Nonparametric Bounds on the Income Distribution in the Presence of Item Nonresponse

Vazquez-Alvarez, R.; Melenberg, B.; van Soest, A.H.O.

Publication date:
1999

Citation for published version (APA):
Vazquez-Alvarez, R., Melenberg, B., & van Soest, A. H. O. (1999). *Nonparametric Bounds on the Income Distribution in the Presence of Item Nonresponse*. (CentER Discussion Paper; Vol. 1999-33). Econometrics.

Nonparametric Bounds on the Income Distribution in the Presence of Item Nonresponse[1]

Rosalia Vazquez Alvarez,
Bertrand Melenberg,
Arthur van Soest

Corresponding author:
Rosalia Vazquez Alvarez
Department of Econometrics
Tilburg University
P.O.Box 90153
5000 LE Tilburg
The Netherlands
E-mail: r.vazquez-alvarez@kub.nl

Abstract

Item nonresponse in micro surveys can lead to biased estimates of the parameters of interest if such nonresponse is nonrandom. Selection models can be used to correct for this, but parametric and semiparametric selection models require additional assumptions. Manski has recently developed a new approach, showing that, without additional assumptions, the parameters of interest are identified up to some bounding interval. In this paper, we apply Manski's approach to estimate the distribution function and quantiles of personal income, conditional on given covariates, taking account of item nonresponse on income. Nonparametric techniques are used to estimate the bounding intervals. We consider worst case bounds, as well as bounds which are valid under nonparametric assumptions on monotonicity or under exclusion restrictions.

Key words: nonparametrics, bounds and identification, sample non-response
JEL Classification: C14, D31.

[1] We are grateful to Rob Alessie, Myoung-Jae Lee, Erwin Charlier and Bas Donkers for their help. Gauss programs used in this paper are available from the corresponding author.

1 Introduction

A problem often encountered in the collection of survey data, is that of nonresponse or missing values. Nonresponse occurs if individuals do not answer fully to the questionnaire - item nonresponse -, or if some of the individuals who are asked to fill in the questionnaire, do not answer at all - unit nonresponse. This paper focuses on item nonresponse. In household surveys, typical variables that suffer from item nonresponse are income, earnings, or measures of wealth. While people are usually willing and able to disclose information on family composition, labour market status, etc., many people do not provide full information on the level of their earnings, income or wealth.

Item nonresponse implies that for a non-negligible number of respondents, the realization of the variable of interest is either missing, or is registered as missing by the researcher, because the information given is inconsistent with other information provided by the respondent. We focus on the case of item nonresponse in one variable *Y*, for which we are interested in some feature of the conditional distribution $F_{Y|X}$, where *X* is a set of covariates. We neither address unit nonresponse, nor item nonresponse on *X*.

Item nonresponse can be seen as an example of the sample selection problem. If item nonresponse is not completely random, the full response sample is not representative for the population of interest. The traditional approach until about 20 years ago was to avoid this problem by assuming that nonresponse was completely random. This has changed since the seminal work by Heckman (Heckman, 1979, for example). Since then, a huge literature on parametric and semiparametric selection models has appeared. See Vella (1998) for a recent overview. A classical example of how the selection bias can affect the results is found in Mroz (1987). He analyzes various models for females' hours of work. The results show that using selection models to control for selectivity bias can lead to wage and income effects which are substantially different from those obtained with models which do not account for selectivity.

In most applications of selection models, the assumption is made that some location measure $m(Y|X)$ of *Y* conditional on *X* is a linear combination $X'\beta$ of the covariates. Usually, $m(Y|X)$ is the conditional mean $E[Y|X]$ or some conditional quantile. The slope coefficients in β are then the parameters of interest. Mroz (1987) and most other applied studies use parametric selection models, in which distributional assumptions are made on the error terms. If the distributional assumptions are violated, estimates of β will in general still be biased. Semiparametric estimators have been developed to obtain consistent estimates of β under less stringent assumptions on the errors. Examples are Newey et al. (1990) and Ahn and Powell (1993). Both assume that $E[Y|X] = X'\beta$ and focus on estimating β. Both also need the exclusion restriction assumption that at least one given variable affects the selection probability but not $E[Y|X]$. Approaches to the sample selection problem, therefore, allow for weaker distributional assumptions than parametric models, but still retain various restrictive assumptions on the data generating process.

Since the early 1990's, a new approach to deal with the selection problem has been developed. It focuses on nonparametric identification without additional assumptions such as those in parametric or semiparametric selection models. This approach is usually concerned with the full conditional distribution function of *Y* given *X*. See Manski (1989, 1990, 1994, 1995, 1997), but also, for example, Heckman (1990). The idea is to use nonparametrics, imposing no assumptions, or much weaker assumptions than in the parametric or semiparametric literature, together with the concept of *identification up to a bounding interval.* Manski (1995) shows that, without additional assumptions, the sampling process fails to fully identify most features of the conditional distribution of *Y* given *X*, but that in many cases a lower bound and an upper bound for the feature of interest can be derived. For example, suppose we are interested in $F(y) = P(Y \leq y)$ for some given $y \in \mathbb{R}$ (no conditioning variables). Let δ be a binary random variable that takes the value 1 if *Y* is observed, and 0 otherwise. Then we can write

$$F(y) \;=\; P(Y \leq y | \delta = 1) P(\delta = 1) \;+\; P(Y \leq y | \delta = 0) P(\delta = 0) \qquad \textbf{(1)}$$

The data can identify $P(Y \leq y|\delta = 1)$, $P(\delta = 1)$ and $P(\delta = 0) = 1 - P(\delta = 1)$. These population parameters can be estimated straightforwardly from the sub-sample with $\delta = 1$ or from the complete sample, respectively. But the data are not informative about $P(Y \leq y|\delta = 0)$, the distribution function of *Y* for the non-respondents. If we assume completely random nonresponse, then $P(Y \leq y|\delta = 1) = P(Y \leq y|\delta = 0)$ and the identification problem is solved. If we are not prepared to make this or other assumptions, however, all we know is that $0 \leq P(Y \leq y|\delta = 0) \leq 1$. This leads to the following lower and upper bounds on $F(y)$.

$$P(Y \leq y | \delta = 1) P(\delta = 1) \;\leq\; F(y) \;\leq\; P(Y \leq y | \delta = 1) P(\delta = 1) + P(\delta = 0) \qquad \textbf{(2)}$$

These are Manski's 'worst case' bounds on the distribution function, which can easily be extended for conditioning on covariates *X* (see below). Manski (1995) shows how these worst case bounds can be improved upon by adding nonparametric assumptions of monotonicity or exclusion restrictions. In Manski (1994), he also shows how the same ideas can be used to derive bounds on (conditional) quantiles of *Y*, or on the (conditional) mode of *Y*.

