Instrumental Variables regression

606EC Econometria - Mod. II

2023-2023

Martin Magris

Overview I

Introduction I

Definition (IV regression)

Instrumental variables (IV) regression is a general way to obtain a consistent estimator of the unknown causal coefficients when the regressor, X, is correlated with the error term, u

Think of the variation in X as having two parts:

- one part that, for whatever reason, is correlated with u (problem)
- a second part that is uncorrelated with u

Idea: isolate the second part, and focus on those variations in X that are uncorrelated with u, disregarding the variations in X that bias the OLS estimates.

Introduction II

The information about the movements in X that are uncorrelated with u is gleaned from one or more additional variables called **instrumental variables** or simply **instruments**.

$\rightarrow$ instruments are tools to isolate the movements in X that are uncorrelated with u, which allow the consistent estimation of the regression coefficients.

The IV Estimator with a Single Regressor and a Single Instrument

- Single regressor, X, which might be correlated with the error, u
- If X and u are correlated, the OLS estimator is inconsistent

Correlation between X and u can stem from various sources

- omitted variables
- errors in variables
- simultaneous causality

$\rightarrow$ if there is a valid instrument Z available the, effect on Y of a unit change in X can be estimated using the instrumental variables estimator.

The IV Model Assumptions I

Let β_1 be the causal effect of X on Y, the model is

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \quad i = 1, \ldots, n$$

- u_i is the error term representing omitted factors that determine Y_i.
- If X_i and u_i are correlated, the OLS estimator is inconsistent.
- (i denotes the observation).
- Instrumental variables estimation uses an additional, (instrumental) variable Z to isolate that part of X that is uncorrelated with u.

The IV Model Assumptions II

Definition (endogenous and exogenous variables)

Variables correlated with the error term are called **endogenous** variables, while variables uncorrelated with the error term are called **exogenous** variables.

$\rightarrow$ Thus in IV we deal with regressors that are endogenous, through instruments Z that are thus exogenous.

The IV Model Assumptions III

Question: what makes an instrument valid?

Definition (Instrument)

A valid instrumental variable Z must satisfy two conditions, known as the instrument **relevance condition** and the instrument **exogeneity condition**:

1. Instrument relevance: $\mathbb{C}\text{orr}(Z_i, X_i) \neq 0$
2. Instrument exogeneity: $\mathbb{C}\text{orr}(Z_i, u_i) = 0$

Recall that $\mathbb{C}\text{ov}(Z_i, X_i) = \mathbb{C}\text{orr}(Z_i, X_i)\sigma_{Z_i}\sigma_{X_i}$. Since σ_{Z_i},σ_{X_i} are both non-zero:

- Instrument relevance implies $\mathbb{C}\text{ov}(Z_i, X_i) \neq 0$
- Instrument exogeneity implies $\mathbb{C}\text{ov}(Z_i, u_i) = 0$

The IV Model Assumptions IV

Interpretation:

- If an instrument is relevant, then variation in the instrument is related to variation in X_i.
- If the instrument is exogenous, then that part of the variation of X_i captured by the instrumental variable is exogenous.
- Thus an instrument that is relevant and exogenous can capture movements in X_i that are exogenous.

Two stage Least Squares Estimation I

If the instrument Z satisfies the conditions of instrument relevance and exogeneity, the coefficient can be estimated using an IV estimator called **two-stage least squares** (TSLS).

1. The first stage decomposes X into two components: a problematic component that may be correlated with the regression error and another, problem-free component that is uncorrelated with the error
2. The second stage uses the problem-free component to estimate β_1.

