4.8 Instrumental Variables

A major complication that is emphasized in microeconometrics is the possibility of inconsistent parameter estimation due to endogenous regressors. Then regression estimates measure only the magnitude of association, rather than the magnitude and direction of causation which is needed for policy analysis.

The instrumental variables estimator provides a way to nonetheless obtain consistent parameter estimates. This method, widely used in econometrics and rarely used elsewhere, is conceptually difficult and easily misused.

We provide a lengthy expository treatment that defines an instrumental variable and explains how the instrumental variables method works in a simple setting.

4.8.1 Inconsistency of OLS

Consider the scalar regression model with dependent variable y and single regressor x. The goal of regression analysis is to estimate the conditional mean function $E[y|x]$. A linear conditional mean model, without intercept for notational convenience, specifies

$$E[y|x] = \beta x. \tag{4.42}$$

This model without intercept nests the model with intercept if dependent and regressor variables are measured as deviations from their respective means. Interest lies in obtaining a consistent estimate of β as this gives the change in the conditional mean given an *exogenous* change in x. For example, interest may lie in the effect in earnings due to an increase in schooling due to exogenous reasons, such as an increase in the minimum school leaving age, that are not a choice of the individual.

The OLS regression model specifies

$$y = \beta x + u, \tag{4.43}$$

where u is an error term. Regression of y on x yields OLS estimate $\widehat{\beta}$ of β.

Standard regression results make the assumption that the regressors are uncorrelated with the errors in the model (4.43). Then the only effect of x on y is a direct effect via the term βx. We have the following path analysis diagram

$$x \quad \longrightarrow \quad y$$
$$\nearrow$$
$$u$$
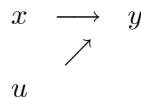
where there is no association between x and u.

But in some situations there may be an association between regressors and errors. For example, consider regression of earnings (y) on years of schooling (x). The error term u embodies all factors other than schooling that determine earnings, such as ability. Suppose a person has a high level of u, due to high (unobserved) ability. This increases earnings, since $y = \beta x + u$. But it may also lead to higher levels of x, since schooling is likely to be higher for those with high ability. A more appropriate path diagram is then the following

$$
\begin{array}{ccc}
x & \longrightarrow & y \\
\uparrow & \nearrow & \\
u & &
\end{array}
$$
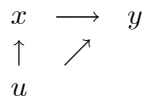
where now there is an association between x and u.

What are the consequences of this correlation between x and u? Now higher levels of x have two effects on y. From (4.43) there is both a direct effect via βx and an indirect effect via u effecting x which in turn effects y. The goal of regression is to estimate only the first effect, yielding an estimate of β. The OLS estimate will instead combine these two effects, giving $\widehat{\beta} > \beta$ in this example where both effects are positive. Using calculus, we have $y = \beta x + u(x)$ with total derivative

$$
\frac{dy}{dx} = \beta + \frac{du}{dx}. \tag{4.44}
$$

The data give information on dy/dx, so OLS estimates the total effect $\beta + du/dx$ rather than β alone. The OLS estimator is therefore biased and inconsistent for β, unless there is no association between x and u.

A more formal treatment of the linear regression model with K regressors leads to the same conclusion. From subsection 7.1 a necessary condition for consistency of OLS is that $\text{plim}\, N^{-1}\mathbf{X}'\mathbf{u} = \mathbf{0}$. Consistency requires that the regressors are asymptotically uncorrelated with the errors. From (4.37) the magnitude of the inconsistency of OLS is $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$, the OLS coefficient from regression of u on $\mathbf{x}$. This is just the OLS estimate of $du/d\mathbf{x}$, confirming the intuitive result in (4.44).

4.8.2   Instrumental Variable

The inconsistency of OLS is due to endogeneity of x, meaning that changes in x are associated not only with changes in y but also changes in the error u. What is needed is a method to generate only exogenous variation in x. An obvious way is through an experiment, but for most economics applications experiments are too expensive or even infeasible.

