JOURNAL OF APPLIED ECONOMETRICS
J. Appl. Econ. (2011)
Published online in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/jae.1232

ESTIMATING ENGEL CURVES UNDER UNIT AND ITEM NONRESPONSE

GIUSEPPE DE LUCA[a] AND FRANCO PERACCHI[b,c]*

[a] *ISFOL, Rome, Italy*
[b] *University of Rome Tor Vergata, Rome, Italy*
[c] *Einaudi Institute for Economics and Finance, Rome, Italy*

SUMMARY

This paper estimates food Engel curves using data from the first wave of the Survey on Health, Aging and Retirement in Europe (SHARE). Our statistical model simultaneously takes into account selectivity due to unit and item nonresponse, endogeneity problems, and issues related to flexible specification of the relationship of interest. We estimate both parametric and semiparametric specifications of the model. The parametric specification assumes that the unobservables in the model follow a multivariate Gaussian distribution, while the semiparametric specification avoids distributional assumptions about the unobservables.

1. INTRODUCTION

Starting with the pioneering work of Engel (1857), the link between household food expenditure and household income, or food Engel curve, has been one of the most investigated economic relationships. The study of Engel curves, however, is subject to a number of problems that are still unsettled.

First, there are theoretical and empirical reasons for avoiding the assumption that observed household income is exogenous. According to economic theory, household income is an outcome of the utility maximization problem faced by the household because it reflects choices, such as labor and saving decisions by its members, that are jointly made with consumption expenditure decisions (see, for example, Blundell *et al*., 2007). Problems of endogeneity also arise because measuring household income is not easy and may be subject to error (Hausman *et al*., 1991; Newey, 2001). In either case, estimation of the parameters of an Engel curve requires the availability of a suitable set of instruments to control for endogeneity bias.

Second, recent developments in the nonparametric literature have emphasized the importance of relaxing strong parametric assumptions about the shape of Engel curves. Popular models of consumer demand, such as the Almost Ideal Demand System (AIDS) of Deaton and Muellbauer (1980) and the Translog of Jorgenson *et al*. (1982), force budget shares to be linear in the logarithm of household income. The Quadratic AIDS (QUAIDS) of Banks *et al*. (1997) generalizes the above systems by allowing the effect of household income to be nonmonotonic but still parametric. More recent studies by Blundell *et al*. (2003, 2007) and Imbens and Newey (2009) focus on nonparametric methods to capture the observed patterns of consumer behavior for certain categories of consumption expenditure.

A third problem that plagues empirical studies based on survey data is nonresponse. It is important to distinguish between two types of nonresponse. The first—unit nonresponse—occurs

* Correspondence to: Franco Peracchi, Faculty of Economics, Tor Vergata University, via Columbia 2, Rome 00 133, Italy.
E-mail: franco.peracchi@uniroma2.it

when eligible sample units fail to participate in a survey because of noncontact or explicit refusal to cooperate. The second—item nonresponse—occurs when responding units do not provide useful answers to particular items of the survey instrument. This is often the case with household income and expenditure, as both variables are typically collected through a number of open-ended and retrospective questions that are sensitive and difficult to answer precisely.

The distinction between unit and item nonresponse is important for data users because they may improve model specification by exploiting the different information available, at least in principle, for the two types of nonresponse. For unit nonresponse this information is necessarily confined to that obtained from the sampling frame or the data collection process, whereas for item nonresponse one can use the additional information collected during the interview.

The distinction is also important at the survey design stage, where resources have to be allocated efficiently to reduce the various sources of nonsampling error. Well-designed surveys aim to reduce unit nonresponse by appropriately choosing fieldwork procedures, interview modes, interviewer training and incentive schemes. Several studies (see, for example, Groves and Couper, 1998; Groves *et al*., 2002; Riphahn and Serfling, 2005) show that these characteristics help predict the response rates attained in a survey. For item nonresponse, aspects of questionnaire design (e.g. length of the interview, wording and reference period for the questions, etc.) may also be important.

This paper is mainly concerned with problems of nonresponse in the first wave of a panel survey, where nonresponse rates are typically much higher than in subsequent waves. Despite its relevance, nonresponse in the first wave of a panel has received little attention in the literature relative to panel attrition, largely because of the lack of information on unit nonrespondents and on the interview process. The data analyzed in this paper, namely the first wave of the Survey on Health, Aging and Retirement in Europe (SHARE), represent an important exception because of the richness of the information provided on both unit nonrespondents and the interview process.

Unfortunately, despite the preventive measures adopted in many sample surveys, response rates are rarely close to 100%. This may explain why most of the survey nonresponse literature focuses on statistical methods for *ex post* adjustments (see Lessler and Kalsbeek, 1992; Little and Rubin, 2002). These adjustments typically require assumptions about the nature of the missing data mechanism. Following Rubin (1976), we say that data on an outcome of interest are missing completely at random (MCAR) if missingness depends on neither the observed outcome nor the observed covariates, are missing at random (MAR) if after conditioning on the observed covariates there is no relation between missingness and the observed outcome, and are not missing at random (NMAR) if missingness and the observed outcome are related even after conditioning on the observed covariates.

The MAR assumption is the basis of most common ways of handling unit and item nonresponse. Weighting procedures, which involve assigning weights to sample respondents to compensate for systematic differences relative to nonrespondents, have typically been used to deal with unit nonresponse. Imputation procedures, which fill in missing values to produce a completed dataset, have typically been used to deal with item nonresponse. Although weighting and imputation procedures represent the standard practice, they are not immune from criticism. First, relying on the MAR assumption when the underlying missing data mechanism is NMAR may lead to invalid inference about the population parameters of interest. The MAR assumption may be particularly restrictive in the case of unit nonresponse because of the limited information available to construct sample weights. Second, as pointed out by Heckman and Navarro (2004), the available procedures offer little guidance on how to pick the covariates that should account for sample selection and are not robust to the choice of the conditioning set.

Our paper differs from previous studies in several respects. First, we consider problems of selectivity due to both unit and item nonresponse, and we analyze these problems jointly. Second, we allow the missing data mechanism underlying the two types of nonresponse to be NMAR.

DOI: 10.1002/jae

Third, we simultaneously address issues of flexible specification of the regression function of interest and problems of nonresponse and endogeneity by using an extended sample selection model where a partially linear specification of the food Engel curve is subject to selectivity due to unit and item nonresponse and endogeneity of household income. Our model is closely related to the sample selection model analyzed by Das *et al*. (2003) (henceforth DNV). The main difference with respect to DNV is partial observability of one of the selection indicators, as item response behavior cannot be observed for those who are unit nonrespondents. We focus attention on two alternative specifications of the model, one parametric and the other semiparametric. The parametric specification is easy to estimate but relies on the strong assumption that the unobservables in the model follow a multivariate Gaussian distribution with zero mean and nondiagonal covariance matrix. The semiparametric specification is more appealing, for it avoids distributional assumptions.

The remainder of this paper is organized as follows. Section 2 describes our data. Section 3 presents the statistical model that we use to estimate the food Engel curve under endogeneity of household income and selectivity due to unit and item nonresponse. Section 4 describes our empirical results. Finally, Section 5 offers some conclusions.

2. DATA

Our data are from Release 2.1 of the first wave of the Survey of Health, Aging and Retirement in Europe (SHARE), a multidisciplinary and cross-national biannual household panel survey coordinated by the Mannheim Research Institute for the Economics of Aging (MEA). The survey collects data on health, socio-economic status, and social and family networks for nationally representative samples of elderly people in the participating countries.

The first wave, conducted in 2004, covers about 19,500 households and about 28,500 individuals in 11 European countries (Austria, Belgium, Denmark, France, Germany, Greece, Italy, the Netherlands, Spain, Sweden and Switzerland). The target population consists of people aged 50 and older living in residential households, plus their (possibly younger) spouse/partner. National samples are selected through probability sampling, but sampling procedures are not completely standardized across countries. We only consider the countries (Denmark, Italy, the Netherlands, Spain, and Sweden) for which the sampling frame contains basic information on the individuals selected for interview, namely their gender and year of birth. The interview mode adopted by SHARE is computer-assisted personal interviewing (CAPI), supplemented by show-cards and a self-administered paper-and-pencil questionnaire. Except for Denmark, all national samples consist of a main sample (on average about 80% of the total sample) and a vignette sample (the remaining 20%).[1] For the vignette sample, who were interviewed later than the main sample, a section of the self-administered questionnaire was replaced by one with anchoring vignette questions.[2]

Like most other sample surveys, the first wave of SHARE is affected by substantial unit nonresponse. Table I presents summary statistics on survey participation. Of the 15,895 households selected for interview, only 8750 agreed to participate, corresponding to an unweighted household response rate of 55%. Household response rates vary considerably by country and sample type, ranging from a minimum of 47% for the Swedish main sample to a maximum of 67% for the Danish main sample. As for the reasons for household nonresponse, refusal to participate represents about three-quarters of the cases, noncontact about one-quarter.

[1] In Sweden, the main sample also included a supplementary sample which was fielded to increase the number of completed interviews.
[2] Additional methodological details about survey organization, sampling design, response rates, weighting and imputation strategies can be found in Börsch-Supan and Jürges (2005).

