UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Scienze Statistiche
SCUOLA DI DOTTORATO DI RICERCA IN SCIENZE STATISTICHE
CICLO XXIII

MEASUREMENT ERROR ISSUES IN CONSUMPTION DATA

Direttore della Scuola: Prof. ALESSANDRA SALVAN

Supervisore: Prof. ERICH BATTISTIN

Dottorando: MICHELE DE NADAI

31 Gennaio 2011

Acknowledgements

First and foremost I would like to thank my family for their invaluable, monetary but most importantly moral, support during these years in Padova. I would have never had the possibility to write this thesis without them so I would like to dedicate this work to them.

I will always be immensely grateful to my supervisor Professor Erich Battistin, without whose help and encouragements I would have never undertaken a Ph.d. program in Statistics. The joy and enthusiasm he has for research have been contagious and motivational for me, even during tough times.

This thesis was partly developed during my visiting period at Boston College. I owe my deepest gratitude to Professor Arthur Lewbel for his comments and suggestions and for all of the useful discussions we had that greatly contributed to clarify my mind.

I am indebted to my friends and colleagues Slavica and Nicola for their, sometimes too encouraging, support. It has been a pleasure to share this great experience with them and I will always remember the great fun we had during these years. This thesis could also not have been possible without the countless coffee breaks with Bustra, Giovanna and Fany and the endless nights with Sandro, Nava, Mattia and Jackie who also contributed to giving me a place to stay during these last two years of mobility between Padova and Boston. Lastly I would like to thank all of the other Ph.d. students and friends I had the honor to meet during my stay both at the University of Padova and Boston College, to all of you goes my deepest gratitude.

Abstract

Data available in commonly employed consumer surveys, like the *Consumer Expenditure Survey* in the US, is widely known to be affected by measurement errors. Ignoring the effect of such errors in the estimation of consumption models may result in severely biased estimates of the quantities of interest. In this thesis I consider identification of three different models of consumption behavior allowing for the presence of measurement errors.

Identification is particularly difficult to achieve due to the high non-linearity of the specifications involved and to peculiarities of consumptions models. In fact in many instances, allowing for mismeasured covariates also implies correlated measurement errors also in the dependent variable. This further complicates the identification of the model, invalidating most of the non-linear errors in variables results in the literature.

The core of the thesis is made of three Chapters. In the first Chapter I consider identification of a particular specification of Engel curves when unobserved expenditure is endogenous and measured with error. In the second Chapter I study identification of a general non-linear errors in variables model allowing for correlated measurement errors on both sides of the equation. In the third Chapter I derive identification and estimation of the distribution of consumption when only expenditure and the number purchases are observed.

Sommario

I dati disponibili nelle piú comuni indagini sui consumatori, come la *Consumer Expenditure Survey* negli Stati Uniti, sono noti per essere affetti da errori di misura. Ignorare l'effetto di questi errori nella stima di modelli di consumo puó portare a stime distorte delle quantitá di interesse. Questa tesi discute l'identificazione di tre differenti modelli di comportamento dei consumatori in presenza di errori di misura.

L'identificazione risulta particolarmente difficile a causa della elevata non-linearitá delle specificazioni utilizzate e di alcune peculiaritá proprie dei modelli con dati di consumo. In molti casi infatti, la presenza di covariate misurate con errore implica errori di misura correlati nella variabile dipendente. Questo complica ulteriormente lídentificazione del modello, invalidando la maggior parte dei risultati presenti nella letteratura su errori di misura in modelli non-lineari.

Il contenuto della tesi é discusso in tre Capitoli. Il primo Capitolo discute l'identificazione di una particolare specificazione di curve di Engel quando la spesa totale non osservata é endogena e misurata con errore. Il secondo Capitolo studia l'identificazione di un modello non-lineare molto generale con errori di misura correlati su entrambi i lati dell'equazione. Il terzo Capitolo ottiene identificazione e stima della distribuzione di consumo quando solo la spesa e il numero di acquisti sono osservati.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

Consumption data are an important source of information for investigating consumer behavior. They allow the estimation of a variety of models useful for testing the restrictions imposed by the economic theory, making welfare comparisons or constructing counterfactual scenarios which are relevant to answer policy relevant questions.

An important branch of the analysis of consumer behavior is the specification and estimation of demand functions. They are defined as the relationship between consumption on a single good and total consumption, prices, and demographics. Examples of interest in this relationship date back at least to Stone (1954) who first fit this kind of models. Since his pioneering work a lot of specifications have been proposed for demand functions, either based on empirical findings or derived from restrictions imposed by the assumption of utility maximizing consumers. A worth mentioning example notable example of a popular parametric specification is the Almost Ideal Demand System (AIDS) of Deaton and Muellbauer (1980), though recently a growing part of the literature is focusing on more flexible semi-parametric estimation procedures to tackle more general specifications (see for instance Blundell, Chen, and Kristensen 2007).

The relevant information for the estimation of demand functions is usually collected by means of either recall data or diaries. In the former case, households are asked to report what was their level of expenditure for some commodities over a relatively long time span, while in the latter case respondents are given a diary to fill in with their purchases that tipically covers a very limited period of time (a one or two week period).

Both these sources of information are widely known to be affected by measurement errors (see Bound, Brown, and Mathiowetz 2001). When asked about their level of expenditure over a long period of time, household may be unable to remember the exact amount or they may willfully misreport

for a variety of reasons. On the other hand in diary data some commodities, like durables, are infrequently recorded, due to the limited time-span of the survey.

The first part of the thesis deals with the identification and estimation of Engel curves, which are defined as demand functions when prices are held constant. It is common practice to ignore the presence of measurement errors when estimating Engel curves. However, it is well understood that failing to account for their presence, even in a simple linear setting, would in general result in biased estimates of the parameters of interest.

A variety of methods have been proposed to identify and consistently estimate parameters of general non-linear errors in variables models, though they may not always be readily applied in practice. The proposed models are estimated by either making assumptions about the distribution of measurement error, using validation data to estimate features of its distribution or exploiting the availability of repeated measurements or instrumental variables (IVs). In this thesis I will follow this last stream of the literature, and will provide identification results using additional information provided by IVs. This is motivated by the general availability in econometrics of IVs as opposed to repeated measurements and by the purpose of avoiding any assumption regarding the distribution of measurement errors.

The estimation of Engel curves presents challenges which are not usually encountered in the measurement error literature. In particular they exhibit non-linearities and, more importantly, present correlated measurement errors on both the dependent and independent variables, hence invalidating the usual approaches. The only estimator proposed in the literature which explicitly takes into account these difficulties is the one proposed by Lewbel (1996). However he restricts his attention to one particular framework in which budget shares are polynomials in the logarithms of total expenditure and unobserved total expenditure may be assumed exogenous.

While this framework is one that has received a lot of attention in empirical work, recent advancements in demand analysis (see for instance Banks, Blundell, and Lewbel 1997, Lewbel and Pendakur 2009, Blundell, Chen, and Kristensen 2007) suggest that the estimation of more realistic models would require extensions of the available identification results to more flexible specifications of Engel curves and to the endogeneity of unobserved total expenditure. The main objective of chapters 3 and 4 is to derive identification and estimation results in these settings.

It is widely accepted that household consumption is a far better indicator of households well-being than income. This is because, while income is highly affected, at least in the short term, by exogenous shocks, consumption, being chosen by individuals, is thought of as a long term households' own assessment of their economic position, and hence is a much more reliable indicator of households well-being. However, as reported above, data usually record informations about household expenditures. Consumption

is in fact an unobserved variable and for this reason expenditure is widely accepted to be a good approximation for consumption and is then treated accordingly. Nonetheless there may be situations in which expenditure does not capture all the features of consumption we are interested in.

The discrepancies between consumption and expenditure are mainly due to the presence of storage costs and indivisibility of commodities. Consumption is, in principle, a continuous process while expenditure is a discrete process which occurs on a certain number of purchase occasions. If commodities were infinitely divisible and storage costs were particularly high, households would consume immediately after purchasing and expenditure would be equal to consumption. In reality storage costs are not very high and most of the commodities are indivisible by nature, hence the difference between expenditure and consumption may be substantial.

While it would be interesting to analyze household inequality in terms of consumption inequality measures are routinely computed from expenditure data. This problem is particularly severe when using data from diary surveys since, for infrequently purchased goods, recorded expenditure may either overstate or understate the underlying consumption. The objective of Chapter 5 is then that of providing an estimator of consumption from knowledge of expenditure data.

The remainder of this thesis is organized as follows. Chapter 2 serves as a review, while Chapters 3, 4 and 5 offer the main contributions. In particular, Chapter 3 proposes a method for parametrically identifying Engel curves when unobserved total expenditure is allowed to be endogenous. This applies to Engel curves in which budget shares are polynomials in the logarithms of total expenditure. Chapter 4 provides non-parametric identification for non-linear specifications of more general forms. I also provide conditions for the identification of the conditional distribution of expenditure for single goods on total expenditure. Chapter 5 derives an estimator for features of the unobserved distribution of consumption when only expenditure is observed. This is done by specifically modeling the purchase pattern of households, hence providing a way to estimate consumption of infrequently purchased goods from diary data. Simulation studies to compare the performance of the proposed estimators are also provided.

1.2 Main Contributions of the Thesis

The core of the thesis is represented by Chapters 3, 4 and 5.

In Chapter 3 I derive an estimator for a particular specification of Engel curves in the presence of measurement errors in total expenditures. I consider Engel curves where budget shares are polynomials in the logarithm of total expenditure. I first derive the expression for the asymptotic bias of the simple IV estimator, which has already been shown to be inconsistent

in this specification. The result shows that, when unobserved total expenditure is allowed to be jointly determined with expenditure on single goods, and so is endogenous in the model, the parameters of interest are identified and easily estimated by employing a control function approach. The estimator takes the form of a *Generalized Method of Moments* (GMM) estimator with linear in the parameters moment conditions and is therefore readily computed by standard statistical softwares. In order to evaluate the performances of the proposed estimator a simulation study has been conducted. The results show improvements over alternative estimators especially when the extent of both measurement error and endogeneity of the unobserved total expenditure is particularly severe.

A second identification result I provide is reported in Chapter 4. Here unobserved total expenditure is treated as exogenous, and the focus is instead on the specification of Engel curves. I provide identification for a very general error-in-variables model in which the conditional mean is very flexibly specified and correlated measurement errors are present on both the dependent and the independent variable. This result extends the previous result by Schennach (2007) to allow for correlated measurement error in the dependent variable. In the special case of Engel curves an additional result is obtained, namely the identification of the conditional distribution of unobserved expenditure on one good and unobserved total expenditure. This is done by exploiting the particular dependence structure between measurement errors entailed by the very nature of the variables. This result is of particular interest since it allows to separate the variability on the expenditure on a single good due to measurement error from that due to heterogeneity in preferences. This is an issue of major concern in the empirical literature on the topic. The identification result is applied to the estimation of generally non-linear Engel curves in a parametric setting. The estimator I propose is based on the method of *Simulated Moments* (SM), already adopted in the literature by Newey (2001).

A third contribution is presented in Chapter 5 which deals with the estimation of the distribution of consumption from expenditure data. The estimator I propose is based on a generalization of the models developed by Kay, Keen, and Morris (1984) and Meghir and Robin (1992). This is done by modeling the purchasing process of households, which in turn depends on the unobserved frequency of purchase of each indivudual. The estimator builds on the assumption of utility functions which are separable in allocation of expenditure and frequency of purchase. Under this assumption the distribution of consumption, including the proportion of non-consumers, is identified and readily estimated. The above argument is then applied to estimate the distribution of consumption for commodities using diary data from the *Consumer's Expenditure Survey*.

Chapter 2

Literature Review

The objective of this Chapter is that of providing the relevant background literature on the topics that will be covered in the remainder of the thesis. In the following Sections I will discuss previous findings about specification of Engel curves, non-linear errors in variables models and frequency of purchases issues. I will not provide an exhaustive list of works on these topics, but I will only selectively review the literature that is useful to have a better understanding of what will be discussed later in the thesis.

2.1 Engel Curves Estimation

Very few topics in economics do not involve knowledge of consumer behavior. An important branch of this is the so called demand analysis, which is mainly concerned with the specification and estimation of demand functions or, more generally, demand systems. These are defined as the structural relationship between consumption on single goods, on the one hand, and total consumption, prices and demographics on the other hand. Knowledge of these functions is of central interest in economics since they allow the evaluation of the impact of different tax policies, welfare analyses or the construction of equivalence scales (see Blundell 1988 and Lewbel 2010 for a review).

Engel curves are defined as demand functions in which prices are held constant. These functions have received particular attention in the literature since they may be seen as building blocks in the construction of more general demand systems, and because they are easier to estimate since they do not require variability in prices. Engel curves are also of interest in their own for welfare analysis and the characterization of goods as inferior, necessities or luxuries.

The estimation of Engel curves dates back at least to Engel (1895), who first studied the relationship between household food expenditure and income. The interest there was merely that of comparing different specifica-

tions on the basis of the goodness of fit. Since then, several works followed along the same lines, see for example Ogburn (1919), Working (1943), Leser (1963) and Prais and Houthakker (1971). Among these, special attention has received the so called Working-Leser specification, in which budget shares, which is the proportion of total expenditure allocated to each good, are linear in the logarithm of total expenditure. This is still one of the most used Engel curves specifications, since it seems to fit reasonably well for a wide variety of goods.

However, when the objective of the analysis is that of performing welfare analyses a theory consistent demand system needs to be specified. This leads to a more structural approach, in which demand functions, and hence Engel curves, are specified on the basis of the restrictions imposed on consumer preferences and of the assumption of utility maximizing individuals. Such approach gave rise to several "theory consistent" specifications in which the model estimated has a *structural*, as opposed to *statistical* interpretation (see the discussion in Lewbel 2001).

To clarify this point, suppose individuals possess an indirect utility function $V(x, \mathbf{p})$, where x is logarithm of total expenditure and $\mathbf{p}$ is a vector of prices. Standard consumer theory states that the budget share for the i-th good (w_i) of a utility maximizing individual is defined by the Roy's identity as:

$$w_i(x, \mathbf{P}) = -\frac{\partial V(x, \mathbf{p})/\partial \log p_i}{\partial V(x, \mathbf{p})/\partial x}.$$

Thus different models arise depending on the specification of consumer preferences which are summarized by the indirect utility function $V(x, \mathbf{p})$. On the other hand there is no guarantee that for any functional form specification there exists a function $V(x, \mathbf{p})$ which satisfies the usual properties of an indirect utility function.

A very general class of indirect utility functions which have received great attention in the literature is the *translog* family. Several routinely employed demand system specifications, like the *linear expenditure system* or the *quadratic expenditure system*, where expenditure on one good is either a linear or a quadratic function of income, may be derived from preferences which belong to this general class (see Gorman 1961 and Howe, Pollack, and Wales 1979). A notable example in this class is the log translog indirect utility function, which is given by the very flexible functional form:

$$V(x, \mathbf{p}) = -\sum_{i=1}^{I} \alpha_i (\log(p_i/e^x) - x) - \frac{1}{2}\sum_{j=1}^{I}\sum_{i=1}^{I} \beta_{ij} \log(p_j/e^x) \log(p_i/e^x), \tag{2.1}$$

where $\beta_{ij} = \beta_{ji}$, $\sum_{i=1}^{I}\sum_{j=1}^{I} \beta_{ij} = 0$ and $\sum_{i=1}^{I} \alpha_i = 1$. This form of indirect utility function yields demand functions in which budget shares are linear in x, hence belonging to the *Price Independent Generalized Logarithmic*

(PIGLOG) class of demand functions. This is not the only function $V(x, \mathbf{p})$ which exhibits these properties. Arguably the most cited example of demand system which is both a member of the translog family and of the PIGLOG class is the *Almost Ideal Demand System* (AIDS) of Deaton and Muellbauer (1980). It follows from the specification of a very general functional form for $V(x, \mathbf{p})$ and it implies Engel curves in which budget shares are linear in the logarithm of total expenditure.

In general, however, there are no theoretical restrictions which impose budget shares to be linear in x. While for some goods this specification provides a good approximation (see Banks, Blundell, and Lewbel 1997) empirical findings show that a linear specification is not enough to characterize Engel curves for some categories of goods (see Atkinson, Gomulka, and Stern 1990, Bierens and Pott-Buter 1990, Blundell, Pashardes, and Weber 1993, Hausman, Newey, and Powell 1995, and Lewbel 1991). In this spirit Banks, Blundell, and Lewbel (1997) propose a generalization of the AIDS model and show that, if additional terms were to be included in the specification of Engel curves, then theory would restrict this term to be a quadratic term in the logarithm of income. Their *Quasi Almost Ideal Demand System* (QAIDS) demand system then implies Engel curves in which budget shares are quadratic in the logarithm of income. These theoretical results only apply to exactly aggregable demand systems and are very popular in empirical works since they are quite general and easy to estimate. Nonetheless, in the most recent years, a growing stream of research has focused on semi and non-parametric estimation of Engel curves to encompass a wider range of alternatives: see Hardle and Marron (1990), Pinkse and Robinson (1995) and Blundell, Browning, and Crawford (2003).

All models discussed so far imply that households deterministically set their level of consumption on each good given their level of income and prices. In practice, however, data exhibit great variability among households with the same observed characteristics. This is usually considered to be the result of, both observable and unobservable, heterogeneity in preferences and measurement errors. While measurement errors only affect the estimation of the models and do not pose theoretical issues, heterogeneous preferences raise the question of how (un)observed demographics should enter the utility function and which kind of restrictions these imply on consumers' behavior.

In theory one would need to specify an indirect utility function of the form $V(x, \mathbf{p}, \varepsilon)$, where ε is a vector of (un)observed characteristics of the household, and solve the maximization problem. In order to keep the model as parsimonious as possible, it would be tempting to include heterogeneity as an additive component to the reference Engel curve $g_i(x)$, the latter being defined as the Engel curve associated with the household for whom there is $\varepsilon = 0$. This would define the following relationship:

$$w_i = g_i(x) + \varepsilon, \tag{2.2}$$

which implies that heterogeneity acts as a vertical shifter for the Engel curve of the reference household. This, as showed by Blundell, Duncan, and Pendakur (1998), is in general inconsistent with utility maximization, meaning that there exists no sufficiently general indirect utility function which generates Engel curves of the form (2.2).

A more attractive solution to the problem of allowing for heterogeneity are shape invariant Engel curves. As discussed by Blundell, Duncan, and Pendakur (1998) a class of Engel curves which is consistent with utility maximization is:

$$w_i = h_i(\varepsilon) + g_i(x - m(\varepsilon)),$$

for some functions $h_i(\cdot)$ and $m(\cdot)$. These functions are shape invariant in the sense that, up to location and scale transformations, they look identical as ε varies. Shape invariant Engel curves are generated from demand systems which are derived from indirect utility functions associated with independent of base equivalence scales[1], that is:

$$V(x, \mathbf{p}, \varepsilon) = T(x - m(\varepsilon), \mathbf{p}),$$

for some functions $T(\cdot)$ and $m(\cdot)$. Heterogeneity is then allowed to enter the utility function only by suitably scaling total income. These results are of particular importance if one is willing to coherently specify unobserved heterogeneity, see Brown and Walker (1989), and I will consider this issue in Chapter 4.

2.2 Non-linear Errors in Variables Models

Identification and estimation of models when variables are observed with error has been a long-standing problem in both economics and statistics. Since the very pioneering work of Adcock (1878) a variety of approaches have been proposed to deal with this difficult task. The proposed approaches may be broadly divided into three main categories:

1. *The parametric approach*: it generally relies on assumptions on the measurement error distribution, allowing the estimation of a fully parametric model. In this framework the nature of the distribution considered is known to play a major role: in linear models, for instance, the parameters are not in general identified if the distribution of measurement error is normal. However, if one is willing to assume parametric distributions for the (unobserved) random variables involved, this is the classical approach to identification. See, for instance, Hsiao (1989), Hsiao and Wang (2000) and Carroll, Ruppert, Stefanski, and Crainiceanu (2006).

[1]This is not a characterization result, but it may be shown that "all shape invariant Engel curves are either derived from independent of base indirect utility function or from a restrictive class of alternative models" (Lewbel 2010).

2. *Repeated measurements*: this approach makes use of additional mismeasured observations of the variable of interest. This may be enough to provide identification under the assumption that measurement errors in the two repeated measurements are somewhat unrelated. This may be seen as a particular case of the IV approach below. Several works have been developed on the topic, see Hausman, Newey, Ichimura, and Powell (1991), Hausman, Newey, and Powell (1995), Li (2002) and Schennach (2004). The limitations here are given by the fact that data on repeated measurements of the quantities of interest are not always readily available.

3. *Instrumental variables*: this approach exploits additional information provided by an instrumental variable which is assumed to be correlated with the (unobserved) variable of interest but independent of the measurement error. This is arguably the most traditional approach in economics and the one which has received the most attention in recent years, see for instance Hausman, Newey, and Powell (1995), Lewbel (1996), Newey (2001) and Schennach (2007).

In Chapters 3 and 4 I will consider the instrumental variables approach. This is motivated by the fact that, while repeated measurements are not always available, the application of instrumental variables techniques is now common practice in econometrics.

