1 IV estimation and Two-Stage Least Squares

Let’s start with the simplest model for the outcome variable $y_1$ where we have only one (potentially endogenous) explanatory variable, $y_2$, and one instrumental variable, $z$:

$$y_1 = \beta y_2 + e.$$  \hfill (1)

The concern is that $y_2$ is correlated with $e$ ($\text{cov}(y_2, e) \neq 0$) which results in inconsistent OLS estimates of $\beta$. By contrast, $\text{cov}(z, e) = 0$. It is then easy to see how $\beta$ can be estimated by instrumental variables (IV) methods by simply looking at the covariance between $y_1$ and $z$:

$$\text{cov}(y_1, z) = \beta \text{cov}(y_2, z) + \text{cov}(e, z).$$ \hfill (2)

Since $\text{cov}(e, z) = 0$, it follows that:

$$\beta = \frac{\text{cov}(y_1, z)}{\text{cov}(y_2, z)}.$$ \hfill (3)

A consistent estimate $\hat{\beta}_{IV}$ of $\beta$ can then be obtained by replacing the covariances by their empirical counterparts:

$$\hat{\beta}_{IV} = \frac{\hat{\text{cov}}(y_1, z)}{\hat{\text{cov}}(y_2, z)}.$$ \hfill (4)

Note that this model is exactly identified as we have exactly one IV for exactly one endogenous variable. In the exactly identified case, other popular estimators such as
two-stage least-squares (TSLS) or limited information maximum likelihood (LIML) yield numerically identical estimates of $\tilde{\beta}$. Consider, for example, the TSLS estimator. To implement the estimator, consider the linear relationship between $y_2$ and the IV $z$:

$$y_2 = \pi z + u_2,$$

where $\pi$ is the regression coefficient

$$\pi = \frac{\text{cov}(y_2, z)}{\text{var}(z)}$$  \hspace{1cm} (5)

The mechanics of TSLS works as follows. In the first stage, we regress $y_2$ on $z$, which yields the estimate of the regression coefficient $\hat{\pi}$, and then compute the predicted value of $y_2$:

$$\hat{y}_2 = \hat{\pi} z.$$  \hspace{1cm} (6)

In the second stage, $\beta$ is estimated by running a regression of $y_1$ on $\hat{y}_2$. The resulting estimate $\hat{\beta}_{TSLS}$ is:

$$\hat{\beta}_{TSLS} = \frac{\hat{\text{cov}}(y_1, \hat{y}_2)}{\hat{\text{var}}(\hat{y}_2)} = \frac{\hat{\text{cov}}(y_1, \hat{\pi} z)}{\hat{\pi}^2 \text{var}(z)} = \frac{\hat{\text{cov}}(y_1, z)}{\hat{\pi} \text{var}(z)}$$  \hspace{1cm} (7)

$$= \frac{\hat{\text{cov}}(y_1, z)}{(\hat{\text{cov}}(y_2, z)/\text{var}(z))\text{var}(z)} = \frac{\hat{\text{cov}}(y_1, z)}{\hat{\text{cov}}(y_2, z)},$$  \hspace{1cm} (8)

which is identical to $\hat{\beta}_{IV}$. In practice, it is more convenient to use TSLS than IV in the case where there are also exogenous regressors $x$ in the equation for $y_1$. The TSLS estimator is similar except that $x$ is used as an additional regressor in both stage 1 and 2. Furthermore, TSLS can also be readily computed when the model is overidentified, for example when we have two instruments $z_1$ and $z_2$ instead of just $z$.

A useful way of understanding what happens in an overidentified model is to look at the reduced forms for $y_1$ and $y_2$. The reduced form for $y_2$ is simply the equation $y_2 = \pi z + u_2$ defined earlier, which is often called the first-stage equation in TSLS. The reduced form for $y_1$ is obtained by substituting this equation into the structural equation $y_1 = \beta y_2 + e$:

$$y_1 = \beta y_2 + e = \beta (\pi z + u_2) + e = \theta z + u_1$$  \hspace{1cm} (9)

where $\theta = \beta \pi$ and $u_1 = e + \beta u_2$. Note that the reduced form coefficient $\theta$ is just the
regression coefficient:

\[
\theta = \frac{\text{cov}(y_1, z)}{\text{var}(z)}
\]  

(11)

