

CHAPTER 1

METHODS OF MOMENTS FOR SINGLE LINEAR EQUATION MODELS

Method-of-moment (*MOM*) estimator for single linear equation models is introduced here, whereas MOM for multiple linear equations will be examined in the next chapter. Least squares estimator (LSE) is reviewed to estimate the conditional mean (i.e., regression function) in a model with *exogenous* regressors. Not just conditional mean, but conditional variance also matters, and it is discussed under the headings “heteroskedasticity/homoskedasticity” and generalized LSE (GLS). Instrumental variable estimator (IVE) and generalized method-of-moment (GMM) estimator allow *endogenous* regressors; IVE and GMM include LSE as a special case. Endogeneity matters greatly for policy variables, as the “*ceteris paribus*” effect of a policy is of interest but endogenous regressors lead to biased effect estimates. In addition to MOM estimation, testing linear hypotheses with “Wald test” is studied.

1 Least Squares Estimator (LSE)

This section introduces standard linear models with exogenous regressors, and then reviews least squares estimator (LSE) for regression functions, which is a “bread-and-butter” estimator in econometrics. Differently from the conventional approach, however, LSE will be viewed as a MOM. Also differently from the conventional approach, we will adopt a large sample framework and invoke only a few assumptions.

1.1 LSE as a Method of Moment (MOM)

1.1.1 Linear Model

Consider a linear model

$$y_i = x_i' \beta + u_i, \quad i = 1, \dots, N$$

where x_i is a $k \times 1$ “regressor” vector with its first component being 1 (i.e., $x_i = (1, x_{i2}, \dots, x_{ik})'$), $\beta \equiv (\beta_1, \dots, \beta_k)'$ is a $k \times 1$ parameter vector reflecting effects of x_i on y_i , and u_i is an “error” term. In β , β_1 is called the “intercept” whereas β_2, \dots, β_k are called the “slopes.” The left-hand side variable y_i is the “dependent” or “response” variable, whereas components of x_i are

“regressors,” “explanatory variables,” or “independent variables.” Think of x_i as a collection of the observed variables affecting y_i through $x'_i\beta$, and u_i as a collection of the unobserved variables affecting y_i . Finding β with data (x'_i, y_i) , $i = 1, \dots, N$, is the main goal in regression analysis. Assume that (x'_i, y_i) , $i = 1, \dots, N$, are *independent and identically distributed (iid)* unless otherwise noted, which means that each (x'_i, y_i) is an independent draw from a common probability distribution. We will often omit the subscript i indexing individuals.

The linear model is linear in β , but not necessarily linear in x_i , and it is more general than it looks. For instance, x_3 may be x_2^2 , in which case $\beta_2x_2 + \beta_3x_2^2$ depicts a quadratic relationship between x_2 and y : the “effect” of x_2 on y is then $\beta_2 + 2\beta_3x_2$ —the first derivative of $\beta_2x_2 + \beta_3x_2^2$ with respect to (wrt) x_2 . For instance, with y monthly salary and x_2 age, the effect of age on monthly salary may be quadratic: going up to a certain age and then declining after. Also x_4 may be x_2x_3 , in which case

$$\beta_2x_2 + \beta_3x_3 + \beta_4x_2x_3 = (\beta_2 + \beta_4x_3)x_2 + \beta_3x_3 :$$

the effect of x_2 on y is $\beta_2 + \beta_4x_3$. For instance, x_3 can be education level: the effect of age on monthly salary is not the constant slope β_2 , but $\beta_2 + \beta_4x_3$ which varies depending on education level. The display can be written also as $\beta_2x_2 + (\beta_3 + \beta_4x_2)x_3$ to be interpreted analogously. The term x_2x_3 is called the *interaction term* between x_2 and x_3 , and its coefficient is the interaction effect. By estimating β with data (x'_i, y_i) , $i = 1, \dots, N$, we can find these effects.

1.1.2 LSE and Moment Conditions

The *least squares estimator (LSE)* for β is obtained by minimizing

$$\frac{1}{N} \sum_i (y_i - x'_ib)^2$$

wrt b , where $y_i - x'_ib$ can be viewed as a “prediction error” in predicting y_i with the linear function x'_ib . LSE is also often called *ordinary LSE (OLS)*, relative to “generalized LSE” to appear later.

The first-order condition for the LSE b_{lse} is

$$\frac{1}{N} \sum_i x_i(y_i - x'_ib_{lse}) = 0 \iff \frac{1}{N} \sum_i x_iy_i = \frac{1}{N} \sum_i x_ix'_i \cdot b_{lse}.$$

Assuming that $N^{-1} \sum_i x_ix'_i$ is invertible, solve this for b_{lse} to get

$$b_{lse} = \left(\frac{1}{N} \sum_i x_ix'_i \right)^{-1} \cdot \frac{1}{N} \sum_i x_iy_i = \left(\sum_i x_ix'_i \right)^{-1} \cdot \sum_i x_iy_i.$$

The *residual* $\hat{u}_i \equiv y_i - x'_ib_{lse}$, which is an estimator for u_i , has zero sample mean and zero sample covariance with the regressors due to the first-order condition:

$$\frac{1}{N} \sum_i x_i (y_i - x_i' b_{lse}) = \left(\frac{1}{N} \sum_i \hat{u}_i, \frac{1}{N} \sum_i x_{i2} \hat{u}_i, \dots, \frac{1}{N} \sum_i x_{ik} \hat{u}_i \right)' = 0.$$

Instead of minimizing $N^{-1} \sum_i (y_i - x_i' b)^2$, LSE can be motivated directly from a moment condition. Observe that the LSE first-order condition at $b = \beta$ is $N^{-1} \sum_i x_i u_i = 0$, and its population version is

$$E(xu) = 0 \iff \begin{bmatrix} E(u) \\ E(x_2 u) \\ \vdots \\ E(x_k u) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$\iff E(u) = 0, \text{ } COV(x_j, u) = 0 \text{ (or } COR(x_j, u) = 0), j = 2, \dots, k$$

as $COV(x_j, u) = E(x_j u) - E(x_j)E(u)$, where COV and COR stand for covariance and correlation, respectively.

Replacing u with $y - x'\beta$ yields

$$E\{x(y - x'\beta)\} = 0 \iff E(xy) = E(xx')\beta$$

which is a restriction on the joint distribution of (x', y) . Assuming that $E(xx')$ is invertible, we get

$$\beta = \{E(xx')\}^{-1} \cdot E(xy).$$

LSE b_{lse} is just a *sample analog* of this expression of β , obtained by replacing $E(xx')$ and $E(xy)$ with their sample versions $N^{-1} \sum_i x_i x_i'$ and $N^{-1} \sum_i x_i y_i$. Instead of identifying β by minimizing the prediction error, here β is identified by the “information” (i.e., the assumption) that the observed x is “orthogonal” to the unobserved u .

For any $k \times 1$ constant vector γ ,

$$\gamma' E(xx') \gamma = E(\gamma' x x' \gamma) = E\{(x' \gamma)' (x' \gamma)\} = E\{(x' \gamma)^2\} \geq 0.$$

Hence $E(xx')$ is positive semidefinite (p.s.d.). Assume that

$$E(xx') \text{ is of full rank.}$$

As $E(xx')$ is p.s.d., this full rank condition is equivalent to $E(xx')$ being positive definite (p.d.) and thus being invertible. Note that $E(xx')$ being p.d. is equivalent to $E^{-1}(xx')$ being p.d. where $E^{-1}(xx')$ means $\{E(xx')\}^{-1}$.

1.1.3 Zero Moments and Independence

The assumption $E(xu) = 0$ is the weakest for the LSE to be a valid estimator for β as can be seen in the next subsection. In econometrics, the following two assumptions have been used as well for LSE:

(i) $E(u|x) = 0 \{ \iff E(y|x) = x'\beta \text{ for the linear model} \}$

(ii) u is independent of x and $E(u) = 0$.

Note that $E(u|x) = 0$ implies $E(u) = E\{E(u|x)\} = 0$. For the three assumptions, the following implications hold:

$$\text{independence of } u \text{ from } x \text{ and } E(u) = 0 \implies E(u|x) = 0 \implies E(xu) = 0;$$

the last implication holds because $E(xu) = E\{xE(u|x)\} = 0$.

We will use mainly $E(u|x) = 0$ from now on unless otherwise mentioned, because $E(xu) = 0$ would not take us much farther than LSE while the independence is too strong to be realistic. The regressor vector x is often said to be *exogenous* if any one of the three conditions holds. The function $E(y|x) = x'\beta$ is called the (*mean*) *regression function*, which is nothing but a location measure in the distribution of $y|x$. We can also think of other location measures, say quantiles, in the distribution of $y|x$, which then yield “quantile regression functions.”

In β , the intercept β_1 shows the level of y , and each slope represents the effect of its regressor on $E(y|x)$ while controlling for (i.e., holding constant) the other regressors. This may be understood in

$$\frac{\partial E(y|x)}{\partial x_j} = \beta_j, \quad j = 1, \dots, k.$$

The condition of “holding the other regressors constant”—reflected here with the partial differentiation symbol ∂ —may be better understood when “partial regression” is explained later. The formal causal interpretation of regarding x_j as a cause and β_j as its effect on the response y requires a little deeper reasoning (see, e.g., Lee, 2005, and the references therein). This is because LSE is a MOM which depends only on the covariances of the variables involved, and the covariances per se do not designate any variable as a cause or the response.