The well known traditional framework in which the availability of IVs allows identification and consistent estimation of the parameters of interest is the classical linear model. Suppose:

$$Y_i^* = \theta_0 + \theta_1 X_i^* + \xi_i, \qquad E[\xi_i | X_i^*] = 0,$$

where only the couple (Y_i^*, X_i) is observed, while $X_i = X_i^* + W_i$ with $E[W_i|X_i^*] = 0$. The feasible regression of Y_i^* on the observed X_i may then be written as:

$$Y_i^* = \theta_0 + \theta_1 X_i + \xi_i - \theta_1 W_i,$$

which shows that *Ordinary Least Squares* (OLS) does not provide consistent estimates of $\theta = (\theta_0, \theta_1)$, since by construction $E[W_i|X_i] \neq 0$. A general result is that, let $\hat{\theta}_1$ be the OLS estimator for θ_1, then

$$\operatorname*{plim}_{n\to\infty} \hat{\theta}_1 = \frac{\sigma^2_{X^*}}{\sigma^2_{X^*} + \sigma^2_W} \theta_1,$$

showing that $\hat{\theta}_1$ is downward biased in magnitude.

In this case standard results show that the availability of a valid instrumental variable which is correlated with X_i^* but uncorrelated with W_i, allows the consistent estimation of the parameters by straightforwardly applying *Two Stage Least Squares* (2SLS) estimator.

However, as first pointed out by Amemiya (1985), this argument breaks down when the regression function is non-linear. If the dependent variable is a general function of X_i^*, i.e. $g(X_i^*;\theta)$, then the feasible regression of Y_i^* on X_i is no longer separable in the measurement error W_i and X_i:

$$Y_i^* = g(X_i + W_i;\theta) + \xi_i,$$

hence, even if an instrument with the above properties were available, θ could not be consistently estimated by 2SLS.

This lack of identifiability prompted a long search for identification in the so called non-linear errors in variables models when an instrument is available. Hausman, Newey, Ichimura, and Powell (1991) and Hausman, Newey, and Powell (1995) in a series of papers proved identification of polynomials specifications through instrumental variables and derived a root n consistent *Generalized Method of Moments* (GMM) estimator[2]. More general specifications have been discussed by Wang and Hsiao (1995), who unfortunately restrict their attention to functions which are absolutely integrable[3], hence ruling out most of the empirically relevant specifications, including polynomials for instance. An interesting result, for our purposes, is the one proposed by Lewbel (1996) who considers identification of a specific functional form, which is particularly attractive in the context of Engel curves estimation. The estimator he proposes is also based on a GMM procedure. A more general non-parametric identification result for a broad class of non-linear functions has been recently developed by Schennach (2007), who also provides a consistent, root n and asymptotically normal estimator based on the properties of the Fourier transform of $g(\cdot)$.

In Chapters 3 and 4 we will deal with the estimation of Engel curves when expenditures are measured with error. As we have seen in Section 2.1, Engel curves exhibit non-linear behaviors, thus allowing for measurement errors greatly complicates the identification of the model. Futhermore Engel curves possess a peculiar feature: allowing for measurement errors in the independent variable, that is total expenditure, implies that the dependent variable, that is expenditures on a single good (or possibly budget shares), is also measured with error, since by definition total expenditure is the sum of expenditures over all goods. Also note that these measurement errors are correlated by construction, a feature which invalidates most of the identification results for non-linear errors in variables models. The only identification results, I am aware of, which explicitly account for the presence of correlated measurement errors on both sides of the equation are those derived in Hausman, Newey, and Powell (1995) and Lewbel (1996).

[2]They also showed that a general non-linear specification $g(\cdot)$ is identified from additional information provided by repeated measurements.

[3]They assume that $g(\cdot)$ is such that $\int |g(x)|dx < \infty$.

2.3 Frequency of Purchase

Data on purchase have been largely used in the literature as a proxy for consumption. However it is well known that in large surveys, such as the *Family Expenditure Survey* in the UK, a non-negligible fraction of respondents report zero-expenditure on some commodities like drink or clothing. It is hard to interpret these data as zero-consumption, instead it seems quite natural to interpret expenditures as an error ridden measure of true consumption. This interpretation was first discussed in a series of papers related to the permanent income hypothesis (Summers 1959; Prais 1959; Liviatan 1961) and subsequently rather neglected until the 80s.

One of the main explanations for the presence of so many zeros in expenditure surveys is the so called frequency of purchase problem, which arises when the short time window over which the survey takes place prevents the observation of the complete pattern of purchases for each household. Such a problem is particularly severe in diary surveys, where the time window is of one or two weeks. On the other hand diary data are an important source of information on the consumption for a variety of commodities.

When dealing with this kind of data two main issues arise. Firstly distributional measures of consumption based on observed expenditure may give a misleading picture of the real distribution of consumption. Secondly, estimates of Engel curves based on expenditure data are in general inconsistent. To consider these issues a number of models have been developed (Kay, Keen, and Morris 1984; Deaton and Irish 1984; Keen 1986; Blundell and Meghir 1987; Meghir and Robin 1992; Robin 1993). In particular Kay, Keen, and Morris (1984) develop a general framework to link observed expenditure with underlying consumption showing that mean expenditure is an unbiased estimate of the mean of consumption under fairly general conditions, but estimates of the variances may be severely biased. Thus they introduce a consistent estimator for the variance of consumption under quite restrictive assumptions, such as linearity of the Engel curves and knowledge of its parameters.

Deaton and Irish (1984) make use of a special case of the above framework to test the hypotesis that zero-expenditures arise from other sources than that implyed by no-consumption. Their findings are controversial, the results being that the observed zeros are *less* than those explained by the model.

Keen (1986) shows that OLS estimation of Engel curves with observed expediture as a proxy for consumption is inconsistent under the presence of infrequency of purchase problems and develop a simple IV estimator which is consistent under the condition of linear Engel curves.

Meghir and Robin (1992) consider for the first time an explicit frequency of purchase model also including the potential effect of multiple purchases. Their goal is to construct an identification strategy to recover estimates for

the parameters of the Engel curves.

Chapter 3

Endogenous Total Expenditure

The objective of this Chapter is that of providing identification and estimation of Engel curves when total expediture is endogenous and measured with error. I consider Engel curves in which budget shares are polynomials in the logarithm of total expenditure. Identification is obtained exploiting additional information provided by an instrumental variable through control functions building on previous results by Lewbel (1996), which account for the presence of classical measurement error on the logarithms of total expenditure.

3.1 Introduction

Consider the following Engel curve:

$$W_{ih}^* = b_{i0} + b_{i1} \log X_h^* + \varepsilon_{ih}, \tag{3.1}$$

where $W_{ih}^* \equiv Y_{ih}^*/X_h^*$ is the budget share on the i-th good for household h, Y_{ih}^* being expenditure on the i-th good, while $X_h^* \equiv \sum_i^I Y_{ih}^*$ is total expenditure and I is the number of goods considered.

In empirical applications one would tipically estimate (3.1) using instrumental variables (IV) (see for instance Blundell, Chen, and Kristensen 2007 and Attanasio, Battistin, and Mesnard 2009 among others). This follows from the fact that there is no reason to assume that $\log X_h^*$ and ε_{ih} are uncorrelated, as in general X_h^* and Y_{ih}^* may be simultaneously chosen by individuals thus leading to endogeneity of X_h^*. In some cases restrictions on household's behaviour, such as two-stage budgeting, are introduced in order to avoid this source of endogeneity (Pollack and Wales, 1995), but in general there is no clear economic justification to assume that X_h^* is exogenous.

Another source of endogeneity in the estimation of (3.1) arises when X_h^* is measured with error. Then using X_h as a proxy for X_h^* would in

general introduce endogeneity. As Amemiya (1985) first pointed out, in this context IV would no longer be consistent, due to non-linearities in the specification. The result follows as measurement error is no longer additively separable, thus invalidating the use of IV. Moreover equation (3.1) exhibits highly correlated (by construction) measurement errors on both sides of the equation, a feature not usually encountered in the errors-in-variables literature (Lewbel, 1996).

The literature has addressed separately these two sources of endogeneity, with few exceptions (see for instance Blundell, Chen, and Kristensen 2007). A solution to the bias induced by measurement error in the same setup that I consider in this Chapter has been proposed by Lewbel (1996). He shows that the model in (3.1) is identified from knowledge of the conditional moments $E[X_h^l W_{ih}|Z_h]$, $E[X_h^l|Z_h]$ and $E[X_h^l \log X_h|Z_h]$ for $l = 1, 2$, where Z_h is a valid instrument for total expenditure. His main identifying assumptions are classical measurement error on $\log X_h^*$ and exogenous total expenditure ($E[\varepsilon_{ih}|X_h^*] = 0$). In what follows I provide a more general result which also allows for endogenous X_h^*.

I first characterize the asymptotic bias of the simple *two-stage-least-squares* (2SLS) estimator. I show that the IV estimator for b_i in equation (3.1) is upward biased in magnitude, a result which contrasts with the usual attenuation bias found in the literature in the case of linear specifications. I then propose an estimator which makes use of control functions to correct for the endogeneity of X_h^*. I rely throughout on standard assumptions from the control function literature and estimate Engel curve parameters via a generalized method of moments procedure. The result I provide may be extended to allow for exogenous error-free regressors, thus providing a practical way to estimate more general demand functions. The finite sample properties of the proposed estimator are evaluated with a Monte Carlo simulation study and compared to those of alternative estimators. Finally, I provide an empirical application to show how the method we propose can be applied to real data.

3.2 Identification

3.2.1 Exogenous Total Expenditure

I will consider identification of the Engel curve for the i-th good, since the whole system of I equations is recovered by treating each good separately and discarding the I-th equation which is uniquely identified by the summing-up properties of demand functions. Assume that:

$$W_{ih}^* = \sum_{j=0}^{K} b_{ij} (\log X_h^*)^j + \varepsilon_{ih}, \tag{3.2}$$

where K is a positive integer. To ease notation, in the remainder of the Chapter I will omit the subscript h from all variables. Identification and estimation of (3.2) is of particular interest in economics since it defines the shape of the Engel curves for the subset of goods under study. This specification is quite general and underpins many empirical specifications found in the literature. In particular, most budget shares models entail Engel curves which are polynomials in $\log X^*$. Notable examples are the AIDS (Deaton and Muellbauer, 1980) for $K = 1$, or the Quadratic AIDS (Banks, Blundell, and Lewbel, 1997) which corresponds to (3.2) with $K = 2$. I derive my result for a general K-th order polynomial for the sake of completeness, though most of the application would deal with $K \leq 2$. This follows from the fact that, if I restrict the attention to exactly aggregable demand systems, Gorman (1981) proved that the rank is at most three.

Identification of (3.2) is trivial when Y_i^* is observed since, when $E[\varepsilon_i|X^*] \neq 0$, a standard 2SLS estimator would provide consistent estimates of $b = (b_0, \ldots, b_K)$. Here I consider the case in which Y_i^*, and thus X^*, is unobserved. Instead its mismeasured counterpart Y_i is observed, such that:

$$Y_i = Y_i^* + X^* \nu_i. \tag{3.3}$$

This is consistent with (possibly correlated) measurement errors on all goods. The rationale behind such a specification for the measurement error follows from the fact that summing up over goods I obtain a classical measurement error on $\log X^*$, as:

$$X = \sum_{i=1}^{I} Y_i = X^* \left(1 + \sum_{i=1}^{I} \nu_i \right) = X^* V, \tag{3.4}$$

with $V = 1 + \sum_{i=1}^{I} \nu_i$, and then:

$$\log X = \log X^* + \log V. \tag{3.5}$$

This also allows for the variance of measurement error on expenditure levels to increase with total expenditure, a feature usually encountered in the data. Note that (3.3) together with (3.2) implies:

$$W_i = \frac{W_i^* + \nu_i}{V}, \tag{3.6}$$

which shows how the measurement error enters non-linearly on the left hand side of equation (3.2). The following set of assumptions will provide the basis for the identification of the Engel curve parameters when total expenditure is exogenous.

Assumption 3.1. *Let $(X^*, X, Y_i, Z, \varepsilon_i, \nu_i)$ be a vector of i.i.d. random variables such that:*

(i) $E[X|Z] \neq 0$,

(ii) $E[\varepsilon_i|Z] = 0$,

(iii) $E[\nu_i] = 0$ *with* $\nu_i \perp (X^*, Z, \varepsilon_i)$.

Assumptions (i) and (ii) are standard and ensure the validity of the instrument, while (iii) implies that the measurement errors are independent of total expenditure.[1] Full independence is required due to the non-linearities in the relationships considered. Note that (iii) also implies $E[V] = 1$.

Under Assumption 3.1 by multiplying either side of (3.6) by $X = X^*V$ and taking conditional expectations with respect to Z it is (see Lewbel 1996):

$$\begin{aligned} E[XW_i|Z] &= E\left[\frac{X^*VW_i^*}{V}|Z\right] + E\left[\frac{X^*V\nu_i}{V}\right], \\ &= \sum_{j=0}^{K} b_{ij} E[X^*(\log X^*)^j|Z] + E[X^*\varepsilon_i|Z]. \end{aligned} \tag{3.7}$$

Lewbel (1996) shows that, under Assumption 3.1, unobservable moments on the right hand side of the above equation may be expressed in terms of moments that involve observable variables. To see this consider identification for the $K = 1$ case, which I will maintain throughout as running example (a general identification result will be given in Theorem 3.1). There is:

$$E[X^*|Z] = E[X|Z], \tag{3.8}$$

$$E[X^* \log X^*|Z] = E[X \log X|Z] - E[X|Z]E[V \log V], \tag{3.9}$$

and by using $X = X^*V$:

$$E[X^*\varepsilon_i|Z] = E[X\varepsilon_i|Z]. \tag{3.10}$$

Substitution of (3.8), (3.9) and (3.10) into (3.7) yields:

$$E[Y_i|Z] = (b_{i0} - b_{i1}E[V \log V])E[X|Z] + b_{i1}E[X \log X|Z] + E[X\varepsilon_i|Z].$$

When X^* is exogenous (that is when $E[\varepsilon_i|X^*] = 0$) the last term in the above equation is zero. Thus b_{i1} is identified and consistent estimates may be obtained through a 2SLS regression of Y_i on X and $X \log X$ with no constant terms using Z as an instrument. Identification of b_{i0} follows along the same lines exploiting similar expressions for $E[X^l W_i|Z]$, with $l \geq 1$.

Example 1
In order to fix ideas suppose the data are generated from model (3.1) where:

$$\varepsilon_i \sim N(0, \sigma_\varepsilon^2),$$

[1] In general Z may represent a vector of valid instruments, as it will be considered in the empirical example below.

X^* is log-normally distributed and an instrument is available such that:

$$\log Z \sim N(\mu_Z, \sigma_Z^2),$$

with $E[\varepsilon_i|Z] = 0$. If X^* is exogenous, under Assumption 3.1 the random variable $\log \xi$ is independent of ε_i in the following first stage regression:

$$\log X^* = \gamma_0 + \gamma_1 \log Z + \log \xi, \tag{3.11}$$

with $\log \xi \sim N(0, \sigma_\xi^2)$. Measurement error V is defined as in equation (3.5) and independent of X^*, such that:

$$\log V \sim N(-\sigma_V^2/2, \sigma_V^2). \tag{3.12}$$

Note that $E[\log V] < 0$ to ensure that $E[V] = 1$, which is an assumption I maintain in what follows. This setting is of particular interest since there is widespread evidence that consumption is log-normally distributed (as documented by Battistin, Blundell, and Lewbel 2009). This, under the assumption that the measurement error is independent of total expenditure, implies that both X^* and V are log-normally distributed. Here X^* is independent of ε_i, hence $E[X^*\varepsilon_i|Z] = 0$ and Lewbel's (1996) Theorem is readily applied to provide identification of the parameters of interest b_{i0} and b_{i1}. ■

3.2.2 Endogenous Total Expenditure

In what follows I will relax the exogeneity assumption $E[\varepsilon_i|X^*] = 0$, and will consider conditions which allows us to express $E[X\varepsilon_i|Z]$ in terms of observable moments.

To this end I use a control function approach following Imbens and Newey (2009). As shown by Hahn and Ridder (2010), it is always possible to write X as a function of the instrument Z and an independent error term η such that $X = g(Z, \eta)$. In particular let $F_{X|Z}(x|z) \equiv Pr(X \leq x|Z = z)$ be the conditional distribution of X given Z, and define the following random variable:

$$\eta \equiv F_{X|Z}(X|Z), \tag{3.13}$$

so that η is uniformly distributed and independent of Z. By defining $g(z, e) \equiv F^{-1}(e|z)$, where the inverse is computed with respect to the first argument, it is:

$$X = g(Z, \eta), \qquad Z \perp \eta.$$

Identification under endogeneity and possible mismeasurement of total expenditure will build upon the following assumption:

Assumption 3.2. *Let (ε_i, Z, η) be a vector of iid random variables such that:*

$$E[\varepsilon_i|Z, \eta] = E[\varepsilon_i|\eta] \equiv \lambda_i(\eta).$$

This is a standard assumption in the control function literature. This for instance would hold if Z was independent of ε_i, but it is slightly weaker than that, allowing ε_i to be heterosckedastic.[2]

Using the law of iterated expectations, it follows from Assumption 3.2 that:

$$\begin{aligned} E[X\varepsilon_i|Z] &= E[g(Z,\eta)\varepsilon_i|Z], \\ &= E\Big\{g(Z,\eta)E[\varepsilon_i|Z,\eta]|Z\Big\}, \\ &= E[g(Z,\eta)\lambda_i(\eta)|Z], \end{aligned}$$

where $E[\lambda_i(\eta)] = 0$. Note that, since η is uniformly distributed in [0,1], it is:

$$E[g(Z,\eta)\lambda_i(\eta)|Z] = \int_0^1 g(z,\eta)\lambda_i(\eta)d\eta. \tag{3.14}$$

Now consider three possible specifications for the functional $g(z,\eta)$:

(i) $g(z,\eta)$ is additively separable in its arguments, that is $g(z,\eta) = h_1(z) + h_2(\eta)$. It follows from (3.14) and the fact that $E[\lambda_i(\eta)] = 0$ that:

$$\begin{aligned} E[X\varepsilon_i|Z] &= f(Z)\int_0^1 \lambda_i(\eta)d\eta + \int_0^1 h(\eta)\lambda_i(\eta)d\eta, \\ &= Cov(h(\eta),\varepsilon_i), \end{aligned} \tag{3.15}$$

which is a constant with respect to Z.

(ii) $g(z,\eta)$ is multiplicatively separable in its arguments, that is $g(z,\eta) = f(z)h(\eta)$. Then there is:

$$\begin{aligned} E[X\varepsilon_i|Z] &= f(Z)\int_0^1 h(\eta)\lambda_i(\eta)d\eta, \\ &= f(Z)Cov(h(\eta),\varepsilon_i). \end{aligned}$$

(iii) $g(z,\eta)$ is non-separable. In this case a close form is no longer available since:

$$\begin{aligned} E[X\varepsilon_i|Z] &= \int_0^1 g(Z,\eta)\lambda_i(\eta)d\eta, \\ &= \tau(Z), \end{aligned}$$

where $\tau(\cdot)$ is an unknown function of Z which depends on the form of $\lambda_i(\cdot)$.[3]

[2]Note that this is not exactly the same of assuming that $\delta \perp Z$ in the error-free first stage $X^* = Z + \delta$, though in practice these two assumptions are very close.

[3]Since $\lambda_i(\cdot)$ is unknown, formal identification when $g(z,\eta)$ is non-separable in its arguments would require regularity conditions on the form of $\lambda_i(\eta)$, in order for this to be suitably approximated by some basis functions. However this is beyond the scope of the present thesis and left for future research, since in most of the applications the multiplicative separability will be a reasonable assumption to make.

Example 1 (continued)
Suppose:

$$Cov(\varepsilon_i, \log \xi) = \sigma_{\varepsilon\xi} \neq 0,$$

then it immediately follows that:

$$E[X^* \varepsilon_i | Z] \neq 0,$$

and the the usual results in Lewbel (1996) may not be applied. However, in this fully parameterized model, using (3.5) and (3.11) we may write:

$$\log X = \gamma_0 + \gamma_1 \log Z + \log \xi + \log V,$$

or alternatively:

$$X = \exp\{\gamma_0 + \gamma_1 \log Z\} V \xi,$$

with $\log(V\xi) \sim N(-\sigma_V^2/2, \sigma_V^2 + \sigma_\xi^2)$. Now according to equation (3.13) there is:

$$\eta \equiv F_{X|Z}(X|Z) = \Phi\left(\frac{X}{\exp\{\gamma_0 + \gamma_1 \log Z\}}\right),$$

where $\Phi(\cdot)$ is the conditional distribution function of a lognormal distribution with parameters $-\sigma_V^2/2$ and $\sigma_V^2 + \sigma_\xi^2$. It thus follows that X can be rewritten in terms of η as:

$$X = \exp\{\gamma_0 + \gamma_1 \log Z\} \Phi^{-1}(\eta),$$

implying that $g(z, \eta)$ is multiplicatively separable and $f(z) = \exp\{\gamma_0 + \gamma_1 \log z\}$, which in turn yields:

$$E[X\varepsilon_i|Z] \quad = \quad \exp\{\gamma_0 + \gamma_1 \log Z\} Cov\left(\Phi^{-1}(\eta), \varepsilon_i\right).$$

■

I can now state the main identification result for the generic K-th order polynomial case.