Rewriting \( \theta = \beta \pi \) as \( \beta = \theta / \pi \) shows that \( \beta \) is simply the ratio of the coefficients in the reduced form equations for \( y_1 \) and \( y_2 \). This suggests yet another estimate \( \hat{\beta}_{RBF} \) (RRF for ratio of reduced forms) of \( \beta \) which turns out to be numerically equivalent to \( \hat{\beta}_{IV} \) and \( \hat{\beta}_{TSL} \) in the exactly identified case since:

\[
\hat{\beta}_{RBF} = \frac{\theta}{\pi} = \frac{\text{cov}(y_1, z) / \text{var}(z)}{\text{cov}(y_2, z) / \text{var}(z)} = \frac{\text{cov}(y_1, z)}{\text{cov}(y_2, z)}.
\]  

(12)

Now, when the model is overidentified, the reduced form equation for \( y_2 \) is instead \( y_2 = \pi_1 z_1 + \pi_2 z_2 + u_2 \). Substituting into the equation for \( y_1 \) again, the system of reduced form equations can now be written as:

\[
y_1 = \beta_1 \pi_1 z_1 + \beta_2 \pi_2 z_2 + u_1
\]  

(13)

\[
y_2 = \pi_1 z_1 + \pi_2 z_2 + u_2
\]  

(14)

or as:

\[
y_1 = \theta_1 z_1 + \theta_2 z_2 + u_1
\]  

(15)

\[
y_2 = \pi_1 z_1 + \pi_2 z_2 + u_2
\]  

(16)

where \( \theta_1 = \beta_1 \pi_1 \) and \( \theta_1 = \beta \pi_1 \). The model is overidentified because the four reduced form parameters, \( \theta_1, \theta_2, \pi_1, \) and \( \pi_2, \) depend on only three underlying parameters, \( \beta, \pi_1, \) and \( \pi_2 \). This also means there is no longer a unique estimator for \( \beta \) since we could either (for example) solve for \( \hat{\beta} = \hat{\theta}_1 / \hat{\pi}_1 \) or \( \hat{\beta} = \hat{\theta}_2 / \hat{\pi}_2 \). Different estimators (TSL, LIML, etc.) will now yield different estimates of \( \beta \) depending on what exactly is the criterion used to “best fit” the system of equations for \( y_1 \) and \( y_2 \). The system of equations being fitted is:

\[
y_1 = \hat{\beta}_1 \hat{\pi}_1 z_1 + \hat{\beta}_2 \hat{\pi}_2 z_2 + \hat{u}_1
\]  

(17)

\[
y_2 = \hat{\pi}_1 z_1 + \hat{\pi}_2 z_2 + \hat{u}_2
\]  

(18)

As it turns out, what TSLS does is to pick the (OLS) estimates of \( \hat{\pi}_1 \) and \( \hat{\pi}_2 \) that
minimizes the sum of the square of the residuals $\hat{u}_2$, and then to pick the value of $\hat{\beta}$, holding $\hat{\pi}_1$ and $\hat{\pi}_2$ fixed, that minimizes the sum of the square of the residuals $\hat{u}_1$. One could think, however, of alternative values of $\hat{\pi}_1$ and $\hat{\pi}_2$ that would just worsen a bit the fit of the second equation (sum of squares of $\hat{u}_2$) but greatly improve the fit of the first equation (sum of squares of $\hat{u}_1$). As it turns out, LIML is an example of such an alternative estimator that tries to find the best fit for both equations considered simultaneously. Many studies have found that LIML works better in practice than TSLS for this reason.

2 Overidentification test, Hausman test, and weak instruments.

2.1 Overidentification test

One benefit of an overidentified model is that it provides a (limited way) of testing the validity of the instruments. Remember that in the exactly identified model, we need the instruments to satisfy the conditions $\text{cov}(z, e) = 0$ and $\text{cov}(z, y_2) \neq 0$. While the latter is testable (see weak instruments discussion below), the former is not, which is why theory or institutional details are so important to credibly hope that the condition $\text{cov}(z, e) = 0$ is indeed satisfied. The reason why we cannot test whether the covariance is really equal to zero is that the error term $e$, that can be written as:

$$e = y_1 - \beta y_2,$$

cannot be estimated unless we also have an estimate of $\beta$. The problem is that the OLS estimate of $\beta$ is presumed to be inconsistent, while the IV/TSLS estimate is only consistent if $\text{cov}(z, e) = 0$ holds in the first place. This means that if we compute

$$\hat{e} = y_1 - \hat{\beta}_{IV} y_2$$

and then look at the empirical covariance $\text{cov}(z, \hat{e})$, we will get by construction that $\text{cov}(z, \hat{e}) = 0$ because:

$$\text{cov}(\hat{e}, z) = \text{cov}(z, y_1) - \hat{\beta}_{IV} \text{cov}(z, y_2) = \text{cov}(z, y_1) - \frac{\text{cov}(z, y_1)}{\text{cov}(z, y_2)} \text{cov}(z, y_2) = 0$$

