Submitted to Econometrica

On the Use of Instrumental Variables to Identify the Effect of a Mis-measured, Binary Regressor

Francis J. DiTraglia and Camilo Garcia-Jimeno

July 4, 2015

Submitted to Econometrica

ON THE USE OF INSTRUMENTAL VARIABLES TO IDENTIFY THE EFFECT OF A MIS-MEASURED, BINARY REGRESSOR[1]

FRANCIS J. DITRAGLIA[a] AND CAMILO GARCIA-JIMENO[a]

Abstract goes here.

KEYWORDS: Instrumental Variables, Measurement Error, Binary Regressor, Endogeneity.

1. INTRODUCTION

Introduction goes here.

2. NOTES ON MAHAJAN (2006)

Mahajan (2006) considers regression models of the form

$$E\left[y - g(x^*, z)\right] = 0 \tag{1}$$

where x^* is an unobserved binary regressor and z is a $d_z \times 1$ vector of control regressors. Rather than x^* we observe a noisy measure x called the "surrogate" and an additional variable v that acts, in essence, as an instrumental variable. Since v does not, strictly speaking, meet the traditional requirements for an instrument, Mahajan refers to it as an "instrument-like variable" or ILV for short. Throughout the paper, Mahajan assumes that v is binary although he claims that the same idea applies to arbitrary discrete variables. The paper considers two main cases: one in which x^* is assumed to be exogenous, and another in which it is not.

[1]We thank...

[a]Dept. of Economics, University of Pennsylvania, 3718 Locust Walk, Philadelphia, PA 19104, U.S.A.; fditra@econ.upenn.edu; gcamilo@econ.upenn.edu

2.1. *The Case of Exogenous x^**

The first is based on the restriction

$$E\left[y - g(x^*, z) \mid x^*, x, z, v\right] = 0 \tag{2}$$

2.2. *The Case of Endogenous x^**

While the preceding case required x^* to be exogenous, Mahajan claims (page 640) that his identification results can be extended to account for endogeneity provided that one is willing to restrict attending to additively separable models of the form

$$y = g^*(x^*, z) + \varepsilon \tag{3}$$

In this case, the ILV is assumed to satisfy the usual instrumental variables mean independence assumption

$$E\left[\varepsilon | z, v\right] = 0 \tag{4}$$

and Equation 2 is replaced by

$$E\left[y | x^*, x, z, v\right] = E\left[y | x^*, z\right] \tag{5}$$

Unfortunately, Mahajan's proof is incorrect and the model in Equation 3 is unidentified. The mistake stems from a false analogy with the identification proof in the case of exogenous x^*. In A.2 Mahajan argues, correctly, that under 3–5 knowledge of the mis-classification rates is sufficient to identify the model even when x^* is endogenous. He then appeals to Theorem 1 to argue that the mis-classification rates are indeed identified. The proof of Theorem 1, however, depends crucially on the assumption that x^* is exogenous. Without this assumption, the mis-classification rates are unidentified,

as we now show. For ease of exposition we consider the case without covariates. Equivalently, one can interpret all of the expressions that follow as implicitly conditioned on $z = z_a$ where z_a is a value in the support of z.[1]

Without covariates we can write

$$y = \alpha + \beta x^* + \varepsilon \tag{6}$$

where $\alpha = g^*(0)$ and $\beta = g^*(1) - g^*(0)$ and the mis-classification rates become $\eta_0 = P(x = 1|x^* = 0)$ and $\eta_1 = P(x = 1|x^* = 1)$. Now define

$$m_{jk} = E\left[\varepsilon | x^* = j, v = k\right] \tag{7}$$

3. CONCLUSION

Conclusion goes here.

[1] Because the covariates are held fixed throughout the proof of Mahajan's Theorem 1, there is no loss of generality.