1.2 Asymptotic Properties of LSE

As $N \rightarrow \infty$, the sample will be “close” to the population, and we would want b_{lse} to converge to β in some sense. This is necessary for b_{lse} to be a “valid” estimator for β . Going further, to be a “good” estimator for β , b_{lse} should converge fast to β . For instance, both N^{-1} and N^{-2} converge to 0, and they are valid “estimators” for 0, but N^{-2} is better than N^{-1} because N^{-2} converges to 0 faster. This subsection discusses these issues in the names “consistency” and “asymptotic distribution.” The first-time readers may want to only browse this subsection instead of reading every detail, to come back later when better motivated theoretically. The upshot of this subsection is the display (*) showing the asymptotic distribution of b_{lse} (with its variance estimator in (**)) and its practical version (**) showing that b_{lse} will degenerate (i.e., converge) to β as $N \rightarrow \infty$. The main steps will also appear in the instrumental variable estimator (IVE) section.

1.2.1 LLN and LSE Consistency

A *law of large numbers (LLN)*, for an iid random variable (rv) sequence z_1, \dots, z_N with $E(z) < \infty$, holds that

$$\frac{1}{N} \sum_i z_i \rightarrow^p E(z) \text{ as } N \rightarrow \infty$$

where “ \rightarrow^p ” denotes convergence in probability:

$$P\left(\left|\frac{1}{N} \sum_i z_i - E(z)\right| < \varepsilon\right) \rightarrow 1 \text{ as } N \rightarrow \infty \text{ for any constant } \varepsilon > 0;$$

(the estimator) $\bar{z}_N \equiv N^{-1} \sum_i z_i$ is said to be “*consistent*” for (the parameter) $E(z)$.

If \bar{z}_N is a matrix, the LLN applies to each component. This element-wise convergence in probability of \bar{z}_N to $E(z)$ is equivalent to $|\bar{z}_N - E(z)| \rightarrow^p 0$ where $|A| \equiv \{tr(A'A)\}^{1/2}$ for a matrix A —the usual matrix norm—in the sense that the element-wise convergence implies the norm convergence and vice versa. As “ $\bar{z}_N - E(z) \rightarrow^p 0$ ” means that the difference between \bar{z}_N and $E(z)$ converges to 0 in probability, for two rv matrix sequences W_N and M_N , “ $W_N - M_N \rightarrow^p 0$ ” (or $W_N \rightarrow^p M_N$) means that the difference between the two rv matrix sequences converges to zero in probability.

Substitute $y_i = x_i' \beta + u_i$ into b_{lse} to get

$$b_{lse} = \beta + \left(\frac{1}{N} \sum_i x_i x_i'\right)^{-1} \frac{1}{N} \sum_i x_i u_i.$$

Clearly, $b_{lse} \neq \beta$ due to the second term on the right-hand side (rhs) which shows that each $x_i u_i$ contributes to the deviation $b_{lse} - \beta$. Using the LLN, we have

$$\frac{1}{N} \sum_i x_i u_i \rightarrow^p E(xu) = 0 \quad \text{and} \quad \frac{1}{N} \sum_i x_i x_i' \rightarrow^p E(xx').$$

Substituting these into the preceding display, we can get $b_{lse} \rightarrow^p \beta$, but we need to deal with the inverse: for a square random matrix W_N , when $W_N \rightarrow^p W$, will W_N^{-1} converge to W^{-1} in probability?

It is known that, for a rv matrix W_N and a constant matrix W_o ,

$$f(W_N) \rightarrow^p f(W_o) \quad \text{if } W_N \rightarrow^p W_o \text{ and } f(\cdot) \text{ is continuous at } W_o.$$

The inverse $f(W) = W^{-1}$ of W , when it exists, is the adjoint of W divided by the determinant $\det(W)$. Because $\det(W)$ is a sum of products of elements of W and the adjoint consists of determinants, both $\det(W)$ and the adjoint are continuous in W , which implies that W^{-1} is continuous in W (see, e.g.,

Schott, 2005). Thus, W^{-1} is continuous at W_o so long as W_o^{-1} exists, and using the last display, we get $A_N^{-1} \rightarrow^p A^{-1}$ if $A_N \rightarrow^p A$ so long as A^{-1} exists; note that A_N^{-1} exists for a large enough N because $\det(A_N) \neq 0$ for a large enough N . Hence,

$$\left(\frac{1}{N} \sum_i x_i x_i' \right)^{-1} \rightarrow^p E^{-1}(xx') < \infty \text{ as } N \rightarrow \infty.$$

Therefore, b_{lse} is β plus a product of two terms, one consistent for a zero vector and the other consistent for a bounded matrix; thus the product is consistent for zero, and we have $b_{lse} \rightarrow^p \beta$: *b_{lse} is consistent for β .*

1.2.2 CLT and \sqrt{N} -Consistency

For the asymptotic distribution of the LSE, a *central limit theorem (CLT)* is needed that, for an iid random vector sequence z_1, \dots, z_N with finite second moments,

$$\frac{1}{\sqrt{N}} \sum_i \{z_i - E(z)\} \rightsquigarrow N(0, E[\{z - E(z)\}\{z - E(z)\}']) \text{ as } N \rightarrow \infty$$

where “ \rightsquigarrow ” denotes *convergence in distribution*; i.e., letting $\Psi(\cdot)$ denote the df of $N(0, E[\{z - E(z)\}\{z - E(z)\}'])$,

$$\lim_{N \rightarrow \infty} P \left\{ \frac{1}{\sqrt{N}} \sum_i \{z_i - E(z)\} \leq t \right\} = \Psi(t) \quad \forall t.$$

When $w_N \rightarrow^p 0$, it is also denoted as $w_N = o_p(1)$; “ $o_p(1)$ ” is the probabilistic analog for $o(1)$ where $o(1)$ is a sequence converging to 0. For \bar{z}_N , we thus have $\bar{z}_N - E(z) = o_p(1)$. In comparison to $w_N = o_p(1)$, “ $w_N = O_p(1)$ ” means that $\{w_N\}$ is *bounded in probability (or stochastically bounded)*—i.e., “not explosive as $N \rightarrow \infty$ ” (even if it does not converge to anything) in the probabilistic sense. Note that $o_p(1)$ is also $O_p(1)$. Formally, $w_N = O_p(1)$ is that, for any constant $\varepsilon > 0$, there exists a constant δ_ε such that

$$\sup_N P\{|w_N| > \delta_\varepsilon\} < \varepsilon.$$

A single rv z always satisfies $P\{|z| > \delta_\varepsilon\} < \varepsilon$, because we can capture “all but ε ” probability mass by choosing δ_ε large enough. The last display means that we can capture all but ε probability mass with δ_ε for any rv in the sequence w_1, w_2, \dots . Any random sequence converging in distribution is $O_p(1)$, which implies $N^{-1/2} \sum_i \{z_i - E(z)\} = O_p(1)$.

To understand O_p better, consider N^{-1} and N^{-2} , both of which converge to 0. Observe $N^{-1}/N^{-1} = 1$, but $N^{-1}/N^{-1+\varepsilon} = 1/N^\varepsilon \rightarrow 0$ whereas $N^{-1}/N^{-1-\varepsilon} = N^\varepsilon \rightarrow \infty$ for any constant $\varepsilon > 0$. Thus the “(fastest) convergence rate” is N^{-1} which, when divided into N^{-1} , makes the resulting ratio

bounded. Analogously, the convergence rate for N^{-2} is N^{-2} . Now consider $z_N \equiv z/\sqrt{N}$ where z is a rv. Then $\sqrt{N}z_N = z = O_p(1)$ (or $z_N = O_p(1/\sqrt{N})$) because we can choose δ_ε for any constant $\varepsilon > 0$ such that

$$\sup_N P\left(|\sqrt{N}z_N| > \delta_\varepsilon\right) = \sup_N P(|z| > \delta_\varepsilon) = P(|z| > \delta_\varepsilon) < \varepsilon.$$

For an estimator a_N for a parameter α , in most cases, we have $\sqrt{N}(a_N - \alpha) = O_p(1)$: a_N is “ \sqrt{N} -consistent.” This means that $a_N \xrightarrow{p} \alpha$, and that the convergence rate is $N^{-1/2}$ which, when divided into $a_N - \alpha$, makes the resulting product bounded in probability. For most cases in our discussion, it would be harmless to think of the \sqrt{N} -consistency of a_N as $\sqrt{N}(a_N - \alpha)$ converging to a normal distribution as $N \rightarrow \infty$.

Analogously to $o(1)O(1) = o(1)$ —“a sequence converging to zero” times “a bounded sequence” converges to zero—it holds that $o_p(1)O_p(1) = o_p(1)$. Likewise, $o_p(1) + O_p(1) = O_p(1)$. *Slutsky Lemma* shows more: if $w_N \rightsquigarrow w$ (thus $w_N = O_p(1)$) and $m_N \xrightarrow{p} m_o$, then

$$(i) \quad m_N w_N \rightsquigarrow m_o w$$

$$(ii) \quad m_N + w_N \rightsquigarrow m_o + w.$$

Slutsky Lemma (i) states that, not just the product $m_N w_N$ is $O_p(1)$, its asymptotic distribution is that of w times the constant m_o . Slutsky Lemma (ii) can be understood analogously.

