# Semi-parametric Models for Satisfaction with Income

Bellemare, C.; Melenberg, B.; van Soest, A.H.O.

*Publication date:*
2002



*Citation for published version (APA):*
Bellemare, C., Melenberg, B., & van Soest, A. H. O. (2002). *Semi-parametric Models for Satisfaction with Income*. (CentER Discussion Paper; Vol. 2002-87). Econometrics.

CentER

Discussion paper

No. 2002-87

# SEMI-PARAMETRIC MODELS FOR SATISFACTION WITH INCOME

By Charles Bellemare, Bertrand Melenberg, Arthur van Soest

October 2002

ISSN 0924-7815

1 2

# Semi-parametric Models for Satisfaction with Income

Charles Bellemare, Bertrand Melenberg, Arthur van Soest

This version: March 2002

**Abstract**

An overview is presented of some parametric and semi-parametric models, estimators, and specification tests that can be used to analyze ordered response variables. In particular, limited dependent variable models that generalize ordered probit are compared to regression models that generalize the linear model. These techniques are then applied to analyze how self-reported satisfaction with household income relates to household income, family composition, and other background variables. Data are drawn from the 1998 wave of the German Socio-Economic Panel. The results are used to estimate equivalence scales and the cost of children. We find that the standard ordered probit model is rejected, while some semi-parametric specifications survive specification tests against nonparametric alternatives. The estimated equivalence scales, however, are often similar for the parametric and semi-parametric specifications.

**Key words:** Semi–parametric estimation, Ordered Response, Equivalence scales

**JEL codes:** C14, C35, D12

*Tilburg University, Department of Econometrics, P.O. Box 90153, 5000 LE Tilburg, Netherlands

†We are grateful to an anonymous referee, Colin Cameron, and to participants of a CeMMAP/ESG workshop at University College London and seminars at CentER (Tilburg University) and Humboldt University Berlin for useful comments.

# 1 Introduction

How much additional income does a family with two children need to attain the same welfare level as a married couple without children? And how much does a single person need compared to a childless couple? The answers to these questions, so–called equivalence scales, are important for public policy concerning social benefits and child allowances. See, for example, Browning (1992), Nelson (1993), and Lewbel (1989). Chakrabarty (2000) uses equivalence scales to analyze gender bias in children in rural India. Equivalence scales are also required for an analysis of income inequality within and between countries that corrects for differences in household composition (see Jenkins, 1991) and for the analysis of poverty (see, for example, De Vos and Zaidi, 1997).

The most common approach to estimate equivalence scales is via a consumer demand system, relying on variation in expenditure on commodities such as food or typical adult goods across families with different composition (see Browning, 1992). Pollak and Wales (1979) already showed the main limitation of this approach: expenditure data alone are not sufficient to identify the equivalence scales. Identification can be achieved by making the assumption of independence of base utility, but this assumption has been rejected numerous times in empirical work.[1]

Results of Blundell and Lewbel (1991) imply that the informational content of demand systems about equivalence scales is limited, and that estimating equivalence scales could proceed in two steps. First, the levels of the equivalence scales in a given reference price setting should be estimated using other than demand data. Second, information on demand data can be used to identify the effects of price changes on the equivalence scales. An enormous literature is devoted to this second step (see Browning, 1992). The current paper is concerned with the first step only. Equivalence scales in a fixed price setting are analyzed, avoiding the complications and specification

[1]An exception is the analysis of Pendakur (1998). Using a semi–parametric model estimated on Canadian expenditure data, he cannot reject independence of base utility.

choices involved with estimating a demand system.

There are two types of non–demand data that have been used for this purpose in the literature. Both are subjective data, reflecting either the income level needed to attain a certain utility level (see van Praag, 1968, 1991, and Kapteyn, 1994, for example) or reflecting satisfaction with actual family income. We will use the latter type. It has been used before by, for example, Vaughan (1984) and Charlier (2002). The latter analyzes parametric cross-section as well as panel data models for Germany. Van den Bosch (1996), and Melenberg and van Soest (1996a) compare equivalence scales based upon the two types of subjective information. The latter study finds that equivalence scales based on the first type of subjective data are implausibly low. One of the possible explanations of this is that heads of households tend to underestimate household income components when reporting total household income (see Kapteyn, Kooreman and Willemse, 1988). If the head of the household underestimates required income in a similar way as actual household income, this could explain the low equivalence scales derived from this type of subjective data. Most of the studies on this issue use parametric models only. Exceptions are Melenberg and van Soest (1996a), who compare some parametric and semi–parametric estimates of equivalence scales for Dutch data, and Stewart (2002), who uses parametric and semi-parametric models explaining self-reported financial well-being to estimate equivalence scales for pensioners in the UK.

The current paper provides an overview of some parametric and semi–parametric techniques for estimating and comparing models that can be used to analyze ordered response variables such as satisfaction with income and to estimate functions of the parameters and non-parametric features of the model such as equivalence scales. Not only the estimation techniques will be described and applied, but also some tests that can be used to select the most appropriate model. The techniques will then be applied to compare a number of models of varying degrees of flexibility that explain satisfaction with household income from household income, family size and other family composition and age variables, and regional dummies. Data are drawn from the 1998 wave of the German Socio-Economic Panel, which has information on the household representative's satisfaction with household income on the discrete scale 0,1,2,...,10.

The semi–parametric models that we consider differ in several dimensions. Some are direct generalizations of the standard ordered probit model, relaxing distributional assumptions on the error terms. Others can be seen as generalizations of linear models, allowing for a flexible, non–linear, specification of the systematic part. These models and the corresponding estimation and testing techniques will be discussed in Section 2. The data used for the empirical analysis are presented in Section 3. Empirical results are discussed in Section 4. Section 5 concludes.

# 2 Models, Estimation Techniques, and Specification Tests

The standard model to explain an ordered discrete choice variable is the parametric ordered probit model:

$$y_i^* = x_i'\beta + u_i\,, \tag{1}$$

$$y_i = j \quad \text{if} \quad m_{j-1} < y_i^* \leq m_j, \quad j = 0, \ldots, 10\,, \tag{2}$$

$$u_i|x_i \sim N(0, \sigma^2)\,. \tag{3}$$

The index $i$ denotes the household; $x_i$ is a vector of explanatory variables including a constant term, $\beta$ is the vector of parameters of interest, and $u_i$ is the error term. We assume $m_{-1} = -\infty$ and $m_{10} = \infty$. The variance $\sigma^2$ and the bounds $m_0, \ldots, m_9$ can be seen as nuisance parameters. For identification, location and scale have to be fixed by imposing two parameter restrictions. This will be discussed below. Throughout, we assume that the observations $(y_i, x_i)$ are a random sample from the population of interest. The standard way to estimate this model is maximum likelihood (ML).

Based upon moments involving generalized residuals, this specification can be tested against models with heteroskedasticity and non-normality of the error terms in the underlying latent variable equation (see Chesher and Irish, 1987). If the tests reject the standard ordered probit specification, parametric extensions allowing for heteroskedasticity and/or non-normality can be used. See, for example, Horowitz (1993) and Melenberg and van Soest (1996b) for applications in the binary choice case.

The standard ordered probit model has the property that the conditional distribution of the dependent variable given the regressors $x_i$ depends on $x_i$ only through some linear index $x_i'\beta$, making it a special case of the following single index model as presented by Ichimura (1993):

$$E[y_i|x_i] = G(h(x_i, \beta)) \,, \tag{4}$$

where $h$ is given but $G$ is an unknown function, referred to as the link function. In this model, $x_i$ affects $E[y_i|x_i]$ only through the single index $h(x_i, \beta)$. The most common case applied in practice is the case of a linear index, with $h(x_i, \beta) = x_i'\beta$:

$$E[y_i|x_i] = G(x_i'\beta) \,. \tag{5}$$

It is easy to see that the standard ordered probit model is a special case of (5), with a link function that is known up to the auxiliary parameters $\sigma^2$ and $m_0, \ldots, m_9$. If in the ordered probit model the normality assumption in (3) is replaced by the assumption that $u_i$ and $x_i$ are independent, (5) is still satisfied, but with an unknown link function that depends also on the distribution of $u_i$. Thus, (5) is a natural semi–parametric generalization of the standard ordered probit model. Identifying $\beta$ in (5) (without imposing restrictions on $G$) requires normalizations of location and scale. Location is fixed by excluding the constant term from $x_i$. The scale is normalized by fixing one of the slope parameters to 1 or $-1$. This makes the assumption that the effect of the corresponding variable is known to be non–zero.