The purpose of this paper is to apply the approach of Manski and to examine this approach in an empirical application. We study the conditional distribution of gross personal income using a survey of households in the Netherlands, drawn in 1993. Our sample consists of 2207 adult respondents - heads of households and their partners ; 8% of them do not declare their personal income. We look at the conditional distribution function and at conditional quantiles. We also derive two sets of bounds for the conditional mode. We do not only present point estimates of the bounds, but also construct confidence bands, allowing us to compare the imprecision due to the nonresponse problem and the imprecision due to finite sample error. We nonparametrically estimate worst case bounds, bounds under a monotonicity assumption, and bounds under exclusion restrictions. We focus particularly on the latter, since this has received little or no attention in earlier applications. In particular, we find that in many cases, imposing exclusion restrictions leads to lower and upper bounds which are not compatible with each other, implying that the exclusion restrictions are not supported by the data. This leads to an informal way of testing the exclusion restrictions.

The remainder of this paper is organized as follows. Section 2 reviews Manski's framework. Section 3 describes the estimation method. Section 4 describes the data. Section 5 presents the empirical results. Section 6 concludes the paper.

2 Theoretical framework

2.1 Bounds on the distribution function

In this section we review the theory of Manski (1994,1995) on bounds for a (conditional) distribution function. The aim is to obtain the value of the conditional distribution function defined by,

$$F_{Y|x}(y) \;\equiv\; P(Y \leq y|x) \qquad \textbf{(3)}$$

at given $y \in \mathbb{R}$, and given $X{=}x \in \mathbb{R}^k$. Introduce a dummy variable that models item nonresponse (or, in other words, sample selection):

$$\begin{aligned}&\delta=1 \;\; \textit{if } \; Y \; \textit{is observed}\\ &\delta=0 \;\; \textit{if } \; Y \; \textit{is missing}\end{aligned} \qquad \textbf{(4)}$$

The conditional distribution of Y can be expressed as follows.

$$F_{Y|x}(y) \;=\; F_{Y|(x,\delta=1)}(y)P(\delta=1|x) \;+\; F_{Y|(x,\delta=0)}(y)P(\delta=0|x) \qquad \textbf{(5)}$$

where $F_{Y|(x,\delta=1)}(y) = P(Y \le y|x,\delta=1)$ and $F_{Y|(x,\delta=0)}(y) = P(Y \le y|x,\delta=0)$.We assume that item nonresponse on *Y* is the only problem; there is no nonresponse in *X* and no unit nonresponse, and there are no measurement errors such as under or over reporting the value of *Y*. This means that for all *x* in the support of *X*, $F_{Y|(x,\delta=1)}(y)$ is identified. If *X* is continuous, $F_{Y|(x,\delta=1)}(y)$ can be estimated using a nonparametric regression estimator; see Section 3. Similarly, $P(\delta=1|x)$ and $P(\delta=0|x)$ are identified and can be estimated consistently, since by our assumptions, there is complete response on δ and *X*.

If δ is independent of *Y* conditional on *X*, then $F_{Y|(x,\delta=1)}(y) = F_{Y|(x,\delta=0)}(y)$ and all expressions in the right hand side of (5) are identified. This is the case of conditional independence of nonresponse and variable of interest, also referred to as *exogenous sampling* or *exogenous nonresponse*. It is the basis of the traditional approach to selection models and imputation methods, but also for the matching literature (see, for example, Rosenbaum and Rubin 1984). In general, however, δ can be related to *Y*, and $F_{Y|(x,\delta=0)}(y)$ is not identified, so that $F_{Y|x}(y)$ is not identified either.

The method proposed here, aims at bounding $F_{Y|x}$, using various types of prior assumptions: no additional assumptions (i.e., 'worst case'), monotonicity, or exclusion restrictions.

Worst case bounds

With no additional assumptions, all we know is

$$0 \le F_{Y|x,\delta=0}(y) \le 1 \qquad \textbf{(6)}$$

With (5) this implies

$$F_{Y|(x,\delta=1)}(y)P(\delta=1|x) \le F_{Y|x}(y) \le F_{Y|(x,\delta=1)}(y)P(\delta=1|x)+P(\delta=0|x) \qquad \textbf{(7)}$$

Manski shows that the lower and upper bound in (7) cannot be improved upon without making additional assumptions which is why he named them worst case bounds. The width of the interval between the bounds is $P(\delta=0|x)$, the conditional percentage of nonresponse. Thus, as intuitively expected, the larger the probability of nonresponse, the less information can be retrieved from the data, and the wider the interval. The other bounds use additional information to reduce the distance between the bounds.

Bounds under a monotonicity assumption

In many cases, it may be reasonable to impose a priori that those with a high value of *Y* are more likely to be non respondents than those with low values of the dependent variable.[2] For example, suppose *Y* is income. It is often claimed that item nonresponse on income is positively correlated with income, since high income earners are less willing to disclose their income. This monotonicity assumption implies that $P(\delta=1|Y\leq y,x)\geq P(\delta=1|x)$ and $P(\delta=0|Y\leq y,x)\leq P(\delta=0|x)$, and thus, using Bayes' rule

$$P(Y\leq y|x,\delta=0) \;\leq\; P(Y\leq y|x,\delta=1) \qquad \textbf{(8)}$$

or

$$0 \;\leq\; F_{Y|(x,\delta=0)}(y) \;\leq\; F_{Y|(x,\delta=1)}(y) \qquad \textbf{(9)}$$

Applying (9) to (5) leads to the following upper and lower bounds under monotonicity

$$F_{Y|(x,\delta=1)}(y)P(\delta=1|x) \;\leq\; F_{Y|x}(y) \;\leq\; F_{Y|(x,\delta=1)}(y) \qquad \textbf{(10)}$$

Compared to (7), the upper bound is reduced by $P(\delta=0|x)[1-F_{Y|(x,\,\delta=1)}(y)]$.The reduction of the width between upper and lower bound due to imposing this assumption of monotonicity on the conditional distribution function is largest in the left tail of the income distribution.

