Available online at www.sciencedirect.com
ScienceDirect
Statistics and Probability Letters 78 (2008) 1473–1479
STATISTICS & PROBABILITY LETTERS
www.elsevier.com/locate/stapro

A note on the closed-form identification of regression models with a mismeasured binary regressor

Xiaohong Chen[a,1], Yingyao Hu[b,*], Arthur Lewbel[c,2]

[a] *Department of Economics, Yale University, Box 208281, New Haven, CT 06520-8281, USA*
[b] *Department of Economics, Johns Hopkins University, 440 Mergenthaler Hall, 3400 N. Charles Street, Baltimore, MD 21218, USA*
[c] *Department of Economics, Boston College, 140 Commonwealth Avenue, Chestnut Hill, MA 02467, USA*

Received 4 November 2007; received in revised form 12 December 2007; accepted 26 December 2007
Available online 1 January 2008

Abstract

This note considers the identification of a nonparametric regression model with an unobserved 0–1 dichotomous regressor. The sample consists of a dependent variable and a 0–1 dichotomous proxy of the unobserved regressor. We obtain nonparametric identification of every element in the model as a closed-form function of the observed moments or densities. Our identification strategy does not require any additional sample information, such as instrumental variables or a secondary sample. The closed-form solution may be used to construct estimators of the unknowns.

1. Introduction

Binary variables are widely used in statistical studies. Just drawing from economics, some well known examples include employment status, union status, and education level (diploma or not). When a model containing such variables is estimated, one major concern is that these variables may be subject to reporting errors. For example, self-reported smoking behavior may not be accurate because an individual may not want others to know that he or she smokes. Ignoring such reporting errors in regressors generally leads to inconsistent model estimates.

A well known strategy for dealing with such misreporting or misclassification errors in binary variables is to use a secondary measurement or an instrumental variable. See Aigner (1973), or more recently Mahajan (2006) and Lewbel (2007). Such additional sample information can yield parametric or nonparametric identification of the latent model. Without additional sample information, one usually can only identify bounds on features of the model, as in Klepper (1988) and Bollinger (1996). In contrast, we provide full nonparametric identification without using additional sample information.

* Corresponding author. Tel.: +1 410 516 7610.
E-mail addresses: xiaohong.chen@yale.edu (X. Chen), yhu@jhu.edu (Y. Hu), lewbel@bc.edu (A. Lewbel).
[1] Tel.: +1 203 432 5852.
[2] Tel.: +1 617 522 3678.

0167-7152/$ - see front matter
doi:10.1016/j.spl.2007.12.024

We consider the nonparametric regression model

$$Y = m\left(X^*\right) + \eta, \qquad E[\eta|X^*] = 0 \tag{1.1}$$

where Y is a scalar dependent variable, X^* is a 0–1 dichotomous regressor, and η is the regression error. The variables X^* and η are not observed. We observe a random sample of Y and a 0–1 dichotomous scalar X, where X is a proxy of the unobserved X^*.

Define $m_j = m(j)$ for $j = 0, 1$. Note that since X^* is binary, identifying the function $m(X^*)$ is equivalent to identifying the constants m_0 and m_1. We could alternatively define $a = m_0$ and $b = m_1 - a$ and without loss of generality rewrite the model as $Y = a + bX^* + \eta$. In addition to identifying m_0 and m_1 or equivalently a and b, we also identify the conditional distributions of Y (and hence η) conditional on X^* and the probability mass function of X given X^*. As we note later, our results readily extend to the case of $Y = m(X^*, W) + \eta$ where W is a vector of additional regressors that are observed without error.

Our identification relies on some assumptions regarding the regression model instead of on additional sample information. The key assumption is that the first three moments of the regression error are independent of the latent regressor. We show that the latent regression function is nonparametrically identified as a known function of observed moments. Our identification is constructive in the sense that it can directly lead to a consistent estimator. Other examples of obtaining identification in measurement error models without additional sample information include exploiting model restrictions as in Huwang and Hwang (2002) or the use of higher moment error restrictions as in Lewbel (1997) and Erickson and Whited (2002).

This note is organized as follows: Section 2 provides the main identification results and Section 3 summarizes the note and discusses extensions. All the proofs are in the Appendix.

2. Nonparametric identification

We now show how to obtain identification of the regression model (1.1). We first assume

Assumption 2.1. $X \perp \eta|X^*$.

