.6	−1.2	−2.6	−7.9	−5.3	−10.6	−16.0	−8.0	−13.4	−18.7	−24.0
$Q_{0.05}$	−2.0	0.3	−1.5	−3.3	−9.7	−6.6	−13.0	−19.2	−9.9	−16.2	−22.4	−28.4
$Q_{0.25}$	−1.6	1.3	−1.4	−3.2	−9.6	−6.5	−12.8	−19.1	−9.8	−16.0	−22.2	−28.2
$Q_{0.75}$	−0.9	1.8	−1.0	−2.2	−6.4	−4.4	−8.7	−13.4	−6.5	−11.0	−16.0	−20.9
$Q_{0.95}$	−0.4	2.2	−0.4	−0.8	−2.5	−1.6	−3.5	−5.9	−2.5	−4.7	−7.3	−10.5
							Local logit					
Mean	−1.2	1.6	−1.9	0.0	−2.0	−4.0	−8.9	−11.7	−12.2	−11.8	−27.1	−19.3
$Q_{0.05}$	−3.8	0.2	−7.8	−6.6	−6.4	−14.6	−18.9	−24.5	−22.3	−22.1	−37.7	−36.7
$Q_{0.25}$	−1.9	1.3	−4.9	−2.5	−3.2	−6.8	−14.0	−17.9	−15.9	−15.4	−33.9	−28.7
$Q_{0.75}$	−0.4	2.0	0.3	2.7	−0.6	−0.6	−3.9	−4.3	−8.2	−8.3	−22.0	−9.6
$Q_{0.95}$	0.6	2.6	4.8	5.6	2.3	2.8	2.1	2.8	−2.3	−0.4	−5.7	2.9
							Klein–Spady					
Mean	−1.5	1.5	−0.7	−2.1	−4.5	−4.2	−7.3	−10.5	−7.1	−10.3	−13.1	−15.7
$Q_{0.05}$	−6.2	−2.3	−2.3	−8.1	−14.8	−14.3	−19.2	−24.7	−18.8	−24.2	−29.1	−31.6
$Q_{0.25}$	−3.1	−0.5	−1.5	−5.1	−9.5	−8.9	−14.0	−18.4	−13.6	−18.2	−20.1	−24.1
$Q_{0.75}$	0.5	3.1	0.3	1.4	0.6	0.9	−3.7	−3.7	−3.5	−3.8	−5.4	−8.7
$Q_{0.95}$	2.9	5.8	1.4	4.8	5.9	6.0	4.7	0.3	4.6	0.1	2.9	4.2

Note: Changes in employment probability (in %-points) due to a change in one of the characteristics. *Mean* provides the sample mean of the estimated marginal effects; $Q_{0.05}$, $Q_{0.25}$, $Q_{0.75}$ and $Q_{0.95}$ represent their 5, 25, 75 and 95 percentiles in the population. For the continuous variables, the effects refer to a 5% increase in this variable. The effects for the different children compositions are always relative to the base category of zero children.

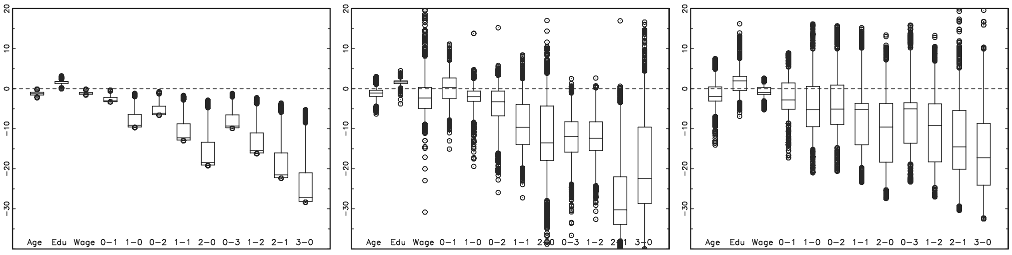

Figure 4. *Distribution of marginal effects in the population* Distribution of estimated marginal effects according to parametric logit (left-hand side), local logit (middle) and Klein–Spady (right-hand side). 0–1 denotes the effect of zero young and one older children relative to the base category of no children. 2–1 denotes the effect of two younger and one older children. Effects correspond to Table 4. The boxes represent the median and the 25 and 75 percentiles (see $Q_{0.25}$ and $Q_{0.75}$ in Table 4) and extend to the 2.5 and 97.5 percentiles. Estimated effects below the 2.5 or above the 97.5 percentile are marked by circles.