Bounds with Exclusion Restrictions

In parametric and semiparametric selection models, it is usually assumed that the conditional distribution of *Y* given *X* depends on a subset of the covariates only. Assume that the vector *x* can be decomposed into two sets of variables, $x=(m\,,\nu)$. An exclusion restriction on ν means that $P(Y\leq y|(m\,,\nu))$ does not vary with ν , so that it can be written as $P(Y\leq y|m)$. Applying this to (7) for given *m* and *y* but for all values of ν results in the following bounds under the exclusion restrictions

[2] There might also be examples where the opposite is a reasonable assumption, of course. This can be treated analogously. We do not work this out here since it seems less relevant for our empirical application.

$$\begin{gathered} sup_{\nu}[F_{Y|(m,\nu,\delta=1)}(y)P(\delta=1|(m,\nu))] \\ \leq\ F_{Y|(m)}(y)\ \leq \\ inf_{\nu}[F_{Y|(m,\nu,\delta=1)}(y)P(\delta=1|(m,\nu))+P(\delta=0|(m,\nu))] \end{gathered} \qquad \textbf{(11)}$$

Again, these bounds use prior assumptions, and, therefore, generally result in tighter bounds than (7). Note that even if the probability of response $P(\delta=1|(m,\nu))$ does not depend on ν , the bounds in (11) may still be more informative than those in (7), as long as $F_{Y|(m,\nu,\delta=1)}(y)$ - and thus also $F_{Y|(m,\nu,\delta=0)}(y)$ - vary with ν . This is in contrast with the situation in semiparametric selection models, where it is usually assumed that ν does play a role in the selection mechanism. Nothing tells us whether the bounds in (11) are tighter or less tight than those in (10). This will depend on the empirical application considered.

Combining exclusion restrictions and monotonicity

If both types of prior assumptions are imposed simultaneously, it is straightforward to derive the following bounds

$$\begin{gathered} sup_{\nu}[F_{Y|(m,\nu,\delta=1)}(y)P(\delta=1|(m,\nu))] \\ \leq\ F_{Y|(m)}(y)\ \leq \\ inf_{\nu}[F_{Y|(m,\nu,\delta=1)}(y)] \end{gathered} \qquad \textbf{(12)}$$

2.2 Bounds on conditional quantiles

Income distributions are often described in terms of quantiles. It is therefore interesting to apply the same framework to identify conditional quantiles in case of item nonresponse. In what follows, expressions (7), (10), (11) and (12) are used to obtain analogous expressions for the conditional quantiles of the distribution. This draws on Manski (1994).

For $\alpha \in [0,1]$, the α -quantile of the conditional distribution of Y given $X=x$, is the smallest number $q(\alpha, x)$ that satisfies $F_Y[q(\alpha, x)] \geq \alpha$,:

$$q(\alpha,x)\ \equiv\ inf\ \{y\text{: } F_{Y|x}(y) \geq \alpha\ \} \qquad \textbf{(13)}$$

For $\alpha > 1$, $q(\alpha, x) = \infty$, and for $\alpha < 0$, $q(\alpha, x) = -\infty$. The α-quantile of the conditional distribution of Y given $X = x$ and $\delta = 1$ will be denoted by $q_1(\alpha, x)$.

The bounds for the quantiles follow from those for the distribution functions by ‘inverting’

(7), (10), (11) and (12).These can all be written as

$$L(y,x) \;\leq\; F_{Y|x}(y) \;\leq\; U(y,x) \tag{14}$$

for different choices of *L(y, x)* and *U(y, x),* all of them non-decreasing functions of *y*. Inverting this gives:

$$inf\{y{:}L(y,x)\geq\alpha\} \;\geq\; inf\{y{:}F_{Y|x}(y)\geq\alpha\}\geq inf\{y{:}U(y,x)\geq\alpha\} \tag{15}$$

Worst Case Bounds on Conditional Quantiles

Applying (15) for *L(y, x)* and *U(y, x)* given in (7) and using the quantiles of $F_{Y|x,\delta=1}$ gives the following worst case bounds.

$$q_1\left(1-\frac{(1-\alpha)}{P(\delta=1|x)},x\right) \;\leq\; q(\alpha,x) \;\leq q_1\left(\frac{\alpha}{P(\delta=1|x)},x\right) \tag{16}$$

The lower bound is informative only if $(1\text{-}\alpha)\leq P(\delta=1|x)$ and it is -∞ otherwise. Similarly, the upper bound is informative only if $\alpha\leq P(\delta=1|x)$. The width of the bounding interval for the quantiles varies with α and depends on the slope of $F_{Y|(x,\ \delta=1).}$ It is no longer simply determined by the probability of nonresponse as was the case in (7).

Bounds for conditional quantiles under monotonicity

Applying (15) to (10) leads to

$$q_1(\alpha,x) \;\leq\; q(\alpha,x) \;\leq\; q_1\left(\frac{\alpha}{P(\delta=1|x)},x\right) \tag{17}$$

Note that the lower bound in (17) exceeds the lower bound in (16) since $\alpha > \left[1-\frac{1-\alpha}{P(\delta=1|x)}\right]$. Thus, imposing monotonicity helps to tighten the bounds.

Bounds for conditional quantiles under exclusion restrictions

Applying (15) to (11) gives

$$\begin{aligned} & sup_{\nu}\ q_1\left(1-\frac{(1-\alpha)}{P(\delta=1|m,\nu)},(m,\nu)\right) \le \\ & \le\ q(\alpha,x)\ \le \\ & \le\ inf_{\nu}\ q_1\left(\frac{\alpha}{P(\delta=1|m,\nu)},(m,\nu)\right) \end{aligned} \qquad \textbf{(18)}$$

Combining exclusion restrictions and monotonicity

Finally, applying (15) to (12) gives

$$sup_{\nu}\ q_1\big(\alpha,(m,\nu)\big)\ \le q(\alpha,x) \le\ inf_{\nu}\ q_1\left(\frac{\alpha}{P(\delta=1|m,\nu)},(m,\nu)\right) \qquad \textbf{(19)}$$

2.3 Bounds on the conditional mode

Drawing from Manski (1994, p.153-156) we derive bounds for the so called η-mode of the conditional distribution function $F_{Y|x}$. Define the loss function $h_\eta(y, b)=I[\ |y-b|>\eta\ ]$, for $b\in\mathbb{R}$ and $\eta>0$. The conditional expectation of $h_\eta(y, b)$ is given by

$$E[h_\eta(Y,b)|x]=P(|Y-b|>\eta|x) \qquad \textbf{(20)}$$

The η-mode of $F_{Y|x}$, denoted by $\bar{b}(\eta,x)$, is the value of b for which this conditional expectation is minimized (see also Lee, 1996)

$$\bar{b}(\eta,x)=argmin_b E[h_\eta(y,b)|x] \qquad \textbf{(21)}$$

If $F_{Y|x}$ has a unimodal density $f_{Y|x}$, and if η is sufficiently small, then $\bar{b}(\eta,x)$ will approximate the mode of the conditional distribution function.

