Gregory Kordas
Last update: June 6, 2023

LECTURE 7

Weighted Least Squares

1. Introduction.

Consider again the classical linear regression model given by

$$\boldsymbol{y}_{n\times 1} = X_{n\times k}\boldsymbol{\beta}_{k\times 1} + \boldsymbol{u}_{n\times 1}.$$

We have seen that under the following assumptions:

(i) Exogeneity of the Regressors: $E(\boldsymbol{u}|X) = \mathbf{0}$.
(ii) Spherical Errors: $E(\boldsymbol{u}\boldsymbol{u}') = \sigma_u^2 I_n$.
(iii) Full Rank: $X'X$ is a symmetric positive definite matrix.
(iv) Normal Errors: $\boldsymbol{u} \sim N(\mathbf{0}, \sigma_u^2 I_n)$.

the OLS coefficients

$$\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\boldsymbol{y},$$

are normal, $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma_u^2(X'X)^{-1})$, and efficient (BLUE).

In what follows we will discuss various violations of the classical assumptions that will make OLS inefficient or even inconsistent. We will see that optimal estimators are solutions to 'weighted' least squares problems, so we will refer to them collectively as WLS. The "weight" will, of course, depend in the situation at hand.

2. Eigenvalue Decomposition

Many mathematical objects can be understood better by breaking them into constituent parts, or finding some properties of them that are universal, not caused by the way we choose to represent them. For example, integers can be decomposed into prime factors. The way we represent the number 12 will change depending on whether we write it in base ten or in binary, but it will always be true that $12 = 2 \times 2 \times 3$. From this representation we can conclude useful properties, such as that 12 is not divisible by 5, or that any integer multiple of 12 will be divisible by 3.

Much as we can discover something about the true nature of an integer by decomposing it into prime factors, we can also decompose matrices in ways that show us information about

their functional properties that is not obvious from the representation of the matrix as an array of elements.

One of the most widely used kinds of matrix decomposition is called eigendecomposition, in which we decompose a matrix into a set of eigenvectors and eigenvalues.

An *eigenvector* of a square matrix A is a non-zero vector $\boldsymbol{v}$ such that multiplication by A alters only the scale of $\boldsymbol{v}$:

$$A\boldsymbol{v} = \lambda\boldsymbol{v}.$$

The scalar λ is known as the *eigenvalue* corresponding to this eigenvector. (One can also find a *left eigenvector* such that $\boldsymbol{v}A = \lambda\boldsymbol{v}$, but we are usually concerned with right eigenvectors).

If $\boldsymbol{v}$ is an eigenvector of A, then so is any rescaled vector $s\boldsymbol{v}$ for $s \in \mathbb{R}, s \neq 0$. Moreover, $s\boldsymbol{v}$ still has the same eigenvalue. For this reason, we usually only look for unit-length eigenvectors.

Suppose that a matrix A has n linearly independent eigenvectors, $\{\boldsymbol{v}^{(1)}, ..., \boldsymbol{v}^{(n)}\}$, with corresponding eigenvalues $\{\lambda_1, ..., \lambda_n\}$. We may concatenate all of the eigenvectors to form a matrix V with one eigenvector per column: $V = [\boldsymbol{v}^{(1)}, ..., \boldsymbol{v}^{(n)}]$. Likewise, we can concatenate the eigenvalues to form a vector $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_n)'$. The *eigendecomposition* of A is then given by

$$A = V \operatorname{diag}(\boldsymbol{\lambda}) V^{-1}.$$

We have seen that constructing matrices with specific eigenvalues and eigenvectors allows us to stretch space in desired directions. However, we often want to decompose matrices into their eigenvalues and eigenvectors. Doing so can help us analyze certain properties of the matrix, much as decomposing an integer into its prime factors can help us understand the behavior of that integer.

Not every matrix can be decomposed into eigenvalues and eigenvectors. In some cases, the decomposition exists, but may involve complex rather than real numbers. Fortunately, in this lecture, we usually need to decompose only a specific class of matrices that have a simple decomposition. Specifically, every real symmetric matrix can be decomposed into an expression using only real-valued eigenvectors and eigenvalues:

$$A = Q\Lambda Q',$$

where Q is an orthogonal matrix composed of eigenvectors of A, and Λ is a diagonal matrix. The eigenvalue $\Lambda_{i,i}$ is associated with the eigenvector in column i of Q, denoted as $Q_{:,i}$. Because Q is an orthogonal matrix, we can think of A as scaling space by λ_i in direction $\boldsymbol{v}^{(i)}$. See Figure 1 for an example.

While any real symmetric matrix A is guaranteed to have an eigendecomposition, the eigendecomposition may not be unique. If any two or more eigenvectors share the same eigenvalue,

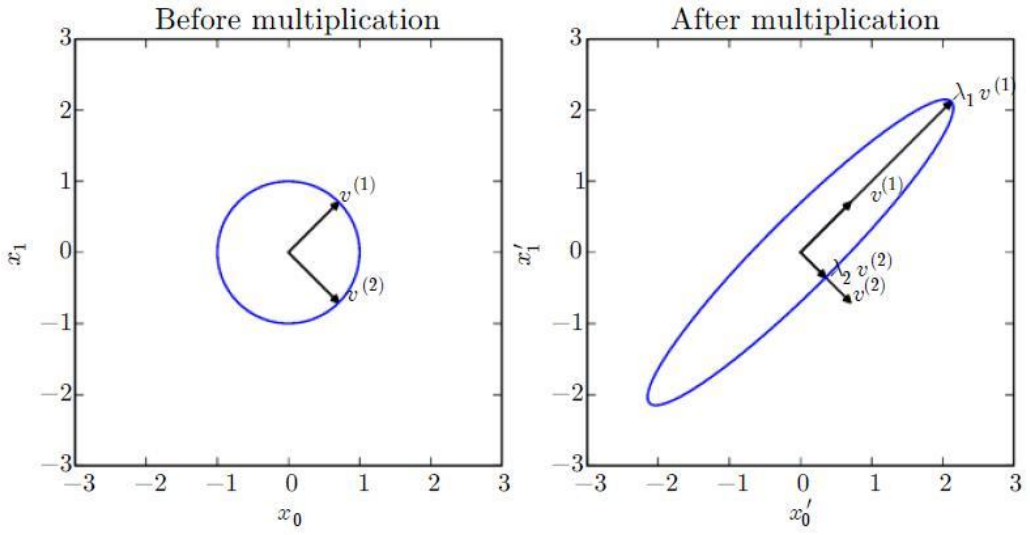


FIGURE 1. An example of the effect of eigenvectors and eigenvalues. Here, we have a matrix A with two orthonormal eigenvectors, $\boldsymbol{v}^{(1)}$ with eigenvalue λ_1 and $\boldsymbol{v}^{(2)}$ with eigenvalue λ_2. (Left) We plot the set of all unit vectors $\boldsymbol{u} \in \mathbb{R}^2$ as a unit circle. (Right) We plot the set of all points $A\boldsymbol{u}$. By observing the way that A distorts the unit circle, we can see that it scales space in direction $\boldsymbol{v}^{(i)}$ by λ_i.

then any set of orthogonal vectors lying in their span are also eigenvectors with that eigenvalue, and we could equivalently choose a Q using those eigenvectors instead. By convention, we usually sort the entries of Λ in descending order. distinct. The eigendecomposition of a matrix tells us many useful facts about the matrix. The matrix is singular if and only if any of the eigenvalues are zero.

The eigendecomposition of a real symmetric matrix can also be used to optimize quadratic expressions of the form maximize $f(\boldsymbol{x}) = \boldsymbol{x}'A\boldsymbol{x}$ subject to $||\boldsymbol{x}|| = 1$. Whenever $\boldsymbol{x}$ is equal to an eigenvector of A, f takes on the value of the corresponding eigenvalue. The maximum value of f within the constraint region is the maximum eigenvalue and its minimum value within the constraint region is the minimum eigenvalue.

A matrix whose eigenvalues are all positive is called *positive definite*. A matrix whose eigenvalues are all positive or zero is called *positive semidefinite*. Likewise, if all eigenvalues are negative, the matrix is *negative definite*, and if all eigenvalues are negative or zero, it is *negative semidefinite*. Positive semidefinite matrices are interesting because they guarantee that $\forall \boldsymbol{x}$, $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} \geq 0$. Positive definite matrices additionally guarantee that $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0} \Rightarrow \boldsymbol{x} = \boldsymbol{0}$.

3. The Triangular Factorization of a Symmetric Positive Definite Matrix

Any square $n \times n$ positive definite symmetric matrix $\mathbf{\Omega}$ has a unique representation of the form

$$\Omega = ADA',$$

where A is an $(n \times n)$ lower triangular matrix with 1s along the principal diagonal,

$$A = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ a_{21} & 1 & 0 & \cdots & 0 \\ a_{31} & a_{32} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & 1 \end{bmatrix},$$

and D is an $(n \times n)$ diagonal matrix,

$$D = \begin{bmatrix} d_{11} & 0 & 0 & \cdots & 0 \\ 0 & d_{22} & 0 & \cdots & 0 \\ 0 & 0 & d_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & d_{nn} \end{bmatrix},$$

with $d_{ii} > 0$ for all $i = 1, ..., n$. This is known as the *triangular factorization* of Ω.

3.1. Calculating the The Triangular Factorization

To see how the triangular factorization can be calculated, consider

$$\Omega = \begin{bmatrix} \omega_{11} & \omega_{12} & \omega_{13} & \cdots & \omega_{1n} \\ \omega_{21} & \omega_{22} & \omega_{23} & \cdots & \omega_{2n} \\ \omega_{31} & \omega_{32} & \omega_{33} & \cdots & \omega_{3n} \\ \vdots & \vdots & \vdots & & \vdots \\ \omega_{n1} & \omega_{n2} & \omega_{n3} & \cdots & \omega_{nn} \end{bmatrix}.$$

We assume that $\mathbf{\Omega}$ is positive definite, meaning that $\boldsymbol{x}'\mathbf{\Omega}\boldsymbol{x} > 0$ for any nonzero $(n \times 1)$ vector $\boldsymbol{x}$. We also assume that $\mathbf{\Omega}$ is symmetric, so that $\omega_{ij} = \omega_{ji}$ for all $i, j = 1, ..., n$.