Definition of an Instrument

A crude experimental or treatment approach is still possible using observational data, provided there exists an **instrument** z that has the property that changes in z are associated with changes in x but do not led to change in y (aside from the indirect route via x). This leads to the following path diagram

$$z \quad \longrightarrow \quad x \quad \longrightarrow \quad y$$
$$\uparrow \quad \nearrow$$
$$u$$
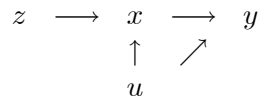
which introduces a variable z that is associated with x but not u. It is still the case that z and y will be correlated, but the only source of such correlation is the indirect path of z being correlated with x which in turn determines y. The more direct path of z being a regressor in the model for y is ruled out.

More formally, a variable z is called an **instrument** or **instrumental variable** for the regressor x in the scalar regression model $y = \beta x + u$ if (1) z is uncorrelated with the error u;and (2) z is correlated with the regressor x.

The first assumption excludes the instrument z from being a regressor in the model for y, since if instead y depended on both x and z and y is regressed on x alone then z is being absorbed into the error so that z will then be correlated with the error. The second assumption requires that there is some association between the instrument and the variable being instrumented.

Examples of an Instrument

In many microeconometric applications it is difficult to find legitimate instruments. Here we provide two examples.

Suppose we want to estimate the response of market demand to exogenous changes in market price. Quantity demanded clearly depends on price, but prices are not exogenously given since they are determined in part by market demand. A suitable instrument for price is a variable that is correlated with price but does not directly effect quantity demanded. An obvious candidate is a variable that effects market supply, since this also effect prices, but is not a direct determinant of demand. An example is a measure of favorable growing conditions if an agricultural product is being modelled. The choice of instrument here is uncontroversial, provided favorable growing conditions do not directly effect demand, and is helped greatly by the formal economic model of supply and demand.

Next suppose we want to estimate the returns to exogenous changes in schooling. Most observational data sets lack measures of individual ability, so regression of earnings on schooling has error that includes unobserved ability and hence is

correlated with the regressor schooling. We need an instrument z that is correlated with schooling, uncorrelated with ability and more generally is uncorrelated with the error term which means that it cannot directly determine earnings.

One popular candidate for z is proximity to college or university (Card, 1995). This clearly satisfies condition 2 as, for example, people whose home is a long way from a community college or state university are less likely to attend college. It most likely satisfies 1, though since it can be argued that people who live a long way from a college are more likely to be in low-wage labor markets one needs to estimate a multiple regression for y that includes as additional regressors controls such as indicators for non-metropolitan area.

A second candidate for the instrument is month of birth (Angrist and Krueger, 1991). This clearly satisfies condition 1 as there is no reason to believe that month of birth has a direct effect on earnings if the regression includes age in years. Surprisingly condition 2 may also be satisfied, as birth month determines age of first entry into school which in turn may effect years of schooling due to laws that specify a minimum school leaving age. Bound et al. (1995) provide a critique of this instrument.

The consequences of choosing poor instruments are considered in detail in section 9.

4.8.3   Instrumental Variables Estimator

For regression with scalar regressor x and scalar instrument z, the **instrumental variables (IV) estimator** is defined as

$$\widehat{\beta}_{\text{IV}} = (\mathbf{z}'\mathbf{x})^{-1}\mathbf{z}'\mathbf{y}, \tag{4.45}$$

where in the scalar regressor case $\mathbf{z}$, $\mathbf{x}$ and $\mathbf{y}$ are $N \times 1$ vectors. This estimator provides a consistent estimator for the slope coefficient β in the linear model $y = \beta x + u$ if z is correlated with x and uncorrelated with the error term.

There are several ways to derive (4.45). We provide an intuitive derivation, one that differs from derivations usually presented such as that in chapter 6.2.5.

Return to the earnings-schooling example. Suppose a one unit change in the instrument z is associated with 0.2 more years of schooling and with a \$500 increase in annual earnings. This increase in earnings is a consequence of the indirect effect that increase in z led to increase in schooling which in turn increases income. Then it follows that 0.2 years additional schooling are associated with a \$500 increase in earnings, so that a one year increase in schooling is associated with a \$500/0.2 = \$2,500 increase in earnings. The causal estimate of β is therefore 2500. In mathematical notation we have estimated the changes dx/dz and dy/dz

and calculated the causal estimator as

$$\beta_{\text{IV}} = \frac{dy/dz}{dx/dz}. \tag{4.46}$$

This approach to identification of the causal parameter β is given in Heckman (2000, p.58); see also the example in chapter 2.4.2.