To derive the bounds on the η-mode in case of item nonresponse, rewrite the expected loss function as

$$E[h_\eta(Y,b)|x] = E[h_\eta(Y,b)|x,\delta=1]P(\delta=1|x) + E[h_\eta(Y,b)|x,\delta=0]P(\delta=0|x) \qquad \textbf{(22)}$$

The data does not provide any information on $E[h_\eta(Y,b)|x,\delta=0]$. All we know is that it is between 0 and 1. This implies

$$\begin{aligned} &E[h_\eta(Y,b)|x,\delta=1]P(\delta=1|x) \;\le \\ &\le\; E[h_\eta(Y,b)|x] \;\le \\ &\le\; E[h_\eta(Y,b)|x,\delta=1]P(\delta=1|x) + P(\delta=0|x) \end{aligned} \qquad \textbf{(23)}$$

Combining (21) and (23) shows that $\overline{b}(\eta,x)$ has to satisfy

$$E[h_\eta(Y,\overline{b}(\eta,x))|x,\delta=1] \;\le\; inf_b\; E[h_\eta(Y,b)|x,\delta=1] + \frac{P(\delta=0|x)}{P(\delta=1|x)} \qquad \textbf{(24)}$$

Condition (24) defines some subset of possible $\overline{b}(\eta,x)$. It can be called the worst case subset for the η-modes; it is not necessarily an interval.

The monotonicity assumption discussed in Sections 2.1 and 2.2 does not provide additional information on the η-modes since monotonicity says nothing about the slope of the distribution function. On the other hand, the idea of using exclusion restrictions does lead to a new subset of possible η-modes. As in Sections 2.1 and 2.2, assume that $x=(m,\nu)$, and that $F_{Y|(m,\nu)}$ does not depend on the vector ν. From (23) we then get

$$\begin{aligned} &sup_\nu \left(E[h_\eta(Y,b)|m,\nu,\delta=1]P(\delta=1|m,\nu)\right) \;\le \\ &\le\; E[h_\eta(Y,b)|m] \;\le \\ &\le\; inf_\nu \left(E[h_\eta(Y,b)|m,\nu,\delta=1]P(\delta=1|m,\nu) + P(\delta=0|m,\nu)\right) \end{aligned} \qquad \textbf{(25)}$$

This implies that under the exclusion restriction on ν, $\overline{b}(\eta,x) = \overline{b}(\eta,m)$ has to satisfy

$$\begin{aligned} & sup_{\nu}\ E[h_{\eta}(Y,\overline{b}(\eta,m))|m,\nu,\delta=1] \ \leq \\ & \leq\ inf_{\nu,b}\left(E[h_{\eta}(Y,b)|m,\nu,\delta=1]+\frac{P(\delta=0|m,\nu)}{P(\delta=1|m,\nu)}\right) \end{aligned} \qquad \textbf{(26)}$$

For a given η and m, the subset of possible η-modes defined by (26) is a subset of the set defined by (24).

3 Estimation Methods

3.1 Estimating the bounds on the distribution function

The bounds on values of the distribution function in Section 2.1 are all in terms of characteristics of the population. We need to estimate them using the available sample. In general, the bounds are functions of conditional expectations of observed quantities, which can be estimated by nonparametric regression estimators. For example, (7) contains three conditional expectations to be estimated: $F_{Y|x,\,\delta=1}(y) = E[I(Y\leq y)|\ x,\ \delta=1]$, $P(\delta=1|x) = E[\delta|x]$ and $P(\delta=0|x) = E[1\text{-}\delta|x]$. For all cases, we use kernel estimators (see Härdle and Linton, 1994, for example), either based upon the sub-sample with $\delta=1$ or upon the whole sample. The vector of covariates x typically contains discrete variables with a finite number of possible outcomes, as well as continuous variables. This implies that the kernel estimator is basically a nonparametric regression on the continuous variables for each separate cell determined by the values of the discrete variables. The rate of convergence only depends upon the number of continuous variables (see Bierens, 1987, for example). We use kernels which are products of Gaussian kernels. The bandwidth is determined by cross-validation following Härdle and Marron (1985).[3] Similar techniques are applied to obtain estimates of (10), (11) and (12). For the latter two expressions upper bounds are minimized and lower bounds are maximized with respect to the variables chosen as exclusion restrictions.

The bounds in (7) and (10) can also be written directly as conditional expectations of

[3] We have replaced $I[Y\leq\ y]$ by $\Phi[(y\text{-}Y)/h_y]$ in the nonparametric regression determining $E[I(Y\leq\ y)|x]$, where Φ is the standard normal cumulative distribution function and h_y is a smoothness parameter set equal to $0.05\hat{\sigma}(y)n^{(-0.2)}$, where $\hat{\sigma}(y)$ stands for the sample standard deviation of the dependent variable. This replacement does not affect the estimation results but leads to smoother curves in the figures.

appropriate functions of Y and δ.[4] Therefore, it is straightforward to derive analytical expressions for their (pointwise) asymptotic distributions, and to construct explicit consistent estimators for the asymptotic biases and asymptotic covariance matrices (see Härdle and Linton, 1994, for example). This is not the case for the bounds in Section 2.1, given in (11) and (12): these expressions require taking the maximum and minimum over a collections of nonparametric estimates and the sampling distribution of these estimates is not yet well understood. We therefore use a naive bootstrap procedure to find sets of confidence bands. This particular bootstrap method consists of re-sampling randomly 500 times from the original sample with replacement, to obtain two sided 95% pointwise confidence intervals for each of the estimated upper and lower bounds. Notice that these confidence bands are not measuring the error of estimating the unknown distribution function but the error of estimating the upper bound and the lower bound. This means that the vertical distance between upper confidence band of the upper bound and the lower confidence band of the lower bound is an overestimation of the total measurement error for the unknown distribution function. For (7) and (10), we have compared the bootstrapped confidence intervals with confidence intervals based upon the analytical expressions. The results were virtually identical and therefore we only present the bootstrapped intervals for all expressions in Section 2.

3.2 Estimating the bounds on conditional quantiles

The bounds on the conditional quantiles in (16), (17), (18) and (19) can be estimated in two ways. One way is to use estimates $\hat{L}(y,x)$ and $\hat{U}(y,x)$ of the bounds on the distribution function in (14), and determine $inf\ \{y\colon\ \hat{L}(y,x)\geq\alpha\}$ and $inf\ \{y\colon\ \hat{U}(y,x)\geq\alpha\}$. These can be used to replace the population quantiles in (15) and thus provide estimates of the upper and lower bounds on the quantiles of the distribution. Another way is to use that (16)-(19) are based upon conditional quantiles $q_1(\beta,x)$ of the complete response sub-population, where β is some function of the given α and the response probability $P(\delta=1|x)$. Replacing the latter by its nonparametric estimate yields a consistent estimate $\hat{\beta}$ for β . Then $q_1(\beta,x)$ can be estimated after plugging in $\hat{\beta}$ of β and using an existing nonparametric quantile estimator (see Härdle and Linton, 1994). For example, the estimator based upon minimizing a weighted sum of absolute deviations can be used, originating from Koenker and Bassett (1978) and developed further by Chaudhuri (1991). It is given by

$$\hat{q}_1[\hat{\beta},x]=argmin_q \sum_{i=1}^{n} \delta_i K_h(x-x_i)[|y_i-q|+(2\hat{\beta}-1)(y_i-q)] \qquad \textbf{(27)}$$

[4] For example, the right hand side of (7) can also be written as $E[I(\delta=1,\ Y\leq y)+I(\delta=0)|x]$.

