redundant instruments. The third (12.7) is often called the **relevance condition** and is essential for the identification of the model, as we discuss later. A necessary condition for (12.7) is that $\ell \geq k$.

Condition (12.5) – that the instruments are uncorrelated with the equation error – is often described as that they are **exogenous** in the sense that they are determined outside the model for Y.

Notice that the regressors X_1 satisfy condition (12.5) and thus should be included as instrumental variables. They are therefore a subset of the variables Z. Notationally we make the partition

$$Z = \left(\begin{array}{c} Z_1 \\ Z_2 \end{array} \right) = \left(\begin{array}{c} X_1 \\ Z_2 \end{array} \right) \begin{array}{c} k_1 \\ \ell_2 \end{array} . \tag{12.8}$$

Here, $X_1 = Z_1$ are the **included exogenous variables** and Z_2 are the **excluded exogenous variables**. That is, Z_2 are variables which could be included in the equation for Y (in the sense that they are uncorrelated with e) yet can be excluded as they have true zero coefficients in the equation. With this notation we can also write the structural equation (12.4) as

$$Y_1 = Z_1'\beta_1 + Y_2'\beta_2 + e. \tag{12.9}$$

This is useful notation as it clarifies that the variable Z_1 is exogenous and the variable Y_2 is endogenous.

Many authors describe Z_1 as the "exogenous variables", Y_2 as the "endogenous variables", and Z_2 as the "instrumental variables".

We say that the model is **just-identified** if $\ell = k$ and **over-identified** if $\ell > k$.

What variables can be used as instrumental variables? From the definition $\mathbb{E}[Ze] = 0$ the instrument must be uncorrelated with the equation error, meaning that it is excluded from the structural equation as mentioned above. From the rank condition (12.7) it is also important that the instrumental variables be correlated with the endogenous variables Y_2 after controlling for the other exogenous variables Z_1. These two requirements are typically interpreted as requiring that the instruments be determined outside the system for $\vec{Y}$, causally determine Y_2, but do not causally determine Y_1 except through Y_2.

Let's take the three examples given above.

Measurement error in the regressor. When X is a mis-measured version of Z a common choice for an instrument Z_2 is an alternative measurement of Z. For this Z_2 to satisfy the property of an instrumental variable the measurement error in Z_2 must be independent of that in X.

Supply and Demand. An appropriate instrument for price P in a demand equation is a variable Z_2 which influences supply but not demand. Such a variable affects the equilibrium values of P and Q but does not directly affect price except through quantity. Variables which affect supply but not demand are typically related to production costs.

An appropriate instrument for price in a supply equation is a variable which influences demand but not supply. Such a variable affects the equilibrium values of price and quantity but only affects price through quantity.

Choice Variable as Regressor. An ideal instrument affects the choice of the regressor (education) but does not directly influence the dependent variable (wages) except through the indirect effect on the regressor. We will discuss an example in the next section.

12.6 Example: College Proximity

In a influential paper David Card (1995) suggested if a potential student lives close to a college this reduces the cost of attendence and thereby raises the likelihood that the student will attend college. However, college proximity does not directly affect a student's skills or abilities so should not have a direct effect on his or her market wage. These considerations suggest that college proximity can be used as an instrument for education in a wage regression. We use the simplest model reported in Card's paper to illustrate the concepts of instrumental variables throughout the chapter.

Card used data from the National Longitudinal Survey of Young Men (NLSYM) for 1976. A baseline least squares wage regression for his data set is reported in the first column of Table 12.1. The dependent variable is the log of weekly earnings. The regressors are *education* (years of schooling), *experience* (years of work experience, calculated as *age* (years) less *education*+6), *experience*^{2}/100, *Black*, *south* (an indicator for residence in the southern region of the U.S.), and *urban* (an indicator for residence in a standard metropolitan statistical area). We drop observations for which *wage* is missing. The remaining sample has 3,010 observations. His data is the file `Card1995` on the textbook website.

The point estimate obtained by least squares suggests an 7% increase in earnings for each year of education.

Table 12.1: Instrumental Variable Wage Regressions

	OLS	IV(a)	IV(b)	2SLS(a)	2SLS(b)	LIML
education	0.074	0.132	0.133	0.161	0.160	0.164
	(0.004)	(0.049)	(0.051)	(0.040)	(0.041)	(0.042)
experience	0.084	0.107	0.056	0.119	0.047	0.120
	(0.007)	(0.021)	(0.026)	(0.018)	(0.025)	(0.019)
experience^2/100	−0.224	−0.228	−0.080	−0.231	−0.032	−0.231
	(0.032)	(0.035)	(0.133)	(0.037)	(0.127)	(0.037)
Black	−0.190	−0.131	−0.103	−0.102	−0.064	−0.099
	(0.017)	(0.051)	(0.075)	(0.044)	(0.061)	(0.045)
south	−0.125	−0.105	−0.098	−0.095	−0.086	−0.094
	(0.015)	(0.023)	(0.0284)	(0.022)	(0.026)	(0.022)
urban	0.161	0.131	0.108	0.116	0.083	0.115
	(0.015)	(0.030)	(0.049)	(0.026)	(0.041)	(0.027)
Sargan				0.82	0.52	0.82
p-value				0.37	0.47	0.37

Notes:

1. IV(a) uses *college* as an instrument for *education*.
2. IV(b) uses *college*, *age*, and *age*^{2}/100 as instruments for *education*, *experience*, and *experience*^{2}/100.
3. 2SLS(a) uses *public* and *private* as instruments for *education*.
4. 2SLS(b) uses *public*, *private*, *age*, and *age*^{2} as instruments for *education*, *experience*, and *experience*^{2}/100.
5. LIML uses *public* and *private* as instruments for *education*.

As discussed in the previous sections it is reasonable to view years of education as a choice made by an individual and thus is likely endogenous for the structural return to education. This means that least squares is an estimate of a linear projection but is inconsistent for coefficient of a structural equation representing the causal impact of years of education on expected wages. Labor economics predicts that ability, education, and wages will be positively correlated. This suggests that the population projection coefficient estimated by least squares will be higher than the structural parameter (and hence upwards biased). However, the sign of the bias is uncertain since there are multiple regressors and there are other potential sources of endogeneity.

To instrument for the endogeneity of education, Card suggested that a reasonable instrument is a dummy variable indicating if the individual grew up near a college. We will consider three measures:

college	Grew up in same county as a 4-year college
public	Grew up in same county as a 4-year public college
private	Grew up in same county as a 4-year private college.

12.7 Reduced Form

The reduced form is the relationship between the endogenous regressors Y_2 and the instruments Z. A linear reduced form model for Y_2 is

$$Y_2 = \Gamma' Z + u_2 = \Gamma_{12}' Z_1 + \Gamma_{22}' Z_2 + u_2 \tag{12.10}$$

This is a multivariate regression as introduced in Chapter 11. The $\ell \times k_2$ coefficient matrix Γ is defined by linear projection:

$$\Gamma = \mathbb{E}\left[ZZ'\right]^{-1} \mathbb{E}\left[ZY_2'\right] \tag{12.11}$$

This implies $\mathbb{E}\left[Zu_2'\right] = 0$. The projection coefficient (12.11) is well defined and unique under (12.6).

We also construct the reduced form for Y_1. Substitute (12.10) into (12.9) to obtain

$$\begin{aligned} Y_1 &= Z_1'\beta_1 + \left(\Gamma_{12}' Z_1 + \Gamma_{22}' Z_2 + u_2\right)' \beta_2 + e \\ &= Z_1'\lambda_1 + Z_2'\lambda_2 + u_1 \qquad (12.12) \\ &= Z'\lambda + u_1 \qquad (12.13) \end{aligned}$$

where

$$\begin{aligned} \lambda_1 &= \beta_1 + \Gamma_{12}\beta_2 \qquad (12.14) \\ \lambda_2 &= \Gamma_{22}\beta_2 \qquad (12.15) \\ u_1 &= u_2'\beta_2 + e. \end{aligned}$$

We can also write

$$\lambda = \overline{\Gamma}\beta \tag{12.16}$$

where

$$\overline{\Gamma} = \left[\begin{array}{cc} \boldsymbol{I}_{k_1} & \Gamma_{12} \\ 0 & \Gamma_{22} \end{array}\right] = \left[\begin{array}{cc} \boldsymbol{I}_{k_1} & \Gamma \\ 0 & \end{array}\right].$$

Together, the reduced form equations for the system are

$$\begin{aligned} Y_1 &= \lambda' Z + u_1 \\ Y_2 &= \Gamma' Z + u_2. \end{aligned}$$

or

$$\vec{Y} = \left[\begin{array}{cc} \lambda_1' & \lambda_2' \\ \Gamma_{12}' & \Gamma_{22}' \end{array}\right] Z + u \tag{12.17}$$

where $u = (u_1, u_2)$.

The relationships (12.14)-(12.16) are critically important for understanding the identification of the structural parameters β_1 and β_2, as we discuss below. These equations show the tight relationship between the structural parameters (β_1 and β_2) and the reduced form parameters (Γ and λ).

The reduced form equations are projections so the coefficients may be estimated by least squares (see Chapter 11). The least squares estimators of (12.11) and (12.13) are

$$\widehat{\Gamma} = \left(\sum_{i=1}^{n} Z_i Z_i' \right)^{-1} \left(\sum_{i=1}^{n} Z_i Y_{2i}' \right) \tag{12.18}$$

$$\widehat{\lambda} = \left(\sum_{i=1}^{n} Z_i Z_i' \right)^{-1} \left(\sum_{i=1}^{n} Z_i Y_{1i} \right) \tag{12.19}$$

12.8 Identification

A parameter is **identified** if it is a unique function of the probability distribution of the observables. One way to show that a parameter is identified is to write it as an explicit function of population moments. For example, the reduced form coefficient matrices Γ and λ are identified since they can be written as explicit functions of the moments of the variables (Y, X, Z). That is,

$$\Gamma = \mathbb{E}\left[ZZ'\right]^{-1} \mathbb{E}\left[ZY_2'\right] \tag{12.20}$$

$$\lambda = \mathbb{E}\left[ZZ'\right]^{-1} \mathbb{E}\left[ZY_1\right]. \tag{12.21}$$

These are uniquely determined by the probability distribution of (Y_1, Y_2, Z) if Definition 12.1 holds, since this includes the requirement that $\mathbb{E}\left[ZZ'\right]$ is invertible.

We are interested in the structural parameter β. It relates to (λ, Γ) through (12.16). β is identified if it uniquely determined by this relation. This is a set of ℓ equations with k unknowns with $\ell \geq k$. From linear algebra we know that there is a unique solution if and only if $\overline{\Gamma}$ has full rank k.

$$\operatorname{rank}\left(\overline{\Gamma}\right) = k. \tag{12.22}$$

Under (12.22) β can be uniquely solved from (12.16). If (12.22) fails then (12.16) has fewer equations than coefficients so there is not a unique solution.

We can write $\overline{\Gamma} = \mathbb{E}\left[ZZ'\right]^{-1} \mathbb{E}\left[ZX'\right]$. Combining this with (12.16) we obtain

$$\mathbb{E}\left[ZZ'\right]^{-1} \mathbb{E}\left[ZY_1\right] = \mathbb{E}\left[ZZ'\right]^{-1} \mathbb{E}\left[ZX'\right] \beta$$

or

$$\mathbb{E}\left[ZY_1\right] = \mathbb{E}\left[ZX'\right] \beta$$

which is a set of ℓ equations with k unknowns. This has a unique solution if (and only if)

$$\operatorname{rank}\left(\mathbb{E}\left[ZX'\right]\right) = k \tag{12.23}$$

which was listed in (12.7) as a condition of Definition 12.1. (Indeed, this is why it was listed as part of the definition.) We can also see that (12.22) and (12.23) are equivalent ways of expressing the same requirement. If this condition fails then β will not be identified. The condition (12.22)-(12.23) is called the **relevance condition**.

It is useful to have explicit expressions for the solution β. The easiest case is when $\ell = k$. Then (12.22) implies $\overline{\Gamma}$ is invertible so the structural parameter equals $\beta = \overline{\Gamma}^{-1} \lambda$. It is a unique solution because $\overline{\Gamma}$ and λ are unique and $\overline{\Gamma}$ is invertible.

When $\ell > k$ we can solve for β by applying least squares to the system of equations $\lambda = \overline{\Gamma}\beta$. This is ℓ equations with k unknowns and no error. The least squares solution is $\beta = \left(\overline{\Gamma}'\overline{\Gamma}\right)^{-1}\overline{\Gamma}'\lambda$. Under (12.22) the matrix $\overline{\Gamma}'\overline{\Gamma}$ is invertible so the solution is unique.

β is identified if $\operatorname{rank}(\overline{\Gamma}) = k$, which is true if and only if $\operatorname{rank}(\Gamma_{22}) = k_2$ (by the upper-diagonal structure of $\overline{\Gamma}$). Thus the key to identification of the model rests on the $\ell_2 \times k_2$ matrix Γ_{22} in (12.10). To see this, recall the reduced form relationships (12.14)-(12.15). We can see that β_2 is identified from (12.15) alone, and the necessary and sufficient condition is $\operatorname{rank}(\Gamma_{22}) = k_2$. If this is satisfied then the solution can be written as $\beta_2 = \left(\Gamma_{22}'\Gamma_{22}\right)^{-1}\Gamma_{22}'\lambda_2$. Then β_1 is identified from this and (12.14), with the explicit solution $\beta_1 = \lambda_1 - \Gamma_{12}\left(\Gamma_{22}'\Gamma_{22}\right)^{-1}\Gamma_{22}'\lambda_2$. In the just-identified case ($\ell_2 = k_2$) these equations simplify to take the form $\beta_2 = \Gamma_{22}^{-1}\lambda_2$ and $\beta_1 = \lambda_1 - \Gamma_{12}\Gamma_{22}^{-1}\lambda_2$.

12.9 Instrumental Variables Estimator

In this section we consider the special case where the model is just-identified so that $\ell = k$.

The assumption that Z is an instrumental variable implies that $\mathbb{E}[Ze] = 0$. Making the substitution $e = Y_1 - X'\beta$ we find $\mathbb{E}\left[Z\left(Y_1 - X'\beta\right)\right] = 0$. Expanding,

$$\mathbb{E}[ZY_1] - \mathbb{E}\left[ZX'\right]\beta = 0.$$

This is a system of $\ell = k$ equations and k unknowns. Solving for β we find

$$\beta = \left(\mathbb{E}\left[ZX'\right]\right)^{-1}\mathbb{E}[ZY_1].$$

This requires that the matrix $\mathbb{E}\left[ZX'\right]$ is invertible, which holds under (12.7) or equivalently (12.23).

The **instrumental variables (IV)** estimator β replaces population by sample moments. We find

$$\begin{aligned}\widehat{\beta}_{\mathrm{iv}} &= \left(\frac{1}{n}\sum_{i=1}^{n} Z_i X_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i Y_{1i}\right)\\ &= \left(\sum_{i=1}^{n} Z_i X_i'\right)^{-1}\left(\sum_{i=1}^{n} Z_i Y_{1i}\right). \qquad (12.24)\end{aligned}$$

More generally, given any variable $W \in \mathbb{R}^k$ it is common to refer to the estimator

$$\widehat{\beta}_{\mathrm{iv}} = \left(\sum_{i=1}^{n} W_i X_i'\right)^{-1}\left(\sum_{i=1}^{n} W_i Y_{1i}\right)$$

as the IV estimator for β using the instrument W.

Alternatively, recall that when $\ell = k$ the structural parameter can be written as a function of the reduced form parameters as $\beta = \overline{\Gamma}^{-1}\lambda$. Replacing $\overline{\Gamma}$ and λ by their least squares estimators (12.18)-(12.19) we can construct what is called the **Indirect Least Squares (ILS)** estimator. Using the matrix algebra representations

$$\begin{aligned}\widehat{\beta}_{\mathrm{ils}} &= \widehat{\overline{\Gamma}}^{-1}\widehat{\lambda}\\ &= \left(\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{X}\right)\right)^{-1}\left(\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{Y}_1\right)\right)\\ &= \left(\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{Y}_1\right)\\ &= \left(\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{Y}_1\right).\end{aligned}$$

We see that this equals the IV estimator (12.24). Thus the ILS and IV estimators are identical.

Given the IV estimator we define the residual $\widehat{e}_i = Y_{1i} - X_i'\widehat{\beta}_{\mathrm{iv}}$. It satisfies

$$\boldsymbol{Z}'\widehat{\boldsymbol{e}} = \boldsymbol{Z}'\boldsymbol{Y}_1 - \boldsymbol{Z}'\boldsymbol{X}\left(\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{Y}_1\right) = 0. \tag{12.25}$$

Since Z includes an intercept this means that the residuals sum to zero and are uncorrelated with the included and excluded instruments.

To illustrate IV regression we estimate the reduced form equations for college proximity, now treating *education* as endogenous and using *college* as an instrumental variable. The reduced form equations for log*(wage)* and *education* are reported in the first and second columns of Table 12.2.

Table 12.2: Reduced Form Regressions

	log(wage)	education	education	experience	experience^2/100	education
experience	0.053	−0.410				−0.413
	(0.007)	(0.032)				(0.032)
experience^2/100	−0.219	0.073				0.093
	(0.033)	(0.170)				(0.171)
black	−0.264	−1.006	−1.468	1.468	0.282	−1.006
	(0.018)	(0.088)	(0.115)	(0.115)	(0.026)	(0.088)
south	−0.143	−0.291	−0.460	0.460	0.112	−0.267
	(0.017)	(0.078)	(0.103)	(0.103)	(0.022)	(0.079)
urban	0.185	0.404	0.835	−0.835	−0.176	0.400
	(0.017)	(0.085)	(0.112)	(0.112)	(0.025)	(0.085)
college	0.045	0.337	0.347	−0.347	−0.073	
	(0.016)	(0.081)	(0.109)	(0.109)	(0.023)	
public						0.430
						(0.086)
private						0.123
						(0.101)
age			1.061	−0.061	−0.555	
			(0.296)	(0.296)	(0.065)	
age^2/100			−1.876	1.876	1.313	
			(0.516)	(0.516)	(0.116)	
F		17.51	8.22	1581	1112	13.87

Of particular interest is the equation for the endogenous regressor *education*, and the coefficients for the excluded instruments – in this case *college*. The estimated coefficient equals 0.337 with a small standard error. This implies that growing up near a 4-year college increases average educational attainment by 0.3 years. This seems to be a reasonable magnitude.

Since the structural equation is just-identified with one right-hand-side endogenous variable the ILS/IV estimate for the education coefficient is the ratio of the coefficient estimates for the instrument *college* in the two equations, e.g. 0.045/0.337 = 0.13, implying a 13% return to each year of education. This is substantially greater than the 7% least squares estimate from the first column of Table 12.1. The IV estimates of the full equation are reported in the second column of Table 12.1. One first reaction is surprise that the IV estimate is larger than the OLS estimate. The endogeneity of educational choice should lead to upward bias in the OLS estimator, which predicts that the IV estimate should have been smaller than the OLS estimator. An alternative explanation may be needed. One possibility is heterogeneous education effects (when the education coefficient β is heterogenous across individuals). In Section 12.34

we show that in this context the IV estimator picks up this treatment effect for a subset of the population, and this may explain why IV estimation results in a larger estimated coefficient on education.

Card (1995) also points out that if *education* is endogenous then so is our measure of *experience* since it is calculated by subtracting *education* from *age*. He suggests that we can use the variables *age* and *age*^{2} as instruments for *experience* and *experience*^{2}. The age variables are exogenous (not choice variables) yet highly correlated with *experience* and *experience*^{2}. Notice that this approach treats *experience*^{2} as a variable separate from *experience*. Indeed, this is the correct approach.

Following this recommendation we now have three endogenous regressors and three instruments. We present the three reduced form equations for the three endogenous regressors in the third through fifth columns of Table 12.2. It is interesting to compare the equations for *education* and *experience*. The two sets of coefficients are simply the sign change of the other with the exception of the coefficient on *age*. Indeed this must be the case because the three variables are linearly related. Does this cause a problem for 2SLS? Fortunately, no. The fact that the coefficient on age is not simply a sign change means that the equations are not linearly singular. Hence Assumption (12.22) is not violated.

The IV estimates using the three instruments *college*, *age*, and *age*^{2} for the endogenous regressors *education*, *experience*, and *experience*^{2} is presented in the third column of Table 12.1. The estimate of the returns to schooling is not affected by this change in the instrument set, but the estimated return to experience profile flattens (the quadratic effect diminishes).

The IV estimator may be calculated in Stata using the `ivregress 2sls` command.

12.10 Demeaned Representation

Does the well-known demeaned representation for linear regression (3.18) carry over to the IV estimator? To see this, write the linear projection equation in the format $Y_1 = X'\beta + \alpha + e$ where α is the intercept and X does not contain a constant. Similarly, partition the instrument as $(1, Z)$ where Z does not contain an intercept. We can write the IV estimator for the i^{th} equation as

$$Y_{1i} = X_i'\widehat{\beta}_{\mathrm{iv}} + \widehat{\alpha}_{\mathrm{iv}} + \widehat{e}_i.$$

The orthogonality (12.25) implies the two-equation system

$$\begin{aligned}\sum_{i=1}^{n}\left(Y_{1i} - X_i'\widehat{\beta}_{\mathrm{iv}} - \widehat{\alpha}_{\mathrm{iv}}\right) &= 0 \\ \sum_{i=1}^{n} Z_i\left(Y_{1i} - X_i'\widehat{\beta}_{\mathrm{iv}} - \widehat{\alpha}_{\mathrm{iv}}\right) &= 0.\end{aligned}$$

The first equation implies $\widehat{\alpha}_{\mathrm{iv}} = \overline{Y_1} - \overline{X}'\widehat{\beta}_{\mathrm{iv}}$. Substituting into the second equation

$$\sum_{i=1}^{n} Z_i\left(\left(Y_{1i} - \overline{Y_1}\right) - \left(X_i - \overline{X}\right)'\widehat{\beta}_{\mathrm{iv}}\right)$$

and solving for $\widehat{\beta}_{\mathrm{iv}}$ we find

$$\begin{aligned}\widehat{\beta}_{\mathrm{iv}} &= \left(\sum_{i=1}^{n} Z_i\left(X_i - \overline{X}\right)'\right)^{-1}\left(\sum_{i=1}^{n} Z_i\left(Y_{1i} - \overline{Y}_1\right)\right) \\ &= \left(\sum_{i=1}^{n}\left(Z_i - \overline{Z}\right)\left(X_i - \overline{X}\right)'\right)^{-1}\left(\sum_{i=1}^{n}\left(Z_i - \overline{Z}\right)\left(Y_{1i} - \overline{Y}_1\right)\right). \qquad (12.26)\end{aligned}$$

Thus the demeaning equations for least squares carry over to the IV estimator. The coefficient estimator $\widehat{\beta}_{\mathrm{iv}}$ is a function only of the demeaned data.

