9 Instrumental Variables

```r
library(fixest)
```

In Section 8, we discussed endogeneity problems that lead to the inconsistency of the ordinary least squares (OLS) estimator. One important solution is the **instrumental variables (IV)** method, which allows for consistent estimation under certain conditions when regressors are endogenous.

9.1 Endogenous Regressors Model

In most applications only a subset of the regressors are treated as endogenous.

Let's assume that we have k endogenous regressors $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ik})'$ and r exogenous regressors $\boldsymbol{W}_i = (1, W_{i2}, \ldots, W_{ir})'$.

In many practical applications the number of endogenous regressors k is small (like 1 or 2). The exogenous regressors $\boldsymbol{W}_i$ include the intercept, which is constant and therefore exogenous, and all control variables for which we do not wish to interpret their coefficients in a causal sense.

Consider the linear model equation:

$$Y_i = \boldsymbol{X}_i'\boldsymbol{\beta} + \boldsymbol{W}_i'\boldsymbol{\gamma} + u_i, \quad i = 1, \ldots, n. \tag{9.1}$$

We have

- the dependent variable Y_i;
- the error term u_i;
- the endogenous regressors $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ik})'$;
- the regression coefficients of interest $\boldsymbol{\beta}$;
- the remaining r regressors $\boldsymbol{W}_i = (1, W_{i2}, \ldots, W_{ir})'$, which are assumed to be exogenous or simply control variables;
- the regression coefficients of the exogenous variables $\boldsymbol{\gamma}$.

Recall (A1), which is in this case given by $E[u_i|\boldsymbol{X}_i, \boldsymbol{W}_i] = 0$ but fails under endogeneity.

Since $\boldsymbol{X}_i$ is endogenous, we have $E[\boldsymbol{X}_i u_i] \neq \boldsymbol{0}$, which is a violation of (A1). Thus, the OLS estimator $\widehat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ is inconsistent.

9.2 Instrumental Variables Model

To consistently estimate $\boldsymbol{\beta}$ in the endogenous regressors model we require additional information. One type of information which is commonly used in economic applications are what we call **instruments**.

A vector of **instrumental variables (IV)** $\boldsymbol{Z}_i = (Z_{i1}, \dots, Z_{im})$ for the endogenous variable X_{ij} is a variable that is

1) **relevant**, meaning that it has a non-zero conditional marginal effect on X_{ij} after controlling for $\boldsymbol{W}_i$. That is, when regressing X_{ij} on $\boldsymbol{Z}_i$ and $\boldsymbol{W}_i$ we have:

$$X_{ij} = \boldsymbol{Z}_i'\boldsymbol{\pi}_{1j} + \boldsymbol{W}_i'\boldsymbol{\pi}_{2j} + v_{ij}, \quad \boldsymbol{\pi}_{1j} \neq \boldsymbol{0}. \tag{9.2}$$

2) **exogenous** with respect to the error term u_i, i.e.:

$$E[\boldsymbol{Z}_i u_i] = \boldsymbol{0}. \tag{9.3}$$

   This means $\boldsymbol{Z}_i$ doesn't have a direct causal effect on Y_i after controlling for $\boldsymbol{W}_i$, only an indirect effect through the endogenous variable X_{ij}.

If there are k endogenous regressors, we need at least k instruments. If $m = k$, we say that the coefficients are exactly identified and if $m > k$ we say that they are overidentified. Then the relevance condition can be expressed jointly as:

$$\operatorname{rank}\Big(E[\tilde{\boldsymbol{Z}}_i \boldsymbol{X}_i']\Big) = k \tag{9.4}$$

where $\tilde{\boldsymbol{Z}}_i := (\boldsymbol{Z}_i', \boldsymbol{W}_i')'$.

Because $\boldsymbol{\pi}_{1j} \neq \boldsymbol{0}$, some part of the variation in X_{ij} can be explained by $\boldsymbol{Z}_i$. Because $\boldsymbol{Z}_i$ is exogenous, that part of the variation in X_{ij} explained by $\boldsymbol{Z}_i$ is exogenous as well and can be used to estimate β_j consistently.