DOI: 10.1002/jae

Table I. Summary statistics on household survey participation by country and sample type

Country	Sample type	Eligible	Interviewed	Response rate	Noncontact rate	Refusal rate
DK	Main	1748	1174	0.67	0.05	0.28
ES	Main	2620	1339	0.51	0.19	0.30
	Vignette	683	414	0.61	0.13	0.26
IT	Main	2502	1376	0.55	0.10	0.35
	Vignette	677	374	0.55	0.12	0.33
NL	Main	2517	1563	0.62	0.09	0.29
	Vignette	657	391	0.60	0.10	0.31
SE	Main	3951	1850	0.47	0.14	0.39
	Vignette	540	269	0.50	0.11	0.40
All	Main	13 338	7302	0.55	0.12	0.33
	Vignette	2557	1448	0.57	0.11	0.32
	Total	15 895	8750	0.55	0.12	0.33

For the households that agreed to participate in the survey, the SHARE interview collects data on household consumption expenditure through retrospective open-ended questions on three consumption categories (food at home, food outside home, and phone) and on total nondurable consumption. These questions are asked of a single 'household respondent', namely the eligible person who is most knowledgeable about housing matters. We focus on the food share, namely the ratio of household food expenditure to total household income.[3] Both the numerator and the denominator of the ratio are generated variables; that is, they are not asked directly but are obtained by aggregating a number of separate components. Food expenditure in the numerator is obtained by adding up household expenditure on food at home and outside home, whereas total household income in the denominator is obtained by aggregating 19 income components collected at the individual level and six income components collected at the household level. Table II summarizes the complex process underlying item nonresponse by providing the unweighted item response rates on food expenditure, total household income, and food share by country and sample type. These generated variables are regarded as missing if any of their components is missing. Based on this definition, the cross-country average of item nonresponse rates is 18% for food expenditure, 64% for household income, and 68% for the food share, with considerable variation across countries. In particular, the item nonresponse rate on food share ranges from a minimum of 57% in Denmark to a maximum of 77% in Spain. Overall, the complete-case subsample consists of 2805 households.

The most important reason for missing data on food share is item nonresponse to questions about income components. This reflects two problems. The first is simply the large number of income components considered in SHARE. The second arises because, according to the SHARE fieldwork rules, a household with two spouses is considered as interviewed if at least one of them agrees to participate. If the other does not, then household income must be imputed because the incomes of the nonresponding spouse are missing. This 'missing spouse problem', which affects 26% of the households in the survey, induces a negative correlation between the indicators of unit and item nonresponse.

The public-use SHARE data include a set of weights to account for unit nonresponse and a set of imputations to account for item nonresponse. The weights are constructed using the calibration methodology of Deville and Särndal (1992), while the imputations are constructed using the multivariate iterative procedure of Buuren *et al*. (2006), which attempts to preserve the correlation

[3] We do not divide food expenditure by total nondurable consumption expenditure because this variable is likely to be severely understated (Browning and Madsen, 2005). Browning *et al*. (2003) and Winter (2004) argue that 'one-shot' retrospective questions do not provide reliable measures of consumption expenditure aggregates. This may explain why total nondurable consumption expenditure has been excluded from the set of imputed variables provided by SHARE.

DOI: 10.1002/jae

Table II. Unweighted item response rates by country and sample type

Country	Sample type	Eligible	Respondents			Item response rate		
			Food	Income	Share	Food	Income	Share
DK	Main	1174	934	565	506	0.80	0.48	0.43
ES	Main	1339	927	382	309	0.69	0.29	0.23
	Vignette	414	290	121	95	0.70	0.29	0.23
IT	Main	1376	1107	526	449	0.80	0.38	0.33
	Vignette	374	298	139	119	0.80	0.37	0.32
NL	Main	1563	1318	512	467	0.84	0.33	0.30
	Vignette	391	329	115	106	0.84	0.29	0.27
SE	Main	1850	1694	690	652	0.92	0.37	0.35
	Vignette	269	261	102	102	0.97	0.38	0.38
All	Main	7302	5980	2675	2383	0.82	0.37	0.33
	Vignette	1448	1178	477	422	0.81	0.33	0.29
	Total	8750	7158	3152	2805	0.82	0.36	0.32

structure of the imputed data. The validity of these *ex post* statistical adjustments relies crucially on the assumption that the missing data mechanism underlying unit and item nonresponse are MAR after conditioning on a suitable set of auxiliary variables. The set of auxiliary variables used to calibrate the weights consists of gender and age class of the sampled household member.[4] For imputations, a much larger set of auxiliary variables is used, as one can exploit the additional information collected in other parts of the interview.

3. STATISTICAL MODEL

Our statistical model is closely related to the extended sample selection model presented in Section 2.4 of DNV and further extended in Klein *et al.* (2010). Both models involve multiple selection mechanisms and allow for endogeneity, with the relationship between the endogenous variables specified as a triangular simultaneous equation system. The main difference between our model and the one in DNV is the nature of the sample selection mechanism. In DNV, all selection indicators are observed. In our case, individuals selected for interview first decide whether to participate in the survey and then, given participation, they decide whether to answer the questions on income and food expenditure. Thus item nonresponse can only be observed for those who are unit respondents. Another difference is that, to avoid curse of dimensionality problems, we restrict the way endogenous variables enter the model by making index function assumptions and by imposing a partially linear structure on the food Engel curve.

3.1. Model Specification

We assume that our data are a random sample from a model with four latent endogenous variables representing, respectively, the willingness to participate in the survey (Y_1^*), the willingness to answer the questions on income and food expenditure (Y_2^*), the logarithm of household income (Y_3^*) and the food share (Y_4^*). The first three latent variables obey linear models of the form

$$Y_j^* = \beta_j^\top X_j + U_j, \quad j = 1, 2, 3$$

[4] In Denmark and Italy, auxiliary variables also include a set of indicators for geographical area.

DOI: 10.1002/jae

where β_j is a vector of k_j unknown parameters, X_j is a vector of exogenous variables and U_j is a random error. The fourth latent variable obeys instead the partially linear model

$$Y_4^* = \beta_4^\top X_4 + g(Y_3^*) + U_4 \tag{1}$$

where β_4 is a vector of k_4 unknown parameters, X_4 is a vector of exogenous variables, g is an unknown smooth function and U_4 is a random error.

Equation (1) allows for flexible income effects and includes as special cases well-known parametric specifications of the Engel curve. For example, when the function g is linear we obtain the specification adopted in the AIDS, and when g is quadratic we obtain the specification adopted in the QUAIDS. Equation (l) could be further generalized to allow for more flexible household composition effects while retaining consistency with the integrability conditions of consumer theory. An example is the extended partially linear model of Blundell *et al*. (2003, 2007), where an index of the socio-demographic variables in X_4 enters the unknown function g.

The observed endogenous variables are related to the latent endogenous variables through different observational rules. The observed indicator of unit response, Y_1, is equal to one for those with a positive willingness to participate in the survey, and is equal to zero otherwise. Similarly, the observed indicator of item response on food share, Y_2, is equal to one for those with a positive willingness to answer the questions on income and food expenditure, and is equal to zero otherwise, but is only available for those with a positive willingness to participate in the survey, namely those with $Y_1 = 1$. Finally, one observes household income Y_3 and food share Y_4 only for those who are willing to participate in the survey and to answer the questions on income and food expenditure; that is, $Y_3 = Y_3^*$ and $Y_4 = Y_4^*$ whenever $Y_1 Y_2 = 1$. Selectivity and endogeneity operate through the correlations between the unobservable errors. In particular, for the Engel curve (1), selectivity due to unit and item nonresponse is captured by the correlation of U_4 with U_1 or U_2, while endogeneity is captured by the correlation of U_4 with U_3.

Our aim is to obtain consistent estimates of the parameters in the Engel curve (1) from the subsample of complete cases, namely those for which $Y_1 Y_2 = 1$. Because of the potential correlation between the unobservables in the model, we have[5]

$$\mathrm{E}(Y_3 | Y_1\ Y_2 = 1) = \mu_3 + h(\mu_1, \mu_2) \tag{2}$$

$$\mathrm{E}(Y_4 | Y_1\ Y_2 = 1,\ Y_3) = \mu_4 + g(Y_3) + l(\mu_1, \mu_2, U_3) \tag{3}$$

where $\mu_j = \beta_j^\top\ X_j,\ j = 1, 2, 3, 4$, and

$$h(\mu_1, \mu_2) = \mathrm{E}(U_3 | U_1 > -\mu_1,\ U_2 > -\mu_2)$$

$$l(\mu_1, \mu_2, U_3) = \mathrm{E}(U_4 | U_1 > -\mu_1,\ U_2 > -\mu_2,\ U_3)$$

The functions h and l are bias correction terms that account, respectively, for sample selection in the equation for household income, and sample selection and endogeneity in the Engel curve relationship. Ignoring these terms would lead to inconsistent estimates of the parameters of interest. Our approach to estimation is a simple generalization of the classical Heckman's two-step procedure (Heckman, 1979) and uses estimates of h and l as control functions to correct for both sample selection and endogeneity.

Although assumptions on the distribution of the error terms are not the main concern when estimating ordinary regression models, they play a crucial role when estimating models with

[5] Throughout this section, conditioning on the set of exogenous covariates is kept implicit to simplify notation.

DOI: 10.1002/jae

sample selection. Semiparametric approaches avoid imposing distributional assumptions by leaving the functions h and l unspecified. The two functions are treated as infinite-dimensional nuisance parameters and are estimated together with the other parameters of interest. Parametric approaches instead assume that h and l are known up to a finite-dimensional parameter vector. Although analytically tedious to derive, the parametric approach provides a useful benchmark for our semiparametric estimators. It also helps in the development of semiparametric specifications which nest the parametric ones and allow easy ways of testing the underlying distributional assumptions.