12.11 Wald Estimator

In many cases including the Card proximity example the excluded instrument is a binary (dummy) variable. Let's focus on that case and suppose that the model has just one endogenous regressor and no other regressors beyond the intercept. The model can be written as $Y = X\beta + \alpha + e$ with $\mathbb{E}[e \mid Z] = 0$ and Z binary.

Take expectations of the structural equation given $Z = 1$ and $Z = 0$, respectively. We obtain

$$\mathbb{E}[Y \mid Z = 1] = \mathbb{E}[X \mid Z = 1]\,\beta + \alpha$$
$$\mathbb{E}[Y \mid Z = 0] = \mathbb{E}[X \mid Z = 0]\,\beta + \alpha.$$

Subtracting and dividing we obtain an expression for the slope coefficient

$$\beta = \frac{\mathbb{E}[Y \mid Z = 1] - \mathbb{E}[Y \mid Z = 0]}{\mathbb{E}[X \mid Z = 1] - \mathbb{E}[X \mid Z = 0]}. \tag{12.27}$$

The natural moment estimator replaces the expectations by the averages within the "grouped data" where $Z_i = 1$ and $Z_i = 0$, respectively. That is, define the group means

$$\overline{Y}_1 = \frac{\sum_{i=1}^n Z_i Y_i}{\sum_{i=1}^n Z_i}, \qquad \overline{Y}_0 = \frac{\sum_{i=1}^n (1 - Z_i)\, Y_i}{\sum_{i=1}^n (1 - Z_i)}$$
$$\overline{X}_1 = \frac{\sum_{i=1}^n Z_i X_i}{\sum_{i=1}^n Z_i}, \qquad \overline{X}_0 = \frac{\sum_{i=1}^n (1 - Z_i)\, X_i}{\sum_{i=1}^n (1 - Z_i)}$$

and the moment estimator

$$\widehat{\beta} = \frac{\overline{Y}_1 - \overline{Y}_0}{\overline{X}_1 - \overline{X}_0}. \tag{12.28}$$

This is the "Wald estimator" of Wald (1940).

These expressions are rather insightful. (12.27) shows that the structural slope coefficient is the expected change in Y due to changing the instrument divided by the expected change in X due to changing the instrument. Informally, it is the change in Y (due to Z) over the change in X (due to Z). Equation (12.28) shows that the slope coefficient can be estimated by a the ratio of a difference in means.

The expression (12.28) may appear like a distinct estimator from the IV estimator $\widehat{\beta}_{\text{iv}}$ but it turns out that they are the same. That is, $\widehat{\beta} = \widehat{\beta}_{\text{iv}}$. To see this, use (12.26) to find

$$\widehat{\beta}_{\text{iv}} = \frac{\sum_{i=1}^n Z_i \left(Y_i - \overline{Y}\right)}{\sum_{i=1}^n Z_i \left(X_i - \overline{X}\right)} = \frac{\overline{Y}_1 - \overline{Y}}{\overline{X}_1 - \overline{X}}.$$

Then notice

$$\overline{Y}_1 - \overline{Y} = \overline{Y}_1 - \left(\frac{1}{n}\sum_{i=1}^n Z_i \overline{Y}_1 + \frac{1}{n}\sum_{i=1}^n (1 - Z_i)\,\overline{Y}_0\right) = \left(1 - \overline{Z}\right)\left(\overline{Y}_1 - \overline{Y}_0\right)$$

and similarly

$$\overline{X}_1 - \overline{X} = \left(1 - \overline{Z}\right)\left(\overline{X}_1 - \overline{X}_0\right)$$

and hence

$$\widehat{\beta}_{\text{iv}} = \frac{\left(1 - \overline{Z}\right)\left(\overline{Y}_1 - \overline{Y}_0\right)}{\left(1 - \overline{Z}\right)\left(\overline{X}_1 - \overline{X}_0\right)} = \widehat{\beta}$$

as defined in (12.28). Thus the Wald estimator equals the IV estimator.

We can illustrate using the Card proximity example. If we estimate a simple IV model with no covariates we obtain the estimate $\widehat{\beta}_{\text{iv}} = 0.19$. If we estimate the group-mean of log wages and education based on the instrument *college* we find

	near college	not near college	difference
log(wage)	6.311	6.156	0.155
education	13.527	12.698	0.829
ratio			0.19

Based on these estimates the Wald estimator of the slope coefficient is (6.311 − 6.156) / (13.527 − 12.698) = $0.155/0.829 = 0.19$, the same as the IV estimator.

12.12 Two-Stage Least Squares

The IV estimator described in the previous section presumed $\ell = k$. Now we allow the general case of $\ell \geq k$. Examining the reduced-form equation (12.13) we see

$$\begin{aligned} Y_1 &= Z'\overline{\Gamma}\beta + u_1 \\ \mathbb{E}\left[Zu_1\right] &= 0. \end{aligned}$$

Defining $W = \overline{\Gamma}'Z$ we can write this as

$$\begin{aligned} Y_1 &= W'\beta + u_1 \\ \mathbb{E}\left[Wu_1\right] &= 0. \end{aligned}$$

One way of thinking about this is that Z is set of candidate instruments. The instrument vector $W = \overline{\Gamma}'Z$ is a k-dimentional set of linear combinations.

Suppose that $\overline{\Gamma}$ were known. Then we would estimate β by least squares of Y_1 on $W = \overline{\Gamma}'Z$

$$\widehat{\beta} = \left(\boldsymbol{W}'\boldsymbol{W}\right)^{-1}\left(\boldsymbol{W}'\boldsymbol{Y}\right) = \left(\overline{\Gamma}'\boldsymbol{Z}'\boldsymbol{Z}\overline{\Gamma}\right)^{-1}\left(\overline{\Gamma}'\boldsymbol{Z}'\boldsymbol{Y}_1\right).$$

While this is infeasible we can estimate $\overline{\Gamma}$ from the reduced form regression. Replacing $\overline{\Gamma}$ with its estimator $\widehat{\Gamma} = \left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{X}\right)$ we obtain

$$\begin{aligned} \widehat{\beta}_{\text{2sls}} &= \left(\widehat{\Gamma}'\boldsymbol{Z}'\boldsymbol{Z}\widehat{\Gamma}\right)^{-1}\left(\widehat{\Gamma}'\boldsymbol{Z}'\boldsymbol{Y}_1\right) \\ &= \left(\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-\mathbf{1}}\boldsymbol{Z}^{'}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{Y}_1 \\ &= \left(\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{Y}_1. \end{aligned} \tag{12.29}$$

This is called the **two-stage-least squares (2SLS)** estimator. It was originally proposed by Theil (1953) and Basmann (1957) and is a standard estimator for linear equations with instruments.

If the model is just-identified, so that $k = \ell$, then 2SLS simplifies to the IV estimator of the previous section. Since the matrices $\boldsymbol{X}'\boldsymbol{Z}$ and $\boldsymbol{Z}'\boldsymbol{X}$ are square we can factor

$$\begin{aligned} \left(\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1} &= \left(\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\left(\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{Z}\right)^{-1} \\ &= \left(\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)\left(\boldsymbol{X}'\boldsymbol{Z}\right)^{-1}. \end{aligned}$$

(Once again, this only works when $k=\ell$.) Then

$$\begin{aligned}\widehat{\beta}_{2\text{sls}} &= \left(\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{Y}_1\\ &= \left(\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)\left(\boldsymbol{X}'\boldsymbol{Z}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{Y}_1\\ &= \left(\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{Y}_1\\ &= \left(\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{Z}'\boldsymbol{Y}_1=\widehat{\beta}_{\text{iv}}\end{aligned}$$

as claimed. This shows that the 2SLS estimator as defined in (12.29) is a generalization of the IV estimator defined in (12.24).

There are several alternative representations of the 2SLS estimator which we now describe. First, defining the projection matrix

$$\boldsymbol{P}_{\boldsymbol{Z}}=\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}' \tag{12.30}$$

we can write the 2SLS estimator more compactly as

$$\widehat{\beta}_{2\text{sls}}=\left(\boldsymbol{X}'\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{Y}_1. \tag{12.31}$$

This is useful for representation and derivations but is not useful for computation as the $n\times n$ matrix $\boldsymbol{P}_{\boldsymbol{Z}}$ is too large to compute when n is large.

Second, define the fitted values for $\boldsymbol{X}$ from the reduced form $\widehat{\boldsymbol{X}}=\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X}=\boldsymbol{Z}\widehat{\Gamma}$. Then the 2SLS estimator can be written as

$$\widehat{\beta}_{2\text{sls}}=\left(\widehat{\boldsymbol{X}}'\boldsymbol{X}\right)^{-1}\widehat{\boldsymbol{X}}'\boldsymbol{Y}_1.$$

This is an IV estimator as defined in the previous section using $\widehat{X}$ as an instrument for X.

Third, since $\boldsymbol{P}_{\boldsymbol{Z}}$ is idempotent we can also write the 2SLS estimator as

$$\widehat{\beta}_{2\text{sls}}=\left(\boldsymbol{X}'\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{Y}_1=\left(\widehat{\boldsymbol{X}}'\widehat{\boldsymbol{X}}\right)^{-1}\widehat{\boldsymbol{X}}'\boldsymbol{Y}_1$$

which is the least squares estimator obtained by regressing Y_1 on the fitted values $\widehat{X}$.

This is the source of the "two-stage" name is since it can be computed as follows.

- Regress X on Z to obtain the fitted $\widehat{X}$: $\widehat{\Gamma}=\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{X}\right)$ and $\widehat{\boldsymbol{X}}=\boldsymbol{Z}\widehat{\Gamma}=\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X}$.

- Regress Y_1 on $\widehat{X}$: $\widehat{\beta}_{2\text{sls}}=\left(\widehat{\boldsymbol{X}}'\widehat{\boldsymbol{X}}\right)^{-1}\widehat{\boldsymbol{X}}'\boldsymbol{Y}_1$.

It is useful to scrutinize the projection $\widehat{\boldsymbol{X}}$. Recall, $\boldsymbol{X}=[\boldsymbol{Z}_1,\boldsymbol{Y}_2]$ and $\boldsymbol{Z}=[\boldsymbol{Z}_1,\boldsymbol{Z}_2]$. Notice $\widehat{\boldsymbol{X}}_1=\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{Z}_1=\boldsymbol{Z}_1$ since $\boldsymbol{Z}_1$ lies in the span of $\boldsymbol{Z}$. Then $\widehat{\boldsymbol{X}}=\left[\widehat{\boldsymbol{X}}_1,\widehat{\boldsymbol{Y}}_2\right]=\left[\boldsymbol{Z}_1,\widehat{\boldsymbol{Y}}_2\right]$. This shows that in the second stage we regress Y_1 on Z_1 and $\widehat{Y}_2$. This means that only the endogenous variables Y_2 are replaced by their fitted values, $\widehat{Y}_2=\widehat{\Gamma}'_{12}Z_1+\widehat{\Gamma}'_{22}Z_2$.

A fourth representation of 2SLS can be obtained using the FWL Theorem. The third representation and following discussion showed that 2SLS is obtained as least squares of Y_1 on the fitted values $(Z_1,\widehat{Y}_2)$. Hence the coefficient $\widehat{\beta}_2$ on the endogenous variables can be found by residual regression. Set $\boldsymbol{P}_1=\boldsymbol{Z}_1\left(\boldsymbol{Z}_1'\boldsymbol{Z}_1\right)^{-1}\boldsymbol{Z}_1'$. Applying the FWL theorem we obtain

$$\begin{aligned}\widehat{\beta}_2 &= \left(\widehat{\boldsymbol{Y}}_2'\left(\boldsymbol{I}_n-\boldsymbol{P}_1\right)\widehat{\boldsymbol{Y}}_2\right)^{-1}\left(\widehat{\boldsymbol{Y}}_2'\left(\boldsymbol{I}_n-\boldsymbol{P}_1\right)\boldsymbol{Y}_1\right)\\ &= \left(\boldsymbol{Y}_2'\boldsymbol{P}_{\boldsymbol{Z}}\left(\boldsymbol{I}_n-\boldsymbol{P}_1\right)\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{Y}_2\right)^{-1}\left(\boldsymbol{Y}_2'\boldsymbol{P}_{\boldsymbol{Z}}\left(\boldsymbol{I}_n-\boldsymbol{P}_1\right)\boldsymbol{Y}_1\right)\\ &= \left(\boldsymbol{Y}_2'\left(\boldsymbol{P}_{\boldsymbol{Z}}-\boldsymbol{P}_1\right)\boldsymbol{Y}_2\right)^{-1}\left(\boldsymbol{Y}_2'\left(\boldsymbol{P}_{\boldsymbol{Z}}-\boldsymbol{P}_1\right)\boldsymbol{Y}_1\right)\end{aligned}$$

since $\boldsymbol{P_Z P}_1 = \boldsymbol{P}_1$.

A fifth representation can be obtained by a further projection. The projection matrix $\boldsymbol{P_Z}$ can be replaced by the projection onto the pair $[\boldsymbol{Z}_1, \widetilde{\boldsymbol{Z}}_2]$ where $\widetilde{\boldsymbol{Z}}_2 = (\boldsymbol{I}_n - \boldsymbol{P}_1)\boldsymbol{Z}_2$ is $\boldsymbol{Z}_2$ projected orthogonal to $\boldsymbol{Z}_1$. Since $\boldsymbol{Z}_1$ and $\widetilde{\boldsymbol{Z}}_2$ are orthogonal, $\boldsymbol{P_Z} = \boldsymbol{P}_1 + \boldsymbol{P}_2$ where $\boldsymbol{P}_2 = \widetilde{\boldsymbol{Z}}_2\left(\widetilde{\boldsymbol{Z}}_2'\widetilde{\boldsymbol{Z}}_2\right)^{-1}\widetilde{\boldsymbol{Z}}_2'$. Thus $\boldsymbol{P_Z} - \boldsymbol{P}_1 = \boldsymbol{P}_2$ and

$$\begin{aligned}
\widehat{\beta}_2 &= \left(\boldsymbol{Y}_2'\boldsymbol{P}_2\boldsymbol{Y}_2\right)^{-1}\left(\boldsymbol{Y}_2'\boldsymbol{P}_2\boldsymbol{Y}_1\right) \\
&= \left(\boldsymbol{Y}_2'\widetilde{\boldsymbol{Z}}_2\left(\widetilde{\boldsymbol{Z}}_2'\widetilde{\boldsymbol{Z}}_2\right)^{-1}\widetilde{\boldsymbol{Z}}_2'\boldsymbol{Y}_2\right)^{-1}\left(\boldsymbol{Y}_2'\widetilde{\boldsymbol{Z}}_2\left(\widetilde{\boldsymbol{Z}}_2'\widetilde{\boldsymbol{Z}}_2\right)^{-1}\widetilde{\boldsymbol{Z}}_2'\boldsymbol{Y}_1\right). \qquad (12.32)
\end{aligned}$$

Given the 2SLS estimator we define the residual $\widehat{e}_i = Y_{1i} - X_i'\widehat{\beta}_{\text{2sls}}$. When the model is overidentified the instruments and residuals are not orthogonal. That is, $\boldsymbol{Z}'\widehat{\boldsymbol{e}} \neq 0$. It does, however, satisfy

$$\begin{aligned}
\widehat{\boldsymbol{X}}'\widehat{\boldsymbol{e}} &= \widehat{\boldsymbol{\Gamma}}'\boldsymbol{Z}'\widehat{\boldsymbol{e}} \\
&= \boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\widehat{\boldsymbol{e}} \\
&= \boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{X}\widehat{\beta}_{\text{2sls}} = 0.
\end{aligned}$$

Returning to Card's college proximity example suppose that we treat experience as exogeneous but that instead of using the single instrument *college* (grew up near a 4-year college) we use the two instruments (*public, private*) (grew up near a public/private 4-year college, respectively). In this case we have one endogenous variable (*education*) and two instruments (*public, private*). The estimated reduced form equation for *education* is presented in the sixth column of Table 12.2. In this specification the coefficient on *public* – growing up near a public 4-year college – is larger than that found for the variable *college* in the previous specification (column 2). Furthermore, the coefficient on *private* – growing up near a private 4-year college – is much smaller. This indicates that the key impact of proximity on education is via public colleges rather than private colleges.

The 2SLS estimates obtained using these two instruments are presented in the fourth column of Table 12.1. The coefficient on *education* increases to 0.161, indicating a 16% return to a year of education. This is roughly twice as large as the estimate obtained by least squares in the first column.

Additionally, if we follow Card and treat *experience* as endogenous and use *age* as an instrument we now have three endogenous variables (*education, experience, experience*^{2}/100) and four instruments (*public, private, age, age*^{2}). We present the 2SLS estimates using this specification in the fifth column of Table 12.1. The estimate of the return to education remains 16% and the return to experience flattens.

You might wonder if we could use all three instruments – *college*, *public*, and *private*. The answer is no. This is because *college=public+private* so the three variables are colinear. Since the instruments are linearly related the three together would violate the full-rank condition (12.6).

The 2SLS estimator may be calculated in Stata using the `ivregress 2sls` command.

12.13 Limited Information Maximum Likelihood

An alternative method to estimate the parameters of the structural equation is by maximum likelihood. Anderson and Rubin (1949) derived the maximum likelihood estimator for the joint distribution of $\vec{Y} = (Y_1, Y_2)$. The estimator is known as **limited information maximum likelihood (LIML)**.

This estimator is called "limited information" because it is based on the structural equation for Y combined with the reduced form equation for X_2. If maximum likelihood is derived based on a structural equation for X_2 as well this leads to what is known as **full information maximum likelihood (FIML)**. The advantage of LIML relative to FIML is that the former does not require a structural model for X_2 and thus allows the researcher to focus on the structural equation of interest – that for Y. We do not describe the FIML estimator as it is not commonly used in applied econometrics.

While the LIML estimator is less widely used among economists than 2SLS it has received a resurgence of attention from econometric theorists.

To derive the LIML estimator recall the definition $\vec{Y} = (Y_1, Y_2)$ and the reduced form (12.17)

$$\begin{aligned}\vec{Y} &= \begin{bmatrix} \lambda_1' & \lambda_2 \\ \Gamma_{12}' & \Gamma_{22}' \end{bmatrix} \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} + u \\ &= \Pi_1' Z_1 + \Pi_2' Z_2 + u \qquad (12.33)\end{aligned}$$

where $\Pi_1 = \begin{bmatrix} \lambda_1 & \Gamma_{12} \end{bmatrix}$ and $\Pi_2 = \begin{bmatrix} \lambda_2 & \Gamma_{22} \end{bmatrix}$. The LIML estimator is derived under the assumption that u is multivariate normal.

Define $\gamma' = \begin{bmatrix} 1 & -\beta_2' \end{bmatrix}$. From (12.15) we find

$$\Pi_2 \gamma = \lambda_2 - \Gamma_{22} \beta_2 = 0.$$

Thus the $\ell_2 \times (k_2 + 1)$ coefficient matrix Π_2 in (12.33) has deficient rank. Indeed, its rank must be k_2 since Γ_{22} has full rank.

This means that the model (12.33) is precisely the reduced rank regression model of Section 11.11. Theorem 11.7 presents the maximum likelihood estimators for the reduced rank parameters. In particular, the MLE for γ is

$$\widehat{\gamma} = \underset{\gamma}{\operatorname{argmin}} \frac{\gamma' \vec{\boldsymbol{Y}}' \boldsymbol{M}_1 \vec{\boldsymbol{Y}} \gamma}{\gamma' \vec{\boldsymbol{Y}}' \boldsymbol{M}_{\boldsymbol{Z}} \vec{\boldsymbol{Y}} \gamma} \qquad (12.34)$$

where $\boldsymbol{M}_1 = \boldsymbol{I}_n - \boldsymbol{Z}_1 \left(\boldsymbol{Z}_1' \boldsymbol{Z}_1\right)^{-1} \boldsymbol{Z}_1'$ and $\boldsymbol{M}_{\boldsymbol{Z}} = \boldsymbol{I}_n - \boldsymbol{Z} \left(\boldsymbol{Z}' \boldsymbol{Z}\right)^{-1} \boldsymbol{Z}'$. The minimization (12.34) is sometimes called the "least variance ratio" problem.

The minimization problem (12.34) is invariant to the scale of γ (that is, $\widehat{\gamma} c$ is equivalently the argmin for any c) so normalization is required. A convenient choice is $\gamma' \vec{\boldsymbol{Y}}' \boldsymbol{M}_{\boldsymbol{Z}} \vec{\boldsymbol{Y}} \gamma = 1$. Using this normalization and the theory of the minimum of quadratic forms (Section A.15) $\widehat{\gamma}$ is the generalized eigenvector of $\vec{\boldsymbol{Y}}' \boldsymbol{M}_1 \vec{\boldsymbol{Y}}$ with respect to $\vec{\boldsymbol{Y}}' \boldsymbol{M}_{\boldsymbol{Z}} \vec{\boldsymbol{Y}}$ associated with the smallest generalized eigenvalue. (See Section A.14 for the definition of generalized eigenvalues and eigenvectors.) Computationally this is straightforward. For example, in MATLAB the generalized eigenvalues and eigenvectors of the matrix $\boldsymbol{A}$ with respect to $\boldsymbol{B}$ is found by the command `eig(`$\boldsymbol{A}$`,`$\boldsymbol{B}$`)`. Once this $\widehat{\gamma}$ is found any other normalization can be obtained by rescaling. For example, to obtain the MLE for β_2 make the partition $\widehat{\gamma}' = \begin{bmatrix} \widehat{\gamma}_1 & \widehat{\gamma}_2' \end{bmatrix}$ and set $\widehat{\beta}_2 = -\widehat{\gamma}_2 / \widehat{\gamma}_1$.