Table 4) are usually larger than for the effects estimated by local logit or logit. This could either imply that the Klein–Spady estimates are most noisy or imply that effect heterogeneity is even more pronounced.

To examine whether the apparent effect heterogeneity detected by the local logit and Klein–Spady estimators is genuine or merely spurious due to a larger variance of these estimators, it is revealing to contrast in a first step the characteristics of the women that display positive effects to those with negative effects. This is done in Table 5, where the upper part is based on the local logit estimates, whereas the lower part refers to the Klein–Spady estimates. The columns labelled as $+$ give the average characteristics of those women for whom the respective effect of children is positive, whereas the columns labelled as $-$ give the characteristics of the women with a negative effect on employment. The first columns refer to the effect of one child versus no children. The following columns refer to the effect of two versus zero children, and finally, the effect of three versus zero children is considered.

According to local logit, the employment effect of one older child is estimated to be positive for 1239 observations and negative for the remaining 1100 observations. These 1239 women with a positive effect are on average 36 years old, with 8.6 years of education and a husbands' log wage income of 11. In contrast, the 1100 women exhibiting a negative effect are on average 41.2 years old, with 5.7 years of education and a husband's log wage of 11.4. When comparing (along the rows) the women with positive to those with negative effects, a striking pattern with respect to education is found for the local logit estimator. Women whose employment probability increases with children are always substantially higher educated than women with decreasing employment probability. For age and husband's wage income, no such regularities can be found. These two variables seem not to be distinguishing characteristics between women with positive and those with negative effects of children.

For the Klein–Spady estimates given in the lower part of Table 5, on the other hand, no such regular patterns can be found. Although education appears to be slightly higher among women with positive effects, this pattern is not stable and the differences are small. The women with positive effects seem not to be systematically different from those with negative effects. The heterogeneity in the effects of children on labour supply detected by the semiparametric Klein–Spady estimator seems to be largely spurious and generated by its larger variance. In contrast, the

Table 5. Comparison of women with positive versus negative employment effects of children.

	One old child		One young child		Two old children		One old & 1 young		Two young children		Three old children		Two old & 1 young		Two young & 1 old		Three young children	
	+	−	+	−	+	−	+	−	+	−	+	−	+	−	+	−	+	−
Local logit																		
No. of obs	1239		471		484		282		325		31		84		4		257	
Age	36.0	41.2	40.2	38.0	37.1	38.8	36.4	38.7	44.4	37.5	32.9	38.5	40.2	38.4			44.4	37.7
Education	8.6	5.7	12.3	5.9	12.9	5.8	13.3	6.4	12.7	6.4	16.3	7.1	16.3	6.9			13.0	6.5
Husb wage	11.0	11.4	11.4	11.1	11.2	11.2	11.2	11.2	11.4	11.2	11.7	11.2	11.4	11.2			11.4	11.2
Klein–Spady																		
No. of obs	776		677		693		292		148		296		148		162		152	
Age	36.6	39.3	36.3	39.3	36.4	39.3	39.0	38.3	38.3	38.4	39.0	38.3	40.1	38.3	48.1	37.7	46.3	37.9
Education	7.7	7.0	7.6	7.1	7.7	7.1	8.4	7.1	7.1	7.2	8.5	7.1	7.1	7.2	4.8	7.4	5.1	7.4
Husb wage	11.2	11.2	11.2	11.2	11.2	11.2	11.2	11.2	11.2	11.2	11.2	11.2	11.2	11.2	11.2	11.2	11.2	11.2

Note: Number of observations and average characteristics of women with increased employment probability (columns +) and with reduced employment probability (−) due to children. Characteristics of groups with less than 20 observations are not displayed. Total number of observations 2339.

