A sample selection model for unit and item nonresponse in cross-sectional surveys[*]

Giuseppe De Luca and Franco Peracchi
University of Rome "Tor Vergata"

This version: March 15, 2006

Abstract

We consider a general sample selection model where unit and item nonresponse simultaneously affect a regression relationship of interest, and both types of nonresponse are potentially correlated. We estimate both parametric and semiparametric specifications of the model. The parametric specification assumes that the errors in the latent regression equations follow a trivariate Gaussian distribution. The semiparametric specification avoids distributional assumptions about the underlying regression errors. In our empirical application, we estimate Engel curves for consumption expenditure using data from the first wave of SHARE (Survey on Health, Aging and Retirement in Europe).

Keywords: Unit nonresponse, item nonresponse, cross-sectional surveys sample selection models, Engel curves.

JEL classification: B12, C14, C31, C34

[*] Preliminary, comments welcome. Financial support for this research was provided through the European Community's Program 'Quality of Life' under the contract No. QLK6-CT-2002-002426 (AMANDA). This paper is based on data from the early Release 1 of SHARE 2004, which is preliminary and may contain errors that will be corrected in later releases. The SHARE data collection has been primarily funded by the European Commission through the 5th framework program (project QLK6-CT-2001-00360 in the thematic program 'Quality of Life'). Additional funding came from the US National Institute on Aging (U01 AG09740-13S2, P01 AG005842, P01 AG08291, P30 AG12815, Y1-AG-4553-01 and OGHA 04-064). Data collection in Austria (through the Austrian Science Fund, FWF) and Switzerland (through BBW/OFES/UFES) was nationally funded.

1 Introduction

Nonresponse is one of the most important and most studied sources of nonsampling error in sample survey. A distinction is usually made between two forms of nonresponse, namely, unit and item nonresponse. Unit nonresponse occurs when an eligible sample unit fails to participate to a survey because of either failure to establish a contact, or because of explicit refusal to cooperate. Item nonresponse occurs instead when a responding unit does not provide useful answers to particular items of the questionnaire. For panel surveys, one can also distinguish another particular form of unit nonresponse, namely, sample attrition. Attrition occurs when a responding unit in one wave of the panel drops out in a subsequent wave.

This paper is mainly concerned with problems of nonresponse in cross-sectional surveys. These problems, however, are also common to the first wave of panel surveys. It is important to emphasize two crucial differences between unit nonresponse in the first wave of a panel and panel attrition. First, the study of unit nonresponse in the first wave of a panel is usually complicated by the lack of adequate information on the units who refuse to participate to the survey, whereas information collected during the previous waves can be used to study panel attrition. This explains why problems of unit nonresponse in the first wave of a panel have received much less attention than problems of attrition. Second, unit response rates in the first wave of a panel are typically lower than those achieved in subsequent waves. Other things being equal, this implies that unit nonresponse in the first wave is more problematic than attrition.

One crucial issue in studying both unit and item nonresponse is to establish whether or not the mechanism generating missing observations is random. Using the terminology introduced by Rubin (1976), one can define three possible missing data mechanisms. The mechanism is missing completely at random (MCAR) if missingness does not depend on the values of the variables in the data matrix. The mechanism is missing at random (MAR) if, after conditioning on a set of observed covariates, there is no relation between missingness and the observed outcome variables. The mechanism is not missing at random (NMAR) if missingness and the observed outcome variables are related even after conditioning on the set of observed covariates. When mechanisms underlying (unit or item) nonresponse are NMAR, analyses that ignore nonresponse errors, or relay on the MAR assumption, may lead to invalid inference about population parameters of interest.

An important strategy in order to reduce nonresponse errors consists of planning preventive measures to cope with nonresponse at the survey design stage. Well-designed surveys aim to reduce unit nonresponse rates by choosing the most appropriate fieldwork period, interview mode, interviewer training, follow-up procedures and incentive schemes. Other aspects of the questionnaire design, like length of the interview, wording of the questions and their reference period, are more likely to affect item nonresponse rates. Empirical studies by Groves and Couper (1998), Groves *et al.* (2002), O'Muircheartaigh and Campanelli (1999) and Riphahn and Serfling (2002), show that all these aspects of survey design are typically crucial to explain response rates achieved in sample surveys. Unfortunately, despite the preventive measures adopted for minimizing nonresponse errors, response rates are rarely close to 100 percent. This explains why most of the survey nonresponse literature focuses on the development of statistical methods for ex-post adjustments of nonresponse errors (see Lessler and Kalsbeek 1992, and Little and Rubin 2002). Weighting adjustment methods, which involve the assignment of weights to sample respondents in order to compensate for their systematic differences relative to nonrespondents, have been traditionally used to deal with problems of unit nonresponse, whereas imputation procedures, which aim to fill in missing values to produce a complete dataset, have been traditionally used to deal with problems of item nonresponse. Although ex-post adjustment techniques have reached a high level of sophistication, such methods commonly assume that the missing data mechanism is MAR, and they do not generally allow compensating simultaneously for errors due to unit and item nonresponse.

This paper differs from previous studies in two respects. First, problems of selectivity due to unit and item nonresponse are analyzed jointly. Second, missing data mechanisms underlying these different types of nonresponse are allowed to be NMAR. In particular, we analyze a general sample selection model where unit and item nonresponse can jointly affect a regression relationship of interest, and the two types of nonresponse can be correlated. Attention focuses on two alternative specifications of the model, one parametric and the other semiparametric. In the parametric specification, errors in the two selection equations (one for unit and one for item nonresponse) and in the equation for the outcome of interest are assumed to follow a trivariate Gaussian distribution. In the semiparametric specification, we avoid distributional assumptions about the errors in the three equations. After discussing issues related to identification and estimation of the two kind of models, we provide an empirical application by using data from the first wave of SHARE (Survey on Health, Aging and Retirement in Europe), a new survey conducted in 2004 across eleven European countries. The aim of this analysis is to investigate the potential selectivity associated with unit and item nonresponse in the estimation of Engel curves for food consumption at home and total nondurable consumption.

The remainder of the paper is organized as follows. Section 2 formalizes the motivation of this study, and presents a general framework to analyze problems of unit and item nonresponse. Sections 3 and 4 consider problems of identification and estimation of the parametric and semi-parametric model respectively. Section 5 discusses the main survey design characteristics of SHARE, and presents results of our empirical study. Finally, Section 6 summarizes our main findings and offers some conclusion.

2 A sample selection model for unit and item nonresponse

Suppose that we are interested in estimating the conditional mean function of a random outcome by using data from a survey, where a set of n_1 units is initially drawn at random from some population. Nonresponse may select the sample at two stages. First, unit nonresponse may reduce the sample size to $n_2 < n_1$ responding units. Second, nonresponse to specific items of the questionnaire may further reduce the number of usable observations to $n_3 < n_2$. The reduction of observations causes, of course, an efficiency loss relative to the ideal situation of complete response. This efficiency loss needs not be the main concern, because lack of independence between the missing data mechanism and the outcome may also generate selectivity in the observed sample and may lead to biased estimates of the population parameters.

To formalize the nature of the problem, we consider a sequential framework where individuals first decide whether to participate to the survey, and then decide whether to answer to each item of the questionnaire. Thus, the indicator of unit response is always observed, while the indicator of item response is only observed for those units that agree to participate in the first stage. Let Y_1 denote the indicator of the event that an eligible sample unit participates to the survey, and let Y_2 denote the indicator of the event that a responding unit provides information on a specific item of interest Y_3. The response process is completely described by two elements: the probability of unit nonresponse, $\pi_0 = \Pr\{Y_1 = 0\}$, and the probability of item nonresponse conditional on unit response, $\pi_{0|1} = \Pr\{Y_2 = 0 \,|\, Y_1 = 1\}$. By the law of iterated expectations

$$\mathrm{E}(Y_3 \,|\, Y_1 = 1) - \mathrm{E}(Y_3) = \pi_0[\mathrm{E}(Y_3 \,|\, Y_1 = 1) - \mathrm{E}(Y_3 \,|\, Y_1 = 0)]. \tag{1}$$

Further

$$\mathrm{E}(Y_3 \,|\, Y_1 = 1) = \mathrm{E}(Y_3 \,|\, Y_1 = 1, Y_2 = 1) + \pi_{0|1}[\mathrm{E}(Y_3 \,|\, Y_1 = 1, Y_2 = 0) - \mathrm{E}(Y_3 \,|\, Y_1 = 1, Y_2 = 1)].$$

Substituting this expression into the left-hand side of (1) and rearranging gives the following expression for the overall nonresponse bias, namely the difference between the conditional mean of

Y_3 for the sub-sample of fully responding units and the unconditional mean of Y_3 for the overall population,

$$\begin{aligned}\mathrm{E}(Y_3 \,|\, Y_1 = 1, Y_2 = 1) - \mathrm{E}(Y_3) = \pi_0[\mathrm{E}(Y_3 \,|\, Y_1 = 1) - \mathrm{E}(Y_3 \,|\, Y_1 = 0)] + \\ + \pi_{0|1}[\mathrm{E}(Y_3 \,|\, Y_1 = 1, Y_2 = 1) - \mathrm{E}(Y_3 \,|\, Y_1 = 1, Y_2 = 0)].\end{aligned}$$

The overall nonresponse bias is zero only if: (i) $\pi_0 = \pi_{0|1} = 0$ (neither unit nor item nonresponse), (ii) $\mathrm{E}(Y_3 \,|\, Y_1 = 1) = \mathrm{E}(Y_3 \,|\, Y_1 = 0)$ and $\mathrm{E}(Y_3 \,|\, Y_1 = 1, Y_2 = 1) = \mathrm{E}(Y_3 \,|\, Y_1 = 1, Y_2 = 0)$, or (iii) the two sources of bias offset each other. If there are both unit and item nonresponse, the overall nonresponse bias depends on two separate components proportional to the probabilities of unit and item nonresponse respectively.