To obtain the MLE for β_1 recall the structural equation $Y_1 = Z_1' \beta_1 + Y_2' \beta_2 + e$. Replace β_2 with the MLE $\widehat{\beta}_2$ and apply regression. This yields

$$\widehat{\beta}_1 = \left(\boldsymbol{Z}_1' \boldsymbol{Z}_1\right)^{-1} \boldsymbol{Z}_1' \left(\boldsymbol{Y}_1 - \boldsymbol{Y}_2 \widehat{\beta}_2\right). \qquad (12.35)$$

These solutions are the MLE for the structural parameters β_1 and β_2.

Many previous econometrics textbooks do not present a derivation of the LIML estimator as the original derivation by Anderson and Rubin (1949) is lengthy and not particularly insightful. In contrast the derivation given here based on reduced rank regression is simple.

There is an alternative (and traditional) expression for the LIML estimator. Define the minimum obtained in (12.34)

$$\widehat{\kappa} = \min_{\gamma} \frac{\gamma' \vec{\boldsymbol{Y}}' \boldsymbol{M}_1 \vec{\boldsymbol{Y}} \gamma}{\gamma' \vec{\boldsymbol{Y}}' \boldsymbol{M}_{\boldsymbol{Z}} \vec{\boldsymbol{Y}} \gamma} \qquad (12.36)$$

which is the smallest generalized eigenvalue of $\vec{\boldsymbol{Y}}' \boldsymbol{M}_1 \vec{\boldsymbol{Y}}$ with respect to $\vec{\boldsymbol{Y}}' \boldsymbol{M}_{\boldsymbol{Z}} \vec{\boldsymbol{Y}}$. The LIML estimator can be written as

$$\widehat{\beta}_{\text{liml}} = \left(\boldsymbol{X}' \left(\boldsymbol{I}_n - \widehat{\kappa} \boldsymbol{M}_{\boldsymbol{Z}}\right) \boldsymbol{X}\right)^{-1} \left(\boldsymbol{X}' \left(\boldsymbol{I}_n - \widehat{\kappa} \boldsymbol{M}_{\boldsymbol{Z}}\right) \boldsymbol{Y}_1\right). \qquad (12.37)$$

We defer the derivation of (12.37) until the end of this section. Expression (12.37) does not simplify computation (since $\widehat{\kappa}$ requires solving the same eigenvector problem that yields $\widehat{\beta}_2$). However (12.37) is important for the distribution theory. It also helps reveal the algebraic connection between LIML, least squares, and 2SLS.

The estimator (12.37) with arbitrary κ is known as a **k-class estimator** of β. While the LIML estimator obtains by setting $\kappa = \widehat{\kappa}$, the least squares estimator is obtained by setting $\kappa = 0$ and 2SLS is obtained by setting $\kappa = 1$. It is worth observing that the LIML solution satisfies $\widehat{\kappa} \geq 1$.

When the model is just-identified the LIML estimator is identical to the IV and 2SLS estimators. They are only different in the over-identified setting. (One corollary is that under just-identification and normal errors the IV estimator is MLE.)

For inference it is useful to observe that (12.37) shows that $\widehat{\beta}_{\text{liml}}$ can be written as an IV estimator

$$\widehat{\beta}_{\text{liml}} = \left(\widetilde{\boldsymbol{X}}'\boldsymbol{X}\right)^{-1}\left(\widetilde{\boldsymbol{X}}'\boldsymbol{Y}_1\right)$$

using the instrument

$$\widetilde{\boldsymbol{X}} = \left(\boldsymbol{I}_n - \widehat{\kappa}\boldsymbol{M}_{\boldsymbol{Z}}\right)\boldsymbol{X} = \left(\begin{array}{c} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 - \widehat{\kappa}\widehat{\boldsymbol{U}}_2 \end{array}\right)$$

where $\widehat{\boldsymbol{U}}_2 = \boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{X}_2$ are the reduced-form residuals from the multivariate regression of the endogenous regressors Y_2 on the instruments Z. Expressing LIML using this IV formula is useful for variance estimation.

The LIML estimator has the same asymptotic distribution as 2SLS. However, they have quite different behaviors in finite samples. There is considerable evidence that the LIML estimator has reduced finite sample bias relative to 2SLS when there are many instruments or the reduced form is weak. (We review these cases in the following sections.) However, on the other hand LIML has wider finite sample dispersion.

We now derive the expression (12.37). Use the normalization $\gamma' = \left[\begin{array}{cc} 1 & -\beta_2' \end{array}\right]$ to write (12.34) as

$$\widehat{\beta}_2 = \underset{\beta_2}{\operatorname{argmin}} \frac{\left(\boldsymbol{Y}_1 - \boldsymbol{Y}_2\beta_2\right)'\boldsymbol{M}_1\left(\boldsymbol{Y}_1 - \boldsymbol{Y}_2\beta_2\right)}{\left(\boldsymbol{Y}_1 - \boldsymbol{Y}\beta_2\right)'\boldsymbol{M}_{\boldsymbol{Z}}\left(\boldsymbol{Y}_1 - \boldsymbol{Y}_2\beta_2\right)}.$$

The first-order-condition for minimization is $2/\left(\boldsymbol{Y}_1 - \boldsymbol{Y}_2\widehat{\beta}_2\right)'\boldsymbol{M}_{\boldsymbol{Z}}\left(\boldsymbol{Y}_1 - \boldsymbol{Y}_2\widehat{\beta}_2\right)$ times

$$\begin{aligned} 0 &= \boldsymbol{Y}_2'\boldsymbol{M}_1\left(\boldsymbol{Y}_1 - \boldsymbol{Y}_2\widehat{\beta}_2\right) - \frac{\left(\boldsymbol{Y}_1 - \boldsymbol{Y}_2\widehat{\beta}_2\right)'\boldsymbol{M}_1\left(\boldsymbol{Y}_1 - \boldsymbol{Y}_2\widehat{\beta}_2\right)}{\left(\boldsymbol{Y}_1 - \boldsymbol{Y}_2\widehat{\beta}_2\right)'\boldsymbol{M}_{\boldsymbol{Z}}\left(\boldsymbol{Y}_1 - \boldsymbol{Y}_2\widehat{\beta}_2\right)}\boldsymbol{X}_2'\boldsymbol{M}_{\boldsymbol{Z}}\left(\boldsymbol{Y}_1 - \boldsymbol{Y}_2\widehat{\beta}_2\right) \\ &= \boldsymbol{Y}_2'\boldsymbol{M}_1\left(\boldsymbol{Y}_1 - \boldsymbol{Y}_2\widehat{\beta}_2\right) - \widehat{\kappa}\boldsymbol{X}_2'\boldsymbol{M}_{\boldsymbol{Z}}\left(\boldsymbol{Y}_1 - \boldsymbol{Y}_2\widehat{\beta}_2\right) \end{aligned}$$

using definition (12.36). Rewriting,

$$\boldsymbol{Y}_2'\left(\boldsymbol{M}_1 - \widehat{\kappa}\boldsymbol{M}_{\boldsymbol{Z}}\right)\boldsymbol{X}_2\widehat{\beta}_2 = \boldsymbol{X}_2'\left(\boldsymbol{M}_1 - \widehat{\kappa}\boldsymbol{M}_{\boldsymbol{Z}}\right)\boldsymbol{Y}_1. \tag{12.38}$$

Equation (12.37) is the same as the two equation system

$$\begin{aligned} \boldsymbol{Z}_1'\boldsymbol{Z}_1\widehat{\beta}_1 + \boldsymbol{Z}_1'\boldsymbol{Y}_2\widehat{\beta}_2 &= \boldsymbol{Z}_1'\boldsymbol{Y}_1 \\ \boldsymbol{Y}_2'\boldsymbol{Z}_1\widehat{\beta}_1 + \left(\boldsymbol{Y}_2'\left(\boldsymbol{I}_n - \widehat{\kappa}\boldsymbol{M}_{\boldsymbol{Z}}\right)\boldsymbol{Y}_2\right)\widehat{\beta}_2 &= \boldsymbol{Y}_2'\left(\boldsymbol{I}_n - \widehat{\kappa}\boldsymbol{M}_{\boldsymbol{Z}}\right)\boldsymbol{Y}_1. \end{aligned}$$

The first equation is (12.35). Using (12.35), the second is

$$\boldsymbol{Y}_2'\boldsymbol{Z}_1\left(\boldsymbol{Z}_1'\boldsymbol{Z}_1\right)^{-1}\boldsymbol{Z}_1'\left(\boldsymbol{Y}_1 - \boldsymbol{Y}_2\widehat{\beta}_2\right) + \left(\boldsymbol{Y}_2'\left(\boldsymbol{I}_n - \widehat{\kappa}\boldsymbol{M}_{\boldsymbol{Z}}\right)\boldsymbol{Y}_2\right)\widehat{\beta}_2 = \boldsymbol{Y}_2'\left(\boldsymbol{I}_n - \widehat{\kappa}\boldsymbol{M}_{\boldsymbol{Z}}\right)\boldsymbol{Y}_1$$

which is (12.38) when rearranged. We have thus shown that (12.37) is equivalent to (12.35) and (12.38) and is thus a valid expression for the LIML estimator.

Returning to the Card college proximity example we now present the LIML estimates of the equation with the two instruments (*public*, *private*). They are reported in the final column of Table 12.1. They are quite similar to the 2SLS estimates.

The LIML estimator may be calculated in Stata using the `ivregress liml` command.

Theodore Anderson

Theodore (Ted) Anderson (1918-2016) was a American statistician and econometrician, who made fundamental contributions to multivariate statistical theory. Important contributions include the Anderson-Darling distribution test, the Anderson-Rubin statistic, the method of reduced rank regression, and his most famous econometrics contribution – the LIML estimator. He continued working throughout his long life, even publishing theoretical work at the age of 97!

12.14 Split-Sample IV and JIVE

The ideal instrument for estimation of β is $W = \Gamma' Z$. We can write the ideal IV estimator as

$$\widehat{\beta}_{\text{ideal}} = \left(\sum_{i=1}^{n} W_i X_i'\right)^{-1}\left(\sum_{i=1}^{n} W_i Y_i\right).$$

This estimator is not feasible since Γ is unknown. The 2SLS estimator replaces Γ with the multivariate least squares estimator $\widehat{\Gamma}$ and W_i with $\widehat{W}_i = \widehat{\Gamma}' Z_i$ leading to the following representation for 2SLS

$$\widehat{\beta}_{2\text{sls}} = \left(\sum_{i=1}^{n} \widehat{W}_i X_i'\right)^{-1}\left(\sum_{i=1}^{n} \widehat{W}_i Y_i\right).$$

Since $\widehat{\Gamma}$ is estimated on the full sample including observation i it is a function of the reduced form error u which is correlated with the structural error e. It follows that $\widehat{W}$ and e are correlated, which means that $\widehat{\beta}_{2\text{sls}}$ is biased for β. This correlation and bias disappears asymptotically but it can be important in applications.

A possible solution to this problem is to replace $\widehat{W}$ with a predicted value which is uncorrelated with the error e. One method is the **split-sample IV (SSIV)** estimator of Angrist and Krueger (1995). Divide the sample randomly into two independent halves A and B. Use A to estimate the reduce form and B to estimate the structural coefficient. Specifically, use sample A to construct $\widehat{\Gamma}_A = \left(\boldsymbol{Z}_A'\boldsymbol{Z}_A\right)^{-1}\left(\boldsymbol{Z}_A'\boldsymbol{X}_A\right)$. Combine this with sample B to create the predicted values $\widehat{\boldsymbol{W}}_B = \boldsymbol{Z}_B\widehat{\Gamma}_A$. The SSIV estimator is $\widehat{\beta}_{\text{ssiv}} = \left(\widehat{\boldsymbol{W}}_B'\boldsymbol{X}_B\right)^{-1}\left(\widehat{\boldsymbol{W}}_B'\boldsymbol{Y}_B\right)$. This has lower bias than $\widehat{\beta}_{2\text{sls}}$.

A limitation of SSIV is that the results will be sensitive to the sample spliting. One split will produce one estimator; another split will produce a different estimator. Any specific split is arbitrary, so the estimator depends on the specific random sorting of the observations into the samples A and B. A second limitation of SSIV is that it is unlikely to work well when the sample size n is small.

A much better solution is obtained by a leave-one-out estimator for Γ. Specifically, let

$$\widehat{\Gamma}_{(-i)} = \left(\boldsymbol{Z}'\boldsymbol{Z} - Z_i Z_i'\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{X} - Z_i X_i'\right)$$

be the least squares leave-one-out estimator of the reduced form matrix Γ, and let $\widehat{W}_i = \widehat{\Gamma}'_{(-i)} Z_i$ be the reduced form predicted values. Using $\widehat{W}_i = \widehat{\Gamma}'_{(-i)} Z_i$ as an instrument we obtain the estimator

$$\widehat{\beta}_{\text{jive1}} = \left(\sum_{i=1}^{n} \widehat{W}_i X_i'\right)^{-1}\left(\sum_{i=1}^{n} \widehat{W}_i Y_i\right) = \left(\sum_{i=1}^{n} \widehat{\Gamma}'_{(-i)} Z_i X_i'\right)^{-1}\left(\sum_{i=1}^{n} \widehat{\Gamma}'_{(-i)} Z_i Y_i\right).$$

This was called the **jackknife instrumental variables (JIVE1)** estimator by Angrist, Imbens, and Krueger (1999). It first appeared in Phillips and Hale (1977).

Angrist, Imbens, and Krueger (1999) pointed out that a somewhat simpler adjustment also removes the correlation and bias. Define the estimator and predicted value

$$\widetilde{\Gamma}_{(-i)} = \left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{X} - Z_i X_i'\right)$$
$$\widetilde{W}_i = \widetilde{\Gamma}'_{(-i)} Z_i$$

which only adjusts the $\boldsymbol{Z}'\boldsymbol{X}$ component. Their **JIVE2** estimator is

$$\widehat{\beta}_{\text{jive2}} = \left(\sum_{i=1}^{n} \widetilde{W}_i X_i'\right)^{-1}\left(\sum_{i=1}^{n} \widetilde{W}_i Y_i\right) = \left(\sum_{i=1}^{n} \widetilde{\Gamma}'_{(-i)} Z_i X_i'\right)^{-1}\left(\sum_{i=1}^{n} \widetilde{\Gamma}'_{(-i)} Z_i Y_i\right).$$

Using the formula for leave-one-out estimators (Theorem 3.7), the JIVE1 and JIVE2 estimators use two linear operations: the first to create the predicted values $\widehat{W}_i$ or $\widetilde{W}_i$, and the second to calculate the IV estimator. Thus the estimators do not require significantly more computation than 2SLS.

An asymptotic distribution theory for the JIVE1 and JIVE2 estimators was developed by Chao, Swanson, Hausman, Newey, and Woutersen (2012).

The JIVE1 and JIVE2 estimators may be calculated in Stata using the `jive` command. It is not a part of the standard package but can be easily added.

12.15 Consistency of 2SLS

We now demonstrate the consistency of the 2SLS estimator for the structural parameter. The following is a set of regularity conditions.

Assumption 12.1

1. The variables (Y_{1i}, X_i, Z_i), $i = 1, \ldots, n$, are independent and identically distributed.
2. $\mathbb{E}\left[Y_1^2\right] < \infty$.
3. $\mathbb{E}\|X\|^2 < \infty$.
4. $\mathbb{E}\|Z\|^2 < \infty$.
5. $\mathbb{E}\left[ZZ'\right]$ is positive definite.
6. $\mathbb{E}\left[ZX'\right]$ has full rank k.
7. $\mathbb{E}[Ze] = 0$.

Assumptions 12.1.2-4 state that all variables have finite variances. Assumption 12.1.5 states that the instrument vector has an invertible design matrix, which is identical to the core assumption about regressors in the linear regression model. This excludes linearly redundant instruments. Assumptions 12.1.6 and 12.1.7 are the key identification conditions for instrumental variables. Assumption 12.1.6 states that the instruments and regressors have a full-rank cross-moment matrix. This is often called the relevance condition. Assumption 12.1.7 states that the instrumental variables and structural error are uncorrelated. Assumptions 12.1.5-7 are identical to Definition 12.1.

Theorem 12.1 Under Assumption 12.1, $\widehat{\beta}_{\text{2sls}} \xrightarrow[p]{} \beta$ as $n \to \infty$.

The proof of the theorem is provided below.

This theorem shows that the 2SLS estimator is consistent for the structural coefficient β under similar moment conditions as the least squares estimator. The key differences are the instrumental variables assumption $\mathbb{E}[Ze] = 0$ and the relevance condition $\operatorname{rank}\left(\mathbb{E}\left[ZX'\right]\right) = k$.

The result includes the IV estimator (when $\ell = k$) as a special case.

The proof of this consistency result is similar to that for least squares. Take the structural equation $\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{e}$ in matrix format and substitute it into the expression for the estimator. We obtain

$$\begin{aligned}
\widehat{\beta}_{\text{2sls}} &= \left(\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\left(\boldsymbol{X}\beta + \boldsymbol{e}\right) \\
&= \beta + \left(\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{e}. \qquad (12.39)
\end{aligned}$$

This separates out the stochastic component. Re-writing and applying the WLLN and CMT

$$\begin{aligned}
\widehat{\beta}_{\text{2sls}} - \beta &= \left(\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\right)\left(\frac{1}{n}\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\left(\frac{1}{n}\boldsymbol{Z}'\boldsymbol{X}\right)\right)^{-1} \\
&\quad \times \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\right)\left(\frac{1}{n}\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\left(\frac{1}{n}\boldsymbol{Z}'\boldsymbol{e}\right) \\
&\xrightarrow[p]{} \left(\boldsymbol{Q}_{XZ}\boldsymbol{Q}_{ZZ}^{-1}\boldsymbol{Q}_{ZX}\right)^{-1}\boldsymbol{Q}_{XZ}\boldsymbol{Q}_{ZZ}^{-1}\mathbb{E}[Ze] = 0
\end{aligned}$$

where

$$\begin{aligned}
\boldsymbol{Q}_{XZ} &= \mathbb{E}\left[XZ'\right] \\
\boldsymbol{Q}_{ZZ} &= \mathbb{E}\left[ZZ'\right] \\
\boldsymbol{Q}_{ZX} &= \mathbb{E}\left[ZX'\right].
\end{aligned}$$

The WLLN holds under the i.i.d. Assumption 12.1.1 and the finite second moment Assumptions 12.1.2-4. The continuous mapping theorem applies if the matrices $\boldsymbol{Q}_{ZZ}$ and $\boldsymbol{Q}_{XZ}\boldsymbol{Q}_{ZZ}^{-1}\boldsymbol{Q}_{ZX}$ are invertible, which hold under the identification Assumptions 12.1.5 and 12.1.6. The final equality uses Assumption 12.1.7.

12.16 Asymptotic Distribution of 2SLS

We now show that the 2SLS estimator satisfies a central limit theorem. We first state a set of sufficient regularity conditions.

Assumption 12.2 In addition to Assumption 12.1,

1. $\mathbb{E}\left[Y_1^4\right] < \infty$.
2. $\mathbb{E}\|X\|^4 < \infty$.
3. $\mathbb{E}\|Z\|^4 < \infty$.
4. $\Omega = \mathbb{E}\left[ZZ'e^2\right]$ is positive definite.

Assumption 12.2 strengthens Assumption 12.1 by requiring that the dependent variable and instruments have finite fourth moments. This is used to establish the central limit theorem.

Theorem 12.2 Under Assumption 12.2, as $n \to \infty$.

$$\sqrt{n}\left(\widehat{\beta}_{\text{2sls}} - \beta\right) \xrightarrow[d]{} \mathrm{N}\left(0, \boldsymbol{V}_\beta\right)$$

where

$$\boldsymbol{V}_\beta = \left(\boldsymbol{Q}_{XZ}\boldsymbol{Q}_{ZZ}^{-1}\boldsymbol{Q}_{ZX}\right)^{-1}\left(\boldsymbol{Q}_{XZ}\boldsymbol{Q}_{ZZ}^{-1}\Omega\boldsymbol{Q}_{ZZ}^{-1}\boldsymbol{Q}_{ZX}\right)\left(\boldsymbol{Q}_{XZ}\boldsymbol{Q}_{ZZ}^{-1}\boldsymbol{Q}_{ZX}\right)^{-1}.$$

This shows that the 2SLS estimator converges at a $\sqrt{n}$ rate to a normal random vector. It shows as well the form of the covariance matrix. The latter takes a substantially more complicated form than the least squares estimator.

As in the case of least squares estimation the asymptotic variance simplifies under a conditional homoskedasticity condition. For 2SLS the simplification occurs when $\mathbb{E}\left[e^2 \mid Z\right] = \sigma^2$. This holds when Z and e are independent. It may be reasonable in some contexts to conceive that the error e is independent of the excluded instruments Z_2, since by assumption the impact of Z_2 on Y is only through X, but there is no reason to expect e to be independent of the included exogenous variables X_1. Hence heteroskedasticity should be equally expected in 2SLS and least squares regression. Nevertheless, under homoskedasticity we have the simplifications $\Omega = \boldsymbol{Q}_{ZZ}\sigma^2$ and $\boldsymbol{V}_\beta = \boldsymbol{V}_\beta^0 \overset{\text{def}}{=} \left(\boldsymbol{Q}_{XZ}\boldsymbol{Q}_{ZZ}^{-1}\boldsymbol{Q}_{ZX}\right)^{-1}\sigma^2$.

The derivation of the asymptotic distribution builds on the proof of consistency. Using equation (12.39) we have

$$\sqrt{n}\left(\widehat{\beta}_{\text{2sls}} - \beta\right) = \left(\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\right)\left(\frac{1}{n}\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\left(\frac{1}{n}\boldsymbol{Z}'\boldsymbol{X}\right)\right)^{-1}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\right)\left(\frac{1}{n}\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\left(\frac{1}{\sqrt{n}}\boldsymbol{Z}'\boldsymbol{e}\right).$$

We apply the WLLN and CMT for the moment matrices involving $\boldsymbol{X}$ and $\boldsymbol{Z}$ the same as in the proof of consistency. In addition, by the CLT for i.i.d. observations

$$\frac{1}{\sqrt{n}}\boldsymbol{Z}'\boldsymbol{e} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i e_i \xrightarrow[d]{} \mathrm{N}(0, \Omega)$$

because the vector $Z_i e_i$ is i.i.d. and mean zero under Assumptions 12.1.1 and 12.1.7, and has a finite second moment as we verify below.

