Kurt Schmidheiny
Unversität Basel

Instrumental Variables

1 Introduction

This handout extends the handout on "The Multiple Linear Regression Model" and refers to its definitions and assumptions in section 2. It discusses the violation of the exogeneity assumption (OLS3), its consequences and the potential solution through the use of instrumental variables.

In many applications of the linear model, we suspect that some regressors are endogenous, i.e. one or more regressors are correlated with the error term, $cov(x_{ik}, u_i) \neq 0$. In this situation, OLS cannot consistently estimate the causal effect of the regressor on the dependent variable.

Sometimes, we are able to find exogenous variables $z_{i\ell}$ which are correlated with the endogenous regressor but not correlated with the error term, i.e. $cov(z_{i\ell}, u_i) = 0$. Such variables $z_{i\ell}$ are called *instrumental variables* or *instruments*. If there are enough good such instruments, we can estimate the causal effect of the regressor on the dependent variable.

2 Canonical Examples

2.1 Omitted Variables

Consider the following regression model

$$y_i = x_{i1}'\beta_1 + x_{i2}\beta_2 + v_i$$

which conforms with standard OLS assumptions. Suppose that the variable x_2 is not observed. The estimated regression model is therefore

$$y_i = x_{i1}'\beta_1 + u_i$$

where $u_i = x_{i2}\beta_2 + v_i$. Regressors x_k in x_1 are therefore correlated with the error term u if they are correlated with the omitted variable x_2. In case x_{i1} and x_{i2} are scalars, $cov(x_{ik}, u_i) = \beta_2 cov(x_{ik}, x_{i2})$.

2.2 Simultaneity and Reversed Causality

Consider the following system of equations

$$y_{i1} = z'_{i1}\beta_1 + y_{i2}\gamma_1 + u_{i1}$$

$$y_{i2} = z'_{i2}\beta_2 + y_{i1}\gamma_2 + u_{i2}$$

where we assume that both z_1 and z_2 are uncorrelated with both u_1 and u_2. This system is called a *structural simultaneous equation system* since y_1 and y_2 are simultenously determined. The regressor y_2 depends on y_1 through the second equation. As y_1 is directly dependent on u_1, the regressor y_2 is also correlated with u_1 and hence endogenous in the first equation. Assuming that u_1 and u_2 are uncorrelated, $cov(y_{i2}, u_1) = [\gamma_2/(1-\gamma_1\gamma_2)]\sigma^2_{u_1}$. The above equation system is also described as *reversed causality* because the dependent variable y_1 has a feedback effect on the regressor y_2.

In the above example z_2 and z_1 are straightforward instruments for IV estimation of the first and second equation, respectively.[1]

2.3 Measurement Errors (Errors in Variables)

Consider the true regression model

$$y_i = \gamma_0 + \beta_1 x^*_i + u^*_i$$

[1] Instead of estimating the single structural equations directly by IV it is possible to formulate and estimate a so-called *reduced form* of the above equation system. The RHS of the reduced form equations consists of exogenous variables only. If the system is *identified*, the parameters in the structural form can be deduced from the estimated parameters in the reduced form.

which conforms the standard OLS assumptions. Suppose that the variable x^* is only observed with an error

$$x_i = x_i^* + v_i$$

where the error v is uncorrelated with x^* and with u_i^*. The estimated regression model uses x as a proxy for x^*

$$y_i = \gamma_0 + \beta_1 x_i + u_i$$

where $u_i = u_i^* - \beta_1 v_i$. The regressor x is therefore correlated with the error term u as both depend on v. Assuming independence between v and u^*, the covariance in the above example is $cov(x, u) = -\beta_1 \sigma_v^2$.

