Working Paper Series, N. 4, April 2012

Identification and Estimation of Engel Curves with Endogenous and Unobserved Expenditures

Erich Battistin
Department of Statistical Sciences
University of Padua
Italy

Michele De Nadai
Department of Statistical Sciences
University of Padua
Italy

Abstract: When dealing with the estimation of Engel curves, measurement errors in expenditure data and simultaneity are likely sources of endogeneity. In this paper we study identification of the parameters that characterize an Engel curve in the presence of both. We consider specifications where budget shares are polynomials in the logarithm of total expenditure, which is the case frequently encountered in empirical applications. We propose an estimation procedure which is an extension of that in Lewbel (1996), and exploits a control function assumption to correct for the endogeneity of the true unobserved total expenditure.

Keywords: Control function, Engel-curves, Errors-in-variables models.

This version (2012-04-24)

Contents

Department of Statistical Sciences
Via Cesare Battisti, 241
35121 Padova
Italy

tel: +39 049 8274168
fax: +39 049 8274170
http://www.stat.unipd.it

Corresponding author:
Michele De Nadai
tel: +39 049 827 4151
denadai@unipd.it

Identification and Estimation of Engel Curves with Endogenous and Unobserved Expenditures

Erich Battistin
Department of Statistical Sciences
University of Padua
Italy

Michele De Nadai
Department of Statistical Sciences
University of Padua
Italy

Abstract: When dealing with the estimation of Engel curves, measurement errors in expenditure data and simultaneity are likely sources of endogeneity. In this paper we study identification of the parameters that characterize an Engel curve in the presence of both. We consider specifications where budget shares are polynomials in the logarithm of total expenditure, which is the case frequently encountered in empirical applications. We propose an estimation procedure which is an extension of that in Lewbel (1996), and exploits a control function assumption to correct for the endogeneity of the true unobserved total expenditure.

Keywords: Control function, Engel-curves, Errors-in-variables models.

1 Introduction

The choice of the most reliable empirical strategy to employ for understanding demand patterns is an issue which is not uncontroversial. There is no general consensus on the specific functional form, on how to address the endogeneity of consumption, on how to model the effect of unobserved prices, and on the estimation approach to employ. The aim of this paper is to address one specific aspect that hampers estimation of the parameters of an Engel curve, by contributing to the literature with an operational strategy to overcome the effects of measurement error in expenditure data.

Significant progress has been made in the recent years to understand the nature of the endogeneity problem arising from error ridden data in demand analysis, and its implications for drawing robust policy conclusions. When considering how expenditure shares vary with total expenditure, even the simplest form of measurement error enters non-linearly in both the right and the left hand side of the equation, thus invalidating the classical assumptions invoked in textbook models. Endogeneity of total expenditure in the regression arises because of the nature of the measurement error, thus invalidating the conventional instrumental variable approach to estimation (see Amemiya 1985).

In this paper we start by making the very simple point that the empirical challenges arising from measurement error come on top of the endogeneity of total expenditure that may be already at work with error free data. The most common interpretation of this problem builds upon a two stage budgeting idea, where in the first step the allocation of total expenditure across time periods is determined, and then the within period allocation is decided. If heterogeneity in preferences is correlated with unobserved taste shifters in the demand system, one would obtain that the residuals of the latter are correlated, across individuals, with the allocation of resources over time, and therefore with total expenditure.

Thus, in empirical applications one would need to follow a strategy which solves for the endogeneity of total expenditure, and at the same time is robust to the presence of measurement error in the data. A bottom up approach to the problem starts by considering estimation of the Engel curve when total expenditure is endogenous, but there is no measurement error in the data. In this situation, identification is achieved through exogenous variability using a standard instrumental variable approach. It is not difficult to show, as we will do further below, that the same procedure will in general yield biased results if expenditure data are measured with error, even if one is willing to make the assumption that the latter is not correlated with the instrumental variable employed.

Similarly, one could deal with measurement error in expenditure data by addressing the difficulties arising because of the nature of the equation being estimated, and hoping that this represents the way to draw correct inference. The procedure developed to this end by Lewbel (1996) works under the assumption that, with error free data, total expenditure is exogenous. As a matter of fact, this assumption is not uncontroversial (see, for example, the discussion in Blundell, Chen, and Kristensen 2007, and in Attanasio, Battistin, and Mesnard 2012).

It follows that the procedures available to estimate Engel curves represent essential tools to solve for specific sources of endogeneity, but do not offer a general solution in the presence of concurrent sources. This is the gap that the paper aims to fill in dealing with measurement error, and marks something of a departure from previous work in the literature.

A problem worth discussing is the type of instrumental variation that is needed to achieve identification for the case at hand. Throughout this paper we will maintain the assumption that a valid instrumental variable is available to deal with the endogeneity of total expenditure that would arise in the absence of measurement error. The validity of the exclusion restriction would of course depend on the expectations about the mechanism at work while estimating the Engel curve. To derive the identification result contained in this paper, it is essential that such instrumental variation is not related to measurement error, and thus serves a “magic bullet” for both the sources of endogeneity that we consider.

In the empirical literature on the estimation of Engel curves, it is common practice to think of measurement error as the main (if not the only) source of endogeneity. In this context income typically serves as an instrumental variable, since it induces variation in expenditure which is arguably not related to measurement error (see, for instance, Hausman, Newey, and Powell 1995, Lewbel 1996, Lyssiotou, Pashardes, and Stengos 1999, Brannlund and Nordstrom 2004, Kedir and Girma 2007). Some other papers, like Blundell, Duncan, and Pendakur (1998), Attanasio and Lechene (2002), Blundell, Chen, and Kristensen (2007) and Attanasio, Battistin, and Mesnard (2012), maintain the assumption that data are error-free, and argue that the endogeneity of total expenditure originates from the underlying economic theory. In the latter case, the existence of a valid instrument has to be discussed depending on the expectations on the nature of the endogeneity, and income does not necessarily offers a valid solution. Attanasio, Battistin, and Mesnard (2012) provide some detailed discussion on the topic.

The point worth making here is that a valid instrument in the absence of measurement error, most likely would also serve as a magic bullet in the sense described above. Quite on the contrary, instrumental variation exploited when endogeneity is totally attributed to measurement error, may not necessarily help identification.

The main results of the paper can be summarized as follows. First, we show that when total expenditure in the Engel curve is treated as endogenous and its measurements are error ridden, the estimation methods usually employed yield biased results for the parameters of interest. A standard instrumental variable approach works only in the absence of measurement error, while the procedure proposed by Lewbel (1996) corrects for measurement error when the latter is considered the only source of endogeneity. If one wants to estimate Engel curves allowing for endogenous expenditures and correcting for measurement error, which we claim is the relevant situation in most empirical applications, neither of these two methods alone provides correct inferential conclusions. Second, we show that, under the conditions stated, a standard instrumental variable approach yields upward biased results for the parameters regulating the shape of the Engel curve. Thus, the effects of measurement error are at odds with the usual attenuation bias found in the literature for the case of linear specifications. Third, we propose a method to estimate the parameters of the Engel curve for the case at hand. We take a control function approach, and derive the conditions under which the availability of an instrument for total expenditure is sufficient to retrieve estimates of the Engel curve parameters that are robust to measurement error in the data. As we will discuss further below, these conditions are very general in nature, or at least are as general as those already presented in other studies that consider estimation of the Engel curves when measurement error is the only source of endogeneity (Lewbel 1996). The results we provide may be extended to allow for exogenous error-free regressors, thus offering a practical way to estimate more general demand systems. The finite sample properties of the proposed estimator are evaluated with a Monte Carlo simulation study and compared to those of alternative estimators. Finally, we provide an empirical application to show how the method we propose can be applied to real data, using information from the Bank of Italy panel survey.

The remainder of the paper is organised as follows. Section 2 defines the model under study. Section 3 derives the identification results for the case of budget shares which are linear or quadratic specifications in the logarithm of total expenditure. Section 4 develops the estimation procedure, whereas Section 5 presents the results from the Monte Carlo study. Section 6 discusses the empirical application, and Section 7 concludes. Proofs and additional materials are made available in the Appendix.