Example 1: Years of schooling -> wage (returns to education). Ability bias: unobserved ability affects both education choices and wages. Possible instruments for years of schooling: distance to nearest colleges, school construction programs, quarter-of-birth, birth order.

Example 2: Market price -> quantity demanded (price elasticity of demand). Simultaneity: quantity demanded feeds back into equilibrium price. Possible instruments for market price: input-costs (e.g., raw materials, energy costs), weather conditions, tax changes.

Example 3: Police presence -> crime (deterrence effect). Reverse causality: more police are deployed to high-crime areas. Possible instruments for police presence: election cycles, sports/large public events, fire-fighters employment.

The idea of instrumental variable regression is to decompose the endogenous regressor X_{ij} into two parts: the "good" exogenous variation explained by the exogenous instruments $\boldsymbol{Z}_i$

and further exogenous control variables, and the “bad” endogenous variation that is correlated with the error term u_i.

This is exactly what is done in Equation 9.2: $\boldsymbol{Z}_i'\boldsymbol{\pi}_{1j}+\boldsymbol{W}_i'\boldsymbol{\pi}_{2j}$ is the part of X_{ij} that is exogenous and v_{ij} is the part of X_{ij} that is endogenous.

9.3 Two Stage Least Squares

The two stage least squares (TSLS) estimator exploits exactly the idea discussed above: first extracting the exogenous part of the endogenous regressors explained by the instruments as described in Equation 9.2 and then use only this exogenous part to estimate the causal relationship of interest.

The *first stage regression* is:

$$X_{ij} = \boldsymbol{Z}_i'\boldsymbol{\pi}_{1j} + \boldsymbol{W}_i'\boldsymbol{\pi}_{2j} + v_{ij}.$$

This equation is sometimes called the *reduced form* equation for X_{ij}. We estimate this regression for $j = 1, \dots, k$ and collect the fitted values

$$\widehat{X}_{ij} = \boldsymbol{Z}_i'\widehat{\boldsymbol{\pi}}_{1j} + \boldsymbol{W}_i'\widehat{\boldsymbol{\pi}}_{2j}, \quad j = 1, \dots, k, \quad i = 1, \dots, n.$$

Let

$$\widehat{\boldsymbol{X}}_i = (\widehat{X}_{i1}, \dots, \widehat{X}_{ik})', \quad i = 1, \dots, n.$$

be the vector of the fitted values for the k endogenous variables from the first stage.

Note that $\widehat{\boldsymbol{X}}_i$ is a function of $\boldsymbol{Z}_i$ and $\boldsymbol{W}_i$ and is therefore exogenous, i.e., uncorrelated with u_i.

Then, the *second stage regression* is

$$Y_i = \widehat{\boldsymbol{X}}_i'\boldsymbol{\beta} + \boldsymbol{W}_i'\boldsymbol{\gamma} + w_i, \quad i = 1, \dots, n. \tag{9.5}$$

Note that w_i absorbs v_{ij}, the part of X_{ij} that is endogenous. Therefore, the second stage regression does not suffer any more from an endogeneity problem and can be used to consistently estimate $\boldsymbol{\beta}$.

The OLS estimator of the second stage (Equation 9.5), denoted as $\hat{\boldsymbol{\beta}}_{TSLS}$ is called the **two-stage least squares estimator** for $\boldsymbol{\beta}$.

9.4 TSLS Assumptions

- (A1-iv) $E[u_i|\boldsymbol{W}_i] = 0$.
- (A2-iv) $(Y_i, \boldsymbol{X}_i', \boldsymbol{W}_i', \boldsymbol{Z}_i')_{i=1}^n$ is an i.i.d. sample.
- (A3-iv) All variables have finite kurtosis.
- (A4-iv) The instrument exogeneity and relevance conditions from Equation 9.3 and Equation 9.4 hold, and $E[\tilde{\boldsymbol{Z}}_i \tilde{\boldsymbol{Z}}_i']$ is invertible

(A1-iv) is the exogeneity condition for the control variables $\boldsymbol{W}_i$.