Our parametric specification assumes that the error vector $U = (U_1, U_2, U_3, U_4)$ follows a multivariate Gaussian distribution with zero mean and nondiagonal covariance matrix. Under this assumption, Poirier (1980) showed that

$$h(\mu_1, \mu_2) = \sigma_3 \ \rho_{13} \ h_1(\mu_1, \mu_2) + \sigma_3 \ \rho_{23} \ h_2(\mu_1, \mu_2) \tag{4}$$

where σ_3 is the standard deviation of U_3, ρ_{jk} is the correlation coefficient between U_j and U_k, and

$$h_j(\mu_1, \mu_2) = \left[\frac{\phi(\mu_j)}{\Phi_2(\mu_1, \mu_2; \rho_{12})}\right] \ \Phi \ \left(\frac{\mu_k - \rho_{12} \ \mu_j}{\sqrt{1 - \rho_{12}^2}}\right), \quad j = 1, 2, \quad k \neq j$$

with $\phi(\cdot)$ and $\Phi(\cdot)$ denoting the density and the distribution function of the $\mathcal{N}(0, 1)$ distribution, and $\Phi_2(\cdot, \cdot; \rho)$ denoting the distribution function of the bivariate Gaussian distribution with zero mean, unit variances and correlation coefficient ρ. Note that $h_j(\mu_1, \mu_2)$ reduces to the usual inverse Mill's ratio when $\rho_{12} = 0$. The form of the function l is slightly more complicated because of the correlation ρ_{34} between U_3 and U_4. One can show that[6]

$$l(\mu_1, \mu_2, U_3) = \sigma_{4|3} \ \rho_{14|3} \ l_1(\mu_1, \mu_2, U_3) + \sigma_{4|3} \ \rho_{24|3} \ l_2(\mu_1, \mu_2, U_3) + \frac{\sigma_4}{\sigma_3}\rho_{34} \ U_3 \tag{5}$$

where $\sigma_{j|k}$ is the standard deviation of the conditional distribution of U_j given U_k, $\rho_{jk|s}$ is the conditional correlation of U_j and U_k given U_s, and l_j has the same form as h_j with ρ_{12} replaced by $\rho_{12|3}$ and μ_j by $\mu_j^* = (\mu_j + \rho_{j3} \ U_3/\sigma_3)/\sigma_{3|j}$.

3.2. Identification

The sequential structure of the model facilitates its identification. It is sufficient that, at each stage, there exists at least one variable that does not affect the outcomes at later stages.

The index μ_1 is always identified up to location and scale, provided the conditions in Manski (1988) are satisfied. Identification of μ_2 in the item response equation requires that X_1 contains at least one variable that is not contained in X_2 (Lee, 1995). The same set of exclusion restrictions is needed for parametric identification of the model (Meng and Schmidt, 1985). As long as μ_1 and μ_2 are identified from the two selection equations, they can be used to identify μ_3 in the equation for household income. The unrestricted form of the bias correction term h requires that X_1 and X_2 each contains at least one variable that is not contained in X_3. As shown by DNV, this condition is sufficient to identify μ_3 up to an additive constant. This set of exclusion restrictions is not necessary for parametric identification of the model, as μ_3 could in principle be identified through the nonlinearity of h_1 and h_2 in (4). Because these terms may be linear over a wide range of their arguments, this way of achieving identification is not very appealing. Thus exclusion restrictions

[6] The proof is available from the authors on request.

DOI: 10.1002/jae

are also crucial in the parametric case. Finally, given μ_1, μ_2 and μ_3, we can identify μ_4 and the function g in the Engel curve up to an additive constant. The unrestricted form of the bias correction term l further requires that X_1, X_2 and X_3 each contains at least one variable that is excluded from X_4.

Our set of exclusion restrictions guarantees identification of all the model parameters except the intercepts. Those in the two selection equations can be absorbed into the unknown distribution functions of the error terms and are not separately identified. This means that some location restriction is needed on either the distributions of U_1 and U_2, or on the systematic part of the two selection equations. The first alternative is complicated when the errors are correlated, so we follow Melenberg and van Soest (1996) and set the intercepts in β_1 and β_2 to their parametric estimates. Under certain conditions, the intercepts in β_3 and β_4 can be identified through the concept of identification at infinity (Heckman, 1990) and estimated through generalizations of the approach developed by Andrews and Schafgans (1998). In our empirical application, we avoid the problem by focusing on the weighted average derivative of the function g. This is an important parameter to consider because it provides a measure of the average slope of the Engel curve.

3.3. Predictors and Exclusion Restrictions

A unique feature of SHARE is the detailed information it offers on the sampling frame, the survey agencies, and the fieldwork. By matching these three sources, the variables available to predict unit nonresponse include background characteristics of the household members selected for interview (years of age and gender), interviewers' characteristics (years of age, gender and years of education), and characteristics of the fieldwork. For households approached by more than one interviewer, we always use the information on the first interviewer. This helps avoid problems of endogeneity of interviewer-level variables that may arise because of the widespread fieldwork strategy of switching to more experienced interviewers when there are difficulties in making contact and gaining respondents' cooperation.

For responding households, we can relate item nonresponse to socio-demographic characteristics and measures of cognitive ability of the respondents, features of the data collection process and characteristics of the interviewers. Our set of socio-demographic characteristics includes age, gender, years of education and marital status of the household respondent, age of the partner, household size, number of children aged less than 18 years, and an indicator for living in small cities. Cognitive abilities of the household respondent are measured by the scores obtained in the mathematical, orientation in time, recall, and fluency tests carried out during the SHARE interview. We also include measures of the burden of the interview, namely indicators for proxy interviews, interviews conducted outside home, and cases where the household respondent often asked for clarification. Identifiability of the parameters in the equation for item nonresponse is achieved by assuming that the fieldwork variables (the dummies for the vignette sample and the Swedish supplementary sample, the measure of delay in the contact process, and the dummy for the presence of an answering machine) help explain household survey participation but not item nonresponse on food share.

The right-hand side variables in the reduced-form equation for household income consist only of socio-demographic characteristics and measures of cognitive ability of the respondents. As suggested by Fitzgerald *et al*. (1998) and Nicoletti and Peracchi (2005), interviewers' characteristics and features of the fieldwork and the interview process provide the required set of exclusion restrictions. Since these variables are external to the individuals under investigation and are not under their control, one may expect them to be irrelevant for household income and food share. On the other hand, results from several validation studies suggest that these variables are important predictors of both unit and item nonresponse.

DOI: 10.1002/jae

Finally, the right-hand-side variables in the Engel curve equation consist only of log household income and socio-demographic characteristics. In this case, identification is attained by excluding the measures of cognitive ability of the household respondent. This corresponds to the reasonable assumption that, after controlling for education of the household respondent, his/her cognitive abilities help predict household income but do not help predict the food share.

Given the high level of comparability of the SHARE data, we pool data from the various countries but include country dummies interacted with NUTS1 regional indicators to capture unobserved heterogeneity at the country and regional level. Pooling the data allows us to increase efficiency of estimation and helps reduce problems of collinearity due to the limited within-country variability of variables such as the characteristics of the fieldwork and the interviewers. Definitions and summary statistics of all relevant variables are presented in Table III. After dropping a few cases with missing data on the covariates, our estimation sample consists of 15,643 households, of which 8565 (55%) agreed to participate in the survey and 2777 (18%) provided the information needed to compute the food share.

3.4. Estimation

Estimation is carried out in three steps. In the first step, we estimate the parameters of the unit and the item nonresponse equations using the semi-nonparametric (SNP) approach proposed by Gallant and Nychka (1987).[7] The resulting estimator generalizes the conventional maximum likelihood (ML) estimator for the bivariate probit model with sample selection by relaxing the assumption that U_1 and U_2 are jointly Gaussian.

In the second step, we estimate equation (2) for household income accounting for sample selection due to unit and item nonresponse. In the Gaussian case, this step coincides with the second step of the procedure developed by Poirier (1980) and Ham (1982), with the bias correction term h specified as in (4). In the semiparametric case, h is instead approximated by a power series estimator. Following DNV, we use series estimators because of their relative simplicity and computational advantage, although kernel regression methods could alternatively be used (Robinson, 1988).

Finally, in the third step, we estimate the conditional mean (3) accounting for both sample selection due to nonresponse and endogeneity of household income. In this case, the parametric approach leads to a simple ordinary least squares (OLS) regression based on a known functional form for g and the specification (5) for l. Our semiparametric approach instead uses series estimators for both g and l. In addition to the fully parametric and semiparametric approaches, we also consider intermediate approaches in which only one of the two functions is estimated nonparametrically. More details on our three-step procedure are provided in the Appendix.

4. EMPIRICAL RESULTS

We present the results separately for each of the three steps of our estimation procedure.

4.1. First Step

Tables IV and V show the estimates of the parameters in the unit and item nonresponse equations. The two equations are estimated jointly by ML probit and by our SNP estimator. For the latter,

[7] An alternative approach is the semiparametric maximum likelihood (SML) estimator of Lee (1995). This estimator is more computationally demanding than the SNP estimator since kernel regression must be conducted at each step of the likelihood maximization process. Furthermore, the Monte Carlo evidence in De Luca (2008) suggests that the SNP estimator has better finite sample performance.