There are many ways in which (5) (with some additional regularity conditions) can be estimated. See, for example, the overview in Powell (1994). If all regressors are continuous, average derivative estimation is a computationally convenient and intuitively attractive estimation procedure, see Powell, Stock and Stoker (1989). Horowitz and Haerdle (1996) show how this technique can be combined with GMM to tackle the case where some regressors are continuous and some other regressors are discrete. See, for example, Dustmann and van Soest (2000) for an application and some simulations exploring the finite sample performance of this estimator. Since this estimator requires non-parametric regressions for each sub-sample of observations with specific values of the discrete regressors, it will not work very well in case the number of discrete outcomes is relatively large (given the size of the sample).

In this paper we focus on the semi–parametric least squares estimator introduced by Ichimura (1993). It has a natural intuitive interpretation. It requires numerical minimization of a non-convex objective function, but this appears to work quite well in practice, at least for the application in our analysis.[2]

## Semi–parametric Least Squares (SLS)

For the true value $\beta_0$ of the parameter $\beta$ in model (5), (5) implies $E[y|x] = E[y|x'\beta_0] = G(x'\beta_0)$. Regularity conditions guaranteeing identification (for example, no multi–collinearity in $x$) imply that for $\beta \neq \beta_0$, there will be some $x$ for which $E[y|x] \neq E[y|x'\beta]$. Together with the equality $E[(y - E[y|x'\beta])^2|x] = E[(y - E[y|x])^2|x] + (E[y|x] - E[y|x'\beta])^2$ (the proof of which is straightforward), this implies that $\beta_0$ is the value of $\beta$ minimizing

$$E[W(x)(y - E[y|x'\beta])^2] \tag{6}$$

[2]Other examples of estimators that require numerical optimization are the maximum rank correlation estimator of Han (1987) and the estimator of Klein and Sherman (2002). The latter is specifically designed for the ordered response model and can also be used to estimate the thresholds $m_1, \ldots, m_9$.

for any weighting function $W(x)$ which is positive for almost all $x$. The standard SLS estimator introduced by Ichimura (1993) minimizes the sample analogue of (6) with $W(x) = 1$, using a sample $(x_1, y_1), \ldots, (x_n, y_n)$. For given $\beta$, $E[y|x'\beta]$ is estimated using a one-dimensional non–parametric kernel regression estimator,

$$\hat{E}[y|z] = \sum_{i=1}^{n} w(x_i'\beta - z)y_i \tag{7}$$

where the $w(x_i'\beta - z), i = 1, \ldots, n$ are kernel weights giving high weight to observations $i$ with $x_i'\beta$ close to $z$. The sample analogue of (6) is then given by

$$\hat{E}[(y - E[y|x'\beta])^2] = 1/n \sum (y_i - \hat{E}[y_i|x_i'\beta])^2. \tag{8}$$

Finding the $\beta$ at which (8) is minimized requires an iterative procedure. If smooth kernel weights are used, the function to be minimized is smooth in $\beta$ and a Newton-Raphson technique can be used to find the optimal $\beta$, i.e., $\hat{\beta}_{SLS}$. Ichimura (1993) shows that, under appropriate regularity conditions, this yields a $\sqrt{n}$ consistent asymptotically normal estimator of $\beta_0$. He also derives the asymptotic covariance matrix of this estimator and shows how it can be estimated consistently.[3]

Implementing the SLS estimator in practice requires a choice of kernel and bandwidth, i.e., a specification of the weights $w(x_i'\beta - z)$. We will work with the Gaussian kernel $K(t) = 1/\sqrt{2\pi}\, exp[-t^2/2]$. For given bandwidth $h > 0$, the weights are then given by

$$w(x_i'\beta - z) = K([x_i'\beta - z]/h) / \sum_{j=1}^{n} K([x_j'\beta - z]/h) \tag{9}$$

For consistency, the bandwidth should tend to zero if $n \to \infty$ at a slow enough rate. Although a large literature on the optimal bandwidth choice exists for the non–parametric regression problem itself, it is not clear how to determine the optimal bandwidth for estimating $\beta_0$. Theoretical results for similar problems suggest that under-smoothing

[3] In general, this estimator is not efficient. Ichimura (1993) mentions that efficiency can be improved by choosing an appropriate weighting function $W(x)$, using a two step procedure. We do not pursue this in the current paper.

will be optimal, i.e., the optimal bandwidth will be smaller than the optimal bandwidth for the non–parametric regression of $y$ on $x'\beta$. The common approach for choosing a bandwidth in a situation like this is to experiment with the bandwidth which would be optimal for the non-parametric regression problem (given plausible values of $\beta$) and with smaller bandwidth values (to under-smooth). In our experiments with such bandwidth choices, the results hardly varied with the bandwidth.

The link function $G$ can be estimated in a second step by regressing $y$ non–parametrically on the estimated index $\beta' x$, using a kernel estimator. The usual asymptotic properties of a kernel estimator apply since $\hat{\beta}_{SLS}$ converges at a faster rate than the non-parametric estimator.

## Smoothed Maximum Score

The parametric model in (1) - (3) assumes that the errors and regressors are independent and thus does not allow for heteroskedasticity. The single index model in (5) only allows for very specific types of heteroskedasticity, where the regressors affect the conditional variance $V[\epsilon|x]$ through the single index $x'\beta$ only. A model that allows for much more general forms of heteroskedasticity is obtained if (3) is replaced by

$$Median[u|x] = 0 \tag{10}$$

This model nests the parametric ordered probit model (1) - (3) but not the single index model in (4), since (10) is a conditional median assumption and not a conditional mean assumption. The reason for using the conditional median is the median preserving property of any increasing function. Lee (1992) uses this property to construct a consistent estimator for the model defined by (1), (2) and (10). These assumptions imply[4]

[4]Here 1[] is the indicator function: $1[A] = 1$ if $A$ is true and $1[A] = 0$ if $A$ is false.

$$Median[y|x] = \sum_{r=0}^{9} 1[x'\beta \geq m_r] \tag{11}$$

Since the conditional median minimizes the conditional expectation $E[|y - a||x]$ over $a$, a consistent extremum estimator for $\beta$ and $m_0, \ldots, m_9$ can be obtained as

$$(\hat{\beta}_{MS}, \hat{m}_0, \ldots, \hat{m}_9) = Argmin_{\beta, m_0, \ldots, m_9}(\sum_{i=1}^{n} |y_i - \sum_{r=0}^{9} 1[x_i'\beta \geq m_r]|) \tag{12}$$

Lee's estimator generalizes the maximum score estimator of Manski (1985) for the binary choice model. It shares the drawback of Manski's estimator: the asymptotic distribution is intractable. For the maximum score estimator, this problem is solved by Horowitz (1992). His 'smoothed maximum score' estimator maximizes a smoothed version of the sum of least absolute deviations. The same idea is applied by Melenberg and van Soest (1996a) to Lee's estimator in (12). See also the clear exposition in Horowitz (1998). The smoothed maximum score estimator is given by

$$(\hat{\beta}_{MS}, \hat{m}_1, \ldots, \hat{m}_9) = Argmin_{\beta, m_1, \ldots, m_9} \sum_{i=1}^{n} |y_i - \sum_{r=0}^{9} K([x_i'\beta - m_r]/\sigma)| \tag{13}$$

where $K$ is some smooth distribution function that is symmetric around zero and $\sigma$ is a bandwidth parameter that tends to zero with the sample size at a slow enough rate. This estimator shares the asymptotic characteristics of the Horowitz (1992) estimator: it is consistent and asymptotically normal. The rate of convergence depends on conditions on smoothness and properties of the kernel, but is always slower than $\sqrt{n}$. Horowitz (1998) and Melenberg and van Soest (1996a) show how the asymptotic covariance matrix can be estimated.

In the application we use a Gaussian distribution function for $K$. Unfortunately, there are no procedures for selecting the optimal bandwidth for this estimator. We experimented with a broad range of bandwidth values and found that the estimation results were similar for a large range of reasonable values. On the other hand, the estimates of the standard errors were more sensitive. Unfortunately, bootstrapped standard errors are not a feasible option here, since the numerical optimization routine to obtain the estimates requires too much computer time.

