Nonparametric regression for binary dependent variables

Markus Frölich

Department of Economics, University of St.Gallen

Last changes: July 15th, 2004

Abstract

Finite-sample properties of nonparametric regression for binary dependent variables are analyzed. Nonparametric 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. Local logit regression, on the other hand, is 10 to 70% more precise than parametric regression. In an application to female labour supply, local logit finds heterogeneity in the effects of children on employment that is not detected by parametric nor semiparametric estimation.

Keywords: Binary choice, local parametric regression, local model, heterogeneous response, heterogeneous treatment effect

JEL classification: C13, C14, C25

The author is also affiliated with the Institute for the Study of Labor (IZA), Bonn. I am grateful for discussions and comments to Michael Gerfin, Francois Laisney, Michael Lechner, Oliver Linton, Ruth Miquel, Winfried Pohlmeier, Patrick Puhani, Jeff Smith, the editor and two anonymous referees. This research was supported by the Swiss National Science Foundation (project NSF 4043-058311) and the Grundlagenforschungsfonds HSG (project G02110112). Address for correspondence: Markus Frölich, Swiss Institute for International Economics and Applied Economic Research (SIAW), University of St. Gallen, Bodanstrasse 8, CH-9000 St. Gallen, Switzerland; markus.froelich@unisg.ch, www.siaw.unisg.ch/froelich

Nonparametric regression for binary dependent variables

Last changes: July 15th, 2004

Abstract

Finite-sample properties of nonparametric regression for binary dependent variables are analyzed. Nonparametric 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. Local logit regression, on the other hand, is 10 to 70% more precise than parametric regression. In an application to female labour supply, local logit finds heterogeneity in the effects of children on employment that is not detected by parametric nor semiparametric estimation.

Keywords: Binary choice, local parametric regression, local model, heterogeneous response, heterogeneous treatment effect

JEL classification: C13, C14, C25

1 Introduction

In this paper, nonparametric regression for binary dependent variables in finite-samples is analyzed. Binary choice models are of great importance in many economic applications, but nonparametric 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, Smith, and Clements (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 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 must 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 nonparametric 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 nonparametric 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

[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).

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 regression, such as local logit, might therefore be better suited for binary dependent variables than local linear regression. In addition, local 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 the curse of dimensionality but also because of 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 logit regression for binary dependent variables with many regressors (relative to the number of observations), of which some are continuous and some are discrete.[4] Local 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, local logit is often more precise than parametric logit. Even when the logit model is globally true, local 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 $P(Y =$

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

[4] Local likelihood estimation has been introduced by Tibshirani and Hastie (1987). Staniswalis (1989) examined local likelihood estimation with a local constant model (i.e. only one location parameter) for estimating hazard functions. Fan, Heckman, and Wand (1995) analyzed the asymptotic properties of local quasi-likelihood estimation of a local model, defined through a link function. Carroll, Ruppert, and Welsh (1998) proposed local estimating equations and derived its asymptotic properties. Gozalo and Linton (2000) developed the asymptotic distribution theory for least squares local parametric regression. However, the small sample properties of local logit regression with many regressors have not been examined so far.

$1|X$), and somewhat smaller for the estimation of marginal effects.

Local logit regression is then applied to analyze the dependence of Portuguese women's labour supply on the number of children.[5] Both the local logit and the Klein Spady estimator detect heterogeneity in the marginal effects of children that is unnoticed by the parametric logit estimator. However, whereas local logit finds some regular patterns, the apparent effect heterogeneity detected by the Klein Spady estimator seems to be an artefact of its larger variability. 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.[6] The standard approach proceeds by specifying a parametric model, e.g. 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 restrictive assumptions. Most semiparametric estimators, such as the Klein and Spady (1993) estimator,[7] rely on a single index restriction, requiring that the conditional mean can be specified as $E[Y|X=x]=\varphi(x'\theta)$ with φ an *unknown* function with range contained in $[0,1]$ and θ an unknown coefficient vector. Although less restrictive than the parametric models, the single index restriction still implies that all in-

[5] 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 below, 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.

[6] For continuous regressors the marginal effect is often defined as $\partial E[Y|X=x]/\partial x_q$. 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.

[7] The Klein Spady estimator is $\sqrt{n}$-asymptotically normal and attains the semiparametric efficiency bound.

dividuals can be aligned on a single dimension. 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 effect of, e.g., 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.

Nonparametric regression is more flexible. Although it is subject to the curse of dimensionality and usually does not achieve $\sqrt{n}$ convergence, it may still perform well in finite samples. 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 modeling 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 *local logit* estimator

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

is convenient, since it imposes the range restriction and is differentiable. For a further discussion see Fan, Heckman, and Wand (1995).[8] If the logit form is closer to the true regression curve than a constant or linear specification, the local logit estimator 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 would be unbiased, see Gozalo and Linton (2000).[9]

[8] Local logit is preferred to local probit because it requires less computation time.

[9] If the true regression curve is constant, also kernel and local linear regression would be unbiased.

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 nonparametric approach. This however is beyond the scope of this paper.

2.1 Local logit estimation

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 $g(x + \Delta x, \hat{\theta}_x) - g(x, \hat{\theta}_x)$. Several approaches to estimate $\hat{\theta}_x$ have been suggested, including local least squares (Gozalo and Linton 2000), local likelihood (Tibshirani and Hastie 1987) and local estimating equations (Carroll, Ruppert, and Welsh 1998). Local least squares estimates $\hat{\theta}_x$ from a sample of n iid observations $\{(Y_i, X_i)\}_{i=1}^n$ as

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

where $K_H(X_i - x)$ is a kernel function and H a vector of bandwidth values. Local likelihood estimates $\hat{\theta}_x$ as

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

where $\ln L\left(Y_i, g(X_i, \theta_x)\right)$ 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.

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

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

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 $q_1+1, ..., q_2$ 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 as

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

where $X_{q,i}$ and x_q denote the q-th 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 nonparametric 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 3 bandwidth values h, δ, λ for all regressors, a different bandwidth could be employed for each regressor. This would increase substantially 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.

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. 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}

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

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

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

3 Finite sample properties

In this section, the finite sample behaviour of local 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 repeatedly 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 $E[Y|X = X_j]$. 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.[10]

The 4 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 $X_1^b, ..., X_{10}^b$ 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 distributed χ^2 with 1,2,3 and 4 degrees of freedom, respectively. The binary variables $X_1^b, ..., X_{10}^b$ 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: $X_1^b \sim Bernoulli(0.5)$ and $X_k^b \sim Bernoulli(p = 0.3{+}0.4{\cdot}\bar{X}_{k-1}^b)$, 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, 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 to 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.

[10] 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 whose effect shall 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.

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

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

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

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

Y-design 3: Linear index model with interaction terms

$$\begin{aligned} Y &= 1 \quad \text{if } Y^* \geq 0, \text{ where} \\ Y^* &= -8 - X_1^c + 2X_2^c - 3X_3^c + 4X_4^c + 2\sum_{k=1}^{10} X_k^b \cdot (-1)^k \\ &\quad - 3X_1^c X_1^b X_2^b + 3X_2^c X_3^b X_4^b - 3X_3^c X_6^b X_7^b + 3X_4^c X_8^b X_9^b + noise \end{aligned}$$

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

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

Y-design 5: Index model with two-regimes

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

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 neither be attractive 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, e.g. affirmative action programmes, preferential tax treatments, exemptions from social security or pension contributions etc.