2 General formulation of the problem

The aim of this Section is twofold. First, we define the channels through which endogeneity of total expenditure operates, and how these affect the various estimation methods employed. We then set out the assumptions about the measurement error model that we will maintain throughout, replicating the same setting already considered by Lewbel (1996).

2.1 Endogeneity of total expenditure

We focus throughout on identification of the Engel curve for a single good. If one is interested in the estimation of a system of I equations, the same procedure applies by treating each good separately and discarding the one equation which is uniquely identified by the summing-up properties of demand functions. We restrict our attention to specifications in which budget shares are polynomials in the logarithm of total expenditure. This approach is quite general, and underpins most of the relevant specifications encountered in the empirical literature. Notable examples are the AIDS Deaton and Muellbauer (1980) or the Quadratic AIDS Banks, Blundell, and Lewbel (1997), which correspond to polynomials of the first and second order, respectively (for a detailed discussion on Engel curves, see Lewbel 2008). To ease notation, in most of the paper we derive identification results for the following equation:

$$W_i^* = b_{i0} + b_{i1} \log X^* + \varepsilon_i, \tag{1}$$

and we discuss separately the extension to higher order polynomials. In the notation employed, $W_i^* \equiv Y_i^*/X^*$ is the budget share on the i-th good, Y_i^* being expenditure on the i-th good, while $X^* \equiv \sum_{i=1}^{I} Y_i^*$ is total expenditure. In what follows, $\boldsymbol{b}_i \equiv (b_{i0}, b_{i1})$ denotes the vector of the parameters of interest, and variables indexed with a star refer to error-free measurements..

In most empirical applications, one would estimate (1) using instrumental variables (see, for instance, Blundell, Chen, and Kristensen 2007 and Attanasio, Battistin, and Mesnard 2012). This is motivated by the fact that there is no clear economic justification to assume exogeneity of expenditure on the right had side of the equation, as in general X^* and Y_i^* may be chosen simultaneously by individuals. Provided that a set of valid instruments (Z) is available, $\boldsymbol{b}_i$ is consistently estimated by 2SLS or by a standard control function estimator. The latter approach would entail considering the following regression:

$$\log X^* = g(Z) + \eta^*, \tag{2}$$

where $g(Z) \equiv E[\log X^*|Z]$. In the remainder of this paper, we make the following control function restriction. This is a standard assumption in the relevant literature and it is also employed in semiparametric approaches (see, for example, Blundell, Duncan, and Pendakur 1998).

Assumption 1. ***(Control Function Restriction).*** *Let η^*, Z and ε_i be such that:*

$$E[\varepsilon_i|Z,\eta^*] = E[\varepsilon_i|\eta^*] = \rho_i\eta^*.$$

Assumption 1 is more restrictive than it is required to identify $\boldsymbol{b}_i$ if a valid instrument is available, but will be needed in the following to handle non-linearities introduced by measurement error. It implies that identification of the parameters of interest can be achieved from the following regression:

$$W_i^* = b_{i0} + b_{i1}\log X^* + \rho_i\eta^* + \xi_i, \tag{3}$$

where there is $E[\xi_i|X^*,\eta^*] = 0$ by *construction.* Using standard arguments, one would estimate $\boldsymbol{b}_i$ from a linear regression of W_i^* on $\log X^*$ and $\hat{\eta}^*$, the latter term denoting the estimated residual from the regression in (2).

2.2 Measurement error

In this Section we spell out the properties of measurement error that will be used later in the paper. If expenditure data are mismeasured, another source of endogeneity in the estimation of (1) arises. Denoting by W_i and X the error ridden measurements of W_i^* and X^*, respectively, the feasible counterpart of equation (1) obtained by regressing W_i on $\log X$ would in general yield biased estimates of the parameters of interest. Moreover, as Amemiya (1985) first pointed out, in non-linear settings instrumental variables do not help identification. The result follows from measurement error being is no longer additively separable in the functional form specification, and entering both sides of equation (1). Such a feature, not usually encountered in the errors-in-variables literature, further complicates identification (see De Nadai and Lewbel 2012, for an example of this).

Suppose that Y_i is observed in place of Y_i^*, defined as:

$$Y_i = Y_i^* + X^*\nu_i, \tag{4}$$

where ν_i is a mean zero random variable independent of Y_j^*, for $j = 1,\dots,I$, and hence from X^*. This definition is consistent with observing (possibly correlated) measurement errors in all goods. Note that this also allows for the variance of measurement error on expenditure levels to increase with total expenditure, a feature usually encountered in the data (see Bound, Brown, and Mathiowetz 2001). The rationale for this specification follows from the fact that summing up over all goods we obtain classical measurement error in $\log X^*$, since there is:

$$X = \sum_{i=1}^{I} Y_i = X^*\left(1 + \sum_{i=1}^{I}\nu_i\right) = X^*V, \tag{5}$$

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

$$\log X = \log X^* + \log V. \tag{6}$$

Equation (5) together with (1) implies:

$$W_i = \frac{W_i^* + \nu_i}{V}, \tag{7}$$

so that measurement error enters non-linearly the left hand side of equation (1). Such a measurement error structure coincides with that considered by Hausman, Newey, and Powell (1995) and Lewbel (1996).

3 Identification

This Section is organised as follows. First, we set out the identification strategy for Engel curves that are *linear* in the logarithm of total expenditure, relying on the control function restriction in Assumption 1. The main result is presented in Theorem 1, where (15) represents the estimating equation that we propose to use in empirical applications. Second, we discuss some threats to the validity of the control function restriction, providing conditions to test it against data. We show that in the worse case scenario, our procedure still retrieves the shape parameter of the Engel curve, which represents the quantity of interest in most applications (see, for example, Attanasio, Battistin, and Mesnard 2012). Finally, we discuss the generalization of the identification result to the case of a *quadratic* specification for the Engel curve. As discussed in the Introduction, this - together with the linear case that we consider as working example - covers most of the empirical applications encountered in the empirical literature.

3.1 Linear Engel curves

Consider the following set of assumptions, which will provide the basis for the identification results derived below.

Assumption 2. ***(Validity of the Instruments).*** *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) $E[\nu_i] = 0$ *and* $\nu_i \perp (X^*, Z, \varepsilon_i)$*.*

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. Full independence is required due to the non-linearities in the functional form considered. Note also that (iii) implies $E[V] = 1$.

Substitute equation (3) into (7) to obtain:

$$W_i = \frac{b_{i0} + b_{i1} \log X^* + \rho_i \eta^* + \xi_i + \nu_i}{V}.$$

Under Assumption 2, by multiplying either side by X and taking conditional expectations with respect to Z, there is:

$$E[XW_i|Z] = b_{i0}E[X^*|Z] + b_{i1}E[X^* \log X^*|Z] + \rho_i E[X^* \eta^*|Z]. \tag{8}$$

Following Lewbel (1996), it is easy to see that:

$$E[X^*|Z] = E[X|Z], \tag{9}$$
$$E[X^* \log X^*|Z] = E[X \log X|Z] - E[X|Z]E[V \log V], \tag{10}$$

so that substitution of (9) and (10) into (8) yields:

$$E[Y_i|Z] = \tilde{\alpha}_{i1} E[X|Z] + b_{i1} E[X \log X|Z] + \rho_i E[X^* \eta^*|Z], \tag{11}$$

where $\tilde{\alpha}_{i1} \equiv b_{i0} - b_{i1} E[V \log V]$.

When $E[\varepsilon_i|X^*] = 0$, then ρ_i is equal to zero and the last term on the right hand side of equation (11) vanishes, implying that b_{i1} is identified through a 2SLS regression of Y_i on X and $X \log X$ without a constant, using Z as instruments. Identification of b_{i0} follows along the same lines exploiting similar expressions for $E[X^l W_i|Z]$, with $l \geq 1$ (see Lewbel 1996). When $\rho_i \neq 0$, this procedure would in general produce incorrect inference for b_{i1} because of an omitted variable problem.

