



Università di Pavia

2010

Instrumental Variables

Eduardo Rossi



Exogeneity Assumption: the explanatory variables which form the columns of \mathbf{X} are exogenous.

It implies that any randomness in the DGP that generated \mathbf{X} is *independent* of the error terms $\boldsymbol{\varepsilon}$ in the DGP for \mathbf{y} .

This independence implies that

$$E(\boldsymbol{\varepsilon}|\mathbf{X}) = \mathbf{0}$$

the mean of the entire vector $\boldsymbol{\varepsilon}$, that is, of every one of the ε_t , is zero conditional on the entire matrix \mathbf{X} .

This assumption is acceptable for cross-section data but not for time series data.



Time series data: each observation might correspond to a year, quarter, month, or day. Even if we are willing to assume that ε_t is in no way related to *current* and *past* values of the regressors, it must be related to *future* values if current values of the dependent variable affect future values of some of the regressors.

In the context of time-series data, the exogeneity assumption is a very strong one that we may often not feel comfortable in making.



The assumption:

$$E[\varepsilon_t | \mathbf{x}_t] = 0$$

is substantially weaker than Exogeneity assumption, because this rules out the possibility that the mean of ε_t may depend on the values of the regressors for any observation, while $E[\varepsilon_t | \mathbf{x}_t] = 0$ rules out the possibility that it may depend on their values for the current observation.

From the point of view of the error terms, it says that they are *innovations*.



PREDETERMINEDNESS

An innovation is a r.v. of which the mean is 0 conditional on the information in the \mathbf{x}_t , and so knowledge of the values taken by the latter is of no use in predicting the mean of the innovation.

From the point of view of the explanatory variables \mathbf{x}_t , Predeterminedness assumption says that they are predetermined with respect to the error terms.



ENDOGENEITY

There is a problem of endogeneity in the linear model

$$y_t = \mathbf{x}'_t \boldsymbol{\beta} + \varepsilon_t$$

if $\boldsymbol{\beta}$ is the parameter of interest and $E(\mathbf{x}_t \varepsilon_t) \neq 0$.



SIMULTANEOUS EQUATIONS

Demand-supply system:

$$q_{d,t} = \alpha_1 p_t + \alpha_2 x_t + \varepsilon_{d,t}$$

$$q_{s,t} = \beta_1 p_t + \varepsilon_{s,t}$$

$$q_{s,t} = q_{d,t} = q_t$$

x_t income exogenous. Prices and quantities are measured in logs and in deviation form their sample means.

Interest centers on estimating the demand elasticity:

$$E[\varepsilon_{d,t} | x_t] = 0$$

$$E[\varepsilon_{s,t} | x_t] = 0$$

$$E[\varepsilon_{d,t}^2 | x_t] = \sigma_s^2$$

$$E[\varepsilon_{s,t}^2 | x_t] = \sigma_d^2$$

$$E[\varepsilon_{s,t} x_t] = E[\varepsilon_{d,t} x_t] = 0$$

$$E[\varepsilon_{s,t} \varepsilon_{d,t} | x_t] = 0$$



SIMULTANEOUS EQUATIONS

Reduced form of the model:

$$p_t = \frac{\alpha_2 x_t}{\beta_1 - \alpha_1} + \frac{\varepsilon_{d,t} - \varepsilon_{s,t}}{\beta_1 - \alpha_1} = \pi_1 x_t + v_{1t}$$
$$q_t = \frac{\beta_1 \alpha_2 x_t}{\beta_1 - \alpha_1} + \frac{\beta_1 \varepsilon_{d,t} - \alpha_1 \varepsilon_{s,t}}{\beta_1 - \alpha_1} = \pi_2 x_t + v_{2t}$$
$$Cov[p_t, \varepsilon_{d,t}] = \frac{\sigma_d^2}{\beta_1 - \alpha_1}$$

Neither the demand nor the supply equation satisfies the assumptions of the classical regression model.

The price elasticity of demand (α_1) cannot be consistently estimated by OLS regression of q on x and p .

Because the endogenous variables are all correlated with the disturbances, the OLS estimators of the parameters of equations with endogenous variables on the RHS are inconsistent.