Two variants of noise are considered: homoskedastic and heteroskedastic noise. The homoskedastic noise is drawn from a *logistic* distribution. The heteroskedastic noise is drawn from a t_2 distribution and multiplied by $0.14\sqrt{\sum X_k^c \sum X_k^b}$. Hence, for Y-design 1 with 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.[11]

3.1 Implementation of the estimators

All estimators use as regressors $X_1^c, ..., X_4^c, X_1^b, ..., X_{10}^b$ and a constant, but no interaction or higher-order terms. The benchmark parametric logit estimator is estimated by maximum likelihood.[12] The semiparametric Klein Spady estimator is implemented as in Gerfin (1996) with $\varphi(\cdot)$ estimated by one-dimensional kernel regression and the bandwidth selected by generalized cross-validation bandwidth selection.[13]

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

[12] If the parametric logit did not converge (collinearity, perfect prediction), a new sample is drawn.

[13] The bandwidth is chosen from the set of values: $\{0.02, 0.04, ..., 0.60\}$.

For Nadaraya-Watson, local linear and local logit regression, the bandwidths are chosen according to the least-squares criterion CV_{LS} (7) or according to the likelihood criterion CV_{ML} (8).[14] The kernel weights $K(X_i - x)$ for given bandwidth values h, λ are computed as

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

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 $g(x, \theta_x)$. Since the local linear estimates may lie outside the interval $[0, 1]$, they are capped at zero and at one. 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.

With local linear and local logit regression it can happen that for small bandwidth values the estimate of the coefficients θ_x is undefined at some locations x due to *local multicollinearity*. This occurs particularly with a compact kernel, such as the Epanechnikov kernel. Nevertheless also for the Gaussian kernel near-multicollinearity can render the estimate undefined (due to numerical inaccuracies in matrix inversion). Three procedures to cope with such situations are examined.

Variant 1: If the estimate of θ_x is undefined, the estimate of the conditional mean $\hat{E}[Y|X = x]$ is simply set to the (unconditional) mean of the full sample: $\frac{1}{n}\sum Y_i$.

Variant 2: If the number of observations in the local neighbourhood is smaller than the number of regressors (which are 14 plus the constant), all bandwidths are locally increased by 10% repeatedly until the local neighbourhood includes at least 15 observations. Then all regressors that cause local multicollinearity are dropped until a valid estimate of θ_x is obtained. For detecting (nearly) linear dependencies in the regressor matrix, the pivotal orthogonal-triangular (QR) decomposition is used, see Judd (1998, p. 58 f) or Press, Flannery, Teukolsky, and Vetterling (1986, p. 357 ff). This decomposition decompounds a regressor or moment matrix into an orthogonal matrix Q and an upper triangular matrix R, where diagonal elements of

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

R that are close to zero indicate (nearly) linear dependencies attributable to the corresponding columns. All regressors associated with a diagonal element in R smaller than 10^{-5} are dropped in the local regression.[15]

Variant 3: All bandwidths are locally increased by 10% repeatedly until a valid estimate of θ_x is obtained.

Variant 1 attempts to penalize small bandwidth values by using an uninformed estimate, which will lead to the selection of larger bandwidth values in the cross-validation routine. Variant 2 reduces the complexity of the local model,[16] whereas Variant 3 reduces the localization of the model. Variant 3 is similar to a proposal in Seifert and Gasser (1996) to locally increasing the bandwidth, but different in spirit. Whereas the main motivation for the above procedures is to obtain well-defined estimates in situations of almost exact multicollinearity, Seifert and Gasser (1996) rather pursue a stabilization of the variance of the local linear estimator. For an unbounded outcome variable Y, they show that the unconditional variance of local linear regression is infinite and the conditional variance unbounded, because the estimators' denominator can be arbitrarily close to zero. To bound the variance of the estimator, they suggest to increase the bandwidth locally or to use ridge regression. Alternatively, using Gaussian weights instead of Epanechnikov weights can help to stabilize the variance of local linear regression. For a binary dependent variable Y, however, the variance of the estimator is bounded and variance reduction is therefore of a lesser concern (and is left for future research).[17]

Local logit regression with Epanechnikov kernel is examined for variants 1, 2 and 3. In addition, local logit with Gaussian kernel is analyzed for variant 3. Notice that with variant 3, the marginal effects of local logit regression are computed from within the local model. For

[15] Different threshold values have been tried and did not affect the results very much. 10^{-5} is a rather conservative choice, in the sense that when in doubt about near-linear dependencies rather more than less regressors are dropped, to spare local degrees of freedom for estimating the remaining coefficients.

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

[17] In a similar spirit, ridge or shrinkage regression attempts to reduce variance at the cost of a larger bias by shrinking the coefficient estimates towards zero, see Stein (1981), Judge, Hill, Griffiths, Lütkepohl, and Lee (1982, p. 878 ff), Seifert and Gasser (1996) and Mittelhammer, Judge, and Miller (2000, Chapter 18.7). However, because the coefficients themselves are not of interest here, dropping some of the regressors as in Variant 2 appears more promising than deliberately biasing the estimates.

variants 1 and 2, this is not always possible since the local model may no longer include all regressors (e.g. because of the elimination of linearly dependent regressors). With variants 1 and 2, marginal effects are obtained by estimating separately the conditional means at x and at $x + \Delta x$. For local linear regression only variant 1 with Epanechnikov kernel is examined.

3.2 Simulation results

The out-of-sample prediction performance for the conditional mean $E[Y|X]$ and for the marginal effects are assessed by their simulated mean absolute error, median absolute error, mean squared error and median squared error. The performance is measured *relative* to the benchmark parametric logit estimator. (The results are always given in percent. Hence, numbers below 100 indicate an improvement over parametric logit regression, whereas numbers above indicate a worse performance.) Among these four error measures, generally the relative performance of local logit is worst with respect to mean squared error (MSE) and best with respect to the median squared error (MdSE). The relative performance with respect to the two other error measures is usually between these two. Therefore, in the following, only the results for the MSE and the MdSE are given.

The results are summarized through box-plots in Figures 3.1 and 3.2 and in Table 3.1. Detailed simulation results for the conditional mean predictions are given in Tables A.1 to A.3 in the appendix. Figure 3.1 illustrates the relative performance of Klein Spady, Nadaraya-Watson and local linear regression. The first row refers to sample size 250 with logistic noise, the second row contains the results for sample size 250 with heteroskedastic noise, and the last row for sample size 500 with heteroskedastic noise.[18] The first figure in each row gives the results for Klein Spady regression (6 box-plots), the second figure for Nadaraya-Watson regression (12 box-plots) and the third figure for local linear regression (12 box-plots).