In what follows, we will express $E[X^* \eta^*|Z]$ in terms of observable moments. Define η as the residual term from the regression of $\log X$ on the set of instruments Z. That is, η is the analogue of η^* when $\log X$ is substituted for $\log X^*$ into equation (2). It follows from the measurement error structure in equation (6) that:

$$\eta = \eta^* + \log V - E[\log V]. \tag{12}$$

Now consider the conditional expectation:

$$E[X\eta|Z] = E[X^* \eta^*|Z] + E[X^*|Z](E[V \log V] - E[\log V]),$$

where we exploit once again the independence of V from (X^*, η^*) and the fact that $E[V] = 1$. This, together with equation (9), implies:

$$E[X^* \eta^*|Z] = E[X\eta|Z] - E[X|Z](E[V \log V] - E[\log V]). \tag{13}$$

Hence substituting (13) into (11) and rearranging terms we obtain:

$$E[W_i X|Z] = \alpha_{i1} E[X|Z] + b_{i1} E[X \log X|Z] + \rho_i E[X\eta|Z], \tag{14}$$

with $\alpha_{i1} = b_{i0} - b_{i1} E[V \log V] - \rho_i Cov(V, \log V)$.

This result can be seen as the particular case of the following theorem, that generalizes the above argument to the conditional expectation $E[X^l W_i|Z]$, for any l, and whose proof is given in the Appendix A.

Theorem 1. ***(Identification of Linear Curves).*** *Let equations* (1) *and* (4) *hold. Under Assumptions 1 and 2, for any integer l for which* $E[V^l \log V]$ *and* $E[\nu_i V^{l-1}]$ *are finite there is:*

$$E[X^l W_i|Z] = \alpha_{il} E[X^l|Z] + \beta_{il} E[X^l \log X|Z] + \tilde{\rho}_{il} E[X^l \eta|Z], \tag{15}$$

where η *is defined as in equation* (12)*, and:*

$$\alpha_{il} = b_{i0} \frac{E[V^{l-1}]}{E[V^l]} - b_{i1} \frac{E[V^{l-1}]E[V^l \log V]}{E[V^l]^2} - \rho_i \frac{E[V^{l-1}]Cov(V^l, \log V)}{E[V^l]^2} + \frac{E[V^{l-1}\nu_i]}{E[V^l]},$$

$$\beta_{il} = b_{i1} \frac{E[V^{l-1}]}{E[V^l]}, \qquad \tilde{\rho}_{il} = \rho_i \frac{E[V^{l-1}]}{E[V^l]}.$$

The moments restrictions in (15) imply that a 2SLS regression of $X^l W_i$ on X^l, $X^l \log X$ and $X^l \hat{\eta}$, using Z as instruments, would consistently estimate α_{il}, β_{il} and $\tilde{\rho}_{il}$. As before, $\hat{\eta}$ represents the empirical analogue of (12) obtained from the feasible regression of $\log X$ on Z.

As we will discuss in the next Section, Theorem 1 defines a set of moment conditions corresponding to different values of l that can be used to estimate all the parameters of the Engel curve, as well as ρ_i. More in general, the theorem offers an important insight on the bias resulting from the application of a standard instrumental variable strategy to estimate the parameters of the curve. The result is easily obtained upon discussing the properties of the estimating equation defined by (15), once l is set to zero.

Corollary 1. ***(Failure of the Instrumental Variable Estimator).*** *Under the Assumptions of Theorem 1, there is for $l = 0$:*

$$E[W_i|Z] = \alpha_{i0} + \beta_{i0} E[\log X|Z], \tag{16}$$

where:

$$\begin{aligned} \alpha_{i0} &= E[v_i V^{-1}] + E[V^{-1}] \left(b_{i0} - b_{i1} E[\log V] \right), \\ \beta_{i0} &= b_{i1} E[V^{-1}]. \end{aligned} \tag{17}$$

Equation (16) implies that a 2SLS regression of W_i on $\log X$, using Z as instruments, yields biased results for b_{i1}. This proves that instrumenting for endogeneity without adjusting for the non-linearities introduced by measurement error will in general result in incorrect inference on the parameters of interest.

Three implications of practical relevance are worth noting from Corollary 1, which were left sort of implicit in the discussion by Lewbel (1996). First, from Jensen's inequality there is $E[V^{-1}] > E[V]^{-1} = 1$, hence the *naive* instrumental variable estimator is biased upward. Second, by taking a second order Taylor series expansion of $E[V^{-1}]$ around its mean, there is:

$$E[V^{-1}] \approx E[V] + Var[V] = 1 + Var[V], \tag{18}$$

this implying that the magnitude of the bias is approximately proportional to the variance of the measurement error. Note that, when V is log-normally distributed, which may be a sensible assumption to make in practice, the above approximation is exact. Finally, Corollary 1 offers an intuitive explanation for the informational content brought by the set of moment conditions defined by different values of l, and how such information helps to the identification of important features of the model. For example, one could combine equations (14) and (17) to jointly estimate b_{i1} and $E[V^{-1}]$, and thus the variance of measurement error if one is willing to assume that the latter is log-normally distributed. We will come back to this point in the Section about estimation.

3.2 Validity of the control function restriction

The estimation strategy brought forward through equation (14) requires some careful discussion about the nature of the control function term $E[X\eta|Z]$. It is crucial for identification that this term is not collinear with the remaining terms that enter the moment equation. It turns out that to achieve identification of all parameters of the Engel curve there must a certain degree of dependence between η^* and the set of instruments Z. Note that η^* is uncorrelated by construction with the instruments Z, therefore dependence between η^* and Z might only be due to higher order moments, for example through heteroskedasticity of η^* with respect to the instruments Z. To see this, rewrite equation (2) as $X^* = e^{g(Z)}e^{\eta^*}$, so that there is:

$$X^* = E[X^*|Z]\frac{e^{\eta^*}}{E[e^{\eta^*}|Z]}.$$

If η^* is stochastically independent of Z, it is easy to see that:

$$E[X^*\eta^*|Z] = \frac{E[e^{\eta^*}\eta^*]}{E[e^{\eta^*}]}E[X^*|Z] = \delta E[X^*|Z].$$

It is then immediately clear that, when substituting the above expression back into equation (11), and using (9), the equation in (14) becomes:

$$E[W_iX|Z] = (b_{i0} - b_{i1}E[V\log V] + \rho_i\delta)\, E[X|Z] + b_{i1}E[X\log X|Z]. \tag{19}$$

It follows that, when η^* is independent of Z, Lewbel's (1996) estimator provides consistent estimates for b_{i1} in the presence of endogenous unobserved total expenditure X^*. However, b_{i0} is no longer identified from knowledge of moments of the form $E[X^lW_i|Z]$, for $l \geq 2$, since b_{i0} and δ cannot be disentangled without additional information.

It is worth noting that, under Assumption 2, the independence condition required to avoid collinearity can be tested against data, as η^* is independent of Z if and only if η is independent of Z. In the remainder of this paper we will work as if this condition is met in the data, and we will test for this in the empirical application.

3.3 Quadratic Engel curves

The generalization to the case of quadratic Engel curves is readily obtained at the cost of complicating the algebra. Consider the following specification:

$$W_i^* = b_{i0} + b_{i1}\log X^* + b_{i2}(\log X^*)^2 + \varepsilon_i. \tag{20}$$

The same arguments employed above allow us to state the following theorem.