Two stage Least Squares Estimation II

The first stage begins with a population regression linking X and Z:

$$\begin{aligned} X_i &= \pi_0 + \pi_1 Z_i + v_i \\ &= [\pi_0 + \pi_1 Z_i] + [v_i] \end{aligned} \tag{1}$$

This regression decomposes X_i in two parts:

- $[\pi_0 + \pi_1 Z_i]$: the part of X_i that can be predicted by Z_i.
  Because Z_i is exogenous, $\rightarrow \pi_0 + \pi_1 Z_i$ is exogenous (π_0, π_1 after all are numbers) $\rightarrow$ this part of X is exogenous: $[\pi_0 + \pi_1 Z_i]$ can be interpreted as a part of X_i that is not correlated with $u \rightarrow$ this part is *not* problematic for OLS regression
- $[v_i]$: the problematic part of X_i that is correlated with u_i.

Two stage Least Squares Estimation III

Main idea:

1. "Extract" the problem-free component of X_i, $[\pi_0 + \pi_1 Z_i]$ and disregard v_i. π_0 and π_1 are however unknown and need to be estimated.
2. Run the regression of Y on this problem-free part. As this problem-free part is indeed uncorrelated with the error u_i, OLS are valid.

Two stage Least Squares Estimation IV

In practice,

1. The first stage of TSLS applies OLS to eq. (1) and keeps the predicted value from the OLS regression,

$$\hat{X}_i = \hat{\pi}_1 + \hat{\pi}_2 Z_i$$

2. The second stage of TSLS regresses Y_i on $\hat{X}_i$: the resulting estimator for the second stage are the TSLS estimators β_0^{TSLS} ,β_1^{TSLS}

The Sampling Distribution of the TSLS Estimator I

TSLS estimator with a single instrument is **[PROOF]**:

$$\hat{\beta}_1^{\text{TSLS}} = \frac{s_{ZY}}{s_{ZX}}$$

I.e., is the ratio of the sample covariance between Z and Y to the sample covariance between Z and X.

Moreover the TSLS estimator is consistent, **[PROOF]**:

$$\hat{\beta}_1^{\text{TSLS}} \xrightarrow{\text{p}} \beta_1$$

Proof the TSLS estimator I

To Show: that $\hat{\beta}_1^{\text{TSLS}} = s_{ZY}/s_{ZX}$.

Remember that for a simple linear regression of X_i on Y_i, $Y_i = a + bX_i + u_i$, the OLS estimate of b is obtained as the ratio of the sample covariance between X and Y and the sample variance of X, i.e.

$$\hat{b} = s_{XY}/s_X^2 \tag{2}$$

Also, recall the following are general relations.

$$\mathbb{C}\text{ov}(a + bX, Y) = b\,\mathbb{C}\text{ov}(X, Y), \qquad \mathbb{V}(aX) = a^2\,\mathbb{V}(X),$$

from which, at a sample level,

$$s_{(a+bX)Y} = b\,s_{XY} \tag{3}$$

$$s_X^2 = a^2 s_X^2 \tag{4}$$

Proof the TSLS estimator II

The second stage of the TSLS procedure is to regress Y_i on $\hat{X}_i$ by OLS. Therefore,

$$\hat{\beta}_1^{\text{TSLS}} = \hat{\beta}_1^{\text{OLS}} \overset{(2)}{=} \frac{s_{\hat{X}Y}}{s_{\hat{X}}^2} \tag{5}$$

Since $\hat{X}_i$ is predicted from the first-stage regression,

$$\hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i \tag{6}$$

we have that

$$\hat{\pi}_1 \overset{(1)+(2)}{=} \frac{s_{ZX}}{s_Z^2}, \qquad s_{\hat{X}Y} \overset{(6)+(3)}{=} \hat{\pi}_1 s_{ZY}, \qquad s_{\hat{X}}^2 \overset{(6)+(4)}{=} \hat{\pi}_1^2 s_Z^2$$

Therefore, using the above three in eq.(5):

$$\hat{\beta}_1^{\text{TSLS}} = \frac{s_{\hat{X}Y}}{s_{\hat{X}}^2} = \frac{\hat{\pi}_1 s_{ZY}}{\hat{\pi}_1^2 s_Z^2} = \frac{s_{ZY}}{\hat{\pi}_1 s_Z^2} = \frac{s_{ZY}}{\frac{s_{ZX}}{s_Z^2} s_Z^2} = \frac{s_{ZY}}{s_{ZX}}$$