The matrix $\mathbf{\Omega}$ can be transformed into a matrix with zero in the $(2, 1)$ position by multiplying the first row of $\mathbf{\Omega}$ by $\omega_{21}\omega_{11}^{-1}$ and substracting the resulting row from the second. A zero can be put in the $(3, 1)$ position by multiplying the first row by $\omega_{31}\omega_{11}^{-1}$ and substracting the resulting

row from the third. We proceed in this fashion down the first column. This set operations can be summarized as premultiplying $\boldsymbol{\Omega}$ by the matrix

$$E_1 = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ -\omega_{21}\omega_{11}^{-1} & 1 & 0 & \cdots & 0 \\ -\omega_{31}\omega_{11}^{-1} & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ -\omega_{n1}\omega_{11}^{-1} & 0 & 0 & \cdots & 1 \end{bmatrix}.$$

This matrix always exists provided that $\omega_{11} \neq 0$. This is ensured in the present case, because ω_{11} is equal to $\boldsymbol{e}_1'\boldsymbol{\Omega}\boldsymbol{e}_1$, where $\boldsymbol{e}_1' = (1, 0, 0, ..., 0)$. Since $\boldsymbol{\Omega}$ is positive definite, $\boldsymbol{e}_1'\boldsymbol{\Omega}\boldsymbol{e}_1$ must be greater than zero.

When $\boldsymbol{\Omega}$ is premultiplied by $\boldsymbol{E}_1$ and postmultiplied by $\boldsymbol{E}_1'$ the result is

$$\boldsymbol{H} = \boldsymbol{E}_1\boldsymbol{\Omega}\boldsymbol{E}_1',$$

where

$$\begin{aligned} H &= \begin{bmatrix} h_{11} & 0 & 0 & \cdots & 0 \\ 0 & h_{22} & h_{23} & \cdots & h_{2n} \\ 0 & h_{32} & h_{33} & \cdots & h_{3n} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & h_{n2} & h_{n3} & \cdots & h_{nn} \end{bmatrix} \\ &= \begin{bmatrix} \omega_{11} & 0 & 0 & \cdots & 0 \\ 0 & \omega_{22} - \omega_{21}\omega_{11}^{-1}\omega_{12} & \omega_{23} - \omega_{21}\omega_{11}^{-1}\omega_{13} & \cdots & \omega_{2n} - \omega_{21}\omega_{11}^{-1}\omega_{1n} \\ 0 & \omega_{32} - \omega_{31}\omega_{11}^{-1}\omega_{12} & \omega_{33} - \omega_{31}\omega_{11}^{-1}\omega_{13} & \cdots & \omega_{3n} - \omega_{31}\omega_{11}^{-1}\omega_{1n} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & \omega_{n2} - \omega_{n1}\omega_{11}^{-1}\omega_{12} & \omega_{n3} - \omega_{n1}\omega_{11}^{-1}\omega_{13} & \cdots & \omega_{nn} - \omega_{n1}\omega_{11}^{-1}\omega_{1n} \end{bmatrix}. \end{aligned}$$

We next proceed in exactly the same way with the second column of $\boldsymbol{H}$. The approach will now be to multiply the second row of $\boldsymbol{H}$ by $h_{32}h_{22}^{-1}$ and subtract the result from the third row. Similarly, we multiply the second row of $\boldsymbol{H}$ by $h_{42}h_{22}^{-1}$ and subtract the result from the fourth row, and so on down through the second column of $\boldsymbol{H}$. These operations can be represented

as premultiplying $\boldsymbol{H}$ by the following matrix

$$\boldsymbol{E}_2 = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & -h_{32}h_{22}^{-1} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & -h_{n2}h_{22}^{-1} & 0 & \cdots & 1 \end{bmatrix}.$$

This matrix always exists provided that $h_{22} \neq 0$. But h_{22} can be calculated as $h_{22} = \boldsymbol{e}_2'\boldsymbol{H}\boldsymbol{e}_2$, $\boldsymbol{e}_2' = (0, 1, 0, ..., 0)$. Moreover, $\boldsymbol{H} = \boldsymbol{E}_1\boldsymbol{\Omega}\boldsymbol{E}_1'$, where $\boldsymbol{\Omega}$ is positive definite. Since $\boldsymbol{E}_1$ is lower triangular, its determinant is the product of terms along the principal diagonal, which are all unity. Thus $\boldsymbol{E}_1$ is nonsingular, meaning that $\boldsymbol{H} = \boldsymbol{E}_1\boldsymbol{\Omega}\boldsymbol{E}_1'$ is positive definite and so $h_{22} = \boldsymbol{e}_2'\boldsymbol{H}\boldsymbol{e}_2$ must be strictly positive. Thus $\boldsymbol{E}_2$ can always be calculated.

If $\boldsymbol{H}$ is premultiplied by $\boldsymbol{E}_2$ and postmultiplied by $\boldsymbol{E}_2'$, the result is

$$\boldsymbol{K} = \boldsymbol{E}_2\boldsymbol{H}\boldsymbol{E}_2',$$

where

$$\boldsymbol{K} = \begin{bmatrix} h_{11} & 0 & 0 & \cdots & 0 \\ 0 & h_{22} & 0 & \cdots & 0 \\ 0 & 0 & h_{33} - h_{32}h_{22}^{-1}h_{23} & \cdots & h_{3n} - h_{32}h_{22}^{-1}h_{2n} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & h_{n3} - h_{n2}h_{22}^{-1}h_{23} & \cdots & h_{nn} - h_{n2}h_{22}^{-1}h_{2n} \end{bmatrix}.$$

Again, since $\boldsymbol{H}$ is positive definite and since $\boldsymbol{E}_2$ is nonsingular, $\boldsymbol{K}$ is positive definite and in particular k_{33} is positive. Proceeding through each of the columns with the same apprach, we see that for any positive definite symmetric matrix $\boldsymbol{\Omega}$ there exist matrices $\boldsymbol{E}_1, \boldsymbol{E}_2, ..., \boldsymbol{E}_{n-1}$ such that

$$\boldsymbol{E}_{n-1} \cdots \boldsymbol{E}_2\boldsymbol{E}_1\boldsymbol{\Omega}\boldsymbol{E}_1'\boldsymbol{E}_2' \cdots \boldsymbol{E}_{n-1}' = \boldsymbol{D},$$

where

$$\boldsymbol{D} = \begin{bmatrix} \omega_{11} & 0 & 0 & \cdots & 0 \\ 0 & \omega_{22} - \omega_{21}\omega_{11}^{-1}\omega_{12} & 0 & \cdots & 0 \\ 0 & 0 & h_{33} - h_{32}h_{22}^{-1}h_{23} & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & p_{nn} - p_{n,n-1}p_{n-1,n-1}^{-1}p_{n-1,n} \end{bmatrix},$$

with all the diagonal elements of $\boldsymbol{D}$ strictly positive. The matrices $\boldsymbol{E}_1$ and $\boldsymbol{E}_2$ are as above. In general, $\boldsymbol{E}_j$ is a matrix with nonzero values in the jth column below the principal diagonal, 1s along the principal diagonal, and zeros everywhere else.

Thus each $\boldsymbol{E}_j$ is lower triangular with unit determinant. Hence $\boldsymbol{E}_j^{-1}$ exists, and the following matrix exists:

$$\boldsymbol{A} = (\boldsymbol{E}_{n-1} \cdots \boldsymbol{E}_2 \boldsymbol{E}_1)^{-1} = \boldsymbol{E}_1^{-1} \boldsymbol{E}_2^{-1} \cdots \boldsymbol{E}_{n-1}^{-1}.$$

If $\boldsymbol{D}$ is premultiplied by $\boldsymbol{A}$ and postmultiplied by $\boldsymbol{A'}$, the result is

$$\boldsymbol{\Omega} = \boldsymbol{A}\boldsymbol{D}\boldsymbol{A'},$$

Recall that $\boldsymbol{E}_1$ represents the operation of multiplying the first row of $\boldsymbol{\Omega}$ by certain numbers and substracting from each of the sunsequent rows. Its inverse $\boldsymbol{E}_1^{-1}$ undoes this operation, which would be acheived by multiplying the first row by the same numbers and *adding* the results to the subsequent rows. Thus

$$\boldsymbol{E}_1^{-1} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ \omega_{21}\omega_{11}^{-1} & 1 & 0 & \cdots & 0 \\ \omega_{31}\omega_{11}^{-1} & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ \omega_{n1}\omega_{11}^{-1} & 0 & 0 & \cdots & 1 \end{bmatrix},$$

as may be verified directly by multiplying $\boldsymbol{E}_1$ by $\boldsymbol{E}_1^{-1}$ to obtain the identity matrix. Similarly,

$$\boldsymbol{E}_2^{-1} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & h_{32}h_{22}^{-1} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & h_{n2}h_{22}^{-1} & 0 & \cdots & 1 \end{bmatrix},$$

and so on. Because of this special structure, the series of multiplications defining $\boldsymbol{A}$ above turns out to be trivial to carry out and yields

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ \omega_{21}\omega_{11}^{-1} & 1 & 0 & \cdots & 0 \\ \omega_{31}\omega_{11}^{-1} & h_{32}h_{22}^{-1} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ \omega_{n1}\omega_{11}^{-1} & h_{n2}h_{22}^{-1} & k_{n3}k_{33}^{-1} & \cdots & 1 \end{bmatrix}.$$

That is, the jth column of $\boldsymbol{A}$ is just the jth column of $\boldsymbol{E}_j^{-1}$.

Since $\boldsymbol{A}$ is lower triagular with 1s along the principal diagonal, and $\boldsymbol{D}$ is diagonal with all its diagonal elements strictly positive, $\boldsymbol{\Omega = ADA'}$ is indeed the triangular fasctorization of $\boldsymbol{\Omega}$ we were looking for.

3.2. Uniqueness of the Triangular Factorization

THEOREM 1. *The triangular factorization $\boldsymbol{ADA'}$ of a symmetric positive definite matrix $\boldsymbol{\Omega}$ is unique.*

Proof. □

3.3. The Cholesky Factorization

A closely related factorization of a symmetric positive definite matrix Ω is obtained as follows. Define $D^{1/2}$ to be the $(n \times n)$ diagonal matrix whose diagonal elements are the square roots of the corresponding elements of the matrix D in the triangular factorization:

$$D^{1/2} = \begin{bmatrix} \sqrt{d_{11}} & 0 & 0 & \cdots & 0 \\ 0 & \sqrt{d_{22}} & 0 & \cdots & 0 \\ 0 & 0 & \sqrt{d_{33}} & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & \sqrt{d_{nn}} \end{bmatrix}.$$

Since the matrix D is unique and has strictly positive diagonal elements, the matrix $\boldsymbol{D}^{1/2}$ exists and is unique. Then the triangular factorization can be written as

$$\Omega = AD^{1/2}D^{1/2}A' = (AD^{1/2})(AD^{1/2})',$$

or

$$\Omega = PP',$$

where

$$\begin{aligned} P &= AD^{1/2} \\ &= \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ a_{21} & 1 & 0 & \cdots & 0 \\ a_{31} & a_{32} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & 1 \end{bmatrix} \begin{bmatrix} \sqrt{d_{11}} & 0 & 0 & \cdots & 0 \\ 0 & \sqrt{d_{22}} & 0 & \cdots & 0 \\ 0 & 0 & \sqrt{d_{33}} & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & \sqrt{d_{nn}} \end{bmatrix} \\ &= \begin{bmatrix} \sqrt{d_{11}} & 0 & 0 & \cdots & 0 \\ a_{21}\sqrt{d_{11}} & \sqrt{d_{22}} & 0 & \cdots & 0 \\ a_{31}\sqrt{d_{11}} & a_{32}\sqrt{d_{22}} & \sqrt{d_{33}} & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ a_{n1}\sqrt{d_{11}} & a_{n2}\sqrt{d_{22}} & a_{n3}\sqrt{d_{33}} & \cdots & \sqrt{d_{nn}} \end{bmatrix}. \end{aligned}$$

The expression $\Omega = PP'$ is known as the *Cholesky factorization* of Ω. Note that P, like A, is lower triangular, and is unique, i.e., for each symmetric positive definite matrix Ω there is a unique lower triangular matrix P such that $\Omega = PP'$. This follows directly from the uniqueness of the triangular factorization of Ω.

