Econ 551: Lecture Notes
Endogenous Regressors and Instrumental Variables
Professor John Rust

0. Introduction

These notes introduce students to the problem of endogeneity in linear models and the method of *instrumental variables* that under certain circumstances allows consistent estimation of the structural coefficients of the endogenous regressors in the linear model. Sections 1 and 2 review the linear model and the method of ordinary least squares (OLS) in the abstract (L^2) setting, and the concrete (R^N) setting. The abstract setting allows us to define the "theoretical" regression coefficient to which the sample OLS estimator converges as the sample size $N \longrightarrow \infty$. Section 3 discusses the issue of non-uniqueness of the OLS coefficients if the regressor matrix does not have full rank, and describes some ways to handle this. Seftion 4 reviews the two key asymptotic properties of the OLS estimator, consistency and asymptotic normality. It derives a heteroscedasticity-consistent covariance matrix estimator for the limiting normal asymptotic distribution of the standardized OLS estimator. Section 5 introduces the problem of endogeneity, showing how it can arise in three different contexts. The next three sections demonstrate how the OLS estimator may not converge to the true coefficient values when we assume that the data are generated by some "true" underlying structural linear model. Section 6 discusses the problem of omitted variable bias. Section 7 discusses the problem of measurement error. Section 8 discusses the problem of simultaneous equations bias. Section 9 introduces the concept of an *instrumental variable* and proves the optimality of the *two stage least squares* (2SLS) estimator.

1. The Linear Model and Ordinary Least Squares (OLS) in L^2: We consider regression first in the abstract setting of the Hilbert space L^2. It is convenient to start with this infinite-dimensional space version of regression, since the least squares estimates can be viewed as the limiting result of doing OLS in R^N, as $N \to \infty$. In L^2 it is more transparent that we can do OLS under very general conditions, without assuming non-stochastic regressors, homoscedasticity, normally distributed errors, or that the true regression function is linear. *Regression is simply the process of orthogonal projection of a dependent variable $\tilde{y} \in L^2$ onto the linear subspace space spanned by K random variables $\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_K)$.* To be concrete, let $\tilde{y}$ be a 1×1 dependent variable and $\tilde{X}$ is a $1 \times K$ vector of explanatory variables. Then as long as $E\{\tilde{X}'\tilde{y}\}$ and $E\{\tilde{X}'\tilde{X}\}$ exist and are finite, and as long as $E\{\tilde{X}'\tilde{X}\}$ is a nonsingular matrix, then we have the identity:

$$\underset{1\times 1}{\tilde{y}} = \underset{1\times K}{\tilde{X}} \underset{K\times 1}{\beta^*} + \underset{1\times 1}{\tilde{\epsilon}} \tag{1}$$

where β^* is the least squares estimate given by:

$$\beta^* = \underset{\beta}{argmin}\, E\{(\tilde{y} - \tilde{X}\beta)^2\} = [E\{\tilde{X}'\tilde{X}\}]^{-1}E\{\tilde{X}'\tilde{y}\}. \tag{2}$$

Note by construction, the residual term $\tilde{\epsilon}$ is orthogonal to the regressor vector $\tilde{X}$,

$$E\{\tilde{X}'\tilde{\epsilon}\} = 0, \tag{3}$$

where $\langle \tilde{y}, \tilde{x} \rangle = E\{\tilde{y}, \tilde{x}\}$ defines the inner product between two random variables in L^2. The orthogonality condition (3) implies the *Pythagorean Theorem*

$$\|\tilde{y}\|^2 = \|\tilde{X}\beta^*\|^2 + \|\tilde{\epsilon}\|^2, \tag{4}$$

where $\|\tilde{y}\|^2 \equiv \langle \tilde{y}, \tilde{y} \rangle$. From this we define the R^2 as

$$R^2 = \frac{\|\tilde{y}\|^2}{\|\tilde{X}\beta^*\|^2}. \tag{5}$$

Conceptually, R^2 is the cosine of the angle θ between the vectors $\tilde{y}$ and $\tilde{X}\beta^*$ in L^2. The main point here is that the linear model (1) holds "by construction", regardless of whether the true relationship between $\tilde{y}$ and $\tilde{X}$, the conditional expectation $E\{\tilde{y}|\tilde{X}\}$ is a linear or nonlinear function of $\tilde{X}$. In fact, the latter is simply the result of projecting $\tilde{y}$ into a larger subspace of L^2, the space of all measurable functions of $\tilde{X}$. The second point is that definition of β^* insures the $\tilde{X}$ matrix is "exogenous" in the sense of equation (3), i.e. the error term $\tilde{\epsilon}$ is uncorrelated with the regressors $\tilde{X}$. In effect, we define β^* in such a way so the regressors $\tilde{X}$ are *exogenous by construction.* It is instructive to repeat the simple mathematics leading up to this second conclusion. Using the identity (1) and the definition of β^* in (2) we have:

$$\begin{aligned}
E\{\tilde{X}'\tilde{\epsilon}\} &= E\{\tilde{X}'(\tilde{y} - \tilde{X}\beta^*)\} && (6)\\
&= E\{\tilde{X}'\tilde{y}\} - E\{\tilde{X}'\tilde{X}\}\beta^* && (7)\\
&= E\{\tilde{X}'\tilde{y}\} - E\{\tilde{X}'\tilde{X}\}[E\{\tilde{X}'\tilde{X}\}]^{-1}E\{\tilde{X}'\tilde{y}\} && (8)\\
&= E\{\tilde{X}'\tilde{y}\} - E\{\tilde{X}'\tilde{y}\} && (9)\\
&= 0. && (10)
\end{aligned}$$

2. The Linear Model and Ordinary Least Squares (OLS) in R^N: Consider regression in the "concrete" setting of the Hilbert space R^N. The dimension N is the number of observations, where we assume that these observations are *IID* realizations of the vector of random variables $(\tilde{y}, \tilde{X})$. Define $y = (y_1, \ldots, y_N)'$ and $X = (X_1, \ldots, X_N)'$, where each y_i is 1×1 and each x_i is i$1 \times K$. Note y is now a vector in R^N. We can represent the $N \times K$ matrix X as K vectors in R^N: $X = (X^1, \ldots, X^K)$, where X^j is the j^{th} column of X, a vector in R^N. *Regression is simply the process of orthogonal projection of the dependent variable $y \in R^N$ onto the linear subspace spanned by the K columns of $X = (X^1 \ldots, X^K)$.* This gives us the identity:

$$\underset{N\times 1}{y} = \underset{N\times K}{X}\underset{K\times 1}{\hat{\beta}} + \underset{N\times 1}{\epsilon} \tag{11}$$

where $\hat{\beta}$ is the least squares estimate given by:

$$\hat{\beta} = \underset{\beta}{argmin}\, \frac{1}{N}\sum_{i=1}^{N}(y_i - X_i\beta)^2 = \left[\frac{1}{N}\sum_{i=1}^{N} X_i'X_i\right]^{-1}\left[\frac{1}{N}\sum_{i=1}^{N} X_i'y_i\right] \tag{12}$$

and by construction, the $N \times 1$ residual vector ϵ is orthogonal to the $N \times K$ matrix of regressors:

$$\frac{1}{N}\sum_{i=1}^{N} X_i'\epsilon_i = 0. \tag{13}$$

where $\langle y, x\rangle = \sum_{i=1}^{N} y_i x_i/N$ defines the inner product between two random variables in the Hilbert space R^N. The orthogonality condition (13) implies the *Pythagorean Theorem*

$$\|y\|^2 = \|X\hat{\beta}\|^2 + \|\epsilon\|^2, \tag{14}$$

where $\|y\|^2 \equiv \langle y, y\rangle$. From this we define the (uncentered) R^2 as

$$R^2 = \frac{\|y\|^2}{\|X\hat{\beta}\|^2}. \tag{15}$$

Conceptually, R^2 is the cosine of the angle θ between the vectors y and $X\beta^*$ in R^N.