1.2.3 LSE Asymptotic Distribution

Observe

$$\sqrt{N}(b_{lse} - \beta) = \left(\frac{1}{N} \sum_i x_i x_i'\right)^{-1} \cdot \frac{1}{\sqrt{N}} \sum_i x_i u_i.$$

From the CLT, we have

$$\frac{1}{\sqrt{N}} \sum_i x_i u_i \rightsquigarrow N\{0, E(xx'u^2)\}.$$

Using Slutsky Lemma (i),

$$\text{if } B_N \rightsquigarrow N(0, C) \text{ and } A_N \xrightarrow{p} A, \text{ then } A_N B_N \rightsquigarrow N(0, ACA').$$

Apply this to

$$B_N = \frac{1}{\sqrt{N}} \sum_i x_i u_i \text{ and } A_N = \left(\frac{1}{N} \sum_i x_i x_i'\right)^{-1} \xrightarrow{p} E^{-1}(xx')$$

to get

$$\sqrt{N}(b_{lse} - \beta) \rightsquigarrow N(0, \Omega) \quad \text{where } \Omega \equiv E^{-1}(xx')E(xx'u^2)E^{-1}(xx') : \quad (*)$$

$\sqrt{N}(b_{lse} - \beta)$ is *asymptotically normal with mean 0 and variance Ω* . Often this convergence in distribution (or “in law”) of $\sqrt{N}(b_{lse} - \beta)$ is informally stated as

$$b_{lse} \sim N \left\{ \beta, \frac{1}{N} E^{-1}(xx') E(xx'u^2) E^{-1}(xx') \right\} \quad (*)$$

The asymptotic variance Ω of $\sqrt{N}(b_{lse} - \beta)$ can be estimated consistently with (this point will be further discussed later)

$$\Omega_N \equiv \left(\frac{1}{N} \sum_i x_i x_i' \right)^{-1} \left(\frac{1}{N} \sum_i x_i x_i' \hat{u}_i^2 \right) \left(\frac{1}{N} \sum_i x_i x_i' \right)^{-1}. \quad (**)$$

Alternatively (and informally), the asymptotic variance of b_{lse} is estimated consistently with

$$\frac{\Omega_N}{N} = \left(\sum_i x_i x_i' \right)^{-1} \left(\sum_i x_i x_i' \hat{u}_i^2 \right) \left(\sum_i x_i x_i' \right)^{-1}.$$

Equipped with Ω_N and the asymptotic normality, we can test hypotheses involving β as to be seen later.

1.3 Matrices and Linear Projection

It is sometimes convenient (for computation) to express b_{lse} using matrices. Define $Y \equiv (y_1, \dots, y_N)'$, $U \equiv (u_1, \dots, u_N)'$, and $X \equiv (x_1, \dots, x_N)'$ where $x_i = (x_{i1}, \dots, x_{ik})'$ so that

$$X_{N \times k} \equiv \begin{bmatrix} x_1' \\ \vdots \\ x_N' \end{bmatrix} = \begin{bmatrix} x_{11}, & x_{12}, & \dots, & x_{1k} \\ & \vdots & & \\ x_{N1}, & x_{N2}, & \dots, & x_{Nk} \end{bmatrix};$$

the numbers below X denote its dimension. In this matrix notation, the N linear equations $y_i = x_i' \beta + u_i$, $i = 1, \dots, N$, become $Y = X\beta + U$, and

$$\frac{1}{N} \sum_i (y_i - x_i' \beta)^2 = \frac{1}{N} \sum_i u_i^2 = \frac{1}{N} U' U = \frac{1}{N} (Y - X\beta)' (Y - X\beta).$$

Differentiating this, the LSE first-order condition is $N^{-1} X' (Y - Xb_{lse}) = 0$, which is also a moment condition for MOM. This yields

$$b_{lse} = (X' X)^{-1} X' Y.$$

The parts $X' X$ and $X' Y$ are the same as $\sum_i x_i x_i'$ and $\sum_i x_i y_i$, respectively. For example, with $k = 2$ and $x_{i1} = 1 \forall i$,

$$\begin{aligned} X' X &= \begin{bmatrix} 1 & \cdots & 1 \\ x_{12} & \cdots & x_{N2} \end{bmatrix} \begin{bmatrix} 1 & x_{12} \\ \vdots & \vdots \\ 1 & x_{N2} \end{bmatrix} = \begin{bmatrix} N & \sum_i x_{i2} \\ \sum_i x_{i2} & \sum_i x_{i2}^2 \end{bmatrix}, \\ \sum_i x_i x_i' &= \sum_i \begin{pmatrix} 1 \\ x_{i2} \end{pmatrix} (1, x_{i2}) = \sum_i \begin{bmatrix} 1 & x_{i2} \\ x_{i2} & x_{i2}^2 \end{bmatrix} = \begin{bmatrix} N & \sum_i x_{i2} \\ \sum_i x_{i2} & \sum_i x_{i2}^2 \end{bmatrix}. \end{aligned}$$

Define the $N \times N$ “(linear) projection matrix on X ”

$$P_X \equiv X(X'X)^{-1}X'$$

to get

$$\begin{aligned}\hat{Y} &\equiv Xb_{lse} = X(X'X)^{-1}X'Y = P_XY \quad (\text{“fitted value of } Y\text{”}), \\ \hat{U} &\equiv Y - Xb_{lse} = Y - P_XY = Q_XY, \quad \text{where } Q_X \equiv I_N - P_X;\end{aligned}$$

$\hat{U} = (\hat{u}_1, \dots, \hat{u}_N)'$ is the $N \times 1$ residual vector. We may think of Y comprising X and the other components. Then P_X *extracts* the X part of Y , and Q_X *removes* the X part of Y (or Q_X extracts the non- X part of Y). The fitted value $\hat{Y} = Xb_{lse}$ is the part of Y explained by X , and the residual \hat{U} is the part of Y unexplained by X as clear in the decomposition

$$Y = I_NY = P_XY + (I_N - P_X)Y = P_XY + Q_XY = Xb_{lse} + \hat{U}.$$

The LSE $(X'X)^{-1}X'Y$ is called the *sample (linear) projection coefficients of Y on X* . The population versions of the linear projection and linear projection coefficient are, respectively,

$$x'\beta \quad \text{and} \quad \beta \equiv E^{-1}(xx')E(xy).$$

The matrices P_X and Q_X are symmetric and idempotent:

$$P_X' = P_X, \quad P_XP_X = P_X \quad \text{and} \quad Q_X' = Q_X, \quad Q_XQ_X = Q_X.$$

Also note

$$P_XX = X \quad \text{and} \quad Q_XX = 0_N :$$

extracting the X part of X gives X itself, and removing the X part of X yields 0.

Suppose we use 1 as the only regressor. Defining 1_N as the $N \times 1$ vector of 1's and denoting Q_{1_N} just as Q_1 ,

$$\begin{aligned}Q_1Y &= \left(I_N - 1_N(1_N'1_N)^{-1}1_N' \right) Y = \left(I_N - 1_N \frac{1}{N} 1_N' \right) Y \\ &= \left(I_N - \frac{1}{N} 1_N 1_N' \right) Y \\ &= \left\{ \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 \end{bmatrix} - \frac{1}{N} \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \cdots & 1 & 1 \end{bmatrix} \right\} \cdot \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \\ &= \begin{bmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \vdots \\ y_N - \bar{y} \end{bmatrix}.\end{aligned}$$

The part $(1'_N 1_N)^{-1} 1'_N Y = \bar{y}$ demonstrates that the LSE with 1 as the sole regressor is just the sample mean \bar{y} . Q_1 may be called the “mean-deviation” or “mean-subtracting” matrix.

1.4 R^2 and Two Examples

Before we present two examples of LSE, we introduce some terminologies frequently used in practice. Recall the LSE asymptotic variance estimator $\Omega_N/N \equiv [\omega_{N,hj}]$, $h, j = 1, \dots, k$; i.e., the element of Ω_N/N in row h and column j is denoted as $\omega_{N,hj}$. The *t-values* (*t-ratios*, or *z-values*) are defined as

$$\frac{b_{lse,j}}{\sqrt{\omega_{N,jj}}}, \quad j = 1, \dots, k, \quad \text{where } b_{lse} = (b_{lse,1}, \dots, b_{lse,k})'.$$

Since the diagonal of Ω_N/N is the asymptotic variances of $b_{lse,j}$, $j = 1, \dots, k$, the j th *t-value* asymptotically follows $N(0, 1)$ under the $H_0 : \beta_j = 0$, and hence it is a test statistic for $H_0 : \beta_j = 0$. The off-diagonal terms of Ω_N/N are the asymptotic covariances for $b_{lse,j}$, $j = 1, \dots, k$, and they are used for hypotheses involving multiple parameters.

The “*standard error* (of model)” s_N and “*R-squared*” R^2 are defined as

$$\begin{aligned} s_N &\equiv \left(\frac{\sum_i \hat{u}_i^2}{N - k} \right)^{1/2} \rightarrow^p SD(u), \\ R^2 &\equiv 1 - \frac{N^{-1} \sum_i \hat{u}_i^2}{N^{-1} \sum_i (y_i - \bar{y})^2} \rightarrow^p 1 - \frac{V(u)}{V(y)} = \frac{V(x'\beta)}{V(y)}, \text{ as} \\ V(y) &= V(x'\beta + u) = V(x'\beta) + V(u) \quad \text{because } COV(x'\beta, u) = 0. \end{aligned}$$

R^2 shows the proportion of $V(y)$ that is explained by $x'\beta$, and R^2 measures the “*model fitness*.” In general, the higher the R^2 is the better, because the less is buried in the unobserved u . But this statement should be qualified, because R^2 keeps increasing by adding more regressors into the model. Using fewer regressors to explain y is desirable, which is analogous to using fewer shots to hit a target.