All that remains is consistent estimation of dy/dz and dx/dz. The obvious way to estimate dy/dz is by OLS regression of y on z with slope estimate $(\mathbf{z'z})^{-1}\mathbf{z'y}$. Similarly estimate dx/dz by OLS regression of x on z with slope estimate $(\mathbf{z'z})^{-1}\mathbf{z'x}$. Then

$$\widehat{\beta}_{\text{IV}} = \frac{(\mathbf{z'z})^{-1}\mathbf{z'y}}{(\mathbf{z'z})^{-1}\mathbf{z'x}} = (\mathbf{z'x})^{-1}\mathbf{z'y}. \tag{4.47}$$

4.8.4 Wald Estimator

A leading simple example of IV is one where the instrument z is a **binary instrument**. Denote the sub-sample averages of y and x by $\bar{y}_1$ and $\bar{x}_1$ when $z = 1$ and by $\bar{y}_0$ and $\bar{x}_0$ when $z = 0$. Then $\Delta y/\Delta z = (\bar{y}_1 - \bar{y}_0)$ and $\Delta x/\Delta z = (\bar{x}_1 - \bar{x}_0)$, and (4.46) yields

$$\widehat{\beta}_{\text{Wald}} = \frac{(\bar{y}_1 - \bar{y}_0)}{(\bar{x}_1 - \bar{x}_0)}. \tag{4.48}$$

This estimator is called the **Wald estimator**, after Wald (1940), or the **grouping estimator**.

The Wald estimator can also be obtained from the formula (4.45). For the no-intercept model variables are measured in deviations from means, so $\mathbf{z'y} = \sum_i (z_i - \bar{z})(y_i - \bar{y})$. For binary z this yields $\mathbf{z'y} = N_1(\bar{y}_1 - \bar{y}) = N_1 N_0 (\bar{y}_1 - \bar{y}_0)/N$, where N_0 and N_1 are the number of observations for which $z = 0$ and $z = 1$. This result uses $\bar{y}_1 - \bar{y} = (N_0 \bar{y}_1 + N_1 \bar{y}_1)/N - (N_0 \bar{y}_0 + N_1 \bar{y}_1)/N = N_0(\bar{y}_1 - \bar{y}_0)/N$. Similarly $\mathbf{z'x} = N_1 N_0 (\bar{x}_1 - \bar{x}_0)/N$. Combining, (4.45) yields (4.48).

For the earnings-schooling example it is being assumed that we can define two groups where group membership does not directly determine earnings, though does affect level of schooling and hence indirectly affects earnings. Then the IV estimate is the difference in average earnings across the two groups divided by the difference in average schooling across the two groups.

4.8.5 Covariance and Correlation Analysis

The IV estimator can also be interpreted in terms of covariances or correlations.

For covariances we have directly from (4.45) that

$$\widehat{\beta}_{\text{IV}} = \frac{\text{Cov}[z, y]}{\text{Cov}[x, y]}. \tag{4.49}$$

For correlations, note that the OLS estimator for the model (4.43) can be written as $\widehat{\beta}_{\text{OLS}} = r_{xy}\sqrt{\mathbf{y'y}}/\sqrt{\mathbf{x'x}}$ where $r_{xy} = \mathbf{x'y}/\sqrt{(\mathbf{x'x})(\mathbf{y'y})}$ is the sample correlation between x and y. This leads to the interpretation of the OLS estimator as implying that a one standard deviation change in x is associated with an r_{xy} standard deviation change in y. The problem is that the correlation r_{xy} is contaminated by correlation between x and u. An alternative approach is to measure the correlation between x and y indirectly by the correlation between z and y divided by the correlation between z and x. Then

$$\widehat{\beta}_{\text{IV}} = \frac{r_{zy}}{r_{zx}} \frac{\sqrt{\mathbf{y'y}}}{\sqrt{\mathbf{x'x}}}, \tag{4.50}$$

which can be shown to equal $\widehat{\beta}_{\text{IV}}$ in (4.45).