3.4. The Spectral Factorization of a Positive Definite Matrix

Theorem 2. *Suppose that the $n \times 1$ vector $\boldsymbol{y} \sim N(\boldsymbol{\mu}, \Sigma)$, where $\boldsymbol{\mu}$ is an $n \times 1$ vector and Σ is an $n \times n$ positive definite, symmetric matrix, and let $w = (\boldsymbol{y} - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{y} - \boldsymbol{\mu})$. Then $w \sim \chi^2_n$.*

Proof. It suffices to show that $w = \boldsymbol{z'z}$, where the $(n \times 1)$ vector $\boldsymbol{z}$ is distributed $N(0, I_n)$. We make the following steps:

(i) Since Σ is a positive definite matrix, we can write

$$\Sigma = C\Lambda C'$$

where Λ is the diagonal matrix whose diagonal elements are the eigenvalues of Σ, and C is the matrix whose columns are the corresponding eigenvectors of Σ. Now, $\boldsymbol{C}$ is an *orthonormal* matrix, i.e., $CC' = C'C = I$ so that $C^{-1} = C'$, i.e., its inverse is equal to its transpose. We can write

$$\begin{aligned} \Sigma &= C\Lambda C' = (C\Lambda^{1/2})(\Lambda^{1/2}C') \\ &= (C\Lambda^{1/2}C')(C\Lambda^{1/2}C')' = \Sigma^{1/2}(\Sigma^{1/2})', \end{aligned}$$

where $\Sigma^{1/2} = C\Lambda^{1/2}C'$ and $\Lambda^{1/2}$ is the diagonal matrix whose diagonal element is the square root of the corresponding diagonal element of Λ. The matrix $\Sigma^{1/2}$ is *a square root matrix* of Σ. There are, of course, many square roots of Σ, the Cholesky decomposition discussed above being another.

(ii) Since $\boldsymbol{C}$ is orthonormal, $C^{-1} = C'$, which means that

$$\Sigma^{-1} = (C\Lambda C')^{-1} = C\Lambda^{-1}C'.$$

This is a very convenient way of computing Σ^{-1} since Λ is a diagonal matrix and Λ^{-1} is simply the diagonal matrix whose elements are the reciprocals of the corresponding elements of Λ. Now, let $\Lambda^{-1/2}$ be the diagonal matrix whose diagonal elements are the reciprocal square roots of the corresponding diagonal elements of Λ. Clearly, then $\Lambda^{-1/2}(\Lambda^{-1/2})' = \Lambda^{-1}$, which justifies us in calling this matrix the inverse square root matrix of Λ.

(iii) Let $K = \Sigma^{-1/2} = C\Lambda^{-1/2}C'$ be a square root matrix of Σ^{-1}, i.e., a matrix for which $K = K'$, $KK' = C\Lambda^{-1}C = \Sigma^{-1}$.

(iv) Let $\boldsymbol{\epsilon} = \boldsymbol{y} - \boldsymbol{\mu}$. Then $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \Sigma)$.

(v) Let $\boldsymbol{z} = K\boldsymbol{\epsilon}$. Then $\boldsymbol{z} \sim N(\mathbf{0}, K'\Sigma K) = N(\mathbf{0}, I_n)$.

(vi) $w = \boldsymbol{\epsilon}'\Sigma^{-1}\boldsymbol{\epsilon} = \boldsymbol{\epsilon}'K'K\boldsymbol{\epsilon} = (K\boldsymbol{\epsilon})'(K\boldsymbol{\epsilon}) = \boldsymbol{z}'\boldsymbol{z} \sim \chi^2_n$.

□

Example 1. Let

$$\Sigma = \begin{bmatrix} 2 & -1 & 1 \\ -1 & 5 & 1 \\ 1 & 1 & 3 \end{bmatrix}.$$

Then

$$C = \begin{bmatrix} -0.1809726 & 0.5932347 & 0.7844243 \\ 0.9364571 & -0.1397551 & 0.3217400 \\ 0.3004946 & 0.7928059 & -0.5302470 \end{bmatrix}$$

and

$$\Lambda = \begin{bmatrix} 5.514137 & 0 & 0 \\ 0 & 3.571993 & 0 \\ 0 & 0 & 0.9138698 \end{bmatrix}.$$

The eigenvalues of $\boldsymbol{\Sigma}$ in the diagonal of $\boldsymbol{\Lambda}$ are here arranged in decreasing order, and to each of them corresponds the eigenvector of $\boldsymbol{\Sigma}$ contained in the matching column of $\boldsymbol{C}$. For example, to the first eigenvalue $\lambda_1 = 5.514137$ corresponts the eigenvector $\boldsymbol{c}_1 = (-0.1809726, 0.9364571, 0.3004946)'$.

The diagonal matrix $\Lambda^{-1/2}$ is then given by

$$\begin{aligned}\Lambda^{-1/2} &= \begin{bmatrix} 1/\sqrt{5.514137} & 0 & 0 \\ 0 & 1/\sqrt{3.571993} & 0 \\ 0 & 0 & 1/\sqrt{0.9138698} \end{bmatrix} \\ &= \begin{bmatrix} 0.4258545 & 0 & 0 \\ 0 & 0.5291084 & 0 \\ 0 & 0 & 1.046063 \end{bmatrix}.\end{aligned}$$

Clearly, $\Lambda^{-1/2}\Lambda^{-1/2\prime} = \Lambda^{-1}$. We can now compute the matrix $K = \Sigma^{-1/2}$ as

$$K = C\Lambda^{-1/2}C' = \begin{bmatrix} 0.8438200 & 0.1479681 & -0.2094063 \\ 0.1479681 & 0.4920731 & -0.1172490 \\ -0.2094063 & -0.1172490 & 0.6651328 \end{bmatrix} = K'.$$

We verify that

$$KK' = \Sigma^{-1} = \begin{bmatrix} 7/9 & 2/9 & -1/3 \\ 2/9 & 5/18 & -1/6 \\ -1/3 & -1/6 & 1/2 \end{bmatrix},$$

and that

$$K'\Sigma K = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = I_3.$$

■

3.5. Block Triangular Factorization

Suppose now we have observations on two sets of variables. The first set is collected in an $(n_1 \times 1)$ vector $\boldsymbol{Y}_1$ and the second set in an $(n_2 \times 1)$ vector $\boldsymbol{Y}_2$. Their second-moment matrix can be written in partitioned form as

$$\boldsymbol{\Omega} = \begin{bmatrix} E(\boldsymbol{Y}_1\boldsymbol{Y}_1') & E(\boldsymbol{Y}_1\boldsymbol{Y}_2') \\ E(\boldsymbol{Y}_2\boldsymbol{Y}_1') & E(\boldsymbol{Y}_2\boldsymbol{Y}_2') \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22} \end{bmatrix}.$$

where $\boldsymbol{\Omega}_{11}$ is an $(n_1 \times n_1)$ matrix, $\boldsymbol{\Omega}_{22}$ is an $(n_2 \times n_2)$ matrix, and the $(n_1 \times n_2)$ matrix $\boldsymbol{\Omega}_{12}$ is the transpose of the $(n_2 \times n_1)$ matrix $\boldsymbol{\Omega}_{21}$.

We can put zeros in the lower left $(n_2 \times n_1)$ block of $\mathbf{\Omega}$ by premultiplying $\mathbf{\Omega}$ by the following matrix:

$$\bar{\boldsymbol{E}}_1 = \begin{bmatrix} \boldsymbol{I}_{n_1} & \mathbf{0} \\ -\mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1} & \boldsymbol{I}_{n_2} \end{bmatrix}.$$

If $\mathbf{\Omega}$ is premultiplied by $\bar{\boldsymbol{E}}_1$ and postmultiplied by $\bar{\boldsymbol{E}}_1'$, the result is

$$\begin{bmatrix} \boldsymbol{I}_{n_1} & \mathbf{0} \\ -\mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1} & \boldsymbol{I}_{n_2} \end{bmatrix} \begin{bmatrix} \mathbf{\Omega}_{11} & \mathbf{\Omega}_{12} \\ \mathbf{\Omega}_{21} & \mathbf{\Omega}_{22} \end{bmatrix} \begin{bmatrix} \boldsymbol{I}_{n_1} & -\mathbf{\Omega}_{11}^{-1}\mathbf{\Omega}_{21} \\ \mathbf{0} & \boldsymbol{I}_{n_2} \end{bmatrix} = \begin{bmatrix} \mathbf{\Omega}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Omega}_{22} - \mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1}\mathbf{\Omega}_{12} \end{bmatrix}.$$

4. Non-Spherical Errors - Generalized Least Squares

In some situations we have reason to believe that the spherical error assumption is violated in our model.

(ii)′ $E(\boldsymbol{u}\boldsymbol{u}'|X) = \Sigma$.