Theorem 2. ***(Identification of Quadratic Curves).*** *Let equations* (20) *and* (4) *hold. Under Assumptions 1 and 2, for any integer l for which $E[V^l\log V]$ and $E[\nu_iV^{l-1}]$ are finite, there is:*

$$E[X^lW_i|Z] = \alpha_{il}E[X^l|Z] + \beta_{il}E[X^l\log X|Z] + \gamma_{il}E[X^l(\log X)^2|Z] + \tilde{\rho}_{il}E[X^l\eta|Z], \tag{21}$$

where η is defined as in equation (12)*, and:*

$$\begin{aligned}
\alpha_{il} &= b_{i0}\frac{E[V^{l-1}]}{E[V^l]} - b_{i1}\frac{E[V^{l-1}]E[V^l \log V]}{E[V^l]^2} \\
&\quad -b_{i2}E[V^{l-1}]\left\{\frac{E[V^l(\log V)^2]}{E[V^l]^2} - 2\frac{E[V^l \log V]^2}{E[V^l]^3}\right\} \\
&\quad -\rho_i\frac{E[V^{l-1}]Cov(V^l, \log V)}{E[V^l]^2} + \frac{E[V^{l-1}\nu_{ih}]}{E[V^l]}, \\
\beta_{il} &= b_{i1}\frac{E[V^{l-1}]}{E[V^l]} - 2b_{i2}\frac{E[V^{l-1}]E[V^l \log V]}{E[V^l]^2}, \\
\gamma_{il} &= b_{i2}\frac{E[V^{l-1}]}{E[V^l]}, \\
\tilde{\rho}_{il} &= \rho_i\frac{E[V^{1-1}]}{E[V^l]}.
\end{aligned}$$

Theorem 2, whose proof is reported in Appendix A, provides moment conditions for the estimation of $\boldsymbol{b}_i$ in a way completely similar to Theorem 1. In particular consider the following result:

Corollary 2. *Under the Assumptions of Theorem 2, when $l = 1$ there is:*

$$E[XW_i|Z] = \alpha_{i1}E[X|Z] + \beta_{i1}E[X \log X|Z] + \gamma_{i1}E[X(\log X)^2|Z] + \tilde{\rho}_{i1}E[X\eta|Z],$$

where:

$$\begin{aligned}
\alpha_{i1} &= b_{i0} - b_{i1}E[V \log V] - b_{i2}\left\{E[V(\log V)^2] - 2E[V \log V]^2\right\} - \rho_i Cov(V, \log V), \\
\beta_{i1} &= b_{i1} - 2b_{i2}E[V \log V], \\
\gamma_{i1} &= b_{i2}, \\
\tilde{\rho}_{i1} &= \rho_i.
\end{aligned}$$

This implies that a 2SLS regression of XW on X, $X \log X$, $X(\log X)^2$ and $X\hat{\eta}$, using Z as instruments, would consistently estimate the quadratic coefficient b_{i2} through the coefficient on $X(\log X)^2$. As for the linear case, identification of the remaining components of $\boldsymbol{b}_i$ is achieved through the additional moment restrictions implied by equation (21) for different values of l.

4 Estimation

The results in Section 3 imply that equation (15) can be used for $l = 1$ to estimate b_{i1} and ρ_i. Although η is not observed, it may be estimated through a (non)parametric regression of the observed $\log X$ on the instruments Z, and then plugged into the main regression. In what follows, we discuss how b_{i0} can be retrieved from raw data. The general setting arising here is similar to that considered in Lewbel (1996), the only difference being the additional term $\rho_i Cov(V, \log V)$.

First, note that the entire distribution of V is identified by assuming the existence of its moment generating function. The result follows using the additional restrictions provided by equation (15) in Theorem 1 when $l \neq 1$, as one could identify any moment of the distribution of V from knowledge of $\boldsymbol{\beta}_i = (\beta_{i0}, \beta_{i1}, \ldots)$. For instance, it is easy to see that $\beta_{i2}/\beta_{i1} = E[V^2]$. This in turn allows to nonparametrically estimate both $E[V \log V]$ and $Cov(V, \log V)$ following the arguments in Lewbel (1996).

In empirical applications, however, there is little scope for using large values of l because of the drawbacks on the standard errors of the β_{il}'s. One can get around this problem by imposing parametric assumptions on the distribution of V. Suppose, for instance, that V is log-normally distributed. This assumption is rather appealing, as there is empirical regularity from various surveys worldwide suggesting that total expenditure is log-normally distributed (see Battistin, Blundell, and Lewbel 2009 for a discussion on the reasons for this pattern). Using the Cramér's (1936) characterization result for normal distributions (see Johnson, Kotz, and Balakrishnan 1994, pag. 102-103) and the fact that X^* is independent of V, log-normality of raw expenditure data implies that V must be itself log-normal.

The ratio of β_{i1} to β_{i0}, obtained from equations (14) and (16), respectively, identifies $E[V^{-1}]$. Thus, using (18) and the assumption of log-normality of V, a method of moments estimate for the variance of $\log V$ (σ_V^2) is obtained through:[1]

$$\sigma_V^2 = \log\left(\frac{\beta_{i0}}{\beta_{i1}}\right),$$

using the fact that $1 + Var[V] = e^{\sigma_V^2}$. This approach is to be preferred in general to the one proposed by Lewbel (1996) which is based on knowledge of β_{i1} and β_{i2} since the variance of β_{i2} is generally much larger than that of β_{i0}.[2]

With the distribution of V at hand, one may estimate $E[V \log V]$ and $Cov(V, \log V)$ and substitute these back into the expression for α_{i1}, hence determining b_{i0}. For example, under log-normality of V there is:

$$\begin{aligned} E[V \log V] &= \frac{\sigma_V^2}{2}, \\ Cov(V, \log V) &= \sigma_V^2. \end{aligned}$$

5 Monte Carlo Simulation

To assess the finite sample properties of the proposed estimator, a 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 was given in Section 3, and to the one proposed by Lewbel (1996). We consider the following model:

$$\begin{aligned} W_i^* &= 1 - 0.05 \log X^* + \varepsilon_i, \\ \log X^* &= 1.45 + 0.93 \log Z + -0.03 (\log Z)^2 + \eta^*, \end{aligned}$$

[1] Note that, if V is log-normal, the fact that $E[V] = 1$ implies that $E[\log V] = -\sigma_V^2/2$, meaning that the only parameter to be estimated is σ_V^2.

[2] Combining estimates of β_{il} for several values of l in a GMM framework, in a manner similar to that discussed in Lewbel (1996), would in general increase the efficiency of the resulting estimate.

where we set $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$ and $\log Z \sim N(10.5, 2.5^2)$. Endogeneity of the unobserved X^* is induced by generating η^* according to:

$$\begin{aligned} \varepsilon_i &= \theta_1 \eta^* + \xi, \\ \eta^* &\sim N(0, kZ^{0.3}), \end{aligned}$$

with $\xi \sim N(0, \sigma_\xi^2)$. The parameters θ_1, k and σ_ξ^2 were chosen to get $Corr(\varepsilon_i, \log \eta^*) =$ 0, 0.3, 0.5 and 0.8, and to keep the R^2 of the first and second stage regressions, in the case of no endogeneity, at about 0.75. The parameters of the Engel curve and of the first stage were calibrated such that the marginal distributions of $\log X^*$ and $\log Z$ roughly match the observed distributions in the data used for the application in the next Section, while retaining sufficient variability in W_i^*.

Measurement error of the form outlined in Section 3 is introduced, so that the observed pair (Y_i, X) is given by:

$$\begin{aligned} Y_i &= Y_i^* + X^* v_i, \\ \log X &= \log X^* + \log V, \end{aligned}$$

where $V = 1 + v_i$, hence assuming that only the expenditure of the good under study Y_i^* is measured with error. The amount of measurement error is decided by setting the noise to signal ratio, that is $\frac{Var(\log V)}{Var(\log X^*)}$, to 0, 0.1, 0.3 or 0.5.

We compare the performance of the proposed estimator with OLS, 2SLS and Lewbel's (1996) estimator using 10,000 replications from the process sketched above. To ensure comparability with the study by Lewbel (1996) the set of instruments is defined as: Z, $\log Z$, $(\log Z)^2$, $Z \log Z$, Z^2 and $Z^2 \log Z$. The proposed estimator is computed as in Section 4, by constructing a control function which is the interaction between the observed X and the residuals of the regression of $\log X$ on the set of instruments Z.

The results of the simulation are summarized in the tables below. Two scenarios are considered, defined by values of the sample size equal to 1,000 (Table 1) and 5,000 (Table 2). Finite sample properties were also investigated for samples of 500 and 10,000 observations, for which results are reported in Appendix B. The presentation of the results is organised as follows. The left hand side of the top panel of each table considers the case of no measurement error and exogenous expenditures, for which OLS estimates should be preferred. By moving to the right of the same panel endogeneity of expenditures is added, so that IV estimates should be preferred. By moving down in the table, increasingly larger measurement error is added for scenarios defined by values of the noise to signal ratio set at 10%, 30% and 50% of the variability in observed expenditure. Thus, figures in the first column of each table are derived under the setup considered by Lewbel (1996). Results at the bottom end of each table are derived for the case considered in this paper, for different combinations of measurement error and endogeneity of expenditure.