For the kernel function K_h, we again use a Gaussian product kernel, and the bandwidth h is determined by cross-validation in an identical way as the choice of bandwidth for the product kernel of the estimated bounds on the distribution function. Using Härdle (1984, Theorem 2.3) it is possible to derive the asymptotic distribution of this quantile estimator for given β. Since β is also estimated here, the limit distribution is considerably more complicated, and, therefore, we use bootstrapped confidence bands applying the same bootstrap technique as described above.

We estimated the quantiles using both techniques described above and found virtually identical results.[5] We present the results based upon the first technique, based upon (14) and (15).

3.3 Estimating bounds on the conditional mode

The conditions which determine possible values of the conditional mode, presented in Section 2.3, are built upon conditional expectations $E[h_\eta(Y,b)|x,\delta=1]$ and the conditional probabilities $P(\delta=1|x)$ and $P(\delta=0|x)$. These can be estimated using the same kernel regression estimators as used for estimating the bounds on the distribution function. The results can be used to obtain estimates for the subset of feasible conditional modes. Since we are not estimating points but sets, we will not aim at estimating the precision with which these sets are determined.

4 The data

The data set used is taken from the 1993 wave of the VSB panel. This panel is a joint venture between the VSB foundation and CentER for Economic Research at Tilburg University.[6] It aims at providing a better understanding of household savings and household financial decision making in the Netherlands. The questions are classified in five categories, namely household characteristics, income and wealth, accommodation and mortgages, assets and loans and finally, a section on psychological questions on attitudes, personality, etc. The panel contains approximately 3000 households with around 9000 respondents of ages 16 and over. It is divided into two sub-panels. One sub-panel contains approximately 2000 households and is designed to be representative of the Dutch population with respect to certain socio-economic variables. The other sub-panel, with approximately 1000 households, should represent households in the top decile of the income distribution. Households in this sub-panel are drawn from high income areas. Since the second sub-panel is obviously not a random sample, we only use the first, representative, sub-panel. The information in both sub-panels is collected by a computerized system. The

[5] The differences are due to the fact that we smooth the distribution function in the first technique. Without this smoothing and using the same kernels, the results would be identical.

[6] For detailed information on the VSB panel, see Nyhus (1996).

participants in the representative sub-panel supplied answers on a weekly basis.

The 2000 households in the representative sub-panel contain about 4500 individuals of working age (age 16 or older). We select only heads of households (including singles) and their permanent partners (married or unmarried). From this selection, we retain a total of 2207 individuals from 1415 households. The remaining 585 households were not included since neither heads nor partners in these households answered the psychological section of the survey; this section contains the conditioning variables used for exclusion restrictions. We include all individuals of working age, i.e. part-time and full-time employees, self-employed, unemployed, students, disabled, pensioners and housewives.

Our dependent variable of interest (Y) is gross personal income. It includes gross earnings for employees, gross profits for self-employed, various government transfers and benefits, and capital income. With this definition, 13.3% of the 2207 individuals have zero income; these are treated as genuine zeros, and should not be confused with income nonresponse. A total of 171 individuals did not provide information on the level of one or more of their income components. Thus the (unconditional) sample probability of item nonresponse $\hat{P}(\delta=0)$ is 7.7%. Table 1 below shows how the 171 nonresponse individuals and the 293 who declare to have zero income are categorized by labour market state.

Table 1: Non respondents and zero incomes by labour market state

	Total % of nonresponse	**Males % of nonresponse**	**Female % of nonresponse**	**Total % with zero income**	**Males %with zero income**	**Female % with zero inc.**
Employed	64.3	36.8	27.5	0	0	0
Self-empl.	5.9	3.5	2.3	1.4	0	1.4
Unemployed	4.1	2.3	1.8	0	0	0
Disabled	2.9	2.9	0	0.7	0	0.7
Pensioners	8.8	7.6	1.2	1.0	0	1.0
Housewives	4.7	2.3	2.3	88.7	0.7	88.1
Students	2.3	1.2	1.2	2.0	0.7	1.4
Volunteers	7.0	4.1	2.9	6.1	0	6.1
TOTAL	**171 units**	**104 units**	**67 units**	**293 units**	**4 units**	**289 units**

The table shows that the large majority of nonresponse individuals are males who are either employed, self-employed or pensioner; on the other hand, zero income will be associated with housewives. Table 1 shows that the nonresponse is associated with employed individuals.

This suggests that nonresponse is associated with the earnings of individuals rather than with any other type of income such as capital income or net transfers.

The covariates (*X*) are age, education measured by an ordered categorical variable, and family size. The psychological section of the questionnaire contains a variety of questions which may affect the individuals' response tendency, without directly determining income. Some of these variables could be used as exclusion restrictions (ν in Section 2). On the basis of some preliminary probits, explaining item nonresponse, we selected the variables WORR, REFG, RISK and DWRK.[7] WORR is based upon a variable which measures the self-perception of how easily the respondent gets worried, in general. The variable REFG is based upon a question on someone's reference group for the household's financial situation. The variable RISK is a measure of risk aversion based upon information on how often the respondent buys lottery tickets. The fourth (DWRK) is a dummy variable measuring whether the individual completely responds to the section of the questionnaire called 'work and pensions' and stands as a general indicator of the respondent's carefulness in answering the questions.

Table 2 is a statistical summary of the conditioning variables and exclusion restriction variables mentioned above for the selected sample of heads and partners. From this table we see that, on average, non-respondents are younger than respondents and are more often male and single. Non-respondents also have higher educational achievement than respondents. People that do not easily get worried (WORR=1) have a larger tendency to respond. This suggests that nonresponse might be related to worrying about privacy. People that do not identify their reference group (REFG=0) and people who do not answer all the questions in the work and pensions questionnaire (DWRK=0) also have a larger tendency not to respond to the income questions. Finally, people who reveal risk aversion in the sense that they do not often play the lottery (RISK=0) are relatively likely not to respond.

Table 2: Means (standard deviations) and percentages (standard errors) for covariates and exclusion restrictions variables.

[7] Appendix A explains how these variables are constructed and presents the exact wording of the underlying questions.