Recall $\hat{Y} = X b_{lse}$, $Y = \hat{Y} + \hat{U}$, and the idempotent mean-subtracting matrix Q_1 to observe

$$\begin{aligned} Q_1 \hat{U} &= \hat{U} \text{ (because the sample mean of } \hat{U} \text{ is already zero),} \\ Y' Q_1 Q_1 Y &\left\{ = \sum_i (y_i - \bar{y})^2 \right\} = Y' Q_1 Y \\ &= (\hat{Y} + \hat{U})' Q_1 (\hat{Y} + \hat{U}) = \hat{Y}' Q_1 \hat{Y} + \hat{U}' \hat{U} = \hat{Y}' Q_1 \hat{Y} + \sum_i \hat{u}_i^2 \end{aligned}$$

$$\begin{aligned} \text{because } \hat{Y}' Q_1 \hat{U} &= b'_{lse} X' Q_1 \hat{U} = b'_{lse} X' \hat{U} = b'_{lse} X' Q_X Y = 0 \text{ for} \\ &X' Q_X = (Q_X X)' = 0. \end{aligned}$$

The last line also implies $\hat{Y}'Q_1Y = \hat{Y}'Q_1(\hat{Y} + \hat{U}) = \hat{Y}'Q_1\hat{Y}$. The key point of this display is the well-known decomposition

$$(Y'Q_1Y =) \sum_i (y_i - \bar{y})^2 = \underbrace{\hat{Y}'Q_1\hat{Y}}_{\text{explained (by } x) \text{ variation}} + \underbrace{\sum_i \hat{u}_i^2}_{\text{unexplained variation}}.$$

R^2 is defined as the ratio of the explained variation to the total variation:

$$\begin{aligned} R^2 &\equiv \frac{\hat{Y}'Q_1\hat{Y}}{Y'Q_1Y} = \frac{\hat{Y}'Q_1Y \cdot \hat{Y}'Q_1\hat{Y}}{\hat{Y}'Q_1Y \cdot Y'Q_1Y} = \frac{\hat{Y}'Q_1Y \cdot \hat{Y}'Q_1\hat{Y}}{\hat{Y}'Q_1\hat{Y} \cdot Y'Q_1Y} \\ &= \frac{\{\sum_i (\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})\}^2}{\sum_i (\hat{y}_i - \bar{\hat{y}})^2 \cdot \sum_i (y_i - \bar{y})^2} = (\text{sample correlation of } Y \text{ and } \hat{Y})^2 \end{aligned}$$

R^2 falls in $[0, 1]$, being a squared correlation.

EXAMPLE: HOUSE SALE. A data set of size 467 was collected from the State College District in Pennsylvania for year 1991. State College is a small college town with the population of about 50,000. The houses sold during the year were sampled, and the sale prices and the durations until sale since the first listing in the market were recorded.

The dependent variable is the discount (DISC) percentage defined as 100 times the natural log of list price (LP) over sale price (SP) of a house:

$$100 \cdot \ln\left(\frac{LP}{SP}\right) = 100 \cdot \ln\left(1 + \frac{LP - SP}{SP}\right) \simeq 100\left(\frac{LP - SP}{SP}\right) = \text{discount \%}.$$

LP and SP are measured in \$1000. Since LP is the initial list price, given LP, explaining DISC is equivalent to explaining SP. The following is the list of regressors—the measurement units should be kept in mind: the number of days on the market until sold (T), years built minus 1900 (YR), number of rooms (ROOM), number of bathrooms (BATH), dummy for heating by electricity (ELEC), property tax in \$1,000 (TAX), dummy for spring listing (L1), summer listing (L2), and fall listing (L3), sale-month interest rate in % (RATE), dummy for sale by a big broker (BIGS), and number of houses on the market divided by 100 in the month when the house is listed (SUPPLY).

In Table 1, examine only the first three columns for a while. $\ln(T)$ appears before 1, because $\ln(T)$ is different from the other regressors—it is determined nearly simultaneously with DISC—and thus needs a special attention. Judging from the t-values in “tv-het,” most regressors are statistically significant at 5% level, for their absolute t-values are greater than 1.96; “tv-ho” will be used in the next subsection where the qualifiers “het” and “ho” will be explained. A longer $\ln(T)$ implies the bigger DISC: with $\partial \ln T \simeq \partial T/T$, an increase of $\partial \ln T = 1$ (i.e., 100% increase in T) means 4.6% increase in DISC, which in turn means 1% increase in T leading to

Table 1: LSE for House-Sale Discount %

	b_{lse}	tv-het (tv-ho)	b_{lse} (T)	tv-het (tv-ho) (T)
$\ln(T)$	4.60	7.76 (12.2)	0.027	8.13 (13.9)
1	-2.46	-0.23 (-0.24)	8.73	0.82 (0.86)
BATH	0.11	0.18 (0.17)	0.31	0.51 (0.51)
ELEC	1.77	2.46 (2.60)	1.84	2.67 (2.80)
ROOM	-0.18	-0.67 (-0.71)	-0.26	-0.95 (-1.04)
TAX	-1.74	-1.28 (-1.65)	-1.92	-1.49 (-1.88)
YR	-0.15	-3.87 (-5.96)	-0.15	-4.11 (-6.17)
$\ln(LP)$	6.07	2.52 (3.73)	5.71	2.51 (3.63)
BIGS	-2.15	-2.56 (-3.10)	-1.82	-2.25 (-2.72)
RATE	-2.99	-3.10 (-3.25)	-2.12	-2.31 (-2.41)
SUPPLY	1.54	1.06 (1.02)	1.89	1.36 (1.30)
s_N, R^2	6.20, 0.34		5.99, 0.39	

Variable	Mean	SD
DISC	7.16	7.64
L1	0.29	0.45
L2	0.31	0.46
L3	0.19	0.39
SP	115	57.7
T	188	150
BATH	2.02	0.67
ELEC	0.52	0.50
ROOM	7.09	1.70
TAX	1.38	0.65
YR	73.0	15.1
LP	124	64.9
BIGS	0.78	0.42
RATE	9.33	0.32
SUPPLY	0.62	0.19

0.046% increase in DISC. A newer house commands the less DISC: one year newer causes 0.15% less DISC, and thus 10 year newer causes 1.5% less DISC. A higher RATE means the lower DISC (1% increase in RATE causing 2.99% DISC drop); this finding seems, however, counter-intuitive, because a higher mortgage rate means the lower demand for houses. $R^2 = 0.34$ shows that 34% of the DISC variance is explained by $x'b_{lse}$, and $s_N = 6.20$ shows that about 95% of u_i 's fall in the range $\pm 1.96 \times 6.20$ if u_i 's follow $N(0, V(u))$.

As just noted, 1% increase in T causes 0.046% increase in DISC. Since this may not be easy to grasp, T is used instead of $\ln(T)$ for the LSE in the last two columns of the table. The estimate for T is significant with the

magnitude 0.027, meaning that 100 day increase in T leads to 2.7% DISC increase, which seems reasonable. This kind of query—whether the popular logged variable $\ln(T)$, level T , or some other function of T should be used—will be addressed later when we deal with “transformation of variables” in nonlinear models.

EXAMPLE: INTEREST RATE. As another example of LSE, we use time-series data on three month US treasury bill rates monthly from 01/1982 to 12/1999 ($N = 216$). To see what extent the past interest rates can explain the current one, the LSE of y_i on 1, y_{i-1} and y_{i-2} was done (since two lags are used, the sample size becomes $N = 214$) with the following result:

$$\begin{array}{lcl} y_i & = & 0.216 + 1.298 \cdot y_{i-1} - 0.337 \cdot y_{i-2}, \quad s_N = 0.304, \\ \text{t-values:} & & (2.05) \quad (10.29) \quad (-2.61) \\ & & R^2 = 0.980. \end{array}$$

Both y_{i-1} and y_{i-2} are statistically significant; i.e., $H_0 : \beta_2 = 0$ and $H_0 : \beta_3 = 0$ are rejected. The R^2 indicates that the two past rates predict very well the current rate. Since the unit of measurements are the same across the regressors and dependent variable, there is no complication in interpreting the estimates as in the house-sale example.

One curious point is that the estimate for y_{i-2} is significantly negative and differs too much from the estimate for y_{i-1} , casting some doubt over the linear model. One reason could be the “truncation bias”: the other lagged regressors (y_{i-3}, y_{i-4}, \dots) were omitted from the regressors to become part of u_i , which means $COR(y_{i-1}, u_i) \neq 0$ and $COR(y_{i-2}, u_i) \neq 0$, violating the basic tenet of LSE. One counter argument, however, is $COR(\hat{u}_i, \hat{u}_{i-1}) = 0.104$ which means that $COR(u_i, u_{i-1})$ would not be far from zero. If omitting y_{i-3}, y_{i-4}, \dots really matters, then one would expect $COR(\hat{u}_i, \hat{u}_{i-1})$ to be higher than 0.104. Having $COR(\hat{u}_i, \hat{u}_{i-1}) = 0.104$ is also comforting for the iid assumption for u_i ’s. This data as well as the house sale data will be used again.

1.5 Partial Regression

Suppose x consists of two sets of regressors of dimension $k_f \times 1$ and $k_g \times 1$, respectively: $x = (x'_f, x'_g)'$ and $k = k_f + k_g$. Partition X and b_{lse} accordingly:

$$X = [X_f, X_g] \text{ and } b_{lse} = \begin{bmatrix} b_f \\ b_g \end{bmatrix} \implies X b_{lse} = X_f b_f + X_g b_g.$$

X_f can be written as

$$X_f = \underset{N \times k_f}{X} \cdot \underset{k \times k_f}{S_f}$$

where S_f is a “selection matrix” consisting only of 1’s and 0’s to select the components of X for X_f ; analogously we can get $X_g = X \cdot S_g$. For example,

with $N = 3$, $k = 3$, and $k_f = 2$, the preceding display is

$$\begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \end{bmatrix} \times \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}.$$

Observe

$$\begin{aligned} P_{X_f} P_X &= P_{X_f} \text{ and } Q_{X_f} Q_X = Q_X \text{ because} \\ P_{X_f} P_X &= X_f (X_f' X_f)^{-1} X_f' \cdot X (X' X)^{-1} X' \\ &= X_f (X_f' X_f)^{-1} S_f' X' \cdot X (X' X)^{-1} X' \\ &= X_f (X_f' X_f)^{-1} S_f' X' = X_f (X_f' X_f)^{-1} X_f' = P_{X_f}, \\ Q_{X_f} Q_X &= (I_N - P_{X_f})(I_N - P_X) = I_N - P_{X_f} - P_X + P_{X_f} = Q_X. \end{aligned}$$

In words, for $P_{X_f} P_X = P_{X_f}$, extracting first the X part (with P_X) and then its subset X_f part (with P_{X_f}) is the same as extracting only the X_f part. As for $Q_{X_f} Q_X = Q_X$, removing first the X part and then its subset X_f part is the same as removing the whole X part.