For Klein Spady regression, the first box-plot gives the distribution of the relative MSE of Klein Spady in predicting the conditional mean $E[Y|X]$ over the 20 different simulation designs (4 X-designs times 5 Y-designs). For the 20 different designs, the MSE was simulated and the 20 values themselves are given in Tables A.1 to A.3. It can be seen that the relative MSE for the different designs is between 100% and 250%, i.e. the Klein Spady estimator performed in

[18] Results for sample size 500 with logistic noise not shown.

all designs worse than parametric logit regression. The second box-plot summarizes the results for the marginal effects of the 4 continuous regressors and the third box-plot represents the results for the marginal effects of the 10 binary regressors. The second box-plot is based on 80 values (4 continuous regressors times 20 designs) and the third box-plot is based on 200 values (10 binary regressors times 20 designs). For both the continuous and the discrete regressors, the relative performance is much worse and the median over all designs is about 300%, i.e. the MSE of Klein Spady regression is 3 times as large as the MSE of parametric logit regression. The following box-plots 4 to 6 are analogous to the first three box-plots but refer to the MdSE instead. The results are very similar to those for the MSE.

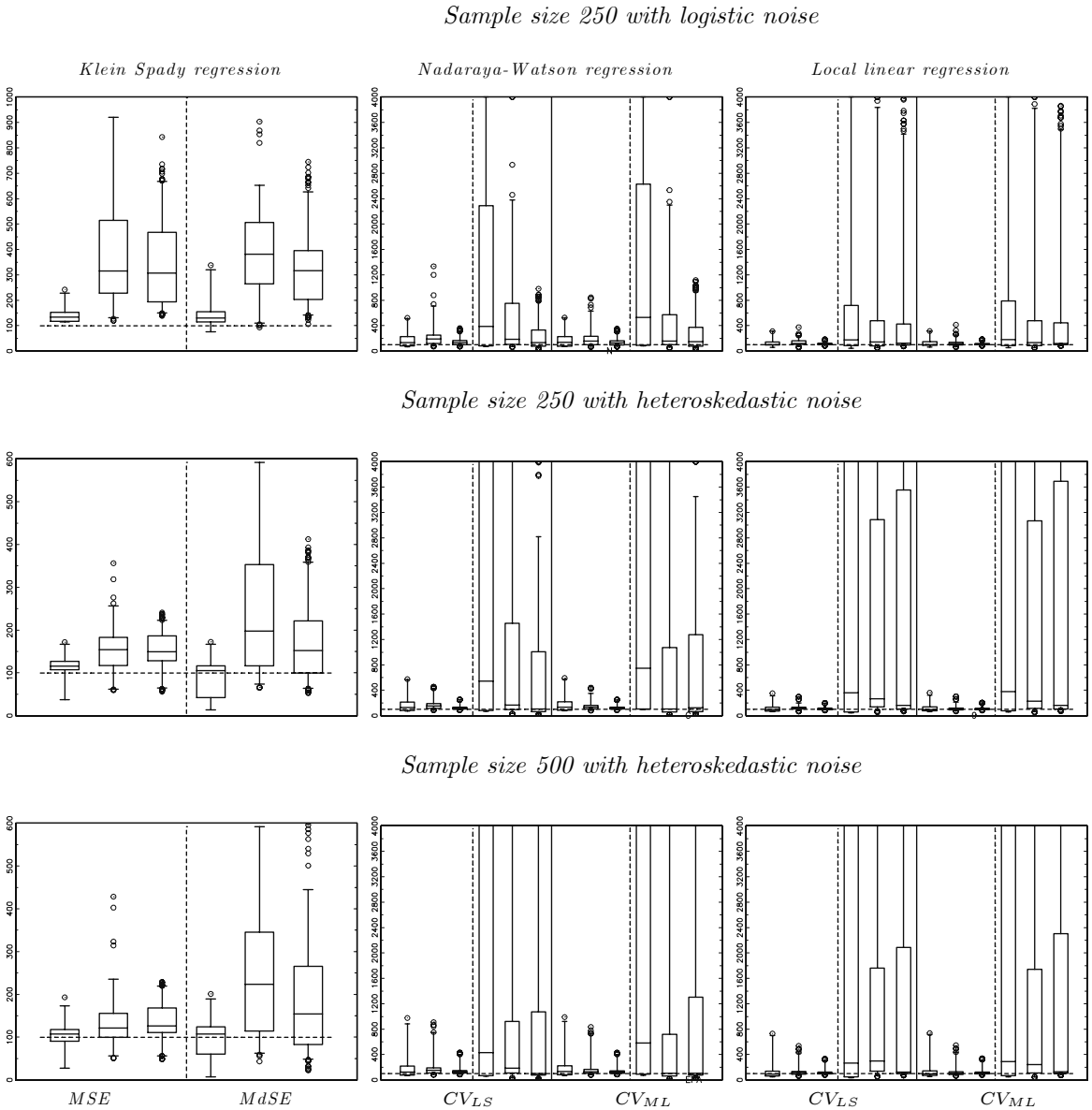
For heteroskedastic noise and for a larger sample size (graphs below), the relative performance of Klein Spady regression improves somewhat. Nevertheless, it still performs worse than parametric logit in most designs.

The second figure in each row illustrates the results for Nadaraya-Watson regression, which are given through 12 box-plots: The first 6 box-plots refer to Nadaraya-Watson regression with bandwidth choice based on the least squares CV_{LS} criterion, the second 6 box-plots refer to bandwidth choice based on CV_{ML}. In all other respects the box-plots are identically defined as for the Klein Spady estimator. With respect to MSE, the relative performance of Nadaraya-Watson is sometimes below 100%, but in most designs it behaves worse than parametric logit. Its MSE can be up to ten times larger than for parametric logit, even with 500 observations. In terms of MdSE, however, it can behave even worse. With heteroskedastic noise, for example, in more than one fourth of the 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 20 designs, the MdSE is about 400. These findings are similar, but somewhat less extreme, with respect to median absolute error (not shown). 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 behaves slightly better than Nadaraya-Watson and its MSE varies less with the simulation design. In terms of MdSE, however, it can be even more variable when estimating marginal effects. The finding that local linear performs extremely badly with respect to MdSE but somewhat less bad in terms of MSE, indicates that local linear regression produces disproportion-

ately 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.

Figure 3.1: Out-of-sample prediction performance of Klein Spady, Nadaraya-Watson and local linear regression



Note: Distribution over the 20 different designs of the out-of-sample prediction performance (relative to parametric logit). First row: n=250 with logistic noise. Second row: n=250 with heteroskedastic noise. Third row: n=500 with heteroskedastic noise. Results for Klein Spady (left), for Nadaraya-Watson (middle) and local linear regression (right). First six box-plots refer to CV_{LS}-based bandwidth choice, second six box-plots refer to CV_{ML}-based bandwidth choice (except for Klein Spady). 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 4 continuous regressors and the third box-plot for the marginal effects of the 10 binary regressors. The box-plots 4 to 6 give the analogous results for the MdSE. The boxes represent the median and the 25 and 75 percentiles and extend to the 5 and 95 percentiles. Results are capped at 4000.

Figure 3.2 summarizes analogously the results for 4 variants of local logit regression. The first three graphs in each row refer to local logit regression with Epanechnikov kernel and variants 1, 2 and 3, respectively, for handling local multicollinearity. The fourth graph represents local logit regression with Gaussian kernel. The results are generally more favourable than they were for Klein Spady, Nadaraya-Watson and local linear regression.