Our objective is to construct consistent estimates of the mean function of Y_3 (conditional on covariates) allowing for selectivity generated by unit and item nonresponse. Notice that distinguishing between unit and item nonresponse is of considerable practical importance for at least two reasons. First, it can help improving the specification of the model, because different information is usually available for studying the two types of nonresponse. In fact, the information available to study unit nonresponse is usually confined to the information obtained from the sampling frame or from the data collection process, whereas additional information collected during the interview can be used to study item nonresponse. Second, understanding the different types of error generated by unit and item nonresponse plays a key rule at the survey design stage, where resources have to be allocated efficiently to reduce nonresponse errors. For instance, improving incentive schemes and follow-up procedures can be useful fieldwork strategies to reduce unit nonresponse, while reducing the complexity of the questionnaire can help reduce item nonresponse.

Our way of modelling the effects of both forms of selectivity is based on a straightforward generalization of the classical sample selection model proposed by Heckman (1979). Our model has the following form:

$$Y_{ij}^* = \beta_j^\top X_{ij} + \sigma_j U_{ij}, \qquad j = 1, 2, 3, \qquad i = 1, \dots, N, \tag{2}$$

$$Y_{i1} = 1\{Y_{i1}^* \geq 0\}, \qquad i = 1, \dots, n_1, \tag{3}$$

$$Y_{i2} = 1\{Y_{i2}^* \geq 0\}, \qquad \text{if } Y_{i1} = 1, \qquad i = 1, \dots, n_2, \tag{4}$$

$$Y_{i3} = Y_{i3}^*, \qquad \text{if } Y_{i1}Y_{i2} = 1, \qquad i = 1, \dots, n_3, \tag{5}$$

where the Y_{ij}^*, $j = 1, 2, 3$, are latent continuous random variables representing respectively the propensity to participate to the survey, the propensity to answer to the item of interest, and the outcome variable in the uncensored sample. The latent variables are related to their observed counterparts through the observation rules (3)–(5), where $1\{A\}$ is the indicator function of the event A. The X_{ij}, $j = 1, 2, 3$, are k_j-vectors of fully observable exogenous predictors and β_j are their associated parameters. The U_{ij} are latent regression errrors with zero mean and unit variance, and the σ_j are nuisance scale parameters. As usual, σ_1 and σ_2 are normalized to one in order to identify coefficients of the binary response equations.

The primary interest of the analysis is to estimate the parameter β_3 of the population regression function from the sub-sample of fully observed units, for which:[1]

$$\mathrm{E}(Y^*_{i3}\,|\,Y_{i1} = 1, Y_{i2} = 1) = \mu_{i3} + \sigma_3\,\mathrm{E}(U_{i3}\,|\,U_{i1} > -\mu_{i1}, U_{i2} > -\mu_{i2}), \tag{6}$$

where $\mu_{ij} = \beta_j^\top X_{ij}$. If one of the two nonresponse mechanisms is NMAR, then the conditional expectation on the right hand side of (6) is different from zero, and traditional estimation methods lead to inconsistent estimates of the population parameter β_3. Consistent estimates can in general be obtained through generalizations of the classical Heckman two-step procedure. Parametric and semiparametric versions of this estimation procedure are presented in Sections 3 and 4 respectively.

3 A parametric model

In this section we consider a parametric framework in which the latent regression errors are assumed to follow a trivariate Gaussian distribution with zero mean and correlation matrix

$$\Sigma = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{bmatrix},$$

where ρ_{jk} is the correlation between the errors in the j-th and the k-th equation.

In this parametric setting, the vector $\beta = (\beta_1, \beta_2, \beta_3)$ of model parameters can be estimated consistently through the two-step procedure originally proposed by Poirier (1980) and further developed by Ham (1982). Here, we slightly modify their procedure in order to account for partial observability of Y_2 during the first estimation step. Using results of Tallis (1961), Poirier (1980) shows that the conditional expectation on the right hand side of (6) admits the explicit representation

$$\mathrm{E}(U_{i3}\,|\,U_{i1} > -\mu_{i1}, U_{i2} > -\mu_{i2}) = \rho_{13}\lambda_{i1}(\theta) + \rho_{23}\lambda_{i2}(\theta), \tag{7}$$

[1] In the following, explicit conditioning on X_1, X_2 and X_3 is suppressed to simplify notation.

where $\theta = (\beta_1, \beta_2, \rho_{12})$ and the $\lambda_{ij}(\theta)$ are bias correction terms given by

$$\lambda_{i1}(\theta) = \frac{\phi(\mu_{i1})\,\Phi(\sigma^{-1}(\mu_{i2} - \rho_{12}\mu_{i1})}{\Phi_2(\mu_{i1}, \mu_{i2}; \rho_{12})},$$
$$\lambda_{i2}(\theta) = \frac{\phi(\mu_{i2})\,\Phi(\sigma^{-1}(\mu_{i1} - \rho_{12}\mu_{i2}))}{\Phi_2(\mu_{i1}, \mu_{i2}; \rho_{12})},$$

with $\sigma = \sqrt{1-\rho_{12}^2}$, $\phi(\cdot)$ and $\Phi(\cdot)$ respectively the density and the distribution function of the standardized Gaussian distribution, and $\Phi_2(\cdot, \cdot; \rho_{12})$ the bivariate Gaussian distribution function with zero means, unit variances and correlation coefficient ρ_{12}. The basic idea of the two-step procedure is to use consistent estimates of the bias correction terms in (7) as additional regressors in a standard OLS procedure.

In the first step, we consider a bivariate probit model with sample selection for (Y_1, Y_2), and estimate the parameter θ by maximum likelihood (ML). Identifiability of the model requires imposing at least one exclusion restriction on the two set of exogenous covariates X_1 and X_2. Subject to the identifiability restrictions, the log-likelihood for a random sample of n_1 units can be written as,

$$L(\theta) = \sum_{i=1}^{n_1} \left[Y_{i1} Y_{i2} \ln \pi_{i11}(\theta) + Y_{i1}(1 - Y_{i2}) \ln \pi_{i10}(\theta) + (1 - Y_{i1}) \ln \pi_{i0}(\theta) \right], \tag{8}$$

where

$$\pi_{i11}(\theta) = \Pr\{Y_{i1} = 1, Y_{i2} = 1\} = \Phi_2(\mu_{i1}, \mu_{i2}; \rho_{12}),$$
$$\pi_{i10}(\theta) = \Pr\{Y_{i1} = 1, Y_{i2} = 0\} = \Phi(\mu_{i1}) - \Phi_2(\mu_{i1}, \mu_{i2}; \rho_{12}),$$
$$\pi_{i0}(\theta) = \Pr\{Y_{i1} = 0\} = 1 - \Phi(\mu_{i1}).$$

A ML estimator $\hat{\theta}$ maximizes (8) over the parameter space $\Theta = \Re^2 \times (-1, 1)$. This estimator is consistent if the bivariate probit model is correctly specified, and is asymptotically normal under general conditions. Within this model, the hypothesis of conditional independence between unit and item nonresponse can be tested through either a Wald test on the significance of ρ_{12}, or a likelihood ratio test that compares the maximized values of the log-likelihood in (8) with the sum of the log-likelihoods of two simple probit models, one for Y_1 and one for Y_2 given $Y_1 = 1$.

In the second step of the procedure, estimates $\hat{\lambda}_{ij} = \lambda_{ij}(\hat{\theta})$, $j = 1, 2$, of the bias correction terms in (7) are used as additional predictors in the augmented regression model

$$Y_{i3} = \beta_3^\top X_{i3} + \sigma_3\rho_{13}\hat{\lambda}_{i1} + \sigma_3\rho_{23}\hat{\lambda}_{i2} + \epsilon_{i3} = \gamma^\top \widetilde{X}_{i3} + \epsilon_{i3}, \tag{9}$$

where $\epsilon_{i3} = U_{i3} - \sigma_3\rho_{13}\hat{\lambda}_{i1} - \sigma_3\rho_{23}\hat{\lambda}_{i2}$ is a heteroscedastic regression error with zero conditional mean, $\gamma = (\beta_3, \sigma_3\rho_{13}, \sigma_3\rho_{23})$ and $\widetilde{X}_{i3} = (X_{i3}, \hat{\lambda}_{i1}, \hat{\lambda}_{i2})$. The parameters γ can be estimated consistently by ordinary least square, even if computation of their standard errors needs to take into account the heteroscedasticity induced by censoring and the additional variability induced by the use of the generated regressors $\hat{\lambda}_{i1}$ and $\hat{\lambda}_{i2}$. Ham (1982) provides consistent estimators of σ_3 and of the asymptotic covariance matrix of the least square estimator of γ. Alternatively, standard errors of the estimates can be obtained via the nonparametric bootstrap.