Measurement error in the regressors. Suppose that (y_t, \mathbf{x}_t^*) are joint random variables

$$E(y_t | \mathbf{x}_t) = \mathbf{x}_t^{*'} \boldsymbol{\beta}$$

is linear. But \mathbf{x}_t^* is not observed. Instead we observe

$$\mathbf{x}_t = \mathbf{x}_t^* + \mathbf{u}_t$$

$$\mathbf{u}_t \quad (K \times 1)$$

measurement error, independent of y_t and \mathbf{x}_t^* .



MEASUREMENT ERROR

$$\begin{aligned}y_t &= \mathbf{x}_t^{*'} \boldsymbol{\beta} + \varepsilon_t \\ &= (\mathbf{x}_t - \mathbf{u}_t)' \boldsymbol{\beta} + \varepsilon_t \\ &= \mathbf{x}_t' \boldsymbol{\beta} + v_t\end{aligned}$$

where

$$v_t = \varepsilon_t - \mathbf{u}_t' \boldsymbol{\beta}.$$

The problem is that

$$E(\mathbf{x}_t v_t) = E[(\mathbf{x}_t^* + \mathbf{u}_t)(\varepsilon_t - \mathbf{u}_t' \boldsymbol{\beta})] = -E(\mathbf{u}_t \mathbf{u}_t') \boldsymbol{\beta} \neq 0$$

If $\boldsymbol{\beta} \neq 0$ and $E(\mathbf{u}_t \mathbf{u}_t') \neq 0$. It follows that if $\hat{\boldsymbol{\beta}}$ is the OLS estimator

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \left[\sum_t (\mathbf{x}_t \mathbf{x}_t') \right]^{-1} \sum_t \mathbf{x}_t v_t$$



$$p \lim \hat{\beta} = \beta + p \lim \left\{ \left[\frac{1}{N} \sum_t (\mathbf{x}_t \mathbf{x}_t') \right]^{-1} \frac{1}{N} \sum_t \mathbf{x}_t v_t \right\}$$

then

$$p \lim \hat{\beta} = \beta - (E(\mathbf{x}_t \mathbf{x}_t'))^{-1} E(\mathbf{u}_t \mathbf{u}_t') \beta \neq \beta.$$



INSTRUMENTAL VARIABLES ESTIMATOR

Let the equation of interest be

$$y_t = \mathbf{x}'_t \boldsymbol{\beta} + \varepsilon_t \quad (1)$$

where \mathbf{x}_t ($K \times 1$), $E(\mathbf{x}_t \varepsilon_t) \neq 0$ so that there is a problem of endogeneity. Eq. (1) is called *structural equation*. In matrix notation, this can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

any solution to the problem of endogeneity requires additional information which we call *instruments*.



INSTRUMENTAL VARIABLES ESTIMATOR

Definition The vector \mathbf{z}_t , $(L \times 1)$, is an *instrumental variable* for eq.(1) if $E(\mathbf{z}_t \varepsilon_t) = 0$.

In a typical set-up, some regressors in \mathbf{x}_t will be uncorrelated with ε_t (for example, at least the intercept). Thus we make the partition

$$\mathbf{x}_t = \begin{bmatrix} \mathbf{x}_{1t} \\ \mathbf{x}_{2t} \end{bmatrix} \begin{matrix} K_1 \\ K_2 \end{matrix}$$

where $E(\mathbf{x}_{1t} \varepsilon_t) = 0$ yet $E(\mathbf{x}_{2t} \varepsilon_t) \neq 0$.

We call \mathbf{x}_{1t} *exogenous* and \mathbf{x}_{2t} *endogenous*. \mathbf{x}_{1t} is an instrument for eq.(1).



INSTRUMENTAL VARIABLES ESTIMATOR

So we have the partition

$$\mathbf{z}_t = \begin{bmatrix} \mathbf{x}_{1t} \\ \mathbf{z}_{2t} \end{bmatrix} \begin{matrix} K_1 \\ L_2 \end{matrix}$$

Three possible cases:

- Just-identified if $L = K$
- Over-identified if $L > K$
- Under-identified if $L < K$.