The main point of these first two sections is that the linear model — viewed either as a linear relationship between a "dependent" random variable $\tilde{y}$ and a $1 \times K$ vector of "independent" random variables $\tilde{X}$ in L^2 as in equation (1), or as a linear relationship between a vector-valued dependent variable y in R^N, and K independent variables making up the columns of the $N \times K$ matrix X in equation (11) — both hold "by construction". That is, regardless of whether the true relationship between y and X is linear, under very general conditions the Projection Theorem for Hilbert Spaces guarantees that there exists $K \times 1$ vectors β^* and $\hat{\beta}$ such that $\tilde{X}\beta^*$ and $X\hat{\beta}$ equal the orthogonal projections of $\tilde{y}$ and y onto the K-dimensional subspace of L^2 and R^N spanned by the K variables in $\tilde{X}$ and X, respectively. These coefficient vectors a constructed in such a way as to force the error terms $\tilde{\epsilon}$ and $\hat{\epsilon}$ to be orthogonal to $\tilde{X}$ and X, respectively. When we speak about the problem of *endogeneity,* we mean a situation where we believe there is a that there is a "true linear model" $\tilde{y} = \tilde{X}\beta_0 + \tilde{\epsilon}$ relating $\tilde{y}$ to $\tilde{X}$ where the "true coefficient vector" β_0 is not necessarily equal to the least squares value β^*, i.e. the error $\tilde{\epsilon} = \tilde{y} - \tilde{X}\beta_0$ is not necessarily orthogonal to $\tilde{X}$. We will provide several examples of how endogeneity can arise after reviewing the asymptotic properties of the OLS estimator.

3. Note on the Uniqueness of the Least Squares Coefficients

The *Projection Theorem* guarantees that in any Hilbert space H (including the two special cases L^2 and R^N discussed above), the *projection* $P(y|X)$ exists, where $P(y|X)$ is the *best linear predictor* of an element $y \in H$. More precisely, if $X = (X_1, \ldots, X_K)$ where each $X_i \in H$, then $P(y|X)$ is the element of the smallest closed linear subspace spanned by the elements of X, $\text{lin}(X)$ that is closest to y:

$$P(y|X) = \underset{\hat{y} \in \text{lin}(X)}{argmin} \|y - \hat{y}\|^2, \tag{16}$$

It is easy to show that $\text{lin}(X)$ is a finite-dimensional linear subspace with dimension $J \leq K$. The projection theorem tells us that $P(y|X)$ is always uniquely defined, even if it can be represented as different linear combinations of the elements of X. However if X has *full rank,* the projection $P(y|X)$ will have a unique representation given by

$$\begin{aligned} P(y|X) &= X\hat{\beta} \\ \hat{\beta} &= \underset{\beta \in R^K}{argmin} \|y - X\beta\|^2 \end{aligned} \tag{17}$$

Definition: We say X has *full rank,* if $J = K$, i.e. if the dimension of the linear subspace $\text{lin}(X)$ spanned by the elements of X equals the number of elements in X.

It is straightforward to show that X has full rank if and only if the K elements of X are linearly independent, which happens if and only if the $K \times K$ matrix $X'X$ is invertible. We use the heuristic notation $X'X$ to denote the matrix whose (i,j) element is $\langle X_i, X_j \rangle$. To see the latter claim, suppose $X'X$ is singular. Then there exists a vector $a \in R^K$ such that $a \neq \mathbf{0}$ and $X'Xa = \mathbf{0}$, where $\mathbf{0}$ is the zero vector in R^K. Then we have $a'X'Xa = 0$ or in inner product notation

$$\langle Xa, Xa \rangle = \|Xa\|^2 = 0. \tag{18}$$

However in a Hilbert space, an element has a norm of 0 iff it equals the 0 element in H. Since $a \neq \mathbf{0}$, we can assume without loss of generality that $a_1 \neq 0$. Then we can rearrange the equation $Xa = 0$ and solve for X_1 to obtain:

$$X_1 = X_2\alpha_2 + \cdots X_K\alpha_K, \tag{19}$$

where $\alpha_i = -a_i/a_1$. Thus, if $X'X$ is not invertible then X can't have full rank, since one of more elements of X are redundant in the sense that they can be exactly predicted by a linear combination of the remaining elements of X. Thus, it is just a matter of convention to eliminate the redundant elements of X to guarantee that it has full rank, which ensures that $X'X$ exists and the least squares coefficient vector $\hat{\beta}$ is uniquely defined by the standard formula

$$\hat{\beta} = [X'X]^{-1}X'y. \tag{20}$$

Notice that the above equation applies to arbitrary Hilbert spaces H and is a shorthand for the $\beta \in R^K$ that solves the following system of linear equations that consistent the *normal equations* for least squares:

$$\begin{aligned} \langle y, X_1 \rangle &= \langle X_1, X_1 \rangle \beta_1 + \cdots + \langle X_K, X_1 \rangle \beta_K \\ &\cdots \\ \langle y, X_K \rangle &= \langle X_1, X_K \rangle \beta_1 + \cdots + \langle X_K, X_K \rangle \beta_K \end{aligned} \tag{21}$$

The normal equations follow from the orthogonality conditions $\langle X_i, \epsilon\rangle = \langle X_i, y - X\beta\rangle = 0$, and can be written more compactly in matrix notation as

$$X'y = X'X\beta \tag{22}$$

which is easily seen to be equivalent to the formula in equation (20) when X has full rank and the $K \times K$ matrix $X'X$ is invertible.

When X does not have full rank there are multiple solutions to the normal equations, all of which yield the same best prediction, $P(y|X)$. In this case there are several ways to proceed. The most common way is to eliminate the redundant elements of X until the resulting reduced set of regressors has full rank. Alternatively one can compute $P(y|X)$ via *stepwise regression* by squentially projecting y on X_1, then projecting $y - P(y|X_1)$ on $X_2 - P(X_2|X_1)$ and so forth. Finally, one can single out one of the many β vectors that solve the normal equations to compute $P(y|X)$. One approach is to use the *shortest* vector β solving the normal equation, and leads to the following formula

$$\hat{\beta} = [X'X]^{+}X'y \tag{23}$$

where $[X'X]^{+}$ is the *generalized inverse* of the square but non-invertible matrix $X'X$. The generalized inverse is computed by calculating the Jordan decomposition of $[X'X]$ into a product of an orthonormal matrix W (i.e. a matrix satisfying $W'W = WW' = I$) and a diagonal matrix D whose diagonal elements are the eigenvalues of $[X'X]$,

$$[X'X] = WDW'. \tag{24}$$

Then the generalized inverse is defined by

$$[X'X]^{+} = WD^{+}W'. \tag{25}$$

where D^{+} is the diagonal matrix whose i^{th} diagonal element is $1/D_{ii}$ if the corresponding diagonal element D_{ii} of D is nonzero, and 0 otherwise.

Exercise: Prove that the generalized formula for $\hat{\beta}$ given in equation (23) does in fact solve the normal equations and results in a valid solution for the best linear predictor $P(y|X) = X\hat{\beta}$. Also, verify that among all solutions to the normal equations, $\hat{\beta}$ has the smallest norm.

4. Asymptotics of the OLS estimator. The sample OLS estimator $\hat{\beta}$ can be viewed as the result of applying the "analogy principle", i.e. replacing the theoretical expectations in (2) with sample averages in (12). The Strong Law of Large Numbers (SLLN) implies that as $N \to \infty$ we have with probability 1,

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - X_i\beta)^2 \longrightarrow E\{(\tilde{y} - \tilde{X}\beta)^2\}. \tag{26}$$

The convergence above can be proven to hold uniformly for β in compact subsets of R^K. This implies a *Uniform Strong Law of Large Numbers* (USLLN) that implies the consistency of the OLS estimator (see Rust's lecture notes on "Proof of the Uniform Law of Large Numbers"). Specifically, assuming β^* is uniquely identified (i.e. that it is the unique minimizer of $E\{(\tilde{y} - \tilde{X}\beta)^2\}$, a result which holds whenever $\tilde{X}$ has full rank as we saw in section 3), then with probability 1 we have

$$\hat{\beta} \longrightarrow \beta^*. \tag{27}$$

Given that we have a closed-form expression for the OLS estimators β^* in equation (2) and $\hat{\beta}$ in equation (12), consistency can be established more directly by observing that the SLLN implies that with probability 1 the sample moments

$$\frac{1}{N}\sum_{i=1}^{N} X_i' X_i \longrightarrow E\{\tilde{X}'\tilde{X}\} \quad \frac{1}{N}\sum_{i=1}^{N} X_i' y_i \longrightarrow E\{\tilde{X}'\tilde{y}\}. \tag{28}$$

So a direct appeal to Slutsky's Theorem establishes the consistency of the OLS estimator, $\hat{\beta} \longrightarrow \beta^*$, with probability 1.