Although the implementation of this two-step estimator is relatively straightforward, one of the major concerns is identifiability of the parameters in model (9). The identification problem is closely related to that arising in the classical Heckman two-step procedure (see Vella 1998 and Puhani 2000 for an extensive discussion). Parameters of the second estimation step may in principle be identified through the nonlinearity of the inverse Mills ratio. However, since the inverse Mills ratio is linear over a wide range of its argument, identification obtained through the nonlinearity of the inverse Mills ratio is often weak. The inclusion of additional variables in the first estimation step can therefore be useful to assist identification in the second estimation step.[2] The above considerations also hold for sample selection models with two censoring equations. Although larger values of the correlation coefficient ρ_{12} increases slightly the nonlinearity of the bivariate Mills ratio, the function is still linear for wide ranges of the two indexes $\mu_{i1} = \beta_1^\top X_{i1}$ and $\mu_{i2} = \beta_2^\top X_{i2}$. Exclusion restrictions (that is, variables which are included in X_1 and X_2 but excluded from X_3) become then crucial to guarantee identifiability of the parameters in the second estimation step.

As suggested by Fitzgerald *et al.* (1998) and Nicoletti and Peracchi (2005), features of the data collection process and socio-demographic characteristics of the interviewers can be promising candidates for this set of exclusion restrictions. Because these variables are external to the subjects under investigation and are not under their control, one should expect them to be irrelevant in explaining the outcome variable of interest. On the other hand, results of several data validation studies have shown that these variables are typically important predictors of both unit and item response.

4 A semiparametric model

One criticism of parametric estimation of sample selection models stresses their sensitivity to incorrect specification of the model. Parametric estimators for this class of models are indeed inconsistent whenever assumptions on the deterministic or the stochastic part of the model are not valid. Dur-

[2] Leung and Yu (1996) show that the quasi-linearity of the inverse Mills ratio causes essentially a problem of collinearity with the other covariates of the second estimation step, which in turn leads to inflated standard errors and unreliable estimates.

ing the last twenty-five years, a large body of the econometric literature has been concerned with finding semiparametric procedures for consistent estimation in the presence of different forms of misspecification. Some of these estimators are also known to be $\sqrt{n}$-consistent and asymptotically normal (see Vella 1998 for a survey).

In this section, we focus on model (2)–(5) and consider semiparametric estimation procedures that are robust to departures from the assumption of Gaussian errors. Once we relax the Gaussian distributional assumption, estimation of the model raises two difficulties. First, one can not invoke distributional assumptions to estimate parameters of the two binary response models. Second, one can not use distributional relationships to find an analytical expression for the bias correction term in equation (6). In this case, the conditional expectation for the outcome variable of interest can be written as the partially linear model

$$\mathrm{E}(Y_{i3} \,|\, Y_{i1} Y_{i2} = 1) = \mu_{i3} + g(\mu_{i1}, \mu_{i2}), \qquad i = 1, \ldots, n_3, \tag{10}$$

where $\beta_3^\top X_{i3}$ is the linear part of the model and g is now an unknown function of the two indexes $\mu_{i1} = \beta_1^\top X_{i1}$ and $\mu_{i2} = \beta_2^\top X_{i2}$. Notice that the model maintains a double index structure. In principle, the index restriction could be relaxed, but the resulting estimators would suffer of the well known curse of dimensionality problem. The double index structure is therefore useful to reduce the dimension of the covariate space, thereby avoiding the curse of dimensionality problem.

Before describing the estimation procedures for model (2)–(5), it is important to mention the conditions under which it is identified. Identifiability of the equations for unit and item nonresponse requires normalizing the location of the underlying distribution functions. Following Melenberg and van Soest (1996), we set the intercept coefficient equal to its probit estimate. As shown by Ichimura and Lee (1991), identifiability of the double-index model (10) also requires that the index of each equation contains at least one continuous variable with a nonzero coefficient which is not contained in the other index.[3] Thus, unlike the parametric specification of the model, exclusion restrictions should now include some continuous variable.

Subject to these identifiability restrictions, semiparametric estimation of model (10) can again be carried out through a two-step procedure. In principle, parameters of the two response equations could be jointly estimated by the semiparametric maximum likelihood approach of Lee (1995), which generalizes to sequential choice models the approach originally proposed for binary choice models by Klein and Spady (1993). In practice, because of both the large sample size and the large number

[3] See Lemmas 2 and 3 in Ichimura and Lee (1991).

of covariates used in our empirical application, the implementation of this estimator would be very time consuming.[4]

To overcome this computational difficulty, we focus on a simpler semiparametric model where errors of the two response equations are assumed to be independent. As shown later in Section 5.4, this conditional independence assumption is strongly supported by the results from parametric estimation of the model. Further, in the first step of the procedure, the parameters of the response equations are estimated by the less computationally demanding semi-nonparametric (SNP) estimator of Gallant and Nychka (1987). Specifically, after a suitable parametrization, the univariate densities of the latent regression errors may be approximated by densities of the Hermite form

$$f_K(u) = \frac{1}{\psi_K}\, \tau_K(u)^2\, \phi(u), \tag{11}$$

where $\tau_K(u) = \sum_{k=0}^{K} \tau_k u^k$ is a polynomial of order K, and $\psi_K = \int_{-\infty}^{\infty} \tau_K(u)^2 \phi(t)\, dt$. Since $f_K(u)$ is invariant to multiplication of $\tau = (\tau_0, \dots, \tau_K)$ by a scalar, τ_0 is normalized to one.

The associated distribution function is of the form

$$F_K(u) = \int_{-\infty}^{u} f_K(t)\, dt = \frac{1}{\psi} \sum_{k=0}^{2K} \tau_k^* I_k(u), \tag{12}$$

where $I_k(u) = \int_{-\infty}^{u} u^k \phi(t)\, dt$ are the truncated moments of the standardized Gaussian distribution and satisfy the recursion $I_k(u) = (k-1)I_{k-2}(u) - u^{k-1}\phi(u)$, with $I_0(u) = \Phi(u)$ and $I_1(u) = -\phi(u)$.

The two vectors of parameters β_1 and β_2 are estimated by maximizing the pseudo log-likelihood functions of two binary choice models (one for Y_1, and one for Y_2 given $Y_1 = 1$) in which the unknown distribution functions are replaced by approximations of the form (12). As shown by Gallant and Nychka (1987), this pseudo-ML estimator is consistent and asymptotically normal provided that the degree K of the polynomial increases with the sample size. For a given sample size, the value of K may be selected either through a sequence of likelihood ratio tests, or by model selection criteria like the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). In particular, Gabler *et al.* (1993) and Stewart (2004) show that semiparametric specifications with $K \leq 2$ are equivalent to a simple probit model. Thus, the first semiparametric model that generalizes the probit model is the specification with $K = 3$. In this semiparametric setting, a test on the joint significance of the coefficients $\tau = (\tau_3, \dots, \tau_K)$ is equivalent to a test on the Gaussian distributional assumption of the error term.

[4] Semiparametric estimators based on kernel density estimation typically require computing n kernel functions at each step of the optimization process. Thus, the computational time required by these estimators depends crucially on both the sample size and the number of covariates. For our empirical application, which involves about 14,700 observations and 42 covariates, the implementation of these estimators would be computationally too demanding.

Given consistent estimates of the two indexes $\mu_{i1} = \beta_1^\top X_{i1}$ and $\mu_{i2} = \beta_2^\top X_{i2}$, the parameters of the model for the outcome of interest can be estimated by the semiparametric approach of Robinson (1988). In particular, model (10) directly implies that:

$$Y_{i3} - \mathrm{E}(Y_{i3}\,|\,\mu_{i1}, \mu_{i2}, Y_{i1}Y_{i2} = 1) = \beta_3^\top [X_{i3} - \mathrm{E}(X_{i3}\,|\,\mu_{i1}, \mu_{i2}, Y_{i1}Y_{i2} = 1)] + \epsilon_{i3}, \tag{13}$$

where $\mathrm{E}(\epsilon_{i3}|\mu_{i1}, \mu_{i2}, Y_{i1}Y_{i2} = 1) = 0$. After replacing the unknown conditional expectations in (13) with their nonparametric estimates, the slope coefficients can be estimated by OLS with no intercept.[5] Robinson (1988) shows that, under mild regularity conditions, the estimator of the slope coefficient β_3 is $\sqrt{n_3}$-consistent and asymptotically normal. The intercept coefficient is instead absorbed in the unknown function g, and is not identified. Finally, the nonlinear function of the model g can be estimated nonparametrically by the residual component:

$$\hat{g}(\mu_{i1}, \mu_{i2}) = \hat{\mathrm{E}}(Y_{i3}\,|\,\mu_{i1}, \mu_{i2}, Y_{i1}Y_{i2} = 1) - \hat{\beta}_3^\top \hat{\mathrm{E}}(X_{i3}\,|\,\mu_{i1}, \mu_{i2}, Y_{i1}Y_{i2} = 1). \tag{14}$$

where $\hat{\mathrm{E}}(Y_{i3}\,|\,\mu_{i1}, \mu_{i2}, Y_{i1}Y_{i2} = 1)$ and $\hat{\mathrm{E}}(X_{i3}\,|\,\mu_{i1}, \mu_{i2}, Y_{i1}Y_{i2} = 1)$ denote respectively the nonparametric estimates of $\mathrm{E}(Y_{i3}\,|\,\mu_{i1}, \mu_{i2}, Y_{i1}Y_{i2} = 1)$ and $\mathrm{E}(X_{i3}\,|\,\mu_{i1}, \mu_{i2}, Y_{i1}Y_{i2} = 1)$.[6] Notice that, since the rate of convergence of $\hat{g}$ depends on the rate of convergence of the nonparametric estimators in (14), this estimator is not $\sqrt{n_3}$-consistent in general.