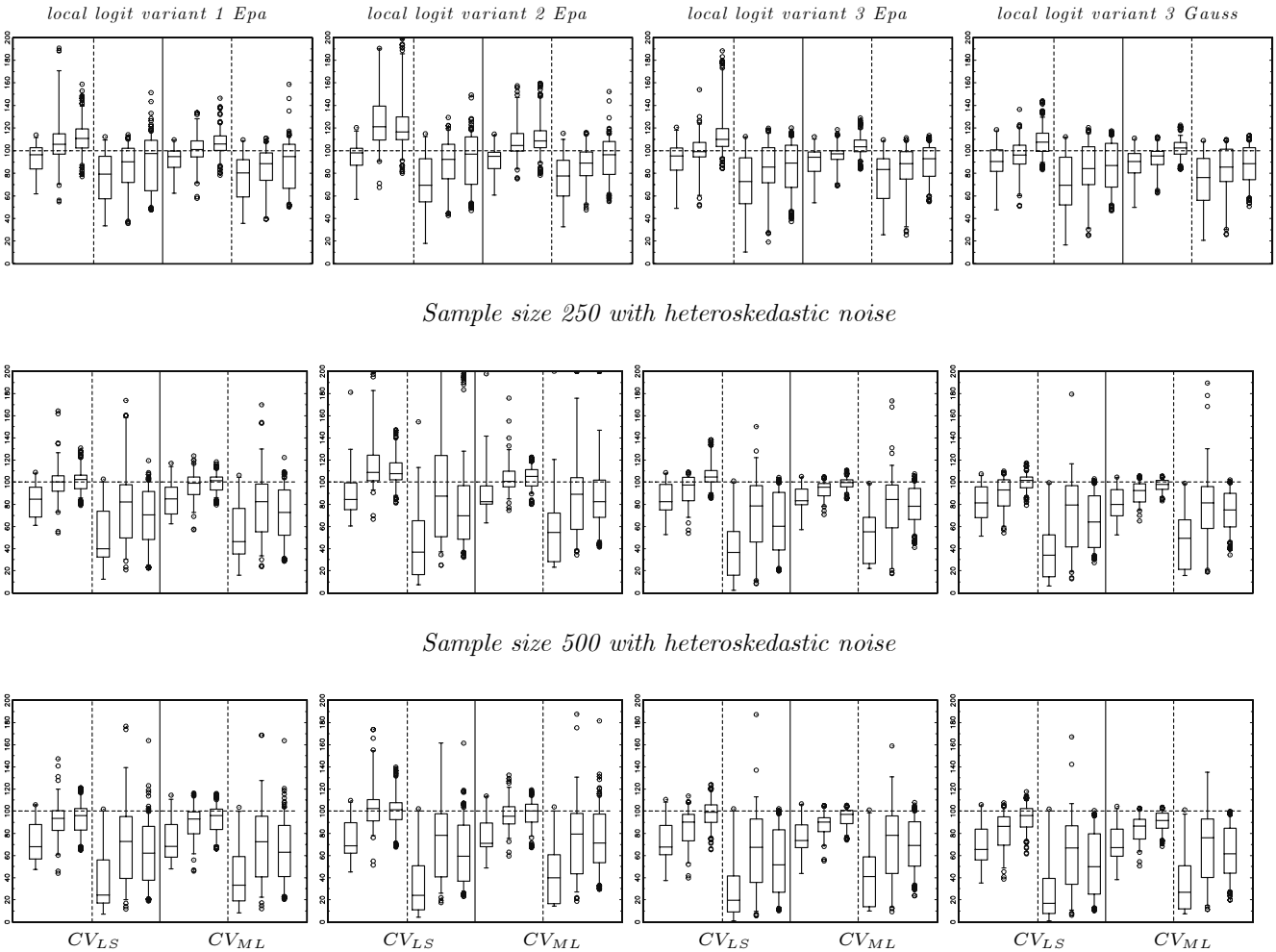
Examining first the performance in predicting the *conditional mean* $E[Y|X]$, which is given by the first, fourth, seventh and tenth box-plot in each graph. Even in the most difficult situation, 250 observations with logistic noise, which favours the parametric logit estimator since it is correctly specified in 4 out of 20 designs, the MSE is in most designs below 100% and the MdSE is usually by 20 to 30% lower than for parametric logit. Of the different variants of the estimator, variant 3 with Epanechnikov or Gaussian kernel and cross-validation criterion CV_{LS} seems to be somewhat superior, but the differences are not very pronounced. With heteroskedastic noise, the relative performance of local logit clearly improves. With Gaussian kernel and variant 3, the MSE is around 20% lower and the MdSE around 60% lower than for parametric logit. For sample size 500, the reduction in MSE is around 35% and in MdSE around 80%. However, local logit does not strictly dominate the parametric logit estimator because its relative performance is worse than 100% in about 2 or 3 of the 20 simulation designs. Nevertheless, the efficiency loss seems to be rather small in particular with the Gaussian kernel, which in the worst case had a 6% (2%) higher MSE (MdSE). The detailed results for the 20 simulation designs are given in Tables A.1 to A.3.

Turning now to the results for the *marginal effects*, which are given by the box-plots 2, 3, 5, 6, 8, 9, 11 and 12 in each graph. A general finding is that the relative precision in predicting marginal effects is worse than for predicting conditional means $E[Y|X]$. This could be related to the well known fact that nonparametric estimation of derivatives is more difficult than estimation of the mean.[19] Among the different variants of local logit regression, variant 2 performs somewhat worse and variant 3 somewhat better. In variant 2, marginal effects are estimated as the difference of two conditional means and in obtaining these two estimates, different regressors might be dropped due to near-collinearity. This might lead to a higher variance, particularly if the regressor, for which the effect is defined, itself is dropped in one

[19] I.e. the convergence rate is lower for derivative estimation than for mean estimation.

but not the other of these two estimates. For variant 3, the type of kernel does not seem to matter much, though the Epanechnikov kernel generates a few more outliers, particularly with n=250 and logistic noise. Using the CV_{ML} criterion leads to less dispersion in the relative performance, whereas CV_{LS} leads to an overall lower MdSE. With sample size 250 and logistic noise, the MdSE of local logit is about 15% lower than for parametric logit, whereas its MSE is of similar size. With n=500, precision gains are around 10% for the MSE and 40% for MdSE.

Figure 3.2: Out-of-sample prediction performance of local logit regression (relative to parametric logit)



Note: Distribution over the 20 different designs of the out-of-sample prediction performance. Local logit regression with Epanechnikov kernel and variants 1, 2 and 3, respectively, and local logit regression with Gaussian kernel and variant 3. See also note below Figure 3.1.

These results are further summarized in Table 3.1, which gives the *average* performance

over the 20 simulation designs, i.e. it gives the average loss if all designs are given equal weights. Since for sample size 250 with logistic noise, the parametric logit model is exactly correct in 4 out of 20 designs, the incidence of correct parametric specification is given implicitly a large weight in this average. The first rows give the average of the MSE and the MdSE, respectively, when predicting the conditional mean $E[Y|X]$. The subsequent rows refer to the estimation of the marginal effects. In each block, the row beginning with 250 L contains the results for sample size 250 with logistic noise, whereas 250 H refers to heteroskedastic noise. To ease the reading of the table, the smallest values of MSE and MdSE in each row are marked in bold.