As expected, departures from standard assumptions have strong negative impact on the properties of the OLS estimator: already at relatively small values of measurement error or of endogeneity, the percentage bias is substantial. Similarly, in the absence of measurement error, the IV estimator does a pretty good job at dealing

Table 1: Percentage Bias for $\hat{\beta}_1$ defined as $(\hat{\beta}_1 - b_1)/|b_1|$. Sample size: 1,000.

N/S		Extent of Endogeneity							
Ratio		0		0.2		0.5		0.8	
0	**OLS**	0.001	(*0.727*)	1.91	(*0.958*)	6.922	(*1.423*)	15.074	(*2.021*)
	IV	-0.002	(*0.776*)	0.032	(*1.013*)	0.061	(*1.411*)	0.055	(*1.668*)
	Lewbel (1996)	0.033	(*2.399*)	3.735	(*3.774*)	13.438	(*8.71*)	29.201	(*17.232*)
	CF	0.002	(*3.878*)	1.766	(*5.891*)	6.395	(*13.086*)	13.924	(*25.63*)
0.1	**OLS**	69.428	(*8.076*)	71.29	(*8.11*)	76.179	(*8.199*)	84.134	(*8.341*)
	IV	–6.273	(*8.005*)	–6.252	(*8.03*)	–6.207	(*8.08*)	–6.092	(*8.111*)
	Lewbel (1996)	7.275	(*35.27*)	10.841	(*35.194*)	20.195	(*35.488*)	35.401	(*37.434*)
	CF	-0.108	(*54.088*)	1.974	(*53.499*)	7.441	(*52.746*)	16.34	(*54.006*)
0.3	**OLS**	156.999	(*12.45*)	158.82	(*12.482*)	163.601	(*12.562*)	171.38	(*12.687*)
	IV	–17.888	(*13.098*)	–17.864	(*13.119*)	–17.801	(*13.165*)	–17.65	(*13.194*)
	Lewbel (1996)	19.986	(*62.453*)	23.335	(*62.07*)	32.134	(*61.366*)	46.461	(*61.18*)
	CF	–14.021	(*83.572*)	–11.33	(*82.566*)	–4.263	(*80.366*)	7.238	(*78.336*)
0.5	**OLS**	204.639	(*14.147*)	206.453	(*14.17*)	211.209	(*14.226*)	218.94	(*14.31*)
	IV	–27.331	(*15.879*)	–27.316	(*15.903*)	–27.263	(*15.927*)	–27.119	(*15.951*)
	Lewbel (1996)	33.116	(*82.642*)	36.3	(*82.069*)	44.649	(*80.768*)	58.214	(*79.324*)
	CF	–23.881	(*94.703*)	–20.856	(*93.687*)	–12.92	(*91.329*)	-0.019	(*88.575*)

Note. Simulation results are presented using the data generating process described in Section 5. The performance of alternative estimators for the shape parameter of a linear Engel curve is evaluated, the true value of the parameter being -0.05. The label "OLS" refers to results from the regression of the budget share on logged total expenditure. The label "IV" refers to results obtained instrumenting expenditure by Z, $\log Z$ and $Z \log Z$. The label "Lewbel (1996)" refers to results obtained following the procedure in Lewbel (1996). The label "CF" refers to results from the control function approach proposed in this paper, which is operationally implemented using the result in Theorem 1. Different combinations of the noise to signal ratio and the extent of endogeneity of total expenditure are considered by row and column, respectively. Standard error of the percentage bias is reported in parenthesis.

Table 2: Percentage Bias for $\hat{\beta}_1$ defined as $(\hat{\beta}_1 - b_1)/|b_1|$. Sample size: 5,000.

N/S		**Extent of Endogeneity**							
Ratio		**0**		**0.2**		**0.5**		**0.8**	
0	**OLS**	0.004	*(0.323)*	1.91	*(0.426)*	6.913	*(0.631)*	15.048	*(0.891)*
	IV	0.005	*(0.343)*	0.015	*(0.446)*	0.02	*(0.619)*	0.013	*(0.732)*
	Lewbel (1996)	0.015	*(1.065)*	3.278	*(1.892)*	11.836	*(4.915)*	25.747	*(10.038)*
	CF	-0.01	*(1.637)*	0.468	*(2.816)*	1.724	*(7.225)*	3.775	*(14.769)*
0.1	**OLS**	69.35	*(3.605)*	71.211	*(3.622)*	76.095	*(3.665)*	84.039	*(3.736)*
	IV	–6.673	*(3.54)*	–6.669	*(3.55)*	–6.656	*(3.58)*	–6.633	*(3.604)*
	Lewbel (1996)	2.149	*(15.887)*	5.366	*(15.897)*	13.804	*(16.349)*	27.52	*(18.265)*
	CF	0.582	*(23.481)*	1.145	*(23.197)*	2.623	*(23.16)*	5.024	*(25.235)*
0.3	**OLS**	156.847	*(5.522)*	158.666	*(5.534)*	163.44	*(5.565)*	171.205	*(5.612)*
	IV	–18.149	*(5.763)*	–18.142	*(5.769)*	–18.12	*(5.791)*	–18.09	*(5.807)*
	Lewbel (1996)	7.005	*(29.854)*	10.145	*(29.759)*	18.378	*(29.722)*	31.76	*(30.322)*
	CF	-0.056	*(42.347)*	0.732	*(41.704)*	2.796	*(40.506)*	6.15	*(40.155)*
0.5	**OLS**	204.17	*(6.447)*	205.978	*(6.456)*	210.723	*(6.479)*	218.438	*(6.517)*
	IV	–27.783	*(6.918)*	–27.778	*(6.928)*	–27.764	*(6.952)*	–27.726	*(6.971)*
	Lewbel (1996)	11.648	*(40.383)*	14.706	*(40.274)*	22.733	*(40.12)*	35.793	*(40.297)*
	CF	–7.501	*(56.512)*	–6.345	*(55.532)*	–3.316	*(53.309)*	1.599	*(50.993)*

Note. See footnote to Table 1.

with endogeneity of expenditures. As documented in Section 4 (Corollary 1), the bias of the IV estimator is unaffected by the extent of endogeneity when measurement error is added to the model. The performance of the estimator worsen as measurement error becomes more important, yielding larger bias which is proportional to the variance of the error.

The estimator proposed by Lewbel (1996) generally outperforms OLS and IV when measurement error comes into play, although this is less so in the presence of sizeable endogeneity of expenditure. The estimator proposed in this paper adjusts for both measurement error and endogeneity, reflecting the properties discussed in Section 3. It is of quality comparable to the estimator previously proposed by Lewbel (1996) when there is no endogeneity, although this result comes at the cost of precision. It is however worth noting that, already for limited extents of endogeneity, the current estimator outperforms its competitors, uniformly across the various scenarios considered for the sample size.

6 Application

In this Section we present an application using data from the 2010 wave of the Bank of Italy's *Survey on Households' Income and Wealth* (SHIW). We select the subsample of couples for which the male is between 30 and 60 years old, resulting in a sample of $2,723$ households. Information is available for expenditures on a variety of commodities, demographics and wages. We consider different groups, and run separate regressions depending on the number of children in the household (couples without children, couples with one child, and couples with more than one child). We decided to focus on the estimation of Engel curves for food, which we model as a linear in logarithms Working (1943) and Leser (1963) budget share specification because of the substantial empirical evidence in support of this (see, for instance, Banks, Blundell, and Lewbel 1997). We control for household regional variation through a set of macro area dummies (North, Center, South), which enter linearly in the specification considered. Such covariates will be assumed exogenous to the model and correctly measured throughout. The main descriptive statistics for the variables employed in the analysis are documented in Appendix B.