In this special case of a bivariate regression, the OLS estimator is "biased towards zero" as

$$|\text{plim } \widehat{\beta}_1| = |\beta_1| \frac{1}{1 + \frac{V(v_i)}{V(x_i)}} < |\beta_1|.$$

3 The Econometric Model

Consider the multiple linear regression model for observations $i = 1, ..., N$

$$y_i = x_i' \beta + u_i$$

where y_i is the dependent variable, x_i' is a $(K + M + 1)$-dimensional row vector of a constant, K endogenous explanatory variables and M exogenous explanatory variables. β is a $(K+M+1)$-dimensional column vector of parameters, and u_i is the error term. Each observation is furthermore described by a $(L+M+1)$-dimensional row vector z_i' of a constant, L additional exogenous variables and the M exogenous regressors. The $(L+M)$ variables in z_i are called *instruments*. The L additional variables in z_i which are not included in x_i are called *excluded instruments*. Sometimes only those L variables are called *instruments*.

The data generation process (dgp) is fully described by the following set of assumptions:

IV1: Linearity

$y_i = x_i'\beta + u_i$ and $E[u_i] = 0$

IV2: Independence

$\{x_i, z_i, y_i\}_{i=1}^N$ i.i.d. (independent and identically distributed)

IV2 means that regressors, instruments and dependent variables are independent across observations. In practice guaranteed by random sampling.

IV3: Exogeneity

$Cov[z_i, u_i] = 0$ (uncorrelated)

IV3 means that the exogenous variables (exogenous regressors and excluded instruments) are uncorrelated with the error term.

IV4: Error Variance

a) $V[u_i|z_i] = \sigma^2 < \infty$ (homoscedasticity)

b) $V[u_i|z_i] = \sigma_i^2 = g(z_i) < \infty$ (conditional heteroscedasticity)

IV5: Identifiability

$Z'X$ and $E[z_i x_i'] = Q_{ZX}$ both have rank $K+M+1 \leq L+M+1 < N$
$rank(Z) = L+M+1$ and $E[z_i z_i'] = Q_{ZZ}$ is positive definite and finite

IV5 is also called *instrument relevance* and requires that there are at least as many excluded instruments as endogenous regressors, $L \geq K$, that all instruments (but the constant) have non-zero variance and not too many extreme values, that the instruments are relevant predictors for the endogenous regressors and that the predicted endogenous regressors are not perfectly collinear, i.e. that different endogenous regressors are differently predicted by the instruments.

4 Estimation with OLS

The OLS estimator of β is biased since $E[u|X] \neq 0$ and inconsistent since $plim \frac{1}{N} X'u \neq 0$.

5 Estimation with IV (2SLS)

The instrumental variables estimator for β is

$$\widehat{\beta}_{IV} = (X'P_Z X)^{-1} X'P_Z y$$

where $P_Z = Z(Z'Z)^{-1}Z'$.

If the number of excluded instruments is larger than the number of endogenous regressors, $L > K$, the IV estimator is called *over-identified*. If the number of excluded instruments equals the number of endogenous regressors, $L = K$, the IV estimator is called *just-identified* and reduces to

$$\widehat{\beta}_{IV} = (Z'X)^{-1} Z'y\,.$$

The IV estimator can always be reformulated as

$$\widehat{\beta}_{IV} = \left(\widehat{X}'X\right)^{-1} \widehat{X}'y = \left(\widehat{X}'\widehat{X}\right)^{-1} \widehat{X}'y$$

where $\widehat{X} = P_Z X = Z(Z'Z)^{-1}Z'X$ and the matrix P_Z is symmetric and idempotent. The columns in $\widehat{X}$ are the predicted values $\widehat{x}_k$ from a regression of x_k on Z. The IV estimator can in principal be calculated by regressing in a first stage each x_k on Z and calculating the predictions $\widehat{x}_k = Z'(Z'Z)^{-1}Z'x_k$ for all $k = 1, ..., K$. The M exogenous regressors are perfectly predicted in this stage $\widehat{x}_k = x_k$ for all $k = K+1, ..., K+M$. In the second stage, y is regressed on $\widehat{X} = [1, \widehat{x}_1, ..., \widehat{x}_{K+M}]$. IV estimation is therefore also called *two-stage least squares* (2SLS).