For predicting the conditional mean $E[Y|X]$, the different variants do not differ very much in their performance. Variant 3 with bandwidth choice based on CV_{LS} seems to perform best. For the estimation of marginal effects, variant 2 sometimes performs poorly and variant 3 with CV_{LS} is again most favourable.

Table 3.1: Average prediction performance of local logit regression (relative to parametric logit)

	Variant 1 Epa				Variant 2 Epa				Variant 3 Epa				Variant 3 Gauss			
	CV_{LS}		CV_{ML}		CV_{LS}		CV_{ML}		CV_{LS}		CV_{ML}		CV_{LS}		CV_{ML}	
	M	md	M	md	M	md	M	md	M	md	M	md	M	md	M	md
Prediction performance for conditional mean																
250 L	93	77	**91**	78	95	**73**	93	78	92	**71**	**91**	77	**90**	**70**	**88**	**73**
250 H	86	51	87	56	92	48	94	64	**84**	**42**	86	58	**81**	**41**	**82**	51
500 H	76	41	77	44	77	37	79	47	74	**34**	77	47	**70**	**32**	**72**	41
Prediction performance for marginal effects of continuous regressors																
250 L	110	**84**	101	**84**	129	89	108	88	99	**82**	**95**	**84**	**95**	**82**	**93**	**83**
250 H	100	81	97	81	119	101	104	101	94	**75**	**93**	80	**90**	**76**	**91**	81
500 H	92	75	90	75	105	81	96	79	**86**	**68**	88	78	**83**	**65**	**84**	78
Prediction performance for marginal effects of binary regressors																
250 L	112	90	106	**88**	125	92	112	93	117	**86**	**104**	**89**	109	**86**	**102**	**88**
250 H	102	69	**99**	72	111	78	104	96	107	**63**	**99**	79	**100**	**66**	**97**	75
500 H	**93**	62	**93**	65	101	64	98	75	98	**54**	95	69	**94**	**52**	**91**	64

Note: Columns labelled M refer to MSE and those labelled md refer to MdSE. The rows 250 L refer to sample size 250 with logistic noise, whereas 250 H refers to sample size 250 with heteroskedastic noise. In each row, the minimum value of MSE and MdSE and all values not exceeding it by more than 3 are marked in bold.

When using variant 3 with Gaussian kernel and CV_{LS}-based bandwidth choice, the precision in estimating $E[Y|X]$ is improved vis-a-vis parametric logit by 10-30% and 30-70% with

respect to MSE and MdSE, respectively. For the estimation of the marginal effects for the continuous regressors, the precision gains are in the order of 5-15% and 20-35%. For the marginal effects of the discrete regressors, however, precision gains with respect to MSE only materialize with sample size 500. For sample size 250, local logit is equal or somewhat worse than parametric logit. In terms of MdSE, nevertheless, the precision gains are around 15-50%.

4 Heterogeneous female labour supply

The previous section indicated that local logit regression 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 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. must be targeted more precisely than if it were largely homogeneous. Therefore, in the analysis of female labour supply, not only mean effects should be estimated but also their distribution.

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 is taken from Martins (2001) and consists 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 B.1 and B.2 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.[20]

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 5 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 5 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 $h_{age} = h \cdot Std(age)$, $h_{education} = h \cdot Std(education)$ and $h_{wage} = h \cdot Std(wage)$, 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 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

[20] Notice 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) or Manski (1993). This would change the interpretation of the estimated effects but not the comparison of the different estimators' ability in detecting heterogeneous effects.

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. Using least squares cross-validation CV_{LS} (7), the two bandwidths h, δ were chosen as (h, δ)=(3.21, 0.35).[21]

The coefficient estimates of the parametric logit and the semiparametric Klein Spady estimator are given in Table 4.1. 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.

Table 4.1: Logit and Klein-Spady estimated coefficients

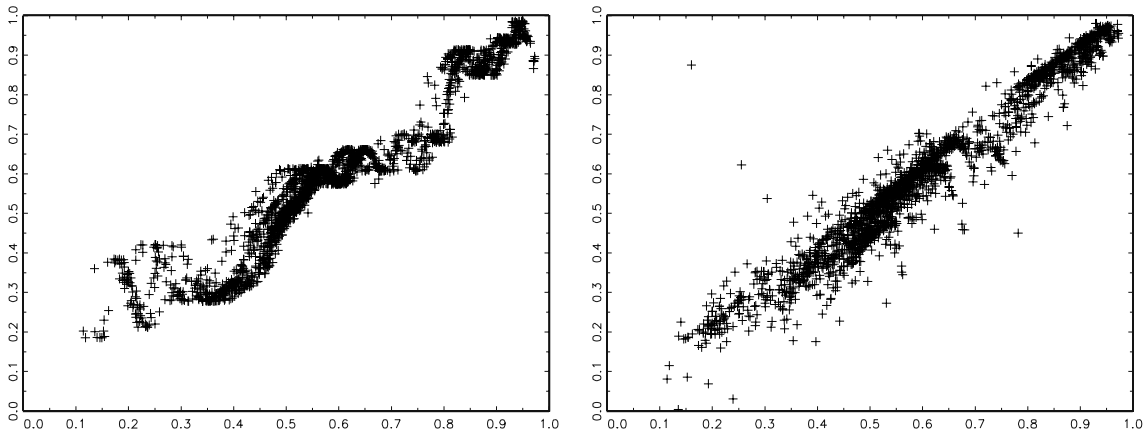
Estimated coefficients	Logit	Klein Spady
Dependent variable: Employment		
Constant	1.34	
Age in years	-0.03^{***}	-0.05
Education in years	0.24^{***}	0.28^{***}
Husband's log wage	-0.10	-0.08
Children 0 to 3 years old	-0.39^{***}	-0.36^{***}
Children 4 to 18 years old	-0.13^{***}	-0.17^{***}

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.

Figure 4.1 shows the estimated employment probabilities of logit versus Klein Spady (left picture) and logit versus local logit (right picture). 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 $h = 3.21$.

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

Figure 4.1: 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) or local logit (right picture); 2339 observations.

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.[22] For the estimation of the children effects, different family compositions are considered and compared to the base category zero children. Table 4.2 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 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 4.2. The effects estimated by parametric logit are not spread

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

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.