Multiply $Y = X_f b_f + X_g b_g + \hat{U}$ by Q_{X_f} to get

$$\begin{aligned} Q_{X_f} Y &= Q_{X_f} X_f b_f + Q_{X_f} X_g b_g + Q_{X_f} \hat{U} = Q_{X_f} X_g b_g + \hat{U}, \text{ because} \\ Q_{X_f} X_f &= O \text{ and } Q_{X_f} \hat{U} = Q_{X_f} (Q_X Y) = Q_X Y = \hat{U}. \end{aligned}$$

Multiply both sides of $Q_{X_f} Y = Q_{X_f} X_g b_g + \hat{U}$ by $X_g' Q_{X_f}$ to get

$$X_g' Q_{X_f} Q_{X_f} Y = X_g' Q_{X_f} Q_{X_f} X_g \cdot b_g + X_g' Q_{X_f} \hat{U}.$$

Because

$$\begin{aligned} X_g' Q_{X_f} \hat{U} &= X_g' Q_{X_f} Q_X Y = X_g' Q_X Y = S_g' X' Q_X Y = 0 \text{ for} \\ X' Q_X &= 0_{k \times N}, \end{aligned}$$

the residual term disappears. Solving for b_g gives

$$b_g = (X_g' Q_{X_f} X_g)^{-1} X_g' Q_{X_f} Y.$$

This expression shows that the LSE b_g for β_g can be obtained in two stages. First, do the LSE of Y on X_f to get the *partial residual* $Q_{X_f} Y$, and then do the LSE of X_g on X_f to get the partial residual $Q_{X_f} X_g$. Second, do the LSE of $Q_{X_f} Y$ on $Q_{X_f} X_g$:

$$(X_g' Q_{X_f} Q_{X_f} X_g)^{-1} X_g' Q_{X_f} Q_{X_f} Y = (X_g' Q_{X_f} X_g)^{-1} X_g' Q_{X_f} Y.$$

This is the *partial regression* interpretation of b_g . The name “partial residual” is appropriate, for only the x_f part of x is used in the first regression. By using only the residuals in the second step, the presence of x_f is nullified, and

thus b_g shows the effect of x_g on y with x_f controlled for. Put it differently, b_g shows the additional explanatory power of x_g for y , over and above what is already explained by x_f . When x_g is a scalar, it is informative to plot $Q_{X_f}Y$ (on the vertical axis) versus $Q_{X_f}X_g$ (on the horizontal axis) to isolate the effect of x_g on y . The correlation between the two residuals is called the *partial correlation* between y and x_g .

As a special case, suppose $x_f = 1$ and $x_g = (x_2, \dots, x_k)'$. Denoting Q_{X_f} simply as Q_1 , we already saw

$$Q_1Y = (y_1 - \bar{y}, \dots, y_N - \bar{y})' \quad \text{and} \quad \begin{matrix} Q_1 & X_g \\ N \times N & N \times (k-1) \end{matrix} = (x_{1g} - \bar{x}_g, \dots, x_{Ng} - \bar{x}_g)'.$$

Using the vector notations, the partial regression for the slopes b_g is nothing but the LSE with the mean-deviation variables: with $x_i = (1, \tilde{x}_i)'$ and $\tilde{x} \equiv N^{-1} \sum_i \tilde{x}_i$,

$$b_g = \left\{ \sum_i \left(\tilde{x}_i - \tilde{x} \right) \left(\tilde{x}_i - \tilde{x} \right)' \right\}^{-1} \sum_i \left(\tilde{x}_i - \tilde{x} \right) (y_i - \bar{y}).$$

The role of “1” is to explain the level of (\tilde{x} and) y .

1.6 Omitted Variable Bias

In the model $y = x'_f\beta_f + x'_g\beta_g + u$, what happens if x_g is not used in estimation? This is an important issue, as we may not have (or use) all relevant regressors in the data. With x_g not used, $x'_g\beta_g + u \equiv v$ becomes the new error term in the model, and the consequence of not using x_g depends on $COR(x_f, x_g)$. To simplify the discussion, assume that the model is written in mean-deviation form, i.e., $E(y) = E(x'_f)\beta_f + E(x'_g)\beta_g + E(u)$ is subtracted from the model to yield

$$y - E(y) = \{x_f - E(x_f)\}'\beta_f + \{x_g - E(x_g)\}'\beta_g + u - E(u)$$

and we redefine y as $y - E(y)$, x_f as $x_f - E(x_f)$ and so on. So long as we are interested in slopes in β_f , the mean deviation model is adequate.

If $COR(x_f, x_g) = 0$ (i.e., if $E(x_fx_g) = 0$), then β_f can still be estimated consistently by the LSE of y on x_f . The only downside is that, in general, $SD(v) > SD(u)$ as v has more terms than u , and thus R^2 will drop. If $COR(x_f, x_g) \neq 0$, however, then $COR(x_f, v) \neq 0$ makes x_f an *endogenous regressor* and the LSE becomes inconsistent. Specifically, the LSE of y on x_f is

$$\begin{aligned} b_f &= \left(\frac{1}{N} \sum_i x_{if} x'_{if} \right)^{-1} \frac{1}{N} \sum_i x_{if} y_i \\ &= \left(\frac{1}{N} \sum_i x_{if} x'_{if} \right)^{-1} \frac{1}{N} \sum_i x_{if} (x'_{if}\beta_f + v_i) \end{aligned}$$

$$\begin{aligned}
&= \beta_f + \left(\frac{1}{N} \sum_i x_{if} x'_{if} \right)^{-1} \frac{1}{N} \sum_i x_{if} v_i \\
&= \beta_f + \left(\frac{1}{N} \sum_i x_{if} x'_{if} \right)^{-1} \frac{1}{N} \sum_i x_{if} (x'_{ig} \beta_g + u_i) \\
&= \beta_f + \left(\frac{1}{N} \sum_i x_{if} x'_{if} \right)^{-1} \frac{1}{N} \sum_i x_{if} x'_{ig} \cdot \beta_g \\
&\quad + \left(\frac{1}{N} \sum_i x_{if} x'_{if} \right)^{-1} \frac{1}{N} \sum_i x_{if} u_i \\
&\quad \text{which is consistent for } \beta_f + E^{-1}(x_f x'_f) E(x_f x'_g) \cdot \beta_g.
\end{aligned}$$

The term other than β_f is called the *omitted variable bias*, which is 0 if either $\beta_g = 0$ (i.e., x_g is not omitted at all) or if $E^{-1}(x_f x'_f) E(x_f x'_g) = 0$ which is the population linear projection coefficient of regressing x_g on x_f . In simple words, if $COR(x_f, x_g) = 0$, then there is no omitted variable bias. When LSE is run on some data and if resulting estimates do not make sense intuitively, in most cases, the omitted variable bias formula will provide a good guide on what might have gone wrong.

One question that might arise when $COR(x_f, x_g) \neq 0$ is what happens if a subvector x_{f2} of x_f is correlated to x_g while the other subvector x_{f1} of x_f is not where $x_f = (x'_{f1}, x'_{f2})'$. In this case, will x_{f1} still be subject to the omitted variable bias? The answer depends on $COR(x_{f1}, x_{f2})$ as can be seen in

$$\begin{aligned}
E^{-1}(x_f x'_f) E(x_f x'_g) &= \begin{bmatrix} E(x_{f1} x'_{f1}) & E(x_{f1} x'_{f2}) \\ E(x_{f2} x'_{f1}) & E(x_{f2} x'_{f2}) \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ E(x_{f2} x'_g) \end{bmatrix} \\
&\quad \text{as } E(x_{f1} x'_g) = 0 \\
&= \begin{bmatrix} 0 \\ E^{-1}(x_{f2} x'_{f2}) E(x_{f2} x'_g) \end{bmatrix} \text{ if } E(x_{f1} x'_{f2}) = 0.
\end{aligned}$$

Hence if $E(x_{f1} x'_{f2}) = 0$, then there is no omitted variable bias for x_{f1} . Otherwise, the bias due to $E(x_{f2} x'_g) \neq 0$ gets channeled to x_{f1} through $COR(x_{f1}, x_{f2})$.

In the case $COR(x_{f1}, x_{f2}) = 0$, $COR(x_{f1}, x_g) = 0$ but $COR(x_{f2}, x_g) \neq 0$, we can in fact use only x_{f1} as regressors—no omitted variable bias in this case. Nevertheless, using x_{f2} as regressors makes the model error term variance smaller, which leads to a higher R^2 and higher t-values for x_{f1} . In this case, we just have to be aware that the estimator for x_{f2} is biased.

As an example for omitted variable bias, imagine a state considering a mandatory seat belt law. Data is collected from N cities in the state, with y_i the yearly traffic fatality proportion per driver in city i , and x_{if} the proportion of drivers wearing seat belt in city i . LSE is run to find $b_f > 0$, which is counter-intuitive however. One possible scenario is that wearing the

seat belt makes the driver go faster, which results in more accidents. That is, driving speed x_g in the error term is correlated with x_f , and the omitted variable bias dominates β_f so that the following sum becomes positive:

$$\underset{\text{negative}}{\beta_f} + \underset{\text{positive}}{E^{-1}(x_f x'_f)E(x_f x'_g)} \cdot \underset{\text{positive}}{\beta_g}$$

In this case, enacting the seat belt law will increase y , not because $\beta_f > 0$ but rather because it will cause x_g to increase.

What the state have in mind is the *ceteris paribus* (“direct”) effect β_f with all the other variables held constant, but what is estimated is the total effect that is the sum of the direct effect β_f and the *indirect effect* of x_f on y through x_g . Both the direct and indirect effects can be estimated consistently using the LSE of y on x_f and x_g , but enacting only the seat belt law will not have the intended effect because the indirect effect will occur. A solution is enacting both the seat belt law and a speed limit law to assure $COR(x_f, x_g) = 0$ after the laws are passed.