Under this assumption

$$E(\hat{\boldsymbol{\beta}}|X) = \boldsymbol{\beta}$$

so the OLS coefficients are still unbiased, but

$$V(\hat{\boldsymbol{\beta}}|X) = (X'X)^{-1}X'\Sigma X(X'X)^{-1}.$$

Let K be the Cholesky decomposition of Σ^{-1}, i.e. $KK' = \Sigma^{-1}$ and $K'\Sigma K = I_n$, and consider the following transformations,

$$\tilde{\boldsymbol{y}} = K'\boldsymbol{y}, \quad \tilde{X} = K'X \quad \tilde{\boldsymbol{u}} = K'\boldsymbol{u}.$$

Then our model becomes

$$\tilde{\boldsymbol{y}} = \tilde{X}\boldsymbol{\beta} + \tilde{\boldsymbol{u}},$$

and for this model

$$E(\tilde{\boldsymbol{u}}|X) = E(K'\boldsymbol{u}|X) = K'E(\boldsymbol{u}|X) = \mathbf{0},$$

and

$$E(\tilde{\boldsymbol{u}}\tilde{\boldsymbol{u}}'|X) = E(K'\boldsymbol{u}\boldsymbol{u}'K|X) = K'E(\boldsymbol{u}\boldsymbol{u}'|X)K = K'\Sigma K = I_n.$$

It follows that the OLS estimate *of the transformed model* will be optimal, since for this model the errors are spherical. This estimator is given by

$$\begin{aligned} \tilde{\boldsymbol{\beta}} &= (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{\boldsymbol{y}} \\ &= (X'KK'X)^{-1}X'K'K\boldsymbol{y} \\ &= (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}\boldsymbol{y}. \end{aligned}$$

We will call this the *Generalized Least Squares (GLS) estimator*. Write

$$\begin{aligned}\tilde{\boldsymbol{\beta}} &= (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{\boldsymbol{y}} \\ &= \boldsymbol{\beta} + (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{\boldsymbol{u}}.\end{aligned}$$

This an unbiased estimator of $\boldsymbol{\beta}$ since

$$\begin{aligned}E(\tilde{\boldsymbol{\beta}}|X) &= \boldsymbol{\beta} + (\tilde{X}'\tilde{X})^{-1}\tilde{X}'E(\tilde{\boldsymbol{u}}|X) \\ &= \boldsymbol{\beta} + (\tilde{X}'\tilde{X})^{-1}\tilde{X}'K'E(\boldsymbol{u}|X) = \boldsymbol{\beta}.\end{aligned}$$

The variance of $\tilde{\boldsymbol{\beta}}$ is given by

$$\begin{aligned}V(\tilde{\boldsymbol{\beta}}|\boldsymbol{X}) &= E[(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})'|X] \\ &= E[(\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{\boldsymbol{u}}\tilde{\boldsymbol{u}}'\tilde{X}(\tilde{X}'\tilde{X})^{-1}|X] \\ &= (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\underbrace{E(\tilde{\boldsymbol{u}}\tilde{\boldsymbol{u}}'|X)}_{I_n}\tilde{X}(\tilde{X}'\tilde{X})^{-1} \\ &= (\tilde{X}'\tilde{X})^{-1} \\ &= (X'\Sigma^{-1}X)^{-1}.\end{aligned}$$

The GLS estimator $\tilde{\boldsymbol{\beta}}$ is optimal (BLUE) under nonspherical errors, which implies that $V(\tilde{\boldsymbol{\beta}}) - V(\hat{\boldsymbol{\beta}})$ is positive definite.

4.1. Heteroskedasticity

In the heteroskedastic error case,

$$\Sigma = \begin{bmatrix}\sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2\end{bmatrix}.$$

The Cholesky decomposition of Σ^{-1} is given by

$$K = \begin{bmatrix}\dfrac{1}{\sigma_1} & 0 & \cdots & 0 \\ 0 & \dfrac{1}{\sigma_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \dfrac{1}{\sigma_n}\end{bmatrix}$$

and it is easy to verify that $K'\Sigma K = I_n$ and $KK' = \Sigma^{-1}$. The transformed model is given by

$$K'\boldsymbol{y} = K'X\boldsymbol{\beta} + K'\boldsymbol{u},$$

which may be written as

$$\frac{y_i}{\sigma_i} = \sum_{j=1}^{k} \frac{x_{ij}}{\sigma_i} b_j + \frac{u_i}{\sigma_i}, \quad i = 1, ..., n,$$

that can now be estimated by OLS.

The only problem remaining is to specify an estimate of Σ. Recall that Σ here is a diagonal matrix with n non-zero elements, $\Sigma = \text{diag}\{\sigma_1^2, \sigma_2^2, ..., \sigma_n^2\}$. There is a problem here: we need to estimate n parameters from n observations! The resolution of this problem is to realize that the estimator of the $n \times n$ matrix Σ need not be consistent, it only needs to be unbiased! Let

$$\hat{\Sigma} = \text{diag}\{\hat{u}_1^2, \hat{u}_2^2, ..., \hat{u}_n^2\},$$

where $\hat{u}_i$ is the OLS residual for the ith observation, be an estimator of Σ. That this estimator can NOT be a consistent estimator of Σ is obvious, but that is fine since it is unbiased and we only need to worry about consistency of the variance matrices $(X'\hat{\Sigma}^{-1}X)^{-1}$ and $(X'X)^{-1}X'\hat{\Sigma}X(X'X)^{-1}$, and not of $\hat{\Sigma}$ itself. The dimension of these matrices are $k \times k$ and can be shown to be consistent for their population analogues $(X'\Sigma^{-1}X)^{-1}$ and $(X'X)^{-1}X'\Sigma X(X'X)^{-1}$, respectively.

Note, however, that we cannot replace Σ in the GLS estimator by $\hat{\Sigma} = \text{diag}\{\hat{u}_1^2, \hat{u}_2^2, ..., \hat{u}_n^2\}$ as this yields an inconsistent estimator (see Section 5.8.6). See Cameron and Trivedi p.82 and p.157-158.

4.1.1. Testing for Heteroskedasticity

Assume that $\sigma_i^2 = E(\varepsilon_i^2) = h(Z\alpha)$ where $Z = [1 z_1 \cdots z_p]$ is a set of variables entering the scale function of ε, $\alpha = [\alpha_1 \alpha_2 \cdots \alpha_p]$, and $h(\cdot)$ is some unspecified function that may take only positive values. The null of homoskedasticity can then be written as

$$H_0 : \alpha_2 = \alpha_3 = \cdots = \alpha_p = 0$$

under which $\sigma_i^2 = h(\alpha_1)$, a constant. The *Breusch Pagan* test of H_0 can now be implemented by regressing the OLS residuals on Z where Z may be the X's in the model and various power's and products of them. The R^2 of this auxiliary regression can then be used to test homoskedasticity since under the null

$$nR^2 \sim \chi_p^2,$$

as $n \to \infty$.

4.2. Autocorrelation

Another well known violation of the spherical error assumption that is often present in time-series data is autocorrelation. There are at least two possibilities. First the error may follow an *Autoregressive process* of order p, $AR(p)$, which is given by

$$\varepsilon_t = \rho_1 \varepsilon_{t-1} + \rho_2 \varepsilon_{t-2} + \cdots + \rho_p \varepsilon_{t-p} + u_t$$

where the ρ's are the coefficients of the AR process and the u_t's are iid disturbances. A second possibility is that the error follows a *Moving Average process* of order q, $MA(q)$, given by

$$\varepsilon_t = u_t + \alpha_1 u_{t-1} + \alpha_2 u_{t-2} + \cdots + \alpha_q u_{t-q}$$

where the α's are the coefficients of the MA process and the u_t's are again iid disturbances.

In what follows we will focus exclusively on the case where ε_t follows an $AR(1)$ process, which is the most common form of autocorrelation in empirical applications. Consider the linear model

$$y_t = x_t' \beta + \varepsilon_t$$

where y_t is the value of the depended variable at time t, x_t is a $k \times 1$ vector of the regressors at time t, β is a $k \times 1$ vector of coefficients, and ε_t is the disturbance at time t. Assume that ε_t follows a nonexplosive $AR(1)$ process, that is

$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t, \quad |\rho| < 1$$

and

$$E(u_t) = 0, \forall t, \quad E(u_t u_s) = \begin{cases} \sigma_u^2, & t = s \\ 0, & t \neq s \end{cases}, \forall t, s.$$

It follows that

$$\begin{aligned} \varepsilon_t &= \rho \varepsilon_{t-1} + u_t \\ &= \rho(\rho \varepsilon_{t-2} + u_{t-1}) + u_t = \rho^2 \varepsilon_{t-2} + \rho u_{t-1} + u_t \\ &= \rho^2(\rho \varepsilon_{t-3} + u_{t-2}) + \rho u_{t-1} + u_t = \rho^2 \varepsilon_{t-3} + \rho^2 \varepsilon_{t-2} + \rho u_{t-1} + u_t \\ &= \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ &= \rho^s \varepsilon_{t-s} + \sum_{j=0}^{s-1} \rho^j u_{t-j}. \end{aligned}$$

Since $|\rho| < 1$, we have that $\lim_{s\to\infty} \rho^s \varepsilon_{t-s} = 0$, while the series $\sum_{j=0}^{s-1} \rho^j u_{t-j}$ converges to a finite limit. Therefore, we can write

$$\varepsilon_t = \sum_{j=0}^{\infty} \rho^j u_{t-j}, \quad t = 1, ..., n.$$

Given this representation of the error term is it easy to compute its expectation and variance as

$$E(\varepsilon_t) = \sum_{j=0}^{\infty} \rho^j E(u_{t-j}) = 0, \quad t = 1, ..., n,$$

and

$$\begin{aligned} E(\varepsilon_t^2) &= E(\sum_{j=0}^{\infty} \rho^j u_{t-j})^2 \\ &= E(\sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \rho^{j+k} u_{t-j} u_{t-k}) \\ &= \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \rho^{j+k} E(u_{t-j} u_{t-k}) \\ &= \sigma_u^2 \sum_{j=0}^{\infty} \rho^{2j} \\ &= \frac{\sigma_u^2}{1-\rho^2}, \quad t = 1, ..., n \end{aligned}$$

where the last line uses $\sum_{j=0}^{\infty} (\rho^2)^j = 1/(1-\rho^2)$. Similarly, the covariance is given by

$$\begin{aligned} E(\varepsilon_t \varepsilon_{t-1}) &= E[(\rho \varepsilon_{t-1} + u_t)\varepsilon_{t-1}] \\ &= \rho E(\varepsilon_{t-1}^2) + E(\varepsilon_{t-1} u_t) \\ &= \rho \frac{\sigma_u^2}{1-\rho^2}, \quad t = 1, ..., n \end{aligned}$$

and generally

$$E(\varepsilon_t \varepsilon_{t-s}) = \rho^s \frac{\sigma_u^2}{1-\rho^2}, \quad t = s+1, s+2, ..., n.$$

From these results we deduce the variance-covariance matrix of the errors,

$$E(\varepsilon\varepsilon') = \sigma_u^2 \frac{1}{1-\rho^2} \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{bmatrix} = \sigma_u^2 W.$$

The inverse of W is

$$W^{-1} = \begin{bmatrix} 1 & -\rho & & & \\ -\rho & 1+\rho^2 & -\rho & 0 & \\ & -\rho & 1-\rho^2 & & \\ & & -\rho & -\rho & \\ & 0 & & 1+\rho^2 & -\rho \\ & & & -\rho & 1 \end{bmatrix},$$

and its Cholesky decomposition is given by

$$K' = \begin{bmatrix} \sqrt{1-\rho^2} & & & & \\ -\rho & 1 & & 0 & \\ & -\rho & 1 & & \\ & & -\rho & & \\ 0 & & & & \\ & & & -\rho & 1 \end{bmatrix}.$$

Transforming the model $y = X\beta + \varepsilon$ we obtain

$$K'y = K'X\beta + K'\varepsilon$$

which reduces to

$$\sqrt{1-\rho^2}\, y_1 = \sqrt{1-\rho^2}\, x_1'\beta + \sqrt{1-\rho^2}\, \varepsilon_1,$$

for the first observation ($t = 1$), and

$$y_t - \rho y_{t-1} = (x_t - \rho x_{t-1})'\beta + \varepsilon_t - \rho\varepsilon_{t-1}, \quad t = 2, ..., n,$$

for the rest of the observations. In applications, the alternative simplified matrix

$$K_1' = \begin{bmatrix} -\rho & 1 & & & \\ & -\rho & 1 & 0 & \\ & & -\rho & & \\ & 0 & & & \\ & & & -\rho & 1 \end{bmatrix}$$

is often used. The resulting transformation disregards the special status of the first observation, but this is inconsequential for large n. The model is given by

$$y_t - \rho y_{t-1} = (x_t - \rho x_{t-1})'\beta + \varepsilon_t - \rho\varepsilon_{t-1}, \quad t = 1, ..., n$$

To operationalize this, we need an estimate of ρ, but this is easy to obtain by first fitting the model by OLS and the estimating ρ by the auxiliary regression

$$\hat{\varepsilon}_t = \rho\hat{\varepsilon}_{t-1} + error$$

where the $\hat{\varepsilon}_t$'s are the OLS residuals.