Theorem 3.1. *Let equations* (3.2) *and* (3.5) *and Assumptions 3.1 and 3.2 hold. For any integer $l \geq 1$ for which $E[V^l \log V]$ and $E[\nu_i V^{l-1}]$ are finite, it is:*

$$E[X^l W_i | Z] = \sum_{t=0}^{K} \beta_{ilt} E[X^l (\log X)^t | Z] + \tau_i(Z),$$

where:

$$\beta_{ilt} = \frac{E[V^{l-1}\nu_i]}{E[V^l]} \mathbf{1}(t = 0) + \sum_{j=0}^{K} b_{ij} E[V^{l-1}] \gamma_{j-t} \mathbf{1}(j - t \geq 0), \tag{3.16}$$

and $\gamma_t = (-1)^t \frac{E[V^l \log V]^t}{E[V^l]^{t+1}}$, *while:*

$$\tau_{il}(Z) \equiv \frac{E[V^{l-1}]}{E[V^l]} \int_0^1 g(Z,\eta)^l \lambda_i(\eta) d\eta,$$

and $\lambda_i(\eta) = E[\varepsilon_i|\eta]$ *is a generic function such that* $\int_0^1 \lambda_i(\eta)d\eta = 0$.

The proof, which is given in the Appendix, is a generalization of Theorem 1 in Lewbel (1996) using Assumption 3.2 to characterize $E[X\varepsilon_i|Z]$. When total expenditure is exogenous it is $E[\varepsilon_i|X^*] = E[\varepsilon_i|\eta] = 0$, which is a special case of Theorem 3.1 with $\lambda_i(\eta) \equiv 0$. It is straightforward to see that the coefficients of interest are identified as in Lewbel (1996), who exploits the fact that b_i is uniquely determined from knowledge of β_{lt} for $l \geq 1$. In order to see this suppose $K = 1$ and $l = 1$. It then follows from equation (3.16) that $b_{i1} = \beta_{11}$, while under the assumption that V is log-normally distributed it may be shown that $b_{i0} = \beta_{10} + \beta_{11} \log(\beta_{11}/\beta_{21})/2$; see Lewbel (1996) for more details.

Theorem 3.1 implies that when $E(\varepsilon_i|X^*) \neq 0$ and X^* is mismeasured, standard OLS and 2SLS estimation would not yield consistent estimates of the parameters of interest. The following theorem, whose proof is derived in the Appendix, gives the expression for the asimptotic bias of the 2SLS estimator.

Theorem 3.2. *Let equation* (3.2), (3.5) *and Assumption 3.1 hold, then*

$$E[W_i|Z] = \sum_{t=0}^{K} \beta_{it}^{2SLS} E[(\log X)^t|Z],$$

where

$$\beta_{it}^{2SLS} = E[\nu_{ih} V_h^{-1}]\mathbf{1}(t=0) + \sum_{j=0}^{K} E[V_h^{-1}] b_{ij} \gamma_{j-t} \mathbf{1}(j - t \geq 0) \qquad (3.17)$$

and $\gamma_t = (-1)^t E[\log V]^t$.

The coefficients β_{it}^{2SLS} in Theorem 3.2 may be interpreted as the probability limit of the 2SLS regression of W_i on a K-th order polynomial in $\log X$ using Z as instrument. An interesting implication of the Theorem 3.2 is the following:

Corollary 3.3. *Equation* (3.17) *implies*

$$\beta_{iK}^{2SLS} = E[V^{-1}] b_{iK}.$$

Corollary 3.3 shows that the 2SLS bias for the highest order coefficient of the polynomial specification (3.2) is proportional to $E[V^{-1}]$. It follows from Jensen's inequality that $E[V^{-1}] > E[V]^{-1} = 1$, hence the 2SLS estimator provides upward biased estimates in magnitude of the coefficient b_{iK}. It is also worth noting that by taking a second order Taylor series expansion of $E[V^{-1}]$ around 1 it is:

$$E[V^{-1}] \approx E[V] + Var[V] = 1 + Var[V]. \tag{3.18}$$

Equation (3.18) shows that the magnitude of the bias is approximately proportional to the variance of the measurement error, with the approximation being exact if V is log-normally distributed. This suggests an alternative identification strategy since $\beta_{iK}^{2SLS}/\beta_{iK} = E[V^{-1}]$. Under the assumption of log-normality of the measurement error, for instance, this is enough to identify the distribution of V and then b_i is also identified.

3.3 Estimation

The estimator I propose follows directly from Theorems 3.1 and 3.2. It is based on a GMM procedure with the addition of a control variable to correct for the endogeneity of the unobserved true regressor. The nature of the control variable will depend crucially on the assumption made on the form of $g(Z, \eta)$. If $g(\cdot)$ is additively separable in its arguments, then b_i may be estimated from the following regression functions:

$$X^l W_i = \alpha_i + \sum_{j=0}^{K} \beta_{ij} X^l (\log X)^j \qquad \text{for } l = 1, 2, \ldots$$

using Z as instrument.

Similarly if $g(\cdot)$ is multiplicatively separable, that is $g(z, \eta) = f(z)h(\eta)$, then the estimation is performed in two stages:

1. Recover an estimate for the conditional mean of X given Z: $\hat{f}(z)$.

2. Run the following regressions:

$$X^l W_i = \alpha_i \hat{f}(Z) + \sum_{j=0}^{K} \beta_{ij} X^l (\log X)^j \qquad \text{for } l = 1, 2, \ldots \tag{3.19}$$

   using Z as an instrument.

Once the β_{ij} coefficients are estimated, the parameters of interest are obtained from equation (3.16) in a way completely similar to Lewbel (1996), using the identification result in Theorem 3.1.

The estimation of the parameters is computationally more demanding if we don't make any assumption regarding the form of $g(\cdot)$. In this case it is worth noting that $g(\cdot)$ is identified from the knowledge of the couple (X, Z) upon inversion of (3.13). In particular let $\hat{F}_{X|Z}(x|z)$ be an estimator for the conditional distribution function of X given Z. An example of a consistent estimator for the conditional distribution function $F(\cdot|\cdot)$ may be found in Imbens and Newey (2009). Consequently an estimator for $g(z, \eta)$ is $\hat{F}^{-1}_{X|Z}(x|z)$. Note that this could also be used to test previous assumptions regarding the functional form of the first stage equation. Now using the fact that η is by construction uniformly distributed and by approximating $\lambda_i(\eta)$ with a polynomial, that is $\lambda_i(\eta) = \sum_{j=0}^{J} \delta_j \eta^j$ for some constant J [4], we obtain the desired control function by numerically integrating

$$\hat{\tau}_{il}(Z) = \int_0^1 \hat{F}^{-1}_{X|Z}(X|Z)^l \lambda_i(\eta) d\eta$$

It follows from Theorem 3.1 that applying 2SLS to the following regression functions

$$X^l W_i = \hat{\tau}_{il}(Z) + \sum_{j=0}^{K} \beta_{ij} X^l (\log X)^j \qquad \text{for } l = 1, 2, \ldots \tag{3.20}$$

would consistently estimate the parameters as above.

Theorem 3.2 provides an alternative, more efficient, way of estimating $E[V^{-1}]$. Under the assumption that V is lognormally distributed, for instence, this would allow the estimation of both the distribution of V and the parameter of interest b_0, without relying on equations (3.3), (3.19) and (3.20) with $l \geq 2$. This result follows from Corollary 3.3, which shows that $E[V^{-1}]$ is identified by the ratio of estimates of b_1 obtained by applying simple 2SLS regression of W_i on a polynomial in $\log X$, and by equation (3.19) or (3.20) with $l = 1$. The parameter b_0 is then obtained as a function of the estimated coefficients in the usual way. The efficiency gain with this procedure comes from the fact that the simple 2SLS estimator, though being biased, has a smaller variance than the proposed estimator for b_1, since the latter are obtained by multiplying each side of the equation by X, hence greatly increasing the variance of the error term. Correcting the bias of IV is then more efficient than using the consistent estimator proposed above. It is finally worth noting that if measurement error is present in only one good, then b_{i0} may be recovered starting from the simple 2SLS estimate of the constant term, which further improves the performance of the estimator.

[4] Note that the fact that $E[\varepsilon_i] = 0$ puts restrictions on the coefficients δ_j, namely

$$E\left[\sum_{j=0}^{J} \delta_j \eta^j | \eta\right] = 0$$

.

Example 1 (continued)
It follows from Theorems 3.1 and 3.2 that

$$\beta_{i1} = b_{i1} \qquad \beta_{i1}^{2SLS} = b_{i1} E[V^{-1}].$$

Then by equation (3.12) it is $E[V^{-1}] = \exp(\sigma_V^2)$ which implies

$$\hat{\sigma}_V^2 = \log\left(\frac{\hat{\beta}_{i1}^{2SLS}}{\hat{\beta}_{i1}}\right). \tag{3.21}$$

Equations (3.17) and (3.21), under the assumption that ν_i is the only source of measurement error, allows us to estimate the constant term b_{i0} as

$$\hat{b}_{i0} = \frac{\hat{\beta}_{i0}^{2SLS} - 1}{e^{\hat{\sigma}^2}} + 1 - \frac{\hat{\sigma}^2 \hat{\beta}_{i1}}{2}.$$

■

3.4 Monte Carlo Simulation

In order to asses the finite sample properties of the proposed estimator a Monte Carlo simulation study is performed. The goal of this exercise is to compare the endogeneity-corrected estimator to the simple IV estimator, for which an expression for the bias is given in Section 3.2, and to the one proposed by Lewbel (1996). A simple model with only two goods is considered, hence the whole set of Engel curves reduces to one equation. Parameters are calibrated on 1995 SHIW food data for married couples without children. The instrument enters the first stage equation in the following way

$$\log X^* = \gamma_0 + \gamma_1 \log Z + \log \xi \tag{3.22}$$

with

$$\log \xi \sim N(0, 0.07) \qquad \log Z \sim N(10.5, 0.34)$$
$$\gamma_0 = 3.65 \qquad \gamma_1 = 0.62.$$

I consider specifications for the true budget shares model up to a quadratic term in $\log X^*$, that is

$$W_i^* = b_{i0} + b_{i1} \log X^* + b_{i2} (\log X^*)^2 + \varepsilon_i \tag{3.23}$$

where

$$\varepsilon_i = \theta_1 \xi + \psi_i$$

with $\psi_i \sim N(0, \sigma_\psi^2)$. The parameters θ_1 and σ_ψ^2 are suitably chosen such that $Corr(\varepsilon_i, \xi) = 0$, 0.3, 0.5 and 0.7 to simulate different extents of endogeneity, while we set $Var(\varepsilon_i) = 0.0025$. Two specifications are considered:

(i) **linear** - that is $b_{i0} = 1.6$, $b_{i1} = -0.11$, $b_{i2} = 0$.

(ii) **quadratic** - with $b_{i0} = -1$, $b_{i1} = 0.42$, $b_{i2} = -0.04$.

Measurement error of the form outlined in Section 3.2 is introduced. In particular only Y_1^* is affected by error and $\log X$ is such that:

$$\log X = \log X^* + \log V \tag{3.24}$$

where $\log V \sim N(-\sigma_V^2/2, \sigma_V^2)$. Different amounts of meaurement errors are considered by setting the noise to signal ratio, that is $\frac{Var(\log V)}{Var(\log X^*)}$, to 0, 0.1, 0.3 and 0.5, which corresponds to values of σ_V of 0, 0.13, 0.26 and 0.4 respectively. I compare the performances of the proposed estimator with the inconsistent alternatives given by OLS, 2SLS and Lewbel's estimator over 10000 replications. Lewbel's estimator is computed as in Lewbel (1996), assuming log-normality of the measurement error and using the same functions of Z as instruments. The proposed estimator is computed under the assumption of multiplicative separability; the first stage is parametrically estimated through a linear regression of $\log X$ on $\log Z$.

The results of the simulation for the linear and quadratic specifications are given in the tables below, showing that the proposed estimator outperforms the other methods especially when the extent of both endogeneity and measurement error is severe.

		0		**0.3**		**0.5**		**0.7**	
0 %	OLS	0.0012	(*0.0925*)	0.3561	(*0.0917*)	0.5893	(*0.0888*)	0.8287	(*0.0864*)
	IV	0.0015	(*0.1128*)	0.0032	(*0.1137*)	0.0037	(*0.1128*)	0.0051	(*0.1129*)
	LEWBEL	-0.0004	(*0.1826*)	0.0093	(*0.1878*)	0.0137	(*0.1862*)	0.0200	(*0.1900*)
	PRESENT	-0.0010	(*0.1751*)	-0.0017	(*0.1721*)	-0.0038	(*0.1609*)	-0.0048	(*0.1489*)
10 %	OLS	0.4345	(*0.1146*)	0.7664	(*0.1117*)	0.9882	(*0.1079*)	1.2082	(*0.1027*)
	IV	-0.0148	(*0.1478*)	-0.0122	(*0.1495*)	-0.0100	(*0.1461*)	-0.0092	(*0.1470*)
	LEWBEL	0.0127	(*0.2548*)	0.0195	(*0.2523*)	0.0261	(*0.2542*)	0.0304	(*0.2491*)
	PRESENT	-0.0047	(*0.2385*)	-0.0070	(*0.2198*)	-0.0076	(*0.2053*)	-0.0068	(*0.1888*)
30 %	OLS	1.4352	(*0.1505*)	1.7206	(*0.1417*)	1.9097	(*0.1382*)	2.0988	(*0.1322*)
	IV	-0.0536	(*0.2283*)	-0.0526	(*0.2260*)	-0.0537	(*0.2283*)	-0.0470	(*0.2282*)
	LEWBEL	0.0468	(*0.3928*)	0.0511	(*0.3913*)	0.0577	(*0.3861*)	0.0668	(*0.3832*)
	PRESENT	-0.0115	(*0.3192*)	-0.0163	(*0.2991*)	-0.0157	(*0.2807*)	-0.0128	(*0.2652*)
50 %	OLS	2.6553	(*0.1893*)	2.8912	(*0.1852*)	3.0426	(*0.1761*)	3.1971	(*0.1715*)
	IV	-0.1354	(*0.3494*)	-0.1303	(*0.3494*)	-0.1233	(*0.3424*)	-0.1329	(*0.3457*)
	LEWBEL	0.0917	(*0.5566*)	0.1019	(*0.5558*)	0.1124	(*0.5438*)	0.1024	(*0.5439*)
	PRESENT	-0.0307	(*0.4082*)	-0.0274	(*0.3924*)	-0.0262	(*0.3628*)	-0.0350	(*0.3515*)

Table 3.1: Empirical Percentage Bias for $\hat{\beta}_1$ $((\hat{\beta}_1 + 0.11)/0.11)$, Linear Specification, N = 500. Extent of endogeneity in columns, noise to signal ratio in rows; standard deviation in parentesis.

		0		**0.3**		**0.5**		**0.7**	
0 %	OLS	0.0007	(*0.0295*)	0.3554	(*0.0285*)	0.5914	(*0.0281*)	0.8284	(*0.0270*)
	IV	0.0010	(*0.0357*)	0.0003	(*0.0352*)	0.0002	(*0.0356*)	0.0006	(*0.0357*)
	LEWBEL	0.0006	(*0.0595*)	0.0022	(*0.0606*)	0.0030	(*0.0613*)	0.0054	(*0.0622*)
	PRESENT	0.0008	(*0.0582*)	-0.0003	(*0.0573*)	-0.0013	(*0.0544*)	-0.0005	(*0.0493*)
10 %	OLS	0.4356	(*0.0366*)	0.7676	(*0.0354*)	0.9889	(*0.0338*)	1.2103	(*0.0323*)
	IV	-0.0168	(*0.0475*)	-0.0157	(*0.0473*)	-0.0160	(*0.0468*)	-0.0157	(*0.0468*)
	LEWBEL	0.0027	(*0.0827*)	0.0070	(*0.0833*)	0.0067	(*0.0834*)	0.0084	(*0.0844*)
	PRESENT	-0.0008	(*0.0788*)	0.0007	(*0.0729*)	-0.0015	(*0.0687*)	-0.0019	(*0.0626*)
30 %	OLS	1.4364	(*0.0472*)	1.7199	(*0.0459*)	1.9082	(*0.0439*)	2.0972	(*0.0420*)
	IV	-0.0685	(*0.0720*)	-0.0666	(*0.0727*)	-0.0657	(*0.0715*)	-0.0672	(*0.0718*)
	LEWBEL	0.0114	(*0.1325*)	0.0150	(*0.1302*)	0.0153	(*0.1313*)	0.0181	(*0.1325*)
	PRESENT	-0.0038	(*0.1057*)	-0.0018	(*0.0999*)	-0.0019	(*0.0921*)	-0.0028	(*0.0862*)
50 %	OLS	2.6528	(*0.0603*)	2.8874	(*0.0580*)	3.0437	(*0.0557*)	3.1989	(*0.0546*)
	IV	-0.1632	(*0.1105*)	-0.1622	(*0.1102*)	-0.1623	(*0.1098*)	-0.1634	(*0.1103*)
	LEWBEL	0.0238	(*0.1984*)	0.0271	(*0.2001*)	0.0270	(*0.1967*)	0.0283	(*0.1972*)
	PRESENT	-0.0088	(*0.1345*)	-0.0093	(*0.1265*)	-0.0084	(*0.1200*)	-0.0102	(*0.1121*)

Table 3.2: Empirical Percentage Bias for $\hat{\beta}_1$ $((\hat{\beta}_1 + 0.11)/0.11)$, Linear Specification, N = 5000. Extent of endogeneity in columns, noise to signal ratio in rows; standard deviation in parentesis.

		0		**0.3**		**0.5**		**0.7**	
0 %	OLS	-0.0005	(*0.0268*)	-0.0001	(*0.0256*)	0.0000	(*0.0247*)	-0.0000	(*0.0234*)
	IV	-0.0009	(*0.0455*)	0.0020	(*0.0457*)	0.0042	(*0.0450*)	0.0048	(*0.0451*)
	LEWBEL	-0.0001	(*0.0689*)	0.0033	(*0.0686*)	0.0079	(*0.0679*)	0.0101	(*0.0667*)
	PRESENT	-0.0310	(*0.8534*)	-0.1589	(*12.5554*)	-0.0404	(*0.9046*)	-0.0336	(*0.5797*)
10 %	OLS	0.0577	(*0.5445*)	0.0511	(*0.5526*)	0.0424	(*0.5461*)	0.0418	(*0.5458*)
	IV	0.0943	(*1.0941*)	0.0829	(*1.1023*)	0.0863	(*1.0929*)	0.0854	(*1.1038*)
	LEWBEL	0.1742	(*1.7296*)	0.2123	(*1.7408*)	0.2060	(*1.7041*)	0.2128	(*1.7156*)
	PRESENT	-0.0782	(*2.6089*)	-0.0220	(*2.4691*)	-0.0950	(*3.4664*)	-0.0515	(*3.4216*)
30 %	OLS	-0.6200	(*0.8177*)	-0.6549	(*0.8305*)	-0.6594	(*0.8144*)	-0.6657	(*0.8125*)
	IV	0.2894	(*2.2772*)	0.3134	(*2.3068*)	0.2594	(*2.2987*)	0.3118	(*2.3140*)
	LEWBEL	0.7549	(*3.2476*)	0.7119	(*3.2870*)	0.6825	(*3.2769*)	0.7518	(*3.2890*)
	PRESENT	-0.0591	(*7.0246*)	-0.1253	(*4.4506*)	-0.0353	(*5.5336*)	-0.0972	(*6.5864*)
50 %	OLS	-2.5250	(*1.0012*)	-2.5564	(*0.9995*)	-2.5704	(*1.0076*)	-2.5610	(*0.9875*)
	IV	0.6974	(*3.8820*)	0.6833	(*3.8224*)	0.7183	(*3.8942*)	0.7272	(*3.9353*)
	LEWBEL	1.7543	(*5.0436*)	1.8294	(*5.0238*)	1.7837	(*5.1104*)	1.7406	(*5.1777*)
	PRESENT	-0.1265	(*7.3929*)	-0.1498	(*6.6778*)	-0.3931	(*27.3196*)	0.0255	(*8.7221*)

Table 3.3: Empirical Percentage Bias for $\hat{\beta}_2$ $((\hat{\beta}_2 + 0.04)/0.04)$, Quadratic Specification, N = 500. Extent of endogeneity in columns, noise to signal ratio in rows; standard deviation in parentesis.

