Journal of Econometrics 191 (2016) 19–32

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Nonparametric errors in variables models with measurement errors on both sides of the equation☆

Michele De Nadai [a], Arthur Lewbel [b,*]

[a] University of New South Wales, Australia

[b] Boston College, United States

ARTICLE INFO

Article history:
Received 18 July 2013
Received in revised form
23 September 2014
Accepted 1 August 2015
Available online 23 October 2015

JEL classification:
C10
C14
D12

Keywords:
Engel curve
Errors-in-variables model
Fourier transform
Generalized function
Sieve estimation

ABSTRACT

Measurement errors are often correlated, as in surveys where respondent's biases or tendencies to err affect multiple reported variables. We extend Schennach (2007) to identify moments of the conditional distribution of a true Y given a true X when both are measured with error, the measurement errors in Y and X are correlated, and the true unknown model of Y given X has nonseparable model errors. After showing nonparametric identification, we provide a sieve generalized method of moments based estimator of the model, and apply it to nonparametric Engel curve estimation. In our application measurement errors on the expenditures of a good Y are by construction correlated with measurement errors in total expenditures X. This problem, which is present in many consumption data sets, has been ignored in most demand applications. We find that accounting for this problem casts doubt on Hildenbrand's (1994) "increasing dispersion" assumption.

1. Introduction

We consider identification and estimation of conditional moments of a dependent variable Y given a regressor X in nonparametric regression models where both Y and X are mismeasured, and the measurement errors in Y and X are correlated. For example, correlated measurement errors are likely in survey data where each respondent's reporting biases or tendencies to err affect multiple variables that he or she self reports.

An example application that we consider empirically is consumer demand estimation, where Y is the quantity or expenditures demanded of some good or service, and X is total consumption expenditures on all goods. 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, and hence be correlated with, the measurement error in X. Similar problems 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.

Our identification procedure allows us to distinguish measurement errors from other sources of error that are due to unobserved structural or behavioral heterogeneity. 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 depend on the distribution of income elasticities in the population. In contrast to our results, most empirical analyses implicitly or explicitly attribute either none or all of estimated model errors to unobserved heterogeneity.

In the consumer demand application, it has long been known that for most goods, empirical estimates of $Var(Y|X)$ are increasing in X. For example, Hildenbrand (1994, Figs. 3.6 and 3.7) documents this phenomenon for a variety of goods in two different countries, calls it the "increasing dispersion" assumption, and exploits it as a behavioral feature that helps give rise to the aggregate law of demand. This property is also often used to justify estimating Engel

☆ This paper benefited from helpful discussions with Erich Battistin and seminar participants at University of Mannheim and University of Padua.

* Correspondence to: School of Economics, University of New South Wales, Australian School of Business, Sydney 2052, Australia and Department of Economics, 140 Commonwealth Avenue, 02467 Chestnut Hill, MA, United States.

E-mail addresses: m.denadai@unsw.edu.au (M. De Nadai), lewbel@bc.edu (A. Lewbel).

http://dx.doi.org/10.1016/j.jeconom.2015.08.005
0304-4076/

curves in budget share instead of quantity form, to reduce the resulting error heteroskedasticity. However, in this paper we find empirically that while this phenomenon clearly holds in estimates of $Var(Y|X)$ on raw data, after nonparametrically accounting for joint measurement error in Y and X, the evidence for increasing dispersion becomes considerably weaker, suggesting that this well documented feature of Engel curve estimates may be in part an artifact of measurement errors rather than a feature of behavior.

Our identification strategy is an extension of Schennach (2007), who provides nonparametric identification of the conditional mean of Y given X (using instruments Q) when X is a classically mismeasured regressor. We extend Schennach (2007) primarily by allowing for a measurement error term in Y that may be correlated with the measurement error in X. An additional extension is that we identify higher moments of the true Y given the true X instead of just the conditional mean. A further extension allows the measurement error in X to take a multiplicative form that is particularly well suited for our Engel curve application. Our proofs make use of recent machinery provided by Zinde-Walsh (2014).

Building on estimators like Newey (2001), Schennach (2007) bases identification on taking Fourier transforms of the conditional means of Y and of XY given instruments. Our main insight is that, if additive measurement errors in X and Y are correlated with each other but otherwise have some of the properties of classical measurement errors, then their presence will only affect the Fourier transform on a finite number of points, so identification will still be possible. Our further extensions exploit similar properties in different measurement error specifications, and our empirical application makes use of some special features of Engel curves to fully identify higher moments.