5 An empirical application

In this section, we use data from the first wave of SHARE (Survey on Health, Aging and Retirement in Europe) to investigate whether selectivity associated with unit and item nonresponse may bias the estimation of Engel curve for household consumption.

There are reasons to believe that the process leading to missing consumption data is NMAR. First, survey nonresponse typically depends on income. For a simple theoretical model of this relation see Korinek, Mistiaen and Ravallion (2004). Second, consumption is a good proxy of permanent income. We focus on two PPP-adjusted consumption expenditure categories: food consumption at home, and total nondurable consumption. Although the measure of primary interest for many economic studies is total nondurable consumption, recent data validation studies by Browning *et al.* (2002), Battistin *et al.* (2003) and Winter (2004) have shown that information collected through

[5] Like for the parametric two-step procedure, computation of the standard errors needs to take into account the heteroscedasticity induced by censoring and the additional variability due to the use of the generated regressors $\hat{\mu}_{i1}$ and $\hat{\mu}_{i2}$. In our empirical application, standard errors are computed via the nonparametric bootstrap.

[6] Alternatively, the nonlinear component of the model can be estimated by a nonparametric regression of $Y_{i3} - \hat{\beta}_3^\top X_{i3}$ on $\hat{\mu}_{i1}$ and $\hat{\mu}_{i2}$.

sub-categories of consumption expenditures is usually more accurate than that collected through a "one-shot" question on total nondurable consumption. In addition, food consumption at home is typically an important component of total nondurable household consumption.

5.1 Country coverage and sampling design

SHARE is a standardized multi-purpose household survey designed to investigate several aspects of the elderly population in Europe. Its first wave, conducted in 2004, covered 15,544 households and 22,431 individuals in eleven European countries (Austria, Belgium, Denmark, France, Germany, Greece, Italy, Netherlands, Spain, Sweden and Switzerland).

In each country, the target population consists of all people living in residential households who have at least 50 years of age, plus their (possibly younger) partners. The target population is further restricted by a number of additional eligibility criteria, which exclude people who currently do not reside at the sampled address, or died before the starting of the field period, or are unable to speak the specific language of the national questionnaire, or are physically or mentally unable to participate to the survey.

All national samples are selected through probability sampling, but sampling procedures are not completely standardized across countries. Here, we distinguish between two groups of countries depending on the nature of the sampling frame adopted. In one group of countries (Denmark, Germany, Italy, Netherlands, Spain, and Sweden), the sampling frame is a population register containing information at least on the age and the gender of the sampled units. In another group of countries, the sampling frame is either a telephone register (like in Austria, Belgium, Greece and Switzerland) or a register of dwellings (like in France), and does not contain information on the background characteristics of the sampled units. In these countries, age-eligibility was assessed through a preliminary screening phase in the field. However, because of nonresponse during the screening phase, it was not possible to determine the eligibility status of about 15 percent of gross sample. For this second group of countries, the analysis of unit nonresponse is therefore complicated by the lack of sampling frame information and unknown eligibility of a fraction of the gross sample. To avoid these problems, we only consider the first group of countries (Denmark, Germany, Italy, Netherlands, Spain, and Sweden).

Table 1 provides the number of eligible households, the unweighted household response rate, and the most common sub-components of the household nonresponse rate (that is, noncontact rate, refusal rate and other non-interview rate) by country.[7] The household response rate ranges between a minimum of 47 percent in Sweden and a maximum of 62 percent in Netherlands, and is equal to 56 percent on average. Focusing attention on the reasons for nonresponse, we find that refusal to participate to the survey is the main reason (34 percent), although in some countries a non negligible fraction of nonresponse is also due to noncontact (13 percent in Spain) and other non-interview reasons (5 percent in Sweden).

Conditional on unit response, SHARE also experienced non-negligible amounts of missing data for open-ended questions on the amounts of income, assets and consumption expenditures. Item response rates (that is, the fraction of eligible respondents with a "Don't know" or "Refusal" answer) for the two consumption expenditure items of interest are reported in Table 2. The cross country average of the item response rates is equal to 86 percent for food consumption at home, and 83 percent for total nondurable consumption. Also in this case there is however a considerable variation across country. The lowest item response rates are in Spain (78 and 77 percent respectively), while the highest are in Sweden (93 and 90 percent respectively).

Although response rates obtained in the first wave of SHARE do not differ considerably from those obtained by other comparable European surveys, results of Tables 1 and 2 suggest that unit and item nonresponse may be two important sources of nonsampling errors.

5.2 Consumption expenditure data and outliers

A preliminary analysis of PPP-adjusted consumption expenditure data reveals clearly the presence of outliers in the tails of the empirical distribution of these variables. This is a typical problem of data collected through retrospective and open-ended questions. On the one hand, there are households who report zero or very low expenditures. Although zero or very low expenditures may be plausible answers for some consumption categories, we believe that these observations are highly suspicious for food consumption at home and total nondurable consumption. To deal with this problem, we trim 1 percent of the observations from the lower tail of the two empirical distributions. On the other hand, we find extremely high expenditures which are presumably due to interviewer's typing errors. To exclude these outliers, we trim 1 percent of the observations from the upper tail of the two empirical distributions. Since outliers can be considered as useless answers,

[7] For each country, the unweighted household response rate is computed as the fraction of eligible households with at least one interviewed person. Further details on the computations of these outcome rates are given in Börsch-Supan and Jürges (2005).

in our empirical application trimmed observations are treated as item nonresponse.[8] Summary statistics (that is, number of nonmissing observations, mean, standard deviation, minimum and maximum) of the two PPP-adjusted and trimmed consumption distributions are showed in Table 3.

5.3 Predictors of the unit and the item response probabilities

In SHARE, predictors of unit nonresponse can be obtained by exploiting the information coming from the sampling frame, the survey agencies and the fieldwork. By matching these different sources of data, we are able to get information on background characteristics of the selected household member (like age and gender), interviewers' characteristics (like age, gender and years of education) and workload (measured by the number of households visited in person), total number of calls and length of the fieldwork (measured by the number of days elapsed between the first and the last call attempt).

Once we focus on the sub-sample of responding households, the additional information collected during the interview can be used to study nonresponse on specific items of the questionnaire. The multidisciplinary nature of the SHARE data offers the unique opportunity of assessing whether item nonresponse on consumption questions is related to different types of economic and health variables, once we control for features of the data collection process and background characteristics of respondents and interviewers.

Since consumption questions are asked to the household member who is most knowledgeable about housing matters (the "household respondent" or HR), a set of variables related to socio-demographic characteristics, cognitive abilities, and health conditions of the HR has been included as predictors of item response. Our set of socio-demographic variables includes age (which enters as a quadratic term), gender, years of education, current job situation, marital status, household size, and a dummy variable for living in small cities. Cognitive abilities are measured through the scores obtained in the mathematical, orientation in time and delayed recall tests performed during the cognitive function (CF) module of the SHARE interview. The set of health variables includes instead the $EURO_D$ depression scale index, and a set of dummies for self-reported problems in managing money, less than good self-perceived health, and at least one ADL limitation.[9]

Although household income and wealth are two obvious predictors, such variables are affected by item nonresponse, measurement errors and outliers. To deal with the first problem, we use

[8] A sensitiveness analysis with 1.5 and 2 percent of trimming does not lead to qualitative different results. Thus, for efficiency reasons, we only present results with 1 percent of trimming.

[9] An accurate description of these health measures can be found in Börsch-Supan *et al.* (2005).

imputed gross annual household income and net financial assets as provided in the SHARE public release database, but include dummies for imputed values.[10] To reduce the impact of measurement errors and outliers, we instead use dummies for income and wealth quartiles.

To control for differences in the interview process, we use a set of measure of the cognitive burden of the interview. These variables include the length of the household respondent interview, and a set of dummies for interviews done by proxy, at the respondent home, and with the presence of other non-household member.

As a measure of the interviewers' computer skill, we also include the length of the Interviewer (IV) module. The IV module contains a set of closed questions on the background characteristics of the interviewers and the conditions of the interview process. More striking, since this module is only completed by the interviewer without involving the respondent, its length provides a proxy measure of the interviewers' computer skill.[11]

Definitions and summary statistics (that is, number of nonmissing observations, mean and standard deviation) of the predictors of unit and item response are provided in Tables 4 and 5 respectively.

5.4 Parametric estimates

To assess the selectivity effects generated by unit and item nonresponse, we estimate and compare five alternative models. Model 1 is a standard linear model estimated for the fully responding units without accounting for selectivity generated by nonresponse. Model 2 is a classical sample selection model estimated for the unit respondents and only accounts for selectivity generated by item nonresponse. Model 3 is a classical sample selection model estimated for the full sample with a single indicator ($D_i = Y_{i1}Y_{i2}$) for unit and item response. Model 4 is a generalized sample selection model which accounts for selectivity generated by unit and item nonresponse, but assumes that errors in the unit and the item response equations are independent. Finaly, Model 5 is a generalized sample selection model which accounts for selectivity generated by unit and item nonresponse, and does not impose independence of the error terms in the two response equations. Parametric estimates of these models are provided in Tables 6 - 9.

All estimated models share two common features. First, given the high comparability of the SHARE data, we pool data from the various countries and introduce country dummies to capture

[10] Imputed values of household income and wealth are based on hotdeck and multiple imputation methodologies. Here, for simplicity reasons, we only use the first of the five imputed distributions.