The reduced form relationship between the variables or "regressors" \mathbf{x}_t and the instruments \mathbf{z}_t is found by *linear projection*. Let

$$\mathbf{\Gamma} = E(\mathbf{z}_t \mathbf{z}_t')^{-1} E(\mathbf{z}_t \mathbf{x}_t')$$

be the $(L \times K)$ matrix of coefficients from a projection of \mathbf{x}_t on \mathbf{z}_t , and define

$$\mathbf{u}_t' = \mathbf{x}_t' - \mathbf{z}_t' \mathbf{\Gamma}$$

as the projection error. Then the reduced form linear relationship between \mathbf{x}_t and \mathbf{z}_t is

$$\mathbf{x}_t = \mathbf{\Gamma}' \mathbf{z}_t + \mathbf{u}_t \tag{2}$$



REDUCED FORM

In matrix notation

$$\mathbf{X} = \mathbf{Z}\mathbf{\Gamma} + \mathbf{u} \quad (3)$$

\mathbf{u} ($N \times K$). By construction

$$E(\mathbf{z}_t \mathbf{u}_t') = 0$$

So (3) is a projection and can be estimated by OLS

$$\begin{aligned} \mathbf{X} &= \mathbf{Z}\hat{\mathbf{\Gamma}} + \hat{\mathbf{u}} \\ \hat{\mathbf{\Gamma}} &= (\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}'\mathbf{X} \end{aligned}$$

Substituting $\mathbf{X} = \mathbf{Z}\mathbf{\Gamma} + \mathbf{u}$ in $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

$$\begin{aligned} \mathbf{y} &= (\mathbf{Z}\mathbf{\Gamma} + \mathbf{u})\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ &= \mathbf{Z}\boldsymbol{\lambda} + \mathbf{v} \end{aligned} \quad (4)$$

where $\boldsymbol{\lambda} = \mathbf{\Gamma}\boldsymbol{\beta}$, $\mathbf{v} = \mathbf{u}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.



REDUCED FORM

Observe that

$$E(\mathbf{z}_t v_t) = E(\mathbf{z}_t \mathbf{u}'_t) \boldsymbol{\beta} + E(\mathbf{z}_t \varepsilon_t) = 0$$

Thus (4) is a projection equation and may be estimated by OLS.

This is

$$\mathbf{y} = \mathbf{Z} \hat{\boldsymbol{\lambda}} + \hat{\mathbf{v}}$$

$$\hat{\boldsymbol{\lambda}} = (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{y}$$

The equation (4) is the reduced form for \mathbf{y} . The system

$$\mathbf{y} = \mathbf{Z} \boldsymbol{\lambda} + \mathbf{v}$$

$$\mathbf{X} = \mathbf{Z} \boldsymbol{\Gamma} + \mathbf{u}$$

OLS yields the reduced-form estimates $(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\Gamma}})$.



The structural parameter β relates to (λ, Γ) through $\lambda = \Gamma\beta$. The parameter is identified, meaning that it can be recovered from the reduced form, if

$$\text{rank}(\Gamma) = K \tag{5}$$

Assume that (5) holds.

If $L = K$, then $\beta = \Gamma^{-1}\lambda$.

If (5) is not satisfied, then β cannot be recovered from (λ, Γ) .

Note that a necessary (although not sufficient) condition for (5) is $L \geq K$.



IDENTIFICATION

Since \mathbf{X} and \mathbf{Z} have the common variables \mathbf{X}_1 , we can rewrite some of the expressions

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_2 \end{bmatrix}$$
$$\mathbf{Z} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{Z}_2 \end{bmatrix}$$

we can partition $\mathbf{\Gamma}$ as

$$\mathbf{\Gamma} = \begin{bmatrix} \mathbf{\Gamma}_{11} & \mathbf{\Gamma}_{12} \\ \mathbf{\Gamma}_{21} & \mathbf{\Gamma}_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{I}_{K_1} & \mathbf{\Gamma}_{12} \\ 0 & \mathbf{\Gamma}_{22} \end{bmatrix} \begin{matrix} L_1 \times K_2 \\ L_2 \times K_2 \end{matrix}$$



$$\begin{aligned}\mathbf{Z}\mathbf{\Gamma} &= \begin{bmatrix} \mathbf{X}_1 & \mathbf{Z}_2 \end{bmatrix} \begin{bmatrix} \mathbf{I}_{K_1} & \mathbf{\Gamma}_{12} \\ 0 & \mathbf{\Gamma}_{22} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_1\mathbf{\Gamma}_{12} + \mathbf{Z}_2\mathbf{\Gamma}_{22} \end{bmatrix}\end{aligned}$$

β is identified if $rank(\mathbf{\Gamma}) = K$, which is true if and only if $rank(\mathbf{\Gamma}_{22}) = K_2$ (by the upper-diagonal structure of $\mathbf{\Gamma}$). The key to identification of the model rests on the $(L_2 \times K_2)$ matrix $\mathbf{\Gamma}_{22}$.



