Journal of Econometrics 185 (2015) 392–408

Contents lists available at ScienceDirect

Journal of Econometrics

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

Closed-form estimation of nonparametric models with non-classical measurement errors

Yingyao Hu *, Yuya Sasaki

Department of Economics, Johns Hopkins University, 440 Mergenthaler Hall, 3400 N. Charles St, Baltimore, MD 21218, United States

ARTICLE INFO

Article history:
Received 7 October 2013
Received in revised form
31 October 2014
Accepted 30 November 2014
Available online 16 December 2014

JEL classification:
C14
C21

Keywords:
Closed form
Non-classical measurement errors
Nonparametric regressions

ABSTRACT

This paper proposes closed-form estimators for nonparametric regressions using two measurements with non-classical errors. One (administrative) measurement has location-/scale-normalized errors, but the other (survey) measurement has endogenous errors with arbitrary location and scale. For this setting of data combination, we derive closed-form identification of nonparametric regressions, and practical closed-form estimators that perform well with small samples. Applying this method to NHANES III, we study how obesity explains health care usage. Clinical measurements and self reports of BMI are used as two measurements with normalized errors and endogenous errors, respectively. We robustly find that health care usage increases with obesity.

1. Introduction

For the increasing availability of combined administrative and survey data (Moffitt and Ridder, 2007), econometric methods that can properly handle matched data with measurement errors have become of great practical importance. For econometric methods to be truly useful no matter how complicated a model is, estimators should ideally be given in a closed form explicitly written in terms of observed data, like the OLS. Unfortunately, such convenient characteristics are rarely shared by nonparametric estimators for non-classical measurement errors.

Identification and estimation of regression models with two measurements of explanatory variables are proposed by Li (2002) and Schennach (2004a,b) among others. A limitation with the existing methods is that they require two measurements with classical errors. In practice, empirical data with two measurements often come from matched administrative, imputed, and/or survey data, where particularly survey data are often subject to non-classical errors (e.g. Bound et al., 2001; Koijen et al., 2013). Ignoring the non-classical nature of errors in measurements may lead to inconsistent estimation, as we demonstrate in our simulations. In this paper, we propose closed-form estimators for nonparametric regression models using two measurements with non-classical errors.

Specifically, we explicitly estimate the nonparametric regression function g for the model

$$Y = g(X^*) + U \qquad E\left[U|X^*\right] = 0,$$

where Y is an observed dependent variable, X^* is an unobserved explanatory variable, and U is the regression residual. While the true explanatory variable X^* is not observed, two measurements, X_1 and X_2, are available from matched data. For simplicity, X^* is assumed to be a scalar and continuously distributed. The relationship between the two measurements and the true explanatory variable X^* is modeled as follows.

$$X_1 = \sum_{p=0}^{P} \gamma_p X^{*p} + \mathcal{E}_1$$

$$X_2 = X^* + \mathcal{E}_2.$$

Unless $\gamma_1 = 1$ and $\gamma_2 = \cdots = \gamma_P = 0$ are true, the first measurement X_1 entails non-classical errors with nonlinearity. Allowing for such non-classical errors is crucial particularly for survey data that are often contaminated by endogenous self-reporting biases. Since the truth X^* is unobserved, the second measurement X_2 is location-/scale-normalized with respect to the unobserved truth X^*. We use alternative independence assumptions on the measurement error $\mathcal{E}_2$ depending on which order P we assume about X_1, but these assumptions are more innocuous than assuming classical errors in any case.

* Corresponding author.
E-mail addresses: yhu@jhu.edu (Y. Hu), sasaki@jhu.edu (Y. Sasaki).

http://dx.doi.org/10.1016/j.jeconom.2014.11.004
0304-4076/

Under assumptions that will be introduced below, we show that the regression function g can be explicitly expressed as a functional of the joint CDF $F_{YX_1X_2}$ in the sense that $g(x^*) = \lambda(x^*|F_{YX_1X_2})$. We provide the concrete expression for this functional $\lambda(x^* \mid \cdot)$. In order to construct a sample-counterpart estimator of $g(x^*)$ given this closed-form identifying solution, it suffices to substitute the empirical distribution $\widehat{F}_{YX_1X_2}$ in this known transformation so we get the closed-form estimator $\widehat{g(x^*)} = \lambda(x^* \mid \widehat{F}_{YX_1X_2})$. We present its theoretical large sample properties as well as its small sample performance. Monte Carlo simulations show that the estimator works quite well with $N = 500$, a very small sample size for nonparametrics.

Measurement error models have been extensively studied in both statistics and econometrics. The statistical literature focuses on cases of classical errors, where measurement errors are independent of the true values — see Fuller (1987) and Carroll et al. (2006) for reviews. The econometric literature investigates nonlinear models and nonclassical measurement errors — see Chen et al. (2011), Bound et al. (2001) and Schennach (2013) for reviews. However, closed-form estimation, nonlinear/nonparametric models, and non-classical measurement errors still remain unsolved, despite their joint practical relevance. Two measurements are known to be useful to correct measurement errors even for external samples if the matched administrative data is known to be true (e.g., Chen et al., 2005). The baseline model of our framework was introduced by Li (2002) and Schennach (2004a), where they consider parametric regression models under two measurements with classical errors. Hu and Schennach (2008) provide general identification results for nonseparable and non-classical measurement errors,[1] but their estimator relies on semi-/non-parametric extremal estimator where nuisance functions are approximated by truncated series.[2] Unlike these existing approaches, we develop a closed-form estimator for nonparametric models involving non-classical measurement errors.

