Using Instrumental Variables to Estimate Models With Mismeasured Regressors

Arthur Lewbel*
Boston College
Intended for inclusion in a forthcoming Chapman and Hall/CRC handbook on Measurement Error Models
June 2019

Abstract

For models with mismeasured regressors, instruments are variables that are correlated with the mismeasured regressors but are uncorrelated with both measurement and model errors. This chapter provides a summary of ways in which instrumental variable methods are used to consistently estimate regression model coefficients in models containing mismeasured regressors. These methods are primarily useful when little is known about the distribution of the measurement error, e.g., when validation data are not available.

*Corresponding address: Arthur Lewbel, Dept of Economics, Maloney Hall, Boston College, 140 Commonwealth Ave., Chestnut Hill, MA, 02467, USA. (617)-552-3678, lewbel@bc.edu, https://www2.bc.edu/arthur-lewbel/
1 Introduction

This chapter discusses the use of instrumental variables (IV) methods for dealing with measurement errors in regression model covariates. In this context, an instrument is defined as an observed variable that correlates with the mismeasured variable, but does not correlate with the measurement error, and does not correlate with the model error. IV methods are used to deal with many sources of endogeneity (i.e., regressors that are correlated with model errors) in regression models, but this chapter will focus on their use with mismeasured regressors. IVs are used in many branches of statistics, but are particularly common in econometrics, where the information required for other methods of dealing with endogeneity is often unavailable. For example, in the case of mismeasured regressors, econometrics applications often lack information such as validation samples or knowledge of the error distribution.

There is some debate regarding the origins of IV methods. An early example of IV estimation was proposed by Sewall Wright (1925). Sewall’s father, Philip Wright (1928) showed that Sewall’s estimator could solve the problem of estimating the coefficient $b$ in a linear regression model $Y = a + bX^* + e$ where the observed $X^*$ and unobserved error $e$ are correlated (in his application of estimating the demand for pig iron, the correlation was due to simultaneity in the determination of $Y$ and $X^*$ rather than measurement error). More about the Wrights’ work in this area can be found in Stock and Trebbi (2003), who also applied a stylometric analysis of their writing to verify that it was indeed Philip and not Sewall who composed this solution to this correlation problem.

Early analyses of regression coefficient estimation in the presence of measurement error are Adcock (1877, 1878) and Kummell (1879), who considered measurement errors in a Deming regression, as popularized in W. Edwards Deming (1943). This is a regression that minimizes the sum of squares of errors measured perpendicular to the fitted line. Gini (1921) gave an example of an estimator that deals with measurement errors in a standard linear regression, but early examples of what are essentially IV estimators are given by Frisch (1934), Koopmans (1937), and Durbin (1954).
2 The IV Idea

Consider estimation of the coefficient $b$ in the linear regression model

$$Y = a + Xb + U,$$

where $Y$ is an outcome variable, $a$ and $b$ are unknown constants, $X$ is a covariate of interest, and $U$ is a random mean zero error. Suppose we do not have a sample of observations of $Y$ and $X$, but our data does contain observations of $Y$ and $X^*$, where $X^* = X + V$. Here $X^*$ is a measure of $X$, where $X$ is the unobserved, correctly measured covariate of interest, and the random variable $V$ is measurement error in $X^*$. For now we maintain the classical measurement error and regression assumptions that $X$, $U$, and $V$ are mutually uncorrelated with each other.

Replacing $X$ with $X^* - V$ in the regression model gives $Y = a + X^*b + e$ where $e = U - Vb$. This linear regression of $Y$ on $X^*$ suffers from the endogeneity problem that $X^*$ and $e$ are correlated with each other, because both depend on $V$. In particular, $\text{cov}(X^*, e) = E[(X + V)(U - Vb)] = E(V^2)b$ which does not equal zero unless there is either no measurement error or the true coefficient $b$ is zero. If we had more information about the distribution of either $X$ or $V$, such as knowing the variance of either one (obtained perhaps from a validation sample), we could use that information along with the data on $Y$ and $X^*$ to recover a consistent estimate of $b$, based on suitably adjusting the ordinary least squares estimate of $b$.

