Statistics and Probability Letters 125 (2017) 202–206

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

Closed-form estimation of a regression model with a mismeasured binary regressor and heteroskedasticity

Yibin Liu [a], Wenbin Wu [b,c,*]

[a] Department of Economics, UCSD, United States
[b] School of Economics, Fudan University, China
[c] Fudan-Oceanwide International School of Finance, Fudan University, China

ARTICLE INFO

Article history:
Received 25 April 2015
Received in revised form 14 February 2017
Accepted 14 February 2017
Available online 24 February 2017

Keywords:
Measurement error
Binary regressor
Heteroskedasticity

ABSTRACT

This paper finds that heteroskedasticity in nonclassical error-in-variable models leads to biased and inconsistent estimates when higher-order moments of data are used. A closed-form estimator is provided to correct this bias based on information from the first three moments.

1. Introduction

Regressions with binary regressors are widely used and studied in economics, such as those concerning employment status, union status, gender, narrative estimation, etc. The major concerns when using binary regressors are misclassification (i.e., measurement error) and heteroskedasticity. For example, self-reported levels of education (e.g., whether or not one has graduated from college) may be inaccurate if respondents misreport. It is also very likely that people who graduate from college have a different probability of cheating than those who do not graduate from college, implying that we need to take into account nonclassical errors. We might also think about monetary policy shock measurement. Rosa (2012) uses a narrative method to measure unconventional monetary policy shocks on Federal Open Market Committee dates by assigning them a value of 0 or 1. This might introduce errors since the classification is somewhat subjective.

A lot of studies focus on the measurement-error problem, see Fuller (1987) and Chen et al. (2011) for a review. Chen et al. (2008a,b) (hereafter CHL) provide a nonparametric way to deal with this problem. However, they do not consider heteroskedasticity in the identification which might severely bias their estimates as shown in this paper. This is different from the classical framework where heteroskedasticity only affects efficiency but not unbiasedness or consistency. The reason is that instead of using instrumental variables (IV) or other additional information, CHL explore moments of data. Hence if the structure of the second moment is misspecified (i.e., heteroskedasticity), their estimates will be invalid. Based on a few simple assumptions, this paper provides an estimator to fix this problem without using instrumental variables or additional sample information.

Consider a nonparametric regression model:

$$Y = m\left(X^*\right) + \varepsilon, \tag{1}$$

* Correspondence to: School of Economics & FOISF, Fudan University, China.
E-mail addresses: yil490@ucsd.edu (Y. Liu), econ.wwb@gmail.com (W. Wu).

http://dx.doi.org/10.1016/j.spl.2017.02.016
0167-7152/

where Y is a scalar dependent variable, X^* is a 0, 1 dichotomous regressor, and ε is the error term. Let X denote the proxy variable for the true variable X^*. Both Y and X are observable, while X^* and ε are not observed. The exogeneity condition is satisfied, since we assume the expected value of ε given the latent X^* is equal to 0. Note that discreteness of X and X^* means that the measurement error $X - X^*$ will be nonclassical, as described in CHL.

Assumption 1. The error term can be decomposed as

$$\varepsilon = \sigma\left(X^*\right)\eta, \qquad E\left[\eta|X^*\right] = 0, \tag{2}$$

where $\sigma\left(X^*\right)$ is a nonparametric function of the true variable X^*.

This assumption holds if ε is a homogeneous function of X^* and η. Usually, we need an IV or additional data to tackle this problem. However, we will see that this model can be identified if we are willing to explore higher-order moments and impose a few restrictions.

Define $m_j = m\left(j\right)$ and $\sigma_j = \sigma\left(j\right)$ for $j = 0, 1$. Since X^* is binary, we only need to identify m_0 and m_1. Then, the conditional distributions of Y and η conditional on X^* and the probability mass function of X given X^* are easily identified. Note that the results can be easily extended to the case of $Y = m\left(X^*, W\right) + \varepsilon$, where W is a vector of additional regressors that are exogenous and correctly measured.

The identification strategy proposed here relies on some assumptions regarding the regression model instead of on additional sample information. The key assumptions are that the first three moments of the separated error term η are independent of the latent regressor, and that the distribution of η is not skewed. These simple assumptions directly enable us to nonparametrically identify the latent regression function as a known function of observed moments.

This paper is organized as follows: Section 2 provides the main identification results and Section 3 concludes the paper.

2. Nonparametric identification

2.1. Identification

Assumption 2. $X \perp \eta \mid X^*$

By Assumption 2, we know that the measurement error $X - X^*$ is independent of the dependent variable Y conditional on the true value X^*. In other words, we have $f_{Y|X^*,X}\left(y|x^*, x\right) = f_{Y|X^*}\left(y|x^*\right)$. Assumption 2 is a standard assumption in the literature, and a similar version of this assumption is also used by CHL. Moreover, it follows from Assumption 2 that there is a relationship between the observed density and the latent density:

$$\begin{aligned} f_{Y|X}\left(y|j\right) &= P_{X^*=0|X=j} f_{Y|X^*}\left(y|0\right) + P_{X^*=1|X=j} f_{Y|X^*}\left(y|1\right) \\ &= P_{X^*=0|X=j} f_{\varepsilon|X^*}\left(y - m_0|0\right) + P_{X^*=1|X=j} f_{\varepsilon|X^*}\left(y - m_1|1\right). \end{aligned} \tag{3}$$

Equality (3) is useful when we want to identify the unobservable density $P_{X^*=i|X=j}$ and $f_{\varepsilon|X^*}\left(y - m_i|j\right)$ for $i, j = 0$ or 1.

Now, define

$$p = P_{X^*=1|X=0}, \qquad q = P_{X^*=0|X=1}, \qquad \mu_0 = E\left[Y|X=0\right], \qquad \mu_1 = E\left[Y|X=1\right],$$

where p and q represent the probability of misreporting. Since $E\left[\eta|X^*\right] = 0$ by Assumption 1, we know from (3) that

$$\mu_0 = \left(1-p\right)m_0 + pm_1, \qquad \mu_1 = qm_0 + \left(1-q\right)m_1. \tag{4}$$

Assumption 3. (i) $\mu_1 > \mu_0$; (ii) $p + q \in [0, 1)$.

Part (i) is not restrictive because we can always redefine X as $1 - X$ if needed. Part (ii) means that the data always contain some information that makes the projection from the data to the true value better than a pure guess. Assumption 3 implies that X^* and X affect the conditional mean of Y in the same direction, which usually holds when the regression is not very severely contaminated by measurement errors.

Solve (4) for p and q provides

$$p = \frac{\mu_0 - m_0}{m_1 - m_0}, \qquad q = \frac{m_1 - \mu_1}{m_1 - m_0}. \tag{5}$$

By Assumption 3, (5) implies that

$$m_1 \geq \mu_1 \geq \mu_0 \geq m_0. \tag{6}$$

Assumption 4. (i) $E\left[\eta^2|X^*\right] = E\left[\eta^2\right]$; (ii) $E\left[\eta^3|X^*\right] = E\left[\eta^3\right] = 0$; (iii) $\sigma_0 = 1$.