The asymptotic distribution of the normalized OLS estimator, $\sqrt{N}(\hat{\beta}-\beta^*)$, can be derived by appealing to the Lindeberg-Levy Central Limit Theorem (CLT) for *IID* random vectors. That is we assume that $\{(y_1, X_1), \ldots, (y_N, X_N)\}$ are *IID* draws from some joint distribution $F(y, X)$. Since $y_i = X_i\beta^* + \epsilon_i$ where $E\{X_i'\tilde{\epsilon}_i\} = 0$ and

$$\mathrm{var}(\tilde{X}\tilde{\epsilon}) = \mathrm{cov}(\tilde{X}'\tilde{\epsilon}, \tilde{X}'\epsilon\} = E\{\tilde{\epsilon}^2\tilde{X}'\tilde{X}\} \equiv \Omega, \tag{29}$$

the CLT implies that

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N} X_i'\epsilon_i \Longrightarrow_d N(0, \Omega). \tag{30}$$

Then, substituting for y_i in the definition of $\hat{\beta}$ in equation (12) and rearranging we get:

$$\sqrt{N}\left[\hat{\beta} - \beta^*\right] = \left[\frac{1}{N}\sum_{i=1}^{N} X_i' X_i\right]^{-1} \frac{1}{N}\sum_{i=1}^{N} X_i'\epsilon_i. \tag{31}$$

Appealing to the Slutsky Theorem and the CLT result in equation (30), we have:

$$\sqrt{N}\left[\hat{\beta} - \beta^*\right] \Longrightarrow N(0, \Lambda), \tag{32}$$

where the $K \times K$ covariance matrix Λ is given by:

$$\Lambda = [E\{\tilde{X}'\tilde{X}\}]^{-1}[E\{\tilde{\epsilon}^2\tilde{X}'\tilde{X}\}][E\{\tilde{X}'\tilde{X}\}]^{-1}. \tag{33}$$

In finite samples we can form a consistent estimator of Λ using the *heteroscedasticity-consistent covariance matrix estimator* $\hat{\Lambda}$ given by:

$$\hat{\Lambda} = \left[\frac{1}{N}\sum_{i=1}^{N} X_i' X_i\right]^{-1} \left[\frac{1}{N}\sum_{i=1}^{N} \hat{\epsilon}_i^2 X_i' X_i\right] \left[\frac{1}{N}\sum_{i=1}^{N} X_i' X_i\right]^{-1}, \tag{34}$$

where $\hat{\epsilon}_i = y_i - X_i\hat{\beta}$. Actually, there is a somewhat subtle issue in proving that $\hat{\Lambda} \longrightarrow \Lambda$ with probability 1. We cannot directly appeal to the SLLN to show that

$$\frac{1}{N}\sum_{i=1}^{N} \hat{\epsilon}_i^2 X_i' X_i \longrightarrow E\{\tilde{\epsilon}^2\tilde{X}'\tilde{X}\}, \tag{35}$$

since the estimated residuals $\hat{\epsilon}_i = \epsilon_i + X_i(\hat{\beta} - \beta^*)$ are not *IID* random variables due to their common dependence on $\hat{\beta}$. To establish the result we must appeal to the Uniform Law of Large Numbers to show that uniformly for β in a compact subset of R^K we have:

$$\frac{1}{N}\sum_{i=1}^{N}[X_i(\beta - \beta^*)]^2 X_i' X_i \longrightarrow E\{\tilde{X}(\beta - \beta^*)\tilde{X}'\tilde{X}\}. \tag{36}$$

Further more we must appeal to the following *uniform convergence lemma:*

Lemma: *If $g_n \longrightarrow g$ uniformly with probability 1 for β in a compact set, and if $\beta_n \longrightarrow \beta^*$ with probability 1, then with probability 1 we have:*

$$g_n(\beta_n) \longrightarrow g(\beta^*). \tag{37}$$

These results enable us to show that

$$\begin{aligned} \frac{1}{N}\sum_{i=1}^{N} \hat{\epsilon}_i^2 X_i' X_i &= \frac{1}{N}\sum_{i=1}^{N} \epsilon_i^2 X_i' X_i + \frac{2}{N}\sum_{i=1}^{N}[\epsilon_i X_i(\hat{\beta}-\beta^*)]X_i'X_i + \frac{1}{N}\sum_{i=1}^{N}[X_i(\hat{\beta}-\beta^*)]^2 X_i'X_i \\ &\longrightarrow E\{\tilde{\epsilon}^2 \tilde{X}'\tilde{X}\} + \mathbf{0} + \mathbf{0}, \end{aligned} \tag{38}$$

where $\mathbf{0}$ is a $K \times K$ matrix of zeros. Notice that we appealed to the ordinary SLLN to show that the first term on the right hand side of equation (38) converges to $E\{\tilde{\epsilon}^2 \tilde{X}'\tilde{X}\}$ and the uniform convergence lemma to show that the remaining two terms converge to $\mathbf{0}$.

Finally, note that under the assumption of conditional independence, $E\{\tilde{\epsilon}|\tilde{X}\} = 0$, and homoscedasticity, $\text{var}(\tilde{\epsilon}|\tilde{X}\} = E\{\tilde{\epsilon}^2|\tilde{X}\} = \sigma^2$, the covariance matrix Λ simplifies to the usual textbook formula:

$$\Lambda = \sigma^2 \left[E\{\tilde{X}'\tilde{X}\}\right]^{-1}. \tag{39}$$

However since there is no compelling reason to believe the linear model is homoscedastic, it is in general a better idea to play it safe and use the heteroscedasticity-consistent estimator given in equation (34).

5. Structural Models and Endogeneity As we noted above, the OLS parameter vector β^* exists under very weak conditions, and the OLS estimator $\hat{\beta}$ converges to it. Further, by construction the residuals $\epsilon = y - X\beta^*$ are orthogonal to X. However there are a number of cases where we believe there is a linear relationship between y and X,

$$y = X\beta_0 + \eta \tag{40}$$

where β_o is not necessarily equal to the OLS vector $\beta*$ and the error term η is not necessarily orthogonal to X. This situation can occur for at least three different reasons:

1. **Omitted variable bias**
2. **Errors in variables**
3. **Simultaneous equations bias**

We will consider omitted variable bias and errors in variables first since they are the easiest cases to understand how endogeneity problems arise. Then in the next section we will consider the simultaneous equations problem in more detail.

6. Omitted Variable Bias

Suppose that the true model is linear, but that we don't observe a subset of variables $\tilde{X}_2$ which are known to affect y. Thus, the "true" regression function can be written as:

$$\tilde{y} = \tilde{X}_1\beta_1 + \tilde{X}_2\beta_2 + \tilde{\epsilon}, \tag{41}$$

where $\tilde{X}_1$ is $1 \times K_1$ and $\tilde{X}_2$ is $1 \times K_2$, and $E\{\tilde{X}_1'\epsilon\} = 0$ and $E\{\tilde{X}_2'\epsilon\} = 0$. Now if we don't observe X_2, the OLS estimator $\hat{\beta}_1 = [X_1'X_1]^{-1}X_1'y$ based on N observations of the random variables $(\tilde{y}, \tilde{X}_1)$ converges to

$$\hat{\beta}_1 = [X_1'X_1]^{-1}X_1'y \longrightarrow \left[E\{\tilde{X}_1'\tilde{X}_1\}\right]^{-1} E\{\tilde{X}_1'\tilde{y}\}. \tag{42}$$

However we have:

$$E\{\tilde{X}_1'\tilde{y}\} = E\{\tilde{X}_1'\tilde{X}_1\}\beta_1 + E\{\tilde{X}_1'\tilde{X}_2\}\beta_2 \tag{43}$$

since $E\{\tilde{X}_2'\epsilon\} = 0$ for the "true regression model" when both $\tilde{X}_1$ and $\tilde{X}_2$ are included. Substituting equation (43) into equation (42) we obtain:

$$\hat{\beta}_1 = \longrightarrow \beta_1 + \left[E\{\tilde{X}_1'\tilde{X}_1\}\right]^{-1} E\{\tilde{X}_1'\tilde{X}_2\}\beta_2. \tag{44}$$