This assumption implies that the measurement error $X - X^*$ is independent of the dependent variable Y conditional on the true value X^*. Define $m_j = m(j)$ for $j = 0, 1$. Assumption 2.1 implies that the relationship between the observed density and the latent ones becomes

$$f_{Y|X}(y|j) = f_{X^*|X}(0|j)\, f_{\eta|X^*}(y - m_0|0) + f_{X^*|X}(1|j)\, f_{\eta|X^*}(y - m_1|1) \quad \text{for } j = 0, 1. \tag{2.1}$$

This equation implies that the observed density $f_{Y|X}(y|j)$ is a mixture of two conditional densities $f_{\eta|X^*}(y - m_0|0)$ and $f_{\eta|X^*}(y - m_1|1)$. Note that we are using f to denote either a probability density function or a probability mass function, so since X and X^* are discrete, $f_{X^*|X}(1|0)$ is equivalent to $\Pr(X^* = 1|X = 0)$ for example.

Given that $E[\eta|X^*] = 0$, we then obtain the ordering of m_j from that of observed $\mu_j \equiv E(Y|X = j)$ under the following assumption:

Assumption 2.2. (i) $\mu_1 > \mu_0$; (ii) $f_{X^*|X}(1|0) + f_{X^*|X}(0|1) < 1$.

Assumption 2.2(i) is not restrictive because one can always redefine X as $1 - X$ if needed. Assumption 2.2(ii) reveals the ordering of m_1 and m_0, by making it the same as that of μ_1 and μ_0 because

$$1 - f_{X^*|X}(1|0) - f_{X^*|X}(0|1) = \frac{\mu_1 - \mu_0}{m_1 - m_0},$$

so $m_1 \geq \mu_1 > \mu_0 \geq m_0$. Assumption 2.2(ii) says that the sum of misclassification probabilities is less than 1, meaning that, on average, the observations X are more accurate predictions of X^* than pure guesses. See Lewbel (2007) for further discussion of this assumption.

Assumption 2.3. $E\left(\eta^k|X^*\right) = E\left(\eta^k\right)$ for $k = 2, 3$.

For identification we only require restrictions on two moments of η in this assumption, because we only need to solve for two unknowns, m_0 and m_1. A sufficient condition for Assumption 2.3 is that η be independent of X^*, which is stronger than necessary because it makes the assumption hold for all k. For $k = 2$, Assumption 2.3 says that the model errors are homoskedastic, and for $k = 3$ the assumption is that the error distributions conditional on X^* have the same skewness. A sufficient condition for $k = 3$ in Assumption 2.3 is symmetry of $\eta|X^*$, which would make $E\left(\eta^k|X^*\right) = 0$ for all odd k. Properties like homoskedasticity and symmetry, or more generally independence, naturally arise in some contexts; for example, these are common assumptions regarding measurement errors (in this case, η would be interpreted as measurement error in Y), or they may arise when errors are unobserved factors that are unrelated to the dichotomy given by X^*.

Identification could also be obtained using alternative restrictions on $\eta|X^*$ including possible restrictions such as quantiles or modes instead of moments. For example, one of the moments in Assumption 2.3 might be replaced with assuming that the density $f_{\eta|X^*=0}$ has zero median. Eq. (2.1) would then imply that

$$0.5 = \frac{\mu_1 - m_0}{\mu_1 - \mu_0}\int_{-\infty}^{m_0} f_{Y|X=0}(y)\mathrm{d}y + \frac{m_0 - \mu_0}{\mu_1 - \mu_0}\int_{-\infty}^{m_0} f_{Y|X=1}(y)\mathrm{d}y$$

which may uniquely identify m_0 under some testable assumptions. An advantage of Assumption 2.3 over alternative restrictions on $\eta|X^*$ is that we obtain a closed-form solution for m_0 and m_1 (see Chen et al. (2007) for the general case).