Sampling distribution of $\hat{\beta}_1^{TSLS}$ with n large I

The variance of $\hat{\beta}_1^{TSLS}$, $\sigma^2_{\hat{\beta}_1^{TSLS}}$ can be estimated from **[PROOF]**

$$\sigma^2_{\hat{\beta}_1^{TSLS}} = \frac{1}{n}\frac{\mathbb{V}[(Z_i - \mu_Z)u_i]}{[\mathbb{C}\text{ov}(Z_i, X_i)]^2}$$

and the square root of the estimate of $\sigma^2_{\hat{\beta}_1^{TSLS}}$ is the standard error of the IV estimator.

$\rightarrow$ Hypothesis tests about β_1 can be performed by computing the t-statics

$\rightarrow$ 95% CI are given by $\hat{\beta}_1^{TSLS} \pm 1.95\text{SE}\left(\hat{\beta}_1^{TSLS}\right)$

General IV Regression Model I

The general IV has four types of variables

1. **Dependent** variable Y
2. **Endogenous** regressors X's (problematic)
3. Additional regressors W's, control variables or **exogenous** variables
4. **Instruments** Z's

In general, we can have multiple endogenous regressors X's, multiple additional regressors Ws', and multiple instruments Z's.

For IV regression to be possible, there must be at least as many instrumental variables (Z's) as endogenous regressors (X's).

For now, we focus on the case where W's variables are exogenous, so that $\mathbb{E}(u_i|W_i) = 0$, later we shall see the precise meaning of W's being control variables.

General IV Regression Model II

The regression coefficients are said

- **Exaclty identified**: if the number of instruments m equals the number of endogenous regressors k: $m = k$.
- **Overidentified**; if the number of instruments m exceeds the number of endogenous regressors k: $m > k$.
- **Underidentified**: if the number of instruments m is less than the number of endogenous regressors k: $m < k$.

$\rightarrow$ The coefficient must be either exactly identified or overidentified if they are to be estimated by VI regression.

General IV Regression Model III

The General Instrumental Variables Regression Model and Terminology

The general IV regression model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \beta_{k+1} W_{1i} + \cdots + \beta_{k+r} W_{ri} + u_i, \qquad (12.12)$$

$i = 1, \ldots, n$, where

- Y_i is the dependent variable;
- $\beta_0, \beta_1, \ldots, \beta_{k+r}$ are unknown coefficients;
- $X_{1i}, \ldots, X_{ki}$ are k endogenous regressors, which are potentially correlated with u_i;
- $W_{1i}, \ldots, W_{ri}$ are r included exogenous regressors, which are uncorrelated with u_i or are control variables;
- u_i is the error term, which represents measurement error and/or omitted factors; and
- $Z_{1i}, \ldots, Z_{mi}$ are m instrumental variables.

The coefficients are overidentified if there are more instruments than endogenous regressors ($m > k$), they are underidentified if $m < k$, and they are exactly identified if $m = k$. Estimation of the IV regression model requires exact identification or overidentification.

TSLS with a single endogenous regressor I

The relevant equation with a signal endogenous regressor X and some additional exogenous variables is

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_{1i} + \cdots + \beta_{1+r} W_{ri} + u_i \tag{7}$$

where X_i might be correlated with the error whereas $W_{1i}, \ldots W_{2i}$ are not. Note:

- 1 endogenous regressor X
- r exogenous regressor W's
- $r + 2$ parameters β's
- m instruments available Z's

TSLS with a single endogenous regressor II

The first-stage regression of TSLS relates X_i to the exogenous variables. I.e., the W's and the instruments Z's:

$$X_i = \pi_0 + \pi_1 Z_{1i} + \ldots \pi_m Z_{mi} + \pi_{m+1} W_{1i} + \cdots + \pi_{m+r} W_{ri} + v_i \qquad (8)$$