We obtain

$$\sqrt{n}\left(\widehat{\beta}_{\text{2sls}}-\beta\right)=\left(\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\right)\left(\frac{1}{n}\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\left(\frac{1}{n}\boldsymbol{Z}'\boldsymbol{X}\right)\right)^{-1}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\right)\left(\frac{1}{n}\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\left(\frac{1}{\sqrt{n}}\boldsymbol{Z}'\boldsymbol{e}\right)$$

$$\underset{d}{\longrightarrow}\left(\boldsymbol{Q}_{XZ}\boldsymbol{Q}_{ZZ}^{-1}\boldsymbol{Q}_{ZX}\right)^{-1}\boldsymbol{Q}_{XZ}\boldsymbol{Q}_{ZZ}^{-1}\mathrm{N}(0,\Omega)=\mathrm{N}\left(0,\boldsymbol{V}_{\beta}\right)$$

as stated.

To complete the proof we demonstrate that Ze has a finite second moment under Assumption 12.2. To see this, note that by Minkowski's inequality (B.34)

$$\left(\mathbb{E}\left[e^4\right]\right)^{1/4}=\left(\mathbb{E}\left[\left(Y_1-X'\beta\right)^4\right]\right)^{1/4}\leq\left(\mathbb{E}\left[Y_1^4\right]\right)^{1/4}+\|\beta\|\left(\mathbb{E}\|X\|^4\right)^{1/4}<\infty$$

under Assumptions 12.2.1 and 12.2.2. Then by the Cauchy-Schwarz inequality (B.32)

$$\mathbb{E}\|Ze\|^2\leq\left(\mathbb{E}\|Z\|^4\right)^{1/2}\left(\mathbb{E}\left[e^4\right]\right)^{1/2}<\infty$$

using Assumptions 12.2.3.

12.17 Determinants of 2SLS Variance

It is instructive to examine the asymptotic variance of the 2SLS estimator to understand the factors which determine the precision (or lack thereof) of the estimator. As in the least squares case it is more transparent to examine the variance under the assumption of homoskedasticity. In this case the asymptotic variance takes the form

$$\begin{aligned}\boldsymbol{V}_{\beta}^{0}&=\left(\boldsymbol{Q}_{XZ}\boldsymbol{Q}_{ZZ}^{-1}\boldsymbol{Q}_{ZX}\right)^{-1}\sigma^2\\&=\left(\mathbb{E}\left[XZ'\right]\left(\mathbb{E}\left[ZZ'\right]\right)^{-1}\mathbb{E}\left[ZX'\right]\right)^{-1}\mathbb{E}\left[e^2\right].\end{aligned}$$

As in the least squares case we can see that the variance of $\widehat{\beta}_{\text{2sls}}$ is increasing in the variance of the error e and decreasing in the variance of X. What is different is that the variance is decreasing in the (matrix-valued) correlation between X and Z.

It is also useful to observe that the variance expression is not affected by the variance structure of Z. Indeed, $\boldsymbol{V}_{\beta}^{0}$ is invariant to rotations of Z (if you replace Z with $\boldsymbol{C}Z$ for invertible $\boldsymbol{C}$ the expression does not change). This means that the variance expression is not affected by the scaling of Z and is not directly affected by correlation among the Z.

We can also use this expression to examine the impact of increasing the instrument set. Suppose we partition $Z=(Z_a,Z_b)$ where $\dim(Z_a)\geq k$ so we can construct a 2SLS estimator using Z_a alone. Let $\widehat{\beta}_a$ and $\widehat{\beta}$ denote the 2SLS estimators constructed using the instrument sets Z_a and (Z_a,Z_b), respectively. Without loss of generality we can assume that Z_a and Z_b are uncorrelated (if not, replace Z_b with the projection error after projecting onto Z_a). In this case both $\mathbb{E}\left[ZZ'\right]$ and $\left(\mathbb{E}\left[ZZ'\right]\right)^{-1}$ are block diagonal so

$$\begin{aligned}\operatorname{avar}\left[\widehat{\beta}\right]&=\left(\mathbb{E}\left[XZ'\right]\left(\mathbb{E}\left[ZZ'\right]\right)^{-1}\mathbb{E}\left[ZX'\right]\right)^{-1}\sigma^2\\&=\left(\mathbb{E}\left[XZ_a'\right]\left(\mathbb{E}\left[Z_aZ_a'\right]\right)^{-1}\mathbb{E}\left[Z_aX'\right]+\mathbb{E}\left[XZ_b'\right]\left(\mathbb{E}\left[Z_bZ_b'\right]\right)^{-1}\mathbb{E}\left[Z_bX'\right]\right)^{-1}\sigma^2\\&\leq\left(\mathbb{E}\left[XZ_a'\right]\left(\mathbb{E}\left[Z_aZ_a'\right]\right)^{-1}\mathbb{E}\left[Z_aX'\right]\right)^{-1}\sigma^2\\&=\operatorname{avar}\left[\widehat{\beta}_a\right]\end{aligned}$$

with strict inequality if $\mathbb{E}\left[XZ_b'\right] \neq 0$. Thus the 2SLS estimator with the full instrument set has a smaller asymptotic variance than the estimator with the smaller instrument set.

What we have shown is that the asymptotic variance of the 2SLS estimator is decreasing as the number of instruments increases. From the viewpoint of asymptotic efficiency this means that it is better to use more instruments (when they are available and are all known to be valid instruments).

Unfortunately there is a catch. It turns out that the finite sample bias of the 2SLS estimator (which cannot be calculated exactly but can be approximated using asymptotic expansions) is generically increasing linearly as the number of instruments increases. We will see some calculations illustrating this phenomenon in Section 12.37. Thus the choice of instruments in practice induces a trade-off between bias and variance.

12.18 Covariance Matrix Estimation

Estimation of the asymptotic covariance matrix $\boldsymbol{V}_\beta$ is done using similar techniques as for least squares estimation. The estimator is constructed by replacing the population moment matrices by sample counterparts. Thus

$$\widehat{\boldsymbol{V}}_\beta = \left(\widehat{\boldsymbol{Q}}_{XZ}\widehat{\boldsymbol{Q}}_{ZZ}^{-1}\widehat{\boldsymbol{Q}}_{ZX}\right)^{-1}\left(\widehat{\boldsymbol{Q}}_{XZ}\widehat{\boldsymbol{Q}}_{ZZ}^{-1}\widehat{\Omega}\widehat{\boldsymbol{Q}}_{ZZ}^{-1}\widehat{\boldsymbol{Q}}_{ZX}\right)\left(\widehat{\boldsymbol{Q}}_{XZ}\widehat{\boldsymbol{Q}}_{ZZ}^{-1}\widehat{\boldsymbol{Q}}_{ZX}\right)^{-1} \tag{12.40}$$

where

$$\begin{aligned}
\widehat{\boldsymbol{Q}}_{ZZ} &= \frac{1}{n}\sum_{i=1}^{n} Z_i Z_i' = \frac{1}{n}\boldsymbol{Z}'\boldsymbol{Z} \\
\widehat{\boldsymbol{Q}}_{XZ} &= \frac{1}{n}\sum_{i=1}^{n} X_i Z_i' = \frac{1}{n}\boldsymbol{X}'\boldsymbol{Z} \\
\widehat{\Omega} &= \frac{1}{n}\sum_{i=1}^{n} Z_i Z_i'\widehat{e}_i^2 \\
\widehat{e}_i &= Y_i - X_i'\widehat{\beta}_{\text{2sls}}.
\end{aligned}$$

The homoskedastic covariance matrix can be estimated by

$$\begin{aligned}
\widehat{\boldsymbol{V}}_\beta^0 &= \left(\widehat{\boldsymbol{Q}}_{XZ}\widehat{\boldsymbol{Q}}_{ZZ}^{-1}\widehat{\boldsymbol{Q}}_{ZX}\right)^{-1}\widehat{\sigma}^2 \\
\widehat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}\widehat{e}_i^2.
\end{aligned}$$

Standard errors for the coefficients are obtained as the square roots of the diagonal elements of $n^{-1}\widehat{\boldsymbol{V}}_\beta$. Confidence intervals, t-tests, and Wald tests may all be constructed from the coefficient estimates and covariance matrix estimate exactly as for least squares regression.

In Stata the `ivregress` command by default calculates the covariance matrix estimator using the homoskedastic covariance matrix. To obtain covariance matrix estimation and standard errors with the robust estimator $\widehat{\boldsymbol{V}}_\beta$, use the "`,r`" option.

Theorem 12.3 Under Assumption 12.2, as $n \to \infty$, $\widehat{\boldsymbol{V}}_\beta^0 \xrightarrow[p]{} \boldsymbol{V}_\beta^0$ and $\widehat{\boldsymbol{V}}_\beta \xrightarrow[p]{} \boldsymbol{V}_\beta$.

To prove Theorem 12.3 the key is to show $\widehat{\Omega} \underset{p}{\longrightarrow} \Omega$ as the other convergence results were established in the proof of consistency. We defer this to Exercise 12.6.

It is important that the covariance matrix be constructed using the correct residual formula $\widehat{e}_i = Y_i - X_i'\widehat{\beta}_{\text{2sls}}$. This is different than what would be obtained if the "two-stage" computation method were used. To see this let's walk through the two-stage method. First, we estimate the reduced form $X_i = \widehat{\Gamma}'Z_i + \widehat{u}_i$ to obtain the predicted values $\widehat{X}_i = \widehat{\Gamma}'Z_i$. Second, we regress Y on $\widehat{X}$ to obtain the 2SLS estimator $\widehat{\beta}_{\text{2sls}}$. This latter regression takes the form

$$Y_i = \widehat{X}_i'\widehat{\beta}_{\text{2sls}} + \widehat{v}_i \tag{12.41}$$

where $\widehat{v}_i$ are least squares residuals. The covariance matrix (and standard errors) reported by this regression are constructed using the residual $\widehat{v}_i$. For example, the homoskedastic formula is

$$\widehat{\boldsymbol{V}}_{\beta} = \left(\frac{1}{n}\widehat{\boldsymbol{X}}'\widehat{\boldsymbol{X}}\right)^{-1}\widehat{\sigma}_v^2 = \left(\widehat{\boldsymbol{Q}}_{XZ}\widehat{\boldsymbol{Q}}_{ZZ}^{-1}\widehat{\boldsymbol{Q}}_{ZX}\right)^{-1}\widehat{\sigma}_v^2$$
$$\widehat{\sigma}_v^2 = \frac{1}{n}\sum_{i=1}^{n}\widehat{v}_i^2$$

which is proportional to the variance estimator $\widehat{\sigma}_v^2$ rather than $\widehat{\sigma}^2$. This is important because the residual $\widehat{v}$ differs from $\widehat{e}$. We can see this because the regression (12.41) uses the regressor $\widehat{X}$ rather than X. Indeed, we calculate that

$$\widehat{v}_i = Y_i - X_i'\widehat{\beta}_{\text{2sls}} + \left(X_i - \widehat{X}_i\right)'\widehat{\beta}_{\text{2sls}} = \widehat{e}_i + \widehat{u}_i'\widehat{\beta}_{\text{2sls}} \neq \widehat{e}_i.$$

This means that standard errors reported by the regression (12.41) will be incorrect.

This problem is avoided if the 2SLS estimator is constructed directly and the standard errors calculated with the correct formula rather than taking the "two-step" shortcut.

12.19 LIML Asymptotic Distribution

In this section we show that the LIML estimator is asymptotically equivalent to the 2SLS estimator. We recommend, however, a different covariance matrix estimator based on the IV representation.

We start by deriving the asymptotic distribution. Recall that the LIML estimator has several representations including

$$\widehat{\beta}_{\text{liml}} = \left(\boldsymbol{X}'\left(\boldsymbol{I}_n - \widehat{\kappa}\boldsymbol{M}_{\boldsymbol{Z}}\right)\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\left(\boldsymbol{I}_n - \widehat{\kappa}\boldsymbol{M}_{\boldsymbol{Z}}\right)\boldsymbol{Y}_1\right)$$

where

$$\widehat{\kappa} = \min_{\gamma}\frac{\gamma'\vec{\boldsymbol{Y}}'\boldsymbol{M}_1\vec{\boldsymbol{Y}}\gamma}{\gamma'\vec{\boldsymbol{Y}}'\boldsymbol{M}_{\boldsymbol{Z}}\vec{\boldsymbol{Y}}\gamma}$$

and $\gamma = (1, -\beta_2')'$. For the distribution theory it is useful to rewrite the slope coefficient as

$$\widehat{\beta}_{\text{liml}} = \left(\boldsymbol{X}'\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X} - \widehat{\mu}\boldsymbol{X}'\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{Y}_1 - \widehat{\mu}\boldsymbol{X}'\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{Y}_1\right)$$

where

$$\widehat{\mu} = \widehat{\kappa} - 1 = \min_{\gamma}\frac{\gamma'\vec{\boldsymbol{Y}}'\boldsymbol{M}_1\boldsymbol{Z}_2\left(\boldsymbol{Z}_2'\boldsymbol{M}_1\boldsymbol{Z}_2\right)^{-1}\boldsymbol{Z}_2'\boldsymbol{M}_1\vec{\boldsymbol{Y}}\gamma}{\gamma'\vec{\boldsymbol{Y}}'\boldsymbol{M}_{\boldsymbol{Z}}\vec{\boldsymbol{Y}}\gamma}.$$

This second equality holds since the span of $\boldsymbol{Z} = [\boldsymbol{Z}_1, \boldsymbol{Z}_2]$ equals the span of $[\boldsymbol{Z}_1, \boldsymbol{M}_1\boldsymbol{Z}_2]$. This implies

$$\boldsymbol{P}_{\boldsymbol{Z}} = \boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}' = \boldsymbol{Z}_1\left(\boldsymbol{Z}_1'\boldsymbol{Z}_1\right)^{-1}\boldsymbol{Z}_1' + \boldsymbol{M}_1\boldsymbol{Z}_2\left(\boldsymbol{Z}_2'\boldsymbol{M}_1\boldsymbol{Z}_2\right)^{-1}\boldsymbol{Z}_2'\boldsymbol{M}_1.$$

We now show that $n\widehat{\mu} = O_p(1)$. The reduced form (12.33) implies that

$$\boldsymbol{Y} = \boldsymbol{Z}_1\Pi_1 + \boldsymbol{Z}_2\Pi_2 + \boldsymbol{e}.$$

It will be important to note that

$$\Pi_2 = [\lambda_2, \Gamma_{22}] = \left[\Gamma_{22}\beta_2, \Gamma_{22}\right]$$

using (12.15). It follows that $\Pi_2\gamma = 0$. Note $\boldsymbol{U}\gamma = \boldsymbol{e}$. Then $\boldsymbol{M_Z Y}\gamma = \boldsymbol{M_Z e}$ and $\boldsymbol{M}_1\boldsymbol{Y}\gamma = \boldsymbol{M}_1\boldsymbol{e}$. Hence

$$\begin{aligned} n\widehat{\mu} &= \min_{\gamma} \frac{\gamma'\vec{\boldsymbol{Y}}'\boldsymbol{M}_1\boldsymbol{Z}_2\left(\boldsymbol{Z}_2'\boldsymbol{M}_1\boldsymbol{Z}_2\right)^{-1}\boldsymbol{Z}_2'\boldsymbol{M}_1\vec{\boldsymbol{Y}}\gamma}{\gamma'\frac{1}{n}\vec{\boldsymbol{Y}}'\boldsymbol{M_Z}\vec{\boldsymbol{Y}}\gamma} \\ &\le \frac{\left(\frac{1}{\sqrt{n}}\boldsymbol{e}'\boldsymbol{M}_1\boldsymbol{Z}_2\right)\left(\frac{1}{n}\boldsymbol{Z}_2'\boldsymbol{M}_1\boldsymbol{Z}_2\right)^{-1}\left(\frac{1}{\sqrt{n}}\boldsymbol{Z}_2'\boldsymbol{M}_1\boldsymbol{e}\right)}{\frac{1}{n}\boldsymbol{e}'\boldsymbol{M_Z e}} \\ &= O_p(1). \end{aligned}$$

It follows that

$$\begin{aligned} \sqrt{n}\left(\widehat{\beta}_{\text{liml}} - \beta\right) &= \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{P_Z X} - \widehat{\mu}\frac{1}{n}\boldsymbol{X}'\boldsymbol{M_Z X}\right)^{-1}\left(\frac{1}{\sqrt{n}}\boldsymbol{X}'\boldsymbol{P_Z e} - \sqrt{n}\widehat{\mu}\frac{1}{n}\boldsymbol{X}'\boldsymbol{M_Z e}\right) \\ &= \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{P_Z X} - o_p(1)\right)^{-1}\left(\frac{1}{\sqrt{n}}\boldsymbol{X}'\boldsymbol{P_Z e} - o_p(1)\right) \\ &= \sqrt{n}\left(\widehat{\beta}_{\text{2sls}} - \beta\right) + o_p(1) \end{aligned}$$

which means that LIML and 2SLS have the same asymptotic distribution. This holds under the same assumptions as for 2SLS.

Consequently, one method to obtain an asymptotically valid covariance estimator for LIML is to use the 2SLS formula. However, this is not the best choice. Rather, consider the IV representation for LIML

$$\widehat{\beta}_{\text{liml}} = \left(\widetilde{\boldsymbol{X}}'\boldsymbol{X}\right)^{-1}\left(\widetilde{\boldsymbol{X}}'\boldsymbol{Y}_1\right)$$

where

$$\widetilde{\boldsymbol{X}} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 - \widehat{\kappa}\widehat{\boldsymbol{U}}_2 \end{pmatrix}$$

and $\widehat{\boldsymbol{U}}_2 = \boldsymbol{M_Z X}_2$. The asymptotic covariance matrix formula for an IV estimator is

$$\widehat{\boldsymbol{V}}_{\beta} = \left(\frac{1}{n}\widetilde{\boldsymbol{X}}'\boldsymbol{X}\right)^{-1}\widehat{\Omega}\left(\frac{1}{n}\boldsymbol{X}'\widetilde{\boldsymbol{X}}\right)^{-1} \tag{12.42}$$

where

$$\begin{aligned} \widehat{\Omega} &= \frac{1}{n}\sum_{i=1}^{n}\widetilde{X}_i\widetilde{X}_i\widehat{e}_i^2 \\ \widehat{e}_i &= Y_{1i} - X_i'\widehat{\beta}_{\text{liml}}. \end{aligned}$$

This simplifies to the 2SLS formula when $\widehat{\kappa} = 1$ but otherwise differs. The estimator (12.42) is a better choice than the 2SLS formula for covariance matrix estimation as it takes advantage of the LIML estimator structure.

12.20 Functions of Parameters

Given the distribution theory in Theorems 12.2 and 12.3 it is straightforward to derive the asymptotic distribution of smooth nonlinear functions of the coefficient estimators.

Specifically, given a function $r\left(\beta\right):\mathbb{R}^k \rightarrow \Theta \subset \mathbb{R}^q$ we define the parameter $\theta = r\left(\beta\right)$. Given $\widehat{\beta}_{\text{2sls}}$ a natural estimator of θ is $\widehat{\theta}_{\text{2sls}} = r\left(\widehat{\beta}_{\text{2sls}}\right)$.

Consistency follows from Theorem 12.1 and the continuous mapping theorem.

Theorem 12.4 Under Assumptions 12.1 and 7.3, as $n \rightarrow \infty$, $\widehat{\theta}_{\text{2sls}} \underset{p}{\longrightarrow} \theta$.

If $r\left(\beta\right)$ is differentiable then an estimator of the asymptotic covariance matrix for $\widehat{\theta}_{\text{2sls}}$ is

$$\widehat{\boldsymbol{V}}_{\theta} = \widehat{\boldsymbol{R}}'\widehat{\boldsymbol{V}}_{\beta}\widehat{\boldsymbol{R}}$$
$$\widehat{\boldsymbol{R}} = \frac{\partial}{\partial\beta} r(\widehat{\beta}_{\text{2sls}})'.$$

We similarly define the homoskedastic variance estimator as $\widehat{\boldsymbol{V}}^0_{\theta} = \widehat{\boldsymbol{R}}'\widehat{\boldsymbol{V}}^0_{\beta}\widehat{\boldsymbol{R}}$.

The asymptotic distribution theory follows from Theorems 12.2 and 12.3 and the delta method.

Theorem 12.5 Under Assumptions 12.2 and 7.3, as $n \rightarrow \infty$,

$$\sqrt{n}\left(\widehat{\theta}_{\text{2sls}} - \theta\right) \underset{d}{\longrightarrow} \mathrm{N}\left(0, \boldsymbol{V}_{\theta}\right)$$

and $\widehat{\boldsymbol{V}}_{\theta} \underset{p}{\longrightarrow} \boldsymbol{V}_{\theta}$ where $\boldsymbol{V}_{\theta} = \boldsymbol{R}'\boldsymbol{V}_{\beta}\boldsymbol{R}$ and $\boldsymbol{R} = \dfrac{\partial}{\partial\beta}\boldsymbol{r}(\beta)'$.

When $q = 1$, a standard error for $\widehat{\theta}_{\text{2sls}}$ is $s(\widehat{\theta}_{\text{2sls}}) = \sqrt{n^{-1}\widehat{\boldsymbol{V}}_{\theta}}$.

For example, let's take the parameter estimates from the fifth column of Table 12.1, which are the 2SLS estimates with three endogenous regressors and four excluded instruments. Suppose we are interested in the return to experience, which depends on the level of experience. The estimated return at *experience*= 10 is $0.047 - 0.032 \times 2 \times 10/100 = 0.041$ and its standard error is 0.003. This implies a 4% increase in wages per year of experience and is precisely estimated. Or suppose we are interested in the level of experience at which the function maximizes. The estimate is $50 \times 0.047/0.032 = 73$. This has a standard error of 249. The large standard error implies that the estimate (73 years of experience) is without precision and is thus uninformative.