The reference model is then:

$$W = b_0 + b_1 \log X + \gamma' Q + \varepsilon, \tag{22}$$

where Q is the vector of dummies. This is formally a shape invariant Engel curve in which demographics are taste shifters inducing heterogeneity in the utility of households (see Blundell, Duncan, and Pendakur 1998). As in Attanasio and Lechene (2002), we decided to instrument total expenditure with the average of male (logged) wages across areas. This is because income is not necessarily the ideal candidate, as endogeneity may be driven by non-separability of labour supply from food in the utility function. Mean wages most likely are not correlated with household unobserved characteristics and measurement errors, and in our data are strongly correlated with total expenditure. Very pragmatically, we decided to increase variability in the instrument by stratifying households using region identifiers and population size of the primary sampling unit, all variables required being available from public use files. This resulted in an instrument defined over 100 cells, comprising on average 27 observations. We however checked the sensitivity of our results to the choice of the instrument, experimenting with total household income in place of, and on top of, male wages. The results that we found proved informationally equivalent to those presented in what follows, and we decided to omit them from the main text.

To ensure comparability with the procedure in Lewbel (1996), the following set of instruments was considered: average of male logged wages, in both levels and logs, and their interaction (for a total of *three* instruments). The first stage regression of logged expenditure on Z and Q yields coefficients on Z which are strongly statistically significant and an F statistic of 115.47. The instruments considered account for about 16% of the total variance of observed total expenditure.

Table 7 presents results from alternative estimation approaches, that would yield correct inference on the parameters of interests depending on the features of the data generating process. To ease readability, we decided to report estimates only for the shape parameter b_1 in (22).

The first set of results refers to estimates obtained from straight OLS, hence ignoring the presence of any source of endogeneity in total expenditure. These were obtained by estimating the empirical counterpart of equation (22) from raw data. Acknowledging endogeneity of total expenditure, we implemented a *naive* 2SLS regression of W on $\log X$ and Q using the first stage regression discussed above. These are the results that we present in the second row of the table. When the two regression outputs are compared, the IV procedure yields point estimates that are in general larger, in absolute terms, than those obtained through OLS. The third set of results is obtained by replicating the procedure in Lewbel (1996), thus adjusting for measurement error. The results presented were obtained by estimating:

$$XW = b_0 X + b_1 X \log X + \gamma' QX + \zeta, \tag{23}$$

through a 2SLS procedure, in which the endogenous variables X, $X \log X$ and QX were instrumented with the Z's and their interactions with Q. Point estimates are lower, in absolute terms, than those obtained with IV. As we have discussed in Section 3, this finding in itself is consistent with having a large extent of measurement error in the data: under the assumption that endogeneity of expenditure is solely determined by error ridden data, the ratio between IV estimates and estimates obtained from equation (23) should speak about the variance of measurement error. We checked preliminarily for evidence against the stochastic independence between η and Z. We run the regression of the square of $\hat{\eta}$ on the instruments, separately for the household types considered, detecting the presence of sizeable heterosckedasticity for two of the three groups. We then estimated the following regression:

$$XW = b_0 X + b_1 X \log X + \gamma' QX + \rho X \hat{\eta} + \omega, \tag{24}$$

$\hat{\eta}$ being the residual term from the regression of $\log X$ on the Z's and their interactions with Q. It turns out that the resulting point estimates are much closer in magnitude to those obtained via IV. Intuitively, under the conditions stated in Section 3, this points to a much lower extent of measurement error in the data than before. Finally, we followed the procedure sketched in Section 4 to estimate the shape parameter of the Engel curve through a GMM system of moment conditions defined from the instrumental variable estimands resulting from equations (22) and (24). The results are presented in row (6) of the table.

To shed light on these sources of endogeneity in the data, reported in the table are estimates of the variance of measurement error obtained assuming log-normality of V. Since estimates of $E[V^{-1}]$ are produced, we employed (18) to derive the quantity of interest. The figures reported suggest that the extent of error in the data is limited. It is worth noting that the size of measurement error that we would have obtained by taking estimates in rows (2) and (3) at face value are much larger. For example, for households with more than one child, the ratio between straight IV estimate and Lewbel's estimate would yield a value for the variance which is almost six times larger than the one reported in the table.

7 Conclusions

In this paper we have proposed an estimator for Engel curves which accounts for the presence of two sources of endogeneity: measurement error on, and endogeneity of total expenditure. The estimator builds upon a standard control function assumption to derive consistent estimates of the parameters of interest. The approach suggested defines a GMM procedure which is readily implementable using standard statistical software. The small sample properties of the estimator have been analysed, and the results point to a significant improvement with respect to its alternative competitors already for small departures form the standard setting. The proposed method was applied to estimate Engel curves for food using data from the Bank of Italy's SHIW Survey. The results suggest that ignoring the presence of endogeneity of unobserved total expenditure may result in severely biased estimates. In particular, the extent of measurement error would be significantly overestimated.

References

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

Attanasio, O., E. Battistin, and A. Mesnard (2012): "Food and cash transfers: evidence from Colombia," *Economic Journal*, 122(559), 92–124.

Attanasio, O., and V. Lechene (2002): "Tests of income pooling in household decisions," *Review of Economic Dynamics*, 5(4), 720–748.

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? Gibrat's Law Revisited," *Journal of Political Economy*, 117(6), 1140–1154.

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.

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

Brannlund, R., and J. Nordstrom (2004): "Carbon Tax Simulations using a Household Demand Model," *European Economic Review*, 48, 211–233.

De Nadai, M., and A. Lewbel (2012): "Nonparametric Errors in Variables Models with Measurement Errors on both sides of the Equation," *Unpublished Manuscript.*

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

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.

Johnson, N. L., S. Kotz, and N. Balakrishnan (1994): *Continuous Univariate Distributions.* Wiley Series in Probability.

Kedir, A., and S. Girma (2007): "Quadratic Engel curves with measurement error: Evidence from a budget survey," *Oxford Bulletin of Economics and Statistics*, 69, 123–138.

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

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

Lewbel, A. (2008): "Engel Curves," entry for The New Palgrave Dictionary of Economics, 2nd Edition.

Lyssiotou, P., P. Pashardes, and T. Stengos (1999): "Testing the Rank of Engel curves with endogenous expenditure," *Economics Letters*, 64, 61–65.

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

Table 3: Estimates of the Engel curve parameters. SHIW 2010 Data.

		No Children			**One Child**			**More Than One Child**		
		b_{i1}	ρ_i	σ^2_V	b_{i1}	ρ_i	σ^2_V	b_{i1}	ρ_i	σ^2_V
(1)	***OLS***	- 0.129			- 0.126			- 0.146		
		(0.01)			*(0.008)*			*(0.006)*		
(2)	***IV***	- 0.284			- 0.262			- 0.259		
		(0.064)			*(0.039)*			*(0.03)*		
(3)	***Lewbel (1996)***	- 0.192			- 0.142			- 0.178		
		(0.031)			*(0.026)*			*(0.021)*		
(4)	***CF***	- 0.283	0.226		- 0.264	0.321		- 0.238	0.2	
		(0.053)	*(0.104)*		*(0.052)*	*(0.149)*		*(0.032)*	*(0.09)*	
(5)	***GMM***	- 0.287	0.243	0.013	- 0.249	0.274	0.043	- 0.231	0.158	0.106
		(0.054)	*(0.114)*	*(0.161)*	*(0.04)*	*(0.1)*	*(0.12)*	*(0.028)*	*(0.075)*	*(0.077)*
	Sample Size	465			870			1387		

Note. Presented are estimates of the shape parameter of the Engel curve in (22). The label “OLS” refers to OLS estimates; “IV” refers to IV estimates obtained without adjusting for measurement error, for which instruments are defined as explained in the text; “Lewbel (1996)” refers to the procedure suggested by Lewbel (1996); “CF” refers to the procedure discussed in Section 4. “GMM” refers to the procedure that makes use of the IV and CF estimating equations jointly. Standard errors reported in parentheses are robust to heterosckedasticity.