- here $\pi_0, \ldots, \pi_{m+r}$ are unknown coefficients, that are estimated by OLS.
- by OLS estimation, we obtain the predicted values from this regression: $\hat{X}_1, \ldots, \hat{X}_n$:

$$\hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_{1i} + \ldots \hat{\pi}_m Z_{mi} + \hat{\pi}_{m+1} W_{1i} + \cdots + \hat{\pi}_{m+r} W_{ri}$$

TSLS with a single endogenous regressor III

In the second stage of TSLS, eq. (7) is estimated by OLS except that X_i is replaced by its predicted value from the first stage. That is, Y_i is regressed on $\hat{X}_i, W_{1i}, \ldots, W_{ri}$ using OLS.

$$Y_i = \beta_0 + \beta_1 \hat{X}_{1i} + \beta_{k+1} W_{1i} + \cdots + \beta_{1+r} W_{ri} + u_i$$

$\rightarrow$ The resulting estimator of $\beta_0, \beta_1, \ldots, \beta_{1+r}$ is the TSLS estimator.

Extension to multiple endogenous regressors I

Now:

- k endogenous regressors X's
- r exogenous regressor W's
- $r + k + 1$ parameters β's
- m instruments available Z's

When there are multiple (k) endogenous regressors $X_{1i}, \ldots, X_{ji}, \ldots, X_{ki}$ each endogenous regressor requires its own first-stage regression.

- Each of these first-stage regressions has the same form as eq. (8). For the j-th regressor,

$$X_{ji} = \pi_0 + \pi_1 Z_{1i} + \cdots + \pi_m Z_{mi} + \pi_{m+1} W_{1i} + \cdots + \pi_{m+r} W_{ri} + v_i$$

from which:

$$\hat{X}_{ji} = \hat{\pi}_0 + \hat{\pi}_1 Z_{1i} + \cdots + \hat{\pi}_m Z_{mi} + \hat{\pi}_{m+1} W_{1i} + \cdots + \hat{\pi}_{m+r} W_{ri}$$

Extension to multiple endogenous regressors II

- In the second stage of TSLS, Equation (12.12) is estimated by OLS except that the endogenous regressors (X's) are replaced by their respective predicted values $\hat{X}_{1i}, \ldots, \hat{X}_{ji}, \ldots, \hat{X}_{ki}$:

$$Y_i = \beta_0 + \beta_1 \hat{X}_{1i} + \cdots + \beta_k \hat{X}_{ki} + \beta_{k+1} W_{1i} + \cdots + \beta_{k+r} W_{ri} + u_i$$

$\rightarrow$ The resulting estimator of $\beta_0, \beta_1, \ldots, \beta_{1+r}$ is the TSLS estimator.

Extension to multiple endogenous regressors III

Two Stage Least Squares

The TSLS estimator in the general IV regression model in Equation (12.12) with multiple instrumental variables is computed in two stages:

1. **First-stage regression(s)**: Regress X_{1i} on the instrumental variables $(Z_{1i}, \ldots, Z_{mi})$ and the included exogenous variables and/or control variables $(W_{1i}, \ldots, W_{ri})$ using OLS, including an intercept. Compute the predicted values from this regression; call these $\hat{X}_{1i}$. Repeat this for all the endogenous regressors $X_{2i}, \ldots, X_{ki}$, thereby computing the predicted values $\hat{X}_{1i}, \ldots, \hat{X}_{ki}$.
2. **Second-stage regression**: Regress Y_i on the predicted values of the endogenous variables $(\hat{X}_{1i}, \ldots, \hat{X}_{ki})$ and the included exogenous variables and/or control variables $(W_{1i}, \ldots, W_{ri})$ using OLS, including an intercept. The TSLS estimators $\hat{\beta}_0^{TSLS}, \ldots, \hat{\beta}_{k+r}^{TSLS}$ are the estimators from the second-stage regression.

In practice, the two stages are done automatically within TSLS estimation commands in econometric software.