6 Small Sample Properties of the IV Estimator

No small sample properties can be analytically established. The IV estimator is in general biased.

7 Asymptotic Properties of the IV Estimator

The following large sample properties can be established under assumptions *IV1* through *IV4*:

- The IV estimator is consistent:

$$\text{plim } \widehat{\beta}_{IV} = \beta$$

- The IV estimator is asymptotically normally distributed:

$$\sqrt{N}(\widehat{\beta}_{IV} - \beta) \xrightarrow{d} N(0, \Sigma)$$

  where $\Sigma = \left[Q_{XZ} Q_{ZZ}^{-1} Q_{ZX}\right]^{-1}$ under *IV4a*.

- The IV estimator is therefore approximately normally distributed:

$$\widehat{\beta}_{IV} \overset{A}{\sim} N\left(\beta, Avar[\widehat{\beta}_{IV}]\right)$$

  where the asymptotic variance $Avar[\widehat{\beta}]$ can be consistently estimated under *IV4a* (homoscedasticity) as

$$\widehat{Avar}[\widehat{\beta}_{IV}] = \widehat{\sigma}^2 \left(X'Z(Z'Z)^{-1}Z'X\right)^{-1} = \widehat{\sigma}^2 (\widehat{X}'\widehat{X})^{-1}$$

  with $\widehat{\sigma}^2 = \widehat{u}'\widehat{u}/N$ and under *IV4b* (heteroscedasticity) as the *robust* or *Eicker-Huber-White* estimator (see handout on "Heteroscedasticity in the linear Model")

$$\widehat{Avar}[\widehat{\beta}_{IV}] = \left(\widehat{X}'\widehat{X}\right)^{-1} \left(\sum_{i=1}^{N} \widehat{u}_i^2 \widehat{x}_i \widehat{x}_i'\right) \left(\widehat{X}'\widehat{X}\right)^{-1}$$

  with $\widehat{u}_i = y_i - x_i'\widehat{\beta}_{IV}$.

Note: The estimated asymptotic variance given in the usual output of the 2nd stage OLS regression is incorrect since $\widehat{\sigma}^2$ will be based on $\widehat{u} = y - \widehat{X}\widehat{\beta}_{IV}$ rather than $\widehat{u} = y - X\widehat{\beta}_{IV}$.

8 What are valid instruments

Valid instruments are typically derived from natural or random experiments (Angrist and Krueger 2001). Instruments are valid if the following two requirements are satisfied:

(1) *Instrument Exogeneity (IV3)*: Valid instruments are uncorrelated with the error term. This requirement needs a strong theoretical argument and can in general not be tested (see section 9). The theoretical argument has to

   (a) convincingly rule out any direct effect of the instruments on the dependent variable or any effect running through omitted variables. This is sometimes called the *exclusion restriction*.

   (b) convincingly rule out any reverse effect of the dependent variable on the instruments.

   (c) convincingly describe why the instruments influence the endogenous regressors. This is the influence after controlling for the effect through exogenous included regressors. If you do not understand why excluded instruments and endogenous regressors are correlated, then this correlation is likely a sign that that either (a) or (b) is violated.

(2) *Instrument Relevance (IV5)*: Valid instruments are highly correlated with the endogenous regressors even after controlling for the exogenous regressors. This requirement can be empirically tested in the first stage regression (see section 10).

In practice the two requirements are often conflicting.

9 Testing for the Exogeneity of the Instruments

The exogeneity of the instruments (*IV3*) can in general *not* be tested.

In case we have more instruments than necessary, $L > K$, we can perform a so-called J-test for *overidentifying restrictions*. This tests whether

all instruments are exogeneous *assuming* that a least one of the instruments is exogenous. The J-Test will therefore not necessarily detect a situation in which all instruments are endogenous.