Additional Results

Appendix A

Proof of Theorem 1

From equation (1) and (7), which directly follows from (4), we have:

$$W_i = \frac{b_{i0} + b_{i1} \log X^* + \varepsilon_i + \nu_i}{V}.$$

Let $g^*(Z)$ be the conditional mean of $\log X^*$ given the instruments Z, then it is:

$$\log X^* = g^*(Z) + \eta^*.$$

Under Assumption 1 we can write:

$$W_i = \frac{b_{i0} + b_{i1} \log X^* + \rho_i \eta^* + \xi_i + \nu_i}{V},$$

with $E[\xi_i|X^*, \eta^*] = 0$. Now multiplying by X^l either side of the equation, using (5) and taking the conditional expectation with respect to Z yields:

$$\begin{aligned}
E[X^l W_i|Z] &= E\left[(X^* V)^l \frac{b_{i0} + b_{i1} \log X^* + \rho_i \eta^* + \xi_i + \nu_i}{V} |Z\right], \\
&= b_{i0} E[V^{l-1} X^{*l}|Z] + b_{i1} E[V^{l-1} X^{*l} \log X^*|Z] + \rho_i E[V^{l-1} X^{*l} \eta^*|Z] + \\
&\quad + E[V^{l-1} X^{*l} \xi_i] + E[V^{l-1} X^{*l} \nu_i|Z], \\
&= b_{i0} E[V^{l-1}] E[X^{*l}|Z] + b_{i1} E[V^{l-1}] E[X^{*l} \log X^*|Z] + \\
&\quad + \rho_i E[V^{l-1}] E[X^{*l} \eta^*|Z] + E[V^{l-1} \nu_i] E[X^{*l}|Z], \qquad (25)
\end{aligned}$$

where the last equality follows from Assumption 2 (iii) and $E[\xi_i|X^*, \eta^*] = 0$. Hence we may write:

$$\begin{aligned}
E[X^l|Z] &= E[X^{*l} V^l|Z], \\
&= E[V^l] E[X^{*l}|Z],
\end{aligned}$$

and:

$$\begin{aligned}
E[X^l \log X|Z] &= E[X^{*l} V^l (\log X^* + \log V)|Z], \\
&= E[V^l] E[X^{*l} \log X^*|Z] + E[V^l \log V] E[X^{*l}|Z].
\end{aligned}$$

Also by defining η as the residual of the linear projection of the observed X on the instruments Z it follows from equation (6) that $\eta = \eta^* + \log V - E[\log V]$, where the last expectation ensures that $E[\eta] = 0$. Thus:

$$\begin{aligned}
E[X^l \eta_h|Z] &= E[X^{*l} V^l (\eta^* + \log V - E[\log V])|Z], \\
&= E[V^l] E[X^{*l} \eta^*|Z] + E[V^l \log V] E[X^{*l}|Z] - E[V_h^l] E[\log V] E[X^{*l}|Z], \\
&= E[V^l] E[X^{*l} \eta^*|Z] + Cov(V^l, \log V) E[X^{*l}|Z].
\end{aligned}$$

The unobservable moments on the right hand side of (25) may then be written in terms of observable ones as:

$$\begin{aligned} E[X^{*l}|Z] &= \frac{E[X^l|Z]}{E[V^l]}, & (26)\\ E[X^{*l}\log X^*|Z] &= \frac{E[X^l\log X|Z]}{E[V^l]} - \frac{E[V^l\log V]E[X^l|Z]}{E[V^l]^2}, & (27)\\ E[X^{*l}\eta^*|Z] &= \frac{E[X^l\eta_h|Z]}{E[V^l]} - \frac{Cov(V^l,\log V)E[X^l|Z]}{E[V^l]^2}. & (28) \end{aligned}$$

Substituting equations (26), (27) and (28) into (25) and rearranging terms yields:

$$\begin{aligned} E[X^lW_i|Z] &= b_{i0}\frac{E[V^{l-1}]}{E[V^l]}E[X^l|Z] - b_{i1}\frac{E[V^{l-1}]E[V^l\log V]}{E[V^l]^2}E[X^l|Z] - \\ &\quad -\rho_i\frac{E[V^{l-1}]Cov(V^l,\log V)}{E[V^l]^2}E[X^l|Z] + b_{i1}\frac{E[V^{l-1}]}{E[V^l]}E[X^l\log X|Z] + \\ &\quad \rho_i\frac{E[V^{l-1}]}{E[V^l]}E[X^l\eta_h|Z] + \frac{E[V^{l-1}\nu_i]}{E[V^l]}E[X^l|Z], \\ &= \alpha_{il}E[X^l|Z] + \beta_{il}E[X^l\log X|Z] + \tilde{\rho}_{il}E[X^l\eta|Z], \end{aligned}$$

where:

$$\begin{aligned} \alpha_{il} &= b_{i0}\frac{E[V^{l-1}]}{E[V^l]} - b_{i1}\frac{E[V^{l-1}]E[V^l\log V]}{E[V^l]^2} - \rho_i\frac{E[V^{l-1}]Cov(V^l,\log V)}{E[V^l]^2} + \frac{E[V^{l-1}\nu_{ih}]}{E[V^l]}, \\ \beta_{il} &= b_{i1}\frac{E[V^{l-1}]}{E[V^l]}, \\ \tilde{\rho}_{il} &= \rho_i\frac{E[V^{1-1}]}{E[V^l]}. \end{aligned}$$

Q.E.D.

Proof of Theorem 2

Combining equations (20) and (7), as above, it is:

$$W_i = \frac{b_{i0} + b_{i1}\log X^* + b_{i2}(\log X^*)^2 + \varepsilon_i + \nu_i}{V},$$

and under Assumption 1 we might write:

$$W_i = \frac{b_{i0} + b_{i1}\log X^* + b_{i2}(\log X^*)^2 + \rho_i\eta^* + \xi_i + \nu_i}{V},$$

with $E[\xi_i|X^*,\eta^*] = 0$. Now multiplying by X^l either side of the equation, using (5) and taking the conditional expectation with respect to Z yields:

$$\begin{aligned} E[X^lW_i|Z] &= b_{i0}E[V^{l-1}]E[X^{*l}|Z] + b_{i1}E[V^{l-1}]E[X^{*l}\log X^*|Z] + \\ &\quad +b_{i2}E[V^{l-1}]E[X^{*l}(\log X^*)^2|Z] + \rho_iE[V^{l-1}]E[X^{*l}\eta^*|Z] + \\ &\quad +E[V^{l-1}\nu_i]E[X^{*l}|Z]. \qquad (29) \end{aligned}$$

Now note that:

$$\begin{aligned}
E[X^l(\log X)^2|Z] &= E[X^{*l}V^l(\log X^*)^2|Z] + E[X^{*l}V^l(\log V)^2|Z] + \\
&\quad +2E[X^{*l}V^l \log X^* \log V|Z], \\
&= E[V^l]E[X^{*l}(\log X^*)^2|Z] + E[V^l(\log V)^2]E[X^{*l}|Z] + \\
&\quad +2E[V^l \log V]E[X^{*l}\log X^*|Z],
\end{aligned}$$

so that by substituting back equations (26) and (27) we obtain:

$$\begin{aligned}
E[X^{*l}(\log X^*)^2|Z] &= \frac{E[X^l(\log X)^2|Z]}{E[V^l]} - 2\frac{E[V^l \log V]}{E[V^l]^2}E[X^l \log X|Z] \\
&\quad -\left\{\frac{E[V^l(\log V)^2]}{E[V^l]^2} - 2\frac{E[V^l \log V]^2}{E[V^l]^3}\right\}E[X^l|Z]. \qquad (30)
\end{aligned}$$