We can see from this equation that the OLS estimator will generally not converge to the true parameter vector β_1 when there are omitted variables, except in the case where either $\beta_2 = 0$ or where $E\{\tilde{X}_1'\tilde{X}_2\} = 0$, i.e. where the omitted variables are orthogonal to the observed included variables $\tilde{X}_1$. Now consider the "auxiliary regression between $\tilde{X}_2$ and $\tilde{X}_1$:

$$\tilde{X}_2 = \tilde{X}_1\gamma + \tilde{\xi}, \tag{45}$$

where $\gamma = [E\{\tilde{X}_1'\tilde{X}_1\}]^{-1}E\{\tilde{X}_1'\tilde{X}_2\}$ is a $(K_1 \times K_2)$ matrix of regression coefficients, i.e. equation (45) denotes a *system* of K_2 regressions written in compact matrix notation. Note that by construction we have $E\{\tilde{X}_1'\xi\} = 0$. Substituting equation (45) into equation (44) and simplifying, we obtain:

$$\hat{\beta}_1 \longrightarrow \beta_1 + \gamma\beta_2. \tag{46}$$

In the special case where $K_1 = K_2 = 1$, we can characterize the omitted variable bias $\gamma\beta_2$ as follows:

1. The asymptotic bias is 0 if $\gamma = 0$ or $\beta_2 = 0$, i.e. if $\tilde{X}_2$ doesn't enter the regression equation $(\beta_2 = 0)$, or if $\tilde{X}_2$ is orthogonal to $\tilde{X}_1$ $(\gamma = 0)$. In either case, the restricted regression $\tilde{y} = \tilde{X}_1\beta_1 + \nu$ where $\nu = \tilde{X}_2\beta_2 + \epsilon$ is a valid regression and $\hat{\beta}_1$ is a consistent estimator of β_1.

2. The asymptotic bias is positive if $\beta_2 > 0$ and $\gamma > 0$, or if $\beta_2 < 0$ and $\gamma < 0$. In this case, OLS converges to a distorted parameter $\beta_1 + \gamma\beta_2$ which overestimates β_1 in order to "soak up" the part of the unobserved $\tilde{X}_2$ variable that is correlated with $\tilde{X}_1$.

3. The asymptotic bias is negative if $\beta_2 > 0$ and $\gamma < 0$, or if $\beta_2 < 0$ and $\gamma > 0$. In this case, OLS converges to a distorted parameter $\beta_1 + \gamma\beta_2$ which underestimates β_1 in order to "soak up" the part of the unobserved $\tilde{X}_2$ variable that is correlated with $\tilde{X}_1$.

Note that in cases 2. and 3., the OLS estimator $\hat{\beta}_1$ converges to a biased limit $\beta_1 + \gamma\beta_2$ to ensure that the error term $\tilde{\eta} = \tilde{X}_1[\beta_1 + \gamma\beta_2]$ is orthogonal to $\tilde{X}_1$.

Exercise: Using the above equations, show that $E\{\tilde{X}_1'\tilde{\eta}\} = 0$.

Now consider how a regression that includes both $\tilde{X}_1$ and $\tilde{X}_2$ automatically "adjusts" to converge to the true parameter vectors β_1 and β_2. Note that the normal equations when we have both $\tilde{X}_1$ and $\tilde{X}_2$ are given by:

$$\begin{aligned} E\{\tilde{X}_1'\tilde{y}\} &= E\{\tilde{X}_1'\tilde{X}_1\}\beta_1 + E\{\tilde{X}_1'\tilde{X}_2\}\beta_2 \\ E\{\tilde{X}_2'\tilde{y}\} &= E\{\tilde{X}_2'\tilde{X}_1\}\beta_1 + E\{\tilde{X}_2'\tilde{X}_2\}\beta_2. \end{aligned} \tag{47}$$

Solving the first normal equation for β_1 we obtain:

$$\beta_1 = [E\{\tilde{X}_1'\tilde{X}_1\}]^{-1}[E\{\tilde{X}_1'\tilde{y}\} - E\{\tilde{X}_1'\tilde{X}_2\}\beta_2]. \tag{48}$$

Thus, the full OLS estimator for β_1 equals the biased OLS estimator that omits $\tilde{X}_2$, $[E\{\tilde{X}_1'\tilde{X}_1\}]^{-1}E\{\tilde{X}_1'\tilde{y}\}$, less a "correction term" $[E\{\tilde{X}_1'\tilde{X}_1\}]^{-1}E\{\tilde{X}_1'\tilde{X}_2\}\beta_2$ that exactly offsets the asymptotic omitted variable bias $\gamma\beta_2$ of OLS derived above.

Now, substituting the equation for β_1 into the second normal equation and solving for β_2 we obtain:

$$\left[E\{\tilde{X}_2'\tilde{X}_2\} - E\{\tilde{X}_2'\tilde{X}_1\}[E\{\tilde{X}_1'\tilde{X}_1\}]^{-1}E\{\tilde{X}_1'\tilde{X}_2\}\right]^{-1}\left[E\{\tilde{X}_2'\tilde{y}\} - E\{\tilde{X}_2'\tilde{X}_1\}[E\{\tilde{X}_1'\tilde{X}_1\}]^{-1}E\{\tilde{X}_1'\tilde{y}\}\right]. \tag{49}$$

The above formula has a more intuitive interpretation: β_2 can be obtained by regressing $\tilde{y}$ on $\tilde{\xi}$, where $\tilde{\xi}$ is the residual from the regression of $\tilde{X}_2$ on $\tilde{X}_1$:

$$\beta_2 = [E\{\tilde{\xi}'\tilde{\xi}\}]^{-1}E\{\tilde{\xi}'\tilde{y}\}. \tag{50}$$

This is just the result of the second step of stepwise regression where the first step regresses $\tilde{y}$ on $\tilde{X}_1$, and the second step regresses the residuals $\tilde{y} - P(\tilde{y}|\tilde{X}_1)$ on $\tilde{X}_2 - P(\tilde{X}_2|\tilde{X}_1)$, where $P(\tilde{X}_2|\tilde{X}_1)$ denotes the projection of $\tilde{X}_2$ on $\tilde{X}_1$, i.e. $P(\tilde{X}_2|\tilde{X}_1) = \tilde{X}_1\gamma$ where γ is given in equation (46) above. It is easy to see why this formula is correct. Take the original regression

$$\tilde{y} = \tilde{X}_1\beta_1 + \tilde{X}_2\beta_2 + \tilde{\epsilon} \tag{51}$$

and project both sides on $\tilde{X}_1$. This gives us

$$P(\tilde{y}|\tilde{X}_1) = \tilde{X}_1\beta_1 + P(\tilde{X}_2|\tilde{X}_1)\beta_2, \tag{52}$$

since $P(\tilde{\epsilon}|\tilde{X}_1) = 0$ due to the orthogonality condition $E\{\tilde{X}_1'\tilde{\epsilon}\} = 0$. Subtracting equation (52) from the regression equation (51), we get

$$\tilde{y} - P(\tilde{y}|\tilde{X}_1) = [\tilde{X}_2 - P(\tilde{X}_2|\tilde{X}_1)]\beta_2 + \tilde{\epsilon} = \tilde{\xi}\beta_2 + \tilde{\epsilon}. \tag{53}$$

This is a valid regression since $\tilde{\epsilon}$ is orthogonal to $\tilde{X}_2$ and to $\tilde{X}_1$ and hence it must be orthogonal to the linear combination $[\tilde{X}_2 - P(\tilde{X}_2|\tilde{X}_1)]$.