(A2-iv) is the standard random sampling assumption for the data.

(A3-iv) is the standard light-tails assumption, meaning large outliers are unlikely

(A4-iv) is the exogeneity and relevance condition for the instruments together with a condition that excludes perfect multicollinearity

9.5 Large-Sample Properties of TSLS

Under assumptions (A1-iv)–(A4-iv), the TSLS estimator is consistent:

$$\hat{\boldsymbol{\beta}}_{TSLS} \xrightarrow{p} \boldsymbol{\beta} \quad (\text{as } n \to \infty).$$

Furthermore, the estimator is asymptotically normal:

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{TSLS} - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \boldsymbol{V}_{TSLS}),$$

where

$$\boldsymbol{V}_{TSLS} = (\boldsymbol{Q}_{XZ}\boldsymbol{Q}_{ZZ}^{-1}\boldsymbol{Q}_{ZX})^{-1}\boldsymbol{Q}_{XZ}\boldsymbol{Q}_{ZZ}^{-1}\boldsymbol{\Omega}\boldsymbol{Q}_{ZZ}^{-1}\boldsymbol{Q}_{ZX}(\boldsymbol{Q}_{XZ}\boldsymbol{Q}_{ZZ}^{-1}\boldsymbol{Q}_{ZX})^{-1},$$

with

$$\boldsymbol{Q}_{XZ} = E[\boldsymbol{X}_i\tilde{\boldsymbol{Z}}_i'], \quad \boldsymbol{Q}_{ZX} = E[\tilde{\boldsymbol{Z}}_i\boldsymbol{X}_i'], \quad \boldsymbol{Q}_{ZZ} = E[\tilde{\boldsymbol{Z}}_i\tilde{\boldsymbol{Z}}_i'], \quad \boldsymbol{\Omega} = E[u_i^2\tilde{\boldsymbol{Z}}_i\tilde{\boldsymbol{Z}}_i'].$$

The asymptotic variance can be estimated as:

$$\widehat{\boldsymbol{V}}_{TSLS} = \frac{n}{n-k-r}\left(\frac{1}{n}\sum_{i=1}^n \widehat{\boldsymbol{X}}_i\widehat{\boldsymbol{X}}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n \hat{u}_i^2\widehat{\boldsymbol{X}}_i\widehat{\boldsymbol{X}}_i'\right)\left(\frac{1}{n}\sum_{i=1}^n \widehat{\boldsymbol{X}}_i\widehat{\boldsymbol{X}}_i'\right)^{-1}$$

This is the HC1 covariance matrix estimator for the TSLS estimator. It can be used to construct confidence intervals, t-tests, and F-tests in the usual way as discussed in previous sections.

9.6 Example: Return of Education

Consider a wage equation for a cross-section of 429 married women:

$$\log(\text{wage}) = \beta_1 + \beta_2 \text{educ}_i + \beta_3 \text{exper}_i + \beta_4 \text{exper}_i^2 + u_i,$$

where

- *wage* is the wife's 1975 average hourly earnings
- *educ* is her educational attainment in years
- *exper* are the actual years of her labor market experience

We use the dataset `mroz` available in this repository: link.

OLS yields:

```r
feols(log(wage) ~ educ + exper + exper^2, data = mroz, vcov = "HC1")
```

```
OLS estimation, Dep. Var.: log(wage)
Observations: 428
Standard-errors: Heteroskedasticity-robust
             Estimate Std. Error  t value   Pr(>|t|)
(Intercept) -0.522041   0.201650 -2.58884 9.9611e-03 **
educ         0.107490   0.013219  8.13147 4.7203e-15 ***
exper        0.041567   0.015273  2.72156 6.7651e-03 **
I(exper^2)  -0.000811   0.000420 -1.93108 5.4139e-02 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 0.663299   Adj. R2: 0.150854
```

If *educ* is correlated with omitted variables like ability or motivation, the estimated coefficient for *educ* does not represent the causal effect of *educ* on *wage*.