4.2.1. Testing for Autocorrelation

Let $\hat{\varepsilon}_t = y_t - x_t'\hat{\beta}$ be again the residuals of the OLS regression. The pth order autocorrelation model is given by

$$\varepsilon_t = \rho_1\varepsilon_{t-1}\rho_2\varepsilon_{t-2} + \cdots \rho_s\varepsilon_{t-p} + u_t.$$

The null for no autocorrelation can be written as

$$H_0 : \rho_1 = \rho_2 = \cdots = \rho_p = 0.$$

Consider again the auxiliary regression

$$\hat{\varepsilon}_t = \rho_1\hat{\varepsilon}_{t-1}\rho_2\hat{\varepsilon}_{t-2} + \cdots \rho_s\hat{\varepsilon}_{t-p} + error$$

that does *not* contain an intercept. Then, under the null

$$nR^2 \sim \chi^2_p,$$

as $n \to \infty$.

5. Endogenous Regressors - Instrumental Variables

Consider again the linear regression model

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{u}$$

with spherical errors u, but this time assume that $X \not\perp u$. We say that X is *endogenous*, and since $E(X'\boldsymbol{u}) \neq \mathbf{0}$, the OLS coefficient will be biased. Assume that we have another set of variables Z, such that $\dim(Z) = \dim(X)$ and $Z \perp \boldsymbol{u}$ or $E(\boldsymbol{u}|Z) = 0$. Imposing the orthogonality condition in our sample we obtain

$$Z \perp \boldsymbol{u} \Rightarrow Z'\boldsymbol{u} = \mathbf{0} \Leftrightarrow Z'(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}_{ILS}) = \mathbf{0},$$

and solving we obtain the *Indirect Least Squares* (ILS) estimator

$$\hat{\boldsymbol{\beta}}_{ILS} = (Z'X)^{-1}Z'\boldsymbol{y}.$$

The ILS estimator is unbiased:

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}_{ILS}|X) &= E[(Z'X)^{-1}Z'\boldsymbol{y}|X,Z] \\ &= E[(Z'X)^{-1}Z'(X\boldsymbol{\beta} + \boldsymbol{u})|X,Z] \\ &= \boldsymbol{\beta} + (Z'X)^{-1}Z'E(\boldsymbol{u}|Z) \\ &= \boldsymbol{\beta}, \end{aligned}$$

and its variance is given by

$$\begin{aligned} V(\hat{\boldsymbol{\beta}}_{ILS}|X,Z) &= E[(\hat{\boldsymbol{\beta}}_{ILS} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_{ILS} - \boldsymbol{\beta})'|X,Z] \\ &= E[(Z'X)^{-1}Z'\boldsymbol{u}\boldsymbol{u}'Z(Z'X)^{-1}|X,Z] \\ &= \sigma_u^2(Z'X)^{-1}Z'Z(Z'X)^{-1}. \end{aligned}$$

It is worth noting that $V(\hat{\boldsymbol{\beta}}_{ILS}) > V(\hat{\boldsymbol{\beta}}_{OLS})$, so there is a cost in not having $X \perp \boldsymbol{u}$. The term $(Z'X)$ measures the (in-sample) correlation between Z and X. When this correlation is small the variance of $\hat{\boldsymbol{\beta}}_{ILS}$ blows up, so we would like to have an instrument that is highly correlated with X. If Z is highly correlated with X we call it a *strong* instrument. Otherwise, we say it is a *weak* instrument.

It is clear that we need at least as many instruments as endogenous regressors, but it doesn't hurt to have more. Consider then the more general situation in which $\dim(Z) \geq \dim(X)$. Our simple algebra above doesn't work anymore because the relevant matrices are not conformable, so we need to follow a 2-step procedure.

First, we will project X onto the space spanned by the columns of Z, i.e. we will estimate the vector-valued regression

$$X = Z\boldsymbol{\delta} + \boldsymbol{e}$$

and obtain fitted values

$$\hat{X} = Z(Z'Z)^{-1}Z'X \equiv P_Z X$$

where P_Z is the projection matrix that carries the X's into the space spanned by the Z's. Since the Z's are orthogonal to $\boldsymbol{u}$, the $\hat{X}$'s will be too, so the OLS estimates of the regression of $\boldsymbol{y}$ on $\hat{X}$ will be consistent. The *Two Stage Least Squares* (2SLS) estimator is defined as

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{2SLS} &= (\hat{X}'\hat{X})^{-1}\hat{X}'\boldsymbol{y} \\ &= (X'P_Z X)^{-1}X'P_Z\boldsymbol{y},\end{aligned}$$

where in the last line we have used the idempotency of the projection matrix P_Z. The expectation of this estimator is given by

$$\begin{aligned}E(\hat{\boldsymbol{\beta}}_{2SLS}|X,Z) &= E[(\hat{X}'\hat{X})^{-1}\hat{X}'\boldsymbol{y}|X,Z] \\ &= E[(\hat{X}'\hat{X})^{-1}\hat{X}'(X\boldsymbol{\beta}+\boldsymbol{u})|X,Z] \\ &= \boldsymbol{\beta} + E[(\hat{X}'\hat{X})^{-1}\hat{X}'\boldsymbol{u}|X,Z] \\ &= \boldsymbol{\beta}\end{aligned}$$

so the estimator is unbiased. It's variance is given by

$$\begin{aligned}V(\hat{\boldsymbol{\beta}}_{2SLS}|X,Z) &= E[(\hat{\boldsymbol{\beta}}_{2SLS}-\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_{2SLS}-\boldsymbol{\beta})'|X,Z] \\ &= E[(\hat{X}'\hat{X})^{-1}\hat{X}'\boldsymbol{u}\boldsymbol{u}'\hat{X}(\hat{X}'\hat{X})^{-1}|X,Z] \\ &= \sigma_u^2(\hat{X}\hat{X})^{-1} \\ &= \sigma_u^2(X'P_Z X)^{-1}.\end{aligned}$$

The 2SLS estimator can be computed in a two-step procedure (thus the name):

(i) Regress each of the endogenous X's on Z to get fitted values $\hat{X}$.

(ii) Regress $\boldsymbol{y}$ on $\hat{X}$ to obtain $\hat{\boldsymbol{\beta}}_{2SLS}$.

The covariance matrix of the second step regression, however, needs correction in that the estimate for σ_u^2 is not appropriate. The correct estimate is given by

$$\hat{\sigma}_u^2 = \frac{1}{n-k}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_{2SLS})^2,$$

which is different from the estimate

$$\frac{1}{n-k}\sum_{i=1}^{n}(y_i - \hat{\boldsymbol{x}}_i'\hat{\boldsymbol{\beta}}_{2SLS})^2$$

that the second step regression produces.

6. Seemingly Unrelated Regressions

Consider a system of m equations

$$\boldsymbol{y}_j = \boldsymbol{X}_j\boldsymbol{\beta}_j + \boldsymbol{u}_j, \qquad j = 1, ..., m$$

where each $\boldsymbol{y}_j$ is an $(n \times 1)$ vector, each $\boldsymbol{X}_j$ is an $(n \times k_j)$ matrix and each $\boldsymbol{\beta}_j$ is an $(k_j \times 1)$ vector of coefficients. To simplify notation we will assume that $k_j \equiv k$ for all $j = 1, ..., m$, but this assumption can be easily relaxed without affecting the developments[1]. Stacking the data

$$\begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_m \end{bmatrix}_{(mn)\times 1} = \begin{bmatrix} \boldsymbol{X}_1 & 0 & \cdots & 0 \\ 0 & \boldsymbol{X}_2 & \cdots & \vdots \\ \vdots & \cdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \boldsymbol{X}_m \end{bmatrix}_{(mn)\times(mk)} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_m \end{bmatrix}_{(mk)\times 1} + \begin{bmatrix} \boldsymbol{u}_1 \\ \boldsymbol{u}_2 \\ \vdots \\ \boldsymbol{u}_m \end{bmatrix}_{(mn)\times 1}$$

we can write the system as

$$\mathscr{Y} = \mathscr{X}\boldsymbol{\beta} + \mathscr{U}.$$

Assume that the $\boldsymbol{u}_j$'s are only contemporaneously correlated, i.e.

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{12} & \sigma_{22} & \cdots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1m} & \sigma_{2m} & \cdots & \sigma_{mm} \end{bmatrix}_{(m\times m)} = [\sigma_{ij}]_{i,j=1,...,m}.$$

Then

$$\begin{aligned} E(\mathscr{U}\mathscr{U}') &= \boldsymbol{\Sigma} \otimes \boldsymbol{I}_n \\ &= \begin{bmatrix} \sigma_{11}\boldsymbol{I}_n & \sigma_{12}\boldsymbol{I}_n & \cdots & \sigma_{1m}\boldsymbol{I}_n \\ \sigma_{12}\boldsymbol{I}_n & \sigma_{22}\boldsymbol{I}_n & \cdots & \sigma_{2m}\boldsymbol{I}_n \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1m}\boldsymbol{I}_n & \sigma_{2m}\boldsymbol{I}_n & \cdots & \sigma_{mm}\boldsymbol{I}_n \end{bmatrix}_{(mn\times mn)}, \end{aligned}$$

[1] Take k to be the total number of explanatory variables appearing in the system, and set to zero the coefficients of the $\boldsymbol{x}$'s that don't appear in a specific equation.

where $\otimes$ denotes the Kronecker product.