In the example, omitted variable bias helped explaining an apparently nonsensical result. But it can also help negating an apparently plausible result. Suppose that there are two types of people, one cautious and the other reckless, with x_g denoting the proportion of the cautious people, and that the cautious people tend to wear seat belts more ($COR(x_f, x_g) > 0$) and have fewer traffic accidents. Also suppose $\beta_f = 0$, i.e., no true effect of seat belt wearing. In this case, the LSE of y on x_f converges to a negative number

$$\underset{0}{\beta_f} + \underset{\text{positive}}{E^{-1}(x_f x'_f)E(x_f x'_g)} \cdot \underset{\text{negative}}{\beta_g}$$

and, due to omitting x_g , we may wrongly conclude that wearing seat belt will lower y to enact the seat belt law. Here the endogeneity problem of x_f leads to an ineffective policy as the seat belt law will have no true effect on y . Note that, differently from the $x_g = \text{speed}$ example, there is no indirect effect of forcing seat belt wearing because seat belt wearing will not change the people’s type.

2 Heteroskedasticity and Homoskedasticity

The assumption $E(u|x) = 0$ for LSE is a restriction on the conditional first moment of $u|x$. We do not need restrictions on higher moments of $u|x$ to estimate β , but whether $E(u^2|x)$ varies or not as x changes matters in the LSE asymptotic inference, which is the topic of this section. $E(u^2|x)$ will also appear prominently later for generalized LSE (GLS).

Observe that $E(u^2|x) = V(u|x)$ because $E(u|x) = 0$, and also that $V(u|x) = V(y|x)$ because $y|x$ is a $x'\beta$ -shifted version of $u|x$. If $V(u|x)$ is a non-constant function of x , then u is “heteroskedastic” (or there is “heteroskedasticity”). If $V(u|x)$ is a constant, say σ^2 , then u is “homoskedastic”

(or there is “*homoskedasticity*”). Although we assume that (u_i, x'_i) are iid across i , $u_i|x_i$ are not iid across i under heteroskedasticity.

2.1 Heteroskedasticity Sources

2.1.1 Forms of Heteroskedasticity

A well-known source for heteroskedasticity is *random coefficients*. Suppose the coefficient vector is β_i that is random around a constant β :

$$y_i = x'_i\beta_i + u_i, \quad \beta_i = \beta + v_i, \quad E(v) = 0, \quad E(vv') \equiv \Lambda, \\ v \text{ is independent of } x \text{ and } u.$$

Substituting the β_i equation yields a constant coefficient model:

$$y_i = x'_i\beta + (x'_iv_i + u_i), \quad E(x'v + u|x) = 0, \quad V(x'v + u|x) = x'\Lambda x + E(u^2|x).$$

Even if $E(u^2|x) = \sigma^2$, still the error term $\varepsilon \equiv x'v + u$ is heteroskedastic. Here the functional form of $V(\varepsilon|x) = x'\Lambda x + \sigma^2$ is known (up to Λ and σ^2) due to the random coefficients and the homoskedasticity of u .

Heteroskedasticity does not necessarily have to be motivated by random coefficients. If x is income and y is consumption per month, we can simply imagine the variation of $y|x$ increasing as x increases. In this case, we may postulate, say,

$$y_i = x'_i\beta + u_i, \quad V(u|x_i) = \exp(x'_i\theta), \\ \text{where } \theta \text{ is an unknown parameter vector;}$$

again, this is heteroskedasticity *of known form* as in the random coefficient model.

The linear model assumption $E(y|x) = x'\beta$ is restrictive because $E(y|x)$ may not be a linear function. Assuming $V(y|x) = \exp(x'\theta)$ additionally is even more restrictive, in which we would have even less confidence than in $E(y|x) = x'\beta$. If we just allow $V(u|x) = V(y|x)$ to be an unknown function of x instead of specifying the functional form of $V(u|x)$, then we allow for a heteroskedasticity *of unknown form*. The consistency and asymptotic distribution results of the LSE hold under heteroskedasticity of unknown form, because we did not impose any assumption on $V(y|x)$ in their derivations.

If $V(u|x) = \sigma^2$, then because

$$E(xx'u^2) = E\{xx'E(u^2|x)\} = \sigma^2 E(xx'),$$

the asymptotic variance of $\sqrt{N}(b_{lse} - \beta)$ is

$$E^{-1}(xx')E(xx'u^2)E^{-1}(xx') = \sigma^2 \cdot E^{-1}(xx')$$

which appears often in introductory econometrics textbooks. The right-hand side (rhs) is valid only under homoskedasticity; the left-hand side (lhs) is called a “*heteroskedasticity-robust (or -consistent) variance*,” which is valid with or without the homoskedasticity assumption.

2.1.2 Heteroskedasticity due to Aggregation

When “averaging with different numbers of observations” takes place, heteroskedasticity can arise without x getting involved. Suppose a model holds at individual level, but we have only a city-level aggregate data:

$$\begin{aligned} y_{j_i} &= x'_{j_i}\beta + u_{j_i}, \quad j_i = 1, \dots, n_i \text{ where } j_i \text{ denotes individual} \\ &\quad j \text{ in city } i = 1, \dots, N \\ \implies y_i &= x'_i\beta + u_i \quad \text{where } y_i \equiv \frac{1}{n_i} \sum_{j_i} y_{j_i}, \quad x_i \equiv \frac{1}{n_i} \sum_{j_i} x_{j_i}, \\ u_i &\equiv \frac{1}{n_i} \sum_{j_i} u_{j_i}. \end{aligned}$$

That is, what is available is a random sample on cities with (n_i, x'_i, y_i) , $i = 1, \dots, N$, where n_i is the total number of people in city i and N is the number of the sampled cities. Suppose that u_{j_i} is independent of x_{j_i} , and that u_{j_i} 's are iid with zero mean and variance σ^2 (i.e., $u_{j_i} \sim (0, \sigma^2)$). Then u_1, \dots, u_N are independent, and $u_i | (x_i, n_i) \sim (0, \sigma^2/n_i)$: the error terms in the city-level model are heteroskedastic wrt n_i , but not wrt x_i . Note that all of n_i , x_i , and y_i are random as we do not know which city gets drawn.

This type of heteroskedasticity can be dealt with by minimizing $\sum_i (y_i - x'_i\beta)^2 n_i$, which is equivalent to applying LSE to the transformed equation

$$y_i^* = x_i^{*\prime}\beta + u_i^*, \quad \text{where } y_i^* \equiv y_i\sqrt{n_i}, \quad x_i^* \equiv x_i\sqrt{n_i} \quad \text{and} \quad u_i^* \equiv u_i\sqrt{n_i}.$$

In the transformed equation, as (x_i, n_i) is “finer” than x_i^* ,

$$\begin{aligned} E(u_i^* | x_i^*) &= E\{ E(u_i\sqrt{n_i} | x_i, n_i) | x_i^* \} = 0 \quad \text{as } E(u_i\sqrt{n_i} | x_i, n_i) = 0, \\ V(u_i^* | x_i^*) &= E\{ V(u_i\sqrt{n_i} | x_i, n_i) | x_i^* \} = \sigma^2 \quad \text{as } V(u_i\sqrt{n_i} | x_i, n_i) = \sigma^2. \end{aligned}$$

Hence u_1^*, \dots, u_N^* are iid $(0, \sigma^2)$. This LSE motivates “weighted LSE” to appear later.

Two remarks on the city-level data example. First, there is no unity in the transformed regressors because 1 is replaced with $\sqrt{n_i}$. This requires a different definition of R^2 which was defined using $Q_1\hat{U} = \hat{U}$. R^2 for the transformed model can be defined as $\{\text{sample } COR(y, \hat{y})\}^2$, not as $\{\text{sample } COR(y^*, \hat{y}^*)\}^2$, where $\hat{y}_i = x'_i b_{lse}^*$, $\hat{y}_i^* = x_i^{*\prime} b_{lse}^*$, and b_{lse}^* is the LSE for the transformed model. This definition of R^2 can also be used for “weighted LSE” below. Second, we assumed above that sampling is done at city level and what is available is the averaged variables y_i and x_i along with n_i . If, instead, all cities are included but n_i individuals get sampled in city i where n_i is a pre-determined (i.e., fixed) constant ahead of sampling, then n_i is not random (but still may vary across i); in contrast, (x'_i, y_i) is still random because it depends on the sampled individuals. In this case, u_i 's are *independent but non-identically distributed (inid)* due to $V(u_i) = \sigma^2/n_i$ where $V(u_i)$ is the marginal variance of u_i . Clearly, how sampling is done matters greatly.

2.1.3 Variance Decomposition

Observe

$$\begin{aligned}
 V(u) &= E(u^2) - \{E(u)\}^2 = E[E(u^2|x)] - [E\{E(u|x)\}]^2 \\
 &= E[V(u|x) + \{E(u|x)\}^2] - [E\{E(u|x)\}]^2 \\
 &= E[V(u|x)] + E[g(x)^2] - [E\{g(x)\}]^2 \quad (\text{defining } g(x) \equiv E(u|x)) \\
 &= E\{V(u|x)\} + V\{g(x)\} \quad (\text{in general}) \\
 \{ &= E\{V(u|x)\} \quad \text{as } g(x) = E(u|x) = 0\}.
 \end{aligned}$$

Under homoskedasticity, $V(u|x) = \sigma^2 \forall x$, and thus $V(u) = E(\sigma^2) = \sigma^2$.