Ability is an unobserved confounder that affects both *educ* and *wage*.

In the following, we assume that mother's education (*mothereduc*) is a valid instrument for *educ* in the wage equation because:

1) *mothereduc* should not appear in a wife's wage equation
2) Instrument relevance: *mothereduc* should be correlated with *educ*: high educated mothers typically have high educated daughters
3) Instrument exogeneity: assume that a woman's ability and motivation are uncorrelated with *mothereduc*

The first stage regression is:

```
firststage = lm(educ ~ mothereduc + exper + I(exper^2), data = mroz)
firststage
```

```

Call:
lm(formula = educ ~ mothereduc + exper + I(exper^2), data = mroz)

Coefficients:
(Intercept)   mothereduc        exper   I(exper^2)
   9.775103     0.267691     0.048862    -0.001281
```

The second stage regression is:

```
Xhat = firststage$fitted.values
secondstage = lm(log(wage) ~ Xhat + exper + I(exper^2), data = mroz)
secondstage
```

```

Call:
lm(formula = log(wage) ~ Xhat + exper + I(exper^2), data = mroz)

Coefficients:
(Intercept)         Xhat        exper   I(exper^2)
  0.1981861    0.0492630    0.0448558   -0.0009221
```

Note that standard errors from these two separate steps should not be used. Instead, the `feols` function gives you the correct standard errors by using the following notation:

```
feols(log(wage) ~ exper + exper^2 | educ ~ mothereduc, data = mroz, vcov = "HC1")
```

```
TSLS estimation - Dep. Var.: log(wage)
                  Endo.    : educ
                  Instr.   : mothereduc
Second stage: Dep. Var.: log(wage)
Observations: 428
Standard-errors: Heteroskedasticity-robust
             Estimate Std. Error  t value  Pr(>|t|)
(Intercept)  0.198186   0.489146  0.405167 0.6855588
```

```
fit_educ     0.049263   0.038040  1.295045 0.1960095
exper        0.044856   0.015604  2.874667 0.0042481 **
I(exper^2)  -0.000922   0.000432 -2.135025 0.0333316 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 0.67642   Adj. R2: 0.116926
F-test (1st stage), educ: stat = 73.9   , p < 2.2e-16 , on 1 and 424 DoF.
              Wu-Hausman: stat =  2.9683, p = 0.085642, on 1 and 423 DoF.
```

- The coefficient for *educ* drops from 0.107 to 0.059
- OLS overestimates the impact of education on wages
- The t-statistic has a p-value of 0.19
- Using *mothereduc* as an instrument, *educ* is no longer significant

To improve the precision of the IV estimator, we can include further instruments like *fathereduc*

```r
feols(log(wage) ~ exper + exper^2 | educ ~ mothereduc + fathereduc, data = mroz, vcov = "HC1"
```

```
TSLS estimation - Dep. Var.: log(wage)
                  Endo.    : educ
                  Instr.   : mothereduc, fathereduc
Second stage: Dep. Var.: log(wage)
Observations: 428
Standard-errors: Heteroskedasticity-robust
             Estimate Std. Error   t value  Pr(>|t|)
(Intercept)  0.048100   0.429798  0.111914 0.9109447
fit_educ     0.061397   0.033339  1.841609 0.0662307 .
exper        0.044170   0.015546  2.841202 0.0047111 **
I(exper^2)  -0.000899   0.000430 -2.090220 0.0371931 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 0.671551   Adj. R2: 0.129593
F-test (1st stage), educ: stat = 55.4     , p < 2.2e-16 , on 2 and 423 DoF.
              Wu-Hausman: stat =  2.79259 , p = 0.095441, on 1 and 423 DoF.
                  Sargan: stat =  0.378071, p = 0.538637, on 1 DoF.
```

- Estimated return to education increases from 0.049 to 0.061
- The t-statistic has a p-value of 0.066
- Stronger instruments leads to more efficient IV estimation: *educ* is now significantly different from zero at least at the 10% level.