## Partially Linear Model

In the more recent econometrics literature, partially linear models and generalized partially linear models have become popular. These models relax the linear index assumption on the conditional mean. Some regressors ($x_1$) are allowed to enter in an arbitrary not necessarily linear way, while others ($x_2$), are assumed to enter linearly. The standard partially linear model assumes

$$E[y|x_1, x_2] = g(x_1) + x_2'\beta \tag{14}$$

where g is an unknown continuous function. Robinson (1988) and Stock (1991) explain how to estimate $\beta$ and $g$, respectively. (14) immediately implies:

$$y - E[y|x_1] = (x_2' - E[x_2'|x_1])\beta + \epsilon, \text{ with } E[\epsilon|x_1] = 0 \tag{15}$$

The first estimation step is to replace the conditional expectations in (15) by their nonparametric (kernel) regression estimates. The second step is to estimate $\beta$ by OLS on (15). This gives $\sqrt{n}$ consistent and asymptotically normal estimates of $\beta$. The third step is to estimate $g$ using a nonparametric regression of $y - x_2'\hat{\beta}$ on $x_1$. This estimator has the same limiting distribution as a usual one step nonparametric regression estimator, since the nonparametric rate of convergence is slower than the rate of convergence of $\hat{\beta}$.

For choosing the bandwidth, similar remarks apply as for the other semi-parametric estimators. There is no theory on how to choose the bandwidth. Bandwidth choices that are optimal for the non-parametric regressions are not necessarily optimal for estimating $\beta$. Our experiments show that such bandwidth choices and bandwidth values that are smaller lead to very similar results.

## Generalized Partially Linear Models

Generalized partially linear models add a link function $G$ to the partially linear model in (14):

$$E[y|x_1, x_2] = G[g(x_1) + x_2'\beta] \tag{16}$$

Horowitz (2001) discusses the case where $G$ is unknown. To identify this model, a sufficient number of continuous variables must be available. Given the limitations of the data with respect to continuous variables, however, we will only consider a special case where $G$ is known. In particular, we will generalize the ordered probit model (1)–(3) as follows:

$$y_i^* = g(x_{1i}) + x_{2i}'\beta + u_i\,, \tag{17}$$

$$y_i = j \quad \text{if} \quad m_{j-1} < y_i^* < m_j, \quad j = 0, \ldots, 10\,, \tag{18}$$

$$u_i|x_i \sim N(0, \sigma^2)\,. \tag{19}$$

Instead of relaxing the distributional assumptions on the error term as for the single index models and the smoothed maximum score estimator, model (17)–(19) retains the normality assumptions but does not impose that the systematic part is linear in $x_{1i}$. The probabilities of the ordered outcomes are given by

$$P[y_i = j|x_i] = \Phi([g(x_{1i}) + x_{2i}'\beta - m_j]/\sigma) - \Phi([g(x_{1i}) + x_{2i}'\beta - m_{j-1}]/\sigma) \tag{20}$$

The model can be estimated by the quasi maximum likelihood technique described by Haerdle, Huet, Mammen and Sperlich (2001). The estimator is based upon the algorithm of Severini and Staniswalis (1994). The nonparametric part ($g(x_1)$) and the parametric part ($\theta = (\beta, m_0, \ldots, m_9, \sigma^2)$) are iteratively updated. For given $\theta$, $g(t)$

is updated by maximizing a weighted likelihood based upon (20), giving weight to observations $i$ with $x_{1i}$ close to $t$ only: $g(t)$ is the value of $\eta$ that maximizes

$$\sum_{i=1}^{n}\sum_{j=0}^{10} 1[y_i = j]\, K([t - x_{1i}]/h)[\Phi([\eta + x_{2i}'\beta - m_j]/\sigma) - \Phi([\eta + x_{2i}'\beta - m_{j-1}]/\sigma)] \quad (21)$$

Substituting this expression for $g(t)$ in the likelihood gives a profile likelihood in terms of $\theta$. Maximizing this profile likelihood over $\theta$ gives the estimates of $\theta$ and $g$.[5] Haerdle et al. (2001) show that the estimator for $\theta$ is $\sqrt{n}$ consistent and asymptotically normal and derive its asymptotic covariance matrix. To determine the limiting distribution of the estimator of $g$, the fact that $\theta$ is estimated can again be ignored, because the non-parametric estimator has a slower rate of convergence than the estimator of $\theta$. We estimated the standard errors using a bootstrap procedure that takes the estimate of $\theta$ as fixed.

## Testing for Misspecification

To test some of the semi–parametric models, we will apply the consistent tests developed by Fan and Li (1996).[6] These can be used to test both the semi–parametric partial linear model and the semi–parametric single-index model. Consider first the semi–parametric single-index model. Define $g(x) = E[y|x]$. Consider the null hypothesis $H_0 : g(x) = G(\beta' x)$, for some function $G$ with domain and range the real line, against the alternative that no $G$ and $\beta$ exist such that $g(x) = G(\beta' x)$ for all $x$ (or, to be precise, almost sure in $x$). Define $u = y - G(\beta' x)$. Then $E[u|x] = 0$ under $H_0$, while under $H_1$, $E[u|x] \neq 0$ for some $x$ (to be precise, $P[E[u|x] \neq 0] > 0$).

For positive weight functions $w_1(x)$ and $w_2(x)$, it is easy to show (using the law of iterated expectations) that under $H_1$, $E[uw_1(x)E[uw_2(x)|x]] = E[w_1(x)w_2(x)(E[u|x])^2] >$

[5] As in the ordered probit model, some normalizations are needed.

[6] Fan and Li (1996, p. 866-867) refer to several alternative tests, but argue that most of these have *ad hoc* features such as sample splitting that probably makes them less powerful.

0, while under $H_0$, $E[uw_1(x)E[uw_2(x)|x]] = 0$. Fan and Li (1996) use this to construct a consistent test for $H_0$ against $H_1$.

Fan and Li use the weighting functions $w_1(x) = f_1(\beta' x) f_2(x)$ and $w_2(x) = f_2(x)$, where $f_1(x)$ is the density of $\beta' x$ and $f_2(x)$ is the density of $x$. This has the advantage that low weight is given to observations in regions where data are sparse and non-parametric estimates are inaccurate, and can thus be seen as some type of trimming. Fan and Li (1996, eq. (14)) show, under some regularity conditions, that an appropriately scaled estimator of $E[uf_1(\beta' x)E[uf_1(\beta' x)|x]f_2(x)]$ yields a test statistic that asymptotically follows a standard normal distribution under $H_0$. Under $H_1$, the probability that the test statistic exceeds the 5% critical value of the standard normal distribution will tend to 1, leading to a one-sided consistent test.

The same idea is also applicable to the semi–parametric partial linear model, see Fan and Li (1996, eq. (11)). In this case the consistent test is again asymptotically $N(0,1)-$distributed under the corresponding null hypothesis.

The asymptotic distributions of the Fan and Li (1996) are derived under the assumption of continuously distributed regressors, while some of our regressors are discrete. We will ignore this problem when we apply the tests in the next section. To investigate whether this is a serious problem, we conducted a small simulation study on the performance of the Fan and Li test in the case when not all regressors are continuous. We sampled data from a standard (homoskedastic) ordered probit model with three possible outcomes and from an ordered probit model with heteroskedasticity of a (separate) single index type. The former satisfies the null that the model is a single index model, the latter does not satisfy the null. We considered two sets of regressors: one with two continuous regressors, and the other with one continuous regressor and one dummy variable. We estimated slope parameters of the (regression) single index applying Ichimura's SLS-estimator, and tested subsequently for misspecification using Fan and Li's approach.

In the Fan and Li test, two bandwidth parameters need to be chosen, one in the non-parametric regression on $\hat{\beta}'x$ and the other in the non-parametric regression on $x$. The results appear to be insensitive for the choice of the first bandwidth but do depend on the second one. We applied Silverman's a rule of thumb to choose the first bandwidth (see Silverman, 1986), and varied the second bandwidth over a fine grid. For each bandwidth choice we performed 100 simulations, with sample size 200, for each of the four models described above.

Figure 1 presents the simulated rejection probabilities. For the two data generating processes that satisfy the null hypothesis, the type I error probability is close to the nominal size of 5% for a large range of bandwidth values, suggesting that the performance of the test is quite good, even in the case of one discrete regressor that does not satisfy Fan and Li's regularity assumptions. For the two data generating processes that do not satisfy the null, the rejection probability (i.e., the power against these specific alternatives) is more sensitive to the chosen bandwidth, particularly for the case with one discrete regressor. The power is not systematically larger or smaller for any of the two cases. We conclude that the test performs well in terms of similarity of actual size and nominal size. On the other hand, the simulations reveal that, as long as there is no theory on how to choose the bandwidth in some optimal way, it seems wise to calculate the test under various bandwidth choices.

