Instrumental Variables

When any of the regressors is correlated with the disturbances, OLS estimates are biased and inconsistent.

Most common solution is **instrumental variables**, or **IV**, estimation.

Suppose there is just one column of $\boldsymbol{X}$, say $\boldsymbol{x}$, that is correlated with $\boldsymbol{u}$.

Let $\boldsymbol{W}$ denote a matrix of instruments with the same dimension as $\boldsymbol{X}$, where $\boldsymbol{x}$ has been replaced by an instrument $\boldsymbol{w}$ that is assumed to be uncorrelated with the disturbances but correlated with $\boldsymbol{x}$.

The remaining columns of $\boldsymbol{W}$ and $\boldsymbol{X}$ are identical.

The **simple IV estimator** is

$$\hat{\boldsymbol{\beta}}_{\text{IV}} = (\boldsymbol{W}^\top \boldsymbol{X})^{-1} \boldsymbol{W}^\top \boldsymbol{y}. \tag{1}$$

This is consistent whenever $\operatorname{plim} N^{-1} \boldsymbol{W}^\top \boldsymbol{u} = \boldsymbol{0}$.

Errors in Variables

Correlations between regressors and disturbances arise for two main reasons. The first is that regressors may be measured with error.

Consider the simple linear regression model

$$y_i^\circ = \beta_1 + \beta_2 x_i^\circ + u_i^\circ, \quad u_i^\circ \sim \text{IID}(0, \sigma^2), \tag{2}$$

where the variables x_i° and y_i° are not actually observed. These are sometimes called **latent variables**. Instead, we observe

$$x_i = x_i^\circ + v_{1i}, \quad \text{and} \quad y_i = y_i^\circ + v_{2i}. \tag{3}$$

Here v_{1i} and v_{2i} are measurement errors which are assumed (not always realistically) to be IID with variances ω_1^2 and ω_2^2, respectively, and to be independent of x_i°, y_i°, and u_i°.

From (3), we see that $x_i^\circ = x_i - v_{1i}$ and $y_i^\circ = y_i - v_{2i}$.

In terms of observables, (2) is replaced by

$$y_i = \beta_1 + \beta_2(x_i - v_{1i}) + u_i^\circ + v_{2i} \tag{4}$$
$$= \beta_1 + \beta_2 x_i + u_i^\circ + v_{2i} - \beta_2 v_{1i} \tag{5}$$
$$= \beta_1 + \beta_2 x_i + u_i, \tag{6}$$

where $u_i = u_i^\circ + v_{2i} - \beta_2 v_{1t}$. Thus

$$\mathrm{Var}(u_i) = \sigma^2 + \omega_2^2 + \beta_2^2 \omega_1^2 > \sigma^2. \tag{7}$$

Because $x_i = x_i^\circ + v_{1i}$, and u_i depends on v_{1i}, u_i must be correlated with x_i whenever $\beta_2 \neq 0$ and $\omega_1^2 > 0$.

Since $\mathrm{E}(u_i \,|\, x_i) = \mathrm{E}(u_i \,|\, v_{1i}) = -\beta_2 v_{1i}$,

$$\mathrm{Cov}(x_i, u_i) = \mathrm{E}(x_i u_i) = \mathrm{E}\big(x_i \mathrm{E}(u_i \,|\, x_i)\big) \tag{8}$$
$$= -\mathrm{E}\big((x_i^\circ + v_{1i})\beta_2 v_{1i}\big) = -\beta_2 \omega_1^2. \tag{9}$$

The correlation between x_i and u_i has sign opposite to that of β_2.

Because of the negative correlation between x_i and u_i, the OLS estimate $\hat{\beta}_2$ is biased towards zero.

$\hat{\beta}_1$ is also biased, as are any other coefficents.

- Friedman's **permanent income hypothesis** is essentially a measurement error story. Consumption depends on permanent income, but we only observe current income.
- Regressing consumption on current income leads to an estimated **marginal propensity to consume** that is much too small.

If we knew how large the measurement errors were, we could remove the bias in $\hat{\beta}_2$.

Of course, this will require the assumption that they are uncorrelated with the x_i°, which we made in (3).

We can also obtain consistent estimates if we can find an instrument w_i that is correlated with x_i° but uncorrelated with v_{1i}.

Simultaneous Equations

Suppose that q_i is quantity and p_i is price, both of which might be in logarithms. A linear (or loglinear) model of demand and supply is

$$q_i = \gamma_d p_i + \boldsymbol{X}_i^d \boldsymbol{\beta}_d + u_i^d \tag{10}$$
$$q_i = \gamma_s p_i + \boldsymbol{X}_i^s \boldsymbol{\beta}_s + u_i^s, \tag{11}$$

where (10) is the demand function and (11) is the supply function. These are the two **structural equations** of the system.

Here $\boldsymbol{X}_i^d$ and $\boldsymbol{X}_i^s$ are row vectors of observations on exogenous or predetermined variables.

Equations (10) and (11) are two linear equations that jointly determine p_i and q_i. They constitute a **linear simultaneous equations model**.