Define $v_j \equiv E\left[\left(Y - \mu_j\right)^2|X = j\right]$, $s_j \equiv E\left[\left(Y - \mu_j\right)^3|X = j\right]$,

$$C_1 \equiv \frac{\left(v_1 + \mu_1^2\right) - \left(v_0 + \mu_0^2\right)}{\mu_1 - \mu_0}, \qquad C_2 \equiv \frac{1}{2}\left(\mu_1 - \mu_0\right)^2 + \frac{3}{2}\left(\frac{v_1 - v_0}{\mu_1 - \mu_0}\right)^2 - \frac{s_1 - s_0}{\mu_1 - \mu_0}.$$

We leave the detailed proof to the appendix and present the results as follows:

Theorem 2.1. *Suppose that Eq. (1.1), Assumptions 2.1–2.3 hold. Then the density $f_{Y|X}$ uniquely determines $f_{Y|X^*}$ and $f_{X^*|X}$. To be specific, we have*

$$m_0 = \frac{1}{2}C_1 - \sqrt{\frac{1}{2}C_2}, \qquad m_1 = \frac{1}{2}C_1 + \sqrt{\frac{1}{2}C_2},$$

$$f_{X^*|X}(1|0) = \frac{\mu_0 - \frac{1}{2}C_1}{\sqrt{2C_2}} - \frac{1}{2}, \qquad f_{X^*|X}(0|1) = \frac{\frac{1}{2}C_1 - \mu_1}{\sqrt{2C_2}} - \frac{1}{2},$$

and

$$f_{Y|X^*=j}(y) = \frac{\mu_1 - m_j}{\mu_1 - \mu_0} f_{Y|X=0}(y) + \frac{m_j - \mu_0}{\mu_1 - \mu_0} f_{Y|X=1}(y).$$

3. Summary and extensions

This note provides a closed-form identification solution for every element in a regression model with a mismeasured dichotomous regressor. Our identification does not use any additional sample information. The key identification assumption is that the first three moments of the regression error are independent of the latent regressor. When such a restriction on the latent model is reasonable in an application, our results suggest that one does not need a secondary measurement or an instrumental variable to identify the latent model. The closed-form solution may directly lead to a consistent estimator.

As noted in the introduction, with a binary X^* our model is equivalent to $Y = a + bX^* + \eta$. It may be possible to extend our method of identification to this linear model, or to other models such as polynomials, with more general distributions of X^*. Our assumptions would not suffice for identification of a linear model with arbitrary distribution with X^* ((Reiersol, 1950) provides a counterexample), but related identification results for linear and polynomial models with continuous regressors based on error moment restrictions exist in the literature, e.g., Lewbel (1997) and Erickson and Whited (2002).

Although we only consider the case where these is a single regressor X^*, the extension to

$$Y = m\left(X^*, W\right) + \eta, \qquad E[\eta|X^*, W] = 0$$

where W is an additional vector of observed error-free covariates is immediate because our assumptions and identification results for model (1.1) can all be restated as conditional upon W.

Appendix. Mathematical proofs

Proof (Theorem 2.1). First, we introduce notation as follows: for $j = 0, 1$

$$\begin{aligned}
&m_j = m(j), \qquad \mu_j = E(Y|X = j),\\
&v_j = E\left[\left(Y - \mu_j\right)^2 |X = j\right], \qquad s_j = E\left[\left(Y - \mu_j\right)^3 |X = j\right],\\
&p = f_{X^*|X}(1|0), \qquad q = f_{X^*|X}(0|1), \qquad f_{Y|X=j}(y) = f_{Y|X}(y|j).
\end{aligned}$$

We start the proof with Eq. (2.1), which is equivalent to

$$\begin{pmatrix} f_{Y|X}(y|0) \\ f_{Y|X}(y|1) \end{pmatrix} = \begin{pmatrix} f_{X^*|X}(0|0) & f_{X^*|X}(1|0) \\ f_{X^*|X}(0|1) & f_{X^*|X}(1|1) \end{pmatrix} \begin{pmatrix} f_{\eta|X^*=0}(y - m_0) \\ f_{\eta|X^*=1}(y - m_1) \end{pmatrix}. \tag{A.1}$$

Using the notation above, we have

$$\begin{pmatrix} f_{Y|X=0}(y) \\ f_{Y|X=1}(y) \end{pmatrix} = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix} \begin{pmatrix} f_{\eta|X^*=0}(y - m_0) \\ f_{\eta|X^*=1}(y - m_1) \end{pmatrix}.$$

Since $E[\eta|X^*] = 0$, we have

$$\mu_0 = (1-p)\, m_0 + p m_1 \quad \text{and} \quad \mu_1 = q m_0 + (1-q)\, m_1.$$

We may solve for p and q as follows:

$$p = \frac{\mu_0 - m_0}{m_1 - m_0} \quad \text{and} \quad q = \frac{m_1 - \mu_1}{m_1 - m_0}. \tag{A.2}$$