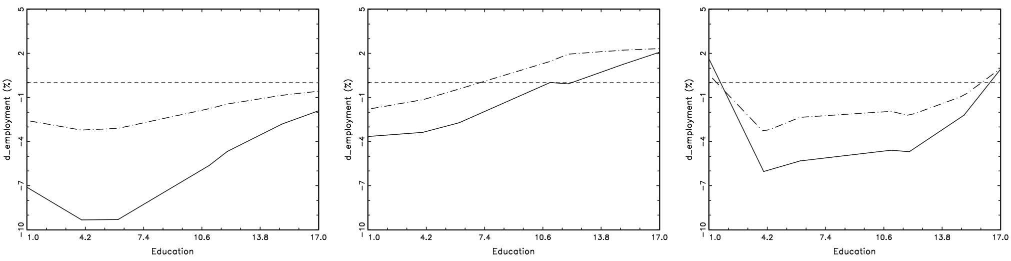

Figure 5. Employment effect of one child conditional on educational level. Change in employment probability (in %-points) due to one older child (dashed line) or one younger child (solid line), relative to no children, according to parametric logit (left-hand side), local logit (middle) and Klein Spady (right-hand side), for different educational levels. Effects for educational levels with less than 20 observations not displayed.

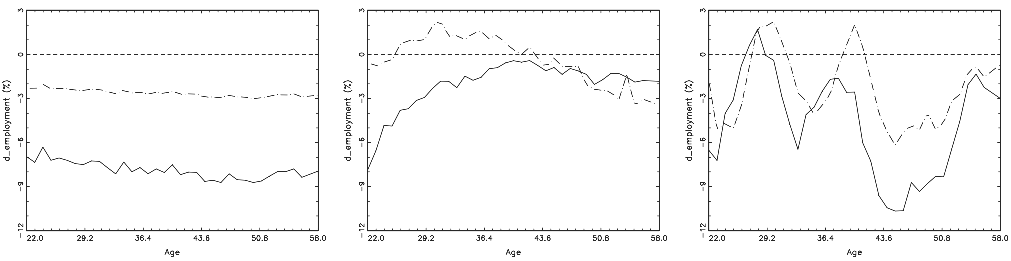

Figure 6. Employment effect of one child conditional on age.

heterogeneity detected by local logit seems to represent an authentic pattern in that the reaction to children varies with the educational level.[37]

This finding is corroborated by Figures 5 and 6, which examine the correlation between the effects of children and education (or age, respectively). Figure 5 plots the estimated employment effect of one older child (dashed line) and of one younger child (solid line) for different educational levels.[38] The left-hand picture corresponds to logit, the middle picture to local logit and the right-hand picture to the Klein–Spady estimates. Figure 6 plots the relationship for age. In Figure 5, the local logit estimates (middle picture) show a strongly positive relationship between education and the employment effect of a child. The effect turns positive at about 7 years of education with respect to an older child and about 12 years of education with respect to a younger child. A somewhat similar relationship emerges for the parametric logit estimator (left-hand picture), but the pattern is less stringent and the effects never become positive. For the Klein–Spady estimator (right-hand picture), the relationship is more noisy and the effects seem first to decrease and then to increase with higher educational levels, turning positive at very low and very high educational levels. This pattern, however, seems to be purely spurious and due to the large variability of

[37] The correlation between the effects estimated by local logit and by Klein–Spady is about 0.1.

[38] This is the average of the estimated effects among those of the 2339 women who have the corresponding educational level. Average effects for educational levels with less than 20 observations are not displayed.

the Klein–Spady estimator. This becomes apparent from the relationship between the effect of a child and age, shown in Figure 6. In this figure, the Klein–Spady estimates exhibit a very erratic behaviour, whereas the local logit estimates display a weak downward trend for an older and a weak upward trend for a younger child. The parametric logit effects are invariant to the age level.