This is a linear model with non-spherical errors, so, by the GLS principle, the optimal *SUR estimator* is given by

$$\hat{\boldsymbol{\beta}}_{SUR} = (\mathscr{X}'(\boldsymbol{\Sigma} \otimes \boldsymbol{I}_n)^{-1}\mathscr{X})^{-1}\mathscr{X}'(\boldsymbol{\Sigma} \otimes \boldsymbol{I}_n)^{-1}\mathscr{Y},$$

which is the same as a GLS estimator of the regression of $\mathscr{Y}$ on $\mathscr{X}$ with covariance matrix $\boldsymbol{\Sigma} \otimes \boldsymbol{I}_n$.

In practise $\boldsymbol{\Sigma}$ is not known, so we will again need to estimate it from the data. Let $\hat{\boldsymbol{u}}_j$ be the residuals from the OLS estimation of the jth equation and define

$$\hat{\sigma}_{ij} = \frac{\sum_{t=1}^{n} \hat{u}_{it}\hat{u}_{jt}}{n-k}, \qquad i,j = 1,...,m.$$

Then $\hat{\boldsymbol{\Sigma}} = [\hat{\sigma}_{ij}]_{i,j=1,...,m}$ is a consistent estimator of $\boldsymbol{\Sigma}$. Replacing $\boldsymbol{\Sigma}$ with $\hat{\boldsymbol{\Sigma}}$ in our estimator above we obtain the *Feasible SUR* (FSUR) estimator.

An important observation is that there is no efficiency gain from reweighting by $(\boldsymbol{\Sigma} \otimes \boldsymbol{I}_n)^{-1}$ if $\mathscr{X} = (\boldsymbol{I}_n \otimes \boldsymbol{X}_0)$. That is, if all equations contain the same set of explanatory variables $\boldsymbol{X}_j = \boldsymbol{X}_0$ for all $j = 1,...,m$ as would be the case in some demand systems, we gain nothing from SUR over what can be accomplished by an equation-by-equation OLS. To see this write

$$(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{I}_n)(\boldsymbol{I}_n \otimes \boldsymbol{X}_0) = (\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{X}_0).$$

Recall that SUR imposes the following orthogonality condition

$$\mathscr{X}'(\boldsymbol{\Sigma} \otimes \boldsymbol{I}_n)^{-1}\hat{\mathscr{U}} = \boldsymbol{0},$$

But if $\mathscr{X} = (\boldsymbol{I}_n \otimes \boldsymbol{X}_0)$, this is equivalent to

$$(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{X}_0')\hat{\mathscr{U}} = \boldsymbol{0},$$

which is implied by the equation-by-equation OLS orthogonality condition

$$\boldsymbol{X}_0'\hat{\boldsymbol{u}}_j = \boldsymbol{0}, \qquad j = 1,...,m.$$

Therefore, when all equations have the same vector of explanatory variables, SUR and OLS are identical.

7. Three Stage Least Squares

Consider a system of m equations

$$y_i = Y_i\delta_i + X_i\gamma_i + \varepsilon_i, \qquad i = 1,...,m.$$

We have to deal with two problems:

(i) the Y_i's are endogenous; and
(ii) the ε_i's are correlated across different equations.

To correct for endogeneity we will use 2SLS, while to correct for correlation across equations will take one more step and do SUR estimation. This is a 3-step procedure and the estimator will be called a Three Stage Least Squares (3SLS) estimator.

Rewrite the above system as

$$y_i = Z_i\beta_i + \varepsilon_i, \qquad i = 1, ..., m.$$

where $Z_i = [Y_i|X_i]$ and $\beta_i = [\delta_i|\gamma_i]$, and stack the observations to obtain

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}_{(mn)\times 1} = \begin{bmatrix} Z_1 & 0 & \cdots & 0 \\ 0 & Z_2 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & Z_m \end{bmatrix}_{(mn)\times(mk)} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix}_{(mk)\times 1} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_m \end{bmatrix}_{(mn)\times 1}$$

or

$$\mathscr{Y} = \mathscr{Z}\beta + \mathscr{E}.$$

Assume again that the ε_j's are only contemporaneously correlated, i.e.

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{12} & \sigma_{22} & \cdots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1m} & \sigma_{2m} & \cdots & \sigma_{mm} \end{bmatrix}_{(m\times m)} = [\sigma_{ij}]_{i,j=1,...,m}$$

so that $E(\mathscr{E}\mathscr{E}') = \Sigma \otimes I_n$.

Start by taking care of the SUR effect. Let Λ be the Cholesky decomposition of Σ^{-1} so that $\Lambda \otimes I_n$ is the Cholesky decomposition of $(\Sigma \otimes I_n)^{-1} = \Sigma^{-1} \otimes I_n$, and transform the data as follows

$$\tilde{\mathscr{Y}} = (\Lambda \otimes I_n)'\mathscr{Y}, \qquad \tilde{\mathscr{Z}} = (\Lambda \otimes I_n)'\mathscr{Z}, \qquad \tilde{\mathscr{E}} = (\Lambda \otimes I_n)'\mathscr{E}.$$

In terms of the transformed data the model is given by

$$\tilde{\mathscr{Y}} = \tilde{\mathscr{Z}}\beta + \tilde{\mathscr{E}},$$

and this model has a spherical error since $E(\tilde{\mathscr{E}}\tilde{\mathscr{E}}') = I_{mn}$.

We next turn to the endogeneity problem and correct for it by using a 2SLS estimator. For instruments we will use the X_i's that are exogenous. Let X be the matrix of all exogenous X_i's across all the equations and define

$$\mathscr{X} = \begin{bmatrix} X & 0 & \cdots & 0 \\ 0 & X & \cdots & \vdots \\ \vdots & \cdots & \ddots & \vdots \\ 0 & \cdots & \cdots & X \end{bmatrix}_{(mn)\times(mk)} = I_m \otimes X.$$

First we project $\tilde{\mathscr{Z}}$ on $\mathscr{X}$ to obtain

$$\hat{\tilde{\mathscr{Z}}} = \mathscr{X}(\mathscr{X}'\mathscr{X})^{-1}\mathscr{X}'\tilde{\mathscr{Z}} = P_{\mathscr{X}}\tilde{\mathscr{Z}}.$$

Finally, we compute the 3SLS estimator by

$$\hat{\beta}_{3SLS} = (\hat{\tilde{\mathscr{Z}}}'\hat{\tilde{\mathscr{Z}}})^{-1}\hat{\tilde{\mathscr{Z}}}'\tilde{\mathscr{Y}}.$$

Unraveling the previous steps, we can open up the 3SLS estimator as

$$\begin{aligned} \hat{\beta}_{3SLS} &= (\tilde{\mathscr{Z}}'P'_{\mathscr{X}}P_{\mathscr{X}}\tilde{\mathscr{Z}})^{-1}\tilde{\mathscr{Z}}'P'_{\mathscr{X}}P_{\mathscr{X}}\tilde{\mathscr{Y}} \\ &= (\mathscr{Z}(\Lambda\otimes I_n)'P_{\mathscr{X}}(\Lambda\otimes I_n)\mathscr{Z})^{-1}\mathscr{Z}'(\Lambda\otimes I_n)'P_{\mathscr{X}}(\Lambda\otimes I_n)\mathscr{Y} \\ &= (\mathscr{Z}\Sigma^{-1}\otimes P_{\mathscr{X}}\mathscr{Z})^{-1}\mathscr{Z}'\Sigma^{-1}\otimes P_{\mathscr{X}}\mathscr{Y}. \end{aligned}$$

To compute the 3SLS estimator we may use the following procedure:

(i) Predict $Z_i, i = 1, ..., m$.
(ii) Compute the 2SLS estimator $\hat{\beta}_{2SLS,i}, i = 1, ..., m$.
(iii) Estimate Σ by $\hat{\Sigma} = [\hat{\sigma}_{ij}]_{i,j=1,...,m}$ where $\hat{\sigma}_{ij}$ are estimated from the 2SLS residuals

$$\hat{\sigma}_{ij} = \frac{1}{n-k}\sum_{t=1}^{n}\hat{\varepsilon}_{it}\hat{\varepsilon}_{jt}.$$

(iv) Compute $\hat{\beta}_{3SLS}$.

We may iterate this procedure by re-estimating $\hat{\Sigma}$ from the residuals of 3SLS regression and recomputing $\hat{\beta}_{3SLS}$ until convergence. Convergence, however, is very fast here, so the first-step estimate will in most cases be nearly optimal and further iterations will only produce minor improvements.

8. Digression Into GMM

All of the above Weighted Least Squares problems have a similar structure in that we find it optimal to reweight the OLS by a weighting matrix W. The general problem may be written as

$$\min_{\beta,W} \ \varepsilon(\beta)'W\varepsilon(\beta)$$

where the minimization is performed both over β and W.

1. GLS: In the case of nonspherical errors our model is given by

$$y = X\beta + \varepsilon$$

with $E(\varepsilon\varepsilon') = \Sigma$. In this case $W = \Sigma^{-1}$ and

$$\hat{\beta}_{GLS} = \underset{\beta,\Sigma}{\operatorname{argmin}} \, (y - X\beta)'\Sigma^{-1}(y - X\beta)$$

where minimization is performed both over β and Σ.

2. 2SLS: In the case of endogenous regressors

$$y = X\beta + \varepsilon$$

with $E(\varepsilon\varepsilon') = \sigma_\varepsilon^2 I_n$, but X is not orthogonal to ε. If Z is a set of valid instruments and P_Z is the projection matric into the space spanned by the Z's, then $W = P_Z$ and

$$\hat{\beta}_{2SLS} = \underset{\beta}{\operatorname{argmin}} \, (y - X\beta)'P_Z(y - X\beta).$$

where minimization is performed only over β since the weighting matrix P_Z is constant here.

3. SUR: In the seemingly unrelated regression context, the model is given by

$$\mathscr{Y} = \mathscr{Z}\beta + \mathscr{E}.$$

with $E(\mathscr{E}\mathscr{E}') = \Sigma \otimes I_n$. This is exactly like GLS, so $W = \Sigma^{-1} \otimes I_n$ here and

$$\hat{\beta}_{SUR} = \underset{\beta,\Sigma}{\operatorname{argmin}} \, (\mathscr{Y} - \mathscr{Z}\beta)'(\Sigma^{-1} \otimes I_n)(\mathscr{Y} - \mathscr{Z}\beta)$$

where minimization is performed over both β and Σ.