As written, quantity depends on price in both equations. But either or both equations could be rewritten so that price depends on quantity.

We can rewrite the two equations in matrix notation as

$$\begin{bmatrix} 1 & -\gamma_d \\ 1 & -\gamma_s \end{bmatrix} \begin{bmatrix} q_i \\ p_i \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_i^d \boldsymbol{\beta}_d \\ \boldsymbol{X}_i^s \boldsymbol{\beta}_s \end{bmatrix} + \begin{bmatrix} u_i^d \\ u_i^s \end{bmatrix}. \tag{12}$$

Provided $\gamma_d \neq \gamma_s$, a solution exists. It is

$$\begin{bmatrix} q_i \\ p_i \end{bmatrix} = \begin{bmatrix} 1 & -\gamma_d \\ 1 & -\gamma_s \end{bmatrix}^{-1} \left(\begin{bmatrix} \boldsymbol{X}_i^d \boldsymbol{\beta}_d \\ \boldsymbol{X}_i^s \boldsymbol{\beta}_s \end{bmatrix} + \begin{bmatrix} u_i^d \\ u_i^s \end{bmatrix} \right). \tag{13}$$

This is the **restricted reduced form** of the system.

Here p_i and q_i depend on both u_i^d and u_i^s, and on every exogenous or predetermined variable in $\boldsymbol{X}_i^d$ and/or $\boldsymbol{X}_i^s$.

Therefore p_i, which appears on the r.h.s. of both (10) and (11), must be correlated with the disturbances in both of those equations.

This is also true of q_i. Rewriting one or both equations will not eliminate the problem.

Instrumental Variables (IV) Estimators

We will focus on the linear regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} = \beta \boldsymbol{y}_2 + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{u}, \quad \mathrm{E}(\boldsymbol{u}\boldsymbol{u}^\top) = \sigma^2 \mathbf{I}, \tag{14}$$

where $\boldsymbol{X}$ is $N \times k$, and $\boldsymbol{Z}$ is $N \times (k-1)$. Here every column of $\boldsymbol{Z}$ is exogenous or predetermined, but $\boldsymbol{y}_2$ is not predetermined.

Let $\boldsymbol{W}$ be an $N \times k$ matrix consisting of $\boldsymbol{Z}$ plus one other column, say $\boldsymbol{w}_2$, that is predetermined and therefore a **valid instrument**.

Since $\mathrm{E}(\boldsymbol{W}^\top \boldsymbol{u}) = \mathbf{0}$, we can employ the moment conditions

$$\boldsymbol{W}^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \mathbf{0}. \tag{15}$$

These lead to the **simple IV estimator**

$$\hat{\boldsymbol{\beta}}_{\mathrm{IV}} = (\boldsymbol{W}^\top \boldsymbol{X})^{-1} \boldsymbol{W}^\top \boldsymbol{y}. \tag{16}$$

The simple IV estimator is consistent whenever

$$\operatorname*{plim}_{N\to\infty} \frac{1}{N} \boldsymbol{W}^\top \boldsymbol{X} = \boldsymbol{S}_{\boldsymbol{W}^\top \boldsymbol{X}} \quad \text{and} \quad \operatorname*{plim}_{N\to\infty} \frac{1}{N} \boldsymbol{W}^\top \boldsymbol{u} = \boldsymbol{0}, \tag{17}$$

where $\boldsymbol{S}_{\boldsymbol{W}^\top \boldsymbol{X}}$ is deterministic and nonsingular. Just replace $\boldsymbol{y}$ in (16) by $\boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{u}$, divide both factors by N, and take plims.

The asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}_{\text{IV}}$ is

$$\operatorname{Var}\left(N^{1/2}(\hat{\boldsymbol{\beta}}_{\text{IV}} - \boldsymbol{\beta}_0)\right) = \sigma^2 \left(\boldsymbol{S}_{\boldsymbol{W}^\top \boldsymbol{X}}\right)^{-1} \boldsymbol{S}_{\boldsymbol{W}^\top \boldsymbol{W}} \left(\boldsymbol{S}_{\boldsymbol{W}^\top \boldsymbol{X}}\right)^{-1}, \tag{18}$$

where $\boldsymbol{S}_{\boldsymbol{W}^\top \boldsymbol{W}} = \operatorname{plim}_{N\to\infty} N^{-1} \boldsymbol{W}^\top \boldsymbol{W}$. In practice, we use

$$\widehat{\operatorname{Var}}(\hat{\boldsymbol{\beta}}_{\text{IV}}) = \hat{\sigma}^2 (\boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X})^{-1}, \tag{19}$$

where

$$\hat{\sigma}^2 = \frac{1}{N} (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{IV}})^\top (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{IV}}). \tag{20}$$

Of course, if we assumed that the u_i are heteroskedastic rather than i.i.d., (19) would be replaced by

$$\widehat{\mathrm{Var}}_{\mathrm{h}}(\hat{\boldsymbol{\beta}}_{\mathrm{IV}}) = (\boldsymbol{W}^\top \boldsymbol{X})^{-1} \left(\sum_{i=1}^{N} \hat{u}_i^2 \boldsymbol{W}_i^\top \boldsymbol{W}_i \right) (\boldsymbol{W}^\top \boldsymbol{X})^{-1}. \tag{21}$$

We could also allow for clustering. Since the i.i.d. assumption leads to some interesting results on efficiency, we will maintain it for now.