Taken together, both local logit and Klein–Spady estimated employment effects of children that are negative for some women but positive for others. However, whereas local logit discerns a regular pattern between educational attainment and the effects of children on labour force participation, the heterogeneity in the Klein–Spady estimates appears to reflect only its higher variability. These conclusions remain unchanged even with different bandwidth choices, $(h, \delta) = (1.5, 0.7)$ or $(1.5, 0.3)$ for local logit and $h = 0.2$ or 0.3 for Klein–Spady.

5. CONCLUSIONS

Defying conventional wisdom, it seems that non-parametric regression can work well even with many regressors if the dependent variable is binary. In the Monte Carlo simulations, local logit regression was about 25 to 55% more precise than parametric logit with respect to mean squared error and median squared error, respectively, with only 250 to 500 observations. In addition, local logit was not much worse than parametric logit in situations where the logit model was correct.

Klein–Spady, Nadaraya–Watson and local linear regression, on the other hand, performed often worse than parametric logit. The weak results of local linear versus local logit regression parallels similar findings for parametric binary choice models, where linear probability models often perform worse than logit or probit models, (see e.g. Hyslop 1999). The local logit specification incorporates several properties that may be appealing in many economic applications (such as decreasing marginal effects and non-saturation). Since in higher-dimensional non-parametric regression, often rather large bandwidth values are selected by cross-validation, an appropriate choice of the local model becomes more relevant. This may explain the poor performance of Nadaraya–Watson regression, which is based on a local constant model.

Local logit regression was then applied to analyze heterogeneity in the effects of children on female labour supply. It was found that highly educated women do not reduce and might even increase their labour force participation with one child (or two older children) compared to no children. This might be due to better access to or higher acceptance of child care outside the home (baby-sitters, boarding schools) among higher educated women, a different division of the child-rearing burden within the family, or other social or psychological reasons. Hence, if economic policy is concerned with facilitating and fostering female labour force participation it should be directed at lower educated women. This heterogeneity in the effects of children was not detected either by parametric logit nor by the semiparametric Klein–Spady estimator.

ACKNOWLEDGMENTS

The author is also affiliated with the Swiss Institute for International Economics and Applied Economic Research (SIAW, St. Gallen), the Institute for the Study of Labor (IZA, Bonn) and the Institute for Labour Market Policy Evaluation (IFAU, Uppsala). I am grateful for discussions and comments to Michael Gerfin, Francois Laisney, Michael Lechner, Ruth Miquel, Winfried Pohlmeier, Patrick Puhani, Jeff Smith, the editor and two anonymous referees. This research was supported by the Marie Curie

Individual Fellowship MEIF-CT-2004-006873, the Swiss National Science Foundation (project NSF 4043-058311) and the Grundlagenforschungsfonds HSG (project G02110112).

REFERENCES

Bebchuk, J. and R. Betensky (2001). Local likelihood analysis of survival data with censored intermediate events. *Journal of American Statistical Association 96*, 449–57.

Blundell, R. and J. Powell (2004). Endogeneity in semiparametric binary response models. *Review of Economic Studies 71*, 655–79.

Carroll, R., D. Ruppert and A. Welsh (1998). Local estimating equations. *Journal of American Statistical Association 93*, 214–27.

Chen, X., O. Linton and I. van Keilegom (2003). Estimation of semiparametric models when the criterion function is not smooth. *Econometrica 71*, 1591–608.

Copas, J. (1995). Local likelihood based on kernel censoring. *Journal of the Royal Statistical Society, Series B 57*, 221–35.

Eguchi, S. and J. Copas (1998). A class of local likelihood methods and near-parametric asymptotics. *Journal of Royal Statistical Society, Series B 60*, 709–24.

Eguchi, S., T. Kim and B. Park (2003). Local likelihood method: a bridge over parametric and nonparametric regression. *Journal of Nonparametric Statistics 15*, 665–83.