		0		**0.3**		**0.5**		**0.7**	
0 %	OLS	-0.0001	(*0.0082*)	-0.0000	(*0.0081*)	-0.0000	(*0.0077*)	0.0000	(*0.0073*)
	IV	-0.0000	(*0.0142*)	0.0003	(*0.0141*)	0.0005	(*0.0141*)	0.0009	(*0.0141*)
	LEWBEL	-0.0001	(*0.0238*)	0.0017	(*0.0240*)	0.0025	(*0.0238*)	0.0036	(*0.0240*)
	PRESENT	-0.0037	(*0.1915*)	-0.0034	(*0.1273*)	-0.0042	(*0.1441*)	-0.0015	(*0.0533*)
10 %	OLS	0.0535	(*0.1698*)	0.0493	(*0.1706*)	0.0430	(*0.1702*)	0.0400	(*0.1714*)
	IV	0.0004	(*0.3412*)	-0.0001	(*0.3443*)	-0.0042	(*0.3423*)	0.0018	(*0.3450*)
	LEWBEL	0.0739	(*0.6243*)	0.0749	(*0.6237*)	0.0745	(*0.6166*)	0.0710	(*0.6183*)
	PRESENT	-0.0111	(*0.7269*)	-0.0061	(*0.7460*)	-0.0191	(*0.6663*)	0.0032	(*0.7835*)
30 %	OLS	-0.6296	(*0.2604*)	-0.6490	(*0.2559*)	-0.6581	(*0.2576*)	-0.6669	(*0.2564*)
	IV	-0.0015	(*0.7165*)	-0.0079	(*0.7150*)	-0.0036	(*0.7228*)	-0.0207	(*0.7269*)
	LEWBEL	0.2440	(*1.2123*)	0.2356	(*1.1990*)	0.2452	(*1.2210*)	0.2282	(*1.2306*)
	PRESENT	-0.0266	(*1.0408*)	-0.0250	(*1.0659*)	-0.0131	(*1.0912*)	-0.0127	(*0.9763*)
50 %	OLS	-2.5341	(*0.3209*)	-2.5546	(*0.3126*)	-2.5636	(*0.3178*)	-2.5758	(*0.3202*)
	IV	-0.0425	(*1.2078*)	-0.0344	(*1.1988*)	-0.0345	(*1.2067*)	-0.0178	(*1.2035*)
	LEWBEL	0.4819	(*1.8184*)	0.4704	(*1.8161*)	0.4715	(*1.7937*)	0.5015	(*1.8047*)
	PRESENT	-0.0433	(*1.3307*)	-0.0494	(*1.3983*)	-0.0310	(*1.2175*)	-0.0053	(*2.1754*)

Table 3.4: Empirical Percentage Bias for $\hat{\beta}_2$ ($(\hat{\beta}_2 + 0.04)/0.04$), Quadratic Specification, N = 5000. Extent of endogeneity in columns, noise to signal ratio in rows; standard deviation in parentesis.

3.5 Application

Section 3.3 proposes an estimator for the coefficients of an Engel curve. Here I apply such an estimator to data from the 2006 Survey on Household's Income and Wealth (SHIW). In order to select a demographically homogeneous sample, so that b_{ij} may be treated as constants, only married couples living in the north of Italy with no children are included, leaving us with an actual sample of 757 households.

In the following X_h is household's total expenditure, while W_{ih} is the proportion of X_h spent on food by household h. The specification we adopt is the usual Working-Leser (Working 1943, Leser 1963) functional form, which corresponds to (3.2) with $K = 1$. This specification is chosen based on the widespread evidence that it is particularly suited for modeling food Engel curves (Banks, Blundell, and Lewbel 1997, Blundell, Duncan, and Pendakur 1998).

Following the existing literature, real-income (Z) is used as an instrument, being highly correlated with total expenditure and unlikely correlated with the measurement error in expenditures. The functions of the instrument Z used are: Z, $Z \log Z$, Z^2 and $Z^2 \log Z$, which are the same functions of the instrument proposed by Lewbel (1996). The first stage regression equation is assumed to be multiplicatively separable, which is equivalent to assume that $\log X$ depends linearly on $\log Z$. Figure 3.1 reports the scatter-plot of $\log X$ on $\log Z$, showing that the conditional mean of $\log X$ is reasonably approximated by a linear function in the logarithm of income.

Table 3.5 reports the estimates for the intercept and the linear coefficient obtained by applying OLS and IV. As expected they are inconsistent, suggesting that measurement error may be an issue here, along with the presence of endogeneity of total expenditure X^*. In order to asses to what extent these sources of bias affect the results I report the estimates of the parameters of interest obtained using Lewbel's (1996) estimator and the estimating procedure proposed in Section 3.3.

OLS		**IV**	
b_{i0}	b_{i1}	b_{i0}	b_{i1}
1.5043	-.1509	1.2676	-.1175
(.0743)	*(.0105)*	*(.1229)*	*(.0174)*

Table 3.5: Standard OLS and IV estimates. Standard errors in parentheses.

The results are given in Table 3.6 along with an estimate of the variance of the measurement error V. The estimates for the linear coefficient seem

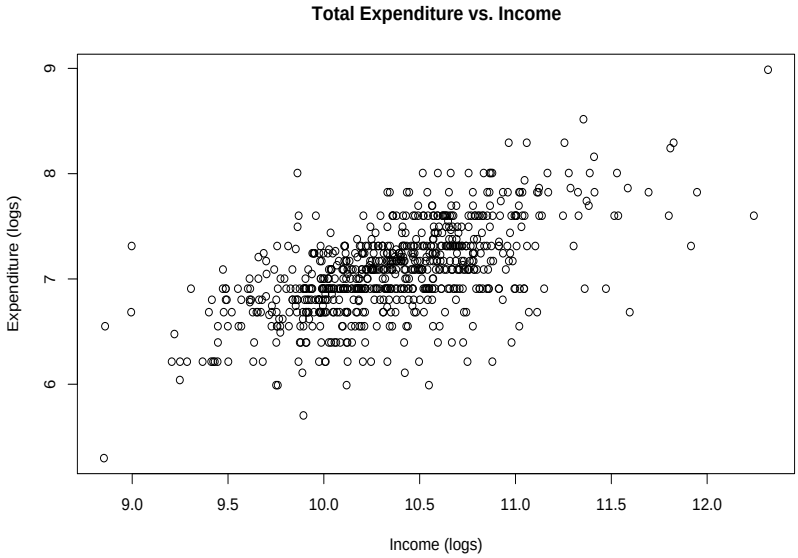

Figure 3.1: Scatterplot of logarithms of total expenditure vs. logarithms of real-income.

to be very close to one another, suggesting that the extent of measurement error is in this case very limited, as reported by the estimated variance of V, which appears to be not significantly different from zero.

Exogenous		**Endogenous**		
b_{i0}	b_{i1}	b_{i0}	b_{i1}	*Var[V]*
1.1696	-.1066	1.2544	-.1101	.067
(.1202)	*(.0164)*	*(.114)*	*(.0202)*	*(.1627)*

Table 3.6: Estimates obtained allowing for exogenous or endogenous total expenditure. Standard errors in parentheses.

3.6 Chapter Summary

In this Chapter I proposed an estimator for Engel curves in which budget shares are polynomials in the logarithms of total expenditure which accounts for the presence of two sources of endogeneity: measurement error on total expenditure and endogeneity of unobserved total expenditure. The estimator makes use of a standard control function assumption to derive consistent

estimates of the parameters of interest and follows a GMM procedure, so that it is readily implementable on available statistical softwares. The proposed method also depends crucially on the assumption we are willing to make on the form of the first stage equation.

The small sample properties of such an estimator have been analized and the results show a significant improvement in terms of reduced bias with respect to alternative (asymptotically biased) available estimators, especially when the extent of both endogeneity and measurement error is severe.

Chapter 4

Errors in Variables Models with Errors on Both Sides of the Equation

In this Chapter I study identification and estimation of general nonparametric regression models where both the dependent variable Y and a regressor X are mismeasured, and the measurement errors in Y and X are correlated. The resulting model is applied to demand estimation, where Y is quantity or expenditures demanded of some good or service, and X is total consumption expenditures on all goods.

4.1 Introduction

In most consumption data sets (e.g., the US Consumer Expenditure Survey or the UK Family Expenditure Survey), total consumption X is constructed as the sum of expenditures on individual goods, so by construction any measurement error in Y will also appear as a component of the measurement error in X. This peculiarity of consumption data invalidates the use of most of the non-linear errors in variables models in the literature. The identification result I provide is then of particular interest in the estimation of Engel curves. This result, however, is not confined to Engel curves. Similar problems may arise in profit, cost, or factor demand equations in production, and in autoregressive or other dynamic models where sources of measurement error are not independent over time.

The identification procedure employed will also allow me to distinguish and separately measure sources of error that are due to preference (or other structural or behavioral) heterogeneity from measurement error. This is important in applications because many policies may depend on the distribution of structural unobserved heterogeneity, but not on measurement error. For example, the effects of an income tax on aggregate demand or

savings depends in general on the complete distribution of income elasticities in the population. Most empirical analyses implicitly or explicitly attribute either none or all of estimated model errors to heterogeneity.

I assume that the measurement errors in X and Y are correlated with each other but are otherwise classical, and so are independent of the underlying true values of these variables, and hence are also independent of unobserved heterogeneity. The identification strategy is an extension of Schennach (2007), who provides identification of nonparametric regression models with a classically mismeasured continuous regressor using information on the conditional expectations of Y and X given an instrument Q. My extension adds a measurement error term to the Y equation, and the fact that it is classical imposes constraints on the second moment of Y given X that provide the additional identifying information required.

4.1.1 The Setup

I consider the following non-linear and non-separable model:

$$Y_i^* = H(X_i^*, U_i), \tag{4.1}$$

where i indexes observations, Y_i^* is a scalar random dependent variable, which is an unknown function $H(\cdot,\cdot)$ of a scalar random regressor X_i^*, and a random scalar or vector of unobservables U_i, which can be interpreted as a regression error or unobserved heterogeneity in the population. The extension to inclusion of other (observed) covariates is straightforward, so they are dropped for now. The primary goal will be identification (and later estimation) of $E[{Y^*}_i^k \mid X_i^*]$ and, by extension, identification of the conditional distribution of Y_i^* given X_i^*.

The difficulty will be that both Y_i^* and X_i^* are measured with error. In particular only Y_i and X_i are observed according to one of the following measurement error specifications:

$$\begin{cases} Y_i = Y_i^* + X^{*l} S_i, \\ X_i = X_i^* + W_i, \end{cases} \qquad \begin{cases} Y_i = Y_i^* + X^{*l} S_i, \\ X_i = X_i^* W_i, \end{cases} \tag{4.2}$$

with S_i and W_i being unobserved measurement errors that may be correlated with each other.

Identification of these models is required in order to account for the specific nature of Engle curves. Equation (4.1) may be interpreted as an Engel curve when Y_i^* is expenditure on one good, X_i^* is total expenditure and U_i is (un)observed preference heterogeneity. In this spirit the measurement error specifications employed encompass most of the specifications used in the literature. In particular $l = 1$ in the second specification corresponds to the classical measurement errors on the logarithm of total expenditure introduced in Chapter 3.

Similar models have been considered by Hausman, Newey, Ichimura, and Powell (1991) and Hausman, Newey, and Powell (1995) and Lewbel (1996), who provide identification for polynomials in levels or logarithms respectively, under the assumption of classical measurement errors in the logarithms of total expenditure. Another closely related work is that of Schennach (2007) who provide identification for general non-linear specifications while assuming additive measurement errors. Moreover her result does not allow for the presence of measurement errors on the dependent variable which are correlated with measurement errors on the covariate. The purpose of this Chapter is that of unifying these results and providing an identification result for general non-linear models under different specifications of the measurement error and explicitly allowing for the presence of correlated measurement errors on both sides of the equation. This is relevant in the spirit of my thesis since consistent estimates for very general specifications of Engel curves with measurement errors can not be obtained by applying Schennach (2007).

The interest in more general specifications of the function $H(\cdot,\cdot)$ follows from the nature of Engel curves. There is no clear economic reason to assume additive separability between total expenditure and the error term. In fact this poses severe restrictions on households behavior and the specification of Engel curves which are consistent with utility maximization often involves error terms which enters non-additively in the specification (see Blundell, Duncan, and Pendakur 1998). Furthermore recently a growing stream of research has focused on the specification of very flexible Engel curves specifications (Blundell, Chen, and Kristensen 2007, Lewbel and Pendakur 2009).

Identification is obtained by assuming the existence of a set of instruments Q_i such that:

$$X_i^* = m(Q_i) + V_i,$$

where the function m is unknown but is defined by $m(Q_i) = E\left(X_i^* \mid Q_i\right)$ and V_i is independent of Q_i. As discussed in the previous Chapter the independence assumption is common in the control function literature when dealing with non-linear models and is crucial to the results below, hence it will be mantained throughout the Chapter. As it is also common in the non-linear models literature I assume mean zero measurement errors which have the classical properties that they are mean zero with $S_i, W_i \perp Y_i^*, X_i^*, Q_i$. However these strong independence assumptions are required only if the object of interest is the entire conditional distribution of Y_i^* given X_i^*. In particular, if one is only interested in the conditional mean, then the assumption of mean independence between S_i, W_i and X_i^* is enough.

4.1.2 Methods

The result I provide is an extension of the identification result in Schennach (2007). In particular she obtains identification of the conditional mean of Y_i^*

given X_i^*, by exploiting properties of the Fourier transform of the observed conditional expectations of Y_i and X_iY_i given Q_i, under the assumption that the measurement error in Y_i^* is uncorrelated with W_i. Such transforms are not proper functions, in general, but more complicated objects called *generalized functions*, being the Fourier Transform of not absolutely integrable functions. A popular example of a generalized function is the Dirac's delta function ($\delta(\zeta)$), which is formally defined as the function such that:

$$\int \phi(\zeta)\delta(\zeta)d\zeta = \phi(0),$$

for any sufficiently smooth function $\phi(\zeta)$. It is easy to see that there exists no ordinary piecewise continuous function which satisfies the integral equation above[1], and the introduction of generalized function is required in order to deal with these objects.

However, for the purpose of the present Chapter it is important to note that a generalized function may always be decomposed into the sum of an ordinary component, that is a well behaved function, and a purely singular one, which may be seen as a linear combination of generalized derivatives of the Dirac's delta function (see Lighthill 1962 and the supplementary material in Schennach 2007 for details). Schennach (2007) studies identification of the conditional mean of Y_i^* given X_i^* and proves that its identification by knowledge of just the ordinary component of the Fourier transforms of the observed conditional moments.

The extension I provide builds on the fact that allowing for correlated measurement error on Y_i^* does not affect the conditional mean of Y_i given Q_i, but shifts the conditional expectaton of X_iY_i given Q_i by the term $E[S_iW_i|Q_i]$. Under the crucial identifying assumption that the covariance between S_i and W_i is uncorrelated with the instruments it is $E[S_iW_i|Q_i] = E[S_iW_i]$, meaning that the observed conditional cross-moment is shifted by a constant with respect to the instruments Q_i. This is important since the Fourier transform of a constant is a purely singular generalized function, implying that the ordinary component of the Fourier transform of $E[X_iY_i|Q_i]$ remains unchanged. Since identification relies on the ordinary components of the Fourier transform of the observed conditional moments considered, the conditional mean of interest is still identified. This argument may be further generalized to identify higher order conditional moments, that is $E[Y_i^k|X_i^*]$.

However, in most applications the objects of interest are conditional moments of the form $E[Y_i^{*k}|X_i^*]$. When dealing with Engel curves we are able to separately identify the effect on the observed Y_i of preferences (U_i) and measurement errors (S_i) and hence obtain identification of the conditional

[1]The Dirac's delta function may also be seen as the limit of a gaussian density as the variance approaches zero.

moments of interest. By extension the conditional conditional distribution of Y_i^* given X_i^* is then identified. This is accomplished by employing two different identification strategies that depend on the measurement error structure considered. The first strategy is based on the insight that any theory consistent Engel curve has the property that $H(0, U_i) \equiv 0$, while the second strategy makes use of the specific dependence structure between W_i and S_i implied by the definition of X_i and Y_i in the Engel curve framework.

Knowledge of the conditional distribution of Y_i^* given X_i^*, and hence of the distribution of the preference heterogeneity in the population is of particular interest in policy analysis. Think for instance at the effect on demand of the introduction of a tax cut which shifts households' consumption level from $\bar{x}^*$ to $\bar{x}_0^*$. This would not only affect households' conditional mean but it would in general alter the entire distribution of demand. While traditional estimation techniques would only allow the estimation of the policy effect on the mean of the distribution, knowledge of the distribution of U_i allows us to focus on the distributional effects of such policies. This information is of particular interest to the construction of important policy indicators, such as the proportion of households below some specified level of consumption under different policy regimes.

The identification results contained in this Chapter will be discussed according to the following plan. First, I will discuss identification of $E[Y_i^k|X_i^*]$ assuming the availability of a set of instruments Q_i. I will then consider identification of $E[Y_i^{*k}|X_i^*]$, which for the problem at hand can be interpreted as the object of interest if one were to work on Engel curves estimation. The proofs of these results will be given in Theorems 4.2 and 4.3 respectively. Section 4.4 finally proposes a consistent estimator for a particular specification of Engel curves.

4.2 Identification of the Conditional Distribution of the Observed Outcome

In the following I discuss identification of the model:

$$Y_i^* = H(X_i^*, U_i), \tag{4.3}$$
$$X_i^* = m(Q_i) + V_i, \tag{4.4}$$

where Q_i is a vector of instruments, V_i is a scalar random variable independent of Q_i and U_i is a vector of disturbances. The function $H(\cdot, \cdot)$ is a general, possibly nonlinear, function of X_i^* and U_i. The scalar random variable Y_i^* is unobserved, but we observe its mismeasured counterpart given by:

$$Y_i = Y_i^* + X_i^{*l} S_i, \tag{4.5}$$

for some non-negative integer l. The empirically most relevant frameworks would entail $l = 0$ or $l = 1$, but we consider identification for a general l for

sake of completeness. The generalization to $l > 0$ is required to deal with measurement errors whose variance increases with X_i^*. Empirical findings suggest that such a situation actually arises in the Engel curves framework as discussed in Section 4.3, but this could also hold in more general contexts.

The random variable X_i^* is also measured with error, with X_i satisfying either:

$$X_i = X_i^* + W_i, \qquad \text{with } E[W_i] = 0, \tag{4.6}$$

or

$$X_i = X_i^* W_i, \qquad \text{with } E[W_i] = 1. \tag{4.7}$$

Equations (4.6) and (4.7) specifically allow measurement error on the covariate X_i^* to enter additively or multiplicative, while retaining the property $E[X_i] = E[X_i^*]$. Let $\mu^k(x_i^*) = E[Y_i^k | X_i^*]$ be the k-th conditional moment of the observed random variable Y_i given X_i^*. The goal of this Section is to provide identification of $\mu^k(x_i^*)$ for $k = 1, \ldots, K$, by only knowledge of (Y_i, X_i, Q_i) where X_i is defined by either (4.6) or (4.7).

The setup here is similar to the one in Schennach (2007), the main difference being the presence of measurement error on the dependent variable which is explicitly allowed to be correlated with W_i. Moreover the focus here is not only on the first conditional moment but also on higher order moments. This is in contrast with the traditional literature on non-linear errors in variables models, which is tipically concerned with the identification of the conditional mean. Knowledge of higher order moments is required to provide identification for the entire conditional distibution of Y_i given X_i^*. The above model may be easily generalized to include correctly measured regressors in the specification of Y_i^* and the identification result below would still apply after a straightforward generalization of the identifying assumptions.

Let us assume the following:

Assumption 4.1. *The random variables Q_i, U_i, V_i, W_i and S_i are jointly i.i.d. and*

(i) $E[W_i^k | Q_i, V_i, U_i] = E[W_i^k]$ *for* $k = 1, \ldots, K$,

(ii) $E[S_i^k | Q_i, V_i, U_i] = E[S_i^k]$ *for* $k = 1, \ldots, K$,

(iii) V_i *is independent of* Q_i,

(iv) $E[W_i S_i | Q_i] = E[W_i S_i]$.

Assumption 4.1 is fairly standard in the literature on non-linear errors in variables models, the main difference being that, since I want to estimate the first K conditional moments, mean independence is in general not enough and (i) and (ii) need to hold for $k = 1, \ldots, K$. Moreover (iv) requires that the correlation between the measurement errors does not depend on the instruments Q_i, this is a crucial assumption for the identification result below.

Independence of the instruments Q_i from the errors (W_i, S_i) would satisfy this assumption, but (iv) is a considerably weaker requirement. Assumption 4.1 (iii) is also standard in the measurement error literature and is mainly required to deal with the non-linearities in the specification of $H(X_i^*, U_i)$.

Note that by Assumption 4.1[2]:

$$E[X_i|Q_i] = m(Q_i),$$

hence $m(Q_i)$ is non-parametrically identified and we may rewrite without loss of generality equation (4.4) as:

$$X_i^* = Z_i - \tilde{V}_i \tag{4.8}$$

in which we set $Z_i \equiv m(Q_i)$ and $\tilde{V}_i \equiv -V_i$. Following Newey (2001) and Schennach (2007) we will show that knowledge of the conditional moments $E[Y_i^k|Z_i]$, for $k = 1, \ldots, K$, and $E[X_iY_i|Z_i]$ is enough to identify $\mu^k(x_i^*)$ for $k = 1, \ldots, K$. In the remainder of the paper, for ease of notation, we will drop the subscript i when not needed.

It follows from Assumption 4.1 that:

$$\begin{aligned} E[Y^k|Z] &= E\left[\left[H(X^*,U) + X^{*l}S\right]^k |Z\right], \\ &= E\left[E\left[\left[H(X^*,U) + X^{*l}S\right]^k |X^*, Z\right] |Z\right], \\ &= E[\mu^k(x^*)|Z], \end{aligned} \tag{4.9}$$

and, if (4.6) holds, we have:

$$\begin{aligned} E[XY|Z] &= E\left[(X^* + W)\left[H(X^*,U) + X^{*l}S\right] |Z\right], \\ &= E[X^*H(X^*,U)|Z] + E[X^{*l}|Z]E[WS|Z], \\ &= E[x^*\mu^1(x^*)|Z] + E[X^{*l}|Z]E[WS]. \end{aligned} \tag{4.10}$$

With a similar argument it may be shown that, if (4.7) holds instead, we have:

$$E[XY|Z] = E[x^*\mu^1(x^*)|Z] + E[X^{*l+1}|Z]E[WS]. \tag{4.11}$$

The proof of identification of $\mu^k(x^*)$ is obtained by exploiting properties of the Fourier trasform of the conditional expectations above. In order to guarantee that well defined objects to deal with, the following assumption is needed:

Assumption 4.2. *$|\mu^k(x^*)|$, $|E[Y^k|Z]|$ and $|E[XY|Z]|$ are defined and bounded by polynomials for x^* and $z \in \mathbb{R}$ and for any $k = 1, \ldots, K$.*

[2]Note that Assumption 4.1 does not assume $E[V_i] = 0$, but equation (4.8) may always be rewritten so that $m(Q_i)$ includes a constant and hence $E[V_i] = 0$.