The simple IV estimator $\hat{\boldsymbol{\beta}}_{\mathrm{IV}}$ can also be written as $(\boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{y}$.

Because $\boldsymbol{X}^\top \boldsymbol{W}$ is square, $(\boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X})^{-1} = (\boldsymbol{W}^\top \boldsymbol{X})^{-1} \boldsymbol{W}^\top \boldsymbol{W} (\boldsymbol{X}^\top \boldsymbol{W})^{-1}$. Thus

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{\mathrm{IV}} &= (\boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{y} && (22)\\
&= (\boldsymbol{W}^\top \boldsymbol{X})^{-1} \boldsymbol{W}^\top \boldsymbol{W} (\boldsymbol{X}^\top \boldsymbol{W})^{-1} \boldsymbol{X}^\top \boldsymbol{W} (\boldsymbol{W}^\top \boldsymbol{W})^{-1} \boldsymbol{W}^\top \boldsymbol{y} && (23)\\
&= (\boldsymbol{W}^\top \boldsymbol{X})^{-1} \boldsymbol{W}^\top \boldsymbol{y} && (24)
\end{aligned}$$

The first expression is also the **generalized IV estimator**, or **GIVE**.

In many cases, we have $l > k$ instruments. The GIVE finds a linear combination of them that is optimal.

Suppose that

$$\boldsymbol{X} = \bar{\boldsymbol{X}} + \boldsymbol{V}, \quad \text{where} \quad \bar{\boldsymbol{X}}_i = \mathrm{E}(\boldsymbol{X}_i \,|\, \Omega_i). \tag{25}$$

Then $\boldsymbol{W}$ should, ideally, be proportional to $\bar{\boldsymbol{X}}$. In this case

$$\operatorname*{plim}_{N\to\infty} \frac{1}{N} \boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X} = \operatorname*{plim}_{N\to\infty} \frac{1}{N} \bar{\boldsymbol{X}}^\top \boldsymbol{P}_{\boldsymbol{W}} \bar{\boldsymbol{X}} = \operatorname*{plim}_{N\to\infty} \frac{1}{N} \bar{\boldsymbol{X}}^\top \bar{\boldsymbol{X}}. \tag{26}$$

Of course, if $\mathrm{Var}(\boldsymbol{V})$ is large, $\bar{\boldsymbol{X}}^\top \bar{\boldsymbol{X}}$ may be much smaller than $\boldsymbol{X}^\top \boldsymbol{X}$, and the IV estimator may be much less efficient than the OLS estimator.

The optimal instruments are given by $\bar{\boldsymbol{X}}$, which we do not observe.

But if we redefine $\boldsymbol{W}$ as an $N \times l$ matrix, we can use $\boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X}$ to estimate $\bar{\boldsymbol{X}}$:

$$\hat{\boldsymbol{\beta}}_{\mathrm{IV}} = (\boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{y}. \qquad \text{[GIVE again]} \tag{27}$$

Standard asymptotic results for GIVE are similar to those for simple IV. $\boldsymbol{S}_{\boldsymbol{W}^\top \boldsymbol{X}}$ is now $l \times k$ instead of $k \times k$, and $\boldsymbol{S}_{\boldsymbol{W}^\top \boldsymbol{W}}$ is now $l \times l$. The former should have rank k, the latter rank l.

It is still essential that

$$\operatorname*{plim}_{N \to \infty} \frac{1}{N} \boldsymbol{W}^\top \boldsymbol{u} = \boldsymbol{0}. \tag{28}$$

Including even one invalid instrument in $\boldsymbol{W}$ will destroy consistency.

The most efficient possible GIV estimator uses $\bar{\boldsymbol{X}}$ as the matrix of instruments. It is less efficient than OLS because

$$\operatorname*{plim}_{N \to \infty} \frac{1}{N} \boldsymbol{X}^\top \boldsymbol{X} = \operatorname*{plim}_{N \to \infty} \frac{1}{N} \bar{\boldsymbol{X}}^\top \bar{\boldsymbol{X}} + \operatorname*{plim}_{N \to \infty} \frac{1}{N} \boldsymbol{V}^\top \boldsymbol{V}. \tag{29}$$

This follows from (25). Terms involving $\boldsymbol{V}^\top \bar{\boldsymbol{X}}$ vanish.

Adding additional (valid) instruments to $\boldsymbol{W}$ brings $\boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X}$ closer to $\bar{\boldsymbol{X}}^\top \bar{\boldsymbol{X}}$ and thus increases asymptotic efficiency. But $\boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X}$ may still be much "smaller" than $\boldsymbol{X}^\top \boldsymbol{X}$, so that $\hat{\boldsymbol{\beta}}_{\text{IV}}$ is much less efficient than $\hat{\boldsymbol{\beta}}_{\text{OLS}}$.