4.8.6   IV Estimation for Multiple Regression

Now consider the multiple regression model with typical observation

$$y = \mathbf{x'}\boldsymbol{\beta} + u,$$

with K regressor variables, so $\mathbf{x}$ and $\boldsymbol{\beta}$ are $K \times 1$ vectors.

Instruments

Assume the existence of an $r \times 1$ vector of **instruments** $\mathbf{z}$, with $r \geq K$, satisfying (1) $\mathbf{z}$ is uncorrelated with the error u; (2) $\mathbf{z}$ is correlated with the regressor vector $\mathbf{x}$; and (3) $\mathbf{z}$ is strongly correlated, rather than weakly correlated, with the regressor vector $\mathbf{x}$.

The first two properties are necessary for consistency and were presented earlier in the scalar case. The third property, defined in subsection 9.1, is a strengthening of the second to ensure good finite sample performance of the IV estimator.

In the multiple regression case $\mathbf{z}$ and $\mathbf{x}$ may share some common components. Some components of $\mathbf{x}$, called **exogenous regressors**, may be uncorrelated with u. These components are clearly suitable instruments as they satisfy condition 1 and 2. Other components of $\mathbf{x}$, called **endogenous regressors**, may be correlated with u. These components lead to inconsistency of OLS and are also clearly unsuitable instruments as they do not satisfy condition 1. Partition $\mathbf{x}$ into $\mathbf{x} = [\mathbf{x}_1' \ \mathbf{x}_2']'$ where

$\mathbf{x}_1$ contains endogenous regressors and $\mathbf{x}_2$ contains exogenous regressors. Then a valid instrument is $\mathbf{z} = [\mathbf{z}_1' \ \mathbf{x}_2']'$, where $\mathbf{x}_2$ can be an instrument for itself, but we need to find as many instruments $\mathbf{z}_1$ as there are endogenous variables $\mathbf{x}_1$.

Identification

Identification in a simultaneous equations model is presented in chapter 2.5. Here we have a single equation. The **order condition** requires that the number of instruments must at least equal the number of endogenous components, so that $r \geq K$. The model is said to be **just-identified** if $r = K$ and **over-identified** if $r > K$.

In many multiple regression applications there is only one endogenous regressor. For example, the earnings on schooling regression will include many other regressors such as age, geographic location and family background. Interest lies in the coefficient on schooling, but this is an endogenous variable that is felt to be correlated with the error because ability is unobserved. Possible candidates for the necessary single instrument for schooling have already been given in subsection 8.2.

If an instrument fails the first condition the instrument is an **invalid instrument**. If an instrument fails the second condition the instrument is an **irrelevant instrument**, and the model may be **unidentified** if too few instruments are relevant. The third condition fails when there is very low correlation between the instrument and the endogenous variable being instrumented. The model is said to be **weakly identified** and the instrument is called a **weak instrument**.

Instrumental Variables Estimator

When the model is just-identified, so $r = K$, the **instrumental variables estimator** is the obvious matrix generalization of (4.45)

$$\widehat{\boldsymbol{\beta}}_{\text{IV}} = \left(\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{Z}'\mathbf{y}, \qquad (4.51)$$

where $\mathbf{Z}$ is an $N \times K$ matrix with i^{th} row $\mathbf{z}_i'$. Substituting the regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ for $\mathbf{y}$ in (4.51) yields

$$
\begin{aligned}
\widehat{\boldsymbol{\beta}}_{\text{IV}} &= \left(\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{Z}'[\mathbf{X}\boldsymbol{\beta} + \mathbf{u}] \\
&= \boldsymbol{\beta} + \left(\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{Z}'\mathbf{u} \\
&= \boldsymbol{\beta} + \left(N^{-1}\mathbf{Z}'\mathbf{X}\right)^{-1}N^{-1}\mathbf{Z}'\mathbf{u}.
\end{aligned}
$$

It follows immediately that the IV estimator is consistent if

$$\text{plim}\, N^{-1}\mathbf{Z}'\mathbf{u} = \mathbf{0}$$
$$\text{plim}\, N^{-1}\mathbf{Z}'\mathbf{X} \neq \mathbf{0}.$$

These are essentially conditions 1 and 2 that $\mathbf{z}$ is uncorrelated with $\mathbf{u}$ and correlated with $\mathbf{x}$. To ensure that the inverse of $N^{-1}\mathbf{Z}'\mathbf{X}$ exists it is assumed that $\mathbf{Z}'\mathbf{X}$ is of full rank K, a stronger assumption than the order condition that $r = K$.