To consistently estimate the coefficient $b$ without such distribution information regarding $X$ or $V$, assume that in addition to observations of $Y$ and $X^*$, we also observe a variable $Z$, which is the instrument. What makes $Z$ be an instrument is that it’s assumed to be correlated with $X$ and uncorrelated with both $V$ and $U$. Just having observations of $Y$, $X^*$, and $Z$ does not provide enough information to recover observations of $X$, or to estimate the distribution of $X$, but it does enable estimation of $b$. In particular

$$\text{cov}(Z, X^*) = \text{cov}(Z, X + V) = \text{cov}(Z, X) + \text{cov}(Z, V) = \text{cov}(Z, X)$$
and
\[
\text{cov}(Z, Y) = \text{cov}(Z, Xb + U) = \text{cov}(Z, X)b + \text{cov}(Z, U) = \text{cov}(Z, X)b
\]
so
\[
\frac{\text{cov}(Z, Y)}{\text{cov}(Z, X^*)} = \frac{\text{cov}(Z, X)b}{\text{cov}(Z, X)} = b.
\]
The key assumptions used for this derivation are the instrument properties \(\text{cov}(Z, U) = 0\), \(\text{cov}(Z, V) = 0\), and \(\text{cov}(Z, X) \neq 0\). The associated IV estimator of \(b\) just replaces these covariances with estimated covariances, so given \(n\) observations \(y_i, x_i, z_i\), the estimator is
\[
\widehat{b} = \frac{\sum_{i=1}^{n} z_i (y_i - \bar{y})}{\sum_{i=1}^{n} z_i (x_i - \bar{x})}
\]
where \(\bar{y}\) and \(\bar{x}\) are the sample averages of \(y\) and \(x\). This \(\widehat{b}\) will then be consistent as long as a law of large numbers can be applied. The constant \(a\) can also be consistently estimated by \(\widehat{a} = \bar{y} - \bar{x} \widehat{b}\).

Another way to think about the above construction is to write the model as \(Y = a + X^*b + e\) where, as shown above, the unobserved error \(e\) is given by \(e = U - bV\). The assumption that \(Z\) is uncorrelated with both the measurement error \(V\) and the underlying model error \(U\) means that \(\text{cov}(e, Z) = 0\). Therefore,
\[
\text{cov}(Z, Y) = \text{cov}(Z, X^*)b + \text{cov}(Z, e) = \text{cov}(Z, X^*)b
\]
and solving for \(b\) gives \(b = \text{cov}(Z, Y)/\text{cov}(Z, X^*)\) as before. The ordinary least squares estimator is just the special case of this IV estimator where \(Z = X^*\), that is, when the regressor \(X^*\) satisfies the conditions required to be an instrument.

This construction shows that the IV method does not just apply to models with measurement errors. Given any regression model \(Y = a + bX^* + e\), all that is required for IV estimation is that we have \(\text{cov}(Z, X^*) \neq 0\) and \(\text{cov}(e, Z) = 0\). Measurement error in \(X^*\) is just one situation where the endogeneity problem \(\text{cov}(e, X^*) \neq 0\) arises, and therefore requires an alternative like IV estimation.

With any such endogeneity problem, we can say that a candidate instrument \(Z\) is valid if \(\text{cov}(e, Z) = 0\), and is useful if \(\text{cov}(Z, X^*) \neq 0\). It is easy to find instruments that are merely valid (such as any \(Z\)’s drawn from a random number generator); the difficulty is finding useful
instruments. An instrument $Z$ is defined to be weak if it’s valid, but its correlation with $X^*$ is close to zero. Weak instruments will be discussed briefly in a later section. Generally, the stronger the instrument, meaning the higher is the correlation between the instrument $Z$ and the true regressor $X$, the lower will be the variance of the IV estimated $\hat{b}$.