12.21 Hypothesis Tests

As in the previous section, for a given function $r\left(\beta\right):\mathbb{R}^k \rightarrow \Theta \subset \mathbb{R}^q$ we define the parameter $\theta = r\left(\beta\right)$ and consider tests of hypotheses of the form $\mathbb{H}_0 : \theta = \theta_0$ against $\mathbb{H}_1 : \theta \neq \theta_0$. The Wald statistic for $\mathbb{H}_0$ is

$$W = n\left(\widehat{\theta} - \theta_0\right)'\widehat{\boldsymbol{V}}^{-1}_{\widehat{\theta}}\left(\widehat{\theta} - \theta_0\right).$$

From Theorem 12.5 we deduce that W is asymptotically chi-square distributed. Let $G_q(u)$ denote the χ_q^2 distribution function.

Theorem 12.6 Under Assumption 12.2, Assumption 7.3, and $\mathbb{H}_0$, then as $n \to \infty$, $W \underset{d}{\longrightarrow} \chi_q^2$. For c satisfying $\alpha = 1 - G_q(c)$, $\mathbb{P}[W > c \mid \mathbb{H}_0] \longrightarrow \alpha$ so the test "Reject $\mathbb{H}_0$ if $W > c$" has asymptotic size α.

In linear regression we often report the F version of the Wald statistic (by dividing by degrees of freedom) and use the F distribution for inference as this is justified in the normal sampling model. For 2SLS estimation, however, this is not done as there is no finite sample F justification for the F version of the Wald statistic.

To illustrate, once again let's take the parameter estimates from the fifth column of Table 12.1 and again consider the return to experience which is determined by the coefficients on *experience* and *experience*$^2/100$. Neither coefficient is statisticially significant at the 5% level and it is unclear if the overall effect is statistically significant. We can assess this by testing the joint hypothesis that both coefficients are zero. The Wald statistic for this hypothesis is $W = 244$ which is highly significant with an asymptotic p-value of 0.0000. Thus by examining the joint test in contrast to the individual tests is quite clear that experience has a non-zero effect.

12.22 Finite Sample Theory

In Chapter 5 we reviewed the rich exact distribution available for the linear regression model under the assumption of normal innovations. There is a similarly rich literature in econometrics for IV, 2SLS and LIML estimators. An excellent review of the theory, mostly developed in the 1970s and early 1980s, is provided by Peter Phillips (1983).

This theory was developed under the assumption that the structural error vector e and reduced form error u_2 are multivariate normally distributed. Even though the errors are normal, IV-type estimators are nonlinear functions of these errors and are thus non-normally distributed. Formulae for the exact distributions have been derived but are unfortunately functions of model parameters and hence are not directly useful for finite sample inference.

One important implication of this literature is that even in this optimal context of exact normal innovations the finite sample distributions of the IV estimators are non-normal and the finite sample distributions of test statistics are not chi-squared. The normal and chi-squared approximations hold asymptotically but there is no reason to expect these approximations to be accurate in finite samples.

A second important result is that under the assumption of normal errors most of the estimators do not have finite moments in any finite sample. A clean statement concerning the existence of moments for the 2SLS estimator was obtained by Kinal (1980) for the case of joint normality. Let $\widehat{\beta}_{\text{2sls},2}$ be the 2SLS estimators of the coefficients on the endogeneous regressors.

Theorem 12.7 If (Y, X, Z) are jointly normal, then for any r, $\mathbb{E}\left\|\widehat{\beta}_{\text{2sls},2}\right\|^r < \infty$ if and only if $r < \ell_2 - k_2 + 1$.

This result states that in the just-identified case the IV estimator does not have any finite order integer moments. In the over-identified case the number of finite moments corresponds to the number of overidentifying restrictions ($\ell_2 - k_2$). Thus if there is one over-identifying restriction 2SLS has a finite mean and if there are two over-identifying restrictions then the 2SLS estimator has a finite variance.

The LIML estimator has a more severe moment problem as it has no finite integer moments (Mariano, 1982) regardless of the number of over-identifying restrictions. Due to this lack of moments Fuller (1977) proposed the following modification of LIML. His estimator is

$$\begin{aligned}\widehat{\beta}_{\text{Fuller}} &= \left(\boldsymbol{X}'\left(\boldsymbol{I}_n - K\boldsymbol{M}_{\boldsymbol{Z}}\right)\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\left(\boldsymbol{I}_n - K\boldsymbol{M}_{\boldsymbol{Z}}\right)\boldsymbol{Y}_1\right)\\ K &= \widehat{\kappa} - \frac{C}{n-k}\end{aligned}$$

for some $C \geq 1$. Fuller showed that his estimator has all moments finite under suitable conditions.

Hausman, Newey, Woutersen, Chao and Swanson (2012) propose an estimator they call HFUL which combines the ideas of JIVE and Fuller which has excellent finite sample properties.

12.23 Bootstrap for 2SLS

The standard bootstrap algorithm for IV, 2SLS, and GMM generates bootstrap samples by sampling the triplets (Y_{1i}^*, X_i^*, Z_i^*) independently and with replacement from the original sample $\{(Y_{1i}, X_i, Z_i) : i = 1, ..., n\}$. Sampling n such observations and stacking into observation matrices $(\boldsymbol{Y}_1^*, \boldsymbol{X}^*, \boldsymbol{Z}^*)$, the bootstrap 2SLS estimator is

$$\widehat{\beta}_{\text{2sls}}^* = \left(\boldsymbol{X}^{*\prime}\boldsymbol{Z}^*\left(\boldsymbol{Z}^{*\prime}\boldsymbol{Z}^*\right)^{-1}\boldsymbol{Z}^{*\prime}\boldsymbol{X}^*\right)^{-1}\boldsymbol{X}^{*\prime}\boldsymbol{Z}^*\left(\boldsymbol{Z}^{*\prime}\boldsymbol{Z}^*\right)^{-1}\boldsymbol{Z}^{*\prime}\boldsymbol{Y}_1^*.$$

This is repeated B times to create a sample of B bootstrap draws. Given these draws bootstrap statistics can be calculated. This includes the bootstrap estimate of variance, standard errors, and confidence intervals, including percentile, BC percentile, BC$_a$ and percentile-t.

We now show that the bootstrap estimator has the same asymptotic distribution as the sample estimator. For overidentified cases this demonstration requires a bit of extra care. This was first shown by Hahn (1996).

The sample observations satisfy the model $Y_1 = X'\beta + e$ with $\mathbb{E}[Ze] = 0$. The true value of β in the population can be written as

$$\beta = \left(\mathbb{E}\left[XZ'\right]\mathbb{E}\left[ZZ'\right]^{-1}\mathbb{E}\left[ZX'\right]\right)^{-1}\mathbb{E}\left[XZ'\right]\mathbb{E}\left[ZZ'\right]^{-1}\mathbb{E}\left[ZY_1\right].$$

The true value in the bootstrap universe is obtained by replacing the population moments by the sample moments, which equals the 2SLS estimator

$$\begin{aligned}&\left(\mathbb{E}^*\left[X^*Z^{*\prime}\right]\mathbb{E}^*\left[Z^*Z^{*\prime}\right]^{-1}\mathbb{E}^*\left[Z^*X^{*\prime}\right]\right)^{-1}\mathbb{E}^*\left[X^*Z^{*\prime}\right]\mathbb{E}^*\left[Z^*Z^{*\prime}\right]^{-1}\mathbb{E}^*\left[Z^*Y_1^*\right]\\ &= \left(\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\right)\left(\frac{1}{n}\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\left(\frac{1}{n}\boldsymbol{Z}'\boldsymbol{X}\right)\right)^{-1}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\right)\left(\frac{1}{n}\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\left[\frac{1}{n}\boldsymbol{Z}'\boldsymbol{Y}_1\right]\\ &= \widehat{\beta}_{\text{2sls}}.\end{aligned}$$

The bootstrap observations thus satisfy the equation $Y_{1i}^* = X_i^{*\prime}\widehat{\beta}_{\text{2sls}} + e_i^*$. In matrix notation for the sample this is

$$\boldsymbol{Y}_1^* = \boldsymbol{X}^{*\prime}\widehat{\beta}_{\text{2sls}} + \boldsymbol{e}^*. \tag{12.43}$$

Given a bootstrap triple $(Y_{1i}^*, X_i^*, Z_i^*) = (Y_{1j}, X_j, Z_j)$ for some observation j the true bootstrap error is

$$e_i^* = Y_{1j} - X_j' \widehat{\beta}_{\text{2sls}} = \widehat{e}_j.$$

It follows that

$$\mathbb{E}^* \left[Z^* e^* \right] = n^{-1} \boldsymbol{Z}' \widehat{\boldsymbol{e}}. \tag{12.44}$$

This is generally not equal to zero in the over-identified case.

This an an important complication. In over-identified models the true observations satisfy the population condition $\mathbb{E}[Ze] = 0$ but in the bootstrap sample $\mathbb{E}^*[Z^* e^*] \neq 0$. This means that to apply the central limit theorem to the bootstrap estimator we first have to recenter the moment condition. That is, (12.44) and the bootstrap CLT imply

$$\frac{1}{\sqrt{n}} \left(\boldsymbol{Z}^{*\prime} \boldsymbol{e}^* - \boldsymbol{Z}' \widehat{\boldsymbol{e}} \right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(Z_i^* e_i^* - \mathbb{E}^* \left[Z^* e^* \right] \right) \underset{d^*}{\longrightarrow} \mathrm{N}(0, \Omega) \tag{12.45}$$

where

$$\Omega = \mathbb{E}\left[Z Z' e^2 \right].$$

Using (12.43) we can normalize the bootstrap estimator as

$$\begin{aligned}
\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{\text{2sls}}^* - \widehat{\boldsymbol{\beta}}_{\text{2sls}}\right) &= \sqrt{n}\left(\boldsymbol{X}^{*\prime}\boldsymbol{Z}^* \left(\boldsymbol{Z}^{*\prime}\boldsymbol{Z}^*\right)^{-1} \boldsymbol{Z}^{*\prime}\boldsymbol{X}^*\right)^{-1} \boldsymbol{X}^{*\prime}\boldsymbol{Z}^* \left(\boldsymbol{Z}^{*\prime}\boldsymbol{Z}^*\right)^{-1} \boldsymbol{Z}^{*\prime}\boldsymbol{e}^* \\
&= \left(\left(\frac{1}{n}\boldsymbol{X}^{*\prime}\boldsymbol{Z}^*\right)\left(\frac{1}{n}\boldsymbol{Z}^{*\prime}\boldsymbol{Z}^*\right)^{-1}\left(\frac{1}{n}\boldsymbol{Z}^{*\prime}\boldsymbol{X}^*\right)\right)^{-1} \\
&\quad \times \left(\frac{1}{n}\boldsymbol{X}^{*\prime}\boldsymbol{Z}^*\right)\left(\frac{1}{n}\boldsymbol{Z}^{*\prime}\boldsymbol{Z}^*\right)^{-1} \frac{1}{\sqrt{n}}\left(\boldsymbol{Z}^{*\prime}\boldsymbol{e}^* - \boldsymbol{Z}'\widehat{\boldsymbol{e}}\right) &\quad (12.46) \\
&\quad + \left(\left(\frac{1}{n}\boldsymbol{X}^{*\prime}\boldsymbol{Z}^*\right)\left(\frac{1}{n}\boldsymbol{Z}^{*\prime}\boldsymbol{Z}^*\right)^{-1}\left(\frac{1}{n}\boldsymbol{Z}^{*\prime}\boldsymbol{X}^*\right)\right)^{-1} \\
&\quad \times \left(\frac{1}{n}\boldsymbol{X}^{*\prime}\boldsymbol{Z}^*\right)\left(\frac{1}{n}\boldsymbol{Z}^{*\prime}\boldsymbol{Z}^*\right)^{-1} \left(\frac{1}{\sqrt{n}}\boldsymbol{Z}'\widehat{\boldsymbol{e}}\right). &\quad (12.47)
\end{aligned}$$

Using the bootstrap WLLN,

$$\begin{aligned}
\frac{1}{n}\boldsymbol{X}^{*\prime}\boldsymbol{Z}^* &= \frac{1}{n}\boldsymbol{X}'\boldsymbol{Z} + o_p(1) \\
\frac{1}{n}\boldsymbol{Z}^{*\prime}\boldsymbol{Z}^* &= \frac{1}{n}\boldsymbol{Z}'\boldsymbol{Z} + o_p(1).
\end{aligned}$$

This implies (12.47) is equal to

$$\sqrt{n}\left(\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{X}\right)\right)^{-1} \boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1} \boldsymbol{Z}'\widehat{\boldsymbol{e}} + o_p(1) = 0 + o_p(1).$$

The equality holds because the 2SLS first-order condition implies $\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\widehat{\boldsymbol{e}} = 0$. Also, combined with (12.45) we see that (12.46) converges in bootstrap distribution to

$$\left(\boldsymbol{Q}_{XZ}\boldsymbol{Q}_{ZZ}^{-1}\boldsymbol{Q}_{ZX}\right)^{-1}\boldsymbol{Q}_{XZ}\boldsymbol{Q}_{ZZ}^{-1}\mathrm{N}(0,\Omega) = \mathrm{N}\left(0, \boldsymbol{V}_\beta\right)$$

where $\boldsymbol{V}_\beta$ is the 2SLS asymptotic variance from Theorem 12.2. This is the asymptotic distribution of $\sqrt{n}\left(\widehat{\beta}_{\text{2sls}}^* - \widehat{\beta}_{\text{2sls}}\right)$.

By standard calculations we can also show that bootstrap t-ratios are asymptotically normal.

Theorem 12.8 Under Assumption 12.2, as $n \to \infty$

$$\sqrt{n}\left(\widehat{\beta}^*_{\text{2sls}} - \widehat{\beta}_{\text{2sls}}\right) \underset{d^*}{\longrightarrow} \mathrm{N}\left(0, \boldsymbol{V}_\beta\right)$$

where $\boldsymbol{V}_\beta$ is the 2SLS asymptotic variance from Theorem 12.2. Furthermore,

$$T^* = \frac{\sqrt{n}\left(\widehat{\beta}^*_{\text{2sls}} - \widehat{\beta}_{\text{2sls}}\right)}{s\left(\widehat{\beta}^*_{\text{2sls}}\right)} \underset{d^*}{\longrightarrow} \mathrm{N}(0,1).$$

This shows that percentile-type and percentile-t confidence intervals are asymptotically valid.

One might expect that the asymptotic refinement arguments extend to the BC$_a$ and percentile-t methods but this does not appear to be the case. While $\sqrt{n}\left(\widehat{\beta}^*_{\text{2sls}} - \widehat{\beta}_{\text{2sls}}\right)$ and $\sqrt{n}\left(\widehat{\beta}_{\text{2sls}} - \beta\right)$ have the same asymptotic distribution they differ in finite samples by an $O_p\left(n^{-1/2}\right)$ term. This means that they have distinct Edgeworth expansions. Consequently, unadjusted bootstrap methods will not achieve an asymptotic refinement.

An alternative suggested by Hall and Horowitz (1996) is to recenter the bootstrap 2SLS estimator so that it satisfies the correct orthogonality condition. Define

$$\widehat{\beta}^{**}_{\text{2sls}} = \left(\boldsymbol{X}^{*\prime}\boldsymbol{Z}^*\left(\boldsymbol{Z}^{*\prime}\boldsymbol{Z}^*\right)^{-1}\boldsymbol{Z}^{*\prime}\boldsymbol{X}^*\right)^{-1}\boldsymbol{X}^{*\prime}\boldsymbol{Z}^*\left(\boldsymbol{Z}^{*\prime}\boldsymbol{Z}^*\right)^{-1}\left(\boldsymbol{Z}^{*\prime}\boldsymbol{Y}^*_1 - \boldsymbol{Z}'\widehat{\boldsymbol{e}}\right).$$

We can see that

$$\begin{aligned}\sqrt{n}\left(\widehat{\beta}^{**}_{\text{2sls}} - \widehat{\beta}_{\text{2sls}}\right) &= \left(\frac{1}{n}\boldsymbol{X}^{*\prime}\boldsymbol{Z}^*\left(\frac{1}{n}\boldsymbol{Z}^{*\prime}\boldsymbol{Z}^*\right)^{-1}\frac{1}{n}\boldsymbol{Z}^{*\prime}\boldsymbol{X}^*\right)^{-1} \\ &\quad\times\left(\frac{1}{n}\boldsymbol{X}^{*\prime}\boldsymbol{Z}^*\right)\left(\frac{1}{n}\boldsymbol{Z}^{*\prime}\boldsymbol{Z}^*\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n\left(Z^*_i e^*_i - \mathbb{E}^*\left[Z^* e^*\right]\right)\right)\end{aligned}$$

which converges to the $\mathrm{N}\left(0, \boldsymbol{V}_\beta\right)$ distribution without special handling. Hall and Horowitz (1996) show that percentile-t methods applied to $\widehat{\beta}^{**}_{\text{2sls}}$ achieve an asymptotic refinement and are thus preferred to the unadjusted bootstrap estimator.

This recentered estimator, however, is not the standard implementation of the bootstrap for 2SLS as used in empirical practice.

12.24 The Peril of Bootstrap 2SLS Standard Errors

It is tempting to use the bootstrap algorithm to estimate variance matrices and standard errors for the 2SLS estimator. In fact this is one of the most common uses of bootstrap methods in current econometric practice. Unfortunately this is an unjustified and ill-conceived idea and should not be done. In finite samples the 2SLS estimator may not have a finite second moment, meaning that bootstrap variance estimates are unstable and unreliable.

Theorem 12.7 shows that under jointly normality the 2SLS estimator will have a finite variance if and only if the number of overidentifying restrictions is two or larger. Thus for just-identified IV, and 2SLS with one degree of overidentification, the finite sample variance is infinite. The bootstrap will be attempting to estimate this value – infinity – and will yield nonsensical answers. When the observations are not jointly normal there is no finite sample theory (so it is possible that the finite sample variance is actually finite) but this is unknown and unverifiable.

In overidentified settings when the number of overidentifying restrictions is two or larger the bootstrap can be applied for standard error estimation. However this is not the most common application of IV methods in econometric practice and thus should be viewed as the exception rather than the norm.

To understand what is going on consider the simplest case of a just-identified model with a single endogenous regressor and no included exogenous regressors. In this case the estimator can be written as a ratio of means

$$\widehat{\beta}_{\mathrm{iv}}-\beta=\frac{\sum_{i=1}^{n} Z_i e_i}{\sum_{i=1}^{n} Z_i X_i}.$$

Under joint normality of (e_i, X_i) this has a Cauchy-like distribution which does not possess any finite integer moments. The trouble is that the denominator can be either positive or negative, and arbitrarily close to zero. This means that the ratio can take arbitrarily large values.

To illustrate let us return to the basic Card IV wage regression from column 2 of Table 12.1 which uses *college* as an instrument for *education*. We estimate this equation for the subsample of Black men which has $n = 703$ observations, and focus on the coefficient for the return to education. The coefficient estimate is reported in Table 12.3, along with asymptotic, jackknife, and two bootstrap standard errors each calculated with 10,000 bootstrap replications.

Table 12.3: Instrumental Variable Return to Education for Black Men

Estimate	0.11
Asymptotic s.e.	(0.11)
Jackknife s.e.	(0.11)
Bootstrap s.e. (standard)	(1.42)
Bootstrap s.e. (repeat)	(4.79)

The bootstrap standard errors are an order of magnitude larger than the asymptotic standard errors, and vary substantially across the bootstrap runs despite using 10,000 bootstrap replications. This indicates moment failure and unreliability of the bootstrap standard errors.

This is a strong message that **bootstrap standard errors should not be computed for IV estimators**. Instead, report percentile-type confidence intervals. A further cautionary remark is that in finite samples percentile confidence intervals also may have poor coverage rates, especially in contexts such as the results of Table 12.3.

12.25 Clustered Dependence

In Section 4.23 we introduced clustered dependence. We can also use the methods of clustered dependence for 2SLS estimation. Recall, the g^{th} cluster has the observations $\boldsymbol{Y}_g=(Y_{1g},...,Y_{n_g g})'$, $\boldsymbol{X}_g=(X_{1g},...,X_{n_g g})'$, and $\boldsymbol{Z}_g=(Z_{1g},...,Z_{n_g g})'$. The structural equation for the g^{th} cluster can be written as the matrix system $\boldsymbol{Y}_g=\boldsymbol{X}_g\beta+\boldsymbol{e}_g$. Using this notation the centered 2SLS estimator can be written as

$$\begin{aligned}\widehat{\beta}_{2\mathrm{sls}}-\beta&=\left(\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{e}\\&=\left(\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\left(\sum_{g=1}^{G}\boldsymbol{Z}_g'\boldsymbol{e}_g\right).\end{aligned}$$

The cluster-robust covariance matrix estimator for $\widehat{\beta}_{2\mathrm{sls}}$ thus takes the form

$$\widehat{\boldsymbol{V}}_{\beta}=\left(\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\widehat{\boldsymbol{S}}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}$$

with

$$\widehat{\boldsymbol{S}} = \sum_{g=1}^{G} \boldsymbol{Z}_g' \widehat{\boldsymbol{e}}_g \widehat{\boldsymbol{e}}_g' \boldsymbol{Z}_g$$

and the clustered residuals $\widehat{\boldsymbol{e}}_g = \boldsymbol{Y}_g - \boldsymbol{X}_g \widehat{\beta}_{\mathrm{2sls}}$.

The difference between the heteroskedasticity-robust estimator and the cluster-robust estimator is the covariance estimator $\widehat{\boldsymbol{S}}$.

12.26 Generated Regressors

The "two-stage" form of the 2SLS estimator is an example of what is called "estimation with generated regressors". We say a regressor is a **generated** if it is an estimate of an idealized regressor or if it is a function of estimated parameters. Typically, a generated regressor $\widehat{W}$ is an estimate of an unobserved ideal regressor W. As an estimate, $\widehat{W}_i$ is a function of the full sample not just observation i. Hence it is not "i.i.d." as it is dependent across observations which invalidates the conventional regression assumptions. Consequently, the sampling distribution of regression estimates is affected. Unless this is incorporated into our inference methods, covariance matrix estimates and standard errors will be incorrect.