9.7 IV Diagnostics

The TSLS estimator relies on the exogeneity and relevance of the instruments. In empirical applications, these assumptions should be critically assessed. This section introduces three diagnostic tools used to evaluate different aspects of the IV strategy:

- The **F-test** for instrument relevance
- The **Sargan test** for instrument exogeneity
- The **Wu-Hausman test** for regressor endogeneity

F-test for instrument relevance

The **first-stage F-test** indicates whether the instruments $\boldsymbol{Z}_i \in \mathbb{R}^m$ contain enough information about the endogenous regressors $\boldsymbol{X}_i \in \mathbb{R}^k$, conditional on the exogenous controls $\boldsymbol{W}_i$.

Consider the one endogenous regressor $k = 1$ case with the first-stage regression,

$$X_i = \boldsymbol{Z}_i'\boldsymbol{\pi}_1 + \boldsymbol{W}_i'\boldsymbol{\pi}_2 + v_i,$$

and test the joint null hypothesis

$$H_0 : \boldsymbol{\pi}_1 = \mathbf{0}.$$

To compute the F-statistic for this hypothesis, we follow the usual procedure and use a suitable robust covariance matrix (e.g., HC1 or cluster-robust), with an F-statistic whose null distribution is asymptotically $F_{m,\infty}$.

If the statistic exceeds its critical value you reject H_0 and conclude the instruments are *relevant*.

Large-n 5% critical values for $F_{m,\infty}$ are 3.84 for $m = 1$, 3.00 for $m = 2$, 2.60 for $m = 3$, etc. (compute with `qf(.95, m, Inf)`).

Weak instruments

Relevance alone is not enough: the instruments may be **weak** if their correlation with X_i is small. Weakness matters because two-stage least squares (2SLS) can then suffer a large finite-sample bias toward OLS. Define the *relative bias*

$$\text{relBias} = \frac{E[\hat{\beta}_{TSLS}] - \beta}{E[\hat{\beta}_{OLS}] - \beta}.$$

Staiger and Stock (1997) and Stock and Yogo (2005) derive critical values for the homoskedastic first-stage statistic that control the null hypothesis "relative bias > 10% of the OLS bias" at the 5% significance level. With one instrument the 5% cut-off is approximately **10**. Hence, the following rule of thumb is established in applied work:

$$\begin{aligned} \text{First-stage } F > 10 \quad &\Rightarrow \quad \text{instruments strong} \\ \text{First-stage } F \leq 10 \quad &\Rightarrow \quad \text{instruments weak} \end{aligned}$$

This is a quick approximation that relies on the homoskedasticity assumption and only works well when m is small.

For heteroskedastic (or cluster-robust) settings, Montiel Olea and Pflueger (2013) replace the standard rule of thumb: To reject the null hypothesis of a relative bias larger than 10% at the 5% level you need a robust F-statistic that exceeds its critical value, which varies between about **11 and 23.1** depending on m and the estimated error-covariance matrix (HC1, cluster-robust, HAC, etc.). The conservative rule

$$\begin{aligned} \text{First-stage robust } F > 23.1 \quad &\Rightarrow \quad \text{instruments strong} \\ \text{First-stage robust } F \leq 23.1 \quad &\Rightarrow \quad \text{instruments weak} \end{aligned}$$

is therefore *sufficient* (but not always necessary) for any number of instruments when $k = 1$.

If several regressors are endogenous ($k \geq 2$), each has its own first-stage equation, and the scalar F no longer summarizes the joint instrument strength. An alternative is the matrix-based Kleibergen–Paap tests of Kleibergen and Paap (2006), which extend the Staiger-Stock-Yogo logic to the multivariate case.

Anderson-Rubin Test

To conduct inference when the first-stage is weak, the usual TSLS t-, F- or Wald tests are unreliable – they tend to over-reject and their confidence intervals undercover.