Where do we find an instrument $Z$? It may be easiest to consider a specific example. Consider a model where $X$ is a person’s true income, $X^*$ is his or her reported income, and $Y$ is how much the individual spends to buy food. One requirement of an instrument $Z$ is that it correlate with $X$. In this example, things we might observe that correlate with a person’s income could be his or her wage rate, or wealth measures like indicators of whether the person owns a car or a house. A second requirement of $Z$ is that it not be correlated with the measurement error $V$. So in these examples, if income and wages or car and home ownership are all self reported by the individual being surveyed, we might be concerned that mismeasurement in both $X^*$ and $Z$ could be correlated, as with an individual who chooses to exaggerate both his or her income and his or her wealth. However, if, e.g., car or home ownership is observed by the surveyer, then that error in observation is unlikely to be correlated with a respondent’s reporting error in income. As this example shows, it is permitted for both $X^*$ and $Z$ to be mismeasured, as long as the measurement errors in the two are not correlated. Finally, the third requirement is that $Z$ be uncorrelated with the model error $U$. So, e.g., if an individual’s expenditures on food $Y$ correlate not only with his or her income but also with whether he or she owns a car (which could affect how often and where he or she shops), then car ownership would not be a valid instrument.

Another possible source of instruments is repeated measures. One measure of $X^*$ might serve as an instrument for another measure of $X^*$. In the above example, the same consumer might report his income both in a survey and on his tax form, yielding two potentially different measures of his income. Or the consumer might be asked his income in two different time periods. Even if his true income changes over time, one measure could still be a valid instrument for the other, since there is typically a high correlation between an individual’s income in one time period and their income in
the next. However, one concern with repeated measures is that, for one to be a valid instrument of
the other, the errors in the two measures need to be uncorrelated. This condition might be violated
if, e.g., the errors are due to making the same kind of reporting mistake in both periods, such as
omitting some source of income in both, or purposely exaggerating one’s income in both.

A convenient feature of IV estimation of the linear model is that instruments do not need to be
continuously distributed. For example, an instrument $Z$ can be both valid and useful even if $X^*$ is
continuous and $Z$ is binary, as in the earlier example of car or home ownership instrumenting for
income.

Suppose that we have more than one instrument available, e.g., suppose both $Q$ and $R$ are valid
instruments for $X^*$. Then letting $Z$ equal $Q$ or letting $Z$ equal $R$ gives two different possible
estimators. Moreover, if we let $Z = c_0 + c_1 Q + c_2 R$ for any constants $c_0$, $c_1$, and $c_2$, then $Z$ will
still generally be a valid instrument. What $Z$ (i.e., what values of $c_0$, $c_1$, and $c_2$) should we choose?

The IV estimator is

$$\hat{b} = \frac{\widehat{\text{cov}}(Z, a + Xb + U)}{\widehat{\text{cov}}(Z, X^*)} = \frac{\widehat{\text{cov}}(Z, a + (X^* - V)b + U)}{\widehat{\text{cov}}(Z, X^*)} = b + \frac{\widehat{\text{cov}}(Z, U - Vb)}{\widehat{\text{cov}}(Z, X^*)}$$

This expression shows that, other things equal, the larger is the estimated covariance between $Z$
and $X^*$, the smaller is the error in the estimator $\hat{b}$. This in turn suggests choosing the constants
$c_0$, $c_1$, and $c_2$ to maximize $\widehat{\text{cov}}(Z, X^*)$. This in turn is equivalent to just linearly regressing $X^*$
on a constant, $Q$, and $R$ using ordinary least squares, and letting $Z$ be the fitted values from
that regression. This estimator is called two stage least squares. Note that simply rescaling $Z$
to increase $\widehat{\text{cov}}(Z, X^*)$ is not helpful, because, e.g., doubling $Z$ would also double the unobserved
sample covariance of $Z$ with $U - Vb$, leaving the estimation error $\hat{b} - b$ unchanged.

3 Linear IV Models

The estimators of the previous section directly generalize to multiple regression with multiple mis-
measured variables and multiple instruments. Consider the model $Y = X^*b + e$, where $Y$ is now an
An $n \times 1$ vector of observations of the dependent variable, $X^*$ is an $n \times k$ matrix of covariate observations (some or all of which may be mismeasured), $b$ is a $k \times 1$ vector of coefficients to be estimated, and $e$ is an $n \times 1$ vector of mean zero errors. Let $Z$ be an $n \times k$ matrix of instruments. Each column of $X^*$ and of $Z$ is a variable. Any variable that is not mismeasured can be an instrument for itself, that is, if a given column $j$ of $X^*$ is not mismeasured, then we can let column $j$ of $Z$ equal column $j$ of $X^*$. Note that the constant term, a vector of ones, would usually be included in both $X^*$ and $Z$.