Assumption 4.2 essentially excludes specifications for $\mu^k(x^*)$ which rapidly approach infinity like the exponential function and is crucial for the following Lemma to hold:

Lemma 4.1. *Under Assumption 4.2, equations* (4.9)*,* (4.10) *and* (4.11) *are equivalent to*

$$\begin{aligned} \varepsilon_{y^k}(\zeta) &= \gamma_k(\zeta)\phi(\zeta), && (4.12)\\ \boldsymbol{i}\dot{\varepsilon}_{xy}(\zeta) &= \dot{\gamma}_1(\zeta)\phi(\zeta) + \lambda \boldsymbol{i}\psi(\zeta)\phi(\zeta), && (4.13) \end{aligned}$$

with $\boldsymbol{i} = \sqrt{-1}$*, overdots denote derivatives with respect to* z *and:*

$$\begin{aligned} \varepsilon_{y^k}(\zeta) &\equiv \int E[Y^k|Z=z]e^{\boldsymbol{i}\zeta z}dz, \qquad \gamma_k(\zeta) \equiv \int \mu^k(x^*)e^{\boldsymbol{i}\zeta x^*}dx^*,\\ \varepsilon_{xy}(\zeta) &\equiv \int E[XY|Z=z]e^{\boldsymbol{i}\zeta z}dz, \qquad \phi(\zeta) \equiv \int e^{\boldsymbol{i}\zeta v}dF(v), \end{aligned}$$

where $F(v)$ *is the cdf of* $\tilde{V}$*,* $\lambda = E[WS]$ *and* $\psi(\zeta) = \int x^{*l}e^{\boldsymbol{i}\zeta x^*}dx^*$ *if* (4.6) *holds or* $\psi(\zeta) = \int x^{*l+1}e^{\boldsymbol{i}\zeta x^*}dx^*$ *if* (4.7) *holds.*

Lemma 4.1, whose proof is given in the Appendix, is a generalization of Lemma 1 in Schennach (2007). The objects of interest in equations (4.12) (4.13) are the functions $\gamma_k(\zeta)$ for $k = 1, \ldots$, while $\phi(\zeta)$ is an unknown function that acts as a nuisance parameter for a fixed ζ. It is also important to note that while $\phi(\zeta)$, being the characteristic function of $\tilde{V}_i$, is a proper function, $\varepsilon_{xy}(\zeta)$, $\varepsilon_{y^k}(\zeta)$, $\gamma_k(\zeta)$ and $\psi(\zeta)$ are more abstract generalized functions. In this sense it is important to recall that the product of two generalized functions is not defined, so that (4.12) and (4.13) cannot manipulated as usual functions to get rid of the nuisance parameter $\phi(\zeta)$. Also note that the unknown quantities here are $\gamma_k(\zeta)$ and $\phi(\zeta)$, while $\psi(\zeta)$ is the Fourier transform of a power function, hence known to be equal to the generalized derivative of the Dirac's delta function. For a more detailed treatment of generalized functions see Lighthill (1962) or the supplementary material in Schennach (2007).

Assumption 4.3. *It is* $E[|\tilde{V}|] < \infty$ *and* $\phi(\zeta) \neq 0$ *for all* $\zeta \in \mathbb{R}$.

Assumption 4.4. *For each* $k = 1, \ldots, K$ *there exists a finite or infinite constant* $\bar{\zeta}_k$ *such that* $\gamma_k(\zeta) \neq 0$ *almost everywhere in* $[-\bar{\zeta}_k, \bar{\zeta}_k]$ *and* $\gamma_k(\zeta) = 0$ *for all* $|\zeta| > \bar{\zeta}_k$.

Assumptions 4.3 and 4.4 are the equivalent of Assumptions 2 and 3 in Schennach (2007) and are quite standard in the deconvolution literature. Since we are seeking non-parametric identification of $\gamma_k(\zeta)$ the characteristic function of $\tilde{V}_i$ needs to be non-vanishing, thus excluding uniform or triangular like distributions, while $\gamma_k(\zeta)$ needs to be either non-vanishing or

to vanish on an infinite interval. This is required in order for $\gamma_k(\zeta)$ to be fully non-parametrically identified. However, when $H(X^*, U)$ is parametrically specified this assumption may be relaxed and the information on a finite number of points of $\gamma_k(\zeta)$ is generally enough for identification. Furthermore Assumption 4.4 essentially only rules out specifications for $\gamma_k(\zeta)$ which exhibit sinusoidal behaviors, a situation which is not usually encountered in practice.

The following theorem states the main identification result:

Theorem 4.2. *Under Assumptions 4.1-4.4, the functions $\mu^k(x^*)$ for $k = 1, \ldots, K$ are non-parametrically identified. Also, if $\bar{\zeta}_1 > 0$ in Assumption 4.4 then:*

$$\mu^k(x^*) = (2\pi)^{-1} \int \gamma_k(\zeta) e^{-i\zeta x^*} d\zeta,$$

where

$$\gamma_k(\zeta) = \begin{cases} 0 & \text{if } \varepsilon_{y^k}(\zeta) = 0, \\ \varepsilon_{y^k}(\zeta)/\phi(\zeta) & \text{otherwise,} \end{cases} \tag{4.14}$$

$\phi(\zeta)$ is the characteristic function of $\tilde{V} \equiv -V$ given, for $|\zeta| < \bar{\zeta}_1$, by:

$$\phi(\zeta) = \exp\left(\int_0^\zeta \frac{i\varepsilon_{(z-x)y,o}(\zeta)}{\varepsilon_{y,o}(\zeta)} d\zeta \right), \tag{4.15}$$

and where $\varepsilon_{y,o}(\zeta)$ and $\varepsilon_{(z-x)y,o}(\zeta)$ denote the ordinary function components of $\varepsilon_y(\zeta) = \int E[Y|Z=z] e^{i\zeta z} dz$ and $\varepsilon_{(z-x)y}(\zeta) = \int E[(Z-X)Y|Z=z] e^{i\zeta z} dz$ respectively.

Theorem 4.2 generalizes Theorem 1 in Schennach (2007). Proof is given in the Appendix and it follows by noting that every generalized function can be written as the sum of an ordinary function and a linear combination of generalized derivatives of the Dirac's delta function, which correspond to the purely singular component (see Schennach 2007). The last term in (4.13) is a purely singular generalized function and then it only affects the singular component of $\varepsilon_{xy}(\zeta)$. Since Schennach (2007) proved that $\phi(\zeta)$ is identified by knowledge of the ordinary components of $\varepsilon_y(\zeta)$ and $\varepsilon_{xy}(\zeta)$, allowing for W and S to be correlated does not alter the identification of $\phi(\zeta)$. The function of interest $\gamma_k(\zeta)$ is then obtained from equation (4.12) as in (4.14), and its inverse Fourier transform gives $\mu^k(x^*)$.

Theorem 4.2 provides non-parametric identification for any conditional moment of Y given X^*. This, under the assumption that Y given X^* has a well-defined moment generating function, ensures the non-parametric identification of the entire distribution. Note that, if $\bar{\zeta}_1 = \infty$, equation (4.15) gives the expression for the characteristic function of $\tilde{V}_i$ over its whole domain. The characteristic function of the unobserved X^* ($\phi_{X^*}(\zeta)$), may then be obtained from equation (4.8) as:

$$\phi_{X^*}(\zeta) = \frac{\phi_Z(\zeta)}{\phi(\zeta)},$$

where $\phi_Z(\zeta)$ is the characteristic function of Z. The assumption of $\bar{\zeta}_1 = \infty$ encompasses most of the empirically relevant specifications for the conditional mean, and then it is not very restrictive. Also note that in the proof of Theorem 4.2 we only considered $\bar{\zeta}_k > 0$ since the case $\bar{\zeta}_k = 0$ only occours if $\mu^k(x^*)$ is a polynomial in X^*, a specification which has already been shown to be identified by Hausman, Newey, Ichimura, and Powell (1991).

4.3 Identification of Engel curves

The previous Section established a set of assumptions under which $\mu^k(x^*)$ is identified for every k. However, in applications, the objects of interest are usually the conditional moments of the unobserved Y^*, that is $\omega^k(x^*) = E[Y^{*k}|X^*]$. The difference between $\omega^k(x^*)$ and $\mu^k(x^*)$ is due to the presence of measurement error in Y^*. While variability in $\mu^k(x^*)$ may be driven by both U and S, $\omega^k(x^*)$ is solely determined by the distribution of U, hence, providing a way to disentangle the effects on Y^* of the structural error U from those of measurement error S would allow the identification of $\omega^k(x^*)$. In this Section we derive a set of assumptions under which $\omega^k(X^*)$, for $k = 1, \ldots, K$, is identified.

To this end, exploiting the additive nature of measurement error in Y^*, we may rewrite the k-th conditional moment of the observed Y as:

$$\begin{aligned} \mu_k(x^*) &= \sum_{j=0}^{k} \binom{k}{j} E[H^j(X^*, U) X^{*l(k-j)} S^{k-j} | X_i^*] \\ &= \sum_{j=0}^{k} \binom{k}{j} \omega^j(x^*) x^{*l(k-j)} E[S^{k-j}], \end{aligned}$$

where the second equality holds because of (ii) in Assumption 4.1. After noting that $\mu^0(x^*) = \omega^0(x^*) = 1$ and by solving for $\omega^k(x^*)$ we obtain:

$$\omega^k(x^*) = \mu^k(x^*) - \sum_{j=0}^{k-1} \binom{k}{j} \omega^j(x^*) x^{*l(k-j)} E[S^{k-j}]. \tag{4.16}$$

Equation (4.16) shows that $\omega^k(x^*)$ is identified from knowledge of $\mu^k(x^*)$, $\omega^j(x^*)$ for $j = 1, \ldots, k-1$ and $E[S^j]$ for $j = 1, \ldots, k$. Since $\mu^k(x^*)$ has already been shown to be identified from Theorem 4.2, we need to consider moments of the distribution of S. These are in general not identified without additional assumptions on the dependence structure between W and S which needs to be treated on a case by case basis.

When dealing with Engel curves, however, the nature of the variables involved implies a specific dependence structure between measurement errors which may be exploited to obtain identification of moments of S. Here we

consider two different measurement error specifications for these kind of models which imply classical measurement errors either in the levels or in the logarithms of total expenditure.

Let Y^* and X^* be unobserved expenditure on a single good and total expenditure respectively. Suppose the observed expenditure on one good is defined as $Y = Y^* + S$, then by the definition of total expenditure it is $X = X^* + W$ and W is a classical measurement error on levels of expenditure with

$$W = S + \tilde{S}.$$

An alternative more used specification (see Lewbel 1996) implies $Y = Y^* + X^*S$, which means $X = X^*W$, so that $\log W$ is a classical measurement error on the logarithms of total expenditures and by the summing up property it is

$$W = S + (1 + \tilde{S}). \tag{4.17}$$

Theorem 4.2 states that, under suitable conditions, $\mu^k(x^*)$ is identified for every k in both of the above measurement error specifications. In particular the former case correspond to the model in (4.3), (4.4), (4.5) and (4.6) with $l = 0$, while the latter may be seen as (4.3), (4.4), (4.5) and (4.7) with $l = 1$. This last setup is the one that has received most of the attention in the literature, see for instance Hausman, Newey, and Powell (1995), Lewbel (1996) and Newey (2001).

Let us start by considering classical measurement errors on levels of total expenditure. Identification of moments of S in this framework follows from the restrictions imposed by utility maximization. Every theory consistent Engel curve, derived from any random utility function, must satisfy $H(0, U) \equiv 0$. This is because when households are given a null amount of income, disregarding of how heterogeneous in preferences they are, the distribution of expenditures on one single good will be degenerate in zero. An example is the very popular AIDS of Deaton and Muellbauer (1980), which implies Engel curves in which budget shares are a linear function of the logarithm of total expenditure. This translates into the fact that $\omega^k(0) = 0$ for any $k \geq 1$ and we may rewrite equation (4.16) to get:

$$\begin{aligned} E[S^k] &= \mu^k(0) - \sum_{j=1}^{k} \omega^j(0) E[S^{k-j}], \\ &= \mu^k(0). \end{aligned} \tag{4.18}$$

Hence substitution of (4.18) in (4.16) yields:

$$\omega^k(x^*) = \mu^k(x^*) - \sum_{j=0}^{k-1} \binom{k}{j} \omega^j(x^*) \mu^{k-j}(0). \tag{4.19}$$

Equation (4.19) may then be used to iteratively compute $\omega^k(x^*)$ for $k = 1, \ldots, K$. By extension, under the assumption of the existence of a well-defined moment generating function for $f_{Y^*|X^*}(y^*|x^*)$, the conditional distribution of Y^* given X^* is identified.

Now consider classical measurement errors in the logarithms of total expenditure. The above argument does not provide identification of the distribution of S anymore, since measurement errors in Y^* depend on X^*. Here, however, we may exploit the particular dependence structure between W and S and note that, under the assumption that $\tilde{S}$ is mean independent of S in equation (4.17), $E[WS] = E[S^2]$. More generally, using (4.17), it is:

$$\begin{aligned} E[W^k S] &= E[[S + (1 + \tilde{S})]^k S], \\ &= E\left[S \sum_{j=0}^{k} \binom{k}{j} S^j (1 + \tilde{S})^{k-j} \right], \\ &= \sum_{j=0}^{k} \binom{k}{j} E[S^{j+1}] E[(1 + \tilde{S})^{k-j}], \end{aligned}$$

which means:

$$E[S^{k+1}] = E[W^k S] - \sum_{j=0}^{k-1} \binom{k}{j} E[S^{j+1}] E[(1 + \tilde{S})^{k-j}]. \tag{4.20}$$

The moments $E[(1 + \tilde{S})^k]$ are obtained with a similar argument:

$$E[(1 + \tilde{S})^k] = E[W^k] - \sum_{j=1}^{k} \binom{k}{j} E[S^j] E[(1 + \tilde{S})^{k-j}]. \tag{4.21}$$

Equations (4.20) and (4.21) show that the k-th moment of the random variable S is identified by knowledge of moments $E[W^j S]$ and $E[W^j]$ for $j = 1, \ldots, k-1$. The following Theorem establishes formal identification for these quantities.

Theorem 4.3. *Let Assumptions* (4.1)-(4.4) *and equations* (4.3), (4.4) *and* (4.6) *hold. Let the first K moments of X exist finite with $E[X^{*k}] \neq 0$ and $E[X^{*k}|Z] \neq 0$ for every $k = 1, \ldots, K$, then the first K moments of W are identified and:*

$$E[W^k] = \frac{E[X^k]}{\sum_{j=0}^{k} \binom{k}{j} (-\boldsymbol{i})^{k-j} E[Z^j] \phi^{(k-j)}(0)}. \tag{4.22}$$

Furthermore if $\bar{\zeta}_1 = \infty$ in Assumption 4.4 then the moments $E[W^k S]$ for $k = 1, \ldots, K - l$ are also identified and:

$$E[W^k S] = \frac{E[X^k Y|Z] - (2\pi)^{-1} E[W^k] \int (-\boldsymbol{i})^k \gamma_1^{(k)}(\zeta) \phi(\zeta) e^{-\boldsymbol{i}\zeta z} d\zeta}{\sum_{j=0}^{k+l} \binom{k+l}{j} z^j (-\boldsymbol{i})^{k+l-j} \phi^{(k+l-j)}(0)}. \tag{4.23}$$

where $\gamma_1^{(k)}(\zeta)$ *is the k-th derivative of* $\gamma_1(\zeta)$ *as defined in equation* (4.14)*, while* $\phi(\zeta)$ *is as in* (4.15)*.*

The requirement for the first K marginal and conditional moments of X^* to be non-zero may seem quite restrictive. However in our setting, being both Y^* and X^* positive random variables, this is always satisfied if X^* is non-degenerate. This is because we are considering raw moments and not central ones, hence we are not for instance ruling out symmetric distributions, for which the third central moment would be zero. Futhermore, the assumption of $\bar{\zeta}_1 = \infty$, though not always satisfied, covers all the empirically relevant frameworks, as discussed in Section 4.2.

Theorem 4.3 proves the identification of $E[W^k S]$ and $E[W^k]$, for $k = 1, \ldots, K$, from knowledge of $\gamma_1(\zeta)$ and $\phi(\zeta)$, which areknown to be identified under the assumptions of Theorem 4.2, $E[X^k Y|Z]$, for $k = 1, \ldots, K$, and observable moments of X and Z. This together with equations (4.20) and (4.21) shows identification of the moments of S, and hence of the conditional moments on interest $\omega^k(x^*)$. Under the additional assumption of the existence of a moment generating function for the conditional distribution function $f_{Y^*|X^*}(y^*|x^*)$, this implies identification of the whole conditional distribution of interest.

4.4 Estimation

I now turn to the estimation of the Engel curve defined by:

$$Y^* = \beta_0(X^*U) + \beta_1(X^*U)\log(X^*U)$$

where as usual Y^* is unobserved expenditure on one good, X^* is unobserved total expenditure while U represents unobserved preferences which enter household's utility function. This specification is known to be consistent with utility maximization and yields shape invariant Engel curves, see Blundell, Duncan, and Pendakur (1998). Functional forms in which the error term enters additively is indeed shown to generate unplausible restrictions on the behavior of the system of demand from which the Engel curves are generated, see Brown and Walker (1989) and Lewbel (2001) for a discussion on the topic.

The measurement error structure imposed is the one implied by equations (4.5) and (4.7) with $l = 1$. This is the usual classical measurement error on the logarithms of total expenditure employed in the literature. This is done in order to allow measurement error to be correlated with the level of expenditures and the variance of measurement errors to increase with total expenditure, a feature usually encountered in the data.

When an instrument is available Theorems 4.2 and 4.3 ensure identifiability of the conditional distribution of Y^* given X^*. In particular any

conditional moment may be written as:

$$\omega^k(x^*) = \sum_{j=0}^{k} \alpha_{jk} x^{*k} \log^j x^*, \tag{4.24}$$

where $\alpha_{jk} = \binom{k}{j} E[(\beta_0 U + \beta_1 U \log U)^{k-j} (\beta_1 U)^j]$ for $k = 1, \ldots, K$ and $j = 0, \ldots, k$. The coefficients α_{jk} are then all identified from the data. An interesting implication of equation (4.24) is that $\alpha_{kk} = \beta_1^k E[U^k]$, hence, since the mean of U is normalized to the unity, it is $\beta_1 = \alpha_{11}$. Futhermore all the moments of U are obtained from consistent estimates of the highest order coefficient of the conditional moment $\omega^k(x^*)$ for $k \geq 1$, and in particular an estimate of the K-th moment of U, γ_K, is

$$\hat{\gamma}_K = \frac{\hat{\alpha}_{KK}}{\hat{\alpha}_{11}^K}, \tag{4.25}$$

where $\hat{\alpha}_{kk}$ is a consitent estimate of α_{kk}. Also note that, since $\alpha_{01} = \beta_0 + \beta_1 E[U \log U]$, a consistent estimator for β_0 is:

$$\hat{\beta}_0 = \hat{\alpha}_{01} - \hat{\alpha}_{11} \int u \log u d\hat{F}_U(u),$$

where $\hat{F}_U(u)$ is a consistent estimate of the distribution function of U, which is obtained from (4.25) for $k = 1, \ldots$, under the assumption of a well-defined moment generating function for $F_U(u)$. Summing up the above results we have that consistent estimates for the parameters of interest β_0 and β_1 are obtained from estimates of the coefficients α_{jk} for $k = 1, \ldots$ and $j = 0, \ldots, k$.

The estimation of the distribution function $F_U(u)$ involves dealing with high order moments, which are widely known for providing not very accurate estimates, due to the increased variability in the estimation. It may then be preferable to approximate the density function $f_U(u)$ to an arbitrary degree of precision, as discussed in Gallant and Nychka (1987), by:

$$f_U(u) \approx \Phi(u) \left[\sum_{j=0}^{P} \gamma_j H_j(u) \right]^2, \tag{4.26}$$

where $H_j(\cdot)$ are Hermite polynomials and $\Phi(\cdot)$ is the standard normal density. Note that considering moments up to the K-th order implies the estimation of $(K+1)(K+2)/2-1$ parameters α_{jk} which may be directly related to the $P+1$ parameters of interest, that is[3] β_0, β_1 and $\gamma_2, \ldots, \gamma_P$. Hence all the parameters of interest are identified if $P \leq (K+1)(K+2)/2 - 2$. Thus if we consider, for instance, the first 2 observed conditional moments, P needs not to be larger than 4.

[3]Note that normalization of $f_U(u)$ implies $\gamma_0 = 1$ and $\gamma_1 = 0$.