DOI: 10.1002/jae

Table III. Definitions and summary statistics for the main variables

Variable	Description	Symbol	Obs.	Mean	SD
particip	Dummy for household survey participation	Y_1	15,895	0.55	0.50
item resp	Dummy for item response on food share	Y_2	8750	0.32	0.47
ln income	PPP-adj0. household income	Y_3	3152	100.08	0.78
food share	Food share	Y_4	2805	0.23	0.20
age SHM	Age of SHM	X_1	15,895	650.18	100.73
female SHM	Dummy for female SHM	X_1	15,895	0.55	0.50
vignette	Dummy for vignette sample	X_1	15,895	0.16	0.37
supplement	Dummy for Swedish supplementary sample	X_1	15,895	0.06	0.23
ans0. machine	Dummy for presence of answering machine	X_1	15,895	0.02	0.15
delay	Measure of delay in the contact process	X_1	15,895	0.46	0.32
female IV	Dummy for female IV	X_1, X_2	15,894	0.74	0.44
age IV	Age of IV	X_1, X_2	15,894	490.26	110.81
educ IV	IV years of education	X_1, X_2	15,828	130.45	20.86
DK	Dummy for Denmark	X_1, X_2, X_3, X_4	15,895	0.11	0.31
ES	Dummy for Spain	X_1, X_2, X_3, X_4	15,895	0.21	0.41
ES nuts_1	Dummy for Spain–Region 1	X_1, X_2, X_3, X_4	15,895	0.02	0.13
ES nuts_2	Dummy for Spain–Region 2	X_1, X_2, X_3, X_4	15,895	0.02	0.14
ES nuts_3	Dummy for Spain–Region 3	X_1, X_2, X_3, X_4	15,895	0.03	0.17
ES nuts_4	Dummy for Spain–Region 4	X_1, X_2, X_3, X_4	15,895	0.03	0.16
ES nuts_6	Dummy for Spain–Region 6	X_1, X_2, X_3, X_4	15,895	0.05	0.21
ES nuts_7	Dummy for Spain–Region 7	X_1, X_2, X_3, X_4	15,895	0.01	0.11
IT	Dummy for Italy	X_1, X_2, X_3, X_4	15,895	0.20	0.40
IT nuts_2	Dummy for Italy–Region 2	X_1, X_2, X_3, X_4	15,895	0.04	0.20
IT nuts_3	Dummy for Italy–Region 3	X_1, X_2, X_3, X_4	15,895	0.04	0.20
IT nuts_4	Dummy for Italy–Region 4	X_1, X_2, X_3, X_4	15,895	0.04	0.20
IT nuts_5	Dummy for Italy–Region 5	X_1, X_2, X_3, X_4	15,895	0.02	0.14
NL	Dummy for Netherlands	X_1, X_2, X_3, X_4	15,895	0.20	0.40
NL nuts_1	Dummy for Netherlands–Region 1	X_1, X_2, X_3, X_4	15,895	0.03	0.17
NL nuts_2	Dummy for Netherlands–Region 2	X_1, X_2, X_3, X_4	15,895	0.03	0.17
NL nuts_4	Dummy for Netherlands–Region 4	X_1, X_2, X_3, X_4	15,895	0.04	0.20
proxy	Dummy for proxy interview	X_2	8750	0.11	0.31
clarif	Dummy for often asked clarifications	X_2	8750	0.08	0.28
outside	Dummy for interview outside home	X_2	8750	0.04	0.20
orient	HR score on orientation in time (1–5)	X_2, X_3	8703	30.76	0.67
math	HR score on math (1–5)	X_2, X_3	8690	30.21	10.18
recall	HR score on delayed recall (0–10)	X_2, X_3	8642	30.23	20.04
fluency	HR score on fluency (0–88)	X_2, X_3	8604	180.40	70.42
age	Age of HR	X_2, X_3, X_4	8750	640.90	100.49
female	Dummy for female HR	X_2, X_3, X_4	8750	0.55	0.50
education	HR years of education	X_2, X_3, X_4	8728	90.14	40.53
single	Dummy for HR living as single	X_2, X_3, X_4	8740	0.32	0.47
age spouse	Age of spouse/partner	X_2, X_3, X_4	8740	630.17	80.26
hsize	Household size	X_2, X_3, X_4	8750	20.16	10.05
children	Number of children	X_2, X_3, X_4	8743	0.09	0.37
small city	Dummy for household living in small city	X_2, X_3, X_4	8750	0.21	0.41

Note: SHM, sampled household member; HR, household respondent; IV, interviewer.

we considered four alternative specifications obtained by varying the degree K of the Hermite polynomial expansion (5) in the Appendix. For brevity, we present the results for the most parsimonious specification with $K = (3, 3)$, which is the one selected by BIC. In Section 4.4 we discuss the sensitivity of our three-step estimator to the choice of K.

The estimated coefficients from the probit and the SNP estimators are not directly comparable, because in the former the variances of U_1 and U_2 are normalized to one, while in the latter they are unconstrained functions of the Hermite polynomial parameters in γ. Thus we compare ratios of the estimated coefficients, dividing the coefficients in the equation for unit nonresponse by the coefficient on the variable that measures delay in the contact process, and the coefficients in the

DOI: 10.1002/jae

Table IV. First-step estimates of the unit response equation

Variable	Bivariate probit	Bivariate SNP
female SHM	−0.001	0.011
age SHM	−0.009**	−0.009**
female IV	0.085	0.300**
age IV	0.013**	0.010**
educ IV	−0.044**	−0.041**
vignette	0.090	0.038
supplement	−0.374**	−0.320**
ans. machine	0.227	0.226*
DK	1.129**	1.247**
ES	−0.020	0.003
ES $nuts_1$	0.578*	0.450*
ES $nuts_2$	1.080**	1.073**
ES $nuts_3$	−0.460*	−0.711*
ES $nuts_4$	1.039**	1.164**
ES $nuts_6$	0.984**	1.123**
ES $nuts_7$	0.043	0.011
IT	−0.064	0.018
IT $nuts_2$	0.387*	0.400**
IT $nuts_3$	0.723**	0.575**
IT $nuts_4$	0.928**	0.947**
IT $nuts_5$	0.473*	0.447*
NL	0.372**	0.473**
NL $nuts_1$	0.333*	0.455
NL $nuts_2$	0.555**	0.643**
NL $nuts_4$	0.590**	0.845**
T_1		18.08**
Skewness		0.16
Kurtosis		1.78**

Note: The models are estimated jointly with those presented in Table V. Asterisks denote a * p-value between 5% and 1%, and a ** p-value below 1%). Results are based on the normalization $|\beta_{\text{delay}}| = 1$. Standard errors of normalized coefficient are computed through the delta method. T_1 is the likelihood ratio statistic for testing Gaussianity and has 1 d.f. Sample size $n_1 = 15,643$.

equation for item nonresponse by the coefficient on the dummy for being single. The standard errors of these ratios are computed through the delta method.

Other things being equal, we find that the probability of survey participation falls with the age of the sampled household member and is significantly lower for households without an answering machine and for those belonging to the Swedish supplementary sample. Interviewers' characteristics are important predictors of survey participation. Being approached by a female interviewer significantly increases the probability of participation. We also find that participation is associated positively with the interviewer's age and negatively with the interviewer's years of education. The estimated coefficients on the country dummies and their interactions with the regional indicators further suggest an important role for unobservable regional differences in sample composition and fieldwork strategy. The assumption that the error in the unit nonresponse equation is Gaussian is rejected at the 1% level by a likelihood ratio test which compares univariate versions of the SNP and probit models, as in Gabler *et al*. (1993). According to our preferred SNP specification, the estimated error density exhibits significantly lower kurtosis than the standard normal distribution.

For item nonresponse on food share, we find significantly lower response probabilities for households living in small cities or with a female household respondent. Response probabilities are also negatively associated with the age of the household respondent and with the household size, and positively associated with the scores on the cognitive ability tests and the age of the

DOI: 10.1002/jae

Table V. First-step estimates of the equation for item response to food share

Variable	Bivariate probit	Bivariate SNP
female	−0.524**	−0.490**
age	−0.012	−0.014**
age spouse	0.030**	0.027**
hsize	−0.149**	−0.149*
children	0.050	0.057
small city	−0.584**	−0.648**
education	−0.022*	−0.024*
orient	0.207**	0.238**
math	0.007	0.008
recall	0.057*	0.036
fluency	0.005	0.007
female IV	−0.002	0.009
age IV	0.002	0.009*
educ IV	0.016	0.003
proxy	0.259*	0.305*
clarif	−0.181	−0.173
outside	−0.430*	−0.340
DK	−0.561	0.064
ES	0.194	0.375
ES nuts_1	−0.810	−0.494
ES nuts_2	−3.133**	−2.717**
ES nuts_3	−1.117*	−0.427
ES nuts_4	−2.693**	−2.276**
ES nuts_6	−1.832**	−1.481**
ES nuts_7	−0.333	−0.485
IT	−0.821**	−0.956**
IT nuts_2	0.568	1.045**
IT nuts_3	0.048	0.712*
IT nuts_4	1.160*	1.776**
IT nuts_5	1.063*	1.522**
NL	−1.042**	−0.758**
NL nuts_1	0.350	0.577*
NL nuts_2	−0.214	0.085
NL nuts_4	−0.072	0.085
T_2		3.38
Skewness		0.68**
Kurtosis		2.44
ρ_{12}	−0.86**	−0.36*

Note: The models are estimated jointly with those presented in Table IV. Asterisks denote a * p-value between 5% and 1%, and a ** p-value below 1%). Results are based on the normalization $|\beta_{\text{single}}| = 1$. Standard errors of normalized coefficient are computed through the delta method. T_2 is the likelihood ratio statistics for testing Gaussianity and has 1 d.f. Sample size $n_2 = 8565$.

partner. As for the characteristics of the interviewer and the interview process, we find that response probabilities are positively associated with the age of the interviewer and are higher for interviews with a proxy respondent. After controlling for our large set of covariates, we still find that unobserved heterogeneity at the regional level plays an important role. The assumption that the error in the item nonresponse equation is Gaussian is rejected at the 5% level for the specification with $K_2 = 4$, but not for the more parsimonious specification with $K_2 = 3$. In our preferred SNP specification, the estimated error density is characterized by significantly positive skewness.

The estimate of the correlation coefficient ρ_{12} between U_1 and U_2 is always negative and significantly different from zero. This is likely to reflect the effect of the missing spouse problem discussed in Section 2. However, point estimates are subject to sizable differences across specifications, ranging between −0.86 in the probit specification and −0.36 in our preferred SNP specification. These differences suggest that departures from Gaussianity push the estimated

DOI: 10.1002/jae

correlation coefficient toward the lower bound of its parameter space. This feature of the ML probit estimator becomes a real problem when estimating the model separately by country. On the other hand, the SNP estimates of ρ_{12} are always reasonably far from their lower bound.