7. Errors in Variables

Endogeneity problems can also arise when there are *errors in variables.* Consider the regression model

$$y^* = x^*\beta + \epsilon \tag{54}$$

where $Ex^*\epsilon = 0$ and the stars denote the true values of the underlying variables. Suppose that we do not observe (y^*, x^*) but instead we observe noisy versions of these variables given by:

$$\begin{aligned} y &= y^* + v \\ x &= x^* + u, \end{aligned} \tag{55}$$

where $E\{v\} = E\{u\} = 0$, $E\{\epsilon u\} = E\{\epsilon v\} = 0$, and $E\{x^* v\} = E\{x^* u\} = E\{y^* u\} = E\{y^* v\} = E\{vu\} = 0$. That is, we assume that the measurement error is unbiased and uncorrelated with the disturbances ϵ in the regression equation, and the measurement errors in y^* and x^* are uncorrelated. Now the regression we actually do is based on the noisy observed values (y, x) instead of the underlying true values (y^*, x^*). Substituting for $y^* x$ and x^* in the regression equation (54), we obtain:

$$y = x\beta + \epsilon - \beta u + v. \tag{56}$$

Now observe that the mismeasured regression equation (56) has a composite error term $\eta = \epsilon - \beta u + v$ that is not orthogonal to the mismeasured independent variable x. To see this, note that the above assumptions imply that

$$E\{x\eta\} = E\{x(\epsilon - \beta u + v\} = -\beta\sigma_u^2. \tag{57}$$

This negative covariance between x and ϵ implies that the OLS estimator of β is asymptotically downward biased when there are errors in variables in the independent variable x^*. Indeed we have:

$$\hat{\beta} = \frac{\frac{1}{N}\sum_{i=1}^N (x_i^* + u_i)(\beta x_i^* + \epsilon_i)}{\frac{1}{N}\sum_{i=1}^N (x_i^* + u_i)^2} \longrightarrow \frac{\beta E\{x^{*2}\}}{[E\{x^{*2}\} + \sigma_u^2]} < \beta. \tag{58}$$

Now consider the possibility of identifying β by the method of moments. We can consistently estimate the three moments σ_y^2, σ_x^2 and σ_{xy}^2 using the observed noisy measures (y, x). However we have

$$\begin{aligned} \sigma_y^2 &= \beta^2 E\{x^{*2}\} + \sigma_\epsilon^2 \\ \sigma_x^2 &= E\{x^{*2}\} + \sigma_u^2 \\ \sigma_{xy}^2 &= \text{cov}(x, y) = \beta E\{x^{*2}\}. \end{aligned} \tag{59}$$

Unfortunately we have 3 equations in 4 unknowns, $(\beta, \sigma_\epsilon^2, \sigma_u^2, E\{x^{*2}\})$. If we try to use higher moments of (y, x) to identify β, we find that we always have more unknowns that equations.

8. Simultaneous Equations Bias

Consider the simple supply/demand example from chapter 16 of Greene. We have:

$$\begin{aligned} q_d &= \alpha_1 p + \alpha_2 y + \epsilon_d \\ q_s &= \beta_1 p + \qquad \epsilon_s \\ q_d &= q_s \end{aligned} \tag{60}$$

where y denotes income, p denotes price, and we assume that $E\{\epsilon_d\} = E\{\epsilon_s\} = E\{\epsilon_d \epsilon_s\} = E\{\epsilon_d y\} = E\{\epsilon_s y\} = 0$. Solving $q_d = q_s$ we can write the *reduced-form* which expresses the *endogenous variables* (p, q) in terms of the *exogenous variable* y:

$$\begin{aligned} p &= \frac{\alpha_2 y}{\beta_1 - \alpha_1} + \frac{\epsilon_d - \epsilon_s}{\beta_1 - \alpha_1} = \pi_1 y + v_1 \\ q &= \frac{\beta_1 \alpha_2 y}{\beta_1 - \alpha_1} + \frac{\beta_1 \epsilon_d - \alpha_1 \epsilon_s}{\beta_1 - \alpha_1} = \pi_2 y + v_2. \end{aligned} \tag{61}$$

By the assumption that y is exogenous in the structural equations (60), it follows that the two linear equations in the reduced form, (61), are valid regression equations; i.e. $E\{yv_1\} = E\{yv_2\} = 0$. However p is not an exogenous regressor in either the supply or demand equations in (60) since

$$\begin{aligned} \text{cov}(p, \epsilon_d) &= \frac{\sigma^2_{\epsilon_d}}{\beta_1 - \alpha_1} > 0 \\ \text{cov}(p, \epsilon_s) &= \frac{-\sigma^2_{\epsilon_s}}{\beta_1 - \alpha_1} < 0. \end{aligned} \tag{62}$$

Thus, the endogeneity of p means that OLS estimation of the demand equation (i.e. a regression of q on p and y) will result in an overestimated (upward biased) price coefficient. We would expect that OLS estimation of the supply equation (i.e. a regression of q on p only) will result in an underestimated (downward biased) price coefficient, however it is not possible to sign the bias in general.

Exercise: Show that the OLS estimate of α_1 converges to

$$\hat{\alpha}_1 \longrightarrow \omega\alpha_1 + (1-\omega)\beta_1, \tag{63}$$

where

$$\omega = \frac{\sigma^2_{\epsilon_s}}{\sigma^2_{\epsilon_s} + \sigma^2_{\epsilon_d}}. \tag{64}$$

Since $\beta_1 > 0$, it follows from the above result that OLS estimator is upward biased. It is possible that when ω is sufficiently small and β_1 is sufficiently large that the OLS estimate will converge to a positive value, i.e. it would lead us to incorrectly infer that the demand equation slopes upwards (Giffen good?) instead of down.

Exercise: Derive the probability limit for the OLS estimator of β_1 in the supply equation (i.e. a regression of q on p only). Show by example that this probability limit can be either higher or lower than β_1.

Exercise: Show that we can identify β_1 from the reduced-form coefficients (π_1, π_2). Which other structural coefficients $(\alpha_1, \alpha_2, \sigma^2_{\epsilon_d}, \sigma^2_{\epsilon_s})$ are identified?

9. Instrumental Variables We have provided three examples where we are interested in estimating the coefficients of a linear "structural" model, but where OLS estimates will produce misleading estimates due to a failure of the orthogonality condition $E\{\tilde{X}'\epsilon\} = 0$ in the linear structural relationship

$$\tilde{y} = \tilde{X}\beta_0 + \tilde{\epsilon}, \tag{65}$$

where β_0 is the "true" vector of structural coefficients. If $\tilde{X}$ is endogenous, then $E\{\tilde{X}'\tilde{\epsilon}\} \neq 0$, then $\beta_0 \neq \beta^* = [E\{\tilde{X}'\tilde{X}\}]^{-1}E\{\tilde{X}'\tilde{y}\}$, and the OLS estimator of the structural coefficients β_0 in equation (65) will be inconsistent. Is it possible to consistently estimate β_0 when $\tilde{X}$ is endogenous? In this section we will show that the answer is yes provided we have access to a sufficient number of *instrumental variables*.

Definition: Given a linear structural relationship (65), we say the $1 \times K$ vector of regressors $\tilde{X}$ is *endogenous* if $E\{\tilde{X}'\tilde{\epsilon}\} \neq 0$, where $\tilde{\epsilon} = \tilde{y} - \tilde{X}\beta_0$, and β_0 is the "true" structural coefficient vector.

Now suppose we have access to a $J \times 1$ vector of *instruments,* i.e. a random vector $\tilde{Z}$ satisfying:

$$\begin{array}{ll} A1) & E\{\tilde{Z}'\tilde{\epsilon}\} = 0 \\ A2) & E\{\tilde{Z}'\tilde{X}\} \neq 0. \end{array} \tag{66}$$

9.1 The exactly indentified case and the simple IV estimator. Consider first the *exactly identified* case where $J = K$, i.e. we have just as many instruments as endogenous regressors in the structural equation (65). Multiply both sides of the structural equation (65) by $\tilde{Z}'$ and take expectations. Using A2) we obtain:

$$\begin{aligned} \tilde{Z}'\tilde{y} &= \tilde{Z}'\tilde{X}\beta_0 + \tilde{Z}'\tilde{\epsilon} \\ E\{\tilde{Z}'\tilde{y}\} &= E\{\tilde{Z}'\tilde{X}\}\beta_0 + E\{\tilde{Z}'\tilde{\epsilon}\} \\ &= E\{\tilde{Z}'\tilde{X}\}\beta_0. \end{aligned} \tag{67}$$

If we assume that the $K \times K$ matrix $E\{\tilde{Z}'\tilde{X}\}$ is invertible, we can solve the above equation for the $K \times 1$ vector β_{SIV}:

$$\beta_{SIV} \equiv [E\{\tilde{Z}'\tilde{X}\}]^{-1} E\{\tilde{Z}'\tilde{y}\}. \tag{68}$$

However plugging in $E\{\tilde{Z}'\tilde{y}\}$ from equation (67) we obtain:

$$\beta_{SIV} = [E\{\tilde{Z}'\tilde{X}\}]^{-1} E\{\tilde{Z}'\tilde{y}\} = [E\{\tilde{Z}'\tilde{X}\}]^{-1} E\{\tilde{Z}'\tilde{X}\}\beta_0 = \beta_0. \tag{69}$$

The fact that $\beta_{SIV} = \beta_0$ motivates the definition of the *simple IV estimator* $\hat{\beta}_{SIV}$ as the sample analog of β_{SIV} in equation (68). Thus, suppose we have a random sample consisting of N *IID* observations of the random vectors $\{\tilde{y}, \tilde{X}, \tilde{Z}\}$, i.e. our data set consists of $\{(y_1, X_1, Z_1), \ldots, (y_N, X_N, Z_N)\}$ which can be represented in matrix form by the $N \times 1$ vector y, and the $N \times K$ matrices Z and X.