A consistent estimator for the coefficients α_{jk}, for $k = 1, 2$ follows from the moment conditions implied by equations (4.9), (4.11) and (4.4):

$$E[q(z)(Y^k - H_k(m(z) - v, v; \theta))] = 0 \qquad \text{for } k = 1, 2 \tag{4.27}$$
$$E[q(z)(XY - \tilde{H}(m(z) - v, v; \theta))] = 0 \tag{4.28}$$
$$E[q(z)(X - m(z; \gamma))] = 0. \tag{4.29}$$

where $q(z)$ is a vector of instruments and

$$\begin{aligned} H_1(x^*, v; \theta) &= \int \alpha_{01} x^* + \alpha_{11} x^* \log x^* dF(v) \\ H_2(x^*, v; \theta) &= \int (\alpha_{02} + \rho) x^{*2} + \alpha_{12} x^{*2} \log x^* + \alpha_{22} x^{*2} \log^2 x^* dF(v) \\ \tilde{H}(x^*, v; \theta) &= \int x^* (\alpha_{11} x^* + \alpha_{21} x^* \log x^*) dF(v) + \rho \int x^{*2} dF(v) \end{aligned}$$

with $\rho = E[S^2]$ and $dF(v)$ as defined by (4.26). The parameters of interest are γ and $\theta = (\alpha, \rho)$ and $dF(v)$. Note that γ could also be infinite dimensional. Since the function $dF(v)$ is unknown and has to be estimated a consistent estimator is the sieve minimum distance estimator of Ai and Chen (2003).

The estimation is carried out in two steps:

1. First estimate γ by any parametric or non-parametric regression of X on the instruments Z to obtain $\hat{m}(z) = m(z; \hat{\gamma})$.

2. Estimate θ and $dF(v)$ by applying a sieve minimum distance estimator after plugging-in $\hat{m}(z)$ in the moment conditions (4.27) and (4.28).

The sieve minimum distance estimator is adopted since it allows to consistently estimate both θ and $dF(v)$ by approximating the unknown function by means of a sequence of known basis functions which gets larger with sample size. The basis function adopted is then a T-th order polynomial defined as in equation (4.26), where T increases with sample size.

The main drawback in the applicability of this procedure is that the integrals involved in the specification of the moment conditions are difficult to evaluate and hence need to be computed numerically. This is analogous to a simulated moments estimator as described by McFadden (1989) and Newey (2001), in which a consistent estimate of the residuals is computed by averaging the integrals over several random draws from a standard normal density, exploiting the nature of the basis function employed.

Under mild regularity conditions Ai and Chen (2003) show that the resulting estimator for $dF(v)$ converges at a rate faster than $n^{-1/4}$, while $\hat{\theta}$ is both $\sqrt{n}$ consistent and asymptotically normal. Note that if $\hat{\gamma}$ is a non-parametric estimator then the rate of convergence of $\hat{\theta}$ is slower than $\sqrt{n}$, while it is unchanged if γ is parametrically estimated.

Once that consistent estimates for α_{jk} are obtained it is straightforward to derive a consistent estimator for β_0, β_1 and $\gamma = (\gamma_2, \ldots, \gamma_P)$ from equations (4.24) and (4.26)

4.5 Chapter Summary

In this Chapter I studied identification of a general non-linear errors in variables with correlated measurement errors on both sides of the equation. Identification of the conditional moments of the observed dependent variable was obtained in the spirit of Schennach (2007). I achieved identification of the entire conditional distribution of interest of the unobserved Y_i^* given X_i^* exploiting the particular dependence structure implied by dealing with Engel curves.

A sieve minimum distance estimator was proposed to consistently estimate moments of the conditional distribution of interest, and by extension the whole distribution. The properties of the estimation procedure, and in particular the asymptotic theory for the estimator proposed, are not considered in this thesis, and their discussion is left to future research.

Chapter 5

Estimating Features of the Distribution of Consumption from Expenditure Data

The objective of this Chapter is that of providing a framework which allows identification and estimation of the distribution of consumption from knowledge of expenditure and the number of purchases. This is achieved by explicitly modeling the household purchase process.

5.1 Introduction

One of the main objectives when collecting data about households' consumption is the analysis of individuals well-being. Traditionally this analysis is carried out by looking at statistics based on individual's consumption or expenditure. In most of the econometric literature these two terms are used interchangeably, but consumption and expenditure represent two conceptually different quantities. The former is essentially a continuous process which takes place at each moment in time, since a commodity is generally consumed by individuals over a period of time, while expenditure is a discrete process which occurs on a finite number of occasions.

Think for instance at an individual buying a car. Then she is going to *consume* a fraction of the car's value each day for several years, while the *expenditure* occasion only took place when she actually purchased the car and is zero in every other moment in time. Therefore consumption and expenditure differ because of the general indivisibility of commodities and the presence of storage costs. If storage costs were zero and commodities infinitely divisible, then households would purchase only what they would immediately consume, thus expenditure would be equal to consumption, but in pratice this is often not the case.

Consumer surveys, such as the US *Consumer Expenditure Survey* or the

UK *Family Expenditure Survey* collect records of households' expenditures over a relatively short time span, such as a month or weeks. Due to the short time span the difference between consumption and expenditure may be substantial. In the above example for instance we could either record the purchase occasion or no purchase occasions. It is easy to see that interpreting the observed expenditure as consumption may be misleading. Furthermore the discrepancy between these two quantities is likely to increase as the frequency of purchase decreases.

While it is clear that the mean level of expenditure, averaging over households, is an unbiased estimator of the mean level of consumption, the effect of considering expenditure instead of consumption when dealing with higher order moments depends on the data generating process. For example, looking at higher order moments is required when assessing inequality and well-being.

In the following I will develop a framework which allows us to estimate the distribution of consumption from expenditure data by modeling purchase occasions. This is done exploiting additional information given by the number of observed purchases. Such information is available when considering data collected through diary surveys, in which households are given a diary on which to record every single purchase over a one week time span. Due to this short time window the difference between expenditure and consumption may be severe, and a procedure which corrects for this discrepances may be required.

5.2 The Model

Suppose we observe expenditure data over a one-week time span (I). Let S_h be the observed total expenditure over this period, while N_h is the number of purchases. The subscript h will refer to the generic household throughout. Denote by C_h the unobserved level of consumption over I. The goal is that of recovering the distribution of consumption C_h from knowledge of S_h and N_h.

The random variables defined in what follows will refer to micro-units, in most empirical applications representing households. Whenever possible, subscripts will be omitted from all expressions to simplify notation. Suppose there exists a time interval $T = [t_1, t_2]$, where I is a randomly chosen subset of T, such that expenditure over T equals consumption over T. Let V be the number of purchases that would be observed over T, while $\Delta^T = t_2 - t_1$ is the length of the time interval. In what follows the length of I will be normalized to one. For ease of exposition it is assumed that every individual is a consumer, that is $V > 0$. This assumption will be relaxed later on by allowing for individuals whose consumption level is zero.

Assumption 5.1. *Times to the next purchase are independent and exponentially distributed with mean* λ^{-1}.

The independence assumption is essentially requiring that the time needed to consume some commodities is independent from the time spent consuming the previous ones. The fact that the distribution of times to the next purchase are identically distributed is also not of major concern since we are considering relatively short time span, so that household preferences may be assumed sufficiently stable over time. Note that λ is allowed to be household specific, hence allowing for heterogeneity across households. Under Assumption 5.1 the observed number of purchases is a Poisson Process with intensity λ, for which there is:

$$Pr[N = n] = e^{-\lambda}\frac{\lambda^n}{n!}. \tag{5.1}$$

The expected number of purchases over any unit time interval is $E[N] = \lambda$. Also, since I is randomly chosen over the interval T, the expected number of purchases over I must be equal to V/Δ^T, so that:

$$\lambda = \frac{V}{\Delta^T}.$$

Now let S^T be the unobserved level of expenditure over T and let the following Assumption hold:

Assumption 5.2. *Expenditure is uniformly distributed across all purchase occurences.*

Assumption 5.2 allows us to express the unobserved expenditure over the hypothetical time interval T in terms of the observed expenditure over I, and in particular:

$$S^T = \frac{V}{N}S. \tag{5.2}$$

Similarly, since consumption is assumed to be uniformly distributed over T, the level of consumption over the interval T, denoted by C^T, may be related to the level of consumption over the observed time span I. It follows that, by the definition of T, we have:

$$C^T = \Delta^T C. \tag{5.3}$$

Also from the definition of T we have $C^T = S^T$ and substituting equations (5.2) and (5.3) into this equation yields:

$$S = \frac{N}{\lambda}C. \tag{5.4}$$

Equation (5.4) gives an expression for the unobserved C in terms of the observed S, N and the unknown λ. If λ was known, we could easily solve

equation (5.4) for C to back up the value of consumption from the observed pair (N, S). However neither V or Δ^T are observable and thus λ is not observed.

Now suppose that λ is distributed according to some probability density function $f(\lambda)$.

Assumption 5.3. *Let C be independent of λ.*

Assumption 5.3, which could be straightforwardly generalized to a conditional indipendence assumption by rewriting any expectation below as conditional on a set of covariates, is satisfied if households' utility functions were separable in consumption and frequency of purchase. This would be the case if the consumer decision is carried out in two steps: first she chooses the desired level of consumption and then, based on variables such as the distance to the stores, the frequency of purchases is chosen. Under Assumption 5.3 the observable moments of S may be written as:

$$\begin{aligned} E[S^k] &= E\left[E\left[\left(\frac{N}{\lambda}\right)^k C^k|\lambda\right]\right] \\ &= E\left[\frac{1}{\lambda^k}E\left[N^k|\lambda\right]E[C^k|\lambda]\right] \\ &= E\left[\frac{\phi^k(\lambda)}{\lambda^k}\right]E[C^k], \end{aligned}$$

where $\phi^k(x)$ is the k-th moment of a Poisson random variable with mean x, which is known, and the last expectation is comupted with respect to the distribution of λ. Therefore the moments of the distribution of consumption are identified up to knowledge of the distribution of λ and are given by:

$$E[C^k] = \frac{E[S^k]}{E\left[\frac{\phi^k(\lambda)}{\lambda^k}\right]} \qquad \text{for } k = 1, \ldots \tag{5.5}$$

By extension the entire distribution of consumption is then identified if the unknown distribution of consumption has a well defined moment generating function. An interesting thing to note about equation (5.5), by setting $k = 1$, is that the mean of S is equal to:

$$\begin{aligned} E[C] &= \frac{E[S]}{E\left[\frac{\phi^1(\lambda)}{\lambda}\right]}, \\ &= E[S], \end{aligned}$$

hence showing that the mean of expenditure equals the mean of consumption in this framework.

Now let us relax the assumption of the support of λ being the positive real line and suppose there exist a non-empty set of non-consumers, that is

households whose consumption is zero. Let these households be the elements of the set defined by $\Upsilon = \{h : C_h = 0\}$. Equation (5.4) shows that observing zero expenditure for a particular household may be the result of $N_h = 0$, which may happen even if $h \notin \Upsilon$, or if $h \in \Upsilon$. In other words the distribution of λ is not in general continuous on the positive real line, but it is a mixed distribution with a positive probability mass in zero. The above results do still apply conditional on λ being positive but, in order to identify the entire distribution of λ the probability of λ being zero, that is $Pr(h \in \Upsilon)$, needs to be identified. This is done by noting that

$$\begin{aligned} Pr(N_n > 0) &= Pr(N_h > 0 | h \in \Upsilon) Pr(h \in \Upsilon) + \\ &+ Pr(N_h > 0 | h \notin \Upsilon) Pr(h \notin \Upsilon), \\ &= Pr(N_h > 0 | h \notin \Upsilon) Pr(h \notin \Upsilon), \end{aligned}$$

where the second equality follows from the fact that $Pr(N_h > 0 | h \in \Upsilon) = 0$. Therefore the proportion of non-consumers is identified by:

$$Pr(h \notin \Upsilon) = \frac{Pr(N_h > 0)}{Pr(N_h > 0 | h \notin \Upsilon)}. \tag{5.6}$$

5.3 Estimation

I now turn to the estimation of the model discussed in Section 5.2. Equation (5.5) states that every moment of the unobserved distribution of consumption can be expressed as the ratio of observed moments of expenditure and the expected value of a known function of λ. Any moment of S is then known up to knowledge of the distribution of λ. In what follows I derive a maximum likelihood estimator for such a distribution.

Suppose λ has a probability density function which belongs to the some parametric familiy of distributions: $f(\lambda; \theta)$, with support on the positive real line. Then equation (5.1) implies that:

$$Pr[N = n] = \int e^{-\lambda} \frac{\lambda^n}{n!} f(\lambda; \theta) d\lambda. \tag{5.7}$$

A maximum likelihood estimator for θ is then obtained from equation (5.7) and is given by:

$$\hat{\theta} = \arg\max_{\theta} \prod_{h=1}^{H} \int e^{-\lambda} \frac{\lambda^{n_h}}{n_h!} f(\lambda; \theta) d\lambda,$$

where H is the number of households. If λ is Gamma distributed then a closed form expression for the integral is available, though in general the integral involved in the estimation can be computed numerically. Note that the discrete nature of N, hence the fact that it contains information only

on a finite number of mass points, does not allow the estimation of $f(\lambda;\theta)$ non-parametrically. If the dimension of the parameter vector θ is less than the number of non-zero mass points in the support of N, it is possible to test for the parametric family of distribution adopted.

If the distribution of λ is allowed to have a non-zero probability mass in zero then the above estimator is inconsistent due to the presence of non-consumers. The distribution function of λ may then be written as

$$\tilde{f}(\lambda) = \begin{cases} 0 & \text{if } h \in \Upsilon, \\ f(\lambda;\theta) & \text{if } h \notin \Upsilon. \end{cases}$$

Therefore, since households for which $N_h > 0$ are consumers by definition, using a similar argument and considering only positive values of N_h a maximum likelihood estimator for θ is:

$$\hat{\theta} = \arg\max_{\theta} \prod_{h=1}^{H} \frac{Pr[N_h = n_h]}{Pr[N_h > 0]}.$$

This, if $f(\lambda;\theta)$ is Gamma distributed, is equivalent to[1]:

$$\hat{\theta} = \arg\max_{\theta} \prod_{h=1}^{H} \frac{\beta^{\alpha}}{(\beta+1)^{\alpha} - \beta^{\alpha}} \frac{\Gamma(\alpha + n_h)}{n_h! \Gamma(\alpha)} \frac{1}{(\beta+1)^{n_h}}, \tag{5.8}$$

where $\theta = (\alpha, \beta)$. Once the distribution of λ is known, the proportion of non-consumers is obtained from equation (5.6) and, in the case of λ Gamma distributed is:

$$\hat{Pr}(h \notin \Upsilon) = \frac{\hat{Pr}(N_h > 0)}{1 - \left(\frac{\hat{\beta}}{\hat{\beta}+1}\right)^{\hat{\alpha}}}. \tag{5.9}$$

Knowledge of the distribution of λ is enough to obtain consistent estimates of moments of the unobserved distribution of consumption. In particular the denominator in equation (5.5) is consistently estimated by:

$$\hat{E}\left[\frac{\phi^k(\lambda)}{\lambda^k}\right] = \int \frac{\phi^k(\lambda)}{\lambda^k} f(\lambda;\hat{\theta}) d\lambda.$$

In the Gamma case for $k = 2$, for instance, using the fact that $\phi^2(\lambda) = \lambda(\lambda + 1)$, the above integral reduces to $1 + \beta/(\alpha - 1)$. This implies, together with equation (5.5), that an estimator for the second moment of consumption ($\hat{E}[C]^2$) is:

$$\hat{E}[C]^2 = \frac{\hat{E}[S]^2}{1 + \frac{\hat{\beta}}{\hat{\alpha}-1}}, \tag{5.10}$$

[1] Here I am using results on the characterization of conjugate distributions, which imply that the integral in equation (5.7) when λ is Gamma distributed is again a Gamma density with different parameters.

hence providing a closed form expression for the variance of expenditure. Note that, since λ is non-negative, the correction term for the variance of expenditure, i.e. $E[1 + 1/\lambda]$, is greater than one, implying that variance of expenditure is an upward biased estimate of the variance of consumption.

The above procedure allows the estimation of all the moments of the distribution of consumption. As discussed in Section 5.2, under the assumption of the existence of a well-defined moment generating function for such a distribution, this means that the entire distribution of consumption can be estimated. Its knowledge is of practical interest since many economic indicators of interest, such as the Gini coefficient or the proportion of households whose consumption level is lower than a certain threshold, are based on the entire distribution of consumption. However, without any additional restriction on its behavior, estimating the entire distribution entails the estimation of an infinite number of moments.

To overcome this difficulty one could assume that consumption is lognormally distributed. Such an assumption is not particularly unrealistic, since severeal empirical findings show that at least the aggregate distribution of non-durable consumption is close to the lognormal (see Battistin, Blundell, and Lewbel 2009). In this case the distribution of consumption is entirely characterized by the first two moments, that is $C \sim logN(\mu, \sigma^2)$ with

$$\begin{aligned} \mu &= \log(E[C]) - \frac{1}{2}\log\left(1 + \frac{Var[C]}{E[C]^2}\right), \\ \sigma^2 &= \log\left(1 + \frac{Var[C]}{E[C]^2}\right). \end{aligned}$$

More generally, let us assume C to be randomly drawn from a distribution which is a member of a sufficiently general class of parametric distributions $f_C(c;\eta)$. Then one could jointly estimate the parameters of both the unobserved distribution of λ and of C via maximum likelihood. Exploiting equations (5.1) and (5.4) and Assumption 5.3 we could express the distribution of the observed couple (C, N) as:

$$f_{S,N}(s,n) = \int f_C\left(\frac{s\lambda}{n};\eta\right) f_{N|\lambda}(n|\lambda) f_\lambda(\lambda;\theta) d\lambda. \tag{5.11}$$

The maximum likelihood estimator for the parameters of interest η and θ is then obtained by maximizing, with respect to the parameters, the above integral.

An alternative and more flexible procedure to estimate the distribution of C makes use of deconvolution techniques. To this end we could rewrite equation (5.4) as:

$$\log S = \log N/\lambda + \log C.$$

Note that the distribution of λ is known once consistent estimates of θ are available. Therefore, under Assumption 5.3, the distribution of $T = \log N/\lambda$

is knonwn and independent of C. Standard deconvolution estimators then allow the estimation of the whole unknown distribution of C[2].

Knowledge of the entire distribution of consumption is of interest when estimating inequality measures that do not depend on a finite number of moments. Examples of these measures are the Gini coefficient or the proportion of poors, defined as the share of households whose consumption level is below a certain threshold. Indeed, once the entire distribution of C is known, the proportion of poors is straightforwardly computed from its cumulative density function, while the Gini coefficient is obtained as:

$$G(f_C(c)) = 1 - 2\int_0^1 \frac{\int_{-\infty}^{q_C(x)} c f_C(c)dc}{\int_{-\infty}^{\infty} c f_C(c)dc} dx,$$

where $q_C(x)$ is the quantile of order x of consumption and $f_C(c)$ is the distribution of consumption. Note that for a lognormal distribution the Gini coefficient is easily computed as $G(f_C(c)) = 2\Phi(\sqrt{\sigma^2/2}) - 1$, where $\Phi(\cdot)$ is the cumulative density function of a standard normal distribution.

5.4 Application to CEX Data

In this Section I apply the estimation procedure described above to recover the unobserved distribution of consumption. I use data from the Diary survey of the US *Consumer Expenditure Survey* for the years 1982 to 2003. I consider Diary data since they are known to be particularly affected by frequency issues due to the short two-week time span they cover. The goods under study are either relatively frequently purchased, such as food, or unfrequently purchased, such as alchool and tobacco. For each category of goods I estimate the latent distribution of the expected frequency of purchase (λ) under the assumption that this is Gamma distributed, hence applying the maximum likelihood estimator described in (5.8) separately for each year. While it may be assumed that every household is a food consumer, there is a non-negligible proportion of households whose consumption level of alcohol or tobacco is zero. In order to account for that I estimated the proportion of non-consumers for alcohol and tobacco consumption by applying equation (5.9). Results, stratified by level of education, are reported in Figure 5.1. The mean of the estimated latent distribution of expected frequency is decreasing over time for all of the goods considered, meaning that households have been changing their purchasing habits during the sample period and, for instance, in 2003 they are buying goods less frequently than in 1982.

[2]Note that if C is lognormally distributed then $\log C$ is normal. It is well known that the the rate of convergence of non-parametric deconvolution estimators is particularly slow when the distribution is normal and, more generally, when the distribution is "supersmooth" (see Fan 1991).

There is also substantial evidence of heterogeneity across groups with different levels of education, though the pattern overtime seems to be the same. The estimated proportion of tobacco non-consumers is around 50 percent, while for alcohol it is roughly around 20 percent. Since estimates of these proportion are obtained through equation (5.9) there is no guarantee that the point estimate is inside the parameter space, that is the interval $(0, 1)$, as is the case in Figure 5.1.