	All Individuals	Respondents	Non respondents
Number of observations	2207	2036	171
Age	47.3 (15.2)	47.8 (15.1)	41.8 (15.3)
Education	2.31 (0.77)	2.3 (0.78)	2.36 (0.73)
% single	19 (0.8)	18.4 (0.9)	26.3 (3.4)
Family size	2.50 (1.28)	2.52 (1.3)	2.3 (1.2)
% male	52.5 (1.1)	51.8 (1.1)	60.8 (3.7)
% home owners	59.2 (1.0)	60 (1.1)	48.5 (3.8)
Gross income	unknown	41,169 (36,621)	unknown
% with zero income	unknown	14.4 (0.8)	unknown
WORR	54.9 (1.1)	55.1 (1.1)	52.0 (3.8)
REFG	47.7 (1.1)	48.6 (1.1)	37.4 (3.7)
RISK	61.8 (1.0)	62.5 (1.1)	53.2 (3.8)
DWRK	81.4 (0.83)	81.9 (0.9)	74.8 (3.3)

5 Results

5.1 Bounds on the distribution function

We present the estimates of the bounds on the income distribution, its quantiles and its mode as discussed in Section 2. In estimating these expressions we use kernels which are products of Gaussian kernels for the three conditioning variables, age, education and family size. The bandwidth for each of these kernels is determined as $h_x = h_o \hat{\sigma}(x) n^{(-0.2)}$ where $\hat{\sigma}(x)$ is the sample standard deviation of the variable and the base bandwidth is h_o = 1.5, determined by least squares cross-validation; the resulting final bandwidth was *h=0.502,* where *h* is calculated as $h = h_{age} h_{education} h_{familysize}$. For any of the estimated sets of bounds, we have conditioned on the mean value of the variables age, education and family size. Figure 1 and Figure 2 present the bounds for the distribution function estimated using expressions (7) and (10), respectively.

*File Contains Data for
PostScript Printers Only*

Figure 1 refers to the worst case bounds - estimated using expression (7) -. The figure contains four curves. The solid curve and the dashed curve are the point estimates of the lower and upper bounds of $F_{Y|\bar{x}}(y)$ respectively, at each income level y, where $\bar{x}$ represents the sample mean of the conditioning set. The dotted curves show the estimated two sided 95% pointwise confidence bands for the upper and lower bound; the figure only shows the upper confidence band for the upper bound and the lower confidence band for the lower bound. The vertical distance between upper and lower bounds at each point of the income distribution is $\hat{P}(\delta=0|\bar{x})=0.057$ and reflects the identification problem due to nonresponse. On the other hand the differences between dotted curves and corresponding bounds reflect imprecision due to finite sampling error. The total vertical distance between the two dotted curves reflects uncertainty due to item nonresponse as well as finite sampling error. The results show that imprecision due to sampling error is certainly as important as the imprecision due to nonresponse. This is different from the example in Manski (1994), where sampling error is relatively unimportant. The difference is due to the limited size of the sample and the three dimensional nonparametric regression, leading to substantial standard errors of the estimates. The sample includes individuals that declare to have zero income; this explains the upward shift of the curves at the zero income point.

File Contains Data for
PostScript Printers Only

Figure 2 presents the bounds on the distribution function under the assumption of monotonicity given in (10). The interpretation of the four curves is similar to that of Figure 1. The lower bound is the same as in Figure 1, only the upper bound differs. Comparison of Figure 1 and Figure 2 clearly shows that the assumption of monotonicity helps to tighten the bounds at the lower end of the income distribution.

File Contains Data for
PostScript Printers Only

Figure 3 shows the bounds of the distribution function according to (11), imposing the exclusion restriction that none of the four variables WORR, RISK, REFG and DWRK affects the income distribution. Thus ν consists of four dummy variables each of wich can take two different values. This implies that in (11) the minimum and the maximum are taken from 16 upper bounds and 16 lower bounds. For each of these 16 cases we have determined a base bandwidth h_0 by cross validation to get the smoothing parameter.

The solid and dashed curves in Figure 3 show that maximizing and minimizing over the potential exclusion restrictions leads to a set of bounds that cross at various points of the distribution. For these values of income the estimated upper bound is below the estimated lower bound and we do not obtain a useful interval for the unknown value of the distribution function. If we only look at the estimated upper and lower bounds, this result suggests that the exclusion restrictions are not supported by the data. On the other hand, the dotted lines suggest that this finding could very well be due to finite sampling error: the upper end points of the confidence band for the upper bound are always above the lower end points of the confidence band for the lower bound. Thus, taking into account the imprecision in the estimates, the conclusion that the data rejects the exclusion restrictions cannot be drawn with sufficient confidence. In other words, we have performed an informal test for the null hypothesis that the exclusion restrictions are valid. The null hypothesis is not rejected. The size of the this test, however, is not clear, since we combine pointwise confidence intervals at different values of income for the lower and the upper bounds. That is why we call the test informal. Although it may be worthwhile to pursue the idea behind this test and develop a formal test for exclusion restrictions in this framework, this is not the aim of this paper.

File Contains Data for
PostScript Printers Only

Figure 4 shows the estimated bounds according to (12) where we have imposed the assumption of monotonicity together with the same exclusion restrictions as in Figure 3. The conclusion is the same as in the previous figure. Point estimates of upper and lower bounds suggest that the joint assumptions of exclusion restrictions and monotonicity are not supported by the data. Taking account of sampling error, however, suggests that this result might not be strong enough to reject the assumptions with large enough confidence.

5.2 Bounds on the quantiles

How informative the bounds in Figure 1 and Figure 2 are, is hard to judge from the figures themselves. Since income distributions are often described in terms of quantiles, it may be easier to interpret the bounds on the quantiles than to interpret the bounds on the distribution function. Figures 5 and 6 present bounds on the quantiles given in (16) and (17) respectively.

File Contains Data for
PostScript Printers Only

As in previous figures the solid and dashed curves represent, respectively, the estimated upper and lower bounds of the quantiles and the dotted curves are the estimated two sided 95% confidence intervals. Again, these figures show that most of the distance between the top and the bottom dotted curves is due to sampling error. The fact that the quantiles are zero for small α is due to the presence of zero incomes in the sample.