# 3 Data and Variables

The data are drawn from the fifteenth wave of the German Socio-Economic Panel (GSOEP) Public use File, drawn in 1998. We have used the full sample, including former East as well as former West Germany and including the refreshment sample drawn in 1998. In each household, one person answers the household specific part of the survey, usually the main earner. This person also reports total household net income, the income measure we use in the empirical models. The dependent variable in our analysis is the answer by the same household representative to the question

*How satisfied are you with your household income?*

*Possible answers: 0 (not satisfied at all) to 10 (very satisfied).*

The total sample consists of 7,274 households. About 7% had a missing value on one of the variables used in the analysis, usually after tax household income. Deleting these gives a sample of 6,755 households that is used for the descriptive statistics and for all the estimations. Definitions and descriptive statistics of the variables used in the analysis are provided in Table 1. About 26.5% of the sample consists of households living in East Germany. Family size varies from 1 to 12, but only 6% of all households consist of more than four persons and only 1.8% of more than five persons. Almost 26% are one person households.

Figure 2 presents the distribution of satisfaction with household income. The sample average is 6.03 but the dispersion is substantial. Figure 3 presents nonparametric kernel density estimates[7] of the distribution of log net household income by family size. As expected, the larger families tend to have the higher incomes. The difference is particularly large between one and two person households, since the third and fourth person in the household are typically children who do not contribute to total household income.

Figure 4 presents nonparametric (Gaussian) kernel regressions[8] of satisfaction with income on log household income for the same family size categories that were distinguished in Figure 3. Figure 3 gives the log income ranges in the data, i.e., the ranges for which the curves are reasonably accurate. For given family size, satisfaction rises with the level of income in the whole income range, except for some regions where data are sparse and estimates are inaccurate. Moreover, for given income, satisfaction falls with family size. This confirms that larger families need more income to be as well off as smaller families.

Figure 4 also illustrates how equivalence scales can be determined in a nonpara-

[7]See, for example, Silverman (1986) or Haerdle and Linton (1994)

[8]See, for example, Haerdle and Linton (1994)

metric setting. A reference satisfaction level has to be set a priori. In Figure 4, the chosen level is represented by the horizontal line at satisfaction level 6.03, the sample average. The intersection of this line with one of the curves gives the typical log income value needed for a family of given size to attain the average satisfaction level. For a one person household, this is log income level 7.60, for a two persons household it is 7.97. Thus, according to these nonparametric estimates, the equivalence scale for a two person household compared to a single living person, is $e^{7.97}/e^{7.60} = 1.45$. Equivalence scales for three and four persons households can be determined in a similar way. We will discuss the results at the end of the next section and compare them to parametric and semi-parametric estimates.

In principle, standard errors on the non-parametric estimates of the equivalence scales can be derived from the asymptotic distribution of the non-parametric estimates. Since the equivalence scales are obtained by inverting the curves, it is not clear how point-wise or uniform confidence bands on the curves could be used directly. Instead, a bootstrapping procedure can be used.

The non-parametric equivalence scales rely on very weak assumptions and may therefore not be very inaccurate. Moreover, they have the drawback that other variables which may affect satisfaction with income (and could be correlated to log income and/or family size) are not taken into account. To control for these additional variables, we use the parametric and semi–parametric models in the previous section.

# 4 Results

The performance of some of the semi-parametric estimators may depend on the number of regressors included in the model. To investigate whether this is indeed the case, we analyzed two different specifications. Both include log income, a dummy for East Germany, and log age of the household respondent, but the specifications differ in the family composition variables. Specification 1 is kept as parsimonious as possible and includes log family size only. Specification 2 includes separately the numbers of children in various age groups.[9] We focus on the second specification but we will compare the equivalence scales according to this model with those according to the more parsimonious model.

The estimation results for the second specification are presented in Table 4. The magnitude of the parameter estimates is not comparable across models, since, due to different link functions, the scale varies. It is possible to compare signs and relative magnitudes, however. In some specifications, the coefficient of log income is normalized to one and in some other specifications the relation between satisfaction with income and income is non-parametric. In the remaining specifications, the log of self-reported income has a strong and significant positive effect on the reported satisfaction with income.

According to all estimates other than smoothed maximum score, East Germans are significantly less satisfied with a given income than West Germans with the same characteristics. The reason may be that satisfaction not only depends on the current real income level but also on the change in purchasing power (cf. Clark and Oswald, 1996). Due to price increases, real wages in East Germany have risen less than in West Germany. Log age is always significantly positive, indicating that the older cohorts tend to be more satisfied with a given income than the younger cohorts. In this cross–section analysis, this may reflect a cohort as well as an age effect. Marital status does not have any effect in the ordered response models and is no longer included in the other models.

According to all estimates other than smoothed maximum score, keeping the income level and other regressors constant, children and other adults in the household reduce satisfaction with income, and increase the family's cost of living. The effects are significant at the 5% level, except, for several models, for the youngest age group.

[9] The two models are non–nested since the first specification uses log family size rather than family size.

The effects are all significant at the 10% level. According to all except the smoothed maximum score estimates, the effect of very young children is much smaller than the effect of children in the older age groups. Moreover, costs of additional adults typically exceed costs of children in all age groups. Only according to the generalized partially linear model, adults and children between 13 and 17 have virtually the same effect.

The smoothed maximum score estimates do not look plausible. They imply insignificantly negative costs of children in the age groups 6-12 and 13-17. Such negative effects do not make sense from an economic point of view. It seems that the rich specification combined with the very weak conditional median assumption makes it very hard in practice to estimate the parameters, in spite of the comparatively large size of the sample. This is confirmed by the estimation results of the more parsimonious specification 1 (not presented). For this specification, the smoothed maximum score estimates look much more plausible. They are also similar to those of other models, with a significantly negative effect of log family size and significantly lower satisfaction levels of East German households, *ceteris paribus*.[10]

Figure 5 presents the estimated link function $G$ for the semi–parametric least squares estimates, together with 95% uniform confidence bounds. This function is obtained by a non–parametric kernel regression of the dependent variable $y_i$ on the estimated index $x_i'\hat{\beta}_{SLS}$. The points $(x_i'\hat{\beta}_{SLS}, y_i)$ are plotted as well. The estimated link function is monotonically increasing on almost the whole range of the index.

In the partially linear model and the generalized partially linear model, log income enters in a non–parametric way. The estimated non–parametric functions of log income ($g(x_1)$ in (14) and (17)) are presented in Figures 6 and 7 for the partially linear model and the generalized partially linear model, respectively. The figures also include the estimates for specification 1, which are very similar to those for specification 2. Figure

[10] Another reason for differences between smoothed maximum score and the other estimates could be the assumption of a zero conditional median instead of a zero conditional mean. See the discussion in section 2.

6 also presents uniform confidence bands for the estimated function in specification 2. The figures show that satisfaction is monotonically rising with income on almost the whole range of observed incomes. Although linearity is formally rejected, the curves are not far from linear, particularly in the partially linear model case.

The parametric ordered probit specification was tested against heteroskedasticity and non-normality using the Lagrange Multiplier tests described in Chesher and Irish (1987). Results are presented in Table 2. The assumption that the error terms are normal is rejected at any reasonable significance level. Moreover, there is evidence of heteroskedasticity, suggesting that the variance of the error term varies with income and the numbers of older children and adults. These results make looking at more general parametric or semi-parametric models worthwhile, since the evidence of mis-specification implies that ordered probit may lead to biased estimates of the parameter estimates. On the other hand, how large this bias is and which sign it has can only be investigated by looking at alternative estimates based upon less stringent model assumptions.

Applying the Fan and Li (1996) test reveals that the estimated single index model fits the data reasonably well. For most values of the bandwidth parameters, the null hypothesis that the single index specification is correct cannot be rejected. Similar results are found for the partially linear regression model, so that the Fan and Li test cannot determine which of the two models should be chosen. In discussing Figure 6, we already showed that the linear model is rejected against the more general partially linear model, since the estimated function $g$ is non-linear in log income. Unfortunately, the tools for testing the generalized partially linear specification are not yet available. We conclude that specification tests show that the two simplest models (ordered probit and linear regression model) are rejected, but are not able to choose among the semi–parametric models.

Estimated equivalence scales according to both specification 1 and specification 2 are presented in Table 4. The single person household is chosen as the benchmark.