The generalized IV estimator can be obtained by minimizing the **IV criterion function**

$$Q(\boldsymbol{\beta}, \boldsymbol{y}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top \boldsymbol{P}_{\boldsymbol{W}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}). \tag{30}$$

The resulting moment conditions are

$$\boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{0}. \tag{31}$$

Intuitively, we only minimize the part of the residuals that can be explained by $\boldsymbol{W}$.

In the i.i.d. case, $\hat{\boldsymbol{\beta}}_{\text{IV}}$ will be asymptotically normal if it is root-N consistent and

$$N^{-1/2} \boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{u} \overset{a}{\sim} \text{N}\left(\boldsymbol{0}, \sigma_0^2 \operatorname*{plim}_{N \to \infty} \frac{1}{N} \boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X}\right). \tag{32}$$

With heteroskedasticity and/or clustering, the covariance matrix of $N^{-1/2} \boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{u}$ will be more complicated.

Two-Stage Least Squares

The (generalized) IV estimator is also commonly known as the **two-stage least-squares**, or **2SLS**, estimator.

It can be calculated using OLS regressions in two stages.

In the first stage, each column $\boldsymbol{x}_i$, $i = 1, \ldots, k$, of $\boldsymbol{X}$ is regressed on $\boldsymbol{W}$, if necessary. If $\boldsymbol{x}_i$ is a valid instrument, it is already one of the columns of $\boldsymbol{W}$, so it serves as its own instrument without a regression.

The second-stage regression is

$$\boldsymbol{y} = \boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X}\boldsymbol{\beta} + \text{residuals}. \tag{33}$$

Because $\boldsymbol{P}_{\boldsymbol{W}}$ is idempotent, the OLS estimate of $\boldsymbol{\beta}$ from this second-stage regression is

$$\hat{\boldsymbol{\beta}}_{\text{2sls}} = (\boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{y}, \tag{34}$$

which is identical to $\hat{\boldsymbol{\beta}}_{\text{IV}}$.

Estimating the Variance of the Disturbances

If 2SLS is used, it is tricky to estimate the standard error of the regression and the covariance matrix of the parameter estimates.

The OLS estimate of σ^2 from (33) is

$$s^2 = \frac{\|\boldsymbol{y} - \boldsymbol{P}_W \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\mathrm{IV}}\|^2}{N - k}. \tag{35}$$

In contrast, the estimate that was used in the estimated IV covariance matrix is

$$\hat{\sigma}^2 = \frac{\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\mathrm{IV}}\|^2}{N}. \tag{36}$$

These two estimates of σ^2 are not asymptotically equivalent, and s^2 is not consistent.

Residuals from (33) do not tend to $\boldsymbol{u}$ as $N \to \infty$.

This happens because the regressors $\boldsymbol{P_W X}$ are not the true explanatory variables $\boldsymbol{X}$.

We divided by N in (36) because the correct IV residuals, $\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{IV}}$, are not necessarily too small.

Cluster-Robust Inference

When the disturbances display intra-cluster correlation, the scores almost always will too.

We need to use the CRVE

$$(\boldsymbol{X}^\top \boldsymbol{P_W X})^{-1} \left(\sum_{g=1}^{G} [\boldsymbol{X}^\top \boldsymbol{P_W}]_g \, \hat{\boldsymbol{u}}_g \hat{\boldsymbol{u}}_g^\top [\boldsymbol{P_W X}]_g \right) (\boldsymbol{X}^\top \boldsymbol{P_W X})^{-1}, \tag{37}$$

where $[\boldsymbol{P_W X}]_g$ is the $N_g \times k$ submatrix of $\boldsymbol{P_W X}$ for cluster g, and $\hat{\boldsymbol{u}}_g$ is the corresponding $N_g \times 1$ subvector of $\hat{\boldsymbol{u}}$.

Heteroskedasticity-robust variance estimator is a special case of (37) with $G = N$.

Tests of Over-Identifying Restrictions

A model with k regressors and l instruments implicitly incorporates $l - k$ **over-identifying restrictions**. We cannot solve l equations uniquely for k unknowns, so we only solve k moment conditions.

We can always write

$$\mathcal{S}(\boldsymbol{W}) = \mathcal{S}(\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X}, \boldsymbol{W}^*), \tag{38}$$

where $\boldsymbol{W}^*$ has $l - k$ columns. Recall that

$$\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X} = \boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top\boldsymbol{X} \tag{39}$$

is just a linear combination of the columns of $\boldsymbol{W}$. Luckily, we don't have to decide which columns of $\boldsymbol{W}$ belong to $\boldsymbol{W}^*$.

Although we cannot test the moment conditions that $\mathrm{E}(\boldsymbol{X}^\top\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{u}) = \boldsymbol{0}$, we can and should test the moment conditions that $\boldsymbol{W}^{*\top}\boldsymbol{u} = \boldsymbol{0}$.