10 Testing for the Relevance of the Instruments

Instruments that have a low correlation with the endogenous regressors after controlling for the exogenous regressors are called *weak instruments*. There is empirical and theoretical evidence that IV estimation with weak instruments has poor statistical properties and may perform even poorer than OLS (surveyed in Stock, Wright and Yogo 2002). In particular, hypothesis tests may not have correct size and confidence intervals may not be correct even in very large samples.

The relevance of the instruments is tested in the first-stage regression. As a rule of thumb, the F-statistic of a joint test whether all excluded instruments (the variables in z_i which are not in x_i) are significantly different from zero should be bigger than 10 in case of a single endogenous regressor. In case of a single instrument and a single endogenous regressor, this implies that the t-value for the instrument should be bigger than $\sqrt{10} \approx 3.2$ or the corresponding p-value below 0.0016. This F-Test should always be reported when reporting IV estimates.

11 Reduced Form Test

In the presence of weak instruments (see section 10), hypothesis tests based on IV estimates are not correct any more. Reduced form estimation offers a simple approach to test the null hypothesis H_0 that all K coefficients β_k related to the *endogenous* explanatory variables (the variables in x_i which are not in z_i) are simultaneously equal to zero.

The reduced form estimation is an OLS regression of the dependent variable y_i on all instruments z_i, i.e. on all excluded instruments and all

exogeneous regressors including a constant

$$y_i = z_i'\delta + v_i$$

where δ is a $(L + M + 1)$-dimensional column vector of parameters, and v_i is the error term. Under H_0, the excluded instruments do not have an effect on the dependent variable. The null hypothesis H_0 can therefore be tested by testing whether the L coefficients in δ related to the excluded instruments (the variables in z_i which are not in x_i) are simultaneously equal to zero in the reduced form regression. This can be tested with a standard joint Wald-test. In case of a single endogenous regressor and a single instrument, it can be tested with a standard t-test. The reduced form test does not involve the first-stage regression(s) and is therefore also correct if the instruments are weak. See Chernozhukov and Hansen (2008) for motivation and generalizations.

12 Testing for the Exogeneity of the Regressors

We may also want to know if there is an endogeneity problem in an application. This is usually tested by a (Durbin-Wu-)Hausmann test. However, the Hausman test is only valid under homoscedasticity and often involves the cumbersome generalized inversion of a non-singular matrix.

Exogeneity of the *regressors* is better tested by running an auxiliary regression (Wooldridge 2010, eq. 6.25)

$$y_i = x_i'\beta + \widehat{v}_i'\delta + e_i$$

where $\widehat{v}_i$ are the residuals from the first stage regressions for all endogenous regressors (the variables x_k which are part of X but not Z). The exogeneity test is then a joint F or Wald-Test that all K coefficients $\delta_1, ..., \delta_K$ are equal to zero. This test is robust to heteroscedasticity if the robust (Eicker-Huber-White) variance estimator is used.

Note: This is a test for the exogeneity of the *regressors* x_i and not for the exogeneity of the *instruments* z_i. If the instruments are not valid, the test is not valid either.

Implementation in Stata 14

Stata calculates the IV (2SLS) estimator by the command

```
ivregress 2sls depvar [varlist1] (varlist2=varlist3)
```

where *varlist1* are exogeneous regressors (hence included in X and Z) , *varlist2* are endogenoues regressors (only included in X) and *varlist3* are excluded instruments (only included in Z). For example, load data

```
webuse hsng2
```

and regress median monthly rents (rent) of census divisions on the share of urban population (pcturban) and the median housing value (hsngval)

```
ivregress 2sls rent pcturban (hsngval = faminc reg2-reg4), vce(robust)
```

Housing values are likely endogeneous and therefore instrumented by median family income (faminc) and 3 regional dummies (reg2, reg4, reg4). The Eicker-Huber-White covariance estimator which is robust to heteroscedasticity is reported with the option `vce(robust)`. The option `first` requests that the first-stage regression results be displayed. First stage results are also provided by the postestimation command

```
estat firststage
```

which includes the F-statistic to assess weak instruments in case of $K = 1$ or the so-called rank F-statistic in case of $K > 1$.