With heteroskedastic errors the IV estimator is asymptotically normal with mean $\boldsymbol{\beta}$ and variance matrix consistently estimated by

$$\widehat{V}[\widehat{\boldsymbol{\beta}}_{IV}] = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\widehat{\boldsymbol{\Omega}}\mathbf{Z}(\mathbf{Z}'\mathbf{X})^{-1}, \tag{4.52}$$

where $\widehat{\boldsymbol{\Omega}}=\text{Diag}[\widehat{u}_i^2]$. This result is obtained in a manner similar to that for OLS given in subsection 4.4.

The IV estimator although consistent leads to a loss of efficiency that can be very large in practice. Intuitively IV will not work well if the instrument $\mathbf{z}$ has low correlation with the regressor $\mathbf{x}$, see subsection 9.3.

4.8.7 Two-Stage Least Squares

The IV estimator in (4.51) requires that the number of instruments equals the number of regressors. For over-identified models the IV estimator can be used, by discarding some of the instruments so that the model is just-identified. But there can be an asymptotic efficiency loss in discarding these instruments.

Instead a common procedure is to use the **two-stage least squares (2SLS) estimator**

$$\widehat{\boldsymbol{\beta}}_{2SLS} = \left[\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}\right]^{-1}\left[\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}\right], \tag{4.53}$$

presented and motivated in chapter 6.4.

The 2SLS estimator is an IV estimator. In a just-identified model it simplifies to the IV estimator given in (4.51) with instruments $\mathbf{Z}$. In an over-identified model the 2SLS estimator equals the IV estimator given in (4.51) if the instruments are $\widehat{\mathbf{X}}$, where $\widehat{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$ is the predicted value of $\mathbf{x}$ from OLS regression of $\mathbf{x}$ on $\mathbf{z}$.

The 2SLS estimator gets its name from the result that it can be obtained by two consecutive OLS regressions: OLS regression of $\mathbf{x}$ on $\mathbf{z}$ to get $\widehat{\mathbf{x}}$ followed by OLS of y on $\widehat{\mathbf{x}}$ which gives $\widehat{\boldsymbol{\beta}}_{2SLS}$. This interpretation does not necessarily generalize to nonlinear regressions; see chapter 6.5.6.

The 2SLS estimator is often expressed more compactly as

$$\widehat{\boldsymbol{\beta}}_{2SLS} = \left[\mathbf{X}'\mathbf{P_Z}\mathbf{X}\right]^{-1}\left[\mathbf{X}'\mathbf{P_Z}\mathbf{y}\right], \tag{4.54}$$

where

$$\mathbf{P_Z} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$$

is a **projection matrix** that satisfies $\mathbf{P_Z} = \mathbf{P'_Z}$, $\mathbf{P_Z P'_Z} = \mathbf{P_Z}$ and $\mathbf{P_Z Z} = \mathbf{Z}$. The 2SLS estimator can be shown to be asymptotically normal distributed with estimated asymptotic variance

$$\widehat{V}[\widehat{\boldsymbol{\beta}}_{2SLS}] = N \left[\mathbf{X'P_Z X}\right]^{-1} \left[\mathbf{X'Z(Z'Z)}^{-1}\widehat{\mathbf{S}}(\mathbf{Z'Z})^{-1}\mathbf{Z'X}\right] \left[\mathbf{X'P_Z X}\right]^{-1},$$
(4.55)

where in the usual case of heteroskedastic errors $\widehat{\mathbf{S}} = N^{-1}\sum_i \widehat{u}_i^2 \mathbf{z}_i \mathbf{z}'_i$ and $\widehat{u}_i = y_i - \mathbf{x}'_i \widehat{\boldsymbol{\beta}}_{2SLS}$. A commonly-used small sample adjustment is to divide by $N - K$ rather than N in the formula for $\widehat{\mathbf{S}}$.