We also have

$$1 - p - q = 1 - \left(\frac{m_1 - m_0 + \mu_0 - \mu_1}{m_1 - m_0}\right) = \frac{\mu_1 - \mu_0}{m_1 - m_0}.$$

Assumption 2.2 implies that

$$m_1 \geq \mu_1 > \mu_0 \geq m_0.$$

and

$$\begin{pmatrix} f_{\eta|X^*=0}(y - m_0) \\ f_{\eta|X^*=1}(y - m_1) \end{pmatrix} = \frac{1}{1-p-q} \begin{pmatrix} 1-q & -p \\ -q & 1-p \end{pmatrix} \begin{pmatrix} f_{Y|X=0}(y) \\ f_{Y|X=1}(y) \end{pmatrix}.$$

Plugging the expressions for p and q into Eq. (A.2), we have

$$\begin{aligned}
&\frac{-p}{1-p-q} = \frac{m_0 - \mu_0}{\mu_1 - \mu_0}, \qquad \frac{-q}{1-p-q} = \frac{\mu_1 - m_1}{\mu_1 - \mu_0},\\
&\frac{1-p}{1-p-q} = 1 - \frac{-q}{1-p-q}, \qquad \frac{1-q}{1-p-q} = 1 - \frac{-p}{1-p-q},
\end{aligned}$$

and

$$\begin{pmatrix} f_{\eta|X^*=0}(y - m_0) \\ f_{\eta|X^*=1}(y - m_1) \end{pmatrix} = \begin{pmatrix} 1 - \dfrac{m_0 - \mu_0}{\mu_1 - \mu_0} & \dfrac{m_0 - \mu_0}{\mu_1 - \mu_0} \\ \dfrac{\mu_1 - m_1}{\mu_1 - \mu_0} & 1 - \dfrac{\mu_1 - m_1}{\mu_1 - \mu_0} \end{pmatrix} \begin{pmatrix} f_{Y|X=0}(y) \\ f_{Y|X=1}(y) \end{pmatrix}$$

$$= \begin{pmatrix} \dfrac{\mu_1 - m_0}{\mu_1 - \mu_0} & \dfrac{m_0 - \mu_0}{\mu_1 - \mu_0} \\ \dfrac{\mu_1 - m_1}{\mu_1 - \mu_0} & \dfrac{m_1 - \mu_0}{\mu_1 - \mu_0} \end{pmatrix} \begin{pmatrix} f_{Y|X=0}(y) \\ f_{Y|X=1}(y) \end{pmatrix}.$$

In other words, we have for $j = 0, 1$

$$f_{\eta|X^*=j}(y) = \frac{\mu_1 - m_j}{\mu_1 - \mu_0} f_{Y|X=0}(y + m_j) + \frac{m_j - \mu_0}{\mu_1 - \mu_0} f_{Y|X=1}(y + m_j). \tag{A.3}$$

In summary, $f_{X^*|X}$ (or p and q) and $f_{\eta|X^*}$ are identified if we can identify m_0 and m_1. Next, we show that m_0 and m_1 are indeed identified. By Assumption 2.3, we have $E\left(\eta^k|X^*\right) = E\left(\eta^k\right)$ for $k = 2, 3$. For $k = 2$, we consider

$$\begin{aligned} v_1 &= E\left[\left(m(X^*) - \mu_1\right)^2 |X = 1\right] + E\left(\eta^2\right) \\ &= E\left[m(X^*)^2 |X = 1\right] - \mu_1^2 + E\left(\eta^2\right) = q m_0^2 + (1 - q) m_1^2 - \mu_1^2 + E\left(\eta^2\right). \end{aligned}$$

Similarly, we have

$$v_0 = (1 - p) m_0^2 + p m_1^2 - \mu_0^2 + E\left(\eta^2\right).$$

We eliminate $E\left(\eta^2\right)$ to obtain

$$(1 - p) m_0^2 + p m_1^2 - \left(v_0 + \mu_0^2\right) = q m_0^2 + (1 - q) m_1^2 - \left(v_1 + \mu_1^2\right).$$

That is

$$\left(v_1 + \mu_1^2\right) - \left(v_0 + \mu_0^2\right) = (1 - p - q)\left(m_1^2 - m_0^2\right).$$

