

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 5, 2015

# ON THE USE OF INSTRUMENTAL VARIABLES TO IDENTIFY THE EFFECT OF A MIS-MEASURED, BINARY REGRESSOR<sup>1</sup>

FRANCIS J. DiTRAGLIA<sup>a</sup> AND CAMILO GARCIA-JIMENO<sup>a</sup>

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

$$(1) \quad E[y - g(x^*, z)] = 0$$

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.

---

<sup>1</sup>We thank...

<sup>a</sup>Dept. of Economics, University of Pennsylvania, 3718 Locust Walk, Philadelphia, PA 19104, U.S.A.; [fditra@econ.upenn.edu](mailto:fditra@econ.upenn.edu); [gcamilo@econ.upenn.edu](mailto:gcamilo@econ.upenn.edu)

### 2.1. *The Case of Exogenous $x^*$*

The first is based on the restriction

$$(2) \quad E[y - g(x^*, z) \mid x^*, x, z, v] = 0$$

### 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

$$(3) \quad y = g^*(x^*, z) + \varepsilon$$

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

$$(4) \quad E[\varepsilon \mid z, v] = 0$$

and Equation 2 is replaced by

$$(5) \quad E[y \mid x^*, x, z, v] = E[y \mid x^*, z]$$

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  $z$ .<sup>1</sup>

Without covariates we can write

$$(6) \quad y = \alpha + \beta x^* + \varepsilon$$

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

$$(7) \quad m_{jk} = E[\varepsilon|x^* = j, v = k]$$

### 3. IDENTIFICATION BY HOMOSKEDASTICITY

This section uses our notation rather than Mahajan's. We'll have to decide what notation we want to use in the paper itself but for the moment I'm trying to avoid confusion by talking about Mahajan's proofs using his own notation while keeping our derivations in the same notation we used on the whiteboard. I think that by assuming the instrument takes on three values (as in Lewbell) and imposing our homoskedasticity assumption we'll get identification in the case where  $T^*$  is endogenous so written out this derivation for arbitrary discrete  $z$ .

Now suppose that one is prepared to assume that

$$(8) \quad E[u^2|z] = E[u^2].$$

When combined with the usual IV assumption,  $E[u|z] = 0$ , this implies  $Var(u|z) = Var(u)$ . Whether this assumption is reasonable, naturally, depends on the application. When  $z$  is the offer of treatment in a randomized

---

<sup>1</sup>Because the covariates are held fixed throughout the proof of Mahajan's Theorem 1, there is no loss of generality.

controlled trial, for example, Equation 8 holds automatically as a consequence of the randomization. Similarly, in studies based on a “natural” rather than controlled experiment one typically argues that the instrument is not merely uncorrelated with  $u$  but *independent* of it, so that Equation 8 follows.

To see why homoskedasticity with respect to the instrument provides additional identifying information, first express the conditional variance of  $y$  as follows

$$(9) \quad \text{Var}(y|z) = \beta^2 \text{Var}(T^*|z) + \text{Var}(u|z) + 2\beta \text{Cov}(T^*, u|z)$$

Under 8,  $\text{Var}(u|z)$  does not depend on  $z$ . Hence the *difference* of conditional variances evaluated at two values  $z_a$  and  $z_b$  in the support of  $z$  is simply

$$(10) \quad \Delta \text{Var}(y|z_a, z_b) = \beta^2 \Delta \text{Var}(T^*|z_a, z_b) + 2\beta \Delta \text{Cov}(T^*, u|z_a, z_b)$$

Where  $\Delta \text{Var}(y|z_a, z_b) = \text{Var}(y|z = z_a) - \text{Var}(y|z = z_b)$ , and we define  $\Delta \text{Var}(T^*|z_a, z_b)$  and  $\Delta \text{Cov}(T^*, u|z_a, z_b)$  analogously.

First we simplify the  $\Delta \text{Var}(T^*|z_a, z_b)$  term. Since  $T$  is conditionally independent of  $z$  given  $T^*$ ,

$$\begin{aligned} P(T = 1|z) &= E_{T^*|z} [E(T|z, T^*)] = E_{T^*|z} [E(T|T^*)] \\ &= P(T^* = 1|z) (1 - \alpha_1) + [1 - P(T^* = 1|z)] \alpha_0 \\ &= \alpha_0 + (1 - \alpha_0 - \alpha_1) P(T^* = 1|z) \end{aligned}$$

Rearranging,

$$(11) \quad P(T^* = 1|z) = \frac{P(T = 1|z) - \alpha_0}{1 - \alpha_0 - \alpha_1}$$

and accordingly,

$$(12) \quad Var(T^*|z) = \frac{[P(T=1|z) - \alpha_0][1 - P(T=1|z) - \alpha_1]}{(1 - \alpha_0 - \alpha_1)^2}$$

Thus, evaluating Equation 12 at  $z_a$  and  $z_b$  and simplifying,

$$(13) \quad \Delta Var(y|z_a, z_b) = \frac{\Delta Var(T|z_a, z_b) + (\alpha_1 - \alpha_0) \Delta E(T|z_a, z_b)}{(1 - \alpha_0 - \alpha_1)^2}$$

Turning our attention to  $\Delta Cov(T^*, u|z_a, z_b)$  first note that

$$(14) \quad Cov(T^*, u|z) = E_{T^*|z} [E(T^* u|z, T^*)] = P(T^* = 1|z) E(u|T^* = 1, z)$$

since  $E[z|u] = 0$ . Combining this with Equation 11 and evaluating at  $z_a$  and  $z_b$  gives

$$(15) \quad \Delta Cov(T^*, u|z_a, z_b) = \frac{[E(T|z_a) - \alpha_0] m_{1a} - [E(T|z_b) - \alpha_0] m_{1b}}{1 - \alpha_0 - \alpha_1}$$

where  $m_{1a} = E[u|T^* = 1, z_a]$  and  $m_{1b} = E[u|T^* = 1, z_b]$ .

#### 4. CONCLUSION

Conclusion goes here.