Definition: Assume that the $K \times K$ matrix $Z'X$ exists. Then the *simple IV estimator* $\hat{\beta}_{SIV}$ is the sample analog of β_{SIV} given by:

$$\hat{\beta}_{SIV} \equiv [Z'X]^{-1} Z'y = \left[\frac{1}{N}\sum_{i-1}^{N} Z_i'X_i\right]^{-1} \left[\frac{1}{N}\sum_{i=1}^{N} Z_i'y_i\right]. \tag{70}$$

Similar to the OLS estimator, we can appeal to the SLLN and Slutsky's Theorem to show that with probability 1 we have:

$$\hat{\beta}_{SIV} \left[\frac{1}{N}\sum_{i-1}^{N} Z_i'X_i\right]^{-1} \left[\frac{1}{N}\sum_{i=1}^{N} Z_i'y_i\right] \Longrightarrow [E\{\tilde{Z}'\tilde{X}\}]^{-1} E\{\tilde{Z}'\tilde{y}\} = \beta_0. \tag{71}$$

We can appeal to the CLT to show that

$$\sqrt{N}[\hat{\beta}_{SIV} - \beta_0] = \left[\frac{1}{N}\sum_{i-1}^{N} Z_i'X_i\right]^{-1} \left[\frac{1}{\sqrt{N}}\sum_{i=1}^{N} Z_i'\epsilon_i\right] \underset{d}{\Longrightarrow} N(0, \Omega), \tag{72}$$

where

$$\Omega = [E\{\tilde{Z}'\tilde{X}\}]^{-1}E\{\tilde{\epsilon}^2\tilde{Z}'\tilde{Z}\}[E\{\tilde{X}'\tilde{Z}\}]^{-1}, \tag{73}$$

where we use the result that $[A^{-1}]' = [A']^{-1}$ for any invertible matrix A. The covariance matrix Ω can be consistently estimated by its sample analog:

$$\hat{\Omega} = \left[\frac{1}{N}\sum_{i=1}^{N} Z_i'X_i\right]^{-1}\left[\frac{1}{N}\sum_{i=1}^{N}\hat{\epsilon}_i^2 Z_i'Z_i\right]\left[\frac{1}{N}\sum_{i=1}^{N} X_i'Z_i\right]^{-1} \tag{74}$$

where $\hat{\epsilon}_i^2 = (y_i - X_i\hat{\beta}_{SIV})^2$. We can show that the estimator (74) is consistent using the same argument we used to establish the consistency of the heteroscedasticity-consistent covariance matrix estimator (34) in the OLS case. Finally, consider the form of Ω in the homoscedastic case.

Definition: We say the error terms ϵ in the structural model in equation (65) are *homoscedastic* if there exists a nonnegative constant σ^2 for which:

$$E\{\tilde{\epsilon}^2\tilde{Z}'\tilde{Z}\} = \sigma^2 E\{\tilde{Z}'\tilde{Z}\} \tag{75}$$

A sufficient condition for homoscedasticity to hold is $E\{\tilde{\epsilon}|\tilde{Z}\} = 0$ and $\text{var}\{\tilde{\epsilon}|\tilde{Z}\} = \sigma^2$. Under homoscedasticity the asymptotic covariance matrix for the simple IV estimator becomes:

$$\Omega = \sigma^2[E\{\tilde{Z}'\tilde{X}\}]^{-1}E\{\tilde{Z}'\tilde{Z}\}[E\{\tilde{X}'\tilde{Z}\}]^{-1}, \tag{76}$$

and if the above two sufficient conditions hold, it can be consistently estimated by its sample analog:

$$\hat{\Omega} = \hat{\sigma}^2\left[\frac{1}{N}\sum_{i=1}^{N} Z_i'X_i\right]^{-1}\left[\frac{1}{N}\sum_{i=1}^{N} Z_i'Z_i\right]\left[\frac{1}{N}\sum_{i=1}^{N} X_i'Z_i\right]^{-1} \tag{77}$$

where $\hat{\sigma}^2$ is consistently estimated by:

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - X_i\hat{\beta}_{SIV})^2. \tag{78}$$

As in the case of OLS, we recommend using the heteroscedasticity consistent covariance matrix estimator (74) which will be consistent regardless of whether the true model (65) is homoscedastic or heteroscedastic rather than the estimator (78) which will be inconsistent if the true model is heteroscedastic.

9.2 The overidentified case and two stage least squares. Now consider the *overidentified case*, i.e. when we have more instruments than endogenous regressors, i.e. when $J > K$. Then the matrix $E\{\tilde{Z}'\tilde{X}\}$ is not square, and the simple IV estimator β_{SIV} is not defined. However we can always choose a subset $\tilde{W}$ consisting of a $1 \times K$ subvector of the $1 \times J$ random vector Z so that $E\{\tilde{W}'\tilde{X}\}$ is square and invertible. More generally we could construct instruments by taking *linear combinations* of the full list of instrumental variables $\tilde{Z}$, where γ is a $J \times K$ matrix.

$$\tilde{W} = \tilde{Z}\gamma. \tag{79}$$

Example 1. Suppose we want our instrument vector $\tilde{W}$ to consist of the first K components of $\tilde{Z}$. Then we set $\gamma = (I|0)'$ where the I is a $K \times K$ identity matrix and 0 is a $K \times J$ matrix of zeros, and $|$ denotes the horizontal concatenation operator.

Example 2. Consider the instruments given by $\tilde{W}^* = \tilde{Z}\gamma^*$ where $\gamma^* = [E\{\tilde{Z}'\tilde{Z}\}]^{-1}E\{\tilde{Z}'\tilde{X}\}$. It is straightforward to verify that this is a $J \times K$ matrix. We can interpret γ^* as the matrix of regression coefficents from regressing $\tilde{X}$ on $\tilde{Z}$. Thus $\tilde{W}^* = P(\tilde{X}|\tilde{Z}) = \tilde{Z}\gamma^*$ is the projection of the endogenous variables $\tilde{X}^*$ onto the instruments $\tilde{Z}$. Since $\tilde{X}$ is a vector of random variables, γ actually represents the horizontal concatenation of K separate $J \times 1$ regression coefficient vectors. We can write all the regressions compactly in vector form as

$$\tilde{X} = P(\tilde{X}|\tilde{Z}) + \tilde{\eta} = \tilde{Z}\gamma^* + \tilde{\eta}, \tag{80}$$

where $\tilde{\eta}$ is $1 \times K$ vector of error terms for each of the K regression equations. Thus, by definition of least squares, each component of $\tilde{\eta}$ must be orthogonal to the regressors $\tilde{Z}$, i.e.

$$E\{\tilde{Z}'\tilde{\eta}\} = \mathbf{0}, \tag{81}$$

where $\mathbf{0}$ is a $J \times K$ matrix of zeros. We will shortly formalize the sense in which $\tilde{W}^* = \tilde{Z}\gamma^*$ are the "optimal instruments" within the class of instruments formed from linear combinations of $\tilde{Z}$ in equation (79). Intuitively, the optimal instruments should be the best linear predictors of the endogenous regressors $\tilde{X}$, and clearly, the instruments $\tilde{W}^* = \tilde{Z}\gamma^*$ from the *first stage regression* (80) are the best linear predictors of the endogenous $\tilde{X}$ variables.