There is a large literature on the estimation of measurement error models. In addition to Schennach (2007), more recent work on measurement errors in nonparametric regression models includes Delaigle et al. (2009), Rummel et al. (2010), Carroll et al. (2010), Meister (2011), and Carroll et al. (2011). Recent surveys containing many earlier references include Carroll et al. (2006) and Chen et al. (2011).[1]

In the literature we find several examples of Engel curve estimation in the presence of measurement errors. Hausman et al. (1991, 1995) provide estimators for polynomial Engel curves with classically mismeasured X, Newey (2001) estimates a nonpolynomial parametric Engel curve with mismeasured X, Blundell et al. (2007) estimate a semi-parametric model of Engel curves that allows X to be endogenous and hence mismeasured, and Lewbel (1996) identifies and estimates Engel curves allowing for correlated measurement errors in X and Y as we do, but does so in the context of a parametric model of Y given X.[2]

The conditional distribution of the true Y given the true X in Engel curves corresponds to the distribution of preference heterogeneity parameters in the population, which can be of particular interest for policy analysis. For example, consider the effect on demand of introducing a tax cut or tax increase that shifts households' total expenditure levels. This will in general affect the entire distribution of demand, not just its mean, both because Engel curves are generally nonlinear and because preferences are heterogeneous. Recovering moments of the distribution of demand is useful because many policy indicators, such as the welfare implication of a tax change, will in turn depend on more features of the distribution of demand than just its mean.

The next two sections show identification of the model with standard additive measurement error and of the specification more specifically appropriate for Engel curve data. We then describe our sieve based estimator, and provide a simulation study. After that is an empirical application to estimating food and clothing expenditures in US Consumer Expenditure Survey data, followed by conclusions and an appendix providing proofs.

2. Overview

Suppose that scalar random variables Y^* and X^* are measured with error, so we only observe Y and X where:

$$Y = Y^* + S,$$
$$X = X^* + W,$$

with S and W being unobserved measurement errors that we assume, for now, to have the classical property of being mean zero with $S, W \perp Y^*, X^*$. This assumption is just made here and now to ease exposition; our formal results will substantially relax these independence assumptions, replacing them with Assumption 1. We will later further generalize the model to include different specifications for the measurement errors. We explicitly allow S and W to be correlated with each other. This might be due to the nature of the variables involved, or caused by the way in which Y and X are collected, as is the case for consumption data as described in the Introduction, or when related reporting biases affect the collection of both Y and X.

The model considered might also arise because of nondifferential properties of the measurement error in X. In the statistics literature, a measurement error W in X is called "differential" if it affects the observed outcome Y, after conditioning on the true X^*, that is, if $Y \mid X^*, W$ does not equal $Y \mid X^*$ (see, e.g., Carroll et al. (2006)). An alternative application of our identification results would be for a model in which Y is not mismeasured, and the additive error S instead represents the effect of differential measurement error W on the true observed outcome Y. In this setup $Y \mid X^*$ is in general different from $Y \mid X^*, W$, with the two distributions being equal only if S and W are independent, so that the amount of correlation between S and W could be thought of as the extent of the departure from the nondifferential assumption on W. Since the nature of S does not affect our identification result, to ease exposition, in the following we will refer to S only as measurement error in Y^*.

Without loss of generality we specify Y^* as

$$Y^* = H(X^*, U),$$

where $H(\cdot,\cdot)$ is an unknown function of a scalar random regressor X^*, and a random scalar or vector of nonseparable unobservables U, which can be interpreted as regression model errors or unobserved heterogeneity in the population. The extension to inclusion of other (observed) covariates will be straightforward, so we drop them for now.

In this setup our primary goal is identification (and later estimation) of the nonparametric regression function $E[H(X^*, U) \mid X^*]$, but we more generally consider identification of conditional moments $E\left[H(X^*, U)^k \mid X^*\right]$ for integers k. Thus our results can be interpreted as separating the impact of unobserved heterogeneity U from the effects of measurement errors on the relationship of Y to X. We do not deal directly with estimation of H and of U, but these

[1] Earlier econometric papers closely related to Schennach (2007), but exploiting repeated measurements, are Hausman et al. (1991), Schennach (2004) and Li (2002). Most of these assume two mismeasures of the true X are available, one of which could have errors correlated with the measurement error Y.

[2] More generally, within econometrics there is a large recent literature on nonparametric identification of models having nonseparable errors (e.g., Chesher, 2003; Meister, 2007; Hoderlein and Mammen, 2007, and Imbens and Newey, 2009), multiple errors (e.g. random coefficient models like Beran et al., 1996 and generalizations like Hoderlein et al., 2011 and Lewbel, 2011) or both (e.g., Matzkin, 2003). This paper contributes to that literature by identifying models that have both additive measurement error and structural nonseparable unobserved heterogeneity.