In the special case that errors are homoskedastic simplification occurs and $\widehat{V}[\widehat{\boldsymbol{\beta}}_{2SLS}] = s^2[\mathbf{X'P_Z X}]^{-1}$. This latter result is given in many introductory treatments, but the more general formula (4.55) is preferred as the modern approach is to treat errors as potentially heteroskedastic.

For over-identified models with heteroskedastic errors an estimator that White (1982) calls the **two-stage instrumental variables estimator** is more efficient than 2SLS. And some commonly-used model specification tests require estimation by this estimator rather than 2SLS. For details see chapter 6.4.2.

4.8.8 IV Example

As an example of IV estimation, consider estimation of the slope coefficient of x for the dgp

$$\begin{aligned} y &= 0 + 0.5x + u \\ x &= 0 + z + v, \end{aligned}$$

where $z \sim \mathcal{N}[2, 1]$ and (u, v) are joint normal with means 0, variances 1 and correlation 0.8.

OLS of y on x yields inconsistent estimates as x is correlated with u since by construction x is correlated with v which in turn is correlated with u. IV estimation yields consistent estimates. The variable z is a valid instrument since by construction it is uncorrelated with u but is correlated with x. Transformations of z, such as z^3, are also valid instruments.

Various estimates and associated standard errors from a generated data sample of size 10,000 are given in Table 4.4. We focus on the slope coefficient.

The OLS estimator is inconsistent, with slope coefficient estimate of 0.902 being more than 50 standard errors from the true value of 0.5. The remaining estimates are consistent and are all within two standard errors of 0.5.

There are several ways to compute the IV estimator. The slope coefficient from OLS regression of y on z is 0.5168 and from OLS regression of x on z is

Table 4.4: Instrumental variables example

	OLS	IV	2SLS	IV-ineff
Constant	-0.804	-0.017	-0.017	-0.014
	(0.014)	(0.022)	(0.007)	
x	0.902	0.510	0.510	0.509
	(0.006)	(0.010)	(0.014)	(0.012)
R-squared	0.709	0.576	0.576	0.574

Note: Generated data for sample size 10,000. OLS is inconsistent and other estimators are consistent. Robust standard errors reported though unnecessary here as errors are homoskedastic. The 2SLS standard errors are incorrect. The data generating process is given in the text.

1.0124, yielding IV estimate $0.5168/1.0124 = 0.510$ using (4.47). In practice one instead directly computes the IV estimator using (4.45) or (4.51), with z used as the instrument for x and standard errors computed using (4.52). The 2SLS estimator, see (4.54), can be computed by OLS regression of y on $\widehat{x}$, where $\widehat{x}$ is the prediction from OLS regression of x on z. The 2SLS estimates exactly equal the IV estimates in this just-identified model, though the standard errors from this OLS regression of y on $\widehat{x}$ are incorrect as explained in chapter 6.4.5.

The final column uses z^3 rather than z as the instrument for x. This alternative IV estimator is consistent, since z^3 is uncorrelated with u and correlated with x. But it is less efficient for this particular dgp, and the standard error of the slope coefficient rises from 0.010 to 0.012.

There is an efficiency loss in IV estimation compared to OLS estimation, see (4.61) for a general result for the case of single regressor and single instrument. Here $\text{Cor}[x, z] = 0.71$ is high so the loss is not great and the standard error of the slope coefficient increases from 0.006 to 0.010. In practice the efficiency loss can be much greater than this.

4.9 Instrumental Variables in Practice

Important practical issues are determining whether IV methods are necessary and, if necessary, determining whether the instruments are valid. The relevant specification tests are presented in chapter 8.4. Unfortunately the validity tests are limited. They require the assumption that in a just-identified model the instruments are valid and test only over-identifying restrictions.

While IV estimators are consistent given valid instruments, as detailed below IV estimators can be much less efficient than the OLS estimator and can have finite