(21)
In the overidentified case, however, it is possible to test whether $\text{cov}(z_2, e) = 0$ under the assumption that $\text{cov}(z_1, e) = 0$ (or vice versa). For example, we can compute $\hat{c} = y_1 - \hat{\beta}_{IV1} y_2$, where $\hat{\beta}_{IV1}$ is the IV estimator obtained using $z_1$ as IV, and then test whether $\text{cov}(z_2, \hat{c})$ is equal to zero. It is easy to show that $\text{cov}(z_2, \hat{c}) = 0$ iff $\hat{\beta}_{IV1} = \widehat{\theta}_1/\widehat{\pi}_1 = \widehat{\theta}_2/\widehat{\pi}_2 = \widehat{\beta}_{IV2}$. A couple of observations are useful here:

- Redoing the test the other way around (compute $\hat{c} = y_1 - \hat{\beta}_{IV2} y_2$, and then test whether $\text{cov}(z_1, \hat{c})$ is equal to zero) yields the same answer. The reason is that all we can really test is whether the two IVs $z_1$ and $z_2$ yield similar estimates of $\beta$. If so, this means both IVs are OK under the assumption that at least one of them is valid. If both of them are invalid ($\text{cov}(z_1, e) \neq 0$ and $\text{cov}(z_2, e) \neq 0$) we are out of luck and the test has zero power.

- When both IVs yield the same estimates, i.e. when $\hat{\beta} = \widehat{\theta}_1/\widehat{\pi}_1 = \widehat{\theta}_2/\widehat{\pi}_2$, this means that we can perfectly “fit” the four reduced form parameters with the three underlying parameters, $\beta$, $\pi_1$, and $\pi_2$. So testing whether the overidentification restrictions are satisfied simply means testing whether the restrictions $\theta_1 = \beta \pi_1$ and $\theta_2 = \beta \pi_2$ in the reduced form are satisfied.

As it turns out, a computationally convenient way of testing whether the overidentification restrictions are satisfied is to follow a simple two-step procedure. In step 1, estimate the model by TSLS (or an alternative method) and keep the estimated residual $\hat{c}$. In step 2, run a regression of $\hat{c}$ on $z_1$ and $z_2$ (and on other $x$ variables if they are in the model) and keep the R-square. Under the null hypothesis that overidentification restrictions are satisfied, we should have $n R^2 \sim \chi^2(q)$, where $q$ is the number of overidentification restrictions (1 in the case discussed here).

### 2.2 Hausman test

Another natural question to ask is whether instrumenting does matter in the sense that TSLS/IV estimates are statistically different from OLS estimates. Hausman and others (Durbin, Wu, there is a debate about whom first “invented” this test) have suggested a simple way of testing for differences between the two estimators. Since doing IV comes at the price a less precision in the estimates, one could argue for using OLS instead of IV unless it can be shown that doing IV significantly changes the results. One standard formula for the Hausman test is:
\[ H = \frac{(\hat{\beta}_{\text{TSLS}} - \hat{\beta}_{\text{OLS}})^2}{\hat{V}(\hat{\beta}_{\text{TSLS}}) - \hat{V}(\hat{\beta}_{\text{OLS}})} \]  

where \( \hat{V}(\hat{\beta}_{\text{TSLS}}) \) and \( \hat{V}(\hat{\beta}_{\text{OLS}}) \) are the estimated variances of the TSLS and OLS estimates, respectively. Under the null hypothesis that OLS is consistent, \( H \) follows a \( \chi^2 \) distribution with one degree of freedom. Though this form of the Hausman test is easily computable, another convenient way of testing the same hypothesis is to estimate the following regression by OLS:

\[ y_1 = \beta y_2 + \gamma \hat{u}_2 + \varepsilon, \]  

where \( \hat{u}_2 \) is the residual from the reduced form (first-stage) equation for \( y_2 \), and do a simple t-test to see whether the estimate of \( \gamma \) is significantly different from zero. One advantage of this test is that the heteroskedasticity robust version of the test is easily obtained by correcting the standard errors for arbitrary forms of heteroskedasticity using White’s formula (“robust” option in Stata, more on this in lecture notes #3). This version of the test is also easily generalized to the case where we have several potentially endogenous regressors \( y_2, y_3, \ldots, y_k \), and a vector of other regressors \( x \) by performing a joint test that \( \gamma_2 = \gamma_3 = \ldots = \gamma_k = 0 \) in the model

\[ y_1 = \beta_2 y_2 + \beta_3 y_3 + \ldots + \beta_k y_k + \gamma_2 \hat{u}_2 + \gamma_3 \hat{u}_3 + \ldots + \gamma_k \hat{u}_k + \delta x + \varepsilon \]  