For a rv y , this display gives the *variance decomposition* of $V(y)$

$$V(y) = E\{V(y|x)\} + V\{E(y|x)\}$$

which can help understand the sources of $V(y)$. Suppose that x is a rv taking on 1, 2, or 3. Decompose the population with x into 3 groups (i.e., subpopulations):

Group	$P(x = 1) = 1/2$	$P(x = 2) = 1/4$	$P(x = 3) = 1/4$
Group mean (level)	$E(y x = 1)$	$E(y x = 2)$	$E(y x = 3)$
(Within-) Group Variance	$V(y x = 1)$	$V(y x = 2)$	$V(y x = 3)$

Each group has its conditional variance, and we may be tempted to think that $E\{V(y|x)\}$ which is the weighted average of $V(y|x)$ with $P(y|x)$ as the weight yields the marginal variance $V(y)$. But the variance decomposition formula demonstrates $V(y) \neq E\{V(y|x)\}$ unless $E(y|x) = 0 \forall x$, although $E(y) = E\{E(y|x)\}$ always. That is, the source of the variance is not just the “*within-group variance*” $V(y|x)$, but also the “*between-group variance*” $V\{E(y|x)\}$ of the group mean $E(y|x)$.

If the variance decomposition is done with an observable variable x , then we may dig deeper by estimating $E(y|x)$ and $V(y|x)$. But a decomposition with an unobservable variable u can be also thought of, as we can choose any variable we want in the variance decomposition: $V(y) = E\{V(y|u)\} + V\{E(y|u)\}$. In this case, the decomposition can help us imagine the sources depending on u . For instance, if y is income and u is ability (whereas x is education), then the income variance is the weighted average of ability-group variances plus the variance between the average group-incomes.

Two polar cases are of interest. Suppose that y is income and x is education group: 1 for “below high school graduation,” 2 for “high school graduation” to “below college graduation,” and 3 for college graduation or above. One extreme case is the same mean income for all education groups:

$$E(y|x) = E(y) \forall x \implies V\{E(y|x)\} = 0 \implies V(y) = E\{V(y|x)\}.$$

The other extreme case is the same variance in each education group:

$$V(y|x) = \sigma^2 \quad \forall x \implies E\{V(y|x)\} = \sigma^2 \implies V(y) = \sigma^2 + V\{E(y|x)\};$$

if $\sigma^2 = 0$, then $V(y) = V\{E(y|x)\}$: the variance comes solely from the differences of $E(y|x)$ across the groups.

2.1.4 Analysis of Variance (ANOVA)*

The variance decomposition formula is the basis for *Analysis of Variance* (ANOVA), where x stands for treatment categories (with one category being no treatment). In ANOVA, the interest is on the “mean treatment effect,” i.e., whether $E(y|x)$ changes across the treatment groups/categories or not. The classical approach—*one-way ANOVA*—assumes normality for y and homoskedasticity across the groups ($V(y|x) = \sigma^2 \quad \forall x$) so that $V(y) = \sigma^2 + V\{E(y|x)\}$. One-way ANOVA decomposes the sample variance into sample versions of σ^2 and $V\{E(y|x)\}$ which are two independent χ^2 rv’s. “ $H_0 : E(y|x)$ is a constant $\forall x$ ” is tested using the ratio of the two sample versions, and the ratio follows a F -distribution.

Let $x^{(j)}$ denote the x -value for group j , and define the group- j mean $\mu_j \equiv E(y|x = x^{(j)})$. In one-way ANOVA, y gets indexed as in y_{ij} , $i = 1, \dots, N_j$, $j = 1, \dots, J$ where j denotes the j th group (category) and i denotes the i th observation in the j th group; there are N_j observations in group j . The model is

$$y_{ij} = \mu_j + u_{ij}, \quad u_{ij} \sim iid N(0, \sigma^2) \text{ across } i \text{ and } j.$$

Define the total sample size, group- j average and the “grand average” as, respectively,

$$N \equiv \sum_{j=1}^J N_j, \quad \bar{y}_j \equiv \frac{1}{N_j} \sum_{i=1}^{N_j} y_{ij} \quad \bar{y} \equiv \frac{1}{N} \sum_{j=1}^J \sum_{i=1}^{N_j} y_{ij}.$$

Then the decomposition $y_{ij} - \bar{y} = (y_{ij} - \bar{y}_j) + (\bar{y}_j - \bar{y})$ is used in one-way ANOVA where $\bar{y}_j - \bar{y}$ is for $V\{E(y|x)\}$.

Specifically, take $\sum_{j=1}^J \sum_{i=1}^{N_j} (y_{ij} - \bar{y})^2 = \{(y_{ij} - \bar{y}_j) + (\bar{y}_j - \bar{y})\}^2$ to see that the cross-product term is zero because

$$\begin{aligned} \sum_{j=1}^J \sum_{i=1}^{N_j} (y_{ij} - \bar{y}_j)(\bar{y}_j - \bar{y}) &= \sum_{j=1}^J (\bar{y}_j - \bar{y}) \sum_{i=1}^{N_j} (y_{ij} - \bar{y}_j) \\ &= \sum_{j=1}^J (\bar{y}_j - \bar{y})(N_j \bar{y}_j - N_j \bar{y}_j) = 0. \end{aligned}$$

Thus we get

$$\begin{array}{ccc} \sum_{j=1}^J \sum_{i=1}^{N_j} (y_{ij} - \bar{y})^2 &= \sum_{j=1}^J \sum_{i=1}^{N_j} (y_{ij} - \bar{y}_j)^2 + \sum_{j=1}^J N_j (\bar{y}_j - \bar{y})^2 \\ \text{total variation} &\quad \text{unexplained variation} \quad \text{explained variation} \end{array}$$

where the two terms on the rhs are for $\sigma^2 + V\{E(y|x)\}$ when divided by N .
The aforementioned test statistic for mean equality is

$$\frac{(J-1)^{-1} \sum_{j=1}^J N_j (\bar{y}_j - \bar{y})^2}{(N-J)^{-1} \sum_{j=1}^J \sum_{i=1}^{N_j} (y_{ij} - \bar{y}_j)^2} \sim F(J-1, N-J).$$

To understand the dof's, note that there are J -many “observations” (\bar{y}_j 's) in the numerator, and 1 is subtracted in the dof because the grand mean gets estimated by \bar{y} . In the denominator, there are N -many observations y_{ij} 's, and J is subtracted in the dof because the group means get estimated by \bar{y}_j 's. Under the H_0 , the test statistic is close to zero as the numerator is so because of $V\{E(y|x)\} = 0$.

The model $y_{ij} = \mu_j + u_{ij}$ can be rewritten as a familiar linear model. Define $J-1$ dummy variables, say x_{i2}, \dots, x_{iJ} , where $x_{ij} = 1$ if observation i belongs to group j and $x_{ij} = 0$ otherwise. Then

$$y_i = x_i' \beta + u_i, \quad \text{where } x_i = (1, x_{i2}, \dots, x_{iJ})' \text{ and} \\ \beta = (\mu_1, \mu_2 - \mu_1, \dots, \mu_J - \mu_1)'.$$

Here the intercept is for μ_1 and the slopes are for the deviations from μ_1 ; group 1 is typically the “control (i.e., no-treatment) group” whereas the other groups are the “treatment groups.” For instance, if observation i belongs to treatment group 2, then

$$x_i' \beta = (1, 1, 0, \dots, 0)' \beta = \mu_1 + (\mu_2 - \mu_1) = \mu_2.$$

Instead of the above F -test, we can test for $H_0 : \mu_1 = \dots = \mu_J$ with “Wald test” to appear later without assuming normality; the Wald test checks out whether all slopes are zero or not.

“Two-way ANOVA” generalizes one-way ANOVA. There are two “factors” now, and we get y_{ijk} where j and k index the group (j, k), $j = 1, \dots, J$ and $k = 1, \dots, K$; group (j, k) has N_{jk} observations. The model is

$$y_{ijk} = \alpha_j + \beta_k + \gamma_{jk} + u_{ijk}, \quad u_{ijk} \sim N(0, \sigma^2) \text{ iid across all indices}$$

where α_j is the factor-1 effect, β_k is the factor-2 effect, and γ_{jk} is the interaction effect between the two factors. The relevant decomposition is

$$y_{ijk} - \bar{y} = (y_{ijk} - \bar{y}_j - \bar{y}_{.k} + \bar{y}) + (\bar{y}_j - \bar{y}) + (\bar{y}_{.k} - \bar{y})$$

where \bar{y} is the grand mean, \bar{y}_j is the average of all observations with j fixed (i.e., $\bar{y}_j \equiv \sum_{k=1}^K \sum_{i=1}^{N_{jk}} y_{ijk} / \sum_{k=1}^K N_{jk}$), and $\bar{y}_{.k}$ is analogously defined. Various F -test statistics can be devised by squaring and summing up this display, but the two-way ANOVA model can be also written as a familiar linear model, to which “Wald tests” can be applied.

2.2 Weighted LSE (WLS)

Suppose $E(u^2|x) = V(u|x) = m'\theta$ where m consists of elements of x and functions of those, and suppose that we know this functional form; e.g., with $k = 4$,

$$m'_i\theta = \theta_1 + \theta_2 x_{i2} + \theta_3 x_{i3} + \theta_4 x_{i2}^2 + \theta_5 x_{i2} x_{i3}.$$

Then we can do “*Weighted LSE (WLS)*”:

- First, apply LSE to $y_i = x'_i\beta + u_i$ to get the residuals \hat{u}_i .
- Second, estimate θ by the LSE of \hat{u}_i^2 on m_i to get the LSE $\hat{\theta}$ for θ ; this is motivated by $E(u^2|x) = m'\theta$.
- Third, assuming $m'_i\hat{\theta} > 0$ for all m_i , estimate β again by minimizing the weighted minimand $N^{-1} \sum_i (y_i - x'_i b)^2 / (m'_i \hat{\theta})$ wrt b .