Now combining equations (26), (27) and (30) into (29) and rearranging terms it is:

$$\begin{aligned}
E[X^lW_i|Z] &= b_{i0}\frac{E[V^{l-1}]}{E[V^l]}E[X^l|Z] - b_{i1}\frac{E[V^{l-1}]E[V^l \log V]}{E[V^l]^2}E[X^l|Z] \\
&\quad -b_{i2}E[V^{l-1}]\left\{\frac{E[V^l(\log V)^2]}{E[V^l]^2} - 2\frac{E[V^l \log V]^2}{E[V^l]^3}\right\}E[X^l|Z] \\
&\quad -\rho_i\frac{E[V^{l-1}]Cov(V^l, \log V)}{E[V^l]^2}E[X^l|Z] \\
&\quad +b_{i1}\frac{E[V^{l-1}]}{E[V^l]}E[X^l \log X|Z] - 2b_{i2}\frac{E[V^{l-1}]E[V^l \log V]}{E[V^l]^2}E[X^l \log X|Z] \\
&\quad +b_{i2}\frac{E[V^{l-1}]}{E[V^l]}E[X^l(\log X)^2|Z] + \rho_i\frac{E[V^{l-1}]}{E[V^l]}E[X^l\eta_h|Z] + \\
&\quad +\frac{E[V^{l-1}\nu_i]}{E[V^l]}E[X^l|Z], \\
&= \alpha_{il}E[X^l|Z] + \beta_{il}E[X^l \log X|Z] + \gamma_{il}E[X^l(\log X)^2|Z] + \tilde{\rho}_{il}E[X^l\eta|Z],
\end{aligned}$$

where:

$$\begin{aligned}
\alpha_{il} &= b_{i0}\frac{E[V^{l-1}]}{E[V^l]} - b_{i1}\frac{E[V^{l-1}]E[V^l \log V]}{E[V^l]^2} \\
&\quad -b_{i2}E[V^{l-1}]\left\{\frac{E[V^l(\log V)^2]}{E[V^l]^2} - 2\frac{E[V^l \log V]^2}{E[V^l]^3}\right\} \\
&\quad -\rho_i\frac{E[V^{l-1}]Cov(V^l, \log V)}{E[V^l]^2} + \frac{E[V^{l-1}\nu_{ih}]}{E[V^l]}, \\
\beta_{il} &= b_{i1}\frac{E[V^{l-1}]}{E[V^l]} - 2b_{i2}\frac{E[V^{l-1}]E[V^l \log V]}{E[V^l]^2}, \\
\gamma_{il} &= b_{i2}\frac{E[V^{l-1}]}{E[V^l]}, \\
\tilde{\rho}_{il} &= \rho_i\frac{E[V^{1-1}]}{E[V^l]}.
\end{aligned}$$

Q.E.D.

Appendix B

Table 4: Descriptive Statistics. SHIW 2010 Data.

	No Children		**One Child**		**More Than One Child**	
	Mean	*S.D.*	*Mean*	*S.D.*	*Mean*	*S.D.*
Food Budget Shares	0.27	0.11	0.28	0.11	0.31	0.12
Total Expenditure (Logs)	10.04	0.43	10.14	0.44	10.15	0.47
Male Wages (Logs)	10.05	0.23	10.02	0.24	9.94	0.26
Dummy - North	0.54	0.5	0.47	0.5	0.35	0.48
Dummy - Center	0.2	0.4	0.25	0.43	0.17	0.38
Dummy - South	0.26	0.44	0.29	0.45	0.48	0.5
Number of Households	465		870		1388	

Note. Summary of descriptive statistics for the variables used in application. These statistics refer to the subsample of couples in the 2010 wave of the SHIW data.

Table 5: Percentage Bias for $\hat{\beta}_1$ defined as $(\hat{\beta}_1 - b_1)/|b_1|$. Sample size: 500.

N/S		**Extent of Endogeneity**							
Ratio		**0**		**0.2**		**0.5**		**0.8**	
0	**OLS**	0.005	(*1.04*)	1.915	(*1.372*)	6.925	(*2.031*)	15.074	(*2.865*)
	IV	0.004	(*1.122*)	0.066	(*1.465*)	0.132	(*2.035*)	0.143	(*2.392*)
	Lewbel (1996)	0.043	(*3.336*)	3.984	(*4.985*)	14.312	(*10.611*)	31.084	(*20.356*)
	CF	0.086	(*5.534*)	2.695	(*7.989*)	9.509	(*15.962*)	20.535	(*29.766*)
0.1	**OLS**	69.678	(*11.549*)	71.548	(*11.602*)	76.454	(*11.744*)	84.43	(*11.981*)
	IV	–5.972	(*11.708*)	–5.941	(*11.755*)	–5.83	(*11.86*)	–5.646	(*11.938*)
	Lewbel (1996)	10.945	(*49.085*)	14.716	(*48.807*)	24.619	(*48.611*)	40.742	(*49.924*)
	CF	–3.477	(*69.432*)	-0.629	(*68.63*)	6.882	(*67.225*)	19.162	(*67.225*)
0.3	**OLS**	157.365	(*17.457*)	159.193	(*17.491*)	163.991	(*17.587*)	171.797	(*17.752*)
	IV	–17.099	(*19.033*)	–17.069	(*19.034*)	–16.95	(*19.07*)	–16.668	(*19.108*)
	Lewbel (1996)	28.164	(*85.518*)	31.694	(*84.765*)	40.968	(*83.093*)	56.071	(*81.395*)
	CF	–13.763	(*98.499*)	–10.589	(*97.475*)	–2.258	(*95.177*)	11.299	(*92.719*)
0.5	**OLS**	204.802	(*20.364*)	206.618	(*20.392*)	211.387	(*20.462*)	219.149	(*20.569*)
	IV	–26.114	(*23.103*)	–26.091	(*23.141*)	–25.985	(*23.214*)	–25.727	(*23.26*)
	Lewbel (1996)	46.787	(*106.938*)	50.029	(*106.027*)	58.52	(*103.85*)	72.301	(*101.043*)
	CF	–15.579	(*112.782*)	–12.262	(*111.73*)	–3.58	(*109.251*)	10.498	(*106.136*)

Note. See footnote to Table 1.

Table 6: Percentage Bias for $\hat{\beta}_1$ defined as $(\hat{\beta}_1 - b_1)/|b_1|$. Sample size: 10,000.

N/S Ratio		**Extent of Endogeneity**							
		0		**0.2**		**0.5**		**0.8**	
0	**OLS**	-0.002	*(0.229)*	1.903	*(0.303)*	6.904	*(0.45)*	15.042	*(0.638)*
	IV	-0.004	*(0.24)*	-0.001	*(0.315)*	0.006	*(0.449)*	0.019	*(0.587)*
	Lewbel (1996)	-0.026	*(1.277)*	2.398	*(2.616)*	8.772	*(7.927)*	19.157	*(16.932)*
	CF	-0.033	*(1.141)*	0.393	*(2.22)*	1.522	*(6.576)*	3.382	*(14.01)*
0.1	**OLS**	69.343	*(2.546)*	71.204	*(2.558)*	76.091	*(2.59)*	84.039	*(2.643)*
	IV	–6.712	*(2.628)*	–6.708	*(2.639)*	–6.698	*(2.662)*	–6.681	*(2.694)*
	Lewbel (1996)	3.161	*(18.637)*	5.479	*(18.75)*	11.568	*(19.984)*	21.481	*(24.342)*
	CF	1.121	*(16.191)*	1.518	*(16.028)*	2.561	*(16.433)*	4.263	*(19.374)*
0.3	**OLS**	156.759	*(3.959)*	158.578	*(3.969)*	163.353	*(3.994)*	171.122	*(4.032)*
	IV	–18.304	*(4.295)*	–18.3	*(4.301)*	–18.289	*(4.315)*	–18.268	*(4.335)*
	Lewbel (1996)	10.014	*(36.147)*	12.301	*(36.053)*	18.307	*(36.437)*	28.081	*(38.929)*
	CF	4.042	*(30.435)*	4.464	*(30.016)*	5.569	*(29.462)*	7.363	*(30.313)*
0.5	**OLS**	204.047	*(4.534)*	205.856	*(4.54)*	210.603	*(4.556)*	218.324	*(4.58)*
	IV	–27.827	*(5.122)*	–27.822	*(5.129)*	–27.809	*(5.145)*	–27.787	*(5.168)*
	Lewbel (1996)	17.008	*(47.907)*	19.177	*(47.708)*	24.866	*(47.64)*	34.115	*(48.943)*
	CF	5.65	*(39.829)*	6.049	*(39.131)*	7.096	*(37.741)*	8.8	*(36.978)*

Note. See footnote to Table 1.

Acknowledgements

This paper benefited from helpful discussions with Ian Crawford, Thomas Crossley, Arthur Lewbel, Franco Peracchi and comments by audiences at the EC^2 Conference - Identification in Econometrics (Toulouse, December 2010), the Italian Conference of Econometrics and Empirical Economics (Pisa, January 2011) and the Econometric Society European Meeting (Oslo, August 2011). Financial support from the University of Padova is gratefully acknowledged.