The econometric theory of generated regressors was developed by Pagan (1984) for linear models and extended to nonlinear models and more general two-step estimators by Pagan (1986). Independently, similar results were obtained by Murphy and Topel (1985). Here we focus on the linear model:

$$\begin{aligned} Y &= W'\beta + v \qquad (12.48)\\ W &= \boldsymbol{A}'Z \\ \mathbb{E}[Zv] &= 0. \end{aligned}$$

The observables are (Y, Z). We also have an estimate $\widehat{\boldsymbol{A}}$ of $\boldsymbol{A}$.

Given $\widehat{\boldsymbol{A}}$ we construct the estimate $\widehat{W}_i = \widehat{\boldsymbol{A}}'Z_i$ of W_i, replace W_i in (12.48) with $\widehat{W}_i$, and then estimate β by least squares, resulting in the estimator

$$\widehat{\beta} = \left(\sum_{i=1}^{n} \widehat{W}_i \widehat{W}_i'\right)^{-1} \left(\sum_{i=1}^{n} \widehat{W}_i Y_i\right). \qquad (12.49)$$

The regressors $\widehat{W}_i$ are called **generated regressors**. The properties of $\widehat{\beta}$ are different than least squares with i.i.d. observations since the generated regressors are themselves estimates.

This framework includes 2SLS as well as other common estimators. The 2SLS model can be written as (12.48) by looking at the reduced form equation (12.13), with $W = \Gamma'Z$, $\boldsymbol{A} = \Gamma$, and $\widehat{\boldsymbol{A}} = \widehat{\Gamma}$.

The examples which motivated Pagan (1984) and Murphy and Topel (1985) emerged from the macroeconomics literature, in particular the work of Barro (1977) which examined the impact of inflation expectations and expectation errors on economic output. Let π denote realized inflation and Z be variables available to economic agents. A model of inflation expectations sets $W = \mathbb{E}[\pi \mid Z] = \gamma'Z$ and a model of expectation error sets $W = \pi - \mathbb{E}[\pi \mid Z] = \pi - \gamma'Z$. Since expectations and errors are not observed they are replaced in applications with the fitted values $\widehat{W}_i = \widehat{\gamma}'Z_i$ and residuals $\widehat{W}_i = \pi_i - \widehat{\gamma}'Z_i$ where $\widehat{\gamma}$ is the coefficient from a regression of π on Z.

The generated regressor framework includes all of these examples.

The goal is to obtain a distributional approximation for $\widehat{\beta}$ in order to construct standard errors, confidence intervals, and tests. Start by substituting equation (12.48) into (12.49). We obtain

$$\widehat{\beta} = \left(\sum_{i=1}^{n} \widehat{W}_i \widehat{W}_i'\right)^{-1} \left(\sum_{i=1}^{n} \widehat{W}_i \left(W_i'\beta + v_i\right)\right).$$

Next, substitute $W_i'\beta = \widehat{W}_i'\beta + \left(W_i - \widehat{W}_i\right)'\beta$. We obtain

$$\widehat{\beta} - \beta = \left(\sum_{i=1}^{n} \widehat{W}_i \widehat{W}_i'\right)^{-1} \left(\sum_{i=1}^{n} \widehat{W}_i \left(\left(W_i - \widehat{W}_i\right)'\beta + v_i\right)\right). \tag{12.50}$$

Effectively, this shows that the distribution of $\widehat{\beta} - \beta$ has two random components, one due to the conventional regression component and the second due to the generated regressor. Conventional variance estimators do not address this second component and thus will be biased.

Interestingly, the distribution in (12.50) dramatically simplifies in the special case that the "generated regressor term" $\left(W_i - \widehat{W}_i\right)'\beta$ disappears. This occurs when the slope coefficients on the generated regressors are zero. To be specific, partition $W_i = (W_{1i}, W_{2i})$, $\widehat{W}_i = \left(W_{1i}, \widehat{W}_{2i}\right)$, and $\beta = \left(\beta_1, \beta_2\right)$ so that W_{1i} are the conventional observed regressors and $\widehat{W}_{2i}$ are the generated regressors. Then $\left(W_i - \widehat{W}_i\right)'\beta = \left(W_{2i} - \widehat{W}_{2i}\right)'\beta_2$. Thus if $\beta_2 = 0$ this term disappears. In this case (12.50) equals

$$\widehat{\beta} - \widehat{\beta} = \left(\sum_{i=1}^{n} \widehat{W}_i \widehat{W}_i'\right)^{-1} \left(\sum_{i=1}^{n} \widehat{W}_i v_i\right).$$

This is a dramatic simplification.

Furthermore, since $\widehat{W}_i = \widehat{\boldsymbol{A}}' Z_i$ we can write the estimator as a function of sample moments:

$$\sqrt{n}\left(\widehat{\beta} - \beta\right) = \left(\widehat{\boldsymbol{A}}' \left(\frac{1}{n}\sum_{i=1}^{n} Z_i Z_i'\right) \widehat{\boldsymbol{A}}\right)^{-1} \widehat{\boldsymbol{A}}' \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i v_i\right).$$

If $\widehat{\boldsymbol{A}} \underset{p}{\longrightarrow} \boldsymbol{A}$ we find from standard manipulations that $\sqrt{n}\left(\widehat{\beta} - \beta\right) \underset{d}{\longrightarrow} \mathrm{N}\left(0, \boldsymbol{V}_\beta\right)$ where

$$\boldsymbol{V}_{\boldsymbol{\beta}} = \left(\boldsymbol{A}'\mathbb{E}\left[ZZ'\right]\boldsymbol{A}\right)^{-1} \left(\boldsymbol{A}'\mathbb{E}\left[ZZ'v^2\right]\boldsymbol{A}\right) \left(\boldsymbol{A}'\mathbb{E}\left[ZZ'\right]\boldsymbol{A}\right)^{-1}. \tag{12.51}$$

The conventional asymptotic covariance matrix estimator for $\widehat{\beta}$ takes the form

$$\widehat{\boldsymbol{V}}_\beta = \left(\frac{1}{n}\sum_{i=1}^{n} \widehat{W}_i \widehat{W}_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} \widehat{W}_i \widehat{W}_i' \widehat{v}_i^2\right) \left(\frac{1}{n}\sum_{i=1}^{n} \widehat{W}_i \widehat{W}_i'\right)^{-1} \tag{12.52}$$

where $\widehat{v}_i = Y_i - \widehat{W}_i'\widehat{\beta}$. Under the given assumptions, $\widehat{\boldsymbol{V}}_\beta \underset{p}{\longrightarrow} \boldsymbol{V}_\beta$. Thus inference using $\widehat{\boldsymbol{V}}_\beta$ is asymptotically valid. This is useful when we are interested in tests of $\beta_2 = 0$. Often this is of major interest in applications.

To test $\mathbb{H}_0 : \beta_2 = 0$ we partition $\widehat{\beta} = \left(\widehat{\beta}_1, \widehat{\beta}_2\right)$ and construct a conventional Wald statistic

$$W = n\widehat{\beta}_2' \left(\left[\widehat{\boldsymbol{V}}_\beta\right]_{22}\right)^{-1} \widehat{\beta}_2.$$

Theorem 12.9 Take model (12.48) with $\mathbb{E}\left[Y^4\right] < \infty$, $\mathbb{E}\|Z\|^4 < \infty$, $\boldsymbol{A}'\mathbb{E}\left[ZZ'\right]\boldsymbol{A} > 0$, $\widehat{\boldsymbol{A}} \underset{p}{\longrightarrow} \boldsymbol{A}$, and $\widehat{W}_i = \left(W_{1i}, \widehat{W}_{2i}\right)$. Under $\mathbb{H}_0 : \beta_2 = 0$, as $n \to \infty$, $\sqrt{n}\left(\widehat{\beta} - \beta\right) \underset{d}{\longrightarrow} \mathrm{N}\left(0, \boldsymbol{V}_\beta\right)$ where $\boldsymbol{V}_\beta$ is given in (12.51). For $\widehat{\boldsymbol{V}}_\beta$ given in (12.52), $\widehat{\boldsymbol{V}}_\beta \underset{p}{\longrightarrow} \boldsymbol{V}_\beta$. Furthermore, $W \underset{d}{\longrightarrow} \chi_q^2$ where $q = \dim(\beta_2)$. For c satisfying $\alpha = 1 - G_q(c)$, $\mathbb{P}\left[W > c \mid \mathbb{H}_0\right] \to \alpha$, so the test "Reject $\mathbb{H}_0$ if $W > c$" has asymptotic size α.

In the special case that $\widehat{\boldsymbol{A}} = \boldsymbol{A}(\boldsymbol{X},\boldsymbol{Z})$ and $v \mid X,Z \sim \mathrm{N}\left(0,\sigma^2\right)$ there is a finite sample version of the previous result. Let W^0 be the Wald statistic constructed with a homoskedastic covariance matrix estimator, and let

$$F = W/q \tag{12.53}$$

be the the F statistic, where $q = \dim(\beta_2)$.

Theorem 12.10 Take model (12.48) with $\widehat{\boldsymbol{A}} = \boldsymbol{A}(\boldsymbol{X},\boldsymbol{Z})$, $v \mid X,Z \sim \mathrm{N}\left(0,\sigma^2\right)$ and $\widehat{W} = \left(W_1, \widehat{W}_2\right)$. Under $\mathbb{H}_0 : \beta_2 = 0$, t-statistics have exact $\mathrm{N}(0,1)$ distributions, and the F statistic (12.53) has an exact $F_{q,n-k}$ distribution where $q = \dim(\beta_2)$ and $k = \dim(\beta)$.

To summarize, in the model $Y = W_1'\beta_1 + W_2'\beta_2 + v$ where W_2 is not observed but replaced with an estimate $\widehat{W}_2$, conventional significance tests for $\mathbb{H}_0 : \beta_2 = 0$ are asymptotically valid without adjustment.

While this theory allows tests of $\mathbb{H}_0 : \beta_2 = 0$ it unfortunately does not justify conventional standard errors or confidence intervals. For this, we need to work out the distribution without imposing the simplification $\beta_2 = 0$. This often needs to be worked out case-by-case or by using methods based on the generalized method of moments to be introduced in Chapter 13. However, in one important set of examples it is straightforward to work out the asymptotic distribution.

For the remainder of this section we examine the setting where the estimators $\widehat{\boldsymbol{A}}$ take a least squares form so for some $\boldsymbol{X}$ can be written as $\widehat{\boldsymbol{A}} = \left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{X}\right)$. Such estimators correspond to the multivariate projection model

$$\begin{aligned} X &= \boldsymbol{A}'Z + u \qquad (12.54)\\ \mathbb{E}\left[Zu'\right] &= 0. \end{aligned}$$

This class of estimators includes 2SLS and the expectation model described above. We can write the matrix of generated regressors as $\widehat{\boldsymbol{W}} = \boldsymbol{Z}\widehat{\boldsymbol{A}}$ and then (12.50) as

$$\begin{aligned} \widehat{\beta} - \beta &= \left(\widehat{\boldsymbol{W}}'\widehat{\boldsymbol{W}}\right)^{-1}\left(\widehat{\boldsymbol{W}}'\left(\left(\boldsymbol{W} - \widehat{\boldsymbol{W}}\right)\beta + \boldsymbol{v}\right)\right)\\ &= \left(\widehat{\boldsymbol{A}}'\boldsymbol{Z}'\boldsymbol{Z}\widehat{\boldsymbol{A}}\right)^{-1}\left(\widehat{\boldsymbol{A}}'\boldsymbol{Z}'\left(-\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{U}\right)\beta + \boldsymbol{v}\right)\right)\\ &= \left(\widehat{\boldsymbol{A}}'\boldsymbol{Z}'\boldsymbol{Z}\widehat{\boldsymbol{A}}\right)^{-1}\left(\widehat{\boldsymbol{A}}'\boldsymbol{Z}'\left(-\boldsymbol{U}\beta + \boldsymbol{v}\right)\right)\\ &= \left(\widehat{\boldsymbol{A}}'\boldsymbol{Z}'\boldsymbol{Z}\widehat{\boldsymbol{A}}\right)^{-1}\left(\widehat{\boldsymbol{A}}'\boldsymbol{Z}'\boldsymbol{e}\right) \end{aligned}$$

where

$$e = v - u'\beta = Y - X'\beta. \tag{12.55}$$

This estimator has the asymptotic distribution $\sqrt{n}\left(\widehat{\beta} - \beta\right) \xrightarrow[d]{} \mathrm{N}\left(0, \boldsymbol{V}_\beta\right)$ where

$$\boldsymbol{V}_\beta = \left(\boldsymbol{A}'\mathbb{E}\left[ZZ'\right]\boldsymbol{A}\right)^{-1}\left(\boldsymbol{A}'\mathbb{E}\left[ZZ'e^2\right]\boldsymbol{A}\right)\left(\boldsymbol{A}'\mathbb{E}\left[ZZ'\right]\boldsymbol{A}\right)^{-1}. \tag{12.56}$$

Under conditional homoskedasticity the covariance matrix simplifies to

$$\boldsymbol{V}_{\boldsymbol{\beta}} = \left(\boldsymbol{A}'\mathbb{E}\left[ZZ'\right]\boldsymbol{A}\right)^{-1}\mathbb{E}\left[e^2\right].$$

An appropriate estimator of $\boldsymbol{V}_\beta$ is

$$\begin{aligned}\widehat{\boldsymbol{V}}_\beta &= \left(\frac{1}{n}\widehat{\boldsymbol{W}}'\widehat{\boldsymbol{W}}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n \widehat{W}_i\widehat{W}_i'\widehat{e}_i^2\right)\left(\frac{1}{n}\widehat{\boldsymbol{W}}'\widehat{\boldsymbol{W}}\right)^{-1} \qquad (12.57)\\ \widehat{e}_i &= Y_i - X_i'\widehat{\beta}.\end{aligned}$$

Under the assumption of conditional homoskedasticity this can be simplified as usual.

This appears to be the usual covariance matrix estimator, but it is not because the least squares residuals $\widehat{v}_i = Y_i - \widehat{W}_i'\widehat{\beta}$ have been replaced with $\widehat{e}_i$. This is exactly the substitution made by the 2SLS covariance matrix formula. Indeed, the covariance matrix estimator $\widehat{\boldsymbol{V}}_\beta$ precisely equals (12.40).

Theorem 12.11 Take model (12.48) and (12.54) with $\mathbb{E}\left[Y^4\right] < \infty$, $\mathbb{E}\|Z\|^4 < \infty$, $\boldsymbol{A}'\mathbb{E}\left[ZZ'\right]\boldsymbol{A} > 0$, and $\widehat{\boldsymbol{A}} = \left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{X}\right)$. As $n \to \infty$, $\sqrt{n}\left(\widehat{\beta} - \beta\right) \xrightarrow[d]{} \mathrm{N}\left(0, \boldsymbol{V}_\beta\right)$ where $\boldsymbol{V}_\beta$ is given in (12.56) with e defined in (12.55). For $\widehat{\boldsymbol{V}}_\beta$ given in (12.57), $\widehat{\boldsymbol{V}}_\beta \xrightarrow[p]{} \boldsymbol{V}_\beta$.

Since the parameter estimators are asymptotically normal and the covariance matrix is consistently estimated, standard errors and test statistics constructed from $\widehat{\boldsymbol{V}}_\beta$ are asymptotically valid with conventional interpretations.

We now summarize the results of this section. In general, care needs to be exercised when estimating models with generated regressors. As a general rule, generated regressors and two-step estimation affects sampling distributions and variance matrices. An important simplication occurs for tests that the generated regressors have zero slopes. In this case conventional tests have conventional distributions, both asymptotically and in finite samples. Another important special case occurs when the generated regressors are least squares fitted values. In this case the asymptotic distribution takes a conventional form but the conventional residual needs to be replaced by one constructed with the forecasted variable. With this one modification asymptotic inference using the generated regressors is conventional.

12.27 Regression with Expectation Errors

In this section we examine a generated regressor model which includes expectation errors in the regression. This is an important class of generated regressor models and is relatively straightforward to characterize. The model is

$$\begin{aligned}Y &= X'\beta + u'\alpha + v\\ W &= \boldsymbol{A}'Z\\ X &= W + u\\ \mathbb{E}[Zv] &= 0\\ \mathbb{E}[uv] &= 0\\ \mathbb{E}\left[Zu'\right] &= 0.\end{aligned}$$

The observables are (Y, X, Z). This model states that W is the expectation of X (or more generally, the projection of X on Z) and u is its expectation error. The model allows for exogenous regressors as in the standard IV model if they are listed in W, X, and Z. This model is used, for example, to decompose the effect of expectations from expectation errors. In some cases it is desired to include only the expectation error u, not the expectation W. This does not change the results described here.

The model is estimated as follows. First, $\boldsymbol{A}$ is estimated by multivariate least squares of X on Z, $\widehat{\boldsymbol{A}} = \left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{X}\right)$, which yields as by-products the fitted values $\widehat{W}_i = \widehat{\boldsymbol{A}}' Z_i$ and residuals $\widehat{u}_i = \widehat{X}_i - \widehat{W}_i$. Second, the coefficients are estimated by least squares of Y on the fitted values $\widehat{W}$ and residuals $\widehat{u}$

$$Y_i = \widehat{W}_i'\widehat{\beta} + \widehat{u}_i'\widehat{\alpha} + \widehat{v}_i.$$

We now examine the asymptotic distributions of these estimators.

By the first-step regression $\boldsymbol{Z}'\widehat{\boldsymbol{U}} = 0$, $\widehat{\boldsymbol{W}}'\widehat{\boldsymbol{U}} = 0$ and $\boldsymbol{W}'\widehat{\boldsymbol{U}} = 0$. This means that $\widehat{\beta}$ and $\widehat{\alpha}$ can be computed separately. Notice that

$$\widehat{\beta} = \left(\widehat{\boldsymbol{W}}'\widehat{\boldsymbol{W}}\right)^{-1}\widehat{\boldsymbol{W}}'\boldsymbol{Y}$$

and

$$\boldsymbol{Y} = \widehat{\boldsymbol{W}}\beta + \boldsymbol{U}\alpha + \left(\boldsymbol{W} - \widehat{\boldsymbol{W}}\right)\beta + \boldsymbol{v}.$$

Substituting, using $\widehat{\boldsymbol{W}}'\widehat{\boldsymbol{U}} = 0$ and $\boldsymbol{W} - \widehat{\boldsymbol{W}} = -\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{U}$ we find

$$\begin{aligned}
\widehat{\beta} - \beta &= \left(\widehat{\boldsymbol{W}}'\widehat{\boldsymbol{W}}\right)^{-1}\widehat{\boldsymbol{W}}'\left(\boldsymbol{U}\alpha + \left(\boldsymbol{W} - \widehat{\boldsymbol{W}}\right)\beta + \boldsymbol{v}\right) \\
&= \left(\widehat{\boldsymbol{A}}'\boldsymbol{Z}'\boldsymbol{Z}\widehat{\boldsymbol{A}}\right)^{-1}\widehat{\boldsymbol{A}}'\boldsymbol{Z}'\left(\boldsymbol{U}\alpha - \boldsymbol{U}\beta + \boldsymbol{v}\right) \\
&= \left(\widehat{\boldsymbol{A}}'\boldsymbol{Z}'\boldsymbol{Z}\widehat{\boldsymbol{A}}\right)^{-1}\widehat{\boldsymbol{A}}'\boldsymbol{Z}'\boldsymbol{e}
\end{aligned}$$

where

$$e_i = v_i + u_i'\left(\alpha - \beta\right) = Y_i - X_i'\beta.$$

We also find

$$\widehat{\alpha} = \left(\widehat{\boldsymbol{U}}'\widehat{\boldsymbol{U}}\right)^{-1}\widehat{\boldsymbol{U}}'\boldsymbol{Y}.$$

Since $\widehat{\boldsymbol{U}}'\boldsymbol{W} = 0$, $\boldsymbol{U} - \widehat{\boldsymbol{U}} = \boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{U}$ and $\widehat{\boldsymbol{U}}'\boldsymbol{Z} = 0$ then

$$\begin{aligned}
\widehat{\alpha} - \alpha &= \left(\widehat{\boldsymbol{U}}'\widehat{\boldsymbol{U}}\right)^{-1}\widehat{\boldsymbol{U}}'\left(\boldsymbol{W}\beta + \left(\boldsymbol{U} - \widehat{\boldsymbol{U}}\right)\alpha + \boldsymbol{v}\right) \\
&= \left(\widehat{\boldsymbol{U}}'\widehat{\boldsymbol{U}}\right)^{-1}\widehat{\boldsymbol{U}}'\boldsymbol{v}.
\end{aligned}$$

Together, we establish the following distributional result.