A simple, robust alternative is the **Anderson–Rubin (AR) test**. The logic is that, under the structural model, the instruments $\boldsymbol{Z}_i$ should contain no information about the structural error

$$u_i = Y_i - \boldsymbol{X}_i'\boldsymbol{\beta} - \boldsymbol{W}_i'\boldsymbol{\gamma}.$$

Hence, if the null hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ holds, the adjusted outcome $Y_i - \boldsymbol{X}_i'\boldsymbol{\beta}_0$ must be uncorrelated with the instruments conditional on the controls. In practice one runs the auxiliary regression

$$Y_i - \boldsymbol{X}_i'\boldsymbol{\beta}_0 = \boldsymbol{Z}_i'\boldsymbol{\pi} + \boldsymbol{W}_i'\boldsymbol{\theta} + e_i$$

and computes the heteroskedastic- or cluster-robust F-statistic, F_{rob}, for the joint null $\boldsymbol{\pi} = \boldsymbol{0}$ (numerator d.f. $= m$). Reject H_0 when

$$F_{\text{rob}} > F_{m,\infty;1-\alpha},$$

where m is the number of instruments. This decision rule delivers correct size regardless of instrument strength, but it has lower power than the TSLS-based tests when instruments are strong.

Repeating the test over a grid of candidate $\boldsymbol{\beta}_0$ values and retaining those not rejected yields a $(1-\alpha)$ Anderson–Rubin confidence region that remains valid even when the first-stage F is very small.

Sargan Test for Instrument Exogeneity

When the set of instruments is overidentified ($m > k$), we can statistically assess whether all instruments satisfy the exogeneity condition $E[\boldsymbol{Z}_i u_i] = 0$.
The classical procedure is the **Sargan test** (also called the test of over-identifying restrictions or the J-test).

Null and alternative hypotheses

- H_0 (all instruments are valid): every instrument is uncorrelated with the structural error term u_i.
- H_1 (at least one instrument is invalid): some instrument is correlated with u_i.

Computation of the Sargan J-statistic

1. **Estimate the structural equation by TSLS** (using *all* m instruments) and obtain the residuals
$$\hat{u}_i^{\text{TSLS}} = Y_i - (\boldsymbol{X}_i'\hat{\boldsymbol{\beta}}_{TSLS} + \boldsymbol{W}_i'\hat{\boldsymbol{\gamma}}_{TSLS}).$$
2. **Regress $\hat{u}_i^{\textbf{TSLS}}$ on the full set of instruments and exogenous controls**
$$\hat{u}_i^{\text{TSLS}} = \delta_0 + \delta_1 Z_{i1} + \cdots + \delta_m Z_{im} + \boldsymbol{W}_i'\boldsymbol{\theta} + e_i.$$
3. Let F be the (homoskedastic-only) F-statistic for the joint null $\delta_1 = \cdots = \delta_m = 0$. The Sargan statistic is
$$J = m \cdot F.$$

Under H_0 and homoskedastic errors, $J \sim \chi^2_{m-k}$ in large samples .

If heteroskedasticity is suspected, the Hansen robust J-statistic should be used.

Decision rule and interpretation

- **Reject** H_0 if J exceeds the critical value of the χ^2_{m-k} distribution (or if the p-value is below the chosen significance level). This implies that the data are inconsistent with the joint exogeneity of the instruments; at least one instrument is likely invalid.
- **Fail to reject** H_0 when J is small. This provides no evidence *against* instrument validity, but does not prove exogeneity.

Practical remarks

- The test cannot be performed when the model is exactly identified ($m = k$); then $J = 0$ by construction and instrument validity must be argued on theoretical grounds.
- A significant J-statistic tells us that something is wrong with the instrument set, but not which instrument(s) are problematic. Empirical judgment and auxiliary tests (e.g. re-estimating with different subsets of instruments) are required.

9.7.1 Wu-Hausman Test for Endogeneity

The **Wu-Hausman test** evaluates whether the regressors $\boldsymbol{X}_i$ are in fact endogenous. That is, it tests the null hypothesis of exogeneity, i.e.: $H_0 : E[\boldsymbol{X}_i u_i] = \mathbf{0}$.