In the i.i.d. case, we can test restrictions on $\boldsymbol{\beta}$ by evaluating the IV criterion function (30) at the restricted and unrestricted estimates to get $Q(\tilde{\boldsymbol{\beta}})$ and $Q(\hat{\boldsymbol{\beta}})$. Then

$$\frac{Q(\tilde{\boldsymbol{\beta}}) - Q(\hat{\boldsymbol{\beta}})}{\sigma^2_{\text{IV}}} \overset{a}{\sim} \chi^2(r). \tag{40}$$

Now consider the IV regression models

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \quad \boldsymbol{u} \sim \text{IID}(\boldsymbol{0}, \sigma^2\mathbf{I}), \quad \text{E}(\boldsymbol{W}^\top \boldsymbol{u}) = \boldsymbol{0}, \text{ and} \tag{41}$$

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{W}^*\boldsymbol{\gamma} + \boldsymbol{u}, \quad \boldsymbol{u} \sim \text{IID}(\boldsymbol{0}, \sigma^2\mathbf{I}), \quad \text{E}(\boldsymbol{W}^\top \boldsymbol{u}) = \boldsymbol{0}. \tag{42}$$

Call the estimates from (41) $\hat{\boldsymbol{\beta}}_{\text{IV}}$ and the ones from (42) $\hat{\boldsymbol{\beta}}_{\text{U}}$. For (41),

$$Q(\hat{\boldsymbol{\beta}}_{\text{IV}}) = \hat{\boldsymbol{u}}_{\text{IV}}{}^\top \boldsymbol{P}_{\boldsymbol{W}} \hat{\boldsymbol{u}}_{\text{IV}}. \tag{43}$$

For (42), by contrast, $Q(\hat{\boldsymbol{\beta}}_{\text{U}}) = \hat{\boldsymbol{u}}_{\text{U}}^\top \boldsymbol{P}_{\boldsymbol{W}} \hat{\boldsymbol{u}}_{\text{U}} = 0$. This holds because $\mathcal{S}(\boldsymbol{W}) = \mathcal{S}(\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X}, \boldsymbol{W}^*)$, so that $\hat{\boldsymbol{u}}_{\text{U}}$ is orthogonal to $\boldsymbol{W}$.

Thus we don't need to estimate the model (42) at all. We only used it to justify a particular test statistic for over-identifying restrictions.

In this case, the test statistic (40) reduces to

$$\frac{Q(\hat{\boldsymbol{\beta}}_{\mathrm{IV}})}{\hat{\sigma}^2_{\mathrm{IV}}} = \frac{\hat{\boldsymbol{u}}_{\mathrm{IV}}{}^\top \boldsymbol{P}_{\boldsymbol{W}} \hat{\boldsymbol{u}}_{\mathrm{IV}}}{\hat{\sigma}^2_{\mathrm{IV}}} \stackrel{a}{=} \frac{\boldsymbol{u}^\top (\boldsymbol{P}_{\boldsymbol{W}} - \boldsymbol{P}_{\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X}})\boldsymbol{u}}{\sigma^2}, \tag{44}$$

because $\hat{\boldsymbol{u}}_{\mathrm{IV}} = \boldsymbol{u} - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{u}$. The numerator is a quadratic form in a matrix that projects onto a space of dimension $l - k$.

The denominator ensures that, asymptotically, the elements of $\boldsymbol{u}/\sigma$ have variance 1. Thus (44) is asymptotically $\chi^2(l-k)$.

The **Sargan statistic** to test for over-identifying restrictions is the middle expression in (44).

It is simply the ESS from a regression of $\hat{\boldsymbol{u}}_{\mathrm{IV}}$ on $\boldsymbol{W}$, divided by $\hat{\sigma}^2_{\mathrm{IV}}$.

This works because $\hat{\boldsymbol{u}}_{\mathrm{IV}}^\top \boldsymbol{P}_{\boldsymbol{W}} \hat{\boldsymbol{u}}_{\mathrm{IV}} = \boldsymbol{u}^\top (\boldsymbol{P}_{\boldsymbol{W}} - \boldsymbol{P}_{\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X}})\boldsymbol{u}$; see (44).

The model (42) represents two conceptually different alternatives. The "true" parameter vector γ could be nonzero in two situations.

1. The model (14) is correctly specified, but some of the instruments are asymptotically correlated with the disturbances and are therefore not valid instruments.
2. The model (14) is misspecified, and some of the instruments (or, possibly, other variables that are correlated with them) have incorrectly been omitted from the regression function.

In either case, the over-identification test statistic should lead us to reject the null hypothesis whenever the sample size is large enough.

The Sargan statistic can be generalized to allow for heteroskedasticity and/or clustering. For GMM, it becomes the **Hansen-Sargan statistic**, which is the minimized value of the GMM criterion function.

These tests may have poor finite-sample properties, so it can be good to bootstrap them; see Davidson and MacKinnon (2015).