For the partially linear and generalized partially linear model, the equivalence scales have been computed numerically, in the same way as for the non-parametric case, described in the previous section. The benchmark satisfaction level is set equal to 6.03, the mean satisfaction level in the data. For given family size and mean values of the other variables, the income required to attain the benchmark satisfaction level is computed using the estimated function $g(x_1)$. The equivalence scales are computed as ratios of required income levels at different values of family size. Standard errors for the partially linear model are bootstrapped. Those for the generalized partially linear model are bootstrapped as well, but taking the first step estimates as given, as explained in the previous section. Even this is extremely time consuming so that only 50 bootstrap replications could be used.[11]

For specification 1, most of the estimated equivalence scales are remarkably close to each other and suggest that the cost of living for a couple are about 32% to 39% higher than the cost of living for a single person. Only the fully non-parametric estimate discussed in the previous section (cf. Figure 4) is substantially larger (45%). This estimate cannot be directly compared to the other estimates since it does not control for age of the household representative or for living in either East or West Germany. A third person raises the household's cost of living by about 37% of the cost of living of a couple according the non-parametric estimates and by about 18% to 22% according to the single index models (ordered probit, SLS, smoothed maximum score and linear model estimates). In the single index models, however, this percentage is directly linked to the cost of living index of a couple, due to the choice of functional form with log family size and log income. This functional form also implies that additional persons lead to lower relative cost increases. The generalized partially linear model yields point estimates of the equivalence scales for two and three person households that are similar to those in the partially linear model, but this model yields particularly high estimates of the costs of a fourth person. Since, however, log family size enters linearly and

[11]Obtaining the estimates already took more than one week of computer time.

only log income enters in a more flexible way, this finding may be due to the chosen benchmark level of income.

Standard errors in the partially linear model are much larger than the standard errors in the single index model or in the linear model. The slower rate of convergence of the nonparametric part seems to play a large role here. According to the standard errors in Table 3, the parametric part is estimated with virtually the same accuracy in linear and partially linear model. The bootstrapped standard errors on the equivalence scales in the generalized partially linear model are somewhat smaller but of similar order of magnitude as those in the partially linear model.

Table 4 also contains some equivalence scales according to the second specification. Since in this specification, the cost of a child or adult can vary with the age of the person, we focus on singles and couples with one or zero children. The results are in line with the estimates in Table 3. The smoothed maximum score estimates lead to negative costs of children in the age groups 6-12 and 13-17 and thus do not make economic sense. The non-parametric estimates are in some cases determined with very little precision only, due to small number of observations with specific family composition. The other estimates are generally in line with the existing literature. They all imply that the cost of a person increases with the person's age.[12] The partially linear model and the generalized partially linear model give somewhat higher equivalence scales than the other models, but the differences are not very large and confidence intervals overlap. The standard errors according to these models are larger than those in the parametric models but much smaller than those of the fully non-parametric estimates. This illustrates once again that the semi-parametric assumptions help to increase precision and avoid the curse of dimensionality, even though the dimension of the non-parametric regression is limited by excluding the region and age variables.

[12]For West Germany 1984–1991, Charlier (2002) finds costs of children of a similar order of magnitude as we do. However, he finds much larger costs of a second adult in the household.

# 5 Conclusions

In this paper we have compared a number of parametric and semi–parametric estimators of the ordered response model. We have discussed theoretical and practical features of the models and the estimators. Moreover, we have presented some consistent ways of testing the underlying model assumptions against general forms of mis–specification. These techniques were applied to estimating the determinants of subjectively measured satisfaction with household income, with emphasis on computing household equivalence scales. This is a particularly attractive application for the single index models, since the parameters of these identify only the ratios of the coefficients, and this is exactly what the equivalence scales refer to.

We find that the specification tests are powerful enough to be of help to evaluate the performance of the various models (to which we have applied these tests). On the other hand, however, the equivalence scales that we find seem to be rather robust for this mis–specification, in the sense that most models give rather similar scales. In particular, this is the case for the estimators that do not depend on smoothness parameters or for which findings are robust for the choice of smoothness parameters. Among the semi–parametric estimators, these are the semi–parametric least squares estimator of Ichimura (1993) and the estimator for the partially linear model taken from Robinson (1988) and Stock (1991). The estimator for the generalized partially linear model recently developed by Haerdle et al. (2001) performs similarly well as far as we can judge, but has the drawback that it requires an enormous amount of computer time. As far as we know, we are the first to apply this estimator, and more refined programming can solve a large part of this problem. We leave this for future work. Obtaining more efficient estimates using a weighted version of Ichimura's SLS estimator is another topic for future work.

The smoothed maximum score estimator is the other estimator that gives some concern about robustness for choice of smoothness parameters and plausibility of the results. This estimator is consistent under weaker conditions than the other single index estimators, but it seems that this theoretical robustness property comes ate the cost of inferior finite sample behavior. Developing methods for choosing appropriate smoothness parameters remains an open issue. This also holds for the tests against non–parametric alternatives that we have considered, since the result of these tests often appears to vary with the smoothness parameters that are chosen.

Overall, we hope to have demonstrated that the applied researcher now has a number of semi–parametric alternatives to the standard parametric ordered probit model and an increasingly large toolbox for testing parametric and semi–parametric assumptions against still more general, non–parametric, alternatives. We hope to have demonstrated that some of these alternative estimators and tests are not only theoretically attractive, but also perform well in practical situations.

# References


- Blundell, R. and A. Lewbel (1991), The information content of equivalence scales, *Journal of Econometrics*, 50, 49-68.
- Browning, M. (1992), Children and household economic behavior, *Journal of Economic Literature*, 30, 1434-1475.
- Chakrabarty, M. (2000), Gender-bias in children in rural Maharashtra – an equivalence scale approach, *Journal of Quantitative Economics*, 16, 51-65.
- Charlier, E. (2002), Equivalence Scales in an intertemporal setting with an application to the former West Germany, *The Review of Income and Wealth*, forthcoming.
- Chesher, A. and M. Irish (1987), Residual analysis in the grouped and censored normal linear model, *Journal of Econometrics*, 34, 33-61.
- Clark, A. and A. Oswald (1996), Satisfaction and comparison income, *Journal of Public Economics*, 61, 359-381.

- Dustmann, C. and A. van Soest (2000), Parametric and Semi-Parametric Estimation of Models with misclassified Categorical Dependent Variables, IZA discussion paper 218, IZA, Bonn.
- De Vos, K. and A. Zaidi (1997), Equivalence scale sensitivity of poverty statistics for the member states of the European Community, *The Review of Income and Wealth*, 43, 319-334.
- Fan, Y. and Q. Li (1996), Consistent Model Specification Tests: Omitted Variables and Semiparametric Functional Forms, *Econometrica*, 64, 865-890.
- Haerdle, W., S. Huet, E. Mammen, and S. Sperlich (2001), Bootstrap inference in semiparametric generalized additive models, working paper, Humboldt University.
- Haerdle, W. and O. Linton (1994), Applied nonparametric methods, in R. Engle and D. McFadden (eds.), *Handbook of Econometrics*, Volume IV, North-Holland, Amsterdam.
- Han, A.K. (1987), Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator, *Journal of Econometrics*, 35, 303-316.
- Horowitz, J., (1992), A smoothed maximum score estimator for the binary response model, *Econometrica*, 60, 505-531.
- Horowitz, J. (1993), Semi-parametric estimation of a work-trip mode choice model, *Journal of Econometrics*, 58, 49-70.
- Horowitz, J. (1998), *Semiparametric Methods in Econometrics*, Springer, New York.
- Horowitz, J. (2001), Nonparametric estimation of a generalized additive model with an unknown link function, *Econometrica*, 69, 499-513.

- Horowitz, J. and W. Haerdle (1996), Direct semi-parametric estimation of single index models with discrete covariates, *Journal of the American Statistical Association*, 91, 1632-1640.

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

- Jenkins, S. (1991), The measurement of income inequality, in: L. Osberg (ed.) *Economic Inequality and Poverty*, M.E. Sharpe, New York.

- Kapteyn, A. (1994), The measurement of household cost functions: revealed preference versus subjective measures, mimeo, Tilburg University.

- Kapteyn, A., P.Kooreman and R. Willemse (1988), Some methodological issues in the implementation of subjective poverty definitions, *Journal of Human Resources*, 23, 222-242.

- Klein, R. and R. Sherman (2002), Shift restrictions and semiparametric estimation in ordered response models, *Econometrica*, 70, 663-691.

- Lee, M.-j., (1992), Median regression for ordered discrete response, *Journal of Econometrics*, 51, 59-77.

- Lewbel, A. (1989), Household equivalence scales and welfare comparisons, *Journal of Public Economics*, 39, 377-391.

- Manski, C. (1985), Semiparametric analysis of discrete response, Journal of Econometrics, 27, 313-333.