Table 4.2: Distribution of marginal effects on employment probability (in %-points)

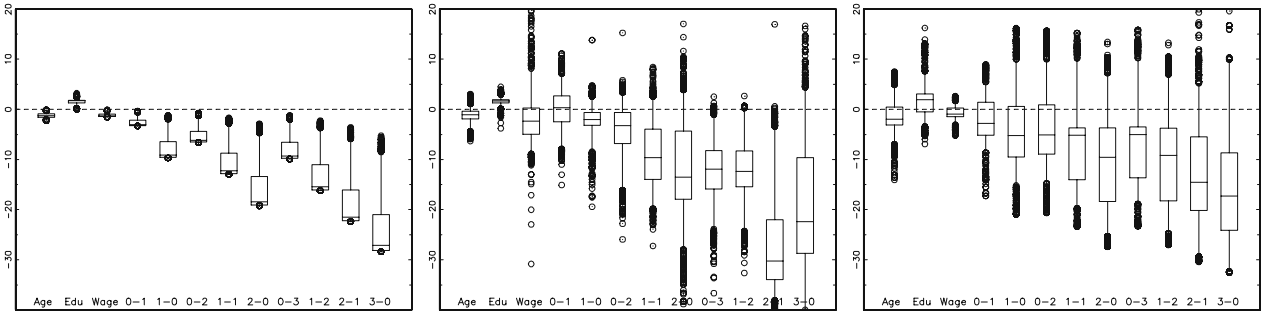
	Age	Education	Husb's wage	1 old child	1 young child	2 older children	1 old & 1 young	2 young children	3 older children	2 old & 1 young	2 young & 1 old	3 young children
						Logit						
Mean	-1.2	1.6	-1.2	-2.6	-7.9	-5.3	-10.6	-16.0	-8.0	-13.4	-18.7	-24.0
$Q_{0.05}$	-2.0	0.3	-1.5	-3.3	-9.7	-6.6	-13.0	-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.

This is different with the semi- and nonparametric 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 2 children. (For the Klein Spady estimator this holds even for 3 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 4.2. The left picture gives the results according to parametric logit, the middle picture for local logit and the right picture 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 4.2 and Table 4.2 that the Klein Spady estimates are generally the most variable. Their interquantile ranges and their standard deviations (not shown in Table 4.2) 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 that effect heterogeneity is even more pronounced.

Figure 4.2: Distribution of marginal effects in the population



Note: Distribution of estimated marginal effects according to parametric logit (left), local logit (middle) and Klein Spady (right picture). 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 4.2. The boxes represent the median and the 25 and 75 percentiles (see $Q_{0.25}$ and $Q_{0.75}$ in Table 4.2), 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.

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 4.3, where the upper part is based on the local logit estimates, whereas the lower part refers to the Klein Spady estimates. The columns labelled by + give the average characteristics of those women for whom the respective effect of children is positive, whereas the columns labelled by - gives 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 4.3, 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 heterogeneity detected by local logit seems to represent an authentic pattern in that the reaction to children varies with the educational level.[23]

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

Table 4.3: Comparison of women with positive versus negative employment effects of children

	1 old child		1 young child		2 old children		1 old & 1 young		2 young children		3 old children		2 old & 1 young		2 young & 1 old		3 young children	
	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-
Local logit																		
#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																		
#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 (#obs) 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.

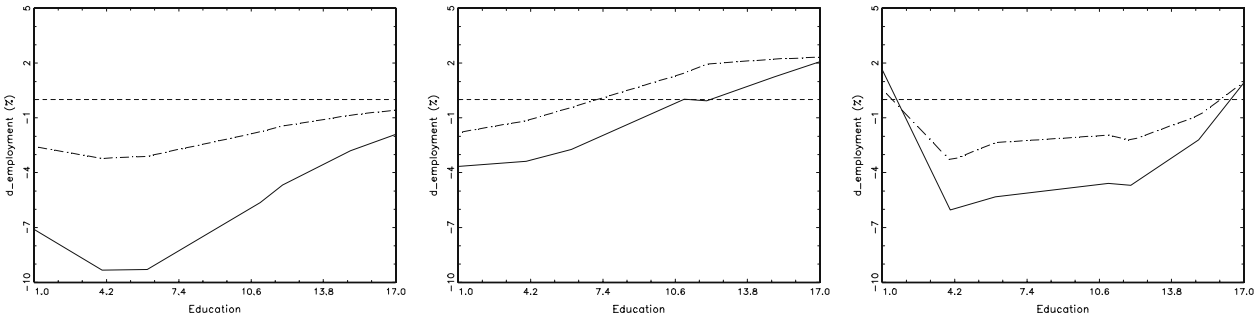
This finding is corroborated by Figures 4.3 and 4.4, which examine the correlation between the effects of children and education (or age, respectively). Figure 4.3 plots the estimated employment effect of one older child (dashed line) and of one younger child (solid line) for different educational levels.[24] The left picture corresponds to logit, the middle picture to local logit and the right picture to the Klein Spady estimates. Figure 4.4 plots the relationship for age. In Figure 4.3, 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 picture), but the pattern is less stringent and the effects never become positive. For the Klein Spady estimator (right 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

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

purely spurious and due to the large variability of the Klein Spady estimator. This becomes apparent from the relationship between the effect of a child and age, shown in Figure 4.4. 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.

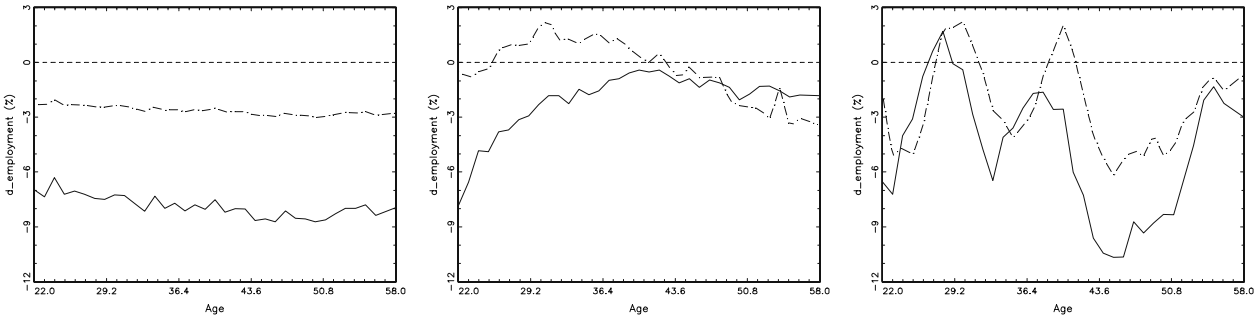
Taken together, both local logit and Klein Spady estimated employment effects of children that are negative for some women but positive for others. However, whereas local logit discerns a regular pattern between educational attainment and the effects of children on labour force participation, the heterogeneity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

Figure 4.3: Employment effect of one child conditional on educational level



Note: 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), local logit (middle) and Klein Spady (right picture), for different educational levels. Effects for educational levels with less than 20 observations not displayed.

Figure 4.4: Employment effect of one child conditional on age



5 Conclusions

Defying conventional wisdom, it seems that nonparametric regression can work well even with many regressors if the dependent variable is binary. With only 250 or 500 observations, local logit regression was about 10-30% (30-70%) more precise than parametric logit with respect to mean squared error (median squared error). In addition, local logit was not much worse than parametric logit in situations where the logit model was correct. When estimating marginal effects, the precision gains of local logit are smaller and around 5-15% (20-50%) in terms of mean (median) squared error.

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 nonparametric 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 by parametric logit nor by the semiparametric Klein Spady estimator.