INSTRUMENTAL VARIABLE ESTIMATION

Suppose that the model is just identified ($L = K$). Then $\beta = \Gamma^{-1}\lambda$.

This suggests the *Indirect Least Squares* estimator:

$$\begin{aligned}\hat{\beta}_{IV} &= \hat{\Gamma}^{-1}\hat{\lambda} \\ &= \left[(\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}'\mathbf{X} \right]^{-1} \left[(\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}'\mathbf{y} \right] \\ &= (\mathbf{Z}'\mathbf{X})^{-1} (\mathbf{Z}'\mathbf{Z}) (\mathbf{Z}'\mathbf{Z})^{-1} (\mathbf{Z}'\mathbf{y}) \\ &= (\mathbf{Z}'\mathbf{X})^{-1} (\mathbf{Z}'\mathbf{y})\end{aligned}$$

$\hat{\beta}_{IV}$ is *the instrumental variables estimator* of β .

For identification by any given sample, the condition is just that $\mathbf{Z}'\mathbf{X}$ should be nonsingular.



CONSISTENCY OF IV ESTIMATOR

Since $(\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\Gamma}}) \xrightarrow{p} (\boldsymbol{\lambda}, \boldsymbol{\Gamma})$ and $\boldsymbol{\Gamma}$ is invertible, $\boldsymbol{\beta}$ is consistent. A more direct way to see consistency is to substitute the equation $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ into the equation for $\widehat{\boldsymbol{\beta}}$ to obtain

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{IV} &= (\mathbf{Z}'\mathbf{X})^{-1} (\mathbf{Z}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})) \\ &= \boldsymbol{\beta} + (\mathbf{Z}'\mathbf{X})^{-1} \mathbf{Z}'\boldsymbol{\varepsilon}\end{aligned}$$

Given:

$$\mathbf{Q}_{ZX} = p \lim \left(\frac{\mathbf{Z}'\mathbf{X}}{N} \right) \text{ deterministic with rank } K \text{ (Asymptotic identification)}$$

the IV estimator is consistent iff:

$$(N^{-1}\mathbf{Z}'\mathbf{X})^{-1} (N^{-1}\mathbf{Z}'\boldsymbol{\varepsilon}) \xrightarrow{p} \mathbf{0}$$

or

$$p \lim_{N \rightarrow \infty} \mathbf{Z}'\boldsymbol{\varepsilon} = \mathbf{0} \text{ asymptotically uncorrelated.}$$



THE ASYMPTOTIC DISTRIBUTION OF IV ESTIMATOR

$$\mathbf{Q}_{ZX} = p \lim \left(\frac{\mathbf{Z}'\mathbf{X}}{N} \right) \text{ deterministic with rank } K \text{ (Asymptotic identification)}$$

$$\mathbf{Q}_{ZZ} = p \lim \left(\frac{\mathbf{Z}'\mathbf{Z}}{N} \right) \text{ positive definite}$$

$$\sqrt{N} \left(\hat{\boldsymbol{\beta}}_{IV} - \boldsymbol{\beta} \right) = \sqrt{N} \left[(\mathbf{Z}'\mathbf{X})^{-1} \mathbf{Z}'\boldsymbol{\varepsilon} \right]$$

Under the hypothesis that $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' | \mathbf{X}, \mathbf{Z}) = \sigma^2 \mathbf{I}_n$, The Central Limit Theorem gives us the limit distribution

$$\sqrt{N} \left(\hat{\boldsymbol{\beta}}_{IV} - \boldsymbol{\beta} \right) \xrightarrow{L} N \left(0, \sigma^2 \mathbf{Q}_{ZX}^{-1} \mathbf{Q}_{ZZ} \mathbf{Q}_{XZ}^{-1} \right)$$

The estimate of σ^2

$$\hat{\sigma}^2 = \frac{1}{N} \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} \right)' \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} \right)$$