[11] Estimates of a regression model for the length of the IV module reveals that this module tends to last longer for interviewers with higher age and lower education (result omitted to save space).

unobserved heterogeneity across countries. Pooling the data allows increasing the number of observations and helps reducing problems of collinearity due to the limited within-country variability of some variable (like characteristics of the fieldwork and the interviewers). Second, identifiability of the model parameters is achieved by imposing a common set of exclusion restrictions. As mentioned in Section (3), our exclusion restrictions are based on characteristics of the fieldwork, the interview process and the interviewers. In particular, characteristics of the fieldwork are used to predict unit nonresponse, features of the interview process are used to predict item nonresponse, and socio-demographic characteristics of the interviewers are used to predict both. If we distinguish between household or personal characteristics and country dummies (V), fieldwork information (Z), interviewer characteristics and characteristics of the interview process (W), then we can write

$$X_1 = (V, W, Z), \qquad X_2 = (V, V^*, W, W^*), \qquad X_3 = (V, V^*).$$

The use of this large set of exclusion restrictions should protect against problems of collinearity, especially during the second estimation step.

Table 6 presents estimates of the probability of unit response for Models 3, 4 and 5. In particular, Model 3 is a probit model with a single indicator for unit and item response, Model 4 is a probit model for unit nonresponse, and Model 5 a bivariate probit model with sample selection which also accounts for the correlation between unit and item nonresponse. We find that the probability of unit response tends to fall with age. Women are less likely to participate than men, but the differences are not strongly significant. The interviewer's gender does not seem to matter, whereas the interviewer's age is positively related to unit response. The interviewer's education, the total number of calls and the length of the fieldwork are negatively related to unit response. This may simply reflect the strategy of increasing the number of calls and switching to more experienced interviewers when there are difficulties in reaching contact and gaining respondents' cooperation. A comparison of the estimates in Models 4 and 5 also shows that relaxing the conditional independence assumption between errors in the two response equations has only negligible effects on parameter estimates in the unit response equation.

Table 7 provides estimates of the probit model (Model 4) and the bivariate probit model (Model 5) for the probabilities of item response on the two consumption items. Estimates of the correlation coefficient ρ_{12} are relatively small, and the corresponding likelihood ratio tests never reject conditional independence between unit and item nonresponse. Accordingly, the differences between estimated coefficients of the probit model and the bivariate probit model are not statistically significant. By focusing on the predictors of item response, we find that the probability of item response tends to fall with the age of the household respondent. Living in a small city, being employed, being single, or being more educated are negatively related to item response, but the estimated effects are only weakly statistically significant. Even after controlling for respondent's background characteristics such as age and education, the cognitive function scores are positively related to item response probabilities, while other health measures are not. The positive coefficients on the income and wealth quartiles suggest that item nonresponse leads to selection of households with higher income and wealth. Furthermore, the negative coefficients on the dummies for income and wealth imputations suggest that nonresponse to income and wealth questions is positively related to nonresponse to consumption questions. Among characteristics of the interview process and the interviewers, we find that allowing the interviewee to be assisted by a proxy respondent and using more experienced interviewers (that is, interviewers with higher workload and better computer skill) both have a positive impact on the probability of item response.

Estimates of the Engel curves for food consumption at home and total nondurable consumption are presented in Tables 8 and 9 respectively. Here, a comparison of the five alternative models allows assessing the selectivity effects of unit and item nonresponse operating through the bias correction terms λ^{unit} and λ^{item}. For both food consumption at home and total nondurable consumption, the selection biases associated to unit and item nonresponse have opposite sign and therefore partly offset each other: the first (unit nonresponse) is positive, the second (item nonresponse) is negative. For food consumption at home, the coefficients on the bias correction terms are not statistically significant, and estimates of the five models are not very different. We conclude that unit and item nonresponse appear to be purely random. For total nondurable consumption, the coefficients on the bias correction terms are statistically significant. Therefore, neither unit nor item nonresponse errors are ignorable, and only the two generalized sample selection models (Models 4 and 5) provide consistent estimates for the model parameters of interest. The estimates of Model 3 are similar to those of Model 1 probably because relevant predictors of item response probability are omitted by the model with a single indicator for unit and item response. The estimates of Model 2 are instead similar to those of Models 4 and 5. Here, the main differences occur in the coefficients of the country dummies that are important predictors of both unit response and total nondurable consumption.

Before turning to the semiparametric specification of the model, we find interesting to investigate the importance of collinearity problems during the second step of the estimation procedure. Since our models are identified through a set of exclusion restrictions, multicollinearity should not be a major concern. Nevertheless, results provided in Table 7 show that most of the variables used as exclusion restrictions are not important predictors of item nonresponse. The lack of suitable exclusion restrictions could therefore be responsible for problems of collinearity which inflate the variance of the bias correction term associated with item nonresponse. In order to assess the relevance of these collinearity problems, we provide at the bottom of Tables 8 and 9 the largest variance inflation factor (VIF) among covariates of the second estimation step. Although the VIF associated with λ^{item} is, as expected, the highest, the informal rules of thumb usually adopted for the analysis of the VIF does not reveal any serious problem of collinearity.[12]

5.5 Semiparametric estimates (Preliminary, to be completed...)

This section presents estimates of a semiparametric two-step procedure that are robust to violations of the Gaussianity of the error terms. As mentioned above in Section 4, this analysis assumes that errors in the unit and the item response equations are independent. In principle, the conditional independence assumption could be relaxed, but the resulting semiparametric estimators would require an unreasonable amount of computer time. Furthermore, in our empirical application, this assumption is strongly supported by the results of the parametric estimation of the model (see Section 5.4).

In the first step of the procedure, parameters of the unit and the item response equations are estimated separately by the SNP estimator of Gallant and Nychka (1987). Since in this approach inference is conducted conditional on K (the degree of the Hermite polynomial used for approximating the unknown distribution function of the error terms), estimation is carried out by varying K from 3 to 5. Specifications underlying these alternative choices of K are then compared through likelihood ratio tests, AIC, and BIC. According to the various model selection criteria in Table 10, the preferred specification has $K = 4$ for unit response and $K = 3$ for item response. Parametric and semiparametric estimates of these models are presented in Tables 11 and 12 respectively. Because of the different scale, estimated coefficients of the probit model and the semiparametric model are not directly comparable. Estimates in Tables 11 and 12 are normalized by setting respectively the coefficients for the length of the fieldwork (*lfield*) and the length of the IV module (*ivlength*) equal to -0.01.[13]

[12] According to these rules, there is evidence of multicollinearity if the largest VIF is greater than 10 (see Chatterjee *et al.* 2000).

[13] Specifically, coefficients of the parametric and semiparametric models are divided by the absolute value of the coefficients associated with these variables time 0.01. Standard errors of the resulting nonlinear combinations are computed by the delta method.

For unit response, we find important differences between the parametric and the semiparametric estimates. In particular, the main differences occur for interviewers' characteristics, features of the fieldwork, and country dummies. The parameters τ of the Hermite polynomial expansion are significantly different from zero at the 1 percent level. Therefore, the probit specification is strongly rejected. The estimated error density exhibits multimodality, positive skewness and greater kurtosis than a standard normal density (see Figure 1).

For item response, the probit model is still rejected for food consumption at home, but not for total nondurable consumption. For food consumption at home, the estimated error density exhibits bimodality. Overall, however, the differences between the parametric and the semiparametric estimates of the regression coefficients are small. Thus, parametric estimates of the item response equations are only marginally affected by violations of the Gaussian distributional assumptions.

In the second step of the procedure, estimates of the two indexes $\hat{\mu}_1 = \hat{\beta}_1^\top X_1$ and $\hat{\mu}_2 = \hat{\beta}_2^\top X_2$ are used to estimate a partially linear model for the outcome variables of interest. Following Robinson (1988), the unknown conditional expectations in (13) are estimated nonparametrically by Nadaraya-Watson kernel regression estimators of the form

$$\hat{\mathrm{E}}(W_i \,|\, \hat{\mu}_{i1}, \hat{\mu}_{i2}, Y_{i1}Y_{i2} = 1) = \frac{A_{n_3}(W_i \,|\, \hat{\mu}_{i1}, \hat{\mu}_{i2}, Y_{i1}Y_{i2} = 1)}{A_{n_3}(1 \,|\, \hat{\mu}_{i1}, \hat{\mu}_{i2}, Y_{i1}Y_{i2} = 1)}, \tag{15}$$

with W_i equal to Y_{i3} or X_{i3}, and

$$A_n(W_i \,|\, \mu_{i1}, \mu_{i2}, Y_{i1}Y_{i2} = 1) = \frac{1}{(n-1)h_n^2} \sum_{j \neq i}^{n} W_j \, \mathcal{K}\left(\frac{\mu_{i1} - \mu_{j1}}{h_n}, \frac{\mu_{i2} - \mu_{j2}}{h_n} \right),$$

$$A_n(1 \,|\, \mu_{i1}, \mu_{i2}, Y_{i1}Y_{i2} = 1) = \frac{1}{(n-1)h_n^2} \sum_{j \neq i}^{n} \mathcal{K}\left(\frac{\mu_{i1} - \mu_{j1}}{h_n}, \frac{\mu_{i2} - \mu_{j2}}{h_n} \right),$$

where $\mathcal{K}(u,v) = K(u)\,K(v)$ is the product of two univariate bias reducing kernels, $K(t) = (3/2 - t^2/2)\,\phi(t)$, and $h_n = n^{-1/p}$ is a bandwidth that goes to zero as $n \to \infty$. In determining nonparametric estimates, we also trim observations for which $A_{n_3}(1 \,|\, \mu_{i1}, \mu_{i2}, Y_{i1}Y_{i2} = 1)$ is less than $n_3^{-1/r}$. After replacing the unknown conditional expectations in (13) with their nonparametric estimates, the vector of parameters β_3 is estimated through standard OLS with no intercept.[14] Standard errors of the OLS estimator are instead computed by the nonparametric bootstrap with 200 replications.