Instrument Relevance and Exogeneity in the General IV Model I

- When there is **one** included endogenous variable but multiple instruments, the condition for instrument relevance is **that at least one** Z is useful for predicting X (given W).
- When there are **multiple** included endogenous variables, this condition is more complicated because we must rule out perfect multicollinearity in the second-stage population regression.

The conditions for valid instruments in the m-instruments read:

Instrument Relevance and Exogeneity in the General IV Model II

The Two Conditions for Valid Instruments

A set of m instruments $Z_{1i}, \ldots, Z_{mi}$ must satisfy the following two conditions to be valid:

1. **Instrument Relevance**

   - *In general*, let $\hat{X}^*_{1i}$ be the predicted value of X_{1i} from the population regression of X_{1i} on the instruments (Z's) and the included exogenous regressors (W's), and let "1" denote the constant regressor that takes on the value 1 for all observations. Then $(\hat{X}^*_{1i}, \ldots, \hat{X}^*_{ki}, W_{1i}, \ldots, W_{ri}, 1)$ are not perfectly multicollinear.

   - *If there is only one* X, then for the previous condition to hold, at least one Z must have a nonzero coefficient in the population regression of X on the Z's and the W's.

2. **Instrument Exogeneity**

   The instruments are uncorrelated with the error term; that is, $\text{corr}(Z_{1i}, u_i) = 0, \ldots, \text{corr}(Z_{mi}, u_i) = 0$.

The IV Regression Assumptions and Sampling Distribution of the TSLS Estimator I

The IV regression assumptions are modifications of the least squares assumptions. Recall that:

The Least Squares Assumptions for Causal Inference in the Multiple Regression Model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i, i = 1, \ldots, n,$$

where $\beta_1, \ldots, \beta_k$ are causal effects and

1. u_i has a conditional mean of 0 given $X_{1i}, X_{2i}, \ldots, X_{ki}$; that is,

$$E(u_i \mid X_{1i}, X_{2i}, \ldots, X_{ki}) = 0.$$

2. $(X_{1i}, X_{2i}, \ldots, X_{ki}, Y_i), i = 1, \ldots, n$, are independently and identically distributed (i.i.d.) draws from their joint distribution.
3. Large outliers are unlikely: $X_{1i}, \ldots, X_{ki}$ and Y_i have nonzero finite fourth moments.
4. There is no perfect multicollinearity.

The IV Regression Assumptions and Sampling Distribution of the TSLS Estimator II

For IV Regression:

The IV Regression Assumptions

The variables and errors in the IV regression model in Key Concept 12.1 satisfy the following:

1. $E(u_i \mid W_{1i}, \ldots, W_{ri}) = 0$;
2. $(X_{1i}, \ldots, X_{ki}, W_{1i}, \ldots, W_{ri}, Z_{1i}, \ldots, Z_{mi}, Y_i)$ are i.i.d. draws from their joint distribution;
3. Large outliers are unlikely: The X's, W's, Z's, and Y have nonzero finite fourth moments; and
4. The two conditions for a valid instrument in Key Concept 12.3 hold.

The IV Regression Assumptions and Sampling Distribution of the TSLS Estimator III

1. The first IV regression assumption modifies the conditional mean assumption in to apply only to the included exogenous variables
2. The second IV regression assumption is that the draws are i.i.d., as they are if the data are collected by simple random sampling.
3. The third IV assumption is that large outliers are unlikely.
4. The fourth IV regression assumption is that the two conditions for the instrument validity hold. The instrument relevance condition subsumes the fourth least squares assumption (no perfect multicollinearity) by assuming that the regressors in the second-stage regression are not perfectly multicollinear.

The IV Regression Assumptions and Sampling Distribution of the TSLS Estimator IV

$\rightarrow$ Under the IV regression assumptions, the TSLS estimator is **consistent** and **normally distributed in large samples**.

$\rightarrow$ Because the sampling distribution of the TSLS estimator is normal in large samples, the general procedures for statistical inference (hypothesis tests and confidence intervals) in regression models extend to TSLS regression

But:

- The standard errors reported by OLS estimation of the second-stage regression are incorrect because they do not recognize that it is the second stage of a two-stage process (use an appropriate TSLS command in an econometric software)
- Nothing prevents the error u from being heteroskedastic: use the heteroskedasticity-robust standard errors.