- Melenberg, B. and A. van Soest (1996a), Measuring the costs of children: parametric and semiparametric estimators, *Statistica Neerlandica*, 50, 171-192.

- Melenberg, B. and A. van Soest (1996b), Parametric and semi-parametric modelling of vacation expenditures, *Journal of Applied Econometrics*, 11, 59-76.

- Nelson, J. (1993), Household equivalence scales: theory versus policy?, *Journal of Labor Economics*, 11, 471-493.
- Pendakur, K. (1998), Semiparametric estimates and tests of base-independent equivalence scales - A theoretical framework, *Journal of Econometrics*, 88, 1-40.
- Pollak, R. and T. Wales (1979), Welfare comparisons and equivalence scales, *American Economic Review, papers and proceedings*, 69, 216-221.
- Powell, J. (1994), Estimation of semiparametric models, in R. Engle and D. McFadden (eds.), *Handbook of Econometrics*, Volume IV, North-Holland, Amsterdam, 2444-2523.
- Powell, J., J. Stock, and T. Stoker (1989), Semi-parametric estimation of index coefficients, *Econometrica*, 57, 1403-1430.
- Robinson, P. (1988), Root-N consistent semi-parametric regression, *Econometrica*, 56, 931-954.
- Severini, T., and J. Staniswalis (1994), Quasi-likelihood estimation in semiparametric models, *Journal of the American Statistical Association*, 89, 501-511.
- Silverman, B. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.
- Stewart, M. (2002), Pensioner financial well-being, equivalence scales and ordered response models, mimeo, University of Warwick.
- Stock, J. (1991), Nonparametric policy analysis: an application to estimating hazardous waste cleanup benefits, in: W.A. Barnett et al. (eds), *Nonparametric and Semi-parametric Methods in Econometrics and Statistics*, Cambridge University Press, Cambridge (Mass.).
- Van den Bosch, K. (1996), Equivalence scales based on subjective income evaluations: are children really cheap, *Recherches Economiques de Louvain*, 62, 203-227.

- Van Praag, B. (1968), *Individual welfare functions and consumer behavior*, North-Holland, Amsterdam.

- Van Praag, B. (1991), Ordinal and Cardinal Utility: an integration of the two dimensions of the welfare concept, *Journal of Econometrics*, 50, 69-89.

- Vaughan, D. (1984), Using subjective assessments of income to estimate family equivalence scales: a report on work in progress, *Proceedings of the Social Section of the American Statistical Association*, 496-501.

**Table 1: Variable Definitions and Sample Statistics**

| Variable | Definition | Mean | Std Dev |
|---|---|---|---|
| Dummy East | 1 if living in East Germany, 0 otherwise | 0.265 | 0.441 |
| Log(age) | log age household representative | 3.778 | 0.360 |
| Log(fam size) | log number of persons in household | 0.761 | 0.537 |
| Log(income) | log household net income (DM per month) | 8.130 | 0.515 |
| DMarried | 1 if married or living together, 0 otherwise | 0.603 | 0.489 |
| NAge06 | number of children 0-5 years old | 0.223 | 0.551 |
| NAge712 | number of children 6-12 years old | 0.214 | 0.525 |
| NAge1317 | number of children 13-17 years old | 0.165 | 0.451 |
| NAdults | number of household members age 18 or older | 1.855 | 0.732 |

Source: German Socio-Economic Panel 1998; 6755 observations

**Table 2: LM Specification Tests Ordered Probit**

| Hypothesis | Test statistic | Critical value |
|---|---|---|
| Non-normality[a] | 29.4036 | 5.9915 |
| Heteroscedasticity[b] | | |
| $\mathbf{z}$ =log(age) | 0.2592 | 3.8415 |
| $\mathbf{z}$ =log(income) | 27.1165 | 3.8415 |
| $\mathbf{z}$ =NAdults | 27.6975 | 3.8415 |
| $\mathbf{z}$ =NAge06 | 0.3562 | 3.8415 |
| $\mathbf{z}$ =NAge712 | 8.5999 | 3.8415 |
| $\mathbf{z}$ =NAge1317 | 6.0481 | 3.8415 |
| $\mathbf{z} = x^c$ | 60.5461 | 14.0671 |

Notes:

[a] $P[\epsilon \leq t] = \Phi\left(t + \beta_1 t^2 + \beta_2 t^3\right)$ ; $H_0 : \beta_1 = \beta_2 = 0$

[b] $V[\epsilon_i | x_i] = \sigma_0^2 \exp\left(\alpha' \mathbf{z}_i\right)$ ; $H_0 : \alpha = 0$

[c] Full specification; includes all regressors ($x_i$) except the constant

**Table 3: Estimation Results (Specification 2)**

| | Ordered Probit | | Ichimura's SLS | | Smoothed Maximum Score | |
|---|---|---|---|---|---|---|
| | Coef. | St.er. | Coef. | St. er. | Coef. | St. er. |
| Constant | -7.738 | 0.283 | - | - | 3.121 | 0.665 |
| Dummy East | -0.253 | 0.029 | -0.193 | 0.021 | -0.025 | 0.865 |
| Log(age) | 0.441 | 0.038 | 0.327 | 0.029 | 0.355 | 0.167 |
| Log(income) | 1.122 | 0.031 | 1 | - | 1 | - |
| DMarried | 0.044 | 0.034 | 0.036 | 0.025 | 0.004 | 0.185 |
| NAge06 | -0.044 | 0.025 | -0.077 | 0.019 | -0.024 | 0.078 |
| NAge712 | -0.156 | 0.026 | -0.127 | 0.018 | 0.169 | 0.139 |
| NAge1317 | -0.179 | 0.028 | -0.168 | 0.022 | 0.151 | 0.142 |
| NAdults | -0.311 | 0.023 | -0.285 | 0.015 | -0.259 | 0.102 |
| | Linear Model (OLS) | | Partially Linear Model | | Gen. Partially Linear Model | |
| | Coef. | St.er. | Coef. | St. er. | Coef. | St. er. |
| Constant | -14.781 | 0.541 | - | - | - | - |
| Dummy East | -0.502 | 0.058 | -0.512 | 0.057 | -0.304 | 0.029 |
| Log(age) | 0.852 | 0.077 | 0.859 | 0.077 | 0.239 | 0.013 |
| Log(income) | 2.268 | 0.059 | - | - | - | - |
| NAge06 | -0.096 | 0.051 | -0.094 | 0.050 | -0.053 | 0.023 |
| NAge713 | -0.301 | 0.050 | -0.303 | 0.050 | -0.104 | 0.025 |
| NAge1317 | -0.365 | 0.057 | -0.362 | 0.057 | -0.127 | 0.029 |
| NAdults | -0.612 | 0.045 | -0.607 | 0.045 | -0.128 | 0.017 |

**Table 4: Equivalence Scales**

| | Ordered Probit | | Ichimura's SLS | | Smoothed Maximum Score | |
|---|---|---|---|---|---|---|
| | Coef. | St.er. | Coef. | St. er. | Coef. | St. er. |
| 1 person | 1 | - | 1 | - | 1 | - |
| *Specification 1* | | | | | | |
| 2 persons | 1.342 | 0.020 | 1.368 | 0.017 | 1.364 | 0.036 |
| 3 persons | 1.593 | 0.037 | 1.644 | 0.033 | 1.636 | 0.068 |
| 4 persons | 1.800 | 0.053 | 1.872 | 0.048 | 1.860 | 0.098 |
| *Specification 2* | | | | | | |
| Single + 1 ch. 0-6 | 1.032 | 0.022 | 1.080 | 0.021 | 1.025 | 0.080 |
| Single + 1 ch. 7-12 | 1.143 | 0.026 | 1.135 | 0.021 | 0.845 | 0.117 |
| Single + 1 ch. 13-17 | 1.169 | 0.030 | 1.182 | 0.026 | 0.860 | 0.123 |
| Couple | 1.302 | 0.020 | 1.283 | 0.027 | 1.290 | 0.117 |
| Couple + 1 ch. 0-6 | 1.344 | 0.034 | 1.386 | 0.035 | 1.322 | 0.111 |
| Couple + 1 ch. 7-12 | 1.488 | 0.040 | 1.457 | 0.036 | 1.090 | 0.119 |
| Couple + 1 ch. 13-17 | 1.522 | 0.042 | 1.517 | 0.041 | 1.109 | 0.230 |