4. 3SLS: Finally in the case of 3SLS estimator the model is given by

$$\mathscr{Y} = \mathscr{Z}\beta + \mathscr{E}.$$

with $E(\mathscr{E}\mathscr{E}') = \Sigma \otimes I_n$, and some of the $\mathscr{Z}$'s not orthogonal to $\mathscr{E}$. Let $\mathscr{X}$ be the part of $\mathscr{Z}$ that is orthogonal to $\mathscr{E}$, and let $W = P_{\mathscr{X}}$ be the matrix projection the $\mathscr{Z}$'s into the space spanned by the $\mathscr{X}$'s. Then the 3SLS estimator solves

$$\min_{\beta, \Sigma=[\sigma_{ij}]} (\mathscr{Y} - \mathscr{Z}\beta)'(\Sigma^{-1} \otimes P_{\mathscr{X}})(\mathscr{Y} - \mathscr{Z}\beta)$$

where minimization is performed over both β and Σ.

All of the above estimators are *Generalized Method of Moments* (GMM) estimators. Let Z be data (for example $Z = (y, X)$ in the regression case) and let $m(\beta; Z)$ be a moment condition such that $E[m(\beta_0; Z)] = 0$, i.e. the population value β_0 solves the FOC. A GMM estimator is one that minimizes a squared Euclidean distance of sample moments from their population counterpart of zero. Let W be a positive semi-definite matrix, so that $(m'Wm)^{1/2}$ is a measure of the distance of m from zero. A GMM estimator solves

$$\min_{\beta, W} \; m(\beta; Z)' \; W \; m(\beta; Z).$$

The matrix W can be chosen in an optimal way, if we set it equal to the covariance matrix of of the moment conditions evaluated at the true parameter β_0, i.e., if we put $W = E[m(\beta_0; Z)m(\beta_0; Z)']$. All of the above estimators have exactly this form, with $m(\beta; Z)$ equal to the residual vector $\varepsilon(\beta; y, x)$ and W equal to the variance covariance of $\varepsilon(\beta; y, x)$.

In practise it may be difficult to maximize both over β and W in one step, so people often employ a 2-step strategy:

(i) Set $W = I$, the identity matrix, and estimate $\hat{\beta}$ by minimizing the criterion function over β.

(ii) Set $\hat{W} = E[m(\hat{\beta}; Z)m(\hat{\beta}, Z)']$ and re-optimize the criterion function over β to get $\hat{\beta}_{GMM}$.

It is possible to iterate this further by refitting W based on the new estimate and re-optimize until convergence, but one step is enough to make the estimator asymptotically optimal.

9. Empirical Application: The Fulton Fish Market

Everybody knows Wall Street in lower Manhattan where the New York Stock Exchange is located, but few have heard of the nearby Fulton Street where a Fish Market used to be.

Graddy (1995) collected data from one supplier at the Fulton Fish Market. The data was collected in order to test if the market was consistent with competitive equilibrium as theory would describe it.

9.1. Endogeneity Bias

The classic illustration of biases created by endogenous regressors was given by Working (1927). In what follows we will describe the problem of endogeneity in terms of the estimation of a classical demand and supply system of equations and derive the IV and 2SLS estimators from first principles.

9.1.1. A simultaneous Equation Model of Market Equilibrium

Consider the following simple model of demand and supply:

$$\begin{aligned} q_t^d &= \alpha_0 + \alpha_1 p_t + u_t, && \text{(demand equation)} \\ q_t^s &= \beta_0 + \beta_1 p_t + \upsilon_t, && \text{(supply equation)} \\ q_t^d &= q_t^s, && \text{(market equilibrium)} \end{aligned}$$

where q_t^d is the quantity demanded for the commodity in question (say coffee) in period t, q_t^s is the quantity supplied, and p_t is the price. The error term u_t in the demand equation represents factors that influence coffee demand other than price, such as the public's mood for coffee. Depending on the value of u_t, the demand curve in the price-quantity plane shifts up or down. Similarly, υ_t represents supply factors other than price. We assume that $E(u_t) = 0$ and $E(\upsilon_t) = 0$ and that $\text{Cov}(u_t, \upsilon_t) = 0$. If we define $q_t = q_t^d = q_t^s$, the three equation system above can be reduced to a two-equation system:

$$\begin{aligned} q_t &= \alpha_0 + \alpha_1 p_t + u_t, && \text{(demand equation)} \\ q_t &= \beta_0 + \beta_1 p_t + \upsilon_t. && \text{(supply equation)} \end{aligned}$$

We say that a regressor is *endogenous* if it is not predetermined (i.e., not orthogonal to the error term), that is, if it does not satisfy the orthogonality condition. When the equation includes an intercept, the orthogonality condition is violated and hence the regressor is endogenous, if and only if the regressor is correlated with the error term. In the present example the regressor p_t is endogenous in both equations. To see why, solve the system of equations for

(p_t, q_t) to obtain,

$$\begin{aligned} p_t &= \frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1} + \frac{\upsilon_t - u_t}{\alpha_1 - \beta_1}, \\ q_t &= \frac{\alpha_1\beta_0 - \alpha_0\beta_1}{\alpha_1 - \beta_1} + \frac{\alpha_1\upsilon_t - \beta_1 u_t}{\alpha_1 - \beta_1}. \end{aligned}$$

We see that price is a function of the two error terms. From the solution for p_t we can calculate the covariance of p_t with the demand shifter u_t and the supply shifter υ_t as

$$\mathrm{Cov}(p_t, u_t) = -\frac{\mathrm{Var}(u_t)}{\alpha_1 - \beta_1}, \qquad \mathrm{Cov}(p_t, \upsilon_t) = \frac{\mathrm{Var}(\upsilon_t)}{\alpha_1 - \beta_1},$$

which are not zero (unless $\mathrm{Var}(u_t) = 0$ and $\mathrm{Var}(\upsilon_t) = 0$). Therefore, assuming that the demand curve is downward-sloping ($\alpha_1 < 0$) and the supply curve is upward-slopping ($\beta_1 > 0$), price is positively correlated with the demand shifter and negatively correlated with the supply shifter.

9.1.2. Endogeneity Bias

So what do we get when we regress quantity on price? Do we estimate the demand or the supply curve? The answer is *neither*, because price is endogenous in both the demand and supply equations. To see this, assume we run the regression

$$q_t = \gamma_0 + \gamma_1 p_t + e_t.$$

Recall that the Least Squares slope coefficient for $\hat{\gamma}_1$ satisfies

$$\mathrm{plim}\, \hat{\gamma}_1 = \frac{\mathrm{Cov}(p_t, q_t)}{\mathrm{Var}(p_t)}.$$

But from the demand equation we have

$$\mathrm{Cov}(p_t, q_t) = \alpha_1 \mathrm{Var}(p_t) + \mathrm{Cov}(p_t, u_t),$$

so by combining the two we get the asymptotic bias of $\hat{\gamma}_1$ in terms of the the price effect in the demand curve as

$$\mathrm{plim}\, \hat{\gamma}_1 - \alpha_1 = \frac{\mathrm{Cov}(p_t, u_t)}{\mathrm{Var}(p_t)}.$$

Similarly, the asymptotic bias of $\hat{\gamma}_1$ in terms of the the price effect in the supply curve is

$$\mathrm{plim}\, \hat{\gamma}_1 - \beta_1 = \frac{\mathrm{Cov}(p_t, \upsilon_t)}{\mathrm{Var}(p_t)}.$$

But since, as we have already seen, $\mathrm{Cov}(p_t, u_t) \neq 0$ and $\mathrm{Cov}(p_t, \upsilon_t) \neq 0$, OLS is inconsistent *both* for α_1 and β_1. This phenomenon is known as the *endogeneity bias*, or as the *simultaneity bias*, because it appears often in systems of simultaneous equations, as in the present example.

Solving $\operatorname{plim}\hat{\gamma}_1 - \beta_1$ for $\text{Var}(p_t)$ and substituting the result in the expression for $\operatorname{plim}\hat{\gamma}_1 - \alpha_1$ we obtain,

$$\operatorname{plim}\hat{\gamma}_1 = \frac{\alpha_1 \text{Cov}(p_t, \upsilon_t) - \beta_1 \text{Cov}(p_t, u_t)}{\text{Cov}(p_t, \upsilon_t) - \text{Cov}(p_t, u_t)}.$$

Finally, using the expressions for $\text{Cov}(p_t, u_t)$ and $\text{Cov}(p_t, \upsilon_t)$ that we have already obtained we get

$$\operatorname{plim}\hat{\gamma}_1 = \frac{\alpha_1 \text{Var}(\upsilon_t) + \beta_1 \text{Var}(u_t)}{\text{Var}(\upsilon_t) + \text{Var}(u_t)}.$$

We see that the OLS coefficient of the regression of quantity on price is a linear combination of the demand and supply price effects with weights equal to the variances of the errors in the system. Of course, if $\text{Var}(u_t) = 0$ $\hat{\gamma}_1$ is consistent for α_1, and if $\text{Var}(\upsilon_t) = 0$ $\hat{\gamma}_1$ is consistent for β_1, but the assumption that one of the equations is measured without error is not realistic.

9.1.3. Observable Demand and Supply Shifters

The reason neither the demand curve nor the supply curve is consistently estimated in the model above is that we cannot infer from the data at hand whether the change in price and quantity is due to a demand shift or a supply shift (in the model above both shifters u_t and υ_t were unobservable and enter the equations as error terms). This suggests that it might be possible to estimate the demand curve if some of the factors shifting the supply curve were observable, and that we could estimate the supply curve if some of the factors shifting the demand curve were observable. So suppose that the supply shifter υ_t can be divided into an observable factor x_t and an unobservable factor ζ_t uncorrelated with x_t, and write the supply equation as

$$q_t = \beta_0 + \beta_1 p_t + \beta_2 x_t + \zeta_t \qquad \text{(supply equation)}.$$

Now imagine that the supply shifter x_t is also uncorrelated with the demand error term u_t. For example, we could think of x_t as the temperature in coffee-growing regions. A variable that is correlated with the endogenous regressor but uncorrelated with the error term is called an *instrumental variable*, or simply an *instrument*. We will show that the presence of such a supply shifter allows us to estimate the demand curve.