If the model is correctly specified the asymptotic covariance matrix:

$$\begin{aligned} \text{Var} \left[p \lim_{N \rightarrow \infty} \sqrt{N} (\hat{\beta}_{IV} - \beta) \right] &= \sigma^2 \mathbf{Q}_{ZX}^{-1} \mathbf{Q}_{ZZ} \mathbf{Q}_{XZ}^{-1} \\ &= \sigma^2 \left[p \lim \left(\frac{\mathbf{Z}'\mathbf{X}}{N} \right) \right]^{-1} \left[p \lim \left(\frac{\mathbf{Z}'\mathbf{Z}}{N} \right) \right] \left[p \lim \left(\frac{\mathbf{X}'\mathbf{Z}}{N} \right) \right]^{-1} \\ &= \sigma^2 \left\{ \left[p \lim \left(\frac{\mathbf{X}'\mathbf{Z}}{N} \right) \right] \left[p \lim \left(\frac{\mathbf{Z}'\mathbf{Z}}{N} \right) \right]^{-1} \left[p \lim \left(\frac{\mathbf{Z}'\mathbf{X}}{N} \right) \right] \right\}^{-1} \\ &= \sigma^2 p \lim \left\{ \left(\frac{\mathbf{X}'\mathbf{Z}}{N} \right) \left(\frac{\mathbf{Z}'\mathbf{Z}}{N} \right)^{-1} \left(\frac{\mathbf{Z}'\mathbf{X}}{N} \right) \right\}^{-1} \\ &= \sigma^2 p \lim \left(\frac{1}{N} \mathbf{X}' \mathbf{P}_Z \mathbf{X} \right)^{-1} \end{aligned}$$

If we have some choice over what instruments to use in the matrix \mathbf{Z} , it makes sense to choose them so as to minimize the above asymptotic covariance matrix.



Optimal instruments: instruments that minimize the asymptotic covariance matrix.

In order to determine the optimal instruments, we must know the data generating process.

Quite generally, we can suppose that the explanatory variables satisfy the relation:

$$\mathbf{X} = \bar{\mathbf{X}} + \mathbf{V}$$

where

$$\bar{\mathbf{x}}_t = E[\mathbf{x}_t | \Omega_t] \quad E[\mathbf{v}_t | \Omega_t] = 0$$

where Ω_t is the set of uncorrelated variables with ε_t .

It turns out that the matrix $\bar{\mathbf{X}}$ provides the optimal instruments.



HAUSMAN SPECIFICATION TEST

It might not be obvious that the regressors are correlated with the disturbances or that the regressors are measured with error.

If not, there would be some benefit to using the OLS rather than the IV estimator.

We can compare the asymptotic variance-covariance matrices under the assumption that both are consistent, that is

$$p \lim \frac{1}{N} \mathbf{X}' \boldsymbol{\epsilon} = \mathbf{0}$$



HAUSMAN SPECIFICATION TEST

The difference between the asymptotic covariance matrices of the two estimators is

$$\begin{aligned} AVar(\hat{\beta}_{IV}) - AVar(\hat{\beta}_{OLS}) &= \frac{\sigma^2}{N} p \lim \left(\frac{\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}}{N} \right)^{-1} \\ &\quad - \frac{\sigma^2}{N} p \lim \left(\frac{\mathbf{X}'\mathbf{X}}{N} \right)^{-1} \\ &= \frac{\sigma^2}{N} p \lim N \left[(\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X})^{-1} - (\mathbf{X}'\mathbf{X})^{-1} \right] \end{aligned}$$



HAUSMAN SPECIFICATION TEST

Compare the inverses

$$(\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}) = \mathbf{X}'\mathbf{P}_Z\mathbf{X} = \mathbf{X}'(\mathbf{I}_N - \mathbf{M}_Z)\mathbf{X} = \mathbf{X}'\mathbf{X} - \mathbf{X}'\mathbf{M}_Z\mathbf{X}$$

since \mathbf{M}_Z is p.s.d. it follows that $\mathbf{X}'\mathbf{M}_Z\mathbf{X}$ is also so. Since $(\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X})$ is smaller, in the matrix sense, than $\mathbf{X}'\mathbf{X}$ then its inverse is larger. If the OLS is consistent is the preferred estimator (asymptotically more efficient).



HAUSMAN SPECIFICATION TEST

We want to test the null hypothesis that the error terms are uncorrelated with all the regressors against the alternative that they are correlated with some of the regressors, although not with the instruments.

The null hypothesis is

$$p \lim \frac{1}{N} \mathbf{X}' \boldsymbol{\epsilon} = \mathbf{0}$$

under the null hypothesis both estimators are consistent.

Under the alternative only $\hat{\boldsymbol{\beta}}_{IV}$ is consistent.