4.2. Second Step

Table VI compares four alternative approaches to estimating the reduced-form equation for household income. The first two approaches correspond to the widespread practice of treating the missing data mechanism as MCAR or MAR. The column labeled Model 1 presents the results for a linear model estimated by OLS using only the complete cases. This model simply ignores unit and item nonresponse by implicitly assuming that the underlying missing data mechanism is MCAR. The column labeled Model 2 presents weighted OLS estimates of a linear regression model estimated on the completed data (complete cases plus imputations). In this case, the underlying missing data mechanism is assumed to be MAR after conditioning on the auxiliary variables employed in constructing the weights and the imputations. The columns labeled Model 3 and Model 4 present the estimates for two alternative specifications where the missing data mechanism is allowed to be NMAR. The two specifications differ in the assumptions about the

Table VI. Estimates of the reduced-form equation for log household income

Variable	Model 1	Model 2	Model 3	Model 4
female	−0.019	0.052*	0.128*	0.084
age	−0.008**	−0.005**	−0.005**	−0.006**
single	−0.473**	−0.521**	−0.698**	−0.626**
age spouse	−0.007**	0.009**	−0.014**	−0.012**
hsize	0.073**	0.148**	0.120**	0.108**
children	−0.027	−0.072*	−0.056	−0.054
small city	−0.064*	−0.011	0.128*	0.095
education	0.039**	0.044**	0.044**	0.042**
orient	−0.023	−0.021	−0.072*	−0.062*
math	0.059**	0.040**	0.056**	0.056**
recall	0.005	−0.004	−0.006	0.002
fluency	0.007**	0.005**	0.005*	0.005*
DK	−0.132**	−0.087**	0.002	−0.099
ES	−0.365**	−0.264**	−0.406**	−0.421**
ES nuts_1	−0.029	−0.113	0.176	0.086
ES nuts_2	0.115	0.055	0.922**	0.674*
ES nuts_3	−0.052	0.119	0.567	0.293
ES nuts_4	−0.125	−0.156*	0.654*	0.438
ES nuts_6	−0.104	−0.015	0.400*	0.258
ES nuts_7	−0.116	−0.220*	0.036	0.042
IT	−0.161*	−0.153**	0.106	0.063
IT nuts_2	−0.072	−0.119*	−0.282*	−0.295*
IT nuts_3	0.021	0.059	−0.068	−0.147
IT nuts_4	−0.228**	−0.252**	−0.616**	−0.591**
IT nuts_5	−0.174	−0.208**	−0.568**	−0.537**
NL	−0.059	0.055	0.211*	0.100
NL nuts_1	−0.019	−0.117*	−0.120	−0.113
NL nuts_2	0.002	−0.122*	0.006	−0.036
NL nuts_4	0.036	−0.025	0.024	0.035
h_1			0.766*	0.194
h_2			−1.099**	−0.494*
constant	0.359**	0.420**	0.351**	0.520**
n_3	2777	8565	2618	2618

Note: Standard errors are computed from 1000 nonparametric bootstrap replications.

DOI: 10.1002/jae

distribution of the unobservables. Model 3 corresponds to the parametric specification that uses the first-step estimates of the bivariate probit model with sample selection and the Gaussian bias correction terms in (4). Model 4 corresponds instead to the semiparametric specification that uses the first-step estimates of the SNP model with $K = (3, 3)$ and the power series expansion (8) in the Appendix (with $R = 1$) to approximate the unknown function h. We explored all the semiparametric specifications that can be obtained by combining powers and interactions of the leading terms h_1 and h_2 up to third order. Using leave-one-out cross-validation as model selection criterion, our preferred specification includes only the leading terms h_1 and h_2. Accordingly, we cannot reject Gaussianity of the conditional distribution of U_3 given U_1 and U_2.

Other things being equal, we find that household income falls with the age of the household respondents and his/her spouse, and is significantly lower if the household respondent is single. We also find that household income is positively associated with the size of the household and with the education and the cognitive abilities of the household respondent. As for country and regional differences, we find significantly lower income levels for Spain and the southern Italian regions.

In the Gaussian specification of our sample selection model, the selectivity effects of unit and item nonresponse are both significantly different from zero and have opposite sign (positive for unit nonresponse and negative for item nonresponse). A Wald test on the joint significance of the two bias correction terms rejects the null at the 5% level. In our preferred semiparametric specification, the selectivity effects of unit and item nonresponse are somewhat weaker, and only the term corresponding to item nonresponse is statistically significant.

4.3. Third Step

Table VII compares four alternative approaches to estimating the relationship of primary interest, namely the food Engel curve (3). The first two approaches correspond to the widespread practice of ignoring endogeneity of household income and treating the missing data mechanism as MCAR or MAR. The column labeled Model 1 presents the results for a partially linear model estimated using only the complete cases under the assumption that the missing data mechanism is MCAR. The column labeled Model 2 presents the results for a partially linear model estimated using the completed data and the survey weights under the assumption that the missing data mechanism is MAR. In both cases, the unknown function g is estimated by a power series estimator with leave-one-out cross-validation as model selection criterion.

The columns labeled Model 3 and Model 4 account for endogeneity of household income and allow the missing data mechanism to be NMAR, but differ in the assumptions about the distribution of the unobservables. Model 3 assumes a multivariate Gaussian distribution for the latent errors and uses a power series estimator for g. Model 4 uses instead the first-step estimates of the SNP model with $K = (3, 3)$, the semiparametric second-step estimates of the reduced form residuals, and power series estimates for both g and l. All power series estimators employ polynomial expansions up to fourth-order and leave-one-out cross-validation as model selection procedure. In Model 4, the additivity restriction on g and l is imposed by excluding interaction terms between log household income Y_3 and each of the leading terms l_1, l_2 and U_3.

Several features of the estimated models are worth noting. First, for all of them, leave-one-out cross-validation leads to the choice of a cubic approximation to the function g, thus rejecting traditional specifications of the food Engel curve, such as the linear and the quadratic adopted by AIDS and QUAIDS. Second, our parametric estimates show little evidence of selectivity or endogeneity, as none of the coefficients on the Gaussian bias correction terms is statistically significant. Third, our semiparametric estimates show instead evidence of selectivity

DOI: 10.1002/jae

Table VII. Estimates of the Engel curve for food share

Variable	Model 1	Model 2	Model 3	Model 4
ln income	−0.123**	−.208**	−0.106**	−0.081*
(ln income)^2	0.040**	0.046**	0.042**	0.044**
(ln income)^3	−0.015**	0.004	−0.018**	−0.013*
female	−0.009*	−0.011*	0.001	−0.004
age	−0.001**	−0.001**	−0.001	−0.000
single	−0.031**	−0.048**	−0.050*	−0.021
age spouse	−0.000	0.000	−0.000	0.000
hsize	0.038**	0.042**	0.038**	0.033**
children	−0.012	−0.009	−0.010	−0.008
small city	−0.019**	−0.018**	−0.003	−0.012
education	0.002**	0.004**	0.002	0.000
DK	−0.021**	−0.023**	0.005	−0.029
ES	.212**	0.166**	.218**	.233**
IT	0.172**	0.126**	0.194**	0.186**
NL	0.031**	0.050**	0.061**	0.035*
ES nuts$_1$	0.007	−0.001	0.033	0.010
ES nuts$_2$	0.003	0.003	0.083	0.004
ES nuts$_3$	0.022	0.015	0.057	0.006
ES nuts$_4$	−0.095**	−0.053**	−0.082	−0.148**
ES nuts$_6$	0.042	0.015	0.101**	0.048
ES nuts$_7$	0.000	0.023	0.009	0.008
IT nuts$_2$	−0.029	0.011	−0.036	−0.032
IT nuts$_3$	−0.043	−0.002	−0.037	−0.056*
IT nuts$_4$	0.013	0.005	0.001	0.003
IT nuts$_5$	−0.056	−0.009	−0.100**	−0.077
NL nuts$_1$	−0.005	−0.011	−0.012	−0.022
NL nuts$_2$	0.008	−0.007	0.021	−0.003
NL nuts$_4$	−0.005	−0.005	−0.004	−0.026
l_1			0.060	0.071
l_2			−0.089	−0.022
u_3			0.093	0.043
l_1^2				0.097*
$l_1 \times l_2$				−0.076
$l_1 \times u_3$				0.040
l_2^2				0.006
$l_2 \times u_3$				−0.097*
u_3^2				0.082**
$l_1^2 \times l_2$				−0.089
$l_1 \times l_2 \times u_3$				−0.056
u_3^3				−0.032*
constant	0.137**	0.155**	0.139**	0.132**
n_4	2724	8396	2591	2591
WAD	−0.16**	−0.20**	−0.14**	−0.11**

Note: WAD is the estimated weighted average derivative. Standard errors are computed using 1000 nonparametric bootstrap replications.

and endogeneity, as the higher-order terms are statistically significant. Thus we also reject Gaussianity of the conditional distribution of U_4 given U_1, U_2 and U_3.

Since interpretation of the polynomial coefficients is difficult, we investigate the implications of the various models for the shape and the average slope of the food Engel curve. Figure 1 plots the estimates of the Engel curve derivatives together with the 95% symmetric confidence bands for each model. Although the estimated derivatives are always negative, as predicted by the Engel law, their level and profile differ across models. In Model 1, the estimated Engel curve derivatives have a concave profile and the estimated weighted average derivative (WAD) is −0.16 with a bootstrap standard error of 0.005. Using instead the household weights and the imputations

DOI: 10.1002/jae

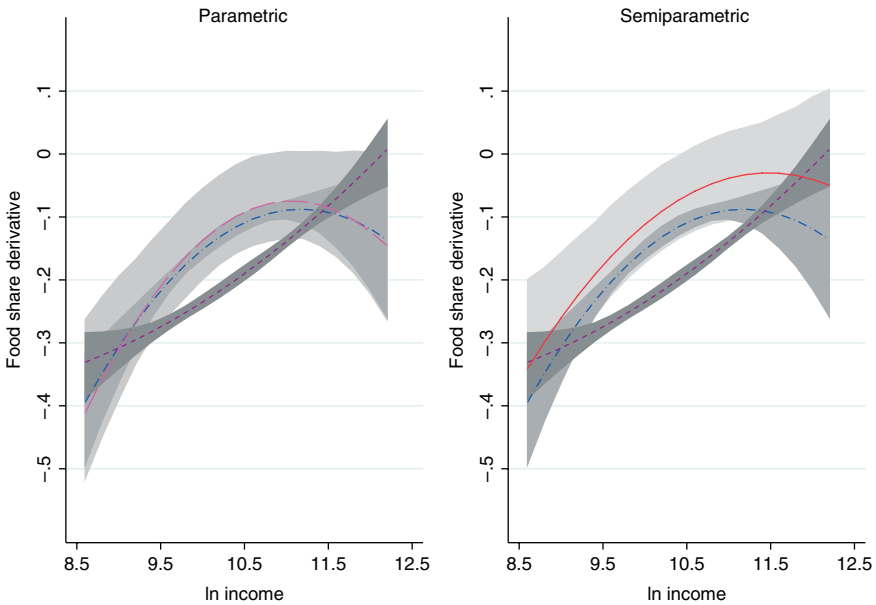

Figure 1. Estimated food share derivatives with 95% symmetric confidence bands. In each panel, the dash-dot line is from Model 1, the short-dash line from Model 2, the long-dash line from Model 3 and the solid line from Model 4. This figure is available in color online at wileyonlinelibrary.com/journal/jae