Recall the first stage regressions

$$X_{ij} = \boldsymbol{Z}_i' \boldsymbol{\pi}_{1j} + \boldsymbol{W}_i' \boldsymbol{\pi}_{2j} + v_{ij}, \quad j = 1, \dots, k,$$

and let $\boldsymbol{v}_i = (v_{i1}, \dots, v_{ik})'$ be the stacked error terms of the first-stage regressions.

As discussed previously, $\boldsymbol{Z}_i' \boldsymbol{\pi}_{1j} + \boldsymbol{W}_i' \boldsymbol{\pi}_{2j}$ represents the exogenous part of X_{ij} and v_{ij} the endogenous part. Thus, $\boldsymbol{v}_i$ is the endogenous part of the full vector of endogenous regressors $\boldsymbol{X}_i$. Therefore,

$$E[\boldsymbol{X}_i u_i] = \mathbf{0} \quad \Leftrightarrow \quad E[\boldsymbol{v}_i u_i] = \mathbf{0}.$$

Consider $\boldsymbol{\delta} = E[\boldsymbol{v}_i \boldsymbol{v}_i']^{-1} E[\boldsymbol{v}_i u_i]$, which is the population regression coefficient of the auxiliary regression

$$u_i = \boldsymbol{v}_i' \boldsymbol{\delta} + \epsilon_i, \quad E[\boldsymbol{v}_i \epsilon_i] = 0. \tag{9.6}$$

From the definition of $\boldsymbol{\delta}$ we see that

$$\boldsymbol{\delta} = \mathbf{0} \quad \Leftrightarrow \quad E[\boldsymbol{v}_i u_i] = \mathbf{0}.$$

Therefore, testing $H_0 : E[\boldsymbol{X}_i u_i] = \mathbf{0}$ is equivalent to testing $\boldsymbol{\delta} = \mathbf{0}$.

Note that Equation 9.6 is an infeasible regression because u_i and $\boldsymbol{v}_i$ are unknown. While $\boldsymbol{v}_i$ can be estimated using the residuals $\hat{\boldsymbol{v}}_i$ from the first-stage regressions, there are no suitable sample counterparts for u_i available under endogeneity.

We may insert Equation 9.6 into the structural equation given by Equation 9.1:

$$Y_i = \boldsymbol{X}_i'\boldsymbol{\beta} + \boldsymbol{W}_i'\boldsymbol{\gamma} + \boldsymbol{v}_i'\boldsymbol{\delta} + \epsilon_i. \tag{9.7}$$

Equation 9.7 is a well defined regression model with regressors $\boldsymbol{X}_i, \boldsymbol{W}_i, \boldsymbol{v}_i$ and regression error ϵ_i. To see this note that

(i) $E[\boldsymbol{v}_i \epsilon_i] = \mathbf{0}$ by Equation 9.6;
(ii) $E[\boldsymbol{W}_i \epsilon_i] = \mathbf{0}$ because $\boldsymbol{W}_i$ are exogenous;
(iii) $E[\boldsymbol{X}_i \epsilon_i] = \mathbf{0}$ because $E[\boldsymbol{X}_i \epsilon_i] = E[\boldsymbol{v}_i \epsilon_i]$.

Therefore, we may apply an F-test on the restriction $\boldsymbol{\delta} = \mathbf{0}$ in Equation 9.7 when $\boldsymbol{v}_i$ is replaced by $\hat{\boldsymbol{v}}_i$, which is known as the Wu-Hausman test.

Wu-Hausman Procedure:

1. Run the first-stage regression for each endogenous regressor X_{ij} and obtain residuals $\hat{v}_{ij}$, $j = 1, \dots, k$.
2. Stack the residuals as $\hat{\boldsymbol{v}}_i = (\hat{v}_{i1}, \dots, \hat{v}_{ik})'$.
3. Run the **augmented regression**:

$$Y_i = \boldsymbol{X}_i'\boldsymbol{\beta} + \boldsymbol{W}_i'\boldsymbol{\gamma} + \hat{\boldsymbol{v}}_i'\boldsymbol{\delta} + \varepsilon_i.$$

4. Test $H_0 : \boldsymbol{\delta} = \mathbf{0}$ using an F-test or Wald test, which has k restrictions.

If the test does not reject H_0, then there is evidence for exogenous regressors with $E[\boldsymbol{X}_i u_i] = 0$, and the conventional OLS without instruments should be used because it is more efficient than TSLS.