The J-Test is reported with the postestimation command

```
estat overid
```

The test for exogeneity of the regressors can be calculated by adding the first stage residuals to an auxiliary regression. For example,

```
regress hsngval pcturban faminc reg2-reg4
predict v, resid
regress rent hsngval pcturban v
test v
```

The reduced form test is performed by

```
regress rent pcturban faminc reg2-reg4, vce(robust)
test faminc reg2 reg3 reg4
```

Implementation in R

The IV (2SLS) estimator is conveniently implemented in the R package `AER` as command

```
> ivreg(y ~ x1 + x2 + w1 + w2 | z1 + z2 + z3 + w1 + w2)
```

where `x1` and `x2` are endogenous regressors, `w1` and `w2` exogeneous regressors, and `z1` to `z3` are excluded instruments. For example, load data

```
> library(foreign)
> hsng2 <- read.dta("http://www.stata-press.com/data/r11/hsng2.dta")
```

and regress median monthly rents (rent) of census divisions on the share of urban population (pcturban) and the median housing value (hsngval)

```
> library(foreign)
> hsng2 <- read.dta("http://www.stata-press.com/data/r11/hsng2.dta")
> fiv <- ivreg(rent~hsngval+pcturban|pcturban+faminc+reg2+reg3+reg4,
   data = hsng2)
> summary(fiv)
```

Housing values are likely endogeneous and therefore instrumented by median family income (faminc) and 3 regional dummies (reg2, reg4, reg4).

The Eicker-Huber-White covariance estimator which is robust to heteroscedastic error terms is reported after estimation with

```
> library(sandwich)
> library(lmtest)
> coeftest(fm, vcov=sandwich)
```

First stage results are reported by explicitly estimating them. E.g,

```
> first <- lm(hsngval~pcturban+faminc+reg2+reg3+reg4, data = hsng2)
> summary(first)
```

In case of a single endogenous variable ($K = 1$), the F-statistic to assess weak instruments is reported after estimating the first stage with e.g.

```
> waldtest(first, .~.-faminc-reg2-reg3-reg4)
```

or in case of heteroscedatistic errors

```
> waldtest(first, .~.-faminc-reg2-reg3-reg4, vcov=sandwich)
```

References

Introductory textbooks

Stock, James H. and Mark W. Watson (2012), Introduction to Econometrics, 3rd ed., Pearson Addison-Wesley, chapter 12.

Wooldridge, Jeffrey M. (2009), Introductory Econometrics: A Modern Approach, 4th ed., South-Western Cengage Learning, chapter 15.

Advanced textbooks

Cameron, A. Colin and Pravin K. Trivedi (2005), Microeconometrics: Methods and Applications, Cambridge University Press, 4.8–4.9.

Wooldridge, Jeffrey M. (2010), Econometric Analysis of Cross Section and Panel Data, MIT Press, chapter 5.

Davidson and MacKinnon (2004), Econometric Theory and Methods, Oxford University Press, chapter 8.

On IV and Measurement Errors

Jerry Hausman (2001), Mismeasured Variables in Econometric analysis: Problem form the Right and from the Left, *Journal of Economics Perspectives*, 15/4, 57–67.

On IV and Omitted Variables

Angrist Joshua and Alan Krueger (2001), Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments, *Journal of Economics Perspectives*, 15/4, 69–85.

On weak instruments

Stock, J. H., J. H. Wright and M. Yogo (2002), A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments, *Journal of Business and Economic Statistics*, 20(4), 518–29.

Chernozhukov, Victor and Christian Hansen (2008), The reduced form: A simple approach to inference with weak instruments, *Economics Letters*, 100, 68–71.