Definition: Assume that $[E\{\tilde{W}'\tilde{X}\}]^{-1}$ exists where $\tilde{W} = \tilde{Z}\gamma$. Then we define β_{IV} by

$$\beta_{IV} = [E\{\tilde{W}'\tilde{X}\}]^{-1}E\{\tilde{W}'\tilde{y}\} = [E\{\gamma'\tilde{Z}'\tilde{X}\}]^{-1}E\{\gamma'\tilde{Z}'\tilde{y}\}. \tag{82}$$

Definition: Assume that $[E\{\tilde{W}^{*\prime}\tilde{W}^*\}]^{-1}$ exists where $\tilde{W}^* = \tilde{Z}\gamma^*$ and $\gamma^* = [E\{\tilde{Z}'\tilde{Z}\}]^{-1}E\{\tilde{Z}'\tilde{y}\}$. Then we define β_{2SLS} by

$$\begin{aligned}
\beta_{2SLS} &= [E\{\tilde{W}^{*\prime}\tilde{X}\}]^{-1}E\{\tilde{W}^{*\prime}\tilde{y}\} \\
&= \left[E\{\tilde{X}'\tilde{Z}\}[E\{\tilde{Z}'\tilde{Z}\}]^{-1}E\{\tilde{Z}'\tilde{X}\}\right]^{-1} E\{\tilde{X}'\tilde{Z}\}[E\{\tilde{Z}'\tilde{Z}\}]^{-1}E\{\tilde{Z}'\tilde{y}\} \\
&= \left[E\{P(\tilde{X}|\tilde{Z})'P(\tilde{X}|\tilde{Z})\}\right]^{-1} E\{P(\tilde{X}|\tilde{Z})'\tilde{y}\}.
\end{aligned} \tag{83}$$

Clearly β_{2SLS} is a special case of β_{IV} when $\gamma = \gamma^* = [E\{\tilde{Z}'\tilde{Z}\}]^{-1}E\{\tilde{Z}'\tilde{X}\}$. We refer to it as *two stage least squares* since β_{2SLS} can be computed in two stages:

Stage 1: Regress the endogenous variables $\tilde{X}$ on the instruments $\tilde{Z}$ to get the linear projections $P(\tilde{X}|\tilde{Z}) = \tilde{Z}\gamma^*$ as in equation (80).

Stage 2: Regress $\tilde{y}$ on $P(\tilde{X}|\tilde{Z})$ instead of on $\tilde{Z}$ as shown in equation (83). The projections $P(\tilde{X}|\tilde{Z})$ essentially "strip off" the endogenous components $\tilde{\eta}$ of $\tilde{X}$, resulting in a valid regression equation for β_0.

We can get some more intuition into the latter statement by rewriting the original structural equation (65) as:

$$\begin{aligned}
\tilde{y} &= \tilde{X}\beta_0 + \tilde{\epsilon} \\
&= P(\tilde{X}|\tilde{Z})\beta_0 + [\tilde{X} - P(\tilde{X}|\tilde{Z})]\beta_0 + \tilde{\epsilon} \\
&= P(\tilde{X}|\tilde{Z})\beta_0 + \tilde{\eta}\beta_0 + \tilde{\epsilon} \\
&= P(\tilde{X}|\tilde{Z})\beta_0 + \tilde{v},
\end{aligned} \tag{84}$$

where $\tilde{v} = \tilde{\eta}\beta_0 + \tilde{\epsilon}$. Notice that $E\{\tilde{Z}'\tilde{v}\} = E\{\tilde{Z}'(\eta\beta_0 + \tilde{\epsilon})\} = 0$ as a consequence of equations (66) and (81). It follows from the projection theorem that equation (84) is a valid regression, i.e. that $\beta_{2SLS} = \beta_0$. Alternatively, we can simply use the same straightforward reasoning as we did for β_{SIV}, substituting equation (65) for $\tilde{y}$ and simplifying equations (82) and (83) to see that $\beta_{IV} = \beta_{2SLS} = \beta_0$. This motivates the definitions of $\hat{\beta}_{IV}$ and $\hat{\beta}_{2SLS}$ as the sample analogs of β_{IV} and β_{2SLS}:

Definition: Assume $W = Z\gamma$ where Z is $N \times J$ and γ is $J \times K$, and $W'X$ is invertible (this implies that $J \geq K$). Then the *instrumental variables estimator* $\hat{\beta}_{IV}$ is the sample analog of β_{IV} defined in equation (82):

$$\begin{aligned}\hat{\beta}_{IV} &\equiv [W'X]^{-1}W'y = \left[\frac{1}{N}\sum_{i-1}^{N} W_i'X_i\right]^{-1}\left[\frac{1}{N}\sum_{i=1}^{N} W_i'y_i\right] \\ &= [\gamma'Z'X]^{-1}\gamma'Z'y = \left[\frac{1}{N}\sum_{i-1}^{N} \gamma'Z_i'X_i\right]^{-1}\left[\frac{1}{N}\sum_{i=1}^{N} \gamma'Z_i'y_i\right]. \end{aligned} \tag{85}$$

Definition: Assume that the $J \times J$ matrix $Z'Z$ and the $K \times K$ matrix $W'W$ are invertible, where $W = Z\hat{\gamma}$ and $\hat{\gamma} = [Z'Z]^{-1}Z'X$. The *two-stage least squares estimator* $\hat{\beta}_{2SLS}$ is the sample analog of β_{2SLS} defined in equation (83):

$$\begin{aligned}\hat{\beta}_{2SLS} &\equiv [W'X]^{-1}W'y \\ &= \left[X'Z[Z'Z]^{-1}Z'X]^{-1}\right]^{-1} X'Z[Z'Z]^{-1}Z'y \\ &= \left[P(X|Z)'P(X|Z)\right]^{-1} P(X|Z)'y \\ &= \left[X'P_Z X\right]^{-1} X'P_Z y, \end{aligned} \tag{86}$$

where P_Z is the $N \times N$ projection matrix

$$P_Z = Z[Z'Z]^{-1}Z'. \tag{87}$$

Using exactly the same arguments that we used to prove the consistency and asymptotic normality of the simple IV estimator, it is straightforward to show that $\hat{\beta}_{IV} \underset{p}{\rightarrow} \beta_0$ and $\sqrt{N}[\hat{\beta}_{IV} - \beta_0] \underset{d}{\Longrightarrow} N(0, \Omega)$, where Ω is the $K \times K$ matrix given by:

$$\Omega = [E\{\tilde{X}'\tilde{W}\}]^{-1}E\{\tilde{\epsilon}^2\tilde{W}'\tilde{W}\}[E\{\tilde{W}'\tilde{X}\}]^{-1} = [E\{\tilde{X}'\tilde{Z}\gamma\}]^{-1}E\{\tilde{\epsilon}^2\gamma'\tilde{Z}'\tilde{Z}\gamma\}[E\{\gamma'\tilde{Z}'\tilde{X}\}]^{-1}. \tag{88}$$

Now we have a whole family of IV estimators depending on how we choose the $J \times K$ matrix γ. What is the optimal choice for γ? As we suggested earlier, the optimal choice should be $\gamma^* = [E\{\tilde{Z}'\tilde{Z}\}]^{-1}E\{\tilde{Z}'\tilde{X}\}$ since this results in a linear combination of instruments $\tilde{W}^* = \tilde{Z}\gamma^*$ that is the best linear predictor of the endogenous regressors $\tilde{X}$.

Theorem: *Assume that the error term $\tilde{\epsilon}$ in the structural model (65) is homoscedastic. Then the optimal IV estimator is 2SLS, i.e. it has the smallest asymptotic covariance matrix among all IV estimators.*

Proof: Under homoscedasticity, the asymptotic covariance matrix for the IV estimator is equal to

$$\Omega = \sigma^2[E\{\tilde{X}'\tilde{W}\}]^{-1}E\{\tilde{W}'\tilde{W}\}[E\{\tilde{W}'\tilde{X}\}]^{-1} = \sigma^2[E\{\tilde{X}Z\gamma\}]^{-1}E\{\gamma'\tilde{Z}'\tilde{Z}\gamma\}[E\{\gamma'\tilde{Z}'\tilde{X}\}]^{-1}. \quad (89)$$

We now show this covariance matrix is minimized when $\gamma = \gamma^*$, i.e. we show that

$$\Omega \geq \Omega^* \quad (90)$$

where Ω^* is the asymptotic covariance matrix for 2SLS which is obtained by substituting $\gamma^* = [E\{\tilde{Z}'\tilde{Z}\}]^{-1}E\{\tilde{Z}'\tilde{X}\}$ into the formula above. Since $A \geq B$ if and only $A^{-1} \leq B^{-1}$, it is sufficient to show that $\Omega^{-1} \leq \Omega^{*-1}$, or

$$E\{\tilde{W}'\tilde{X}\}[E\{\tilde{W}'\tilde{W}\}]^{-1}E\{\tilde{X}'\tilde{W}\} \leq E\{\tilde{Z}'\tilde{X}\}[E\{\tilde{Z}'\tilde{Z}\}]^{-1}E\{\tilde{X}'\tilde{Z}\}. \quad (91)$$