The IV estimator of $b$ is then
\[ \hat{b} = \left( Z^T X^* \right)^{-1} Z^T Y \]

Substituting in $X^* b + e$ for $Y$ shows immediately that the estimation error is $\hat{b} - b = \left( Z^T X^* \right)^{-1} Z^T e$, so $\hat{b}$ is consistent if $\text{plim}_{n \to \infty} \left( Z^T X^* \right)^{-1} Z^T e = 0$. Validity of the instrument matrix $Z$ is the assumption that $E \left( Z^T e \right) = 0$, and usefulness of $Z$ requires that $Z^T X^*$ be nonsingular. $\sqrt{n}$-consistency and asymptotic normality of $\hat{b}$ follows immediately as long as $\text{plim}_{n \to \infty} Z^T X^*/n$ is nonsingular and a central limit theorem can be applied to the average $Z^T e/n$. In particular, letting $\hat{e} = Y - X^* \hat{b}$, and assuming the elements of $e$ are IID (independently and identically distributed) then the limit distribution is
\[ \sqrt{n} \left( \hat{b} - b \right) \xrightarrow{d} N \left( 0, \Psi \right) \]
\[ \Psi = \text{plim}_{n \to \infty} \frac{\hat{e}^T \hat{e}}{n} \left( \frac{Z^T X^*}{n} \right)^{-1} \left( \frac{Z^T Z}{n} \right) \left( \frac{X^* Z}{n} \right)^{-1} \]

The IV estimator can alternatively be derived as a method of moments estimator. The moments we have are $E \left( Z e \right) = 0$. The method of moments replaces expectations with sample averages, giving
\[ \frac{1}{n} Z \left( Y - X^* \hat{b} \right) = 0 \]
which, when one solves for $\hat{b}$, equals the IV estimator.

Suppose we have $L \geq k$ instruments. Let $Q$ be the $n \times L$ matrix of available instruments. Then the two stage least squares estimator is to first construct $Z$ by
\[ Z = X^* \left( Q^T Q \right)^{-1} Q^T X^* \]
and then let \( \hat{b} = \left( Z^T X^* \right)^{-1} Z^T Y \) as before. This construction of \( Z \) is just the fitted values of a linear least squares regression of \( X^* \) on \( Q \).

In this model we have \( L \) moments, given by \( E(Qe) = 0 \). When \( L > k \), we have more moments than parameters, and so cannot directly apply the method of moments. However, we can instead apply Hansen’s (1982) Generalized Method of Moments (GMM) estimator. That estimator takes the form

\[
\hat{b} = \arg \min \left[ \frac{1}{n} Q^T \left( Y - X^* \hat{b} \right) \right]^T \hat{\Omega} \left[ \frac{1}{n} Q^T \left( Y - X^* \hat{b} \right) \right]
\]

where \( \hat{\Omega} \) is an estimated \( L \times L \) weight matrix that is chosen to minimize the asymptotic variance of \( \hat{b} \). When the errors \( e \) (which are linear functions of both the model error and the measurement errors in \( X^* \)) are homoskedastic and uncorrelated across observations, the optimal GMM estimator is numerically identical to the two stage least squares estimator. More generally, GMM can be asymptotically more efficient than two stage least squares, though it may also perform worse in finite samples.