File Contains Data for PostScript Printers Only

Table 3: Estimated bounds and confidence interval,(c.i), on income (in Dutch Guilders) based on (16) and (17)

	Worst case c.i	Worst case lower bound	Monotonicity c.i	Monotonicity lower bound	Upper bound	Upper bound c.i
20th Quantile	**311**	**450**	**450**	**1,577**	**3,360**	**11,100**
25th Quantile	**450**	**2,328**	**1,442**	**9,720**	**11,200**	**17,672**
30th Quantile	**2,328**	**10,380**	**9,114**	**15,908**	**17,200**	**22,798**
40th Quantile	**16,197**	**20,363**	**18,656**	**23,325**	**25,300**	**33,312**
50th Quantile	**23,916**	**32,000**	**27,462**	**36,000**	**39,919**	**46,877**
60th Quantile	**38,521**	**44,830**	**41,735**	**46,877**	**50,221**	**56,344**
70th Quantile	**48,665**	**54,918**	**50,728**	**56,400**	**61,107**	**69,208**
75th Quantile	**55,094**	**60,680**	**56,128**	**61,651**	**67,929**	**73,302**
80th Quantile	**59,753**	**67,192**	**61,366**	**68,559**	**74,000**	**85,000**
90th Quantile	**73,844**	**83,400**	**74,537**	**85,000**	**104,950**	**179,000**

Table 3 shows a selection of estimated quantiles with their corresponding 95% pointwise confidence intervals. For example, according to the worst case bounds, the median is between fl.23,916 and fl.46,877 with 95% confidence. Imposing monotonicity reduces the distance between these bounds on the median by approximately fl.3,500**.** Such an improvement due to imposing monotonicity is visible at all the quantiles considered although the improvement is smaller at the higher income quantiles.

Figure 7 and Figure 8 present the quantiles on the income distribution obtained from estimates of expressions (18) and (19). Figure 7 is estimated imposing the four exclusion restrictions WORR, RISK, REFG and DWRK whereas Figure 8 additionally imposes the assumption of monotonicity. Since these estimates are based on the same estimates used to draw Figure 3 and Figure 4, the conclusion we obtain here is the same as before. The lower bound is above the upper bound at various quantiles of the distribution showing that the data does not support the exclusion restrictions underlying expressions (18) and (19).

File Contains Data for
PostScript Printers Only

File Contains Data for
PostScript Printers Only

5.3 Bounds on the Conditional Mode

The worst case bounds for the conditional mode are implicitly given by (24). In figures 9 and 10 we have drawn the two curves that play a role in (24), namely $E[h_\eta(Y,b)|\bar{x},\delta=1]$ and $E[h_\eta(Y,b)|\bar{x},\delta=1] + \frac{P(\delta=0|\bar{x})}{P(\delta=1|\bar{x})}$. The curves have been drawn for income values ranging from fl.0 to fl.400,000 taking steps of fl.2,000; steps of smaller size only lead to less smoothness in the estimated curves but not to different ranges of values for the conditional mode. The results depend on the choice of η. Estimating (24) for small values of η below fl.7,000 results in curves that are not informative about a possible range of values for the conditional mode.

File Contains Data for
PostScript Printers Only

Figure 9 shows the estimates of (24) for η = fl.10,000; in this figure we see that the infimum of the upper loss function identifies a range between fl.0,00 and fl.67,000 for the conditional mode; nevertheless this range is too wide for any practical purpose. Below we illustrate the results of estimating the same expression (24) but with η = fl.8,000. In this case there are two regions for possible values of the conditional mode; from fl.0 to fl.32,000 and from fl.35,700 to fl.69,000.

File Contains Data for
PostScript Printers Only

The fact that the values between fl.32,000 and fl.35,700 are not feasible, however, seems largely due to finite sampling error. Thus, as in Figure 9, it seems that Figure 10 basically tells us that the mode is less than about fl.70,000. In both figures the results suggest very imprecise conclusions. They show that even a limited nonresponse rate can have dramatic consequences for inference about the conditional mode. Taking account of the finite sampling error would increase the imprecision even further. In principle, this should be possible using some bootstrap procedure but given the imprecise conclusions based on point estimates we did not consider it worthwhile to work this out for our empirical example.

Imposing the four exclusion restrictions and using (26) leads to estimated upper and lower loss functions illustrated in Figure 11. In this figure we see that between fl.0,00 and fl.150,000 the upper loss function lies below the lower loss function so that the inequality in (26) is violated. For values above fl.150,000 the estimated functions coincide: this happens because in that range $E[h_\eta(y|m,\nu,x,\delta=1)]=1$, while for some of the cells defined by the exclusion restriction variables, the estimate of $P(\delta=0|x)$ equals zero.

File Contains Data for
PostScript Printers Only

The results in Figure 11 are similar to those of figures 3-4 and 7-8, in that we find a crossing between upper and lower bounds. In figures 3-4 and 7-8 we concluded that despite the crossings, the confidence bands were so wide that the exclusion restrictions were not rejected. In Figure 11 we have no results for the precision of the estimated upper and lower loss function, and we can only conclude that imposing the four exclusion restrictions simultaneously leads to an empty set of possible conditional modes.

We investigate the possibility of identifying a set for the conditional mode imposing weaker exclusion restrictions. Figures 12 to 15 show the result of estimating (26) when each of

File Contains Data for
PostScript Printers Only

File Contains Data for
PostScript Printers Only

File Contains Data for PostScript Printers Only

File Contains Data for PostScript Printers Only

the four exclusion restrictions is imposed separately; the value of η is set to fl.10,000. Figures 12 and 13 show that estimated upper and lower bounds lead to no values of $\overline{b}$ that satisfy the conditions in (26): this suggests that the exclusion restrictions WORR and RISK are rejected by the data. On the other hand, figures 14 and 15 show that estimates of (26) imposing the exclusion restrictions REFG and DWRK respectively, lead to sets of upper and lower loss functions that satisfy the conditions in expression (26) for a non-empty range of values. Imposing the exclusion REFG identifies a range between fl.50,000 and fl.60,700 whereas if we impose the exclusion DWRK the range becomes fl.42,900-fl.60,700. In either case, imposing these weaker form of exclusion restrictions tightens the bounds on the conditional mode since the ranges in figures 14 and 15 are narrower than the estimated range in Figure 9. On the other hand, imprecision due to sampling error may increase, since estimates will be based on cells with limited numbers of observations.

6 Conclusions

In this paper we have applied the approach by Manski (1994,1995) to deal with item nonresponse in survey data. Compared to existing parametric or semi-parametric models, this approach imposes much weaker assumptions on the data generating process. We have focused on personal incomes in a Dutch cross-section, for which the item nonresponse rate is 7.7%. We have computed bounds for the conditional distribution function and the conditional quantiles of the distribution of personal incomes. Furthermore, we have looked at bounds on the conditional mode. We have considered bounds which do not impose any prior assumptions on the data (worst case bounds), and bounds which add prior assumptions in the form of monotonicity or exclusion restrictions. We have estimated these bounds nonparametrically, and have approximated their small sample distribution using a bootstrap procedure.