Note that $\Omega^{-1} = E\{P(\tilde{X}|\tilde{W})'P(\tilde{X}|\tilde{W})\}$ and $\Omega^{*-1} = E\{P(\tilde{X}|\tilde{Z})'P(\tilde{X}|\tilde{Z})\}$, so our task reduces to showing that

$$E\{P(\tilde{X}|\tilde{W})'P(\tilde{X}|\tilde{W})\} \leq E\{P(\tilde{X}|\tilde{Z})'P(\tilde{X}|\tilde{Z})\}. \quad (92)$$

However since $\tilde{W} = \tilde{Z}\gamma$ fopr some $J \times K$ matrix γ, it follows that the elements of $\tilde{W}$ must span a subspace of the linear subspace spanned by the elements of $\tilde{Z}$. Then the law of Iterated Projections implies that

$$P(\tilde{X}|\tilde{W}) = P(P(\tilde{X}|\tilde{Z})|\tilde{W}). \quad (93)$$

This implies that there exists a $1 \times K$ vector of error terms $\tilde{\xi}$ satisfying

$$P(\tilde{X}|\tilde{Z}) = P(\tilde{X}|\tilde{W}) + \tilde{\xi}, \quad (94)$$

where $\tilde{\xi}$ satisfy the orthogonality relation

$$E\{\tilde{P}(\tilde{X}|\tilde{W})'\tilde{\xi}\} = \mathbf{0}, \quad (95)$$

where $\mathbf{0}$ is an $K \times K$ matrix of zeros. Then using the identity (94) we have

$$\begin{aligned} E\{P(\tilde{X}|\tilde{Z})'P(\tilde{X}|\tilde{Z})\} &= E\{P(\tilde{X}|\tilde{W})'P(\tilde{X}|\tilde{W})\} + E\{\tilde{\xi}'P(\tilde{X}|\tilde{W})\} + E\{P(\tilde{X}|\tilde{W})'\tilde{\xi}\} + E\{\tilde{\xi}'\tilde{\xi}\} \\ &= E\{P(\tilde{X}|\tilde{W})'P(\tilde{X}|\tilde{W})\} + E\{\tilde{\xi}'\tilde{\xi}\} \\ &\geq E\{P(\tilde{X}|\tilde{W})'P(\tilde{X}|\tilde{W})\}. \end{aligned} \quad (96)$$

We conclude that $\Omega^{-1} \leq \Omega^{*-1}$, and hence $\Omega \geq \Omega^*$, i.e. 2SLS has the smallest asymptotic covariance matrix among all IV estimators. •

There is an alternative algebraic proof that $\Omega^{*-1} \geq \Omega^{-1}$. Given a square symmetric positive semidefinite matrix A with Jordan decomposition $A = WDW'$ (where W is an orthonormal matrix and D is a diagonal matrix with diagonal elements equal to the eigenvalues of A) we can define its *square root* $A^{1/2}$ as

$$A^{1/2} = WD^{1/2}W' \quad (97)$$

where $D^{1/2}$ is a diagonal matrix whose diagonal elements equal the square roots of the diagonal elements of D. It is easy to verify that $A^{1/2}A^{1/2} = A$. Similarly if A is invertible we define $A^{-1/2}$ as the matrix $WD^{-1/2}W'$ where $D^{-1/2}$ is a diagonal matrix whos diagonal elements are the inverses of the square roots of the diagonal element of D. It is easy to verify that $A^{-1/2}A^{-1/2} = A^{-1}$. Using these facts about matrix square roots, we can write

$$\Omega^{*-1} - \Omega^{-1} = E\{\tilde{X}'\tilde{Z}\}[E\{\tilde{Z}'\tilde{Z}\}]^{-1/2}M[E\{\tilde{Z}'\tilde{Z}\}]^{-1/2}E\{\tilde{Z}'\tilde{X}\}, \tag{98}$$

where M is the $K \times K$ matrix given by

$$M = \left[I - [E\{\tilde{Z}'\tilde{Z}\}]^{1/2}\gamma[\gamma'[E\{\tilde{Z}'\tilde{Z}\}]^{-1}\gamma]^{-1}\gamma'[E\{\tilde{Z}'\tilde{Z}\}]^{1/2}\right]. \tag{99}$$

It is straightforward to verify that M is idempotent, which implies that the right hand side of equation (98) is positive semidefinite. •

It follows that in terms of the *asymptotics* it is always better to use all available instruments $\tilde{Z}$. However the chapter in Davidson and MacKinnon shows that in terms of the finite sample performance of the IV estimator, using more instruments may not always be a good thing. It is easy to see that when the number of instruments J gets sufficient large, the IV estimator converges to the OLS estimator.

Exercise: Show that when $J = N$ and the columns of Z are linearly independent that $\hat{\beta}_{2SLS} = \hat{\beta}_{OLS}$.

Exercise: Show that when $J = K$ and the columns of Z are linearly independent that $\hat{\beta}_{2SLS} = \hat{\beta}_{SIV}$.

However there is a tension here, since using fewer instruments worsens the finite sample properties of the 2SLS estimator. A result due to Kinal *Econometrica* 1980 shows that the M^{th} moment of the 2SLS estimator exists if and only if

$$M < J - K + 1. \tag{100}$$

Thus, if $J = K$ 2SLS (which coincides with the SIV estimator by the exercise above) will not even have a finite mean. If we would like the 2SLS estimator to have a finite mean and variance we should have at least 2 more instruments than endogenous regressors. See section 7.5 of Davidson and MacKinnon for further discussion and monte carlo evidence.

Exercise: Assume that the errors are homoscedastic. Is it the case that *in finite samples* that the 2SLS estimator dominates the IV estimator in terms of the size of its estimated covariance matrix?

Hint: Note that under homoscedasticity, the inverse of the sample analog estimators of the covariance matrix for $\hat{\beta}_{IV}$ and $\hat{\beta}_{2SLS}$ are is given by:

$$\begin{aligned} [\hat{\Omega}_{IV}]^{-1} &= \frac{1}{\hat{\sigma}^2_{IV}}[X'P_W X], \\ [\hat{\Omega}_{2SLS}]^{-1} &= \frac{1}{\hat{\sigma}^2_{2SLS}}[X'P_Z X]. \end{aligned} \tag{101}$$

If we assume that $\hat{\sigma}^2_{IV} = \hat{\sigma}^2_{2SLS}$, then the relative finite sample covariance matrices for IV and 2SLS depend on the difference

$$X'P_W X - X'P_Z X = X'(P_W - P_Z)X. \tag{102}$$

Show that if $W = Z\gamma$ for some $J \times K$ matrix γ that $P_W P_Z = P_Z P_W = P_W$ and this implies that the difference $(P_W - P_Z)$ is idempotent.

Now consider a structural equation of the form

$$\tilde{y} = \tilde{X}_1\beta_1 + \tilde{X}_2\beta_2 + \tilde{\epsilon}, \tag{103}$$

where the $1 \times K_1$ random vector $\tilde{X}_1$ is known to be exogenous (i.e. $E\{\tilde{X}_1'\tilde{\epsilon}\} = 0$), but the $1 \times K_2$ random vector $\tilde{X}_2$ is suspected of being endogenous. It follows that $\tilde{X}_1$ can serve as instrumental variables for the $\tilde{X}_2$ variables.

Exercise: Is it possible to identify the β_2 coefficients using only $\tilde{X}_1$ as instrumental variables? If not, show why.

The answer to the exercise is clearly no: for example 2SLS based on $\tilde{X}_1$ alone will result in a first stage regression $P(\tilde{X}_2|\tilde{X}_1)$ that is a linear function of $\tilde{X}_1$, so the second stage of 2SLS would encounter multicollinearity. This shows that in order to identify β_2 we need additional instruments W that are *excluded* from the structural equation (103). This results in a full instrument list $\tilde{Z} = (\tilde{W}, \tilde{X}_1)$ of size $J = (J - K_1) + K_1$. The discussion above suggest that in order to identity β_2 we need $J \geq K_2$ *and* $J > K_1$, otherwise we have a multicollinearity problem in the second stage. In summary, to do instrumental variables we need instruments Z which are:

1. Uncorrelated with the error term $\tilde{\epsilon}$ in the structural equation (103),
2. Correlated with the included endogenous variables $\tilde{X}_2$,
3. Contain components $\tilde{W}$ which are *excluded* from the structural equation (103).