provided by SHARE (Model 2), the profile of the estimated Engel curve derivatives is now steeper and convex, and the estimated WAD is lower (-0.20 with a bootstrap standard error of 0.003). The estimated WAD of a similar model which uses the completed data but not the survey weights is -0.18 with a bootstrap standard error of 0.002. This suggests that imputations and survey weights both play some role for the nonresponse adjustments from Model 2. The fact that the standard errors for the estimates from Model 2 are low is due partly to the much larger sample size, as missing values are replaced by imputations, and partly to the fact that we ignore the additional uncertainty caused by imputation. The parametric estimates from Model 3 lead to a profile of the Engel curve derivatives which is not statistically different from that obtained from Model 1. In this case, the estimated WAD is -0.14 with a bootstrap standard error of 0.044. After relaxing the Gaussianity assumption, we find instead a much flatter profile of the Engel curve derivatives. Note that the estimates from Models 1 and 2 are downward biased because of endogeneity of household income. A semiparametric instrumental variable (IV) correction of Model 1 gives indeed an estimated WAD of -0.09 with a bootstrap standard error of 0.034. After correcting for both endogeneity of household income and selectivity of unit and item nonresponse, the estimated WAD from Model 4 is instead -0.11 with a bootstrap standard error of 0.036. Thus our semiparametric corrections for endogeneity and nonresponse have opposite sign and they partly offset each other. The estimates of our sample selection models are then closer to the estimates of Model 1, which completely ignores endogeneity and nonresponse, than to the estimates of Model 2, which ignores endogeneity and uses survey weights and imputations to correct for nonresponse.

4.4. Sensitivity Analysis

We now investigate the sensitivity of our estimates to three issues: the choice of the order of the polynomial expansions in the various steps of the estimation procedure, the implications of using

DOI: 10.1002/jae

a less conservative definition of item nonresponse on income, and the importance of accounting for unobserved country heterogeneity.

As for the first issue, Donald and Newey (1994) suggest that some undersmoothing may help reduce the bias of power series estimators. Thus Figure 2 looks at the effects of undersmoothing by comparing the Engel curve derivatives from the specifications selected by leave-one-out cross-validation with those from less parsimonious specifications based on a fourth-order polynomial expansion of g. For Model 4 only, we also use $K = (4, 4)$ for the SNP estimator of the two selection equations, $R = 2$ for the power series estimator of h, and $S = 3$ for the power series estimator of l. Overall, undersmoothing produces profiles of the Engel curve derivatives that are very similar to those from the specifications selected by leave-one-out cross-validation. The main differences occur for Model 4 where, at high level of income, undersmoothing leads to a steeper profile, thus providing stronger support for the selectivity and endogeneity effects discussed above. However, undersmoothing also leads to larger standard errors. Thus, at these income levels, Engel curve derivatives are not accurately estimated.

As for the second issue, household income is a generated variable obtained by aggregating various income components collected at the individual and the household level. So far, this variable has been regarded as missing if any of its components was missing. This is a very conservative definition of item nonresponse. Income from capital assets is heavily affected by item nonresponse (with item nonresponse rates ranging from 46% for dividend from stocks or shares, to 60% for interest from bank accounts) but, after imputation, income from capital assets represents a relatively unimportant fraction of household income (only 2%). Adopting a less conservative definition, which ignores missing values on this income source, the item nonresponse rate on income decreases from 64% to 45% and the number of complete cases increases from 2805 to 4180. The right-hand-side panel of Figure 3 shows that, with this less conservative definition, our

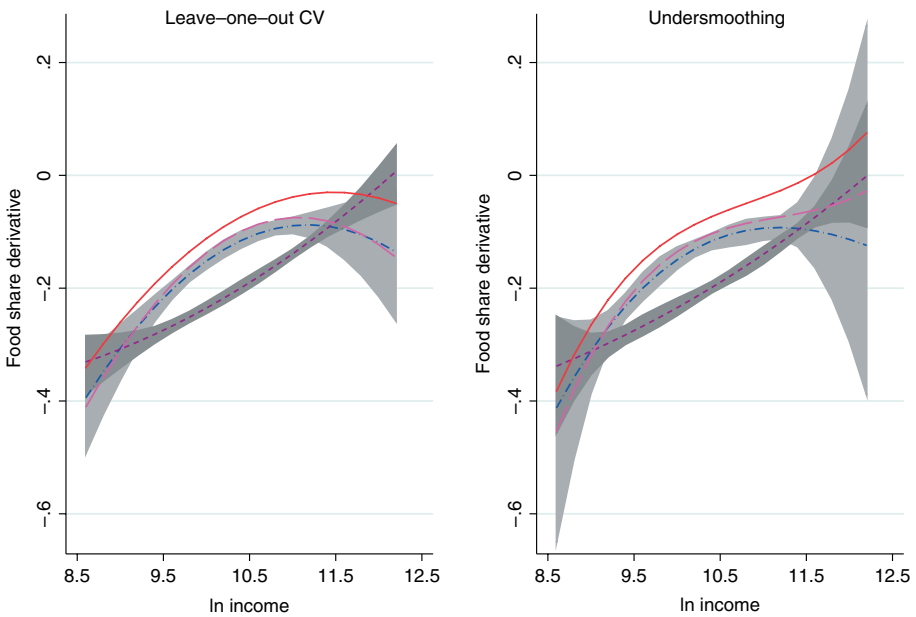

Figure 2. Estimated food share derivatives with 95% symmetric confidence bands under alternative degrees of the polynomial expansions. In each panel, the dash-dot line is from Model 1, the short-dash line from Model 2, the long-dash line from Model 3 and the solid line from Model 4. This figure is available in color online at wileyonlinelibrary.com/journal/jae

DOI: 10.1002/jae

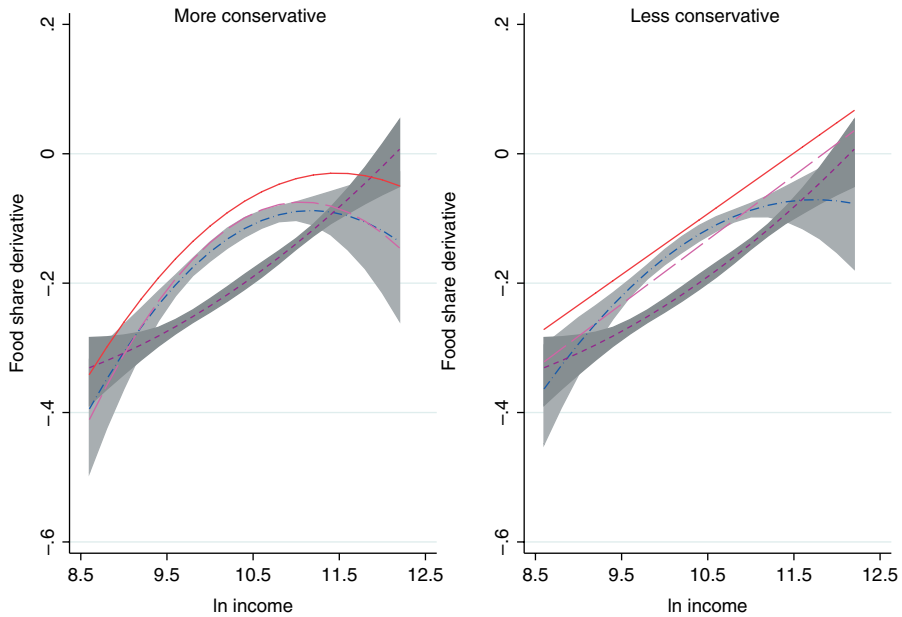

Figure 3. Estimated food share derivatives with 95% symmetric confidence bands under alternative definitions of item nonresponse to household income. In each panel, the dash-dot line is from Model 1, the short-dash line from Model 2, the long-dash line from Model 3 and the solid line from Model 4. This figure is available in color online at wileyonlinelibrary.com/journal/jae

semiparametric estimator leads to a linear profile of the Engel curve derivatives or, equivalently, to a quadratic specification of the food Engel curve. As before, evidence of selectivity and endogeneity effects is only found with the semiparametric specification of the model.

As for the third issue, Figure 4 provides some evidence on the importance of accounting for unobserved country heterogeneity by comparing pooled and country-specific estimates of the Engel curve derivatives. The latter have been obtained by estimating each model separately by country. We omit the fully parametric model (Model 3) because of convergence problems with the ML estimator employed in the first step. The profiles of the food Engel curves derivatives differ both across countries and estimation methods. In particular, our semiparametric approach produces more evidence of linearity than the standard approach (Models 1 and 2). Table VIII presents the estimated WAD and their bootstrap standard errors by model and country. Estimates are always negative but not very precise at the country level because of sample size problems. Interestingly, for all estimated models, the average slope of the Engel curve is steepest in Mediterranean countries (Italy and Spain).

5. CONCLUSIONS

In this paper we consider estimating Engel curves with data from the first wave of a panel survey affected by problems of unit and item nonresponse. Because the first wave of a panel is essentially a pure cross-section, the results that we obtain are valid more generally for cross-sectional data.