Durbin-Wu-Hausman Tests

When do we actually need to use instrumental variables?

Some variables may be measured with error, but how large are the errors? Will they cause enough inconsistency to worry about?

Is a certain explanatory variable actually endogenous? If so, how much inconsistency will this cause?

It may be perfectly reasonable to employ OLS estimation if the disturbances are not very correlated with the regressors.

We can test whether there is correlation by using a test due to Durbin (1954), Wu (1974), and Hausman (1978).

The null and alternative hypotheses for the DWH test are

$$H_0: \quad \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \quad \boldsymbol{u} \sim \text{IID}(\boldsymbol{0}, \sigma^2 \mathbf{I}), \quad \text{E}(\boldsymbol{X}^\top \boldsymbol{u}) = \boldsymbol{0}, \text{ and} \tag{45}$$

$$H_1: \quad \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \quad \boldsymbol{u} \sim \text{IID}(\boldsymbol{0}, \sigma^2 \mathbf{I}), \quad \text{E}(\boldsymbol{W}^\top \boldsymbol{u}) = \boldsymbol{0}. \tag{46}$$

Under H_1, $\hat{\boldsymbol{\beta}}_{\text{IV}}$ is consistent, but $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ is not. Under H_0, both are consistent, but $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ is more efficient.

Evidently, $\text{plim}(\hat{\boldsymbol{\beta}}_{\text{IV}} - \hat{\boldsymbol{\beta}}_{\text{OLS}})$ is zero under the null and nonzero under the alternative.

The **vector of contrasts** is

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{\text{IV}} - \hat{\boldsymbol{\beta}}_{\text{OLS}} &= (\boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{y} - (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} && (47)\\
&= (\boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X})^{-1} \big(\boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{y} - \boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} \big) && (48)\\
&= (\boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \big(\mathbf{I} - \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \big) \boldsymbol{y} && (49)\\
&= (\boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{M}_{\boldsymbol{X}} \boldsymbol{y}. && (50)
\end{aligned}$$

Here $(\boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X})^{-1}$ is a positive definite matrix.

Testing whether $\hat{\boldsymbol{\beta}}_{\text{IV}} - \hat{\boldsymbol{\beta}}_{\text{OLS}} = \mathbf{0}$ is equivalent to testing whether $\text{E}(\boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{M}_{\boldsymbol{X}} \boldsymbol{y}) = \mathbf{0}$. Since $\boldsymbol{M}_{\boldsymbol{X}} \boldsymbol{y} = \boldsymbol{M}_{\boldsymbol{X}} \boldsymbol{u}$, it does if $\boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X}$ is uncorrelated with $\boldsymbol{M}_{\boldsymbol{X}} \boldsymbol{u}$.

Let us partition $\boldsymbol{X}$ as $[\boldsymbol{Z} \;\; \boldsymbol{Y}]$, where the k_1 columns of $\boldsymbol{Z}$ belong to $\boldsymbol{W}$, and the $k_2 = k - k_1$ columns of $\boldsymbol{Y}$ are treated as potentially endogenous.

It is always the case that $\boldsymbol{Z}^\top \hat{\boldsymbol{u}} = \boldsymbol{Z}^\top \boldsymbol{P}_W \boldsymbol{M}_X \hat{\boldsymbol{u}} = \boldsymbol{0}$; here $\hat{\boldsymbol{u}} = \hat{\boldsymbol{u}}_{\text{OLS}}$.

But it will almost never be the case $\boldsymbol{Y}^\top \boldsymbol{P}_W \boldsymbol{M}_X \hat{\boldsymbol{u}} = \boldsymbol{0}$. We can test whether this vector differs significantly from $\boldsymbol{0}$.

The easiest way to test whether $\mathrm{E}(\boldsymbol{Y}^\top \boldsymbol{P}_W \boldsymbol{M}_X \boldsymbol{y}) = \boldsymbol{0}$ is to use an F test for the k_2 restrictions $\boldsymbol{\delta} = \boldsymbol{0}$ in the OLS regression

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{P}_W \boldsymbol{Y}\boldsymbol{\delta} + \boldsymbol{u}. \tag{51}$$

The OLS estimates of $\boldsymbol{\delta}$ are, by the FWL Theorem, the same as those from the FWL regression of $\boldsymbol{M}_X \boldsymbol{y}$ on $\boldsymbol{M}_X \boldsymbol{P}_W \boldsymbol{Y}$, that is,

$$\hat{\boldsymbol{\delta}} = (\boldsymbol{Y}^\top \boldsymbol{P}_W \boldsymbol{M}_X \boldsymbol{P}_W \boldsymbol{Y})^{-1} \boldsymbol{Y}^\top \boldsymbol{P}_W \boldsymbol{M}_X \boldsymbol{y}. \tag{52}$$

Testing whether $\boldsymbol{\delta} = \boldsymbol{0}$ is equivalent to testing whether $\mathrm{E}(\boldsymbol{Y}^\top \boldsymbol{P}_W \boldsymbol{M}_X \boldsymbol{y}) = \boldsymbol{0}$. We can use an F test with k_2 and $N - k - k_2$ degrees of freedom.