Checking Instrument Validity: relevance I

Main idea: whether instrumental variables regression is useful in a given application hinges on whether the instruments are valid:

$\rightarrow$ Invalid instruments produce meaningless results,
$\rightarrow$ A more relevant instrument produces a more accurate estimator "just as a larger sample size produces a more accurate estimator."
$\rightarrow$ If having a more relevant instrument is like having a larger sample size, the more relevant it is, the better the normal approximation to the sampling distribution of the TSLS estimator and its t-statistic.

Definition (Weak instrument)

Instruments that explain little of the variation in X are called **weak instruments.**

... weak instruments are a problem.

Checking Instrument Validity: relevance II

If the instruments are weak, also in large samples:

- The normal distribution provides a poor approximation to the sampling distribution of the TSLS estimator.
- There is no theoretical justification for the usual methods for performing statistical inference.
- The TSLS estimator can be badly biased in the direction of the OLS estimator.

I.e., if instruments are weak, TSLS is no longer reliable.

Checking Instrument Validity: relevance III

Proof:
Consider the special case, of a single included endogenous variable, a single instrument, and no included exogenous regressor.

If the instrument is valid $\hat{\beta}_1^{\text{TSLS}}$ is consistent, because s_{ZY} and s_{ZX} are consistent:

$$\hat{\beta}_1^{\text{TSLS}} \xrightarrow{\text{p}} \frac{\mathbb{C}\text{ov}(Z_i, Y_i)}{\mathbb{C}\text{ov}(Z_i, X_i)} = \beta_1$$

Assume that the instrument is not only weak but indeed irrelevant:

$$s_{ZX} \xrightarrow{\text{p}} \mathbb{C}\text{ov}(Z_i, X_i) = 0$$

As a consequence, the consistency argument of $\hat{\beta}_1^{\text{TSLS}}$ breaks down.

Furthermore, the estimator has a non-normal sampling distribution and is not centered around β_1, even if the sample size is large.

Checking Instrument Validity: relevance IV

Indeed, it can be proved that the distribution of $\hat{\beta}_1^{TSLS} - \beta_1$ is that of the ratio of two correlated standard normal distributions.

How relevant must the instruments be for the normal distribution to provide a good approximation in practice?

- General case: complicated.
- Single endogenous regressor: rule of thumb.

A Rule of Thumb for Checking for Weak Instruments

The first-stage F-statistic is the F-statistic testing the hypothesis that the coefficients on the instruments $Z_{1i}, \ldots, Z_{mi}$ equal 0 in the first stage of two stage least squares. When there is a single endogenous regressor, a first-stage F-statistic less than 10 indicates that the instruments are weak, in which case the TSLS estimator is biased (even in large samples) and TSLS t-statistics and confidence intervals are unreliable.

Checking Instrument Validity: relevance V

It is possible to show that when there are "many" instruments, the bias of the TSLS estimator is approximately

$$\mathbb{E}\left(\hat{\beta}_1^{\text{TSLS}}\right) - \beta_1 \approx \frac{\left(\beta_1^{\text{OLS}} - \beta_1\right)}{\mathbb{E}(F) - 1},$$

where $\mathbb{E}(F)$ is the expectation of the first-stage F-statitsics.

If $\mathbb{E}(F) = 10$, the bias is $1/9 \approx 10\%$ relative to the OLS bias: small enough to be considered acceptable.

I.e., if the bias of the TSLS estimator is less than $1/10$ of the OLS bias, the TSLS estimator is serving its purpose of removing the bias that the OLS estimation has due to the endogeneity of X, thus it is effective in estimating the population parameter β_1, and the entire IV regression procedure leading to it is well-suited.