Our approach differs from traditional adjustment methods in many respects. First, we simultaneously address issues of selectivity due to nonresponse and issues of endogeneity in the structural relationship of interest, namely the Engel curve for food. Second, we treat the underlying missing data mechanism as NMAR. Third, we jointly model the two types of nonresponse. Fourth, we

DOI: 10.1002/jae

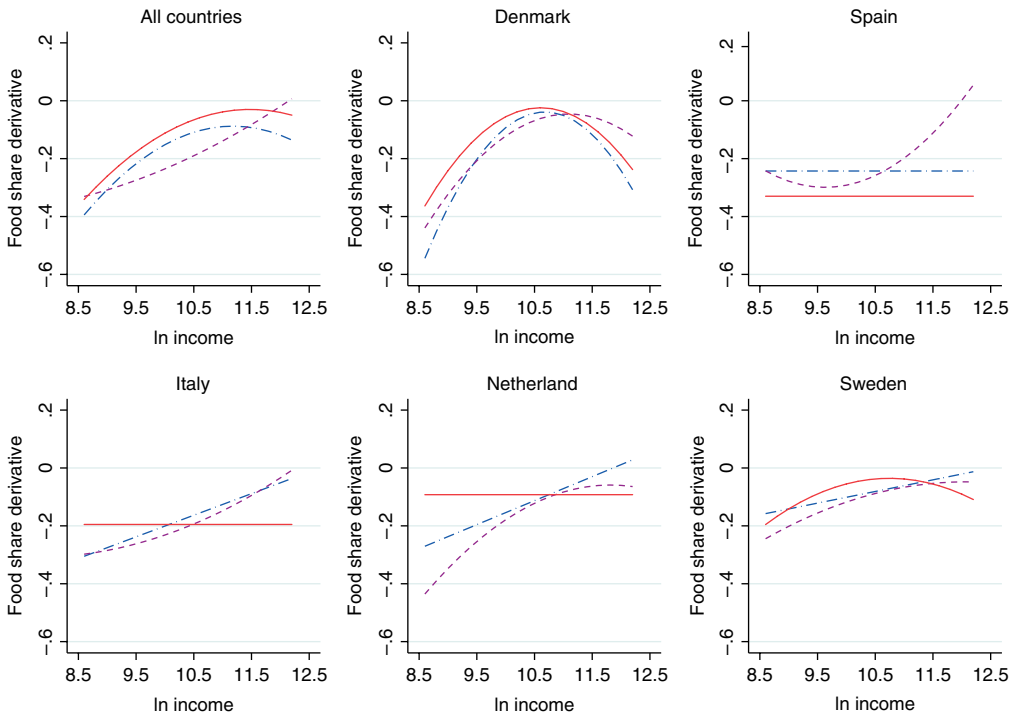

Figure 4. Estimated food share derivatives by country. In each panel, the dash-dot line is from Model 1, the short dash line from Model 2 and the solid line from Model 4. This figure is available in color online at wileyonlinelibrary.com/journal/jae

Table VIII. Estimated weighted average derivatives (WAD) and their bootstrap standard errors (SD) by model and country

Country	Model 1		Model 2		Model 4	
	WAD	SD	WAD	SD	WAD	SD
All	−0.16	0.005	−0.20	0.003	−0.11	0.036
DK	−0.11	0.014	−0.12	0.009	−0.07	0.059
ES	−0.24	0.016	−0.27	0.007	−0.33	0.146
IT	−0.22	0.014	−0.22	0.008	−0.20	0.137
NL	−0.13	0.009	−0.13	0.004	−0.09	0.068
SE	−0.09	0.007	−0.09	0.004	−0.05	0.026

allow unit and item nonresponse to be correlated. Since assumptions about the distribution of the unobservables play a key role when estimating sample selection models, we consider both parametric and semiparametric specifications of our model.

Our empirical results reject the assumption that nonresponse is MAR and therefore question the validity of traditional adjustment methods that rely on that assumption. We provide strong evidence of endogeneity of household income and of country heterogeneity in the shape of the food Engel curve. Our results also confirm the importance of avoiding strong parametric assumptions when estimating models with sample selection. Last, but not least, we illustrate the usefulness of supplementing survey data with information about fieldwork operations and interviewer characteristics. In our approach, this information provides the exclusion restrictions

DOI: 10.1002/jae

through which we identify a sample selection model with NMAR missing data mechanisms for unit and item nonresponse.

This information could also be useful in other contexts. An example is identification and estimation of treatment effects in observational or experimental studies with nonignorable nonresponse among the treatment and the control units. In this case, nonresponse may be viewed as a post-treatment complication that requires the availability of some instruments which, by definition, must be related to the missing data process but unrelated to the outcome of interest. If the instruments are interpreted as additional treatment variables, then they are also required to be randomized conditional on a vector of exogenous variables (Mealli and Pacini, 2008). This assumption is not easily satisfied if the candidate instruments consist of characteristics of the units that may be correlated with either the outcome of interest or the missingness indicator. Unlike other instruments, fieldwork operations and interviewer characteristics have the advantage that they can be controlled by the survey designer and so can easily be randomized. Another argument in favor of these instruments is the possibility of imposing credible monotonicity restrictions on the missing data process to achieve either partial or point identification of the causal effect of interest.

ACKNOWLEDGEMENTS

We thank Manuel Arellano, Joel Horowitz, Chuck Manski, Frank Vella, two anonymous referees and seminar participants at Northwestern and UCL for helpful comments. This paper uses data from release 2 of SHARE 2004. The SHARE data collection has been primarily funded by the European Commission through the 5th and 6th framework programme, with additional funding from the US National Institute on Ageing.

REFERENCES

Andrews DWK, Schafgans MMA. 1998. Semiparametric estimation of the intercept of a sample selection model. *Review of Economic Studies* **65**: 497–517.

Banks J, Blundell RW, Lewbel A. 1997. Quadratic Engel curves and consumer demand. *Review of Economics and Statistics* **79**: 527–539.

Blundell RW, Browning M, Crawford LA. 2003. Nonparametric Engel curves and revealed preferences. *Econometrica* **71**: 205–240.

Blundell R, Chen X, Kristensen D. 2007. Semi-nonparametric IV estimation of shape-invariant Engel curves. *Econometrica* **75**: 1613–1669.

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

Browning M, Madsen E. 2005. Consumption. In *Health, Ageing and Retirement in Europe: First Results from the Survey of Health, Ageing and Retirement in Europe*, Börsch-Supan A, Brugiavini A, Jürges H, Mackenbach J, Siegrist J, Weber G (eds). MEA: Mannheim; 318–324.

Browning M, Crossley TF, Weber G. 2003. Asking consumption questions in general purpose surveys. *Economic Journal* **113**: 540–567.

Buuren VS, Brand JPL, Groothuis-Oudshoorn CGM, Rubin DB. 2006. Fully conditional specification in multivariate imputation. *Journal of statistical Computation and Simulation* **76**: 1049–1064.

Coppejans M, Gallant AR. 2002. Cross-validated SNP density estimates. *Journal of Econometrics* **110**: 27–65.

Das M, Newey WK, Vella F. 2003. Nonparametric estimation of sample selection models. *Review of Economic Studies* **70**: 33–58.

De Luca G. 2008. SNP and SML estimation of univariate and bivariate binary choice models. *Stata Journal* **8**: 190–220.

Deaton AS, Muellbauer J. 1980. An almost ideal demand system. *American Economic Review* **70**: 312–336.

Deville JC, Särndal CE. 1992. Calibration estimators in survey sampling. *Journal of the American Statistical Association* **87**: 376–382.

DOI: 10.1002/jae

Donald SG, Newey WK. 1994. Series estimation of semilinear models. *Journal of Multivariate Analysis* **50**: 30–40.
Engel E. 1857. Die produktions und konsumptionsverhältnisse des Königreichs Sachsen. *Zeitschrift des Statistischen Bureaus des Königlich Sächsischen Ministeriums des Innerm* **8**: 1–54.
Fitzgerald J, Gottschalk P, Moffitt R. 1998. An analysis of sample attrition in panel data: The Michigan Panel Study of Income Dynamics. *Journal of Human Resources* **33**: 251–299.
Gabler S, Laisney F, Lechner M. 1993. Semiparametric estimation of binary-choice models with an application to labor-force participation. *Journal of Business and Economic Statistics* **11**: 61–80.
Gallant AR, Nychka DW. 1987. Semi-nonparametric maximum likelihood estimation. *Econometrica* **55**: 363–390.
Groves RM, Couper MP. 1998. *Nonresponse in Household Interview Surveys*. Wiley: New York.
Groves RM, Dillman DA, Eltinge JL, Little RJA. 2002. *Survey Nonresponse*. Wiley: New York.
Ham JC. 1982. Estimation of a labour supply model with censoring due to unemployment and underemployment. *Review of Economic Studies* **49**: 335–354.
Hausman JA, Newey WK, Ichimura H, Powell JL. 1991. Identification and estimation of polynomial error-in-variables models. *Journal of Econometrics* **50**: 273–295.
Heckman J. 1979. Sample selection bias as a specification error. *Econometrica* **47**: 153–161.
Heckman J. 1990. Varieties of selection bias. *American Economic Review* **80**: 313–318.
Heckman J, Navarro SL. 2004. Using matching, instrumental variables, and control functions to estimate economic choice models. *Review of Economics and Statistics* **86**: 30–57.
Imbens GW, Newey WK. 2009. Identification and estimation of triangular simultaneous equations models without additivity, *Econometrica* **77**: 1481–1512.
Jorgenson DW, Lau LJ, Stoker TM. 1982. The transcendental logarithmic model of aggregate consumer behavior. In *Advances in Econometrics*, Vol. 1, Basmann R, Rhodes G (eds). JAI Press: Greenwich, CT.
Klein R, Shen C, Vella F. 2010.. Triangular semiparametric models featuring two dependent endogenous binary outcomes. Working paper, Georgetown University.
Lee LF. 1995. Semiparametric maximum likelihood estimation of polychotomous and sequential choice models. *Journal of Econometrics* **65**: 381–428.
Lessler JT, Kalsbeek WD. 1992. *Nonresponse Errors in Surveys*. Wiley: New York.
Little RJA, Rubin DB. 2002. *Statistical Analysis with Missing Data* (2nd edn). Wiley: New York.
Manski CF. 1988. Identification of binary response models. *Journal of the American Statistical Association* **83**: 729–738.
Mealli F, Pacini B. 2008. Causal inference with nonignorably missing outcomes: Instrumental and principal stratification. Working paper, University of Florence.
Melenberg B, van Soest A. 1996. Measuring the costs of children: parametric and semiparametric estimators. *Statistica Neerlandica* **50**: 171–192.
Meng CL, Schmidt P. 1985. On the cost of partial observability in the bivariate probit model. *International Economic Review* **26**: 71–85.
Newey WK. 2001. Flexible simulated moment estimation of nonlinear error-in-variables models. *Review of Economics and Statistics* **83**: 616–627.
Nicoletti C, Peracchi F. 2005. Survey response and survey characteristics: microlevel evidence from the European Community Household Panel. *Journal of the Royal Statistical Society, Series A* **168**: 763–781.
Poirier D. 1980. Partial observability in bivariate probit models. *Journal of Econometrics* **12**: 209–217.
Riphahn RT, Serfling O. 2005. Item nonresponse on income and wealth questions. *Empirical Economics* **30**: 521–538.
Robinson PM. 1988. Root-N-consistent semiparametric regression. *Econometrica* **56**: 931–954.
Rubin DB. 1976. Inference and missing data. *Biometrika* **63**: 581–592.
Winter J. 2004. Response bias in survey-based measures of household consumption. *Economics Bulletin* **3**: 1–12.