It is not hard to see that an F test of $\boldsymbol{\zeta} = \mathbf{0}$ in the regression

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{M_W Y\zeta} + \boldsymbol{u} \tag{53}$$

is numerically identical to the F test for $\boldsymbol{\delta} = \mathbf{0}$ in (51).

This follows from the fact that $\mathcal{S}(\boldsymbol{Y},\ \boldsymbol{P_W Y})$ and $\mathcal{S}(\boldsymbol{Y},\ \boldsymbol{M_W Y})$ are the same subspace, so that (51) and (53) fit the same.

If the DWH test rejects, there are two possibilities.

1. At least one column of $\boldsymbol{Y}$ is endogenous.
2. One or more of columns of $\boldsymbol{W}$ have incorrectly been omitted from $\boldsymbol{Z}$; some instruments should have been treated as regressors.

If the DWH test does not reject, then we may feel justified in using OLS. But classic pre-test issues arise when we use this test (Guggenberger, 2010a, 2010b).

Tests Based on Vectors of Contrasts

Hausman (1978) proposed a test that can be used whenever there are two root-N consistent estimators.

One of them, say $\hat{\boldsymbol{\theta}}_{\mathrm{I}}$, is like $\hat{\boldsymbol{\beta}}_{\mathrm{IV}}$. It is inefficient but consistent under relatively weak conditions.

The other, say $\hat{\boldsymbol{\theta}}_{\mathrm{E}}$, is like $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$. It may be inconsistent, but it is more efficient when it is consistent.

In a broad range of cases, we can write

$$N^{1/2}(\hat{\boldsymbol{\theta}}_{\mathrm{I}} - \boldsymbol{\theta}_0) \stackrel{a}{=} N^{1/2}(\hat{\boldsymbol{\theta}}_{\mathrm{E}} - \boldsymbol{\theta}_0) + \boldsymbol{v}, \tag{54}$$

where $\boldsymbol{v}$ is a random k-vector that is uncorrelated with $N^{1/2}(\hat{\boldsymbol{\theta}}_{\mathrm{E}} - \boldsymbol{\theta}_0)$.

Recall the proof of the Gauss-Markov Theorem.

The vector $\boldsymbol{v}$ is asymptotically equal to $N^{1/2}$ times the vector of contrasts, which is just $\hat{\boldsymbol{\theta}}_{\mathrm{I}} - \hat{\boldsymbol{\theta}}_{\mathrm{E}}$.

Because v is not correlated with $N^{1/2}(\hat{\boldsymbol{\theta}}_{\mathrm{E}} - \boldsymbol{\theta}_0)$,

$$\mathrm{Var}(v) \stackrel{a}{=} \mathrm{Var}\left(N^{1/2}(\hat{\boldsymbol{\theta}}_{\mathrm{I}} - \boldsymbol{\theta}_0)\right) - \mathrm{Var}\left(N^{1/2}(\hat{\boldsymbol{\theta}}_{\mathrm{E}} - \boldsymbol{\theta}_0)\right). \tag{55}$$

This is Hausman's key result. Therefore, a suitable test statistic is

$$(\hat{\boldsymbol{\theta}}_{\mathrm{I}} - \hat{\boldsymbol{\theta}}_{\mathrm{E}})^{\top}\left(\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}}_{\mathrm{I}}) - \widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}}_{\mathrm{E}})\right)^{-1}(\hat{\boldsymbol{\theta}}_{\mathrm{I}} - \hat{\boldsymbol{\theta}}_{\mathrm{E}}). \tag{56}$$

However, Hausman forgot that the covariance matrix that is inverted in the middle of (56) may not have full rank.

It does not have full rank in the DWH case if we define $\boldsymbol{\theta}$ as $\boldsymbol{\beta}$. We need to define it as the coefficient vector for $\boldsymbol{Y}$, which has k_2 elements.

Care needs to be taken when using Hausman tests based on (56). Ideally, the rank of $\mathrm{Var}(v)$ should be known. It is the number of degrees of freedom for the test based on (56).

The numerical rank of $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}}_{\mathrm{I}}) - \widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}}_{\mathrm{E}})$ may be misleading.

GIV Estimation Using Control Functions

The estimate of $\boldsymbol{\beta}$ from the DWH test regression (53) is simply $\hat{\boldsymbol{\beta}}_{\text{IV}}$.

Here the regressor(s) $\boldsymbol{M}_{\boldsymbol{W}}\boldsymbol{Y}$ are called **control functions**.

There is no advantage to computing $\hat{\boldsymbol{\beta}}_{\text{IV}}$ in this way for linear regression models, but control function estimators can be useful for nonlinear models, such as logit and probit.

Care must be taken to obtain valid standard errors. The usual standard errors from (51) are not valid, because the control functions are **generated regressors**.

The easiest approach is often to use the bootstrap. The bootstrap samples may be obtained by resampling the rows of $[\boldsymbol{y}, \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{W}]$.