Checking Instrument Validity: exogeneity I

If the instruments are not exogenous, the first stage regression does not "extract" the exogenous part of the X's, and the second stage regression on $\hat{X}$'s will suffer for endogeneity as these are still correlated with the error:

$\rightarrow$ If the instruments are not exogenous, then TSLS is inconsistent: The TSLS estimator converges in probability to something other than the causal coefficient

Checking Instrument Validity: exogeneity II

Case I: exact identification

There are as many instruments as endogenous regressors. Then it is impossible to develop a sta tistical test of the hypothesis that the instruments are, in fact, exogenous.

$\rightarrow$ Assessing whether the instruments are exogenous necessarily requires making an expert judgment based on personal knowledge of the application

Checking Instrument Validity: exogeneity III

Case II: over identification

Suppose you have a single endogenous regressor and two instruments:

- Then you could compute two different TSLS estimators: one using the first instrument and the other using the second. These two estimators will not be the same because of sampling variation, but if both instruments are exoge nous, then they will tend to be close to each other.
- If these two instruments produce very different estimates, one might sensibly conclude that there is some thing wrong with one or the other of the instruments or with both.
- In this latter case, it would be reasonable to conclude that one or the other or both of the instruments are not exogenous.

$\rightarrow$ This is the ideas of the **test of overidentifying restrictions**

Exogeneity of the instruments means that they are not correlated with u_i

Checking Instrument Validity: exogeneity IV

$\rightarrow$ the instruments should be approximately uncorrelated with the TSLS residuals $\hat{u}_i^{TSLS}$:

$$\begin{aligned}\hat{u}_i^{TSLS} &= Y - \hat{Y}_i^{TSLS} \\ &= Y_i - \left(\hat{\beta}_0^{TSLS} + \hat{\beta}_1^{TSLS} X_{1i} + \cdots + \hat{\beta}_k^{TSLS} X_{ki} + \hat{\beta}_{k+1}^{TSLS} W_{1i} + \cdots + \hat{\beta}_{k+r}^{TSLS} W_{ri}\right)\end{aligned}$$

If the instruments are truly exogenous, the coefficients of a regression of $\hat{u}_i^{TSLS}$ on the instruments and the exogenous variables W's should be all statistically zero:

$$\hat{u}_i^{TSLS} = \delta_0 + \delta_1 Z_{1i} + \cdots + \delta_m Z_{mi} + \delta_{m+1} W_{1i} + \cdots + \delta_{m+r} W_{ri} + e_i$$

Indeed, if the instruments are exogenous, they, along with the W's, should be not explain anything about $\hat{u}_i^{TSLS}$, and this should be highlighted in the above regression by obtaining statistically-zero coefficients.

$\rightarrow$ this can be tested

Checking Instrument Validity: exogeneity V

The Overidentifying Restrictions Test (The *J*-Statistic)

Let $\hat{u}_i^{TSLS}$ be the residuals from TSLS estimation of Equation (12.12). Use OLS to estimate the regression coefficients in

$$\hat{u}_i^{TSLS} = \delta_0 + \delta_1 Z_{1i} + \cdots + \delta_m Z_{mi} + \delta_{m+1} W_{1i} + \cdots + \delta_{m+r} W_{ri} + e_i, \quad (12.17)$$

where e_i is the regression error term. Let F denote the homoskedasticity-only F-statistic testing the hypothesis that $\delta_1 = \cdots = \delta_m = 0$. The overidentifying restrictions test statistic is $J = mF$. Under the null hypothesis that all the instruments are exogenous, if e_i is homoskedastic, in large samples J is distributed χ^2_{m-k}, where $m - k$ is the *degree of overidentification*—that is, the number of instruments minus the number of endogenous regressors.

Where Do Valid Instruments Come From? I

The most difficult aspect of IV estimation is finding instruments that are both relevant and exogenous. There are two approaches

- The first approach is to use economic theory to suggest instruments
- The second approach to constructing instruments is to look for some exogenous source of variation in X arising from what is, in effect, a random phenomenon that induces shifts (affect only) in the endogenous regressor.