DOI: 10.1002/jae

APPENDIX: DETAILS OF THE THREE-STEP ESTIMATION PROCEDURE

First Step

The log-likelihood function for a random sample of n observations is

$$L(\beta_1, \beta_2) = \sum_{i=1}^{n} [(1 - Y_{i1}) \ln \pi_{i0}(\beta_1) + Y_{i1}(1 - Y_{i2}) \ln \pi_{i10}(\beta_1, \beta_2) + Y_{i1} Y_{i2} \ln \pi_{i11}(\beta_1, \beta_2)] \quad (6)$$

where, dropping for simplicity the suffix i,

$$\begin{aligned}
\pi_0(\beta_1) &= \Pr\{Y_1 = 0\} = F_1(-\mu_1), \\
\pi_{10}(\beta_1, \beta_2) &= \Pr\{Y_1 = 1,\ Y_2 = 0\} = F_2(-\mu_2) - F(-\mu_1,\ -\mu_2), \\
\pi_{11}(\beta_1, \beta_2) &= \Pr\{Y_1 = 1,\ Y_2 = 1\} = 1 - F_1(-\mu_1) - F_2(-\mu_2) + F(-\mu_1,\ -\mu_2)
\end{aligned}$$

with F_1, F_2 and F denoting, respectively, the unknown marginal distribution functions of U_1 and U_2 and their joint distribution function.

Following Gallant and Nychka (1987), we approximate the joint density f of the latent errors by an Hermite polynomial expansion of the form

$$f^*(u_1, u_2; \gamma) = \frac{1}{\psi_K(\gamma)} \tau_K(u_1, u_2;\ \gamma)^2 \phi(u_1)\phi(u_2) \quad (7)$$

where $\tau_K(u_1, u_2;\ \gamma)$ is a polynomial of order $K = (K_1, K_2)$ in u_1 and u_2, γ is a vector of $K_1 K_2$ unknown parameters, and $\psi_K(\gamma)$ is a normalization factor which ensures that f^* is a proper density. This polynomial expansion can approximate densities with arbitrary skewness and kurtosis, but not violently oscillatory densities or densities with tails that are either too fat or too thin (Gallant and Nychka, 1987). De Luca (2008) shows that, after imposing some identifiability restrictions, integrating the joint density (7) gives the following approximation to the joint distribution function of U_1 and U_2:

$$\begin{aligned}
F^*(u_1, u_2;\ \gamma) = \Phi(u_1)\Phi(u_2) &+ \frac{1}{\psi_K(\gamma)} A_{12}^*(u_1, u_2;\ \gamma)\phi(u_1)\phi(u_2) \\
&- \frac{1}{\psi_K(\gamma)} A_1^*(u_1;\ \gamma)\Phi(u_2)\phi(u_1) - \frac{1}{\psi_K(\gamma)} A_2^*(u_2;\ \gamma)\Phi(u_1)\phi(u_2)
\end{aligned}$$

where $A_{12}^*(u_1, u_2;\ \gamma)$, $A_1^*(u_1;\ \gamma)$ and $A_2^*(u_2;\ \gamma)$ are polynomials in u_1 and u_2. Integrating $F^*(u_1, u_2;\ \gamma)$ one obtains the following approximations to the marginal distribution functions of U_1 and U_2:

$$\begin{aligned}
F_1^*(u_1;\ \gamma) &= \Phi(u_1) - \frac{1}{\psi_K(\gamma)} A_1^*(u_1;\ \gamma)\phi(u_1), \\
F_2^*(u_2;\ \gamma) &= \Phi(u_2) - \frac{1}{\psi_K(\gamma)} A_2^*(u_2;\ \gamma)\phi(u_2)
\end{aligned}$$

The SNP estimator of $(\beta_1, \beta_2, \gamma)$ is obtained by maximizing the pseudo log-likelihood function (6), with the unknown distribution functions replaced by their approximations F^*, F_1^* and F_2^*. Gallant and Nychka (1987) show that the resulting estimator is $\sqrt{n}$-consistent provided that the

DOI: 10.1002/jae

degree K of the polynomial increases with the sample size, but do not provide distributional results. However, if K is treated as known, inference can be conducted as though the model was estimated parametrically. Thus the SNP model is better viewed as a flexible parametric specification for a fixed value of K, with the choice of K as part of the model selection procedure. 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 such as AIC, BIC, or the cross-validation strategies proposed by Coppejans and Gallant (2002).

Second Step

In the second step we estimate (2) using the subsample of complete cases, with the unknown function h approximated by a power series expansion. As argued by DNV, series estimators have lower bias when their leading terms provide a good approximation. Thus, instead of expanding in power series of μ_1 and μ_2, we expand in power series of functions of μ_1 and μ_2 with the Gaussian bias correction (4) as leading term. The proposed approximation to $h(\mu_1, \mu_2)$ is of the form

$$h^*(\mu_1, \mu_2) = \sum_{r=0}^{R} \sum_{s=0}^{R-r} \theta_{rs}\, h_1(\mu_1,\ \mu_2)^r\, h_2(\mu_1,\ \mu_2)^s \tag{8}$$

where R is the degree of the power expansion, θ_{00} is normalized to zero, and the leading terms h_1 and h_2 are exactly equal to the elements of the Gaussian bias correction term (4), while the higher-order terms capture departures from normality. A test of the Gaussian assumption can then be obtained by testing whether these higher-order terms are significantly different from zero.

After replacing μ_1 and μ_2 by their SNP estimates $\hat{\mu}_1$ and $\hat{\mu}_2$, the second step corresponds to a simple OLS regression of Y_3 on X_3 and powers of $\hat{h}_1 = h_1(\hat{\mu}_1,\ \hat{\mu}_2)$ and $\hat{h}_2 = h_2(\hat{\mu}_1,\ \hat{\mu}_2)$ plus their interactions.[8] Under regularity conditions, the resulting estimator of β_3 is consistent and asymptotically normal provided that the degree R of the power series expansion increases with the sample size (DNV). For a given sample size, R can be selected by leave-one-out cross-validation. Note that the regularity conditions for our series estimator require some trimming of the data to guarantee that the estimated indexes from the first step are finite. Accordingly, we symmetrically trim 1% of the complete cases based on the values of $\hat{\mu}_1$ and $\hat{\mu}_2$.

Third Step

In the third step we estimate (3) from the subsample of complete cases, with the unknown functions g approximated by a power series in the logarithm of household income and the function l by a power series in functions of μ_1, μ_2 and $U_3 = Y_3^* - \mu_3$. In the latter case, the leading terms of the power series correspond to the correction for endogeneity and sample selection in the Gaussian case. Thus

$$l^*(\mu_1, \mu_2, U_3) = \sum_{r=0}^{S} \sum_{s=0}^{S-r} \sum_{t=0}^{S-r-s} \delta_{rst} l_1(\mu_1,\ \mu_2,\ U_3)^r l_2(\mu_1,\ \mu_2, U_3)^s U_3^t \tag{9}$$

where S is the degree of the power expansion and δ_{000} is normalized to zero. The leading terms l_1, l_2 and U_3 in (9) are exactly equal to the elements of the Gaussian bias correction term (5), while

[8] In constructing the $\hat{h}_j$, the correlation coefficient ρ_{12} is estimated by combining the SNP estimates of the first- and the second-order moments of U_1 and U_2 (De Luca 2008).

DOI: 10.1002/jae

the higher-order terms capture departures from normality. A test of the Gaussian assumption can then be obtained by testing whether these higher-order terms are significantly different from zero.

After replacing μ_1 and μ_2 by their SNP estimates $\hat{\mu}_1$ and $\hat{\mu}_2$ and U_3 by $\hat{U}_3 = Y_3 - \hat{\mu}_3$, the third step corresponds to a simple OLS regression of Y_4 on X_4, powers of log household income, and powers of $\hat{l}_1 = l_1(\hat{\mu}_1,\ \hat{\mu}_2,\ \hat{U}_3)$, $\hat{l}_2 = l_2(\hat{\mu}_1,\ \hat{\mu}_2,\ \hat{U}_3)$ and $\hat{U}_3$ plus their interactions.[9] The higher-order terms included in the power series approximations for g and l are again selected by cross-validation. Again, some trimming of the data is needed to limit the impact of outliers. Thus we symmetrically trim another 1% of the observations on the basis of the values of Y_3 and $\hat{U}_3$. The standard errors of our three-step estimator are computed by the nonparametric bootstrap based on 1000 replications.[10]

[9] In constructing the $\hat{l}_j$, the standard deviation σ_3 is estimated using the procedure proposed by Ham (1982), while the correlation coefficient $\rho_{12|3}$ is estimated by combining the estimates of ρ_{12}, ρ_{13} and ρ_{23} obtained from the first and the second step.

[10] For each replication, we sample with replacement from the original data and re-estimate the overall process (first, second and third step). This approach is time consuming, especially because of the SNP estimator used in the first step. To speed up the process, we use a MATA version of the bivariate SNP routine written in STATA by De Luca (2008).

DOI: 10.1002/jae