The suggestion is to examine

$$\mathbf{d} = \hat{\boldsymbol{\beta}}_{IV} - \hat{\boldsymbol{\beta}}_{OLS}$$



HAUSMAN SPECIFICATION TEST

Under the null hypothesis

$$p \lim \mathbf{d} = \mathbf{0}$$

We might test this hypothesis using a Wald statistic

$$H = \mathbf{d}' [AVar(\hat{\boldsymbol{\beta}}_{IV} - \hat{\boldsymbol{\beta}}_{OLS})]^{-1} \mathbf{d}$$



HAUSMAN SPECIFICATION TEST

$$AVar(\hat{\beta}_{IV} - \hat{\beta}_{OLS}) = AVar(\hat{\beta}_{IV}) + AVar(\hat{\beta}_{OLS}) - 2ACov(\hat{\beta}_{IV}, \hat{\beta}_{OLS})$$

We do not have an expression for the covariance term. Hausman shows that

$$Cov[\hat{\beta}_{OLS}, \hat{\beta}_{OLS} - \hat{\beta}_{IV}] = \mathbf{0}$$

$$\begin{aligned} Cov(X, X - Y) &= E[(X - E(X))(X - Y - E(X) + E(Y))] \\ &= E[(X - E(X))^2] - E[(X - E(X))(Y - E(Y))] \\ &= Var[X] - Cov(X, Y) \end{aligned}$$

$$Cov[\hat{\beta}_{OLS}, \hat{\beta}_{OLS} - \hat{\beta}_{IV}] = Var[\hat{\beta}_{OLS}] - Cov[\hat{\beta}_{IV}, \hat{\beta}_{OLS}] = \mathbf{0}$$

then

$$Var[\hat{\beta}_{OLS}] = Cov[\hat{\beta}_{IV}, \hat{\beta}_{OLS}]$$



HAUSMAN SPECIFICATION TEST

$$AVar(\hat{\beta}_{IV} - \hat{\beta}_{OLS}) = AVar(\hat{\beta}_{IV}) - AVar(\hat{\beta}_{OLS})$$

Inserting this useful result into the Wald statistics and reverting to empirical estimates

$$H = (\hat{\beta}_{IV} - \hat{\beta}_{OLS})' [AVar(\hat{\beta}_{IV}) - AVar(\hat{\beta}_{OLS})]^{-1} (\hat{\beta}_{IV} - \hat{\beta}_{OLS})$$

Under the null hypothesis we are using two consistent estimators of σ^2 . If we use s^2 as the common estimator, then the test statistic

$$H = \frac{(\hat{\beta}_{IV} - \hat{\beta}_{OLS})' [(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1} - (\mathbf{X}'\mathbf{X})^{-1}]^{-1} (\hat{\beta}_{IV} - \hat{\beta}_{OLS})}{s^2}$$

where

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



HAUSMAN SPECIFICATION TEST

Unless \mathbf{X} and \mathbf{Z} have no variables in common, the rank of the matrix in the statistic is less than K , and the ordinary inverse will not even exist.

In most cases some of the variables in \mathbf{X} appear in \mathbf{Z} (at least the constant term!).

Let

$$\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$$

where \mathbf{X}_2 are the suspected endogenous variables.

\mathbf{X}_1 is included in \mathbf{Z} .

A subset of K_1 variables is uncorrelated with the disturbances. The quadratic form in the statistic is used to test only $K_2 = K - K_1$ hypotheses. In this case H is quadratic form of rank K_2 .



HAUSMAN SPECIFICATION TEST

Since $\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$ is an idempotent matrix

$$\widehat{\mathbf{X}} = \mathbf{P}_Z \mathbf{X}$$

$$\widehat{\mathbf{X}}'\widehat{\mathbf{X}} = \mathbf{X}'\mathbf{P}_Z\mathbf{P}_Z\mathbf{X} = \mathbf{X}'\mathbf{P}_Z\mathbf{X} = \widehat{\mathbf{X}}'\mathbf{X}.$$

Using this result and expanding \mathbf{d}

$$\begin{aligned}\mathbf{d} &= (\widehat{\mathbf{X}}'\widehat{\mathbf{X}})^{-1}\widehat{\mathbf{X}}'\mathbf{y} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= (\widehat{\mathbf{X}}'\widehat{\mathbf{X}})^{-1}[\widehat{\mathbf{X}}'\mathbf{y} - (\widehat{\mathbf{X}}'\widehat{\mathbf{X}})(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}] \\ &= (\widehat{\mathbf{X}}'\widehat{\mathbf{X}})^{-1}\widehat{\mathbf{X}}'[\mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}] \\ &= (\widehat{\mathbf{X}}'\widehat{\mathbf{X}})^{-1}\widehat{\mathbf{X}}'\widehat{\boldsymbol{\epsilon}}\end{aligned}$$