We have shown that

$$1 - p - q = \frac{\mu_1 - \mu_0}{m_1 - m_0}.$$

Thus, m_1 and m_0 satisfy the following linear equation:

$$m_1 + m_0 = \frac{\left(v_1 + \mu_1^2\right) - \left(v_0 + \mu_0^2\right)}{\mu_1 - \mu_0} \equiv C_1.$$

This means that we need one more restriction to identify m_1 and m_0. We consider

$$\begin{aligned} s_1 &= E\left[(Y - \mu_1)^3 |X = 1\right] = E\left[\left(m\left(X^*\right) - \mu_1\right)^3 |X = 1\right] + E\left[\eta^3\right] \\ &= q (m_0 - \mu_1)^3 + (1 - q)(m_1 - \mu_1)^3 + E\left[\eta^3\right] \end{aligned}$$

and

$$s_0 = (1 - p)(m_0 - \mu_0)^3 + p (m_1 - \mu_0)^3 + E\left[\eta^3\right].$$

We eliminate $E\left(\eta^3\right)$ in the two equations above to obtain

$$(1 - p)(m_0 - \mu_0)^3 + p (m_1 - \mu_0)^3 - s_0 = q (m_0 - \mu_1)^3 + (1 - q)(m_1 - \mu_1)^3 - s_1.$$

Plugging in the expressions for p and q in Eq. (A.2), we have

$$-(m_1 - \mu_0)(m_0 - \mu_0)(m_1 + m_0 - 2\mu_0) - s_0 = -(m_1 - \mu_1)(m_0 - \mu_1)(m_1 + m_0 - 2\mu_1) - s_1.$$

Since $m_1 + m_0 = C_1$, we have

$$(C_1 - m_0 - \mu_0)(m_0 - \mu_0)(C_1 - 2\mu_0) + s_0 = (C_1 - m_0 - \mu_1)(m_0 - \mu_1)(C_1 - 2\mu_1) + s_1,$$

that is,

$$-\left(m_0^2-\mu_0^2\right)(C_1-2\mu_0)+(m_0-\mu_0)C_1(C_1-2\mu_0)+s_0$$
$$=-\left(m_0^2-\mu_1^2\right)(C_1-2\mu_1)+(m_0-\mu_1)C_1(C_1-2\mu_1)+s_1.$$

Moreover, we have

$$-2m_0^2(\mu_1-\mu_0)+2C_1(\mu_1-\mu_0)m_0$$
$$=\mu_1^2(C_1-2\mu_1)-\mu_0^2(C_1-2\mu_0)-\mu_1C_1(C_1-2\mu_1)+\mu_0C_1(C_1-2\mu_0)+s_1-s_0$$
$$=\left(\mu_1^2-\mu_0^2\right)C_1-2\left(\mu_1^3-\mu_0^3\right)-(\mu_1-\mu_0)C_1^2+2\left(\mu_1^2-\mu_0^2\right)C_1+s_1-s_0.$$

Since $(\mu_1-\mu_0)>0$, we have

$$-2m_0^2+2C_1m_0=3(\mu_1+\mu_0)C_1-2\frac{\mu_1^3-\mu_0^3}{\mu_1-\mu_0}-C_1^2+\frac{s_1-s_0}{\mu_1-\mu_0}.$$