For worst case bounds on the distribution function, we find that imprecision due to item nonresponse is substantial, although smaller than the imprecision due to sampling error. Imposing monotonicity helps to tighten the bounds and to reduce the first type of imprecision substantially, particularly at the lower end of the distribution. Imposing joint exclusion restrictions based upon four psychological variables in the data set, leads to sets of point estimates of lower bounds which exceed point estimates of upper bounds. These estimated bounds are not useful for determining the unknown distribution function. We have shown how the upper and lower bounds can be used to construct an informal test of the exclusion restrictions. This is interesting since in this case, semiparametric selection model estimators typically do not yield a test of the exclusion restrictions. In our example finite sampling error is so large that the exclusion restrictions cannot be rejected.

Bounds for quantiles can easily be derived from bounds on the distribution function. We find that both item nonresponse and sampling error lead to substantial imprecision in estimating the conditional median or other quantiles. Imposing monotonicity helps to reduce the imprecision due to item nonresponse for the quantiles at the middle and lower end of the distribution.

For the conditional mode, the worst case bounds are informative for estimated loss functions when the smoothing parameter is greater than fl.8,000. Nevertheless, the range of potential values of the mode is too wide to be useful. Imposing four exclusion restrictions jointly leads to an estimated upper loss function that lies below the estimated lower loss function so that no useful range for the conditional mode is found. When each of the exclusion restriction variables is used separately, in two out of four cases we do find a feasible range for the conditional mode that improves upon the range obtained with the worst case bounds.

Our overall conclusion is that Manski’s approach works reasonably well for the distribution function and quantiles of the distribution, although a limited item nonresponse rate of less than 8% already leads to substantial uncertainty on the income quantiles, even in large samples. Moreover, this approach offers new ways of checking exclusion restrictions, and allows for an informal test of a single exclusion restriction which cannot be tested in a standard semi-parametric framework.

References

Ahn, H.T., and J.L Powell, (1993), Semiparametric estimation of censored selection models with a nonparametric selection mechanism, *Journal of Econometrics*, 58, 3-29.

Bierens, H.J., (1987), Kernel estimation of regression function, in *Advances of Econometrics: Fifth World Congress*, Vol.1, T. F. Bewley (ed.), Cambridge University Press, Cambridge, 102-115.

Chaudhuri, P., (1991), Global nonparametric estimation of conditional quantile functions and their derivatives, *Journal of Multivariate analysis,* 39, 246-269.

Härdle, W., (1984), Robust regression function estimation, *Journal of Multivariate Analysis,* 14, 169-180.

Härdle, W., (1990), *Applied nonparametric regression*. Econometric Society Monographs 19, Cambridge University Press, Cambridge.

Härdle, W., and O. Linton, (1994), *Applied nonparametric methods, in: Handbook of Econometrics IV,* R.F. Engle and D.L. McFadden (eds.), North-Holland, Amsterdam, 2297-2339.

Härdle, W., and J.S. Marron, (1985), Optimal bandwidth selection in nonparametric regression functions estimation, *Annals of Statistics*, 13, 1465-1481.

Heckman, J.J., (1979), Sample selection bias as a specification error, *Econometrica*, 47, 153-161.

Heckman, J.J., (1990), Varieties of selection bias, *American Economic Review, Papers & Proceedings*, 80, 313-318.

Koenker, R., and G. Bassett, (1978), Regression quantiles, *Econometrica,* 46, 33-50.

Lee, M.J., (1996), *Methods of moment in semi-parametric econometrics for limited dependent variables models,* Springer, New York.

Manski, C.F., (1989), Anatomy of the selection problem, *Journal of Human resources*, 24, 343-360

Manski, C.F., (1990), Nonparametric bounds on treatment effects, *American Economic Review, Papers & Proceedings*, 80, 319-323.

Manski, C.F., (1994), *The selection problem in: Advances in Econometrics*, C. Sims (ed.), Cambridge University Press, 143-170.

Manski, C.F., (1995), *Identification problems in the social science*, Harvard University Press.

Manski, C.F., (1997), Monotone treatment response, *Econometrica*, 65, 1311-1334.

Mroz, T.A., (1987), The sensitivity of empirical models with sample selection bias, *Econometrica*, 55, 765-799.

Newey, W., J.L.Powell and J.R. Walker, (1990), Semiparametric estimation of selection models: some empirical results, *American Economic Review, Papers and Proceedings*, 80, 324-328.

Nyhus, E.K., (1996), The VSB-CENTER Savings Project: Data collection methods, questionnaires and sampling procedures, *VSB progress report,* 42, Center for Economic Research (CentER), Tilburg University.

Rosenbaum, P.R and D.B. Rubin, (1984), Comment: Estimating the effects caused by treatment, *Journal of the American Statistical Association*, 79, 26-28.

Vella, F., (1998), Estimating models with sample selection bias, Journal of Human Resources 33, 127-172.

Appendix A

Four variables are used as exclusion restrictions. DWRK is constructed using the response behavior of each individual to the section in the panel called 'work and pensions'. REFG is constructed using only one question in the Psychological section of the panel named 'Group1', and the variables WORR and RISK are two ordered response variables that measure the psychological characteristics of the individual; for these two latter variables the information used is in the form as provided by the individual.

WORR is constructed from the answer to the following survey question,

> *"Now, we would like to know how would you describe your personality. Below we have mentioned a number of personal qualities in pairs. The qualities are not in every case opposites. Please indicate for each of the pairs of qualities which number would best describe your personality"*

Quality: easily get worried------------------Don't easily get worried.

Easily get worried............................ 1
.... 2
.... 3
.... 4
.... 5
.... 6
Don't easily get worried.................. 7
Don't know.................................... -9

We define WORR as 0 if the answer to this is below 4 - including '-9' - and 1 otherwise.

The variable RISK is based upon the following question,

> *"The following questions concern your readiness to take risks. First, some questions about games of chance"*

How often do you buy lottery tickets, do you play the lottery, or something of the kind?

Every week	1
A few times per month	2
Once a month	3
Six to ten times per year	4
One to six times per year	5
Rearly	6
Never/hardly ever	7
Don’t know	-9

The variable RISK is 1 if the answer is 1, 2 or 3, and 0 otherwise.

REFG is based upon the question,

Which group is most important to you, with respect to the financial situation of your household?

The neighbors	1
Friends and acquaintances	2
Colleagues at work	3
People with my level of education	4
People about the same age as myself	5
People with a similar job as myself	6
Brothers, sisters and other relatives	7
People known from newspapers and TV	8
Others	9
Don’t know	-9

The variable REFG is 0 if the answer is -9 and 1 otherwise.

Finally, the variable DWRK is constructed using all the questions in the ‘works and pensions’ section of the survey. These questions refer to conditions in the workplace, thoughts about pension plans, etc. None of these questions are directly related to income. If the individual answers all the questions in this section, DWRK is set equal to 1. Otherwise it is set to 0.