4 IV Estimators for Nonlinear Models With Measurement Error

Return to the case where \( X \) is a scalar, but suppose now we have a quadratic model

\[
Y = a + Xb + X^2c + U.
\]

The goal now is to estimate the coefficients \( a, b, \) and \( c \). We again observe scalars \( Y, X^* \), and \( Z \) where \( X^* = X + V \), and \( Z \) is an instrument. To deal with nonlinearity, we now strengthen the instrument validity and usefulness assumptions. Let \( X = \gamma + Z\delta + W \) for some constants \( \gamma \) and \( \delta \) with \( \delta \neq 0 \), assume \( U, V \) and \( W \) have mean zero, and assume \( U, V, W, \) and \( Z \) are mutually independent (these conditions are stronger than necessary but are convenient for illustrating the estimation method).
Let $S = \gamma + Z\delta$. Then

$$E (X^* \mid Z) = \gamma + Z\delta = S.$$  

So we can linearly regress $X^*$ on a constant and on $Z$ using ordinary least squares, and the fitted value will be an estimate of $S$. Next, letting $\mu_j = E (W^j)$ where $j$ is a positive integer, observe that

$$E (Y \mid S) = a + Sb + (S^2 + \mu_2) c$$

$$= (a + \mu_2 c) + Sb + S^2 c$$

So if we linearly regress $Y$ on a constant, $S$, and $S^2$ using ordinary least squares, the coefficients will provide estimates of $(a + \mu_2 c)$, $b$, and $c$. This already gives us estimates of the desired coefficients $b$ and $c$. To obtain the remaining coefficient $a$, we have

$$X^* Y = Xa + X^2 b + X^3 c + UX + (a + Xb + X^2 c + U) V$$

from which we can calculate

$$E (X^* Y \mid S) = Sa + (S^2 + \mu_2) b + (S^3 + 3S\mu_2 + \mu_3) c$$

$$= \zeta + S(a + 3\mu_2 c) + S^2 b + S^3 c.$$  

where $\zeta = \mu_2 b + \mu_3 c$. This shows that if we linearly regress $X^* Y$ on a constant, $S$, $S^2$, and $S^3$ using ordinary least squares, the coefficient of $S$ will provide an estimate of $a + 3\mu_2 c$, which along with the previous coefficient estimates allows us to recover an estimate of the remaining coefficient $a$.  

$\sqrt{n}$-consistency and asymptotic normality of these estimators follows immediately from standard conditions that suffice for $\sqrt{n}$-consistency of linear regression coefficients.

The above procedure of running repeated regressions is inefficient. To obtain more efficient estimates, we can recast the above equations as the following collection of unconditional moments

$$E \left[ (X^* - \gamma - Z\delta) Z^j \right] = 0 \text{ for } j = 0, 1$$

$$E \left[ (Y - (a + \mu_2 c) - (\gamma + Z\delta) b - (\gamma + Z\delta)^2 c) Z^j \right] = 0 \text{ for } j = 0, 1, 2$$

$$E \left[ (X^* Y - \zeta - (\gamma + Z\delta)(a + 3\mu_2 c) - (\gamma + Z\delta)^2 b - (\gamma + Z\delta)^3 c) Z^j \right] = 0 \text{ for } j = 0, 1, 2, 3$$
Under standard conditions we may then apply the GMM estimator to this set of nine moments to
directly construct $\sqrt{n}$- consistent, asymptotically normal estimates of the desired parameters $a$, $b$, and $c$, along with the additional parameters $\gamma$, $\delta$, $\mu_2$, and $\zeta$. The GMM estimator will also provide a consistent estimate of the asymptotic covariance matrix of these parameters.

In this quadratic regression case, the instrument $Z$ could not be binary, because we require $S$, $S^2$, and $S^3$ to not be perfectly correlated with each other. However, $Z$ and therefore $S$ could still be discretely distributed. We require $S$ and therefore $Z$ to take on at least four different values. Note that the above IV estimation method could be immediately extended to allow $Z$ to be a vector of instruments, by just letting $S = \gamma + Z^T \delta$ for a vector of coefficients $\delta$. So, e.g., if $Z$ was a two element vector where both elements were binary (such as the earlier example of an indicator of car ownership and an indicator of home ownership), then that would suffice to make $S$ four valued, as required.