Theorem 12.12 For the model and estimators described in this section, with $\mathbb{E}\left[Y^4\right]<\infty$, $\mathbb{E}\|Z\|^4<\infty$, $\mathbb{E}\|X\|^4<\infty$, $\boldsymbol{A}'\mathbb{E}\left[ZZ'\right]\boldsymbol{A}>0$, and $\mathbb{E}\left[uu'\right]>0$, as $n\to\infty$

$$\sqrt{n}\begin{pmatrix}\widehat{\beta}-\beta\\ \widehat{\alpha}-\alpha\end{pmatrix}\xrightarrow[d]{}\mathrm{N}(0,\boldsymbol{V}) \tag{12.58}$$

where

$$\boldsymbol{V}=\begin{pmatrix}\boldsymbol{V}_{\beta\beta} & \boldsymbol{V}_{\beta\alpha}\\ \boldsymbol{V}_{\alpha\beta} & \boldsymbol{V}_{\alpha\alpha}\end{pmatrix}$$

and

$$\begin{aligned}
\boldsymbol{V}_{\beta\beta}&=\left(\boldsymbol{A}'\mathbb{E}\left[ZZ'\right]\boldsymbol{A}\right)^{-1}\left(\boldsymbol{A}'\mathbb{E}\left[ZZ'e^2\right]\boldsymbol{A}\right)\left(\boldsymbol{A}'\mathbb{E}\left[ZZ'\right]\boldsymbol{A}\right)^{-1}\\
\boldsymbol{V}_{\alpha\beta}&=\left(\mathbb{E}\left[uu'\right]\right)^{-1}\left(\mathbb{E}\left[uZ'ev\right]\boldsymbol{A}\right)\left(\boldsymbol{A}'\mathbb{E}\left[ZZ'\right]\boldsymbol{A}\right)^{-1}\\
\boldsymbol{V}_{\alpha\alpha}&=\left(\mathbb{E}\left[uu'\right]\right)^{-1}\mathbb{E}\left[uu'v^2\right]\left(\mathbb{E}\left[uu'\right]\right)^{-1}.
\end{aligned}$$

The asymptotic covariance matrix is estimated by

$$\begin{aligned}
\widehat{\boldsymbol{V}}_{\beta\beta}&=\left(\frac{1}{n}\widehat{\boldsymbol{W}}'\widehat{\boldsymbol{W}}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\widehat{W}_i\widehat{W}_i'\widehat{e}_i^2\right)\left(\frac{1}{n}\widehat{\boldsymbol{W}}'\widehat{\boldsymbol{W}}\right)^{-1}\\
\widehat{\boldsymbol{V}}_{\alpha\beta}&=\left(\frac{1}{n}\widehat{\boldsymbol{U}}'\widehat{\boldsymbol{U}}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\widehat{u}_i\widehat{W}_i'\widehat{e}_i\widehat{v}_i\right)\left(\frac{1}{n}\widehat{\boldsymbol{W}}'\widehat{\boldsymbol{W}}\right)^{-1}\\
\widehat{\boldsymbol{V}}_{\alpha\alpha}&=\left(\frac{1}{n}\widehat{\boldsymbol{U}}'\widehat{\boldsymbol{U}}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\widehat{U}_i\widehat{U}_i'\widehat{v}_i^2\right)\left(\frac{1}{n}\widehat{\boldsymbol{U}}'\widehat{\boldsymbol{U}}\right)^{-1}
\end{aligned}$$

where

$$\begin{aligned}
\widehat{W}_i&=\widehat{\boldsymbol{A}}'Z_i\\
\widehat{u}_i&=\widehat{X}_i-\widehat{W}_i\\
\widehat{e}_i&=Y_i-X_i'\widehat{\beta}\\
\widehat{v}_i&=Y_i-\widehat{W}_i'\widehat{\beta}-\widehat{u}_i'\widehat{\alpha}.
\end{aligned}$$

Under conditional homoskedasticity, specifically

$$\mathbb{E}\left[\left.\begin{pmatrix}e_i^2 & e_iv_i\\ e_iv_i & v_i^2\end{pmatrix}\right| Z_i\right]=\boldsymbol{C}$$

then $\boldsymbol{V}_{\alpha\beta}=0$ and the coefficient estimates $\widehat{\beta}$ and $\widehat{\alpha}$ are asymptotically independent. The variance components also simplify to

$$\begin{aligned}
\boldsymbol{V}_{\beta\beta}&=\left(\boldsymbol{A}'\mathbb{E}\left[ZZ'\right]\boldsymbol{A}\right)^{-1}\mathbb{E}\left[e_i^2\right]\\
\boldsymbol{V}_{\alpha\alpha}&=\left(\mathbb{E}\left[uu'\right]\right)^{-1}\mathbb{E}\left[v^2\right].
\end{aligned}$$

In this case we have the covariance matrix estimators

$$\begin{aligned}
\widehat{\boldsymbol{V}}_{\beta\beta}^{0}&=\left(\frac{1}{n}\widehat{\boldsymbol{W}}'\widehat{\boldsymbol{W}}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\widehat{e}_i^2\right)\\
\widehat{\boldsymbol{V}}_{\alpha\alpha}^{0}&=\left(\frac{1}{n}\widehat{\boldsymbol{U}}'\widehat{\boldsymbol{U}}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\widehat{v}_i^2\right)
\end{aligned}$$

and $\widehat{\boldsymbol{V}}^0_{\alpha\beta} = 0$.

12.28 Control Function Regression

In this section we present an alternative way of computing the 2SLS estimator by least squares. It is useful in nonlinear contexts, and also in the linear model to construct tests for endogeneity.

The structural and reduced form equations for the standard IV model are

$$\begin{aligned} Y &= X_1'\beta_1 + X_2'\beta_2 + e \\ X_2 &= \Gamma_{12}'Z_1 + \Gamma_{22}'Z_2 + u_2. \end{aligned}$$

Since the instrumental variable assumption specifies that $\mathbb{E}[Ze] = 0$, X_2 is endogenous (correlated with e) if u_2 and e are correlated. We can therefore consider the linear projection of e on u_2

$$\begin{aligned} e &= u_2'\alpha + v \\ \alpha &= \left(\mathbb{E}\left[u_2u_2'\right]\right)^{-1}\mathbb{E}\left[u_2e\right] \\ \mathbb{E}\left[u_2v\right] &= 0. \end{aligned}$$

Substituting this into the structural form equation we find

$$\begin{aligned} Y &= X_1'\beta_1 + X_2'\beta_2 + u_2'\alpha + v \qquad (12.59) \\ \mathbb{E}\left[X_1v\right] &= 0 \\ \mathbb{E}\left[X_2v\right] &= 0 \\ \mathbb{E}\left[u_2v\right] &= 0. \end{aligned}$$

Notice that X_2 is uncorrelated with v. This is because X_2 is correlated with e only through u_2, and v is the error after e has been projected orthogonal to u_2.

If u_2 were observed we could then estimate (12.59) by least squares. Since it is not observed we estimate it by the reduced-form residual $\widehat{u}_{2i} = X_{2i} - \widehat{\Gamma}_{12}'Z_{1i} - \widehat{\Gamma}_{22}'Z_{2i}$. Then the coefficients $(\beta_1, \beta_2, \alpha)$ can be estimated by least squares of Y on $(X_1, X_2, \widehat{u}_2)$. We can write this as

$$Y_i = X_i'\widehat{\beta} + \widehat{u}_{2i}'\widehat{\alpha} + \widehat{v}_i \qquad (12.60)$$

or in matrix notation as

$$\boldsymbol{Y} = \boldsymbol{X}\widehat{\beta} + \widehat{\boldsymbol{U}}_2\widehat{\alpha} + \widehat{\boldsymbol{v}}.$$

This turns out to be an alternative algebraic expression for the 2SLS estimator.

Indeed, we now show that $\widehat{\beta} = \widehat{\beta}_{\text{2sls}}$. First, note that the reduced form residual can be written as

$$\widehat{\boldsymbol{U}}_2 = \left(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{Z}}\right)\boldsymbol{X}_2$$

where $\boldsymbol{P}_{\boldsymbol{Z}}$ is defined in (12.30). By the FWL representation

$$\widehat{\beta} = \left(\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}\right)^{-1}\left(\widetilde{\boldsymbol{X}}'\boldsymbol{Y}\right) \qquad (12.61)$$

where $\widetilde{\boldsymbol{X}} = \left[\widetilde{\boldsymbol{X}}_1, \widetilde{\boldsymbol{X}}_2\right]$ with

$$\widetilde{\boldsymbol{X}}_1 = \boldsymbol{X}_1 - \widehat{\boldsymbol{U}}_2\left(\widehat{\boldsymbol{U}}_2'\widehat{\boldsymbol{U}}_2\right)^{-1}\widehat{\boldsymbol{U}}_2'\boldsymbol{X}_1 = \boldsymbol{X}_1$$

(since $\widehat{\boldsymbol{U}}_2'\boldsymbol{X}_1 = 0$) and

$$\begin{aligned}\widetilde{\boldsymbol{X}}_2 &= \boldsymbol{X}_2 - \widehat{\boldsymbol{U}}_2\left(\widehat{\boldsymbol{U}}_2'\widehat{\boldsymbol{U}}_2\right)^{-1}\widehat{\boldsymbol{U}}_2'\boldsymbol{X}_2 \\ &= \boldsymbol{X}_2 - \widehat{\boldsymbol{U}}_2\left(\boldsymbol{X}_2'\left(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{Z}}\right)\boldsymbol{X}_2\right)^{-1}\boldsymbol{X}_2'\left(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{Z}}\right)\boldsymbol{X}_2 \\ &= \boldsymbol{X}_2 - \widehat{\boldsymbol{U}}_2 \\ &= \boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X}_2.\end{aligned}$$

Thus $\widetilde{\boldsymbol{X}} = [\boldsymbol{X}_1, \boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X}_2] = \boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X}$. Substituted into (12.61) we find

$$\widehat{\beta} = \left(\boldsymbol{X}'\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{Y}\right) = \widehat{\beta}_{\text{2sls}}$$

which is (12.31) as claimed.

Again, what we have found is that OLS estimation of equation (12.60) yields algebraically the 2SLS estimator $\widehat{\beta}_{\text{2sls}}$.

We now consider the distribution of the control function estimator $\left(\widehat{\beta}, \widehat{\alpha}\right)$. It is a generated regression model, and in fact is covered by the model examined in Section 12.27 after a slight reparametrization. Let $W = \overline{\Gamma}'Z$. Note $u = X - W$. Then the main equation (12.59) can be written as $Y = W'\beta + u_2'\gamma + v$ where $\gamma = \alpha + \beta_2$. This is the model in Section 12.27.

Set $\widehat{\gamma} = \widehat{\alpha} + \widehat{\beta}_2$. It follows from (12.58) that as $n \to \infty$ we have the joint distribution

$$\sqrt{n}\begin{pmatrix}\widehat{\beta}_2 - \beta_2 \\ \widehat{\gamma} - \gamma\end{pmatrix} \xrightarrow[d]{} \mathrm{N}\left(0, \boldsymbol{V}\right)$$

where

$$\boldsymbol{V} = \begin{pmatrix}\boldsymbol{V}_{22} & \boldsymbol{V}_{2\gamma} \\ \boldsymbol{V}_{\gamma 2} & \boldsymbol{V}_{\gamma\gamma}\end{pmatrix}$$

$$\begin{aligned}\boldsymbol{V}_{22} &= \left[\left(\overline{\Gamma}'\mathbb{E}\left[ZZ'\right]\overline{\Gamma}\right)^{-1}\overline{\Gamma}'\mathbb{E}\left[ZZ'e^2\right]\overline{\Gamma}\left(\overline{\Gamma}'\mathbb{E}\left[ZZ'\right]\overline{\Gamma}\right)^{-1}\right]_{22} \\ \boldsymbol{V}_{\gamma 2} &= \left[\left(\mathbb{E}\left[u_2u_2'\right]\right)^{-1}\mathbb{E}\left[uZ'ev\right]\overline{\Gamma}\left(\overline{\Gamma}'\mathbb{E}\left[ZZ'\right]\overline{\Gamma}\right)^{-1}\right]_{\cdot 2} \\ \boldsymbol{V}_{\gamma\gamma} &= \left(\mathbb{E}\left[u_2u_2'\right]\right)^{-1}\mathbb{E}\left[u_2u_2'v^2\right]\left(\mathbb{E}\left[u_2u_2'\right]\right)^{-1} \\ e &= Y - X'\beta.\end{aligned}$$

The asymptotic distribution of $\widehat{\gamma} = \widehat{\alpha} - \widehat{\beta}_2$ can be deduced.

Theorem 12.13 If $\mathbb{E}\left[Y^4\right] < \infty$, $\mathbb{E}\|Z\|^4 < \infty$, $\mathbb{E}\|X\|^4 < \infty$, $\boldsymbol{A}'\mathbb{E}\left[ZZ'\right]\boldsymbol{A} > 0$, and $\mathbb{E}\left[uu'\right] > 0$, as $n \to \infty$

$$\sqrt{n}\left(\widehat{\alpha} - \alpha\right) \xrightarrow[d]{} \mathrm{N}\left(0, \boldsymbol{V}_{\alpha}\right)$$

where

$$\boldsymbol{V}_{\alpha} = \boldsymbol{V}_{22} + \boldsymbol{V}_{\gamma\gamma} - \boldsymbol{V}_{\gamma 2} - \boldsymbol{V}_{\gamma 2}'.$$

Under conditional homoskedasticity we have the important simplifications

$$\begin{aligned}
\boldsymbol{V}_{22} &= \left[\left(\overline{\Gamma}'\mathbb{E}\left[ZZ'\right]\overline{\Gamma}\right)^{-1}\right]_{22}\mathbb{E}\left[e^2\right]\\
\boldsymbol{V}_{\gamma\gamma} &= \left(\mathbb{E}\left[u_2u_2'\right]\right)^{-1}\mathbb{E}\left[v^2\right]\\
\boldsymbol{V}_{\gamma 2} &= 0\\
\boldsymbol{V}_{\alpha} &= \boldsymbol{V}_{22}+\boldsymbol{V}_{\gamma\gamma}.
\end{aligned}$$

An estimator for $\boldsymbol{V}_\alpha$ in the general case is

$$\widehat{\boldsymbol{V}}_{\alpha} = \widehat{\boldsymbol{V}}_{22}+\widehat{\boldsymbol{V}}_{\gamma\gamma}-\widehat{\boldsymbol{V}}_{\gamma 2}-\widehat{\boldsymbol{V}}_{\gamma 2}' \tag{12.62}$$

where

$$\begin{aligned}
\widehat{\boldsymbol{V}}_{22} &= \left[\frac{1}{n}\left(\boldsymbol{X}'\boldsymbol{P_Z}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\left(\sum_{i=1}^{n}Z_iZ_i'\widehat{e}_i^2\right)\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{P_Z}\boldsymbol{X}\right)^{-1}\right]_{22}\\
\widehat{\boldsymbol{V}}_{\gamma 2} &= \left[\frac{1}{n}\left(\widehat{\boldsymbol{U}}'\widehat{\boldsymbol{U}}\right)^{-1}\left(\sum_{i=1}^{n}\widehat{u}_i\widehat{W}_i'\widehat{e}_i\widehat{v}_i\right)\left(\boldsymbol{X}'\boldsymbol{P_Z}\boldsymbol{X}\right)^{-1}\right]_{\cdot 2}\\
\widehat{e}_i &= Y_i - X_i'\widehat{\beta}\\
\widehat{v}_i &= Y_i - X_i'\widehat{\beta}-\widehat{u}_{2i}'\widehat{\gamma}.
\end{aligned}$$

Under the assumption of conditional homoskedasticity we have the estimator

$$\begin{aligned}
\widehat{\boldsymbol{V}}_{\alpha}^0 &= \widehat{\boldsymbol{V}}_{\beta\beta}^0+\widehat{\boldsymbol{V}}_{\gamma\gamma}^0\\
\widehat{\boldsymbol{V}}_{\beta\beta} &= \left[\left(\boldsymbol{X}'\boldsymbol{P_Z}\boldsymbol{X}\right)^{-1}\right]_{22}\left(\sum_{i=1}^{n}\widehat{e}_i^2\right)\\
\widehat{\boldsymbol{V}}_{\gamma\gamma} &= \left(\widehat{\boldsymbol{U}}'\widehat{\boldsymbol{U}}\right)^{-1}\left(\sum_{i=1}^{n}\widehat{v}_i^2\right).
\end{aligned}$$

12.29 Endogeneity Tests

The 2SLS estimator allows the regressor X_2 to be endogenous, meaning that X_2 is correlated with the structural error e. If this correlation is zero then X_2 is exogenous and the structural equation can be estimated by least squares. This is a testable restriction. Effectively, the null hypothesis is

$$\mathbb{H}_0:\mathbb{E}\left[X_2e\right]=0$$

with the alternative

$$\mathbb{H}_1:\mathbb{E}\left[X_2e\right]\neq 0.$$

The maintained hypothesis is $\mathbb{E}[Ze]=0$. Since X_1 is a component of Z this implies $\mathbb{E}[X_1e]=0$. Consequently we could alternatively write the null as $\mathbb{H}_0:\mathbb{E}[Xe]=0$ (and some authors do so).

Recall the control function regression (12.59)

$$\begin{aligned}
Y &= X_1'\beta_1+X_2'\beta_2+u_2'\alpha+v\\
\alpha &= \left(\mathbb{E}\left[u_2u_2'\right]\right)^{-1}\mathbb{E}\left[u_2e\right].
\end{aligned}$$

Notice that $\mathbb{E}[X_2 e] = 0$ if and only if $\mathbb{E}[u_2 e] = 0$, so the hypothesis can be restated as $\mathbb{H}_0 : \alpha = 0$ against $\mathbb{H}_1 : \alpha \neq 0$. Thus a natural test is based on the Wald statistic W for $\alpha = 0$ in the control function regression (12.28). Under Theorem 12.9, Theorem 12.10, and $\mathbb{H}_0$, W is asymptotically chi-square with k_2 degrees of freedom. In addition, under the normal regression assumption the F statistic has an exact $F(k_2, n - k_1 - 2k_2)$ distribution. We accept the null hypothesis that X_2 is exogenous if W (or F) is smaller than the critical value, and reject in favor of the hypothesis that X_2 is endogenous if the statistic is larger than the critical value.

Specifically, estimate the reduced form by least squares

$$X_{2i} = \widehat{\Gamma}'_{12} Z_{1i} + \widehat{\Gamma}'_{22} Z_{2i} + \widehat{u}_{2i}$$

to obtain the residuals. Then estimate the control function by least squares

$$Y_i = X'_i \widehat{\beta} + \widehat{u}'_{2i} \widehat{\alpha} + \widehat{v}_i. \tag{12.63}$$

Let W, W^0 and $F = W^0 / k_2$ denote the Wald, homoskedastic Wald, and F statistics for $\alpha = 0$.

Theorem 12.14 Under $\mathbb{H}_0$, $W \xrightarrow[d]{} \chi^2_{k_2}$. Let $c_{1-\alpha}$ solve $\mathbb{P}\left[\chi^2_{k_2} \leq c_{1-\alpha}\right] = 1 - \alpha$. The test "Reject $\mathbb{H}_0$ if $W > c_{1-\alpha}$" has asymptotic size α.

Theorem 12.15 Suppose $e \mid X, Z \sim \mathrm{N}\left(0, \sigma^2\right)$. Under $\mathbb{H}_0$, F $\sim F(k_2, n - k_1 - 2k_2)$. Let $c_{1-\alpha}$ solve $\mathbb{P}[F(k_2, n - k_1 - 2k_2) \leq c_{1-\alpha}] = 1 - \alpha$. The test "Reject $\mathbb{H}_0$ if F > $c_{1-\alpha}$" has exact size α.

Since in general we do not want to impose homoskedasticity these results suggest that the most appropriate test is the Wald statistic constructed with the robust heteroskedastic covariance matrix. This can be computed in Stata using the command `estat endogenous` after `ivregress` when the latter uses a robust covariance option. Stata reports the Wald statistic in F form (and thus uses the F distribution to calculate the p-value) as "Robust regression F". Using the F rather than the χ^2 is not formally justified but is a reasonable finite sample adjustment. If the command `estat endogenous` is applied after `ivregress` without a robust covariance option Stata reports the F statistic as "Wu-Hausman F".

There is an alternative (and traditional) way to derive a test for endogeneity. Under $\mathbb{H}_0$, both OLS and 2SLS are consistent estimators. But under $\mathbb{H}_1$ they converge to different values. Thus the difference between the OLS and 2SLS estimators is a valid test statistic for endogeneity. It also measures what we often care most about – the impact of endogeneity on the parameter estimates. This literature was developed under the assumption of conditional homoskedasticity (and it is important for these results) so we assume this condition for the development of the statistic.

Let $\widehat{\beta} = \left(\widehat{\beta}_1, \widehat{\beta}_2\right)$ be the OLS estimator and let $\widetilde{\beta} = \left(\widetilde{\beta}_1, \widetilde{\beta}_2\right)$ be the 2SLS estimator. Under $\mathbb{H}_0$ and homoskedasticity the OLS estimator is Gauss-Markov efficient so by the Hausman equality

$$\begin{aligned}
\operatorname{var}\left[\widehat{\beta}_2 - \widetilde{\beta}_2\right] &= \operatorname{var}\left[\widetilde{\beta}_2\right] - \operatorname{var}\left[\widehat{\beta}_2\right] \\
&= \left(\left(\boldsymbol{X}'_2 \left(\boldsymbol{P}_{\boldsymbol{Z}} - \boldsymbol{P}_1\right) \boldsymbol{X}_2\right)^{-1} - \left(\boldsymbol{X}'_2 \boldsymbol{M}_1 \boldsymbol{X}_2\right)^{-1}\right) \sigma^2
\end{aligned}$$

where $\boldsymbol{P_Z} = \boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'$, $\boldsymbol{P}_1 = \boldsymbol{X}_1\left(\boldsymbol{X}_1'\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1'$, and $\boldsymbol{M}_1 = \boldsymbol{I}_n - \boldsymbol{P}_1$. Thus a valid test statistic for $\mathbb{H}_0$ is

$$T = \frac{\left(\widehat{\beta}_2 - \widetilde{\beta}_2\right)'\left(\left(\boldsymbol{X}_2'\left(\boldsymbol{P_Z} - \boldsymbol{P}_1\right)\boldsymbol{X}_2\right)^{-1} - \left(\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2\right)^{-1}\right)^{-1}\left(\widehat{\beta}_2 - \widetilde{\beta}_2\right)}{\widehat{\sigma}^2} \tag{12.64}$$

for some estimator $\widehat{\sigma}^2$ of σ^2. Durbin (1954) first proposed T as a test for endogeneity in the context of IV estimation setting $\widehat{\sigma}^2$ to be the least squares estimator of σ^2. Wu (1973) proposed T as a test for endogeneity in the context of 2SLS estimation, considering a set of possible estimators $\widehat{\sigma}^2$ including the regression estimator from (12.63). Hausman (1978) proposed a version of T based on the full contrast $\widehat{\beta} - \widetilde{\beta}$, and observed that it equals the regression Wald statistic W^0 described earlier. In fact, when $\widehat{\sigma}^2$ is the regression estimator from (12.63) the statistic (12.64) algebraically equals both W^0 and the version of (12.64) based on the full contrast $\widehat{\beta} - \widetilde{\beta}$. We show these equalities below. Thus these three approaches yield exactly the same statistic except for possible differences regarding the choice of $\widehat{\sigma}^2$. Since the regression F test described earlier has an exact F distribution in the normal sampling model and thus can exactly control test size, this is the preferred version of the test. The general class of tests are called **Durbin-Wu-Hausman** tests, **Wu-Hausman** tests, or **Hausman** tests, depending on the author.

When $k_2 = 1$ (there is one right-hand-side endogenous variable), which is quite common in applications, the endogeneity test can be equivalently expressed at the t-statistic for $\widehat{\alpha}$ in the estimated control function. Thus it is sufficient to estimate the control function regression and check the t-statistic for $\widehat{\alpha}$. If $|\widehat{\alpha}| > 2$ then we can reject the hypothesis that X_2 is exogenous for β.