K_1 columns of \mathbf{X} are in $\widehat{\mathbf{X}}$. Suppose that these are the first K_1 columns. Thus the first K_1 rows of $\widehat{\mathbf{X}}'\widehat{\boldsymbol{\epsilon}}$ are the same as the first K_1 rows of $\mathbf{X}'\boldsymbol{\epsilon}$ which are 0.



HAUSMAN SPECIFICATION TEST

Thus we can write

$$\mathbf{d} = (\widehat{\mathbf{X}}'\widehat{\mathbf{X}})^{-1} \begin{bmatrix} \mathbf{0} \\ \widehat{\mathbf{X}}_2'\boldsymbol{\epsilon} \end{bmatrix} = (\widehat{\mathbf{X}}'\widehat{\mathbf{X}})^{-1} \begin{bmatrix} \mathbf{0} \\ \mathbf{q}^* \end{bmatrix}$$

the test statistic becomes

$$\begin{aligned} H &= \frac{\mathbf{d}'[(\widehat{\mathbf{X}}'\widehat{\mathbf{X}})^{-1} - (\mathbf{X}'\mathbf{X})^{-1}]^{-1}\mathbf{d}}{s^2} \\ &= \begin{bmatrix} \mathbf{0}' & \mathbf{q}^{*'} \end{bmatrix} (\widehat{\mathbf{X}}'\widehat{\mathbf{X}})^{-1} \mathbf{W} (\widehat{\mathbf{X}}'\widehat{\mathbf{X}})^{-1} \begin{bmatrix} \mathbf{0} \\ \mathbf{q}^* \end{bmatrix} \end{aligned}$$

where

$$\mathbf{W} = s^{-2}[(\widehat{\mathbf{X}}'\widehat{\mathbf{X}})^{-1} - (\mathbf{X}'\mathbf{X})^{-1}]^{-1}$$



HAUSMAN SPECIFICATION TEST

An $(n \times m)$ matrix, denoted by

$$\mathbf{A}^{-}$$

is the *generalized inverse* of the $(m \times n)$ \mathbf{A} matrix if it satisfies

$$\mathbf{A}\mathbf{A}^{-}\mathbf{A} = \mathbf{A}$$

1. \mathbf{A}^{-} is not unique in general.
2. $\mathbf{A}\mathbf{A}^{-}$ and $\mathbf{A}^{-}\mathbf{A}$ are idempotent.
3. $\mathbf{I}_m - \mathbf{A}\mathbf{A}^{-}$ and $\mathbf{I}_n - \mathbf{A}^{-}\mathbf{A}$
4. $\text{rank}(\mathbf{A}) = n \Leftrightarrow \mathbf{A}^{-}\mathbf{A} = \mathbf{I}_n$
5. $\text{rank}(\mathbf{A}) = m \Leftrightarrow \mathbf{A}\mathbf{A}^{-} = \mathbf{I}_m$



HAUSMAN SPECIFICATION TEST

The Moore-Penrose Inverse. The $(n \times m)$ matrix \mathbf{A}^+ is the Moore-Penrose (generalized) inverse of the $(m \times n)$ \mathbf{A} if it satisfies the following four conditions:

1. $\mathbf{A}\mathbf{A}^+\mathbf{A} = \mathbf{A}$,

2. $\mathbf{A}^+\mathbf{A}\mathbf{A}^+ = \mathbf{A}^+$

3. $(\mathbf{A}\mathbf{A}^+)^H = \mathbf{A}\mathbf{A}^+$

4. $(\mathbf{A}^+\mathbf{A})^H = \mathbf{A}^+\mathbf{A}$



Properties

1. \mathbf{A} ($m \times n$): \mathbf{A}^+ exists and is unique.
2. \mathbf{A} ($m \times n$):
 - (a) $\text{rank}(\mathbf{A}) = m \Leftrightarrow \mathbf{A}\mathbf{A}^+ = \mathbf{I}_m$
 - (b) $\text{rank}(\mathbf{A}) = m \Rightarrow \mathbf{A}^+ = \mathbf{A}^H(\mathbf{A}\mathbf{A}^H)^{-1}$
 - (c) $\text{rank}(\mathbf{A}) = n \Leftrightarrow \mathbf{A}^+\mathbf{A} = \mathbf{I}_n$
 - (d) $\text{rank}(\mathbf{A}) = n \Rightarrow \mathbf{A}^+ = (\mathbf{A}^H\mathbf{A})^{-1}\mathbf{A}^H$
3. \mathbf{A} nonsingular ($m \times m$) $\Rightarrow \mathbf{A}^+ = \mathbf{A}^{-1}$.