Semiparametric estimates of the partially linear model for food consumption at home and total nondurable consumption are presented in Tables 13 and 14 respectively. To explore sensitiveness of Robinson's estimator with respect to choice of the bandwidth parameter and the trimming

[14] As mentioned in Section 4, the intercept coefficient is absorbed in the nonlinear function g, and is not identified.

factor, estimation is carried out for alternative combinations of p and r.[15] In particular, results are presented for $p = 5$ and $r = 21$ (column a), $p = 6$ and $r = 13$ (column b), $p = 7$ and $r = 10$ (column c). Parametric estimates of Model 3 are also reported to facilitate comparisons.

To be completed...

6 Conclusions

In this paper we investigate problems of selectivity generated by unit and item nonresponse in cross-sectional surveys. The paper is organized in two parts.

In the first part, we analyze a general sample selection model in which unit and item nonresponse can simultaneously affect a regression relationship of interest through NMAR missing data mechanisms. Issues concerning identification and estimation have been considered for two alternative specifications of this model. In the parametric specification, errors in the two selection equations and in the equation for the outcome of interest are assumed to follow a trivariate Gaussian distribution. In the semiparametric specification, we relax assumptions on the Gaussianity error terms.

In the second part, we use data from the first wave of SHARE to investigate whether selectivity associated with unit and item nonresponse may bias the estimation of Engel curve for food consumption at home and total nondurable consumption.

To be completed...

[15] Combinations of p and r are selected to satisfy conditions imposed on the choice of the bandwidth parameter h_{n_3} and the trimming factor b_{n_3}(see Robinson 1988, Theorem 1).

References

Battistin E., R. Miniaci, and G. Weber (2003), “What Do We Learn from Recall Consumption Data?”, *Journal of Human Resources*, 38: 354-85.

Börsch-Supan A., A. Brugiavini, H. Jürges, J. Mackenbach, J. Siegrist and G. Weber (2005), *Health, Ageing and Retirement in Europe - First Results from the Survey of Health, Ageing and Retirement in Europe*, MEA, Mannheim.

Börsch-Supan A., and H. Jürges (2005), *The Survey of Health, Ageing and Retirement in Europe - Methodology*, MEA, Mannheim.

Browning M., T.F. Crossley, and G. Weber (2002), “Asking Consumption Questions in General Purpose Survey”, *Economic Journal*, 113:540-67.

Chatterjee S., A.S. Hadi, and B. Price (2000), *Regression Analysis by Examples*, 3d edition, New York: John Wiley & Sons.

Fitzgerald J., P. Gottschalk, and R. Moffitt (1998), “An Analysis of Sample Attrition in Panel Data: The Michigan Panel Study of Income Dynamics”, *Journal of Human Resources*, 33: 251-99.

Gabler S., F. Laisney, and M. Lechner (1993), “Semiparametric Estimation of Binary-Choice Models with an Application to Labor-Force Participation”, *Journal of Business and Statistics*, 11: 61-80.

Gallant A.R., and D.W. Nychka (1987), “Semi-Nonparametric Maximum Likelihood Estimation”, *Econometrica*, 55: 363-90.

Groves R.M., and M.P. Couper (1998), *Nonresponse in Household Interview Surveys*, John Wiley & Sons, New York.

Groves R.M., Dillman D. A., Eltinge J.L., Little R.J.A. (2002), *Survey Nonresponse*, John Wiley & Sons, New York.

Ham J.C. (1982), “Estimation of a Labour Supply Model with Censoring due to Unemployment and Underemployment”, *Review of Economic Studies*, 49: 335-54.

Heckman J. (1979), “Sample Selection Bias as a Specification Error”, *Econometrica*, 47: 153-61.

Korinek A., J.A. Mistiaen, and M. Ravallion (2004),“Survey Nonresponse and the Distribution of Income”, ...

Ichimura H., and L.F. Lee (1991), “Semiparametric Least Square of Multiple Index Models: Single Equation Estimation”, In W.A. Barnett, J. Powell, and G. Tauchen (eds.), *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, Cambridge: University Press, Cambridge.

Lessler J. T., and W.D. Kalsbeek (1992), *Nonresponse Errors in Surveys*, John Wiley & Sons, New York.

Leung S.F., and S. Yu (1996), “On the Choice between Sample Selection and Two-Part Models”, *Journal of Econometrics*, 72: 197-229.

Little R.J.A., and D.B. Rubin (2002), *Statistical Analysis with Missing Data*, 2nd edition, New York: John Wiley & Sons.

Melenberg B., and A. van Soest (1996), “Measuring the Costs of Children: Parametric and Semiparametric Estimators”, *Statistica Neerlandica*, 50: 171-92.

Nicoletti C., and F. Peracchi (2005), “Survey Response and Survey Characteristics: Microlevel Evidence from the European Community Household Panel”, *Journal of the Royal Statistical Society*, *Series A*, 168, 763-81.

O'Muircheartaigh C., and P. Campanelli (1999), "A Multilevel Exploration of the Role of Interviewers in Survey Nonresponse", *Journal of the Royal Statistical Society*, *Series B*, 162: 437-46.

Poirier D. (1980), "Partial Observability in Bivariate Probit Models", *Journal of Econometrics*, 12: 209-17.

Puhani P. A. (2000), "The Heckman Correction for Sample Selection and its Critique", *Journal of Economic Surveys*, 14: 53-68.

Riphahn R. T., and O. Serfling (2002), "Item Nonresponse on Income and Wealth Questions", Discussion Paper No. 573, IZA Bonn.

Robinson P.M. (1988), "Root-N-Consistent Semiparametric Regression", *Econometrica*, 56: 931-54.

Rubin D.B. (1976), "Inference and Missing Data", *Biometrika*, 63: 581-92.

Stewart M.B. (2004), "Semi-Nonparametric Estimation of Extended Ordered Probit Models", *The Stata Journal*, 4: 27-39.

Tallis G. M. (1961), "The Moment Generating Function of the Truncated Multi-Normal Distribution", *Journal of the Royal Statistical Society*, *Series B*, 23: 22329.

Vella F. (1998), "Estimating Models with Sample Selection Bias: A Survey", *The Journal of Human Resources*, 33: 127-69.

Winter J. (2004), "Response Bias in Survey-Based Measures of Household Consumption", *Economics Bulletin*, 3: 1-12.

Table 1: Unweighted household response rates.

Country	Eligible	Response rate	Noncontact rate	Refusal rate	Other noninterview rate
Denmark	1742	0.61	0.09	0.29	0.01
Germany	2583	0.60	0.05	0.34	0.01
Italy	2505	0.54	0.08	0.36	0.02
Netherlands	2509	0.62	0.05	0.32	0.01
Spain	2619	0.50	0.13	0.36	0.01
Sweden	3956	0.47	0.06	0.42	0.05
Total	15914	0.56	0.08	0.34	0.02

Table 2: Unweighted item response rates for consumption expenditure questions.

Country	Eligible	Food at home	Total consumption
Denmark	1178	0.81	0.79
Germany	1566	0.88	0.88
Italy	1376	0.85	0.84
Netherlands	1559	0.89	0.77
Spain	1341	0.78	0.77
Sweden	1850	0.93	0.90
Total	8870	0.86	0.83

Table 3: Summary statistics for consumption expenditure questions (Yearly amounts expressed in 100 Euro. Empirical distributions trimmed symmetrically by 2 percent.).

Variable	Obs.	Mean	Std.	Min	Max
Food at home	7496	49.5	39.8	2.3	640.0
Total consumption	7204	118.8	84.2	9.5	960.0

Table 4: Summary statistics for the predictors of unit response.

Variable	Obs.	Mean	Std.	Description
agecl_1	15884	.37	0.48	Age class 50–59
agecl_2	15884	.52	0.50	Age class 60–79
agecl_3	15884	.11	0.32	Age class 80+
female_gs	15893	.54	0.50	Female
iv_age	15900	49.2	11.5	Interviewer age
iv_female	15900	.71	0.45	Interviewer female
iv_yedu	15604	13.5	3.0	Interviewer years of education
iv_wl	15914	43.6	35.1	Interviewer workload (households visited in person)
tot_call	15914	3.7	5.0	Total number of call attempts
lfield	15914	41.1	46.2	Length of fieldwork (days between first and last call)

Table 5: Summary statistics for the predictors of item response.