Our results share much in common with Schennach (2004b) where she develops a closed-form estimator under the restriction, $\gamma_1 = 1$ and $\gamma_2 = \cdots = \gamma_P = 0$, of a classical-error structure. There are notable differences and thus values added by this paper as well. Our method paves the way for non-classical error structures with high degrees of nonlinearity whereas the existing closed-form estimator can handle only classical errors. To this end, we propose a new method to recover and use the characteristic function of the generated latent variable $\sum_{p=1}^{P} \gamma_p X^{*p}$, instead of just X^*, in the framework of deconvolution approaches. Not surprisingly, as we show through simulations, the classical error assumption $\gamma_1 = 1$ and $\gamma_2 = \cdots = \gamma_P = 0$ can severely bias estimates if the true DGP does not conform with this assumption. In our empirical application, we find that $\gamma_1 \neq 1$ is indeed true when people report their physical characteristics, and hence the existing closed-form estimator that assumes classical errors would likely suffer from biased estimates. The contribution of our method is to overcome these practical limitations of the existing closed-form estimators.

For an empirical illustration, we investigate how obesity measured by the Body Mass Index (BMI) explains the health care usage by using a sample of about 1900 observations extracted from the National Health and Nutrition Examination Survey (NHANES III). This data set uniquely matches self-reports and clinical measurements of the BMI. We allow the former measurement to suffer from endogenous biases with arbitrary location and scale, while the latter measurement is location-/scale-normalized with respect to the true BMI. Our results show a robust upward-sloping tendency of the mean health care usage as a function of the true BMI, controlling for the most important health factors, namely gender and age. This tendency is particularly stronger for females.

2. Closed-form identification: a baseline model

Our objective is to derive closed-form identifying formulas for the nonparametric regression function g. For the purpose of intuitive exposition, we first focus on the following simple model:

$$\begin{aligned} Y &= g(X^*) + U, \qquad \mathrm{E}[U \mid X^*] = 0 \\ X_1 &= \gamma_1 X^* + \varepsilon_1 \qquad \mathrm{E}[\varepsilon_1] = \gamma_0 \\ X_2 &= X^* + \varepsilon_2, \qquad \mathrm{E}[\varepsilon_2] = 0 \end{aligned} \tag{2.1}$$

where we observe the joint distribution of (Y, X_1, X_2). The restriction $\mathrm{E}[U \mid X^*] = 0$ means that $g(X^*)$ is the nonparametric regression of Y on X^*. We do not assume $\mathrm{E}[\varepsilon_1]$ to be zero in order to accommodate arbitrary intercept γ_0 for the first measurement X_1. As such, we suppress γ_0 from the equation for X_1, i.e., it is embedded in $\gamma_0 = \mathrm{E}[\varepsilon_1]$. On the other hand, the locational normalization $\mathrm{E}[\varepsilon_2] = 0$ is imposed on the second measurement X_2. A leading example of (2.1) is the case with $\gamma_1 = 1$ often assumed in related papers in the literature. We do not make such an assumption, and thus our model (2.1) accommodates the possibility that the first measurement X_1 is endogenously biased even if $X^* \perp\!\!\!\perp \varepsilon_1$ is assumed, as $\mathrm{E}[X_1 - X^* \mid X^*] = \gamma_0 + (\gamma_1 - 1)X^*$.

We can easily show that γ_1 is identified from the observed data by the closed-form formula

$$\gamma_1 = \frac{\mathrm{Cov}(Y, X_1)}{\mathrm{Cov}(Y, X_2)} \tag{2.2}$$

under the following assumption.

Assumption 1 (*Identification of* γ_1)**.** $\mathrm{Cov}(\varepsilon_1, Y) = \mathrm{Cov}(\varepsilon_2, Y) = 0$ and $\mathrm{Cov}(Y, X_2) \neq 0$.

The first part of this assumption requires that ε_1 and ε_2 are uncorrelated with the dependent variable. These zero covariance restrictions can be implied by a lower-level assumption, such as $\mathrm{E}[U \mid X^*, \varepsilon_1, \varepsilon_2] = 0$, $\varepsilon_1 \perp\!\!\!\perp X^*$, and $\mathrm{E}[\varepsilon_2 \mid X^*] = 0$, which also imply the additional identifying restrictions presented later (Assumption 3). The second part of Assumption 1 is empirically testable with observed data, and implies a non-zero denominator in the identifying Eq. (2.2). We state this auxiliary result below for ease of reference.

Lemma 1 (*Identification of* γ_1)**.** *If Assumption 1 holds, then* γ_1 *is identified with* (2.2)*.*

In some applications, we may simply assume $\gamma_1 = 1$ from the outset, and Assumption 1 need not be invoked. In any case, we hereafter assume that γ_1 is known either by assumption or by the identifying formula (2.2), and that γ_1 is different from zero.

Assumption 2 (*Nonzero* γ_1)**.** $\gamma_1 \neq 0$.

If this assumption fails, then the observed variable X_1 fails to be an informative signal of X^*. Assumption 2 therefore plays the role of letting X_1 be an effective proxy for the latent variable X^*. To complete our definition of the model (2.1), we impose the following independence restrictions.

Assumption 3 (*Restrictions*)**.** (i) $E[U|X_1] = 0$. (ii) $\varepsilon_1 \perp\!\!\!\perp X^*$. (iii) $E[\varepsilon_2|X_1] = 0$.

[1] Also see Mahajan (2006), Lewbel (2007), and Hu (2008) for non-/semi-parametric identification and estimation under non-classical measurement errors with discrete variables.

[2] Our model is also closely related to nonparametric regression models with classical measurement errors, which are extensively studied in the rich literature in statistics. When the error distribution is known, the regression function may be estimated by deconvolution — see Fan and Truong (1993) and Carroll et al. (2006) for reviews. When the error distribution is unknown, Schennach (2004b) uses Kotlarski's identify (see Rao, 1992) to provide a Nadaraya–Watson-type estimator for the regression function.