### 2.3 Weak instruments

As mentioned above, it is possible to directly test whether \( \text{cov}(z, y_2) \neq 0 \) since both \( z \) and \( y_2 \) are observed. Remember that in the simplest case we have the reduced form equation \( y_2 = \pi z + u_2 \) where \( \pi = \text{cov}(y_2, z)/\text{var}(z) \) is regression coefficient. We can thus test whether \( \text{cov}(z, y_2) \neq 0 \) by simply testing whether the estimate \( \hat{\pi} \) of \( \pi \) is significantly different from zero. When the resulting t-statistic is low, we typically say that \( z \) is a weak instrument in the sense that it does not predict \( y_2 \) very well.

The reason why weak instruments are a serious problem is that the IV/TSLS can be expressed as a ratio in which \( \text{cov}(z, y_2) \) or \( \pi = \text{cov}(y_2, z)/\text{var}(z) \) is the denominator. For example, when looking at the ratio of reduced form coefficients we have

\[ \hat{\beta}_{\text{RRF}} = \frac{\hat{\theta}}{\hat{\pi}} \]  

If we cannot rule out that the denominator \( \hat{\pi} \) is equal to zero (which is what happens
when the t-statistic is below 2), this means that we cannot rule out that the ratio is equal
to plus or minus infinity (depending on the value of $\hat{\theta}$). In technical terms, this means
that the confidence intervals for $\beta$ are unbounded and that the estimate is uninformative
about the true value of $\beta$.

There is a rich literature about weak instruments. In terms of econometric practice,
it is mostly important to stress that it is very important to test whether the instruments
are significant predictors of the potentially endogenous regressor ($y_2$ here). This is why
researchers often report the results of this test along with coefficient estimates in tables
that show TSLS or other version of IV estimators.

3 Stata commands

The basic command to perform two-stage least-squares and IV regressions in Stata is
“ivregress”. For example, with one instrumental variable $z$, the dependent variable $y_1$,
the potentially endogenous covariate $y_2$, and three other covariates $x_1$, $x_2$, and $x_3$, just
use the command

\[
\text{ivregress 2sls y1 (y2=z) x1 x2 x3}
\]
to get TSLS estimates. In this setting, it is straightforward to perform a test of the
significance of the instrument in the first stage regression:

\[
\text{reg y2 z x1 x2 x3}
\]

and looking at the t-statistic on $z$.

The Hausman test can be computed manually by first extracting the residual from
the first stage equation, and including it as a regressor in the main model as follows:

\[
\text{reg y2 z x1 x2 x3}
\]

\[
\text{predict u2, residual}
\]

\[
\text{reg y1 y2 x1 x2 x3 u2}
\]

and testing whether $u_2$ is statistically significant.

An alternative approach is to use the “hausman” procedure in Stata. The Hausman
specification test is obtained by comparing IV/TSLS estimates to OLS estimates. OLS
estimates need to be stored before running IV/TSLS estimation. In the example above,
the commands are:
reg y1 y2 x1 x2 x3

estimates store ols_estimate

ivregress 2sls y1 (y2=z) x1 x2 x3

hausman . ols_estimate

Note that in the “hausman” procedure in Stata, we always have to list the consistent (TSLS here) and then the efficient (OLS) estimator. The “.” just means we are using the latest estimates (TSLS) as the consistent estimate. Note also that “ols_estimate” is just an arbitrary name we use to store the estimates.

Now let’s look at the case where the model is overidentified when we have two instrumental variables \( z_1 \) and \( z_2 \). The overidentification test can be done manually by running two-stage least-squares, extracting the residual, and running a regression of the residual on the instrumental variables and the regressors:

\[
\text{ivregress 2sls y1 \( (y2=z1 \ z2) \) x1 x2 x3}
\]

\[
\text{predict res1, residual}
\]

\[
\text{reg res1 z1 z2 x1 x2 x3}
\]

Then just multiply the R-square on the output with the number of observations. This statistic is distributed Chi-square with one degree of freedom.

A simpler approach is to use the post-estimation command “estat overid”. In this case, you just do

\[
\text{ivregress 2sls y1 \( (y2=z1 \ z2) \) x1 x2 x3}
\]

\[
\text{estat overid}
\]