This same IV methodology can be extended to identifying the coefficients of polynomials of
any finite order. For details, see Hausman, Newey, Ichimura, and Powell (1991). This same idea can be further generalized to nonparametric estimation of regression functions as follows. Suppose we maintain the above uncorrelatedness assumptions about $X^*$, $U$, $V$, $W$, and $Z$, but now let

$X = h(Z) + W$ and $Y = g(X) + U$ where $g$ and $h$ are unknown smooth functions. The goal is now estimation of the function $g$. Let $S = h(Z)$, which, given some regularity, can be consistently estimated by a nonparametric regression of $X^*$ on $Z$. Let $F(\cdot)$ be the distribution function of $W$. Then

$$E(Y \mid S) = \int g(S + w) dF(w)$$

$$E(X^*Y \mid Z) = \int (S + w) g(S + w) dF(w)$$

These are two integral equations in two unknown functions $g$ and $F$. Newey (2001) proposed using these equations to identify a parameterized $g$ function, Schennach (2007) and Zinde-Walsh (2014) provided conditions that suffice for nonparametric identification of $g$, and Schennach (2007) proposed a corresponding consistent estimator. Hu and Schennach (2008) further generalized this
model by allowing for other forms of measurement error, for example, they allow $W$ to have median zero rather than mean zero. Lewbel and De Nadai (2016) extended this model to allow $X^*$ and $Y$ to both be mismeasured, with correlated measurement errors (as might result if, e.g., both are responses to questions in the same survey).

5 Additional Results and Issues

It was noted earlier that linear IV estimation allows the instrument $Z$ to be binary. There is a large literature devoted to the estimation of what is called a Local Average Treatment Effect (LATE), by applying the linear IV model when both $X^*$ and $Z$ are binary scalars. See Imbens and Angrist (1984). However, the properties that are required of an instrument $Z$ for LATE estimation are somewhat different from those needed for the estimators discussed in this chapter, and in particular do not apply when $X^*$ is mismeasured.

The estimators discussed here made the classical measurement error assumption that $X^* = X + U$ where the measurement error $U$ is uncorrelated with $X$. That assumption generally requires $X$ to be continuous (which is one reason why the assumptions for LATE do not apply to measurement error). Nevertheless, there do exist IV based estimators for regression models containing discrete mismeasured regressors. Examples include Mahajan (2006), Lewbel (2007), and Hu (2008).

Instruments $Z$ generally come from outside the model, as discussed earlier. However, there exist situations where instruments $Z$ can be constructed just from the observed variables $Y$ and $X^*$, if one is given some additional information regarding the structure of $Y$, $X^*$, and $X$. For example, return to the simple model where $Y$, $X^*$, and $X$ are scalars, $Y = a + bX + U$, $X^* = X + V$, and $U$, $V$, and $X$ are mutually independent. Suppose the measurement error $V$ is symmetrically distributed around zero and $X$ is asymmetrically distributed. Then $Z = [X^* - E(X^*)]^2$ is a valid and useful instrument, which can be estimated by replacing $E(X^*)$ with the sample average. See Lewbel (1997). Related results include Lewbel (2012), Erickson and Whited (2002), and Delaigle and Hall (2016).
A general problem that can arise in IV estimation is weak instruments, which occurs when the correlation of $X^*$ and $Z$ is close to zero. With weak instruments, in moderate sample sizes standard asymptotic theory provides poor approximations to finite sample distributions. For example, in the scalar case a weak instrument $Z$ causes both the numerator and denominator of \( \hat{b} = \frac{\text{cov}(Z, Y)}{\text{cov}(Z, X^*)} \) to be close to zero, making this estimated ratio very noisy. The corresponding elements in the asymptotic variance of $\hat{b}$ will likewise be poorly estimated. To construct better asymptotic approximations in this case, assume that $X^* = \gamma_n Z + \varepsilon$ where $Z$ and $\varepsilon$ are uncorrelated, and $\gamma_n = \gamma n^{-1/2}$ for some constant $\gamma$. This drifting parameter model makes the covariance of $X^*$ and $Z$ shrink as the sample size $n$ grows, and so provides an asymptotic theory that embodies that idea that this covariance is relatively small, no matter how large the sample. See, e.g., Staiger and Stock (1997) and Andrews, Moreira, and Stock (2006) on detecting weak instruments and on applying drifting parameter asymptotics to data with weak instruments.

6 References