| | Linear Model (OLS) | | Partially Linear Model | | Gen. Part. Lin. Model | | Nonparametric[a] Model | | Number of |
|---|---|---|---|---|---|---|---|---|---|
| | Coef. | St.er | Coef. | St. er. | Coef. | St. er. | Coef. | St. er. | Observ. |
| 1 person | 1 | - | 1 | - | 1 | - | 1 | - | |
| *Specification 1* | | | | | | | | 1747 | |
| 2 persons | 1.326 | 0.019 | 1.388 | 0.093 | 1.364 | 0.049 | 1.448 | 0.171 | 2250 |
| 3 persons | 1.564 | 0.036 | 1.688 | 0.102 | 1.636 | 0.084 | 1.981 | 0.248 | 1253 |
| 4 persons | 1.758 | 0.051 | 1.923 | 0.137 | 1.982 | 0.120 | 2.104 | 0.144 | 1102 |
| *Specification 2* | | | | | | | | | |
| Single + 1 ch. 0-6 | 1.033 | 0.051 | 1.059 | 0.084 | 1.104 | 0.062 | 1.807 | 0.331 | 76 |
| Single + 1 ch. 7-12 | 1.135 | 0.056 | 1.173 | 0.106 | 1.228 | 0.091 | 1.980 | 0.701 | 74 |
| Single + 1 ch. 13-17 | 1.169 | 0.067 | 1.207 | 0.079 | 1.299 | 0.079 | 2.129 | 0.417 | 27 |
| Couple | 1.286 | 0.029 | 1.352 | 0.073 | 1.299 | 0.066 | 1.446 | 0.758 | 2036 |
| Couple + 1 ch. 0-6 | 1.329 | 0.034 | 1.414 | 0.094 | 1.462 | 0.082 | 1.683 | 0.112 | 337 |
| Couple + 1 ch. 7-12 | 1.460 | 0.038 | 1.561 | 0.100 | 1.555 | 0.088 | 2.079 | 0.149 | 205 |
| Couple + 1 ch. 13-17 | 1.503 | 0.041 | 1.610 | 0.112 | 1.656 | 0.127 | 2.089 | 0.381 | 186 |

Note: [a] Model without Dummy East, DMarried and Log(age); cf. Figure 4.