Table A.1: Out-of-sample prediction performance for the conditional mean (relative to parametric logit in %), n=250 with logistic noise

		Klein Spady		Nadaraya-Watson (Epa)				local linear (Epa)				local logit variant 1 Epa				local logit variant 2 Epa				local logit variant 3 Epa				local logit variant 3 Gauss			
				CV_{LS}		CV_{ML}		CV_{LS}		CV_{ML}		CV_{LS}		CV_{ML}		CV_{LS}		CV_{ML}		CV_{LS}		CV_{ML}		CV_{LS}		CV_{ML}	
X	Y	MSE	MdSE	MSE	MdSE	MSE	MdSE	MSE	MdSE	MSE	MdSE	MSE	MdSE	MSE	MdSE	MSE	MdSE	MSE	MdSE	MSE	MdSE	MSE	MdSE	MSE	MdSE	MSE	MdSE
1	1	195	250	442	∞	491	∞	312	∞	314	∞	114	112	110	110	115	110	109	105	118	107	106	104	119	112	105	102
2		209	272	435	∞	461	∞	312	∞	316	∞	108	108	107	106	117	115	114	115	121	113	110	110	117	108	110	108
3		242	338	512	2382	520	2649	218	781	223	855	112	109	110	109	120	112	115	109	117	112	112	108	115	112	111	109
4		228	320	523	2291	527	2628	213	717	220	791	109	108	107	108	113	112	111	110	114	106	109	106	111	105	107	105
1	2	116	115	73	87	75	100	59	47	62	57	63	34	62	36	58	20	62	33	49	11	54	26	48	17	50	21
2		116	116	69	72	71	85	58	44	61	51	62	34	62	37	57	18	61	33	49	13	55	27	48	19	50	23
3		128	130	109	178	110	193	64	61	67	68	67	35	68	38	72	29	75	49	60	21	71	46	57	22	65	39
4		125	124	108	146	109	170	64	51	66	59	66	35	66	38	69	29	74	48	60	20	71	48	56	22	63	36
1	3	139	75	216	8130	228	∞	143	5838	146	5784	96	69	96	76	98	63	95	69	96	63	94	73	96	61	94	70
2		141	97	207	9769	220	∞	146	8162	145	8287	96	64	96	72	100	58	96	67	99	62	96	74	98	58	95	71
3		131	145	168	660	172	854	97	278	97	286	84	57	85	59	87	55	84	60	83	53	82	60	82	52	80	56
4		134	153	178	719	183	938	100	296	102	326	86	62	87	65	88	60	86	62	88	56	85	58	86	53	84	57
1	4	115	96	81	82	81	90	97	88	91	88	96	90	91	87	97	85	93	86	94	86	92	87	88	78	86	78
2		117	99	81	84	78	92	95	85	92	86	97	88	92	85	97	85	94	87	94	84	94	87	90	79	89	79
3		150	156	80	81	75	83	103	91	98	90	103	95	97	90	102	92	97	91	102	91	99	92	101	90	97	91
4		149	146	78	81	74	83	106	95	99	93	105	94	99	91	104	91	99	92	103	94	99	93	103	94	99	93
1	5	124	125	136	387	141	530	98	179	99	195	95	79	94	80	97	66	94	78	95	65	93	78	89	59	89	68
2		118	127	128	366	135	501	96	190	100	201	95	77	95	80	98	69	95	77	95	73	95	83	90	69	90	76
3		155	155	229	564	230	637	110	179	113	183	100	96	100	95	101	93	99	93	102	96	99	96	100	95	97	95
4		152	145	232	487	230	537	110	156	111	158	99	92	100	92	101	93	98	91	99	90	97	89	100	90	97	89

Note: The first two columns denote the X and the Y-design. In the following columns the mean squared error (MSE) and median squared error (MdSE) for predicting the conditional mean E[Y|X] is given for the various estimators. The values are relative to the corresponding error measure of parametric logit (in %). Values below 100% indicate a better prediction performance, whereas values above indicate a worse performance. Values larger 10'000 are replaced by ∞. 100 replications.

Table A.2: Out-of-sample prediction performance for the conditional mean (relative to parametric logit in %), n=250 with heteroskedastic noise

		Klein Spady		Nadaraya-Watson (Epa)				local linear (Epa)				local logit variant 1 Epa				local logit variant 2 Epa				local logit variant 3 Epa				local logit variant 3 Gauss			
				cvLS		cvML		cvLS		cvML		cvLS		cvML		cvLS		cvML		cvLS		cvML		cvLS		cvML	
X	Y	MSE	MdSE	MSE	MdSE	MSE	MdSE	MSE	MdSE	MSE	MdSE	MSE	MdSE	MSE	MdSE	MSE	MdSE	MSE	MdSE	MSE	MdSE	MSE	MdSE	MSE	MdSE	MSE	MdSE
1	1	143	136	448	∞	482	∞	352	∞	359	∞	107	100	114	104	130	113	142	121	108	99	105	99	106	100	104	99
2		127	131	377	∞	402	∞	312	∞	315	∞	108	103	117	106	181	154	198	210	106	99	103	99	104	99	102	99
3		172	167	574	∞	593	∞	311	∞	323	∞	105	100	106	100	111	100	114	102	109	101	104	99	108	99	105	99
4		167	173	569	∞	574	∞	304	∞	313	∞	109	96	110	98	107	93	111	95	103	92	104	91	103	92	102	92
1	2	115	117	76	90	77	102	61	55	64	60	66	33	66	35	60	11	63	26	54	4	59	24	52	15	52	16
2		112	110	73	80	75	95	60	50	63	58	63	32	64	36	61	11	64	31	53	4	57	22	51	16	53	19
3		116	114	105	201	105	226	66	98	68	101	65	14	65	17	68	9	72	28	57	3	69	26	55	6	64	21
4		113	110	104	160	106	195	62	71	64	76	61	13	62	16	65	7	70	28	54	2	67	28	51	6	59	20
1	3	127	42	185	∞	196	∞	136	∞	137	∞	96	44	96	53	97	41	95	49	97	37	93	51	95	34	93	47
2		126	51	192	∞	199	∞	141	∞	141	∞	94	40	94	46	97	43	97	55	98	36	94	51	96	35	93	50
3		120	106	157	1337	158	1641	97	709	98	710	82	21	82	25	84	17	82	23	82	16	81	26	80	14	79	21
4		118	93	162	1576	166	2051	102	812	103	824	85	25	85	27	84	17	82	23	82	16	81	26	81	12	80	20
1	4	37	14	80	69	83	95	77	59	80	78	77	50	77	58	79	44	80	65	75	42	80	65	68	37	70	49
2		43	18	81	75	85	104	76	61	82	79	79	54	79	61	79	46	81	65	77	45	82	68	70	38	72	51
3		65	35	83	76	88	110	71	43	79	67	69	33	71	43	75	35	80	60	75	38	82	65	69	34	76	56
4		66	35	84	75	87	105	72	43	80	69	71	35	73	46	77	37	82	63	77	38	83	66	71	35	76	54
1	5	112	87	130	545	136	748	96	318	98	344	94	44	93	46	93	26	91	50	90	23	91	55	84	20	84	34
2		107	75	128	528	135	744	100	358	101	376	93	40	92	43	93	26	92	48	90	25	92	52	85	18	85	32
3		131	115	215	5111	218	6125	117	2182	119	2198	95	77	95	80	100	65	96	73	98	55	93	70	95	52	91	66
4		130	121	230	6738	233	8011	123	2890	123	2854	97	74	97	76	99	67	97	72	97	60	94	68	96	58	94	67