But bootstrap standard errors will be very misleading if $\hat{\boldsymbol{\beta}}$ is exactly identified. They may perform poorly if there are not several over-identifying restrictions. Bootstrap IQR approach should work better.

We can demonstrate numerically that the DWH regression yields $\hat{\boldsymbol{\beta}}_{\text{IV}}$.

It is easiest to prove it if we assume that there is just one endogenous right-hand-side variable and one instrument:

$$\boldsymbol{y}_1 = \beta \boldsymbol{y}_2 + \boldsymbol{u} \tag{57}$$

$$\boldsymbol{y}_2 = \gamma \boldsymbol{w} + \boldsymbol{v}. \tag{58}$$

In this case, it is obvious that

$$\hat{\beta}_{\text{IV}} = (\boldsymbol{y}_2^\top \boldsymbol{P}_w \boldsymbol{y}_2)^{-1} \boldsymbol{y}_2^\top \boldsymbol{P}_w \boldsymbol{y}_1. \tag{59}$$

The DWH regression is

$$\boldsymbol{y}_1 = \beta \boldsymbol{y}_2 + \eta \boldsymbol{M}_w \boldsymbol{y}_2 + \boldsymbol{u}. \tag{60}$$

The OLS estimate of β, which is the control function estimate, is then easily seen to be

$$\hat{\beta}_{\text{CF}} = (y_2^\top M_{M_w y_2} y_2)^{-1} y_2^\top M_{M_w y_2} y_1. \tag{61}$$

Now observe that

$$M_{M_w y_2} = P_{P_w y_2}, \tag{62}$$

because $M_w y_2$ and $P_w y_2$ are orthogonal projection matrices. Projecting off the former is equivalent to projecting onto the latter. Thus

$$\hat{\beta}_{\text{CF}} = (y_2^\top P_{P_w y_2} y_2)^{-1} y_2^\top P_{P_w y_2} y_1. \tag{63}$$

Now observe that

$$P_{P_w y_2} = P_w y_2 (y_2^\top P_w y_2)^{-1} y_2^\top P_w. \tag{64}$$

It follows that

$$y_2^\top P_{P_w y_2} y_2 = y_2^\top P_w y_2 \quad \text{and} \quad y_2^\top P_{P_w y_2} y_1 = y_2^\top P_w y_1. \tag{65}$$

When we substitute (65) into (63), we find that

$$\hat{\beta}_{\text{CF}} = (\boldsymbol{y}_2^\top \boldsymbol{P}_w \boldsymbol{y}_2)^{-1} \boldsymbol{y}_2^\top \boldsymbol{P}_w \boldsymbol{y}_1 = \hat{\beta}_{\text{IV}} \tag{66}$$

Because there is only one instrument, this expression can be simplified further. Observe that

$$(\boldsymbol{y}_2^\top \boldsymbol{P}_w \boldsymbol{y}_2)^{-1} = (\boldsymbol{y}_2^\top \boldsymbol{w})^{-1} \boldsymbol{w}^\top \boldsymbol{w} (\boldsymbol{w}^\top \boldsymbol{y}_2)^{-1}, \text{ and} \tag{67}$$

$$\boldsymbol{y}_2^\top \boldsymbol{P}_w \boldsymbol{y}_1 = \boldsymbol{y}_2^\top \boldsymbol{w} (\boldsymbol{w}^\top \boldsymbol{w})^{-1} \boldsymbol{w}^\top \boldsymbol{y}_1. \tag{68}$$

Thus we see that

$$\hat{\beta}_{\text{CF}} = \hat{\beta}_{\text{IV}} = \frac{\boldsymbol{y}_1^\top \boldsymbol{w}}{\boldsymbol{y}_2^\top \boldsymbol{w}} = (\boldsymbol{w}^\top \boldsymbol{y}_2)^{-1} \boldsymbol{w}^\top \boldsymbol{y}_1. \tag{69}$$

This is, of course, just the simple IV estimator for (57).

This result also holds when the structural equation includes a matrix of exogenous regressors, say $\boldsymbol{Z}$. We just replace $\boldsymbol{w}$ by $\boldsymbol{M}_{\boldsymbol{Z}} \boldsymbol{w}$ in (69).

The simple IV estimator in (69) has an interesting interpretation.

If we regress y_1 on Z and w, we get the coefficient estimate

$$\frac{w^\top M_Z y_1}{w^\top M_Z w}, \tag{70}$$

which is just the OLS estimate for the reduced-form equation for y_1.

If we regress y_2 on Z and w, we get the coefficient estimate

$$\frac{w^\top M_Z y_2}{w^\top M_Z w}, \tag{71}$$

which is just the OLS estimate for the reduced-form equation for y_2.

Since these two OLS estimators have the same denominator, their ratio is just the IV estimator $w^\top M_Z y_1 / w^\top M_Z y_2$.

So the IV estimate will be smaller than the OLS estimate whenever the coefficient (71) is larger than the coefficient (70).