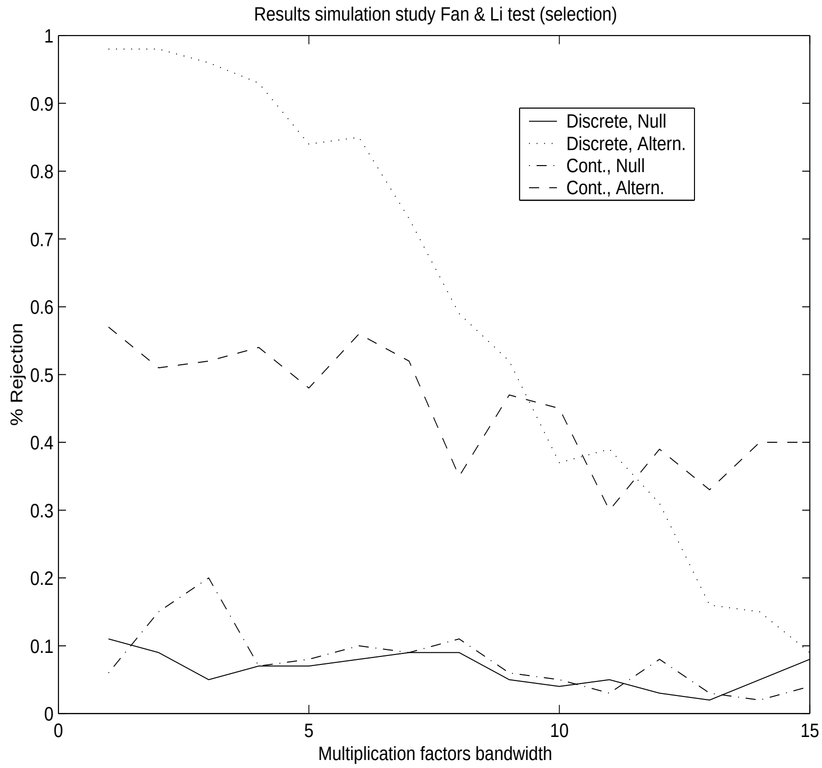


**Figure 1. Simulated rejection probabilities Fan and Li test**

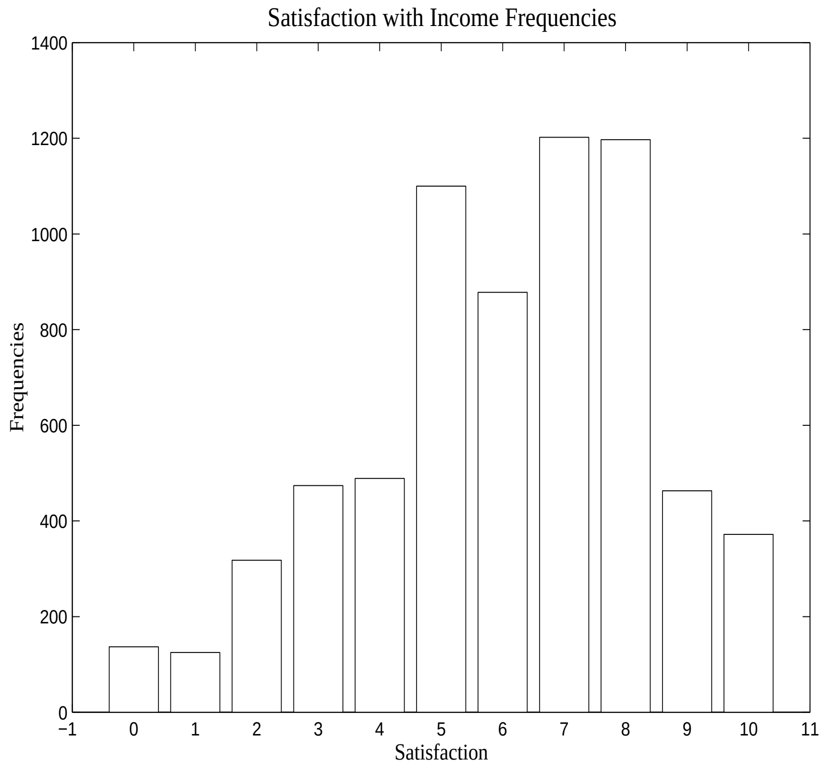


**Figure 2. Distribution of satisfaction with income**

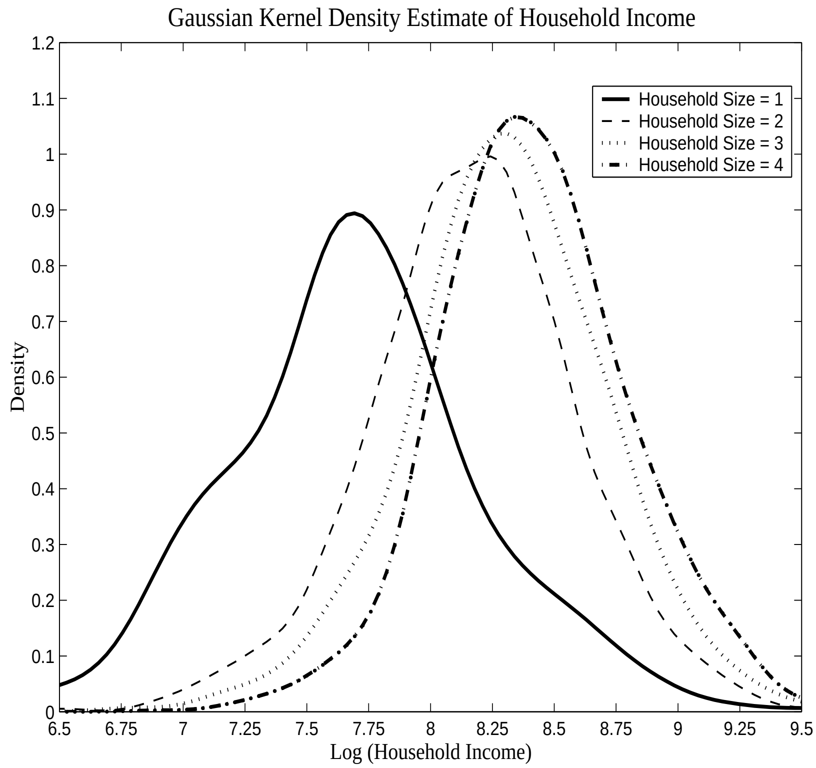


**Figure 3. Distribution of log household income by household size**

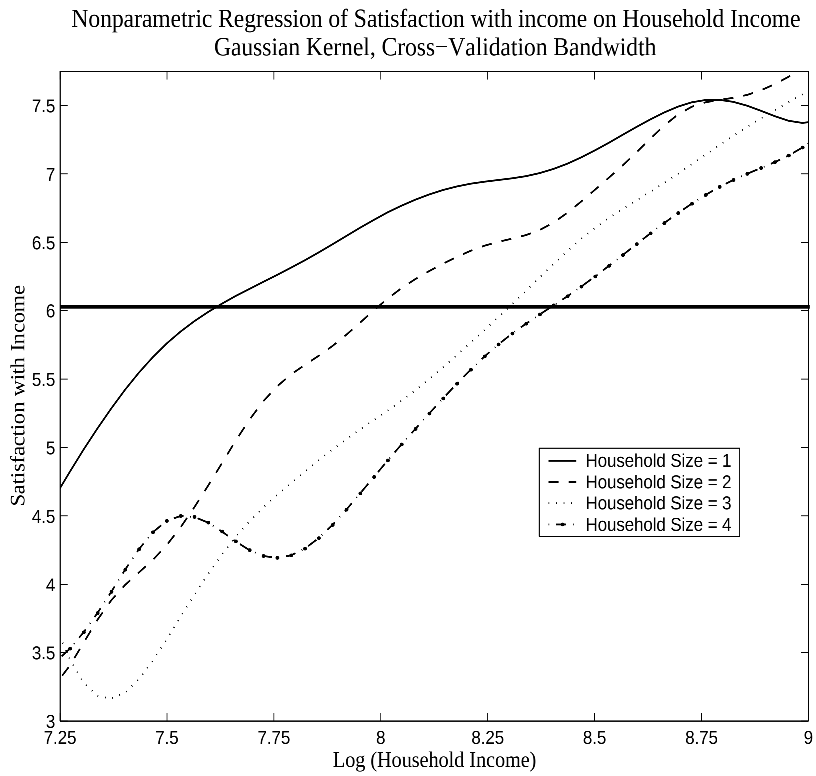


**Figure 4. Nonparametric regression of satisfaction with income on log income by household size**

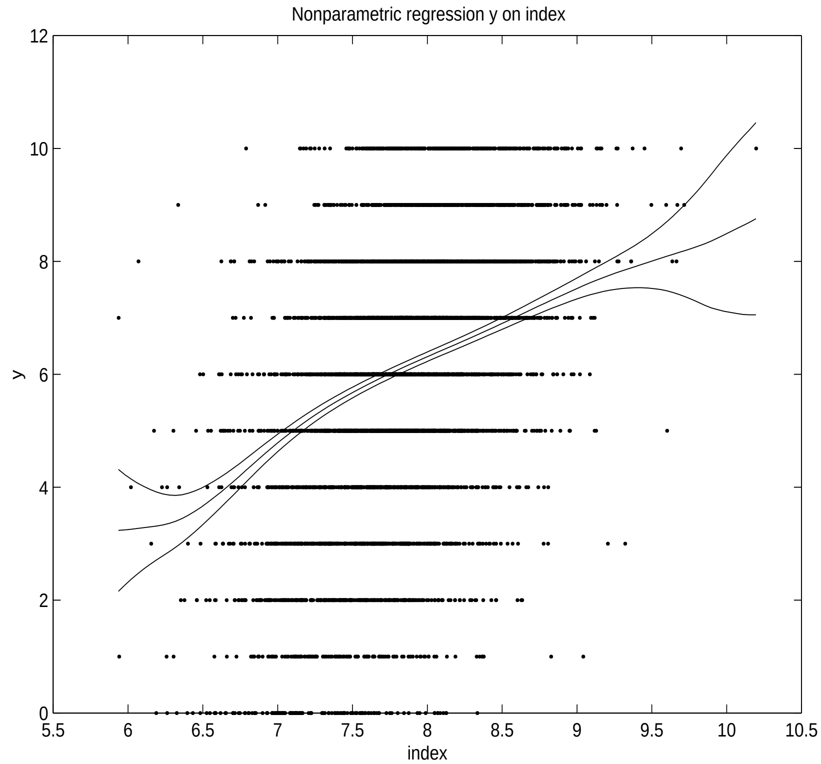


**Figure 5. Non-parametric estimate of link function, with 95% uniform confidence bands. Ichimura's SLS; specification 2**

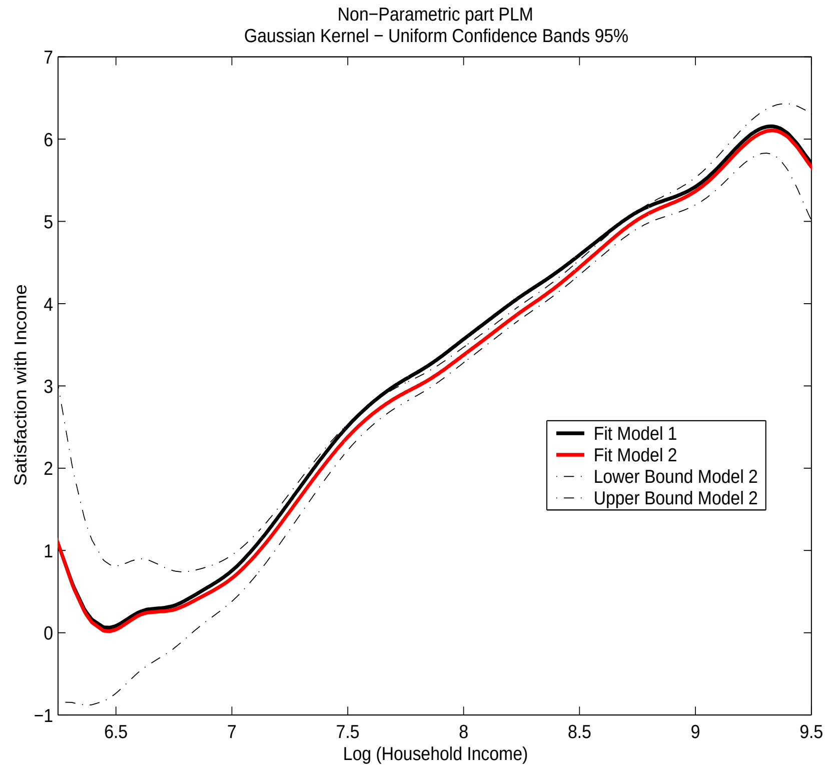


**Figure 6. Non-parametric part partially linear models**

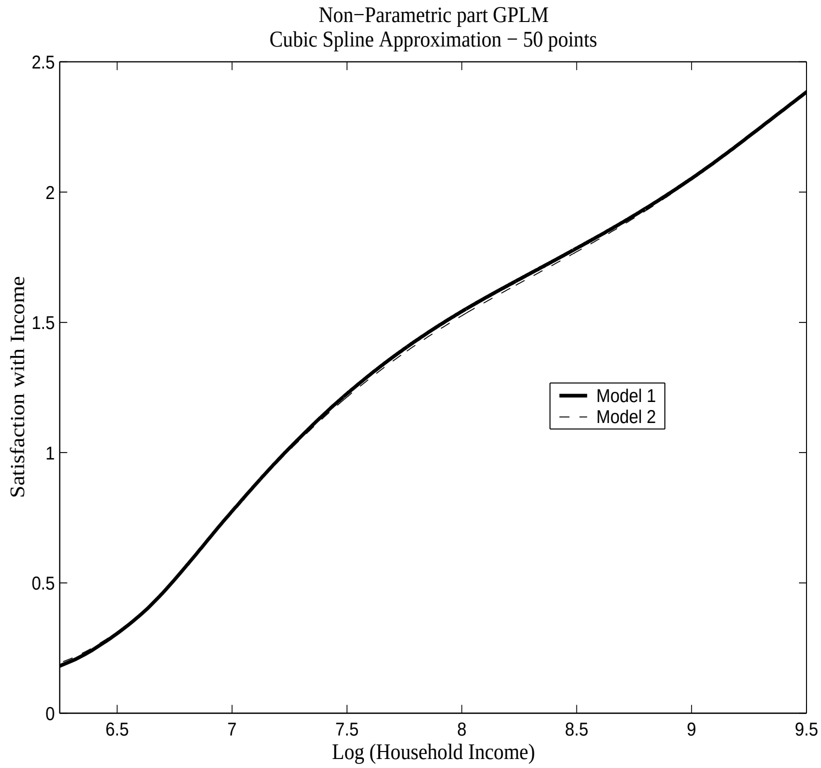


**Figure 7. Non-parametric part generalized partially linear models**