See note below Table A.1

Table A.3: Out-of-sample prediction performance for the conditional mean (relative to parametric logit in %), n=500 with heteroskedastic noise

		Klein Spady		Nadaraya-Watson (Epa)				local linear (Epa)				local logit variant 1 Epa				local logit variant 2 Epa				local logit variant 3 Epa				local logit variant 3 Gauss			
				cvLS		cvML		cvLS		cvML		cvLS		cvML		cvLS		cvML		cvLS		cvML		cvLS		cvML	
X	Y	MSE	MdSE	MSE	MdSE	MSE	MdSE	MSE	MdSE	MSE	MdSE	MSE	MdSE	MSE	MdSE	MSE	MdSE	MSE	MdSE	MSE	MdSE	MSE	MdSE	MSE	MdSE	MSE	MdSE
1	1	173	167	864	∞	920	∞	730	∞	736	∞	106	102	111	103	110	102	114	104	111	102	106	101	104	102	102	101
2		164	154	821	∞	876	∞	700	∞	711	∞	106	100	114	100	108	100	113	100	108	100	107	98	106	100	104	97
3		193	201	976	∞	991	∞	532	∞	541	∞	105	97	110	97	107	98	112	98	108	97	105	97	105	95	101	94
4		167	189	883	∞	899	∞	472	∞	487	∞	103	94	108	93	106	92	107	90	103	92	103	89	101	89	101	87
1	2	107	112	67	69	69	82	51	44	53	48	52	17	52	19	47	6	50	16	39	1	44	10	38	2	39	7
2		107	113	66	60	68	72	50	40	52	46	51	18	51	19	45	5	49	17	38	2	44	13	37	2	38	8
3		107	110	97	165	100	184	57	80	60	85	49	8	50	8	51	4	57	16	38	1	51	12	36	1	46	12
4		106	108	96	130	99	155	55	62	57	67	47	7	48	9	48	4	54	14	37	1	49	13	35	1	43	10
1	3	116	73	186	∞	196	∞	141	∞	143	∞	83	24	84	28	85	22	84	28	84	20	82	30	81	17	81	23
2		115	66	187	∞	197	∞	143	∞	144	∞	83	19	83	21	84	17	83	21	83	14	82	22	81	13	81	16
3		108	110	151	780	155	978	93	420	94	437	68	16	68	16	69	13	69	17	68	12	68	17	65	10	65	14
4		107	107	155	860	160	1155	95	483	96	503	67	13	67	14	68	11	68	15	67	9	67	14	64	8	64	12
1	4	27	7	73	58	80	82	64	48	71	66	64	38	67	47	65	33	70	53	62	33	71	56	57	28	60	41
2		35	11	74	59	80	88	66	49	72	67	66	40	68	48	66	36	70	54	64	34	71	56	59	31	62	42
3		45	18	78	64	84	95	62	33	73	63	57	20	59	30	62	24	70	49	61	25	73	57	56	21	66	45
4		49	21	78	66	83	92	64	34	74	63	60	23	63	35	65	25	71	49	64	27	74	55	60	24	67	47
1	5	99	87	126	432	133	583	92	262	95	283	85	35	85	39	85	24	85	40	82	16	84	41	76	14	77	27
2		90	61	121	348	130	544	93	264	96	290	83	28	83	33	82	16	82	29	80	12	83	35	74	11	75	22
3		118	124	218	3814	223	4542	118	1769	119	1798	88	59	88	61	90	54	89	61	87	44	88	60	84	41	84	53
4		119	137	232	4396	239	5530	126	2181	127	2223	88	56	88	59	90	51	90	61	88	42	88	59	85	39	85	51

See note below Table A.1

B Data appendix: Female labour supply

The data contains observations on 2339 married Portuguese women whose spouses were employed in 1991. Descriptive statistics and correlations of the available variables are given in Table B.1 and B.2. 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 is taken.

Table B.1: 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 B.2: 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

References

Carroll, R., D. Ruppert, and A. Welsh (1998): "Local Estimating Equations," *Journal of American Statistical Association*, 93, 214–227.

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, 141–150.

Gerfin, M. (1996): "Parametric and Semi-Parametric Estimation of the Binary Response Model of Labour Market Participation," *Journal of Applied Econometrics*, 11, 321–339.

Gozalo, P., and O. Linton (2000): "Local Nonlinear Least Squares: Using parametric information in nonparametric regression," *Journal of Econometrics*, 99, 63–106.

Heckman, J. (1990): "Varieties of Selection Bias," *American Economic Review, Papers and Proceedings*, 80, 313–318.

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.

Holland, P. (1986): "Statistics and Causal Inference," *Journal of American Statistical Association*, 81, 945–970.

Hyslop, D. (1999): "State dependence, serial correlation and heterogeneity in intertemporal labor force participation of married women," *Econometrica*, 67, 1255–1294.

Judd, K. (1998): *Numerical Methods in Economics*. MIT Press, Cambridge.

Judge, G., R. Hill, W. Griffiths, H. Lütkepohl, and T.-S. Lee (1982): *Introduction to the Theory and Practice of Econometrics*. Wiley, New York, 2 edn.

Klein, R., and R. Spady (1993): "An Efficient Semiparametric Estimator for Binary Response Models," *Econometrica*, 61, 387–421.

Manski, C. (1993): "The Selection Problem in Econometrics and Statistics," in *Handbook of Statistics*, ed. by G. Maddala, C. Rao, and H. Vinod. Elsevier Science Publishers.

——— (2000): "Identification Problems and Decisions under Ambiguity: Empirical Analysis of Treatment Response and Normative Analysis of Treatment Choice," *Journal of Econometrics*, 95, 415–442.

——— (2004): "Statistical Treatment Rules for Heterogeneous Populations," *Econometrica*, 72, 1221–1246.

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.

Mittelhammer, R., G. Judge, and D. Miller (2000): *Econometric Foundations*. Cambridge University Press, Cambridge.

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

Press, W., B. Flannery, S. Teukolsky, and W. Vetterling (1986): *Numerical Recipes*. Cambridge University Press, Cambridge.

Racine, J., and Q. Li (2004): "Nonparametric Estimation of Regression Functions with Both Categorical and Continuous Data," *Journal of Econometrics*, 119, 99–130.

Rubin, D. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701.

Seifert, B., and T. Gasser (1996): "Finite-Sample Variance of Local Polynomials: Analysis and Solutions," *Journal of American Statistical Association*, 91, 267–275.

Staniswalis, J. (1989): "The kernel estimate of a regression function in likelihood-based models," *Journal of American Statistical Association*, 84, 276–283.

Stein, C. (1981): "Estimation of the Mean of a Multivariate Normal Distribution," *Annals of Statistics*, 9, 1135–51.

Tibshirani, R., and T. Hastie (1987): "Local likelihood estimation," *Journal of American Statistical Association*, 82, 559–567.