TSLS with Control Variables I

So far we assumed the variables W are exogenous, now we consider the case where W is not exogenous, but instead control variables to make Z exogenous.

Basically,

- Z might be endogenous because it depends on some omitted factor in u.
- Then identifying and removing that omitted factor W in u and using it explicitly in the regression (controlling for W) makes Z *given* W truly exogenous **[PROOF]**.
- W is thus an endogenous term in the regression that however makes the instrument Z exogenous: it is analogous to a control variable in OLS regression.

Proof: IV with Control Variables I

Consider the IV regression model with a single X and a single W,

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + u_i. \tag{9}$$

In the IV Regression assumptions we had

$$\mathbb{E}(u_i|W_i) = 0,$$

now we replace it by

$$\mathbb{E}(u_i|W_i, Z_i) = \mathbb{E}(u_i|W_i), \tag{10}$$

stating that conditional on W_i the mean of u_i does not depends on Z_i. Assume that $\mathbb{E}(u_i|W_i)$ is linear in W_i,

$$\mathbb{E}(u_i|W_i) = \gamma_0 + \gamma_1 W_i. \tag{11}$$

Proof: IV with Control Variables II

Then,

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \beta_2 W_i + u_i - \mathbb{E}(u_i|W_i, Z_i) + \mathbb{E}(u_i|W_i, Z_i) \\ &= \beta_0 + \beta_1 X_i + \beta_2 W_i + u_i - \mathbb{E}(u_i|W_i, Z_i) + \gamma_0 + \gamma_1 W_i \\ &= \beta_0 + \beta_1 X_i + \beta_2 W_i + \varepsilon_i + \gamma_0 + \gamma_1 W_i \end{aligned}$$

where $\varepsilon_i = u_i - \mathbb{E}(u_i|W_i, Z_i)$.
It follows that:

$$\begin{aligned} Y_i &= (\beta_0 + \gamma_0) + \beta_1 X_i + (\beta_2 + \gamma_1) W_i + \varepsilon_i \\ &= \delta_0 + \beta_1 X_1 + \delta_i W_i + \varepsilon_i, \end{aligned} \tag{12}$$

Proof: IV with Control Variables III

with $\delta_0 = \beta_0 + \gamma_0$, $\delta_2 = \beta_2 + \gamma_1$. Note that:

$$\begin{aligned}\mathbb{E}(\varepsilon_i|W_i, Z_i) &= \mathbb{E}(u_i - \mathbb{E}(u_i|W_i, Z_i)|W_i, Z_i) \\ &= \mathbb{E}(u_i|W_i, Z_i) - \mathbb{E}(u_i|W_i, Z_i) \\ &= 0.\end{aligned}$$

Thus,

$$\mathbb{E}(\varepsilon_i|Z_i) = \mathbb{E}(\mathbb{E}(\varepsilon_i|W_i, Z_i)) = \mathbb{E}(0) = 0 \quad \rightarrow \quad \mathbb{C}\text{orr}(Z_i, \varepsilon_i) = 0,$$

so Z_i is exogenous w.r.t ε_i in eq.(12).
One can thus estimate eq.(12), with the usual IV regression assumptions but the exogeneity condition replaced by eq.(10).

With endogenous W the relevant regression for which Z is exogenous is eq.(12) and not eq.(9): what is being estimated are δ_0, β_1 and δ_1 and not β_0, β_1 and β_2.

Proof: IV with Control Variables IV

- One obtains the TSLS estimates of δ_0, β_1 and δ_1
- With a strong instrument, the TSLS estimator is consistent for δ_0, β_1 and δ_1
- $\hat{\beta}_1^{\text{TSLS}}$ does have the causal interpretation
- $\hat{\delta}_1^{\text{TSLS}}$ is not the causal effect of W_i on Y_i, as $\delta_1 = \beta_2 + \gamma_1$. Indeed, $\hat{\delta}_1^{\text{TSLS}}$ collects the direct causal effect of W_i on Y_i (eq.(9)) and that of the omitted factors in u_i on W_i eq.(11).