Variable	Obs.	Mean	Std.	Description
hr_age	8856	64.8	10.4	HR age
hr_female	8870	0.54	0.50	HR female
hr_yedu	8842	10.0	4.5	HR years of education
hr_working	8820	0.34	0.47	HR working (1 - paid work in the last 4 weeks)
single	8846	0.33	0.47	HR leaving as single
hsize	8870	2.11	1.02	Household size
s_city	8681	0.23	0.42	Household leaves in a small city
math	8806	3.30	1.18	Score on mathematical test (1–5)
orient	8822	3.76	0.65	Score on orientation in time test (1–5)
recall	8753	3.32	2.03	Score on delayed recall test (1–11)
eurod	8737	2.32	2.26	EURO depression scale index (1–12)
p_money	8870	0.04	0.18	Self-reported problems in managing money
sp_health	8830	0.69	0.46	Less than good self-reported health
adl1	8827	0.10	0.30	At least one ADL limitation
income_q1	8867	0.25	0.43	1st quartile gross annual HH income
income_q2	8867	0.25	0.43	2nd quartile gross annual HH income
income_q3	8867	0.25	0.43	3rd quartile gross annual HH income
income_q4	8867	0.25	0.43	4th quartile gross annual HH income
inc_mis	8870	0.55	0.50	Gross annual income missing
wealth_q1	8870	0.25	0.43	1st quartile net financial assets
wealth_q2	8870	0.25	0.43	2nd quartile net financial assets
wealth_q3	8870	0.25	0.43	3rd quartile net financial assets
wealth_q4	8870	0.25	0.43	4th quartile net financial assets
wea_mis	8870	0.52	0.50	Net financial assets missing
iv_length	8797	1.8	1.5	Length of the IV module (min.)
int_length	8775	71.8	26.1	Length of the HR interview (min.)
f_proxy	8710	0.02	0.13	Full proxy interview (CO module)
p_proxy	8710	0.06	0.25	Partial proxy interview (CO module)
int_home	8708	0.96	0.20	Interview done at the respondent home
int_oper	8870	0.01	0.11	Non-household members present during the interview

Table 6: Parametric estimates for unit response (* denotes an observed significance level between 1% and 5%, ** denotes an observed significance level below 1%.

	Model 3			Model 5	
	Food	Total	Model 4	Food	Total
Variable	at home	consumption		at home	consumption
agecl_1	.0716 **	.0778 **	.0363	.0352	.0362
agecl_3	-.2805 **	-.2623 **	-.2016 **	-.2013 **	-.2021 **
female_gs	-.0175	-.0532 *	-.0409 *	-.0411 *	-.0419 *
iv_agec	.0029 **	.0031 **	.0020 *	.0020 *	.0020 *
iv_agec^2	.0001	.0002 *	.0002 **	.0002 **	.0002 **
iv_female	.0084	.0003	.0131	.0133	.0133
iv_yedu	-.0220 **	-.0174 **	-.0174 **	-.0173 **	-.0173 **
iv_wl	.0011 **	.0008 *	.0005	.0005	.0005
tot_call	-.0216 **	-.0227 **	-.0248 **	-.0249 **	-.0247 **
lfield	-.0025 **	-.0023 **	-.0030 **	-.0030 **	-.0030 **
DK	.0200	.0276	.2481 **	.2478 **	.2481 **
DE	.0592	.0826 *	.1302 **	.1296 **	.1301 **
IT	-.2106 **	-.1767 **	-.0612	-.0616	-.0612
NL	-.0342	-.1586 **	.1191 **	.1188 **	.1197 **
ES	-.3688 **	-.3289 **	-.1751 **	-.1760 **	-.1749 **
_cons	.4467 **	.3629 **	.5253 **	.5253 **	.5249 **
n	15129	15129	15129	15129	15129

Table 7: Parametric estimates for item response. To save space, country dummies are not reported.

	Food at home		Total consumption	
Variable	Model 4	Model 5	Model 4	Model 5
hr_agec	-.0099 **	-.0101 **	-.0141 **	-.0142 **
hr_agec^2	-.0001	-.0001	.0001	.0001
hr_female	.1686 **	.1595 **	-.0440	-.0480
hr_yedu	-.0107	-.0107	-.0089	-.0089
hr_working	-.1164 *	-.1172 *	-.0875	-.0888
single	-.0869	-.0848	-.0957 *	-.0943 *
hsize	.0046	.0046	-.0337	-.0332
s_city	-.0568	-.0535	-.1404 **	-.1365 **
math	.0328	.0323	.0132	.0131
orient	.1361 **	.1327 **	.0570	.0556
recall	.0205	.0200	.0329 **	.0324 **
eurod	.0185 *	.0183 *	.0159	.0158
p_money	-.1856	-.1835	-.1009	-.1007
sp_health	-.0653	-.0639	-.0505	-.0496
adl1	-.0914	-.0898	.0481	.0474
income_q2	.1663 **	.1629 **	.1446 **	.1430 **
income_q3	.2629 **	.2586 **	.2322 **	.2299 **
income_q4	.2066 **	.2028 **	.1745 **	.1726 **
inc_mis	-.5130 **	-.5050 **	-.5405 **	-.5356 **
wealth_q2	.2686 **	.2636 **	.3196 **	.3156 **
wealth_q3	.2778 **	.2734 **	.3400 **	.3366 **
wealth_q4	.2596 **	.2554 **	.3795 **	.3756 **
wea_mis	-.3005 **	-.2968 **	-.4399 **	-.4368 **
int_length	.0003	.0003	-.0001	-.0001
f_proxy	-.1540	-.1539	-.0986	-.1000
p_proxy	.2790 **	.2764 **	.1707 *	.1709 *
int_home	.1757	.1795	.2535 **	.2572 **
iv_length	-.0296 **	-.0291 **	-.0335 **	-.0331 **
iv_agec	.0026	.0029	.0032	.0034
iv_agec^2	-.0001	-.0001	.0000	.0000
iv_female	.0022	.0034	.0213	.0220
iv_yedu	-.0191 **	-.0216 **	-.0036	-.0058
iv_wl	.0025 **	.0026 **	.0023 **	.0024 **
_cons	1.1184 **	.9198 **	1.1510 **	.9934 **
n_2	8343	8343	8343	8343
ρ_{12}		.25		.20
LR stat.		2.04		1.40

Table 8: Parametric estimates for food consumption at home. Standard errors in Models 4 and 5 are computed via the nonparametric bootstrap with 200 replications. The F test for selectivity due to nonresponse has 2 degree of fredom.

Variable	Model 1	Model 2	Model 3	Model 4	Model 5
hr_agec	-.1372 *	-.1141	-.1463 *	-.1154	-.1173
hr_agec^2	-.0014	-.0011	-.0019	-.0014	-.0013
hr_female	-.6723	-1.0175	-.7137	-1.1104	-1.0616
hr_yedu	.6561 **	.6827 **	.6550 **	.6844 **	.6820 **
hr_working	-1.9995	-1.7261	-2.0428	-1.7451	-1.7692
single	-6.1386 **	-5.9074 **	-6.1584 **	-5.9052 **	-5.9320 **
hsize	9.1801 **	9.1529 **	9.1886 **	9.1571 **	9.1591 **
s_city	-3.0464 **	-2.8938 **	-3.0133 **	-2.8498 **	-2.8705 *
math	.9791 *	.9220	.9793 *	.9186	.9252 *
orient	-1.0676	-1.4855	-1.0775	-1.5357	-1.4868
recall	-.1478	-.1895	-.1449	-.1922	-.1881
eurod	-.1864	-.2335	-.1836	-.2355	-.2303
p_money	3.3942	4.1331	3.3062	4.1317	4.0536
sp_health	.3159	.4383	.3383	.4678	.4523
adl1	1.8020	1.9702	1.8164	1.9937	1.9731
income_q2	4.5850 **	4.1464 **	4.5936 **	4.1081 **	4.1586 **
income_q3	6.4019 **	5.7814 **	6.4063 **	5.7211 **	5.7897 **
income_q4	8.6045 **	8.0739 **	8.6165 **	8.0304 **	8.0887 **
inc_mis	-1.4115	-.3109	-1.4428	-.2341	-.3529
wealth_q2	-1.2389	-1.8365	-1.2358	-1.8973	-1.8316
wealth_q3	1.8521	1.2432	1.8691	1.1984	1.2638
wealth_q4	1.7104	1.1334	1.7304	1.0925	1.1546
wea_mis	2.0540 *	2.6909 *	2.0322 *	2.7292 *	2.6614 *
DK	1.4370	2.7389	1.9051	3.6946	3.4180
DE	12.2940 **	12.5080 **	12.6369 **	12.9751 **	12.8819 **
IT	24.5841 **	25.7032 **	24.5809 **	26.1165 **	25.9491 **
NL	16.1005 **	16.8427 **	16.5916 **	17.5761 **	17.3946 **
ES	36.1994 **	37.2255 **	35.9838 **	37.4959 **	37.3591 **
λ^{unit}				2.8767	2.4931
λ^{item}		-7.5852		-8.3263	-6.9457
λ			3.1917		
_cons	8.5574	11.2834 *	5.8704	9.2770	9.2732
n_3	7194	7194	7194	7194	7194
F stat.				1.39	1.19
VIF(λ^{item})				7.55	7.35

Table 9: Parametric estimates for total nondurable consumption.