Fan, J., M. Farmen and I. Gijbels (1998). Local Maximum Likelihood Estimation and Inference. *Journal of the Royal Statistical Society, Series B 60*, 591–608.

Fan, J. and I. Gijbels (1996). *Local Polynomial Modeling and its Applications*. Chapman and Hall, London.

Fan, J., N. Heckman and M. Wand (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *Journal of the American Statistical Association 90*, 141–50.

Friedman, J. and W. Stuetzle (1981). Projection pursuit regression. *Journal of American Statistical Association 76*, 817–23.

Frölich, M. (2004). Finite sample properties of propensity-score matching and weighting estimators. *The Review of Economics and Statistics 86*, 77–90.

Frölich, M. (2005). Matching estimators and optimal bandwidth choice. *Statistics and Computing 15*, 197–215.

Frölich, M. (2006). Nonparametric IV estimation of local average treatment effects with covariates. *Journal of Econometrics* (in press).

Galindo, C., H. Liang, G. Kauermann and R. Carrol (2001). Bootstrap confidence intervals for local likelihood, local estimating equations and varying coefficient models. *Statistica Sinica 11*, 121–34.

Gerfin, M. (1996). Parametric and semi-parametric estimation of the binary response model of labour market participation. *Journal of Applied Econometrics 11*, 321–39.

Gozalo, P. and O. Linton (2000). Local nonlinear least squares: Using parametric information in nonparametric regression. *Journal of Econometrics 99*, 63–106.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica 66*, 315–31.

Han, A. (1987). Non-parametric analysis of a generalized regression model. *Journal of Econometrics 35*, 303–16.

Härdle, W., M. Müller, S. Sperlich and A. Werwatz (2004). *Nonparametric and Semiparametric Models*. Springer, Heidelberg.

Härdle, W. and T. Stoker (1989). Investigating smooth multiple regression by the method of average derivatives. *Journal of American Statistical Association 84*, 986–95.

Hastie, T. and R. Tibshirani (1990). *Generalized Additive Models*. Chapman and Hall, London.

Heckman, J. (1990). Varieties of selection bias. *American Economic Review, Papers and Proceedings 80*, 313–18.

Heckman, J., H. Ichimura and P. Todd (1998). Matching as an econometric evaluation estimator. *Review of Economic Studies 65*, 261–94.

Heckman, J., J. Smith and N. Clements (1997). Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. *Review of Economic Studies 64*, 487–535.

Hirano, K., G. Imbens and G. Ridder (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica 71*, 1161–89.

Hjort, N. and M. Jones (1996). Locally parametric nonparametric density estimation. *Annals of Statistics 24*, 1619–47.

Holland, P. (1986). Statistics and causal inference. *Journal of American Statistical Association 81*, 945–70.

Horowitz, J. and W. Härdle (1996). Direct semiparametric estimation of single-index models with discrete covariates. *Journal of American Statistical Association 91*, 1632–40.

Hyslop, D. (1999). State dependence, serial correlation and heterogeneity in intertemporal labor force participation of married women. *Econometrica 67*, 1255–94.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics 58*, 71–120.

Imbens, G. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics 86*, 4–29.

Klein, R. and R. Spady (1993). An efficient semiparametric estimator for binary response models. *Econometrica 61*, 387–421.

Kuan, C. and H. White (1994). Artifical neural networks: An econometric perspective. *Econometric Reviews 13*, 1–91.

Loader, C. (1996). Local likelihood density estimation. *Annals of Statistics 24*, 1602–18.

Manski, C. (1993). The selection problem in econometrics and statistics. In. Eds. G. Maddala, C. Rao, and H. Vinod. *Handbook of Statistics*. Elsevier Science Publishers, Amsterdam.

Manski, C. (2000). Identification problems and decisions under ambiguity: Empirical analysis of treatment response and normative analysis of treatment choice. *Journal of Econometrics 95*, 415–42.

Manski, C. (2004). Statistical treatment rules for heterogeneous populations. *Econometrica 72*, 1221–46.