Finally, we have

$$-2\left(m_0-\frac{1}{2}C_1\right)^2+C_2=0$$

where

$$\begin{aligned}
C_2 &= \frac{3}{2}C_1^2-3(\mu_1+\mu_0)C_1+2\frac{\mu_1^3-\mu_0^3}{\mu_1-\mu_0}-\frac{s_1-s_0}{\mu_1-\mu_0}\\
&= \frac{3}{2}[C_1-(\mu_1+\mu_0)]^2-\frac{3}{2}(\mu_1+\mu_0)^2+2\left(\mu_1^2+\mu_1\mu_0+\mu_0^2\right)-\frac{s_1-s_0}{\mu_1-\mu_0}\\
&= \frac{3}{2}[C_1-(\mu_1+\mu_0)]^2+\frac{1}{2}(\mu_1-\mu_0)^2-\frac{s_1-s_0}{\mu_1-\mu_0}\\
&= \frac{1}{2}(\mu_1-\mu_0)^2+\frac{3}{2}\left(\frac{\upsilon_1-\upsilon_0}{\mu_1-\mu_0}\right)^2-\frac{s_1-s_0}{\mu_1-\mu_0}\\
s_j &= E\left[\left(Y-\mu_j\right)^3|X=j\right]\\
&= E\left[Y^3|X=j\right]-3E\left[Y^2|X=j\right]\mu_j+3\mu_j^3-\mu_j^3\\
&= E\left[Y^3|X=j\right]-3E\left[Y^2|X=j\right]\mu_j+2\mu_j^3\\
&\equiv \kappa_j-3\upsilon_j\mu_j+2\mu_j^3,\\
C_2 &= \frac{3}{2}C_1^2-3(\mu_1+\mu_0)C_1+2\frac{\mu_1^3-\mu_0^3}{\mu_1-\mu_0}-\frac{s_1-s_0}{\mu_1-\mu_0}\\
&= \frac{3}{2}C_1^2-3(\mu_1+\mu_0)C_1+2\frac{\mu_1^3-\mu_0^3}{\mu_1-\mu_0}-\frac{\kappa_1-3\upsilon_1\mu_1+2\mu_1^3-\left(\kappa_0-3\upsilon_0\mu_0+2\mu_0^3\right)}{\mu_1-\mu_0}\\
&= \frac{3}{2}C_1^2-3(\mu_1+\mu_0)C_1-\frac{\kappa_1-3\upsilon_1\mu_1-(\kappa_0-3\upsilon_0\mu_0)}{\mu_1-\mu_0}\\
&= \frac{3}{2}C_1^2-3(\mu_1+\mu_0)\frac{\upsilon_1-\upsilon_0}{\mu_1-\mu_0}+\frac{3\upsilon_1\mu_1-3\upsilon_0\mu_0}{\mu_1-\mu_0}-\frac{\kappa_1-\kappa_0}{\mu_1-\mu_0}\\
&= \frac{3}{2}\left(\frac{\upsilon_1-\upsilon_0}{\mu_1-\mu_0}\right)^2-3\frac{\upsilon_0\mu_1-\upsilon_1\mu_0}{\mu_1-\mu_0}-\frac{\kappa_1-\kappa_0}{\mu_1-\mu_0}.
\end{aligned}$$

Notice that we also have

$$-2\left(m_1-\frac{1}{2}C_1\right)^2+C_2=0,$$

which implies that m_1 and m_0 are two roots of this quadratic equation. Since $m_1 > m_0$, we have

$$m_0 = \frac{1}{2}C_1 - \sqrt{\frac{1}{2}C_2}, \qquad m_1 = \frac{1}{2}C_1 + \sqrt{\frac{1}{2}C_2}.$$

After we have identified m_0 and m_1, p and q (or $f_{X^*|X}$) are identified from Eq. (A.2), and the density f_η(or $f_{Y|X^*}$) is also identified from Eq. (A.3). Thus, we have identified the latent densities $f_{Y|X^*}$ and $f_{X^*|X}$ from the observed density $f_{Y|X}$. ■

References

Aigner, D.J., 1973. Regression with a binary independent variable subject to errors of observation. Journal of Econometrics 1, 49–59.

Bollinger, C.R., 1996. Bounding mean regressions when a binary regressor is mismeasured. Journal of Econometrics 73, 387–399.

Chen, X., Hu, Y., Lewbel, A., Nonparametric identification and estimation of nonclassical errors-in-variables models without additional information, Cemmap Working Papers CWP18/07. (Centre for Microdata Methods and Practice) 2007.

Erickson, T., Whited, T.M., 2002. Two-step gmm estimation of the errors-in-variables model using high-order moments. Econometric Theory 18, 776–799.

Huwang, L, Hwang, J.T.G., 2002. Prediction and confidence intervals for nonlinear measurement error models without identifiability information. Statistics and Probability Letters 58, 355–362.

Klepper, S., 1988. Bounding the effects of measurement error in regressions involving dichotomous variables. Journal of Econometrics 37, 343–359.

Lewbel, A., 1997. Constructing instruments for regressions with measurement error when no additional data are available, with an application to patents and R&D. Econometrica 65, 1201–1213.

Lewbel, A., 2007. Estimation of average treatment effects with misclassification. Econometrica 75, 537–551.

Mahajan, A., 2006. Identification and estimation of regression models with misclassification. Econometrica 74, 631–665.

Reiersol, O., 1950. Identifiability of a linear relation between variables which are subject to error. Econometrica 18, 375–389.