9.8 Example: Return of Education Revisited

Recall the previous TSLS regression with instrument `mothereduc`

```
feols(log(wage) ~ exper + exper^2 | educ ~ mothereduc, data = mroz, vcov = "HC1")
```

```
TSLS estimation - Dep. Var.: log(wage)
                  Endo.    : educ
                  Instr.   : mothereduc
Second stage: Dep. Var.: log(wage)
Observations: 428
Standard-errors: Heteroskedasticity-robust
             Estimate Std. Error   t value  Pr(>|t|)
(Intercept)  0.198186   0.489146  0.405167 0.6855588
fit_educ     0.049263   0.038040  1.295045 0.1960095
exper        0.044856   0.015604  2.874667 0.0042481 **
I(exper^2)  -0.000922   0.000432 -2.135025 0.0333316 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 0.67642   Adj. R2: 0.116926
F-test (1st stage), educ: stat = 73.9   , p < 2.2e-16 , on 1 and 424 DoF.
              Wu-Hausman: stat =  2.9683, p = 0.085642, on 1 and 423 DoF.
```

The first stage F-statistic is 73.9 indicating that the instrument is strong. The Wu-Hausman statistic has a p-value of 0.08, which indicates that `educ` is significantly endogenous at the 10% level. The Sargan test is not displayed because of exact identification.

We also discussed the TSLS results with two instruments:

```r
feols(log(wage) ~ exper + exper^2 | educ ~ mothereduc + fathereduc, data = mroz, vcov = "HC1"
```

```
TSLS estimation - Dep. Var.: log(wage)
                  Endo.    : educ
                  Instr.   : mothereduc, fathereduc
Second stage: Dep. Var.: log(wage)
Observations: 428
Standard-errors: Heteroskedasticity-robust
             Estimate Std. Error   t value  Pr(>|t|)
(Intercept)  0.048100   0.429798  0.111914 0.9109447
fit_educ     0.061397   0.033339  1.841609 0.0662307 .
exper        0.044170   0.015546  2.841202 0.0047111 **
I(exper^2)  -0.000899   0.000430 -2.090220 0.0371931 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 0.671551   Adj. R2: 0.129593
F-test (1st stage), educ: stat = 55.4     , p < 2.2e-16 , on 2 and 423 DoF.
              Wu-Hausman: stat =  2.79259 , p = 0.095441, on 1 and 423 DoF.
                  Sargan: stat =  0.378071, p = 0.538637, on 1 DoF.
```

Similarly, the F-statistic of 55.4 indicates that the instruments are strong and the Wu-Hausman test gives some statistical evidence of an endogeneity problem. The Sargan test does not reject, which indicates no evidence against instrument validity (but does not prove exogeneity of the instruments).

9.9 R-codes

metrics-sec09.R

References

Kleibergen, Frank, and Richard Paap. 2006. "Generalized Reduced Rank Tests Using the Singular Value Decomposition." *Journal of Econometrics* 133 (1): 97–126.

Montiel Olea, José Luis, and Carolin Pflueger. 2013. "A Robust Test for Weak Instruments." *Journal of Business & Economic Statistics* 31 (3): 358–69.

Staiger, Douglas, and James H Stock. 1997. "Instrumental Variables Regression with Weak Instruments." *Econometrica* 65 (3): 557–86.

Stock, James H., and Motohiro Yogo. 2005. "Testing for Weak Instruments in Linear IV Regression." In *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, edited by Donald W. K. Andrews and James H.Editors Stock, 80–108. Cambridge University Press.

Part V

Big Data Econometrics