We illustrate using the Card proximity example using the two instruments *public* and *private*. We first estimate the reduced form for *education*, obtain the residual, and then estimate the control function regression. The residual has a coefficient −0.088 with a standard error of 0.037 and a t-statistic of 2.4. Since the latter exceeds the 5% critical value (its p-value is 0.017) we reject exogeneity. This means that the 2SLS estimates are statistically different from the least squares estimates of the structural equation and supports our decision to treat education as an endogenous variable. (Alternatively, the F statistic is $2.4^2 = 5.7$ with the same p-value).

We now show the equality of the various statistics.

We first show that the statistic (12.64) is not altered if based on the full contrast $\widehat{\beta} - \widetilde{\beta}$. Indeed, $\widehat{\beta}_1 - \widetilde{\beta}_1$ is a linear function of $\widehat{\beta}_2 - \widetilde{\beta}_2$, so there is no extra information in the full contrast. To see this, observe that given $\widehat{\beta}_2$ we can solve by least squares to find

$$\widehat{\beta}_1 = \left(\boldsymbol{X}_1'\boldsymbol{X}_1\right)^{-1}\left(\boldsymbol{X}_1'\left(\boldsymbol{Y} - \boldsymbol{X}_2\widehat{\beta}_2\right)\right)$$

and similarly

$$\widetilde{\beta}_1 = \left(\boldsymbol{X}_1'\boldsymbol{X}_1\right)^{-1}\left(\boldsymbol{X}_1'\left(\boldsymbol{Y} - \boldsymbol{P_Z}\boldsymbol{X}_2\widetilde{\beta}\right)\right) = \left(\boldsymbol{X}_1'\boldsymbol{X}_1\right)^{-1}\left(\boldsymbol{X}_1'\left(\boldsymbol{Y} - \boldsymbol{X}_2\widetilde{\beta}\right)\right)$$

the second equality since $\boldsymbol{P_Z}\boldsymbol{X}_1 = \boldsymbol{X}_1$. Thus

$$\begin{aligned}\widehat{\beta}_1 - \widetilde{\beta}_1 &= \left(\boldsymbol{X}_1'\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1'\left(\boldsymbol{Y} - \boldsymbol{X}_2\widehat{\beta}_2\right) - \left(\boldsymbol{X}_1'\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1'\left(\boldsymbol{Y} - \boldsymbol{P_Z}\boldsymbol{X}_2\widetilde{\beta}\right)\\ &= \left(\boldsymbol{X}_1'\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1'\boldsymbol{X}_2\left(\widetilde{\beta}_2 - \widehat{\beta}_2\right)\end{aligned}$$

as claimed.

We next show that T in (12.64) equals the homoskedastic Wald statistic W^0 for $\widehat{\alpha}$ from the regression (12.63). Consider the latter regression. Since $\boldsymbol{X}_2$ is contained in $\boldsymbol{X}$ the coefficient estimate $\widehat{\alpha}$ is invariant to replacing $\widehat{\boldsymbol{U}}_2 = \boldsymbol{X}_2 - \widehat{\boldsymbol{X}}_2$ with $-\widehat{\boldsymbol{X}}_2 = -\boldsymbol{P_Z}\boldsymbol{X}_2$. By the FWL representation, setting $\boldsymbol{M_X} = \boldsymbol{I}_n - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'$,

$$\widehat{\alpha} = -\left(\widehat{\boldsymbol{X}}_2'\boldsymbol{M_X}\widehat{\boldsymbol{X}}_2\right)^{-1}\widehat{\boldsymbol{X}}_2'\boldsymbol{M_X}\boldsymbol{Y} = -\left(\boldsymbol{X}_2'\boldsymbol{P_Z}\boldsymbol{M_X}\boldsymbol{P_Z}\boldsymbol{X}_2\right)^{-1}\boldsymbol{X}_2'\boldsymbol{P_Z}\boldsymbol{M_X}\boldsymbol{Y}.$$

It follows that

$$W^0 = \frac{\boldsymbol{Y}'\boldsymbol{M_X}\boldsymbol{P_Z}\boldsymbol{X}_2\left(\boldsymbol{X}_2'\boldsymbol{P_Z}\boldsymbol{M_X}\boldsymbol{P_Z}\boldsymbol{X}_2\right)^{-1}\boldsymbol{X}_2'\boldsymbol{P_Z}\boldsymbol{M_X}\boldsymbol{Y}}{\widehat{\sigma}^2}.$$

Our goal is to show that $T = W^0$. Define $\widetilde{\boldsymbol{X}}_2 = (\boldsymbol{I}_n - \boldsymbol{P}_1)\boldsymbol{X}_2$ so $\widehat{\beta}_2 = \left(\widetilde{\boldsymbol{X}}_2'\widetilde{\boldsymbol{X}}_2\right)^{-1}\widetilde{\boldsymbol{X}}_2'\boldsymbol{Y}$. Then using $(\boldsymbol{P_Z} - \boldsymbol{P}_1)(\boldsymbol{I}_n - \boldsymbol{P}_1) = (\boldsymbol{P_Z} - \boldsymbol{P}_1)$ and defining $\boldsymbol{Q} = \widetilde{\boldsymbol{X}}_2\left(\widetilde{\boldsymbol{X}}_2'\widetilde{\boldsymbol{X}}_2\right)^{-1}\widetilde{\boldsymbol{X}}_2'$ we find

$$\begin{aligned}
\boldsymbol{\Delta} &\overset{\text{def}}{=} \left(\boldsymbol{X}_2'(\boldsymbol{P_Z} - \boldsymbol{P}_1)\boldsymbol{X}_2\right)\left(\widetilde{\beta}_2 - \widehat{\beta}_2\right) \\
&= \boldsymbol{X}_2'(\boldsymbol{P_Z} - \boldsymbol{P}_1)\boldsymbol{Y} - \left(\boldsymbol{X}_2'(\boldsymbol{P_Z} - \boldsymbol{P}_1)\boldsymbol{X}_2\right)\left(\widetilde{\boldsymbol{X}}_2'\widetilde{\boldsymbol{X}}_2\right)^{-1}\widetilde{\boldsymbol{X}}_2'\boldsymbol{Y} \\
&= \boldsymbol{X}_2'(\boldsymbol{P_Z} - \boldsymbol{P}_1)(\boldsymbol{I}_n - \boldsymbol{Q})\boldsymbol{Y} \\
&= \boldsymbol{X}_2'(\boldsymbol{P_Z} - \boldsymbol{P}_1 - \boldsymbol{P_Z}\boldsymbol{Q})\boldsymbol{Y} \\
&= \boldsymbol{X}_2'\boldsymbol{P_Z}(\boldsymbol{I}_n - \boldsymbol{P}_1 - \boldsymbol{Q})\boldsymbol{Y} \\
&= \boldsymbol{X}_2'\boldsymbol{P_Z}\boldsymbol{M_X}\boldsymbol{Y}.
\end{aligned}$$

The third-to-last equality is $\boldsymbol{P}_1\boldsymbol{Q} = 0$ and the final uses $\boldsymbol{M_X} = \boldsymbol{I}_n - \boldsymbol{P}_1 - \boldsymbol{Q}$. We also calculate that

$$\begin{aligned}
\boldsymbol{Q}^* &\overset{\text{def}}{=} \left(\boldsymbol{X}_2'(\boldsymbol{P_Z} - \boldsymbol{P}_1)\boldsymbol{X}_2\right)\left(\left(\boldsymbol{X}_2'(\boldsymbol{P_Z} - \boldsymbol{P}_1)\boldsymbol{X}_2\right)^{-1} - \left(\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2\right)^{-1}\right)\left(\boldsymbol{X}_2'(\boldsymbol{P_Z} - \boldsymbol{P}_1)\boldsymbol{X}_2\right) \\
&= \boldsymbol{X}_2'(\boldsymbol{P_Z} - \boldsymbol{P}_1 - (\boldsymbol{P_Z} - \boldsymbol{P}_1)\boldsymbol{Q}(\boldsymbol{P_Z} - \boldsymbol{P}_1))\boldsymbol{X}_2 \\
&= \boldsymbol{X}_2'\left(\boldsymbol{P_Z} - \boldsymbol{P}_1 - \boldsymbol{P_Z}\boldsymbol{Q}\boldsymbol{P_Z}\right)\boldsymbol{X}_2 \\
&= \boldsymbol{X}_2'\boldsymbol{P_Z}\boldsymbol{M_X}\boldsymbol{P_Z}\boldsymbol{X}_2.
\end{aligned}$$

Thus

$$\begin{aligned}
T &= \frac{\boldsymbol{\Delta}'\boldsymbol{Q}^{*-1}\boldsymbol{\Delta}}{\widehat{\sigma}^2} \\
&= \frac{\boldsymbol{Y}'\boldsymbol{M_X}\boldsymbol{P_Z}\boldsymbol{X}_2\left(\boldsymbol{X}_2'\boldsymbol{P_Z}\boldsymbol{M_X}\boldsymbol{P_Z}\boldsymbol{X}_2\right)^{-1}\boldsymbol{X}_2'\boldsymbol{P_Z}\boldsymbol{M_X}\boldsymbol{Y}}{\widehat{\sigma}^2} \\
&= W^0
\end{aligned}$$

as claimed.

12.30 Subset Endogeneity Tests

In some cases we may only wish to test the endogeneity of a subset of the variables. In the Card proximity example we may wish test the exogeneity of *education* separately from *experience* and its square. To execute a subset endogeneity test it is useful to partition the regressors into three groups so that the structural model is

$$\begin{aligned}
Y &= X_1'\beta_1 + X_2'\beta_2 + X_3'\beta_3 + e \\
\mathbb{E}[Ze] &= 0.
\end{aligned}$$

As before, the instrument vector Z includes X_1. The vector X_3 is treated as endogenous and X_2 is treated as potentially endogenous. The hypothesis to test is that X_2 is exogenous, or $\mathbb{H}_0 : \mathbb{E}[X_2 e] = 0$ against $\mathbb{H}_1 : \mathbb{E}[X_2 e] \neq 0$.

Under homoskedasticity a straightfoward test can be constructed by the Durbin-Wu-Hausman principle. Under $\mathbb{H}_0$ the appropriate estimator is 2SLS using the instruments (Z, X_2). Let this estimator of β_2

be denoted $\widehat{\beta}_2$. Under $\mathbb{H}_1$ the appropriate estimator is 2SLS using the smaller instrument set Z. Let this estimator of β_2 be denoted $\widetilde{\beta}_2$. A Durbin-Wu-Hausman statistic for $\mathbb{H}_0$ against $\mathbb{H}_1$ is

$$T = \left(\widehat{\beta}_2 - \widetilde{\beta}_2\right)' \left(\widehat{\operatorname{var}}\left[\widetilde{\beta}_2\right] - \widehat{\operatorname{var}}\left[\widehat{\beta}_2\right]\right)^{-1} \left(\widehat{\beta}_2 - \widetilde{\beta}_2\right).$$

The asymptotic distribution under $\mathbb{H}_0$ is $\chi^2_{k_2}$ where $k_2 = \dim(X_2)$, so we reject the hypothesis that the variables X_2 are exogenous if T exceeds an upper critical value from the $\chi^2_{k_2}$ distribution.

Instead of using the Wald statistic one could use the F version of the test by dividing by k_2 and using the F distribution for critical values. There is no finite sample justification for this modification, however, since X_3 is endogenous under the null hypothesis.

In Stata, the command `estat endogenous` (adding the variable name to specify which variable to test for exogeneity) after `ivregress` without a robust covariance option reports the F version of this statistic as "Wu-Hausman F". For example, in the Card proximity example using the four instruments *public*, *private*, *age*, and *age*^{2}, if we estimate the equation by 2SLS with a non-robust covariance matrix and then compute the endogeneity test for education we find $F = 272$ with a p-value of 0.0000, but if we compute the test for experience and its square we find $F = 2.98$ with a p-value of 0.051. In this model, the assumption of exogeneity with homogenous coefficients is rejected for education but the result for experience is unclear.

A heteroskedasticity or cluster-robust test cannot be constructed easily by the Durbin-Wu-Hausman approach since the covariance matrix does not take a simple form. To allow for non-homoskedastic errors it is recommended to use GMM estimation. See Section 13.24.

12.31 OverIdentification Tests

When $\ell > k$ the model is **overidentified** meaning that there are more moments than free parameters. This is a restriction and is testable. Such tests are called **overidentification tests**.

The instrumental variables model specifies $\mathbb{E}[Ze] = 0$. Equivalently, since $e = Y - X'\beta$ this is

$$\mathbb{E}[ZY] - \mathbb{E}\left[ZX'\right]\beta = 0.$$

This is an $\ell \times 1$ vector of restrictions on the moment matrices $\mathbb{E}[ZY]$ and $\mathbb{E}\left[ZX'\right]$. Yet since β is of dimension k which is less than ℓ it is not certain if indeed such a β exists.

To make things a bit more concrete, suppose there is a single endogenous regressor X_2, no X_1, and two instruments Z_1 and Z_2. Then the model specifies that

$$\mathbb{E}([Z_1 Y] = \mathbb{E}[Z_1 X_2]\beta$$

and

$$\mathbb{E}[Z_2 Y] = \mathbb{E}[Z_2 X_2]\beta.$$

Thus β solves both equations. This is rather special.

Another way of thinking about this is we could solve for β using either one equation or the other. In terms of estimation this is equivalent to estimating by IV using just the instrument Z_1 or instead just using the instrument Z_2. These two estimators (in finite samples) are different. But if the overidentification hypothesis is correct both are estimating the same parameter and both are consistent for β. In contrast, if the overidentification hypothesis is false then the two estimators will converge to different probability limits and it is unclear if either probability limit is interesting.

For example, take the 2SLS estimates in the fourth column of Table 12.1 which use *public* and *private* as instruments for *education*. Suppose we instead estimate by IV using just *public* as an instrument and then repeat using *private*. The IV coefficient for *education* in the first case is 0.16 and in the second case 0.27. These appear to be quite different. However, the second estimate has a large standard error (0.16) so the difference may be sampling variation. An overidentification test addresses this question.

For a general overidentification test the null and alternative hypotheses are $\mathbb{H}_0 : \mathbb{E}[Ze] = 0$ against $\mathbb{H}_1 : \mathbb{E}[Ze] \neq 0$. We will also add the conditional homoskedasticity assumption

$$\mathbb{E}\left[e^2 \mid Z\right] = \sigma^2. \tag{12.65}$$

To avoid (12.65) it is best to take a GMM approach which we defer until Chapter 13.

To implement a test of $\mathbb{H}_0$ consider a linear regression of the error e on the instruments Z

$$e = Z'\alpha + \nu \tag{12.66}$$

with $\alpha = \left(\mathbb{E}\left[ZZ'\right]\right)^{-1}\mathbb{E}[Ze]$. We can rewrite $\mathbb{H}_0$ as $\alpha = 0$. While e is not observed we can replace it with the 2SLS residual $\widehat{e}_i$ and estimate α by least squares regression, e.g. $\widehat{\alpha} = \left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\widehat{\boldsymbol{e}}$. Sargan (1958) proposed testing $\mathbb{H}_0$ via a score test, which equals

$$S = \widehat{\alpha}'\left(\widehat{\operatorname{var}}\left[\widehat{\alpha}\right]\right)^{-}\widehat{\alpha} = \frac{\widehat{\boldsymbol{e}}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\widehat{\boldsymbol{e}}}{\widehat{\sigma}^2}. \tag{12.67}$$

where $\widehat{\sigma}^2 = \frac{1}{n}\widehat{\boldsymbol{e}}'\widehat{\boldsymbol{e}}$. Basmann (1960) independently proposed a Wald statistic for $\mathbb{H}_0$, which is S with $\widehat{\sigma}^2$ replaced with $\widetilde{\sigma}^2 = n^{-1}\widehat{\boldsymbol{\nu}}'\widehat{\boldsymbol{\nu}}$ where $\widehat{\boldsymbol{\nu}} = \widehat{\boldsymbol{e}} - \boldsymbol{Z}\widehat{\alpha}$. By the equivalence of homoskedastic score and Wald tests (see Section 9.16) Basmann's statistic is a monotonic function of Sargan's statistic and hence they yield equivalent tests. Sargan's version is more typically reported.

The Sargan test rejects $\mathbb{H}_0$ in favor of $\mathbb{H}_1$ if $S > c$ for some critical value c. An asymptotic test sets c as the $1-\alpha$ quantile of the $\chi^2_{\ell-k}$ distribution. This is justified by the asymptotic null distribution of S which we now derive.

Theorem 12.16 Under Assumption 12.2 and $\mathbb{E}\left[e^2 \mid Z\right] = \sigma^2$, then as $n \to \infty$, $S \xrightarrow[d]{} \chi^2_{\ell-k}$. For c satisfying $\alpha = 1 - G_{\ell-k}(c)$, $\mathbb{P}[S > c \mid \mathbb{H}_0] \to \alpha$ so the test "Reject $\mathbb{H}_0$ if $S > c$" has asymptotic size α.

We prove Theorem 12.16 below.

The Sargan statistic S is an asymptotic test of the overidentifying restrictions under the assumption of conditional homoskedasticity. It has some limitations. First, it is an asymptotic test and does not have a finite sample (e.g. F) counterpart. Simulation evidence suggests that the test can be oversized (reject too frequently) in small and moderate sample sizes. Consequently, p-values should be interpreted cautiously. Second, the assumption of conditional homoskedasticity is unrealistic in applications. The best way to generalize the Sargan statistic to allow heteroskedasticity is to use the GMM overidentification statistic – which we will examine in Chapter 13. For 2SLS, Wooldrige (1995) suggested a robust score test, but Baum, Schaffer and Stillman (2003) point out that it is numerically equivalent to the GMM overidentification statistic. Hence the bottom line appears to be that to allow heteroskedasticity or clustering it is best to use a GMM approach.

In overidentified applications it is always prudent to report an overidentification test. If the test is insignificant it means that the overidentifying restrictions are not rejected, supporting the estimated model. If the overidentifying test statistic is highly significant (if the p-value is very small) this is evidence that the overidentifying restrictions are violated. In this case we should be concerned that the model is misspecified and interpreting the parameter estimates should be done cautiously.

When reporting the results of an overidentification test it seems reasonable to focus on very small significance levels such as 1%. This means that we should only treat a model as "rejected" if the Sargan p-value is very small, e.g. less than 0.01. The reason to focus on very small significance levels is because it is very difficult to interpret the result "The model is rejected". Stepping back a bit it does not seem credible that any overidentified model is literally true; rather what seems potentially credible is that an overidentified model is a reasonable approximation. A test is asking the question "Is there evidence that a model is not true" when we really want to know the answer to "Is there evidence that the model is a poor approximation". Consequently it seems reasonable to require strong evidence to lead to the conclusion "Let's reject this model". The recommendation is that mild rejections (p-values between 1% and 5%) should be viewed as mildly worrisome but not critical evidence against a model. The results of an overidentification test should be integrated with other information before making a strong decision.

We illustrate the methods with the Card college proximity example. We have estimated two overidentified models by 2SLS in columns 4 & 5 of Table 12.1. In each case the number of overidentifying restrictions is 1. We report the Sargan statistic and its asymptotic p-value (calculated using the χ_1^2 distribution) in the table. Both p-values (0.37 and 0.47) are far from significant indicating that there is no evidence that the models are misspecified.

We now prove Theorem 12.16. The statistic S is invariant to rotations of $\boldsymbol{Z}$ (replacing $\boldsymbol{Z}$ with $\boldsymbol{ZC}$) so without loss of generality we assume $\mathbb{E}\left[ZZ'\right] = \boldsymbol{I}_\ell$. As $n \to \infty$, $n^{-1/2}\boldsymbol{Z}'\boldsymbol{e} \underset{d}{\longrightarrow} \sigma Z$ where $Z \sim \mathrm{N}(0, \boldsymbol{I}_\ell)$. Also $\frac{1}{n}\boldsymbol{Z}'\boldsymbol{Z} \underset{p}{\longrightarrow} \boldsymbol{I}_\ell$ and $\frac{1}{n}\boldsymbol{Z}'\boldsymbol{X} \underset{p}{\longrightarrow} \boldsymbol{Q}$, say. Then

$$\begin{aligned} n^{-1/2}\boldsymbol{Z}'\widehat{\boldsymbol{e}} &= \left(\boldsymbol{I}_\ell - \left(\frac{1}{n}\boldsymbol{Z}'\boldsymbol{X}\right)\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{P}_Z\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\right)\left(\frac{1}{n}\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\right) n^{-1/2}\boldsymbol{Z}'\boldsymbol{e} \\ &\underset{d}{\longrightarrow} \sigma\left(\boldsymbol{I}_\ell - \boldsymbol{Q}\left(\boldsymbol{Q}'\boldsymbol{Q}\right)^{-1}\boldsymbol{Q}'\right)\mathrm{Z}. \end{aligned}$$

Since $\widehat{\sigma}^2 \underset{p}{\longrightarrow} \sigma^2$ it follows that

$$S \underset{d}{\longrightarrow} \mathrm{Z}'\left(\boldsymbol{I}_\ell - \boldsymbol{Q}\left(\boldsymbol{Q}'\boldsymbol{Q}\right)^{-1}\boldsymbol{Q}'\right)\mathrm{Z} \sim \chi^2_{\ell-k}.$$

The distribution is $\chi^2_{\ell-k}$ since $\boldsymbol{I}_\ell - \boldsymbol{Q}\left(\boldsymbol{Q}'\boldsymbol{Q}\right)^{-1}\boldsymbol{Q}'$ is idempotent with rank $\ell - k$.

The Sargan statistic test can be implemented in Stata using the command `estat overid` after `ivregress 2sls` or `ivregres liml` if a standard (non-robust) covariance matrix has been specified (that is, without the '`,r`' option), or by the command `estat overid, forcenonrobust` otherwise.

Denis Sargan

The British econometrician John Denis Sargan (1924-1996) was a pioneer in the field of econometrics. He made a range of fundamental contributions including the overidentification test, Edgeworth expansions, and unit root theory. He was also influential in his role as dissertation advisor for many LSE-trained econometricians.