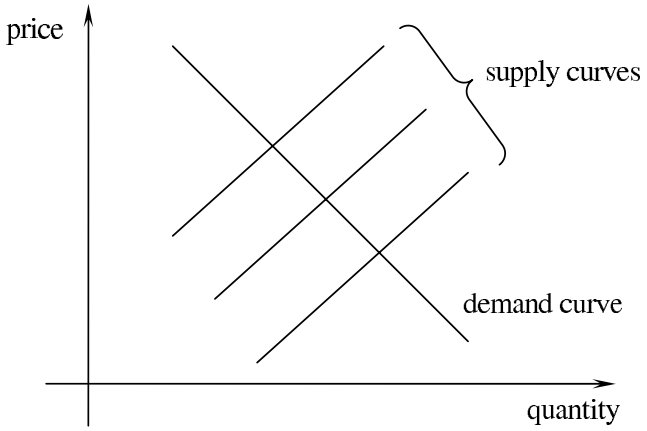
Solve the new system of simultaneous equations to obtain

$$\begin{aligned} p_t &= \frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1} + \frac{\beta_2}{\alpha_1 - \beta_1} x_t + \frac{\zeta_t - u_t}{\alpha_1 - \beta_1}, \\ q_t &= \frac{\alpha_1\beta_0 - \alpha_0\beta_1}{\alpha_1 - \beta_1} + \frac{\alpha_1\beta_2}{\alpha_1 - \beta_1} x_t + \frac{\alpha_1\zeta_t - \beta_1 u_t}{\alpha_1 - \beta_1}. \end{aligned}$$

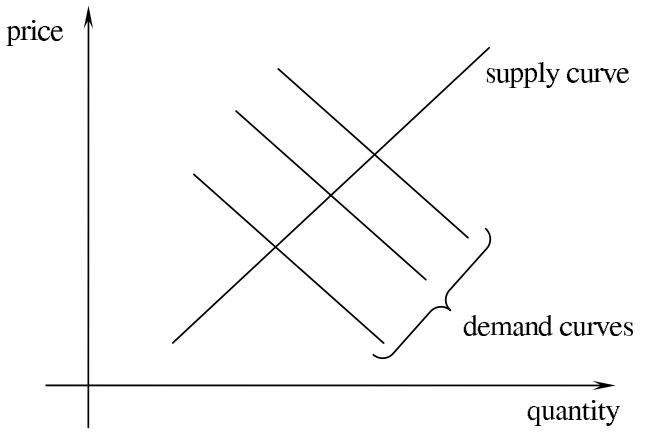
Since $\mathrm{Cov}(x_t, \zeta_t) = 0$ and $\mathrm{Cov}(x_t, u_t) = 0$, it follows from the equation for p_t that

$$\mathrm{Cov}(x_t, p_t) = \frac{\beta_2}{\alpha_1 - \beta_1}\mathrm{Var}(x_t) \neq 0.$$

So x_t is indeed a valid instrument.



(a) No shifts in demand



(b) No shifts in supply

With a valid instrument at hand we can estimate the demand price coefficient α_1 consistently. We have

$$\begin{aligned} \text{Cov}(x_t, q_t) &= \alpha_1 \text{Cov}(x_t, p_t) + \text{Cov}(x_t, u_t) \\ &= \alpha_1 \text{Cov}(x_t, p_t) \end{aligned}$$

since $\text{Cov}(x_t, u_t) = 0$ by assumption and as we have verified $\text{Cov}(x_t, p_t) \neq 0$. So we can divide both sides by $\text{Cov}(x_t, p_t)$ to obtain

$$\alpha_1 = \frac{\text{Cov}(x_t, q_t)}{\text{Cov}(x_t, p_t)}.$$

A natural estimator that suggests itself is thus

$$\hat{\alpha}_{1,IV} = \frac{\widehat{\text{Cov}}(x_t, q_t)}{\widehat{\text{Cov}}(x_t, p_t)},$$

where the hats are the sample covariances. This estimator is called the *instrumental variable (IV) estimator* with x_t as the instrument. We sometimes say "the endogenous regressor p_t is instrumented by x_t".

A similar argument shows that if there is an observable demand shifter then the supply price effect β_1 is identified, and when there are both observable demand and supply shifters then both price effects α_1 and β_1 are identified.

A closely related procedure to estimate α_1 is the *two stage least squares estimator (2SLS)*. In the first stage, endogenous regressor p_t is regressed on a constant and the instrument x_t, to obtain fitted values $\hat{p}_t$. Then the dependent variable q_t is regressed on $\hat{p}_t$ to obtain an estimate of α_1 given by

$$\hat{\alpha}_{1,2SLS} = \frac{\widehat{\text{Cov}}(\hat{p}_t, q_t)}{\widehat{\text{Var}}(\hat{p}_t)},$$

where the hats are again sample covariance and sample variance. To relate the second stage regression to the demand equation rewrite the demand equation as

$$q_i = \alpha_0 + \alpha_1 p_t + [u_t + \alpha_1 (p_t - \hat{p}_t)].$$

The second stage regression estimates this equation, treating the bracketed term as the error term. Since by construction $(p_t - \hat{p}_t)$ is orthogonal to p_t, OLS on the second regression will produce a consistent estimate of α_1. In the present example, the IV and 2SLS estimators are numerically the same.

9.2. The model

We first consider the estimation of the demand equation in the following simultaneous system of supply and demand

$$\begin{aligned} \text{(Demand)} \qquad \log q_t^d &= \alpha_0 + \alpha_1 \log p_t + \delta X_t + u_t \\ \text{(Supply)} \qquad \log q_t^s &= \beta_0 + \beta_1 \log p_t + \beta_2 s_t + \upsilon_t \end{aligned}$$

where (all logs are natural):

- q_t quantity of whiting sold (pounds),
- p_t price of whiting ($/pound),
- X_t a set of regressors affecting the demand for whiting,
- s_t dummy that equals 1 if the weather is stormy and 0 otherwise.

The price p_t in this system is endogenous, so OLS is inconsistent. We can, however, use the supply shifter s_t (stormy weather) as an instrument to identify the demand curve. Stormy weather clearly affects the supply adversely, but there is no reason to beleive that it has any effect on demand.

Consider the simple model where there are no X_t variables in the demand equation (model (1) in the table). The sample variance-covariance matrix of $\log q_t$, $\log p_t$ and s_t is given by

	$\log q_t$	$\log p_t$	s_t
$\log q_t$	.550077		
$\log p_t$	−.078899	.145874	
s_t	−.075134	.069414	.207043

Thus, the OLS estimate of the price elasticity of demand for whiting is given by

$$\hat{\alpha}_{1,OLS} = \frac{\widehat{\text{Cov}}(\log p_t, \log q_t)}{\widehat{\text{Var}}(\log p_t)} = \frac{-0.078899}{0.145874} = -0.54087,$$

while, according to our discussion above, the IV estimate using s_t as an instrument is

$$\hat{\alpha}_{1,IV} = \frac{\widehat{\text{Cov}}(s_t, \log q_t)}{\widehat{\text{Cov}}(s_t, \log p_t)} = \frac{-0.075134}{0.069414} = -1.0824.$$

We see that the OLS estimate of the demand price elasticity for whiting is severely biased towards zero. The IV estimate is close to -1, which is the value of demand elasticity that maximizes revenues. In particular, if sellers have market power, theory predicts that they would choose to operate at the unit elasticity region of the demand curve, as this would maximize

Table 1. OLS and IV estimates of the demand for whiting.

Dependent variable: Log quantity				
Instrument for Log price: Stormy (= 1 if stormy weather)				
	OLS		IV	
Variable	(1)	(2)	(1)	(2)
Log price	−0.54	−0.54	−1.08	−1.22
	(0.18)	(0.18)	(0.48)	(0.55)
Monday		0.03		−0.03
		(0.21)		(0.17)
Tuesday		−0.49		−0.53
		(0.20)		(0.18)
Wednesday		−0.54		−0.58
		(0.21)		(0.20)
Thursday		0.09		0.12
		(0.20)		(0.18)
Weather on shore (= 1 if cold)		−0.06		0.07
		(0.13)		(0.16)
Rain on shore (= 1 if rain)		0.07		0.07
		(0.18)		(0.16)
R^2	0.08	0.23		

Note: $n = 111$ observations. Standard errors in parentheses.

their total revenues. If costs are fixed, which seems to be a reasonable assumption here (the daily catch is sold through an auction each day, the cost of which is not affected by the exact quantity auctioned off each day), revenue maximization is equivalent to profit maximization. The findings of this research thus support both that sellers have market power and that they operate rationally, i.e., they maximize their profits.

Table 1 also reports a second model (Model 2) that includes day-of-the-week dummies, as well as, weather-on-shore dummies that might affect the demand for whiting. We see that the weather dummies are insignificant, and of the day dummies only the Tuesday and Wednesday dummies are significanlty negative (relative the excluded Friday dummy). The estimated price elasticity is a little higher but −1 is included in the 95% confidence interval.

When the instrument is a binary dummy variable, as it is the case here, the IV estimator may also be written as

$$\begin{aligned}\hat{\alpha}_{1,IV} &= \frac{\widehat{\text{Cov}}(\log q_t, s_t)}{\widehat{\text{Cov}}(\log p_t, s_t)} \\ &= \frac{\widehat{E}(\log q_t|s_t=1) - \widehat{E}(\log q_t|s_t=0)}{\widehat{E}(\log p_t|s_t=1) - \widehat{E}(\log p_t|s_t=0)} \\ &= \frac{8.265156 - 8.628047}{0.044929 - (-0.2903333)} \\ &= \frac{-0.362891}{0.3352623} = -1.0824.\end{aligned}$$

In this form, the IV estimator is the classical Wald (1940) estimator, also known as the *grouping estimator*. Grouping the observations into two groups, the stormy-weather-at-sea, $s = 1$, group, and the non-stormy-weather-at-sea, $s = 0$, group, the IV estimate is the difference in average log-quantity across the two groups divided by the difference in average log-price across the two groups.

Koenker likens IV estimation to looking at an object through a *looking glass*. Consider the earnings–schooling example. Suppose a one-unit change in the instrument z is associated with 0.2 more years of schooling and with a \$500 increase in annual earnings. This increase in earnings is a consequence of the indirect effect that increase in z led to increase in schooling, which in turn increases income. Then it follows that 0.2 years additional schooling is associated with a \$500 increase in earnings, so that a one-year increase in schooling is associated with a \$500/0.2 = \$2,500 increase in earnings. The causal estimate of β is therefore 2,500. In mathematical notation we have estimated the changes dx/dz and dy/dz and calculated the causal estimator as

$$\beta_{IV} = \frac{dy/dz}{dx/dz}. \tag{9.1}$$

This approach to identification of the causal parameter β is given in Heckman (2000, p. 58). Therefore, we look at the causal effect of x on y through the "looking glass" of the instrument z.

References

Graddy K., (2006), "The Fulton Fish Market", *Journal of Economic Perspectives*, 20, 207-220.

Heckman J. (2000), "Causal Parameters and Policy Analysis in Economics- A Twentieth Century Retrospective", *Quarterly Journal of Economics*, 115, 45-97.

Wald A. (1940), "The Fitting of Straight Lines if Both Variables are Subject to Error", *Annals of Mathematical Statistics*, 11, 284-300.

Working, E.J., (1927), "What Do 'Statistical Demand' Curves Show?", *Quarterly Journal of Economics*, 41, 212-235.

Every good estimate deserves a standard error.

– *Roger Koenker.*