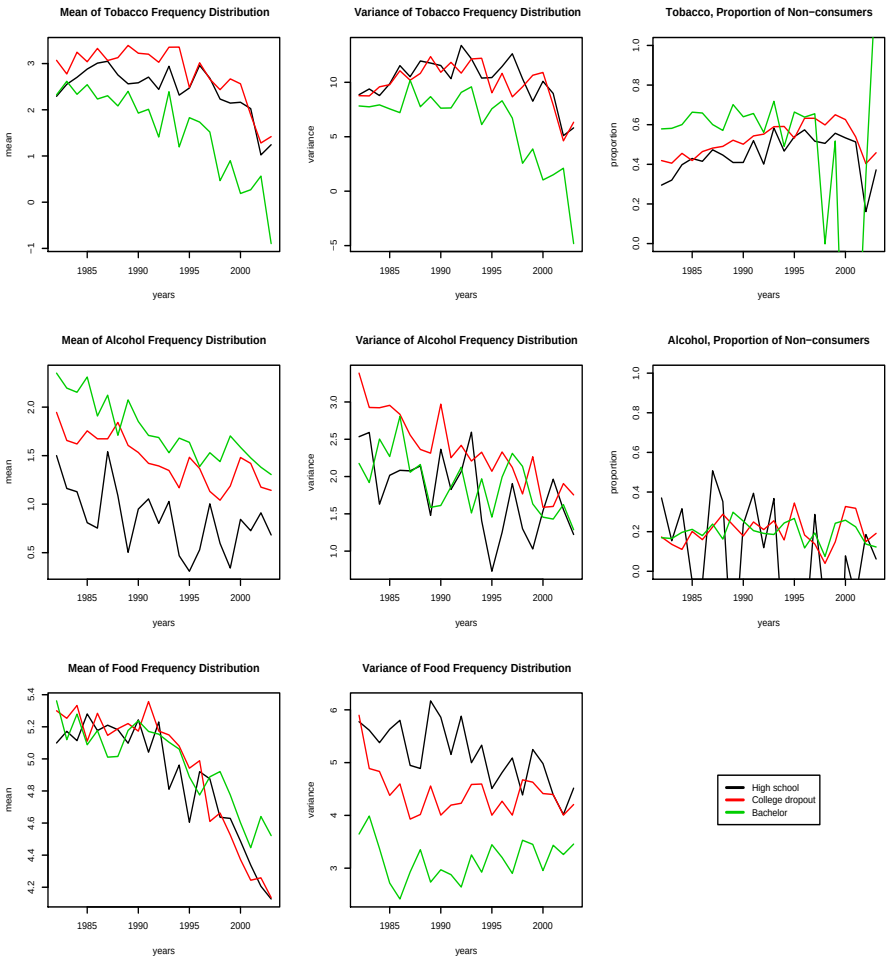

Figure 5.1: Estimates of the distribution of expected frequency by levels of education.

Estimates of the distribution of λ allows the estimation of the variance of the unobserved consumption according to equation (5.10). I assume food consumption is lognormally distributed and estimate its whole distribution via maximum likelihood. The likelihood function is derived from equation (5.11). Estimates of the parameters of the distributon of unobserved food

consumption are then used to obtain consistent estimates of several inequality measures.

Figure 5.2 reports the pattern of the squared coefficient of variation, defined as $Var[C]/E[C]^2$, obtained from observed food expenditure and the estimated moments of unobserved food consumption for the years 1982-2003 stratified by level of education. Results show a severe underestimation of the level of dispersion in food consumption when considering raw expenditure data. The between groups variation seems also to be reduced when considering food consumption.

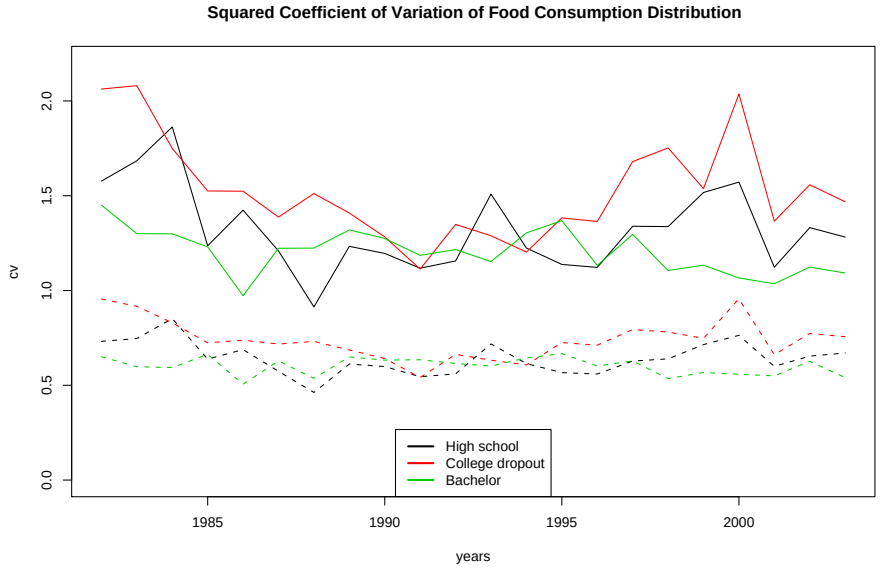

Figure 5.2: Estimates of the squared coefficient of variation of observed expenditure (solid lines) and unobserved consumption (dashed lines) by level of education.

Knowledge of the entire distribution of consumption, and not just the first few moments, is important since it allows the computation of inequality measures such as the Gini coefficient which depend on the whole density function. In order to assess the discrepancies between estimates of these indices obtained from observed expenditure or consumption, Figure 5.3 shows the estimated values of the Gini coefficient for the period under study, while Figure 5.4 reports the computed proportion of poors, being defined as the proportion of households below 0.6 times the median of the marginal distribution considered.

As expected both these graphs show a significant difference in the levels of the measures considered, even though the pattern across years remains essentially unchanged.

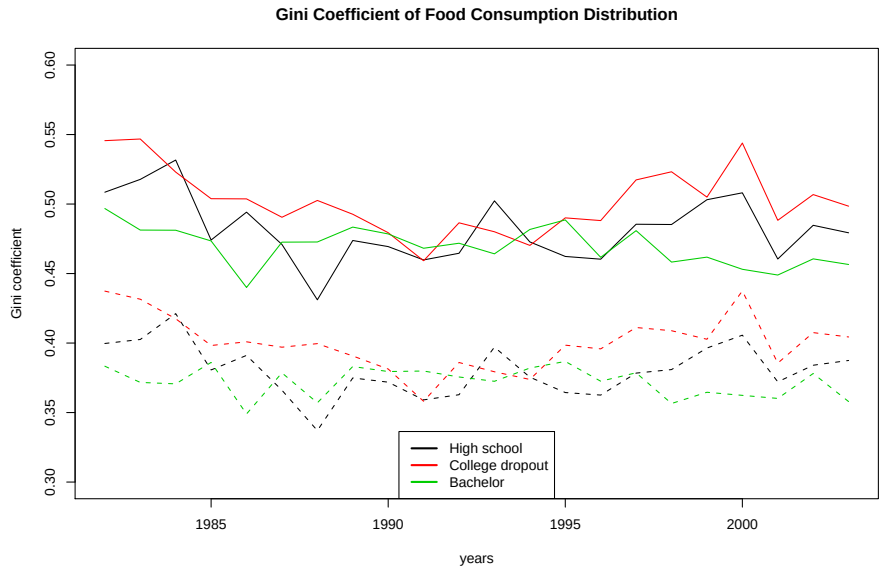

Figure 5.3: Estimates of the Gini coefficient of the observed expenditure (solid lines) and unobserved consumption (dashed lines) by level of education.

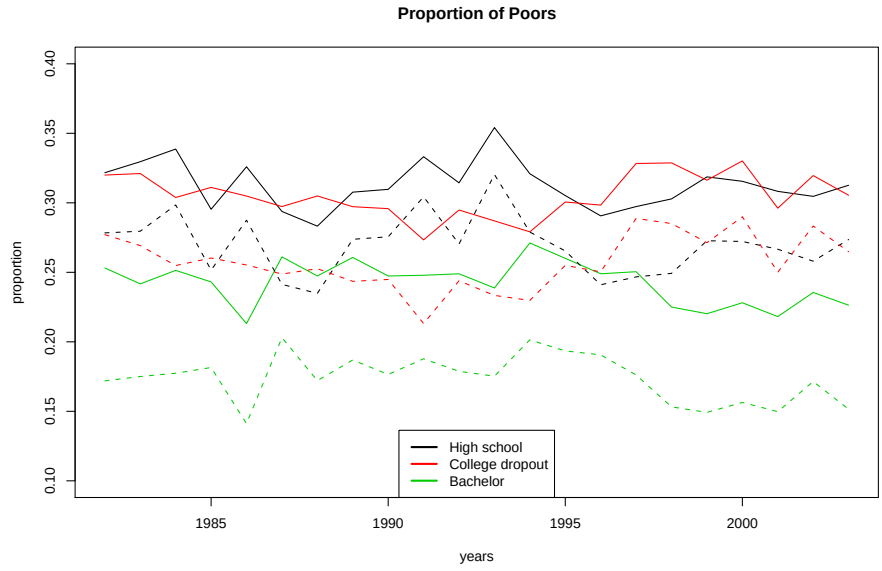

Figure 5.4: Estimates of the proportion of poors according to the observed distribution of expenditure (solid lines) and the estimated distribution of consumption (dashed lines) by level of education.

5.5 Chapter Summary

In this Chapter I discussed a model which accounts for the presence of frequency of purchase issues. I obtained identification of the distribution of consumption from knowledge of expenditure data and the number of observed purchases. The proposed estimator allows for the estimation of all the moments of the unobserved distribution of consumption. By extension the entire distribution of consumption is identified and can be estimated allowing several features of the distribution to be investigated. The results of the application to the CEX data show a substantial overestimation of all of the inequality measures considered when using expenditure data as opposed to the proposed estimator of features of the consumption distribution.

Conclusions

Consumption data are known to be heavily affected by measurement errors. Estimation of most consumption models without accounting for their presence may thus produce severely biased estimates of the parameters of interest. In this thesis I considered three different consumption models which account for the presence of measurement errors of various nature.

Chapter 3 discussed identification of a particular specification of Engel curves when total expenditure is both unobserved and endogenous. Identification was achieved by means of a control function assumption. The resulting estimator is consistent and easily computed with standard statistical software. A Monte Carlo study revealed that the estimator is particularly successfull in reducing the bias of alternative estimators already available in the literature when the extent of endogeneity and measurement error is severe.

Chapter 4 studied identification of general non-linear errors in variables models when correlated measurement errors appear on both sides of the equation. The identification strategy otulined extends the past results by Schennach (2007). In the Engel curves framework I also obtained identification of the conditional distribution of the unobserved Y_i^* given X_i^*. This way I can disentangle variation in expenditures due to measurement errors from variation due to heterogeneity in preferences. The identification result is illustrated by proposing an estimator to consistently estimate shape invariant Engel curves in which the error term, representing heterogeneous preferences, enters non-additively the equation.

Chapter 5 discussed the relationship between expenditure and consumption. Indeed, commonly recorded expenditure data may fail to provide a satisfactory approximation to the desired consumption level because of infrequency of purchasing in some goods. I discussed identification and estimation of features of the distribution of consumption from knowledge of expenditure data and the number of purchases. These results were applied to estimate the unobserved distribution of consumption from CEX expenditure data using the Diary survey. The analysis showed that ignoring infrequency of purchasing may result in non-negligible overestimation of routinely used measures of inequality in food.

Appendix

Proof of Theorem 3.1

From equations (3.2) and (3.6) we have:

$$W_{ih} = \frac{\sum_{j=0}^{K} b_{ij}(\log X_h^*)^j + \varepsilon_{ih} + \nu_{ih}}{V_h}.$$

Multiplying by X_h^l either side of the equation, using (3.4) and taking the conditional expectation with respect to Z_h yields:

$$\begin{aligned}
E[X_h^l W_{ih}|Z_h] &= E\left[(X_h^* V_h)^l \frac{\sum_{j=0}^{K} b_{ij}(\log X_h^*)^j + \varepsilon_{ih} + \nu_{ih}}{V_h}|Z_h\right], \\
&= \sum_{j=0}^{K} b_{ij} E[V_h^{l-1} X_h^{*l} (\log X_h^*)^j|Z_h] + E[V_h^{l-1} X_h^{*l} \varepsilon_{ih}|Z_h] + \\
&\quad + E[V_h^{l-1} X_h^{*l} \nu_{ih}|Z_h], \\
&= \sum_{j=0}^{K} b_{ij} E[V_h^{l-1}] E[X_h^{*l} (\log X_h^*)^j|Z_h] + E[V_h^{l-1}] E[X_h^{*l} \varepsilon_{ih}|Z_h] \\
&+ E[V_h^{l-1} \nu_{ih}] E[X_h^{*l}|Z_h], \qquad (5.12)
\end{aligned}$$

where the last equality follows from Assumption 3.1 (iii). Hence we may write:

$$\begin{aligned}
E[X_h^l (\log X_h)^j] &= E[X_h^{*l} V_h^l (\log X_h^* + \log V_h)^j|Z_h], \\
&= E[X_h^{*l} V_h^l \sum_{s=0}^{j} \binom{j}{s} (\log X_h^*)^{j-s} (\log V_h)^s|Z_h], \\
&= \sum_{s=0}^{j} \binom{j}{s} E[V_h^l (\log V_h)^s] E[X_h^{*l} (\log X_h^*)^{j-s}|Z_h], \\
&= E[X_h^{*l} (\log X_h^*)^j|Z_h] E[V_h^l] + \\
&\quad + \sum_{s=1}^{j} \binom{j}{s} E[V_h^l (\log V_h)^s] E[X_h^{*l} (\log X_h^*)^{j-s}|Z_h],
\end{aligned}$$

which yields:

$$E[X_h^{*l}(\log X_h^*)^j|Z_h] = \frac{E[X_h^l(\log X_h)^j|Z_h]}{E[V_h^l]} + \\ -\frac{\sum_{s=1}^{j}\binom{j}{s}E[V_h^l(\log V_h)^s]E[X_h^{*l}(\log X_h^*)^{j-s}|Z_h]}{E[V_h^l]}.$$

The above equation could be solved recursively to obtain:

$$E[X_h^{*l}(\log X_h^*)^j|Z_h] = \sum_{t=0}^{j}\gamma_{j-t}E[X_h^l(\log X_h)^t|Z_h] \tag{5.13}$$

where

$$\gamma_t \equiv (-1)^t\frac{E[V_h^l\log V_h]^t}{E[V_h^l]^{t+1}}.$$

Substitution of (5.13) into (5.12) yields:

$$\begin{aligned} E[X_h^l W_{ih}|Z_h] &= \sum_{j=0}^{K}b_{ij}E[V_h^{l-1}]\sum_{t=0}^{j}\gamma_{j-t}E[X_h^l(\log X_h)^t|Z_h] + \\ &+ E[V_h^{l-1}]E[X_h^{*l}\varepsilon_{ih}|Z_h] + \frac{E[V_h^{l-1}\nu_{ih}]}{E[V_h^l]}E[X_h^l|Z_h], \\ &= \sum_{j=0}^{K}\sum_{t=0}^{K}b_{ij}E[V_h^{l-1}]\gamma_{j-t}E[X_h^l(\log X_h)^t|Z_h]\mathbf{1}(j-t\geq 0) + \\ &+ E[V_h^{l-1}]E[X_h^{*l}\varepsilon_{ih}|Z_h] + \frac{E[V_h^{l-1}\nu_{ih}]}{E[V_h^l]}E[X_h^l|Z_h], \\ &= \sum_{t=0}^{K}\beta_t E[X_h^l(\log X_h)^t|Z_h] + E[V_h^{l-1}]E[X_h^{*l}\varepsilon_{ih}|Z_h], \end{aligned} \tag{5.14}$$

where:

$$\beta_{ilt} \equiv \frac{E[V_h^{l-1}\nu_{ih}]}{E[V_h^l]}\mathbf{1}(t=0) + \sum_{j=0}^{K}b_{ij}E[V_h^{l-1}]\gamma_{j-t}\mathbf{1}(j-t\geq 0).$$

As there is

$$\begin{aligned} E[X_h^l\varepsilon_{ih}|Z_h] &= E[X_h^{*l}V_h^l\varepsilon_{ih}|Z_h], \\ &= E[V_h^l]E[X_h^{*l}\varepsilon_{ih}|Z_h], \end{aligned} \tag{5.15}$$

and:

$$\begin{aligned} E[X_h^l \varepsilon_{ih}|Z_h] &= E[g(Z_h,\eta_h)^l \varepsilon_{ih}|Z_h], \\ &= E\Big\{ g(Z_h,\eta_h)^l E[\varepsilon_{ih}|Z_h,\eta_h]|Z_h \Big\}, \\ &= E\Big\{ g(Z_h,\eta_h)^l E[\varepsilon_{ih}|\eta_h]|Z_h \Big\}, \\ &= E[g(Z_h,\eta_h)^l \lambda_i(\eta_h)|Z_h], \\ &= \int_0^1 g(Z_h,\eta_h)^l \lambda_i(\eta_h) d\eta_h, \end{aligned} \tag{5.16}$$

Substituting (5.15) and (5.16) into (5.14) yields:

$$E[X_h^l W_{ih}|Z_h] = \sum_{t=0}^{K} \beta_{ilt} E[X_h^l (\log X_h)^t |Z_h] + \tau_{il}(Z_h),$$

where

$$\tau_{il}(Z_h) \equiv \frac{E[V_h^{l-1}]}{E[V_h^l]} \int_0^1 g(Z_h,\eta_h)^l \lambda_i(\eta_h) d\eta_h.$$

Q.E.D.

Proof of Theorem 3.2

Consider the conditional mean of W_{ih} given Z_h, by equation (3.6) it is:

$$\begin{aligned} E[W_{ih}|Z_h] &= E[W_{ih}^* V_h^{-1}|Z_h] + E[\nu_{ih} V_h^{-1}|Z_h] \\ &= E[\nu_{ih} V_h^{-1}] + E[V_h^{-1}] \sum_{j=0}^{K} b_{ij} E\left[(\log X_h^*)^j |Z_h \right]. \end{aligned} \tag{5.17}$$

Now using (5.13) with $l = 0$ and substituting in (5.17) we obtain:

$$\begin{aligned} E[W_{ih}|Z_h] &= E[\nu_{ih} V_h^{-1}] + E[V_h^{-1}] \sum_{j=0}^{K} b_{ij} \sum_{t=0}^{j} \gamma_{j-t} E[(\log X_h)^t |Z_h] \\ &= E[\nu_{ih} V_h^{-1}] + E[V_h^{-1}] \sum_{j=0}^{K} \sum_{t=0}^{K} b_{ij} \gamma_{j-t} E[(\log X_h)^t |Z_h] \mathbf{1}(j-t \geq 0) \\ &= \sum_{t=0}^{K} \beta_{it}^{2SLS} E[(\log X_h)^t |Z_h], \end{aligned}$$

with

$$\beta_{it}^{2SLS} = E[\nu_{ih} V_h^{-1}] \mathbf{1}(t=0) + \sum_{j=0}^{K} E[V_h^{-1}] b_{ij} \gamma_{j-t} \mathbf{1}(j-t \geq 0).$$

Q.E.D.

Assessing what OLS identifies

If $K = 1$ the probability limit of the OLS linear coefficient is:

$$\begin{aligned}
\beta_{i1}^{OLS} &= \frac{Cov(W_{ih}, \log X_h)}{Var(\log X_h)} \\
&= \frac{Cov\Big(\frac{b_{i0}}{V_h} + b_{i1}\frac{\log X_h}{V_h} + \frac{\nu_{ih}+\varepsilon_{ih}-b_{i1}\log V_h}{V_h}, \log X_h^* + \log V_h\Big)}{Var[\log X_h^*] + \lambda} \\
&= \frac{(b_{i0}-1)}{Var[\log X_h^*]+\lambda} Cov\Big(\frac{1}{V_h}, \log V_h\Big) + \\
&\quad + \frac{b_{i1}}{Var[\log X_h^*]+\lambda}\Bigg\{ Cov\Big(\frac{\log X_h^*}{V_h}, \log X_h^*\Big) + \\
&\quad + Cov\Big(\frac{\log X_h^*}{V_h}, \log V_h\Big)\Bigg\} + \frac{1}{Var[\log X_h^*]+\lambda} Cov\Big(\frac{\varepsilon_{ih}}{V_h}, \log V_h\Big),
\end{aligned}$$

where $\lambda = Var(V_h)$. Now using the fact that if $V_h \sim \log N(\mu, \sigma^2)$ it is:

$$E\Big[\frac{\log V_h}{V_h}\Big] = 3\mu e^{-2\mu}.$$

Substituting and rearranging terms we obtain

$$\beta_{i1}^{OLS} = e^{-2\mu}\Bigg\{ b_{i1} + \frac{2\mu[(b_{i0}-1) + b_{i1}(1 + E[\log X^*])]}{\sigma^2_{\log X_h^*} - 2\mu} \Bigg\},$$

or alternatively

$$\beta_{i1}^{OLS} = \beta_{i1}^{2SLS} - (\lambda+1)\frac{\log(\lambda+1)[(b_{i0}-1) + b_{i1}(1 + E[\log X_h^*])]}{Var[\log X_h^*] + \log(\lambda+1)}. \quad (5.18)$$

Proof of Lemma 4.1

Under Assumption 4.1 we may write:

$$\begin{aligned}
E[Y^k|Z] &= \int \mu^k(z-v)dF(v) \\
E[XY|Z] &= \int (z-v)\mu^1(z-v)dF(v) + \lambda,
\end{aligned}$$

where $\lambda = E[WS]$. Now taking the Fourier transform on both sides of the equation we obtain:

$$\begin{aligned}
\varepsilon_{y^k}(\zeta) &= \int\int \mu^k(z-v)dF(v)e^{\mathbf{i}\zeta z}dz \\
&= \int\int \mu^k(x^*)e^{\mathbf{i}\zeta(x^*+v)}dx^*dF(v) \\
&= \int\int \mu^k(x^*)e^{\mathbf{i}\zeta x^*}dx^*e^{\mathbf{i}\zeta v}dF(v) \\
&= \int \mu^k(x^*)e^{\mathbf{i}\zeta x^*}dx^* \int e^{\mathbf{i}\zeta v}dF(v) \\
&= \gamma_k(\zeta)\phi(\zeta),
\end{aligned}$$

and

$$\begin{aligned}
\varepsilon_{xy}(\zeta) &= \int\int (z-v)\mu^1(z-v)dF(v)e^{\mathbf{i}\zeta z}dz + \int \lambda e^{\mathbf{i}\zeta z}dz \\
&= \int\int x^*\mu^1(x^*)e^{\mathbf{i}\zeta(x^*+v)}dx^*dF(v) + \lambda\int e^{\mathbf{i}\zeta z}dz \\
&= \int x^*\mu^1(x^*)e^{\mathbf{i}\zeta x^*}dx^* \int e^{\mathbf{i}\zeta v}dF(v) + \lambda\int e^{\mathbf{i}\zeta z}dz \\
&= \left(-\mathbf{i}\frac{\partial}{\partial\zeta}\int \mu^1(x^*)e^{\mathbf{i}\zeta x^*}dx^*\right)\phi(\zeta) + \lambda\psi(\zeta) \\
&= -\mathbf{i}\dot{\gamma}_1(\zeta)\phi(\zeta) + \lambda\psi(\zeta),
\end{aligned}$$

hence $\mathbf{i}\varepsilon_{xy}(\zeta) = \dot{\gamma}_1(\zeta)\phi(\zeta) + \mathbf{i}\lambda\psi(\zeta)$.