Martins, M. (2001). Parametric and semiparametric estimation of sample selection models: An empirical application to the female labour force in portugal. *Journal of Applied Econometrics 16*, 23–39.

Newey, W. (1994). The asymptotic variance of semiparametric estimators. *Econometrica 62*, 1349–82.

Pagan, A. and A. Ullah (1999). *Nonparametric Econometrics*. Cambridge University Press, Cambridge.

Park, B., W. Kim and M. Jones (2002). On local likelihood density estimation. *Annals of Statistics 30*, 1480–95.

Pearl, J. (2000). *Causality: Models, Reasoning and Inference*. Cambridge University Press, Cambridge.

Powell, J., J. Stock and T. Stoker (1989). Semiparametric estimation of index coefficients. *Econometrica 57*, 1403–30.

Racine, J. and Q. Li (2004). Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics 119*, 99–130.

Robinson, P. (1998). Root-N consistent semiparametric regression. *Econometrica 56*, 931–54.

Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology 66*, 688–701.

Ruppert, D. (1997). Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation. *Journal of American Statistical Association 92*, 1049–62.

Ruppert, D. and M. Wand (1994). Multivariate locally weighted least squares regression. *Annals of Statistics 22*, 1346–70.

Seifert, B. and T. Gasser (1996). Finite-sample variance of local polynomials: Analysis and solutions. *Journal of American Statistical Association 91*, 267–75.

Sherman, R. (1993). The limiting distribution of the maximum rank correlation estimator. *Econometrica 61*, 123–38.

Staniswalis, J. (1989). The kernel estimate of a regression function in likelihood-based models. *Journal of American Statistical Association 84*, 276–83.

Stoker, T. (1991). Equivalence of direct, indirect and slope estimators of average derivatives. In *Nonparametric and semiparametric methods in econometrics and statistics*, pp. 99–118. Eds. W. Barnett, J. Powell and G. Tauchen, Cambridge University Press, Camdridge, UK.

Tibshirani, R. and T. Hastie (1987). Local likelihood estimation. *Journal of American Statistical Association 82*, 559–67.

White, H. (1989). Some asymptotic results for learning in single hidden layer feedforward network models. *Journal of American Statistical Association 84*, 1003–13.

DATA APPENDIX: FEMALE LABOUR SUPPLY

The data contain observations on 2339 married Portuguese women whose spouses were employed in 1991. Descriptive statistics and correlations of the available variables are given in Tables A1 and A2. About 60% of all women were employed and their age ranges from 17 to 59 years. Educational attainment ranges from 0 to 18 years with an average education of 7.2 years. Their log hourly wage rate is observed only for the 1400 employed women and averages 5.8 for them (measured in Portuguese escudos). This variable is not used in the reduced form approach. Husbands' income is in all cases positive and recorded as log *monthly* wage (in escudos). The number of children is subdivided into children up to 3 years old and older children up to 18 years old. On average, each woman has 0.2 younger children and 1.4 older children. More details are found in Martins (2001), from which the data are taken.

Table A1. Descriptive statistics of female labour supply.

Variable	Mean	Stddev.	Min	Max
Employment status	0.60	0.49	0	1
Age in years	38.4	9.4	17	59
Education in years	7.2	3.8	0	18
Wife's log hourly wage	3.5	2.9	0	7.7
Husband's log monthly wage	11.2	0.38	10.3	12.6
Children 0 to 3 years old	0.20	0.44	0	2
Children 4 to 18 years old	1.43	1.11	0	9

Note: 2339 married Portuguese women, of which 1400 employed. Wages measured in Portuguese escudos: log hourly wages for wives, log monthly wages for husbands.

Table A2. Correlation matrix.

	Age	Education	Husb' wage	Children 0−3	Children 4–18
Employment status	−0.16	0.36	0.09	0.03	−0.12
Age in years		−0.14	0.07	−0.41	0.23
Education in years			0.31	0.08	−0.13
Husband's wage				−0.05	−0.02
Children 0 to 3 years old					−0.24