HAUSMAN SPECIFICATION TEST

Denoting

$$\mathbf{P} = (\widehat{\mathbf{X}}' \widehat{\mathbf{X}})^{-1} \mathbf{W} (\widehat{\mathbf{X}}' \widehat{\mathbf{X}})^{-1}$$

$$H = \begin{bmatrix} \mathbf{0}' & \mathbf{q}^{*'} \end{bmatrix} \mathbf{P} \begin{bmatrix} \mathbf{0} \\ \mathbf{q}^* \end{bmatrix} = \mathbf{q}^{*'} \mathbf{P}_{22} \mathbf{q}^*$$

\mathbf{P}_{22} is the lower $K_2 \times K_2$ submatrix of \mathbf{P} .



HAUSMAN SPECIFICATION TEST

An alternative approach, provided by Durbin (1954) and Wu (1973), circumvents the computation of the generalized inverse.

As before, we want to test that

$$p \lim \mathbf{d} = \mathbf{0}.$$

The idea of the DWH test is to check whether the difference $\hat{\beta}_{IV} - \hat{\beta}_{OLS}$ is significantly different from zero in the available sample.

By construction the OLS residuals are orthogonal to the columns of \mathbf{X} , in particular those in \mathbf{X}_1 .



HAUSMAN SPECIFICATION TEST

We know that

$$\mathbf{d} = (\widehat{\mathbf{X}}'\widehat{\mathbf{X}})^{-1}\widehat{\mathbf{X}}'\widehat{\boldsymbol{\epsilon}}$$

testing whether $\widehat{\boldsymbol{\beta}}_{IV} - \widehat{\boldsymbol{\beta}}_{OLS}$ is significantly different from zero is equivalent to testing whether the vector $\widehat{\mathbf{X}}'\widehat{\boldsymbol{\epsilon}}$ is significantly different from zero.

$$\widehat{\mathbf{X}}'\widehat{\boldsymbol{\epsilon}} = \mathbf{X}'\mathbf{P}_Z\widehat{\boldsymbol{\epsilon}}$$

since $\mathbf{X}_1 \in \mathbf{Z}$

$$\mathbf{X}'_1\widehat{\boldsymbol{\epsilon}} = \mathbf{X}'_1\mathbf{M}_X\mathbf{y} = \mathbf{0}$$

The test is thus concerned only with the K_2 elements of $\mathbf{X}'_2\mathbf{P}_Z\widehat{\boldsymbol{\epsilon}}$ (or $\mathbf{X}'_2\mathbf{P}_Z\mathbf{M}_X\mathbf{y}$) which will not in general be identically zero, but should not differ from it significantly under H_0 .



HAUSMAN SPECIFICATION TEST

An F test statistic can be computed to test the joint significance of the elements of γ in the augmented regression

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \widehat{\mathbf{X}}_2\boldsymbol{\gamma} + \boldsymbol{\epsilon}$$

This is equivalent to the Hausman test for this model.

The OLS estimates of $\boldsymbol{\delta}$ are, by the FWL Theorem, the same as those from the FWL regression of $\mathbf{M}_X\mathbf{y}$ on $\mathbf{M}_X\mathbf{P}_Z\mathbf{X}_2$

$$\widehat{\boldsymbol{\delta}} = (\mathbf{X}'_2\mathbf{P}_Z\mathbf{M}_X\mathbf{P}_Z\mathbf{X}_2)^{-1}\mathbf{X}'_2\mathbf{P}_Z\mathbf{M}_X\mathbf{y}$$

Since the inverted matrix is positive definite, we see that testing whether $\boldsymbol{\delta} = \mathbf{0}$ is equivalent to testing whether $\mathbf{X}'_2\mathbf{P}_Z\mathbf{M}_X\mathbf{y} = \mathbf{0}$ as desired.