Q.E.D.

Proof of Theorem 4.2

By manipulating (4.12) and (4.13) we obtain

$$\begin{aligned}
\varepsilon_{y^k}(\zeta) &= \gamma_k(\zeta)\phi(\zeta) \\
\mathbf{i}\varepsilon_{(z-x)y}(\zeta) &= \gamma_1(\zeta)\dot{\phi}(\zeta) - \lambda\mathbf{i}\psi(\zeta).
\end{aligned} \tag{5.19}$$

where $\mathbf{i}\varepsilon_{(z-x)y}(\zeta) = \mathbf{i}\varepsilon_{zy}(\zeta) - \mathbf{i}\varepsilon_{xy}(\zeta)$ and $\mathbf{i}\varepsilon_{zy}(\zeta) \equiv \dot{\varepsilon}_y(\zeta) = \dot{\gamma}_1(\zeta)\phi(\zeta) + \gamma_1(\zeta)\dot{\phi}(\zeta)$. The main point to keep in mind is that $\varepsilon_y(\zeta)$, $\varepsilon_{(z-x)y}(\zeta)$, $\gamma_k(\zeta)$ and $\psi(\zeta)$ are generalized functions (Lighthill, 1962), so that the product operation between two of them is not defined. However any generalized function may be decomposed into the sum of an ordinary function (denoted with subscript o) and a purely singular function (denoted with subscript s), so that

$$\begin{aligned}
\varepsilon_{y,o}(\zeta) + \varepsilon_{y,s}(\zeta) &= [\gamma_{1;o}(\zeta) + \gamma_{1;s}(\zeta)]\,\phi(\zeta) \\
\mathbf{i}\varepsilon_{(z-x)y,o}(\zeta) + \mathbf{i}\varepsilon_{(z-x)y,s}(\zeta) &= [\gamma_{1;o}(\zeta) + \gamma_{1;s}(\zeta)]\,\dot{\phi}(\zeta) - \lambda\mathbf{i}\psi(\zeta).
\end{aligned}$$

Note that we do not put subscripts on $\phi(\zeta)$ or $\psi(\zeta)$, since the former is an ordinary function, being the Fourier transform of an absolutely integrable function, and the latter is a purely singular function as $\psi(\zeta) = 2\pi\delta(\zeta)$. Now applying Lemma 2 in Schennach (2007), which states that the product of an ordinary function with an ordinary function is an ordinary function, whereas the product of a purely singular component with an ordinary function is purely singular, and equating the ordinary components in the above equations we obtain equations (50) and (51) in Schennach (2007), namely

$$\begin{aligned} \varepsilon_{y,o}(\zeta) &= \gamma_{1;o}(\zeta)\phi(\zeta) \\ \mathbf{i}\varepsilon_{(z-x)y,o}(\zeta) &= \gamma_{1;o}(\zeta)\dot{\phi}(\zeta). \end{aligned}$$

These are now all ordinary functions and they may be manipulated as in Schennach (2007) to obtain identification of $\phi(\zeta)$ for $\zeta \in [-\bar{\zeta}_1, \bar{\zeta}_1]$ with

$$\phi(\zeta) = \exp\left(\int_0^\zeta \frac{\mathbf{i}\varepsilon_{(z-x)y,o}(\zeta)}{\varepsilon_{y,o}(\zeta)} d\zeta\right). \tag{5.20}$$

We restrict our attention to the case in which $\bar{\zeta}_1 > 0$, however $\bar{\zeta}_1 = 0$ only when the conditional mean $E[Y^*|X^*]$ is a polynomial in X^*, a case which has already been shown to be identified by Hausman, Newey, Ichimura, and Powell (1991). Substitution of (5.20) in (5.19) leads to

$$\gamma_k(\zeta) = \begin{cases} 0 & \text{if } \varepsilon_y(\zeta) = 0 \\ \varepsilon_{y^k}(\zeta)/\phi(\zeta) & \text{otherwise,} \end{cases}$$

and hence by taking the inverse Fourier transform of $\gamma_k(\zeta)$:

$$\mu^k(X^*) = (2\pi)^{-1} \int \gamma_k(\zeta) e^{-\mathbf{i}\zeta X^*} d\zeta$$

proving the identification of $\mu^k(X^*)$.

Q.E.D.

Proof of Theorem 4.3

Since the first K moments of X exist by assumption then $E[Z^k]$ also exists. Under Assumptions (4.1) to (4.4), by Theorem 4.2, $\phi(\zeta)$ is shown to be identified in a neighborhood of the origin, hence $\phi^{(k)}(0)$ is also identified for $k = 1, \ldots, K$ and, exploiting equation (4.8), we may write:

$$\begin{aligned} E[X^{*k}] &= \sum_{j=0}^{k} \binom{k}{j} E[Z^j] E[(-\tilde{V})^{k-j}] \\ &= \sum_{j=0}^{k} \binom{k}{j} E[Z^j](-1)^{k-j}(-\mathbf{i}^{k-j})\phi^{(k-j)}(0) \\ &= \sum_{j=0}^{k} \binom{k}{j} (-\mathbf{i})^{k-j} E[Z^j]\phi^{(k-j)}(0). \end{aligned} \tag{5.21}$$

The observed k-th moment of X, because of equation (4.7), may be written as $E[X^k] = E[X^{*k}]E[W^k]$, which implies:

$$E[W^k] = \frac{E[X^k]}{E[X^{*k}]} \tag{5.22}$$

which is well defined since by assumption $E[X^{*k}] \neq 0$. Substitution of equation (5.21) into (5.22) yields equation (4.22).

Equation (4.23) follows by noting that from equation (4.7) and by Assumptions 4.1 we have that:

$$\begin{aligned} E[X^k Y|Z] &= E[X^{*k}W^k H(X^*,U)|Z] + E[X^{*k}W^k X^{*l}S|Z] \\ &= E[X^{*k}\mu^1(X^*)|Z]E[W^k] + E[X^{*k+l}|Z]E[W^k S], \end{aligned}$$

which gives

$$E[W^k S] = \frac{E[X^k Y|Z] - E[X^{*k}\mu^1(X^*)|Z]E[W^k]}{E[X^{*k+l}|Z]}. \tag{5.23}$$

The right hand side of equation (5.23) involves functions that are either observable, like $E[X^k Y|Z]$, or already shown to be identified. In particular $E[X^{*k+l}|Z]$ is obtained from knowledge of $\phi(\zeta)$ as

$$\begin{aligned} E[X^{*k+l}|Z] &= \int (z-v)^{k+l} dF(v) = \int \sum_{j=0}^{k+l} \binom{k+l}{j} z^j v^{k+l-j} dF(v) \\ &= \sum_{j=0}^{k+l} \binom{k+l}{j} z^j \int v^{k+l-j} dF(v) \\ &= \sum_{j=0}^{k+l} \binom{k+l}{j} z^j (-\mathbf{i})^{k+l-j} \phi^{(k+l-j)}(0). \end{aligned} \tag{5.24}$$

On the other hand, if $\bar{\zeta}_1 = \infty$ in Assumption 4.4, it is:

$$\begin{aligned} \int E[X^{*k}\mu^1(X^*)|Z] e^{\mathbf{i}\zeta z} dz &= \int\int (z-v)^k \mu^1(z-v) dF(v) e^{\mathbf{i}\zeta z} dz \\ &= (-\mathbf{i})^k \gamma_1^{(k)}(\zeta)\phi(\zeta), \end{aligned}$$

where $\gamma_1(\zeta)$ is defined as in equation (4.14) and by taking the inverse Fourier transform we get:

$$E[X^{*k}\mu^1(X^*)|Z] = (2\pi)^{-1} \int (-\mathbf{i})^k \gamma_1^{(k)}(\zeta)\phi(\zeta) e^{-\mathbf{i}\zeta z} d\zeta. \tag{5.25}$$

Substitution of equations (5.24) and (5.25) into (5.23) gives:

$$E[W^k S] = \frac{E[X^k Y|Z] - (2\pi)^{-1} E[W^k] \int (-\mathbf{i})^k \gamma_1^{(k)}(\zeta)\phi(\zeta) e^{-\mathbf{i}\zeta z} d\zeta}{\sum_{j=0}^{k+l} \binom{k+l}{j} z^j (-\mathbf{i})^{k+l-j} \phi^{(k+l-j)}(0)}$$

Q.E.D.

Bibliography

Adcock, R. (1878): "A problem in least squares," *The Analyst*, 5, 53–54.

Ai, C., and X. Chen (2003): "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions," *Econometrica*, 71(6), 1795–1843.

Amemiya, Y. (1985): "Instrumental variable estimator for the nonlinear errors-in-variables model," *Journal of Econometrics*, 28(3), 273–289.

Atkinson, A. B., J. Gomulka, and N. H. Stern (1990): "Spending on Alcohol: Evidence from the Family Expenditure Survey 1970-1983," *The Economic Journal*, 100(402), 808 – 827.

Attanasio, O., E. Battistin, and A. Mesnard (2009): "Food and cash transfers: evidence from Colombia," Working paper, Institute for Fiscal Studies.

Banks, J., R. Blundell, and A. Lewbel (1997): "Quadratic Engel Curves and Consumer Demand," *The Review of Economics and Statistics*, 79(4), 527 – 539.

Battistin, E., R. Blundell, and A. Lewbel (2009): "Why Is Consumption More Log Normal than Income? Gibrats Law Revisited," *Journal of Political Economy*, 117(6), 1140–1154.

Bierens, H., and H. Pott-Buter (1990): "Specification of household engel curves by nonparametric regression," *Econometric Reviews*, 9(2), 123–184.

Blundell, R. (1988): "Consumer Behaviour: Theory and Empirical Evidence–A Survey," *The Economic Journal*, 98(389), 16 – 65.

Blundell, R., M. Browning, and I. A. Crawford (2003): "Nonparametric Engel Curves and Revealed Preference," *Econometrica*, 71(1), 205 – 240.

Blundell, R., X. Chen, and D. Kristensen (2007): "Semi-Nonparametric IV Estimation of Shape-Invariant Engel Curves," *Econometrica*, 75(6), 1613 – 1669.

Blundell, R., A. Duncan, and K. Pendakur (1998): "Semiparametric Estimation and Consumer Demand," *Journal of Applied Econometrics*, 13(5), 435–461.

Blundell, R., and C. Meghir (1987): "Bivariate Alternatives to the Tobit Model," *Journal of Econometrics*, 34(1), 179–200.

Blundell, R., P. Pashardes, and G. Weber (1993): "What do we Learn About Consumer Demand Patterns from Micro Data?," *The American Economic Review*, 83(3), 570 – 597.

Bound, J., C. Brown, and N. Mathiowetz (2001): "Measurement error in survey data," *Handbook of Econometrics*, 5, 3705–3843.

Brown, B. W., and M. B. Walker (1989): "The Random Utility Hypothesis and Inference in Demand Systems," *Econometrica*, 57(4), 815–829.

Carroll, R., D. Ruppert, L. Stefanski, and C. M. Crainiceanu (2006): *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman & Hall, New York.

Deaton, A., and M. Irish (1984): "Statistical Models for Zero Expenditure in Households Budgets," *Journal of Public Economics*, 23, 59–80.

Deaton, A., and J. Muellbauer (1980): "An Almost Ideal Demand System," *The American Economic Review*, 70(3), 312–326.

Engel, E. (1895): "Die Lebenskosten Belgischer Arbeiter-Familien Fruher and jetzt," *International Statistical Institute Bulletin*, 9, 1–74.

Fan, J. (1991): "On the Optimal Rates of Convergence for Nonparametric Deconvolution Problems," *The Annals of Statistics*, 19(3), 1257 – 1272.

Gallant, A. R., and D. W. Nychka (1987): "Semi-Nonparametric Maximum Likelihood Estimation," *Econometrica*, 55(2), 363 – 390.

Gorman, W. M. (1961): "On a Class of Preference Fields," *Metroeconomica*, (13), 53–56.

——— (1981): *Some Engel Curves*. Cambridge University Press, Cambridge.

Hahn, J., and G. Ridder (2010): "Conditional Moment Restrictions and Triangular Simultaneous Equations," *Review of Economics and Statistics*, (forthcoming).

Hardle, W., and J. S. Marron (1990): "Semiparametric Comparison of Regression Curves," *The Annals of Statistics*, 18(1), 63 – 89.

Hausman, J., W. K. Newey, H. Ichimura, and J. L. Powell (1991): "Identification and estimation of polynomial errors-in-variables models," *Journal of Econometrics*, 50(3), 273–295.

Hausman, J., W. K. Newey, and J. L. Powell (1995): "Nonlinear errors in variables Estimation of some Engel curves," *Journal of Econometrics*, 65(1), 205–233.

Howe, H., R. A. Pollack, and T. J. Wales (1979): "Theory and Time Series Estimation of the Quadratic Expenditure System," *Econometrica*, (47), 1231–1243.

Hsiao, C. (1989): "Consistent estimation for some nonlinear errors-in-variables models," *Journal of Econometrics*, 41(1), 159–185.

Hsiao, C., and Q. K. Wang (2000): "Estimation of Structural Nonlinear Errors-in-Varibles Models by Simulated Least-Squares Method," *International Economic Review*, 41(2), 523–542.

Imbens, G. W., and W. K. Newey (2009): "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," *Econometrica*, 77(5), 1481–1512.

Kay, J., M. J. Keen, and C. N. Morris (1984): "Estimating Consumption from Expenditure data," *Journal of Public Economics*, 23(1-2), 169–181.

Keen, M. (1986): "Zero Expenditure and the Estimation of Engel Curves," *Journal of Applied Econometrics*, 1(3), 277–286.

Leser, C. E. V. (1963): "Forms of Engel Functions," *Econometrica*, 31(4), 694–703.

Lewbel, A. (1991): "The Rank of Demand Systems: Theory and Nonparametric Estimation," *Econometrica*, 59(3), 711 – 730.

——— (1996): "Demand Estimation with Expenditure Measurement Errors on the Left and Right Hand Side," *The Review of Economics and Statistics*, 78(4), 718 – 725.

——— (2001): "Demand Systems with and without Errors," *The American Economic Review*, 91(3), 611 – 618.

——— (2010): "Shape-Invariant Demand Functions," *The Review of Economics and Statistics*, 92(3), 549–556.

Lewbel, A., and K. Pendakur (2009): "Tricks with Hicks: The EASI Demand System," *The American Economic Review*, 99(3), 827–863.

Li, T. (2002): "Robust and consistent estimation of nonlinear errors-in-variables models," *Journal of Econometrics*, 110(1), 1–26.

Lighthill, M. J. (1962): *Introduction to Fourier Analysis and Generalized Functions*. Cambridge University Press, London.

Liviatan, N. (1961): "Errors in Variables and Engel Curves Analysis," *Econometrica*, 29, 336–362.

McFadden, D. (1989): "A Method of Simulated Moments for Estimation of Discrete Response Models Without Numerical Integration," *Econometrica*, 57(5), 995–1026.

Meghir, C., and J.-M. Robin (1992): "Frequency of purchase and the estimation of demand systems," *Journal of Econometrics*, 53(1-3), 53–85.

Newey, W. K. (2001): "Flexible Simulated Moment Estimation of Nonlinear Errors-in-Variables Models," *The Review of Economics and Statistics*, 83(4), 616 – 627.

Ogburn, W. F. (1919): "Analysis of the Standard of Living in the District of Columbia in 1916," *Publications of the American Statistical Association*, 16(126), 374 – 389.

Pinkse, C., and P. Robinson (1995): *Statistical Methods of Econometrics and Quantitative Economics.*

Pollack, R. A., and T. J. Wales (1995): *Demand System Specification and Estimation*. Oxford University Press.

Prais, S., and H. Houthakker (1971): *The Analysis of Family Budgets*. Second edn.

Prais, S. J. (1959): "A comment," *Econometrica*, 27, 127–129.

Robin, J. M. (1993): "Econometric Analysis of the Short-Run Fluctuations of Households Purchases," *The Review of Economic Studies*, 60(4), 923–934.

Schennach, S. M. (2004): "Estimation of Nonlinear Models with Measurement Error," *Econometrica*, 72(1), 33 – 75.

——— (2007): "Instrumental Variable Estimation of Nonlinear Errors-in-Variables Models," *Econometrica*, 75(1), 201 – 239.

Stone, J. R. N. (1954): "Linear Expenditure Systems and Demand Analysis: An Application to the Pattern of British Demand," *Economics Journal*, 64, 511–527.

Summers, R. (1959): "A note on Least Squares Bias in Household Expenditure Analysis," *Econometrica*, 27, 121–126.

Wang, L., and C. Hsiao (1995): "Simulation-Based Seiparametric Estimation of Nonlinear Errors-in-Variables Models," Working paper, University of Southern California.

Working, H. (1943): "Statistical Laws of Family Expenditure," *Journal of the American Statistical Association*, 38(221), 43–56.

Michele De Nadai

CURRICULUM VITAE

Personal Details

Date of Birth: September 28, 1983
Place of Birth: Conegliano (Treviso), Italy
Nationality: Italian

Contact Information

University of Padova
Department of Statistics
via Cesare Battisti, 241-243
35121 Padova. Italy.

Tel. +39 049 827 4174
e-mail: denadai@stat.unipd.it

Current Position

Since January 2008; (expected completion: January 2011)
PhD Student in Statistical Sciences, University of Padova.
Thesis title: Measurement Error Issues in Consumption Data
Supervisor: Prof. Erich Battistin

Research interests

- Measurement Errors
- Microeconometrics
- Demand Analysis
- Panel Data

Education

February 2006 – October 2007
Master (*laurea magistrale*) degree in Statistics and Economics .
University of Padova, Faculty of Statistics
Title of dissertation: "Italian Households' Participation in Risky Assets: a Panel Data Analysis"
Supervisor: Prof. Guglielmo Weber
Final mark: 110/110 *summa cum laude*

October 2002 – February 2006
Bachelor degree (*laurea triennale*) in Statistics.
University of Padova, Faculty of Statistics
Title of dissertation: "Il Controllo di Gestione nelle Imprese Commerciali"
Supervisor: Prof. Fabrizio Cerbioni
Final mark: 101/110.

Visiting periods

October 2009 – December 2009 and October 2010 – November 2010
Boston College, Department of Economics
Chestnut Hill, MA, USA.
Supervisor: Prof. Arthur Lewbel

Further education

August 2008
Summer School in "Empirical Strategies"
CEMFI, Madrid, Spain.
Instructor: Joshua Angrist (M.I.T.)

Awards and Scholarship

October 2008 – present
Italian Ministry of University and Scientific Research: Three-year scholarship for Ph.d. studies at the University of Padova.

Computer skills

- R
- STATA
- LaTex
- Mathematica
- EViews

Language skills

Italian: native; English: fluent.

Conference presentations

Battistin, E., De Nadai, M., (2010). Identification and estimation of Engel curves with endogenous and unobserved expenditures. (poster session) *EC^2: Identification in Econometrics: Theory and Applications.* , Tolouse, France, December 17-18, 2010.

Battistin, E., De Nadai, M., (2011). Identification and estimation of Engel curves with endogenous and unobserved expenditures. (paper presentation) *Fourth Italian Congress of Econometrics and Empirical Economics*, Pisa, Italy, January 19-21, 2011.

References

Prof. Erich Battistin
University of Padova
Department of Statistics
via Cesare Battisti 241-243
35121, Padova, Italy.
Phone: +39 049 827 4165
e-mail: erich.battistin@unipd.it

Prof. Arthur Lewbel
Boston College
Department of Economics
140 Commonwealth Avenue
Chestnut Hill, MA, USA.
e-mail: lewbel@bc.edu