Variable	Model 1	Model 2	Model 3	Model 4	Model 5
hr_agec	-.2433	-.0506	-.2789 *	-.0560	-.0642
hr_agec^2	.0010	.0000	-.0005	-.0021	-.0018
hr_female	-.8914	-.2478	-1.2160	-.5462	-.5430
hr_yedu	3.1308 **	3.2544 **	3.1237 **	3.2550 **	3.2490 **
hr_working	7.0921 **	8.3832 **	6.9384 **	8.2185 **	8.1725 **
single	-11.2351 **	-9.7006 **	-11.3105 **	-9.6729 **	-9.7600 **
hsize	15.0224 **	15.4302 **	15.0539 **	15.5195 **	15.4916 **
s_city	-5.4186 *	-3.4619	-5.3292 *	-3.1144	-3.2555
math	2.5428 *	2.4441 *	2.5509 *	2.4559 *	2.4620 *
orient	.6236	-.3841	.5798	-.5566	-.4916
recall	.5584	.1535	.5697	.1276	.1498
eurod	-.1058	-.3302	-.1001	-.3395	-.3273
p_money	5.8168	8.0690	5.4542	7.7237	7.6177
sp_health	-.1644	.4085	-.0854	.5763	.5327
adl1	1.9367	1.0715	1.9825	1.0358	1.0731
income_q2	7.2244 **	5.0580	7.2468 **	4.8732	5.0038
income_q3	18.8715 **	15.7920 **	18.8726 **	15.4836 **	15.6640 **
income_q4	35.1644 **	32.6011 **	35.2099 **	32.4056 **	32.5547 **
inc_mis	-.1013	6.7949	-.2224	7.2489 *	6.8644
wealth_q2	-1.0916	-5.1350	-1.0841	-5.5318	-5.3022
wealth_q3	5.6637 *	1.4106	5.7229 *	1.1108	1.3469
wealth_q4	15.7174 **	10.9524 **	15.7737 **	10.5943 **	10.8588 **
wea_mis	1.1227	6.6737 *	1.0639	7.0958 *	6.7963 *
DK	-15.6851 **	-9.6490 *	-13.9261 **	-3.8128	-4.7703
DE	7.1788 *	6.1823	8.6411 **	8.9289 **	8.6733 *
IT	29.7056 **	33.0788 **	29.8885 **	35.2981 **	34.9076 **
NL	37.4428 **	45.5221 **	38.2639 **	50.5252 **	49.5840 **
ES	47.2019 **	50.3912 **	46.6226 **	51.8244 **	51.5568 **
λ^{unit}				18.2535 *	16.5190 *
λ^{item}		-39.3742 *		-43.1858 **	-38.3924 *
λ			11.3312		
_cons	6.2336	15.2485	-3.4895	1.7557	2.6268
n_3	6910	6910	6910	6910	6910
F stat.				12.20 **	12.28 **
VIF(λ^{item})				9.37	9.19

Table 10: Model selection criteria for the semiparametric specifications.

K	Log-lik.	LR stat.	AIC	BIC
Unit response				
3	-10040.79		20117.58	20254.82
4	-10013.72	54.13 **	20065.45	20254.82
5	-10013.67	0.10	20067.35	20254.82
Item response (food consumption at home)				
3	-2971.45		6024.91	6313.10
4	-2971.09	0.73	6026.18	6321.40
5	n.c.	n.c.	n.c.	n.c.
Item response (total nondurable consumption)				
3	-3398.12		6878.25	7166.44
4	-3397.90	0.46	6879.79	7175.02
5	-3397.40	0.99	6880.80	7183.05

Table 11: Semiparametric estimates for unit response. The results are based on the scale normalization $\beta_{\texttt{lfield}} = -.01$. Standard errors computed by the delta method. LR stat. is a likelihood-ratio test for the joint significance of τ_3 and τ_4.

Variable	Parametric	Semipar.
agecl_1	.1228	.1448
agecl_3	-.6822 **	-.8047 **
female_gs	-.1386	-.1805
iv_agec	.0068	.0017
iv_agec^2	.0006 **	.0005
iv_female	.0443	-.0077
iv_yedu	-.0589 **	-.0749 **
iv_wl	.0016	.0048 **
tot_call	-.0839 **	-.4090 **
DK	.8395 **	1.2614 **
DE	.4405 **	.5161 *
IT	-.2072	.1360
NL	.4030 **	.8712 **
ES	-.5924 **	-.4439
τ_1		-4.6378 **
τ_2		-3.8454 **
τ_3		1.1993 **
τ_4		.5686 **
_cons	1.7774 **	
n_1	15129	15129
Std. dev.		1.92
Skewness		0.33
Kurtosis		1.90
LR stat.		151.67 **

Table 12: Semiparametric estimates for item response. LR stat. is a likelihood-ratio test for the significance of τ_3.

	Food at home		Total consumption	
Variable	Parametric	Semipar.	Parametric	Semipar.
hr_agec	-.0033 *	-.0039 *	-.0042 **	-.0041 **
hr_agec^2	-.0000	-.0000	.0000	.0000
hr_female	.0569 *	.0625 *	-.0132	-.0123
hr_yedu	-.0036	-.0035	-.0027	-.0026
hr_working	-.0393	-.0519	-.0261	-.0247
single	-.0293	-.0306	-.0286	-.0272
hsize	.0015	-.0003	-.0101	-.0101
s_city	-.0192	-.0214	-.0420 *	-.0402 *
math	.0111	.0129	.0039	.0042
orient	.0460 *	.0386 *	.0170	.0153
recall	.0069	.0099	.0098 *	.0099 *
eurod	.0063	.0055	.0047	.0047
p_money	-.0626	-.0585	-.0301	-.0263
sp_health	-.0220	-.0196	-.0151	-.0143
adl1	-.0308	-.0278	.0144	.0127
income_q2	.0561 *	.0507	.0432 *	.0418 *
income_q3	.0888 *	.0808 *	.0694 *	.0666 *
income_q4	.0698 *	.0672	.0522 *	.0493 *
inc_mis	-.1732 *	-.2076 *	-.1615 **	-.1618 **
wealth_q2	.0907 *	.1069 *	.0955 **	.0951 **
wealth_q3	.0938 *	.1045 *	.1016 **	.0997 **
wealth_q4	.0876 *	.1050 *	.1134 **	.1124 **
wea_mis	-.1015 *	-.1174 *	-.1315 **	-.1278 **
hrint_min	.0001	.0001	-.0000	-.0000
f_proxy	-.0520	-.0598	-.0295	-.0246
p_proxy	.0942 *	.1083 *	.0510	.0518
int_home	.0593	.0639	.0758 *	.0757 *
ivlength	-.0100	-.0100	-.0100	-.0100
iv_agec	.0009	.0012	.0010	.0009
iv_agec2	-.0000	-.0000	.0000	-.0000
iv_female	.0007	.0029	.0064	.0049
iv_yedu	-.0064	-.0069 *	-.0011	-.0012
iv_wl	.0008 *	.0009 *	.0007 *	.0007 *
DK	-.2196 *	-.2635 *	-.1591 **	-.1592 **
DE	-.0647	-.1048 *	-.0016	-.0061
IT	-.1926 *	-.2234 *	-.0917 *	-.0908 *
NL	-.1983 *	-.2416 *	-.2327 **	-.2283 **
ES	-.2076 *	-.2402 *	-.1130 *	-.1109 **
τ_1		.0304		-.0080
τ_2		-.1022 *		-.0498
τ_3		-.0479 *		-.0161
n_2	8343	8343	8343	8343
Std. dev.		1.56		.95
Skewness		0.30		.51
Kurtosis		2.78		5.61
LR Test		14.64 **		1.68

Table 13: Semiparametric estimates of food consumption at home. Bandwidth parameter and trimming factor in semiparametric models are respectively equal to $(n_3)^{-1/p}$ and $(n_3)^{-1/r}$. Hausman 1 is a Hausman-type test computed over all coefficients except the intercept and the bias correction terms (28 d.o.f). Hausman 2 is a Hausman-type test computed for the coefficients on HH size, income and wealth quartiles, and country dummies (12 d.o.f.).

		Semiparametric		
		Model A	Model B	Model C
Variable	Parametric	$p = 5$, $r = 21$	$p = 6$, $r = 13$	$p = 7$, $r = 10$
hr_agec	-.1154	-.1276	-.1302	
hr_agec^2	-.0014	-.0016	-.0017	
hr_female	-1.1104	-1.2704	-1.2905	
hr_yedu	.6844 **	.6901 **	.6805 **	
hr_working	-1.7451	-1.4568	-1.5034	
single	-5.9052 **	-5.8420 **	-5.8689 **	
hsize	9.1571 **	9.0978 **	9.1364 **	
s_city	-2.8498 **	-2.7738 **	-2.8259 **	
math	.9186	.7909	.8383	
orient	-1.5357	-1.9549	-1.7558	
recall	-.1922	-.1650	-.1584	
eurod	-.2355	-.1959	-.1728	
p_money	4.1317	4.1570	4.0126	
sp_health	.4678	.3435	.2908	
adl1	1.9937	1.9136	1.9322	
income_q2	4.1081 **	3.9664 *	4.0258 *	
income_q3	5.7211 **	5.6412 **	5.8064 **	
income_q4	8.0304 **	7.9168 **	8.0922 **	
inc_mis	-.2341	.0578	-.2775	
wealth_q2	-1.8973	-1.5534	-1.5430	
wealth_q3	1.1984	1.0746	1.2595	
wealth_q4	1.0925	.8914	1.0378	
wea_mis	2.7292 *	3.2901 *	3.0450 *	
DK	3.6946	6.9867	6.3047	
DE	12.9751 **	14.5094 **	14.5565 **	
IT	26.1165 **	28.0741 **	27.6258 **	
NL	17.5761 **	19.7250 **	19.5662 **	
ES	37.4959 **	36.8591 **	36.6881 **	
λ^{unit}	2.8767			
λ^{item}	-8.3263			
_cons	9.2770			
n_3	7194	7194	7194	
Hausman 1		11.97	12.73	
Hausman 2		2.83	2.44	

Table 14: Semiparametric estimates of total nondurable consumption.

			Semiparametric	
		Model A	Model B	Model C
Variable	Parametric	$p = 5$, $r = 21$	$p = 6$, $r = 13$	$p = 7$, $r = 10$
n_3				
Hausman				

Figure 1: Semiparametric estimates of density functions.

a) Unit response

b) Item response (food at home)

c) Item response (total consumption)