In Chapter 3.3.3, it will be shown that replacing θ with $\hat{\theta}$ is innocuous, and WLS is asymptotically equivalent to applying LSE to (with $SD(u|x_i) = (m'_i\theta)^{1/2}$)

$$\frac{y_i}{SD(u|x_i)} = \frac{x'_i}{SD(u|x_i)}\beta + \frac{u_i}{SD(u|x_i)}, \quad \text{where}$$

$$V\left\{\frac{u}{SD(u|x)}|x\right\} = \frac{V(u|x)}{SD(u|x)^2} = 1.$$

As in the above averaged data case, we can define $y_i^* \equiv y_i/SD(u|x_i)$ and $x_i^* \equiv x_i/SD(u|x_i)$. The error term in the transformed equation is homoskedastic with known variance 1. Inserting 1 and $x_i/SD(u|x_i)$, respectively, into σ^2 and x in $\sigma^2 E^{-1}(xx')$, we get

$$\sqrt{N}(b_{wls} - \beta) \rightsquigarrow N(0, E^{-1}\left\{\frac{xx'}{V(u|x)}\right\}).$$

The assumption $m'_i\hat{\theta} > 0$ for all m_i can be avoided if $V(u|x) = \exp(m'\theta)$ and if θ is estimated with “nonlinear LSE” that will appear later. The assumption $m'_i\hat{\theta} > 0$ for all m_i is simply to illustrate WLS using LSE in the first step.

An easy practical alternative to guarantee positive estimated weights is adopting a log-linear model $\ln u_i^2 = m'_i\zeta + v_i$ with v_i being an error term. The log-linear model is equivalent to

$$u_i^2 = e^{m'_i\zeta} e^{v_i} = (e^{m'_i\zeta/2} \nu_i)^2 \quad \text{where } \nu_i \equiv e^{v_i/2}$$

and $e^{m'_i\zeta/2}$ may be taken as the scale factor $SD(u|x_i)$ for ν_i (but $e^{m'_i\zeta/2}\nu_i > 0$ and thus the error u_i cannot be $e^{m'_i\zeta/2}\nu_i$ although $u_i^2 = (e^{m'_i\zeta/2}\nu_i)^2$). This suggests using $SD(u|x_i) \simeq e^{m'_i\hat{\zeta}/2}$ for WLS weighting where $\hat{\zeta}$ is the LSE for ζ . Strictly speaking, this “suggestion” is not valid because, for $SD(u|x_i) = e^{m'_i\zeta/2}$ to hold, we need

$$\ln E(u^2|x_i) = m'_i\zeta \iff E(u^2|x_i) = \exp(m'_i\zeta)$$

but $\ln u_i^2 = m'_i \zeta + v_i$ postulates instead $E(\ln u^2 | x_i) = m'_i \zeta$. Since $\ln E(u^2 | x_i) \neq E(\ln u^2 | x_i)$, $\ln u_i^2 = m'_i \zeta + v_i$ is not compatible with $SD(u | x_i) = e^{m'_i \zeta / 2}$. Despite this, however, defining $\hat{u}_i^* \equiv y_i^* - x_i^{*'} b_{wls}$ where b_{wls} is the WLS with weight $\exp(m'_i \hat{\zeta} / 2)$, so long as the LSE of \hat{u}_i^{*2} on m_i returns insignificant slopes, we can still say that the weight $\exp(m'_i \hat{\zeta} / 2)$ is adequate because the heteroskedasticity has been removed by the weight, no matter how it was obtained.

In short, each one of the following has different implications on how we go about LSE.

- heteroskedasticity of unknown form: LSE to use $E^{-1}(xx')$
 $E(xx'u^2)E^{-1}(xx')$
- homoskedasticity: LSE to use $\sigma^2 E^{-1}(xx')$
- heteroskedasticity of known form: WLS to use
 $E^{-1}\{xx'/V(u|x)\}$.

Under homoskedasticity, all three variance matrices agree; otherwise, they differ in general.

Later, we will see that, under the known form of heteroskedasticity, WLS is more efficient than LSE; i.e.,

$$E^{-1}(xx') E(xx'u^2) E^{-1}(xx') \geq E^{-1} \left\{ \frac{xx'}{V(u|x)} \right\}$$

in the matrix sense (for two matrices A and B , $A \geq B$ means that $A - B$ is p.s.d). For instance, if u_i is specified as

$$u_i = w_i \exp(x_i' \theta / 2), \text{ where } w_i \text{ is independent of } x_i \text{ with } V(w) = 1,$$

then $V(u|x) = \exp(x' \theta)$, and we can do WLS with this. This is also convenient in viewing y_i : y_i is obtained by generating x_i and w_i first and then summing up $x_i' \beta$ and $w_i \exp(x_i' \theta)$. But if the specified form of heteroskedasticity $\exp(x' \theta)$ is wrong, then the asymptotic variance of the WLS is no longer $E^{-1}\{xx'/V(u|x)\}$. So, it is safer to use LSE with heteroskedasticity-robust variance. From now on, we will not invoke homoskedasticity assumption, unless it gives helpful insights for the problem at hand, which does happen from time to time.

2.3 Heteroskedasticity Examples

EXAMPLE: HOUSE SALE (continued). In the preceding section, the t-values under heteroskedasticity of unknown form (tv-het) were shown along with the t-values under homoskedasticity (tv-ho). Comparing the two sets of t-values, the differences are small other than for $\ln(T)/T$ and YR, and tv-ho tends to be greater than tv-het. This indicates that the extent of heteroskedasticity would be minor, if any. Three courses of action are conceivable from this observation:

- First, test for the H_0 of homoskedasticity using a test, say, in White (1980). This test does the LSE of \hat{u}_i^2 on 1 and some polynomial functions of regressors to see if the slopes are all zero or not; all zero slopes mean homoskedasticity. $N \cdot R^2 \rightsquigarrow \chi^2_{\#slopes}$ can be used as an asymptotic test statistic where R^2 is the R^2 for the \hat{u}_i^2 equation LSE. If the null is not rejected, then tv-ho may be used.
- Second, if the null is rejected, then model the form of heteroskedasticity using the aforementioned forms (or some others) to do WLS, where the weighted error term $u/SD(u|x)$ should have variance one regardless of x (this can be checked out using the method in the first step).
- Third, instead of testing for the H_0 or modelling heteroskedasticity, simply use tv-het. This is the simplest and most robust procedure. Also, the gain in the above two procedures tends to be small in micro-data; see, e.g., Deaton (1995).

EXAMPLE: INTEREST RATE (continued). Recall the interest rate example:

$$\begin{array}{rclcl} y_i & = & 0.216 & + & 1.298 \quad \cdot y_{i-1} - \quad 0.337 \quad \cdot y_{i-2} \\ t\text{-}v\text{-}l\text{-}a\text{-}u\text{-}e\text{-}s: & & 2.05 \text{ (3.42)} & & 10.29 \text{ (21.4)} \quad \quad -2.61 \text{ (-5.66)} \end{array}$$

We list both tv-het and tv-ho; the latter is in (\cdot) and was computed with $b_{lse,j}/\sqrt{v_{N,jj}}$, $j = 1, \dots, k$, where $V_N \equiv [v_{N,hj}]$, $h, j = 1, \dots, k$, is defined as $s_N^2(\sum_i x_i x_i')^{-1}$. The large differences between the two types of t-values indicate that the homoskedasticity assumption would not be valid for this model. In this time-series data, if the form of heteroskedasticity is correctly modeled, the gain in significance (i.e., the gain in the precision of the estimators) would be substantial. Indeed, such modeling is often done in financial time-series.

3 Testing Linear Hypotheses

3.1 Wald Test

Suppose we have an estimator b_N with

$$\sqrt{N}(b_N - \beta) \rightsquigarrow N(0, C);$$

b_N is said to be a “ \sqrt{N} -consistent asymptotically normal estimator with asymptotic variance C .” LSE and WLS are two examples of b_N and more examples will appear later. Given b_N , often we want to test linear null hypotheses such as

$$H_0 : R'\beta = c, \quad \text{where } \text{rank}(R) = g,$$

R is a $k \times g$ ($g \leq k$) known constant matrix and c is a $g \times 1$ known constant vector. Since $b_N \xrightarrow{p} \beta$, we have $R'b_N \xrightarrow{p} R'\beta$, because $R'b$ is a continuous function of b . If $R'\beta = c$ is true, $R'b_N$ should be close to c . Hence testing for $R'\beta = c$ can be based on the difference $\sqrt{N}(R'b_N - c)$.

As an example of $R'\beta = c$, suppose $k = 4$ and $H_0 : \beta_2 = 0, \beta_3 = 2$. For this, set

$$R' = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \quad c = \begin{bmatrix} 0 \\ 2 \end{bmatrix}$$

to get $R'\beta = c$ equivalent to $\beta_2 = 0$ and $\beta_3 = 2$. If we want to add another hypothesis, say $\beta_1 - \beta_4 = 0$, then set

$$R' = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & -1 \end{bmatrix}, \quad c = \begin{bmatrix} 0 \\ 2 \\ 0 \end{bmatrix}.$$

Typically, we test for some chosen elements of β being zero jointly. In that case, R consists of the column vectors picking up the chosen elements of β (each column consists of $k - 1$ zeros and 1) and c is a zero vector.

Given the above C and R , define H and Λ such that

$$R'CR = H\Lambda H'.$$

H is a matrix whose g columns are orthonormal eigenvectors of $R'CR$ and Λ is the diagonal matrix of the eigenvalues; $H\Lambda H'$ exists because $R'CR$ is real and symmetric. By construction, $H'H = I_g$. Also, pre-multiplying $H'H = I_g$ by H to get $(HH')H = H$, we obtain $HH' = I_g$ because H is of full rank. Observe now

$$\begin{aligned} S &\equiv H\Lambda^{-0.5}H' \implies S'S = (R'CR)^{-1} \quad \text{because} \\ S'S &= H\Lambda^{-0.5}H'H\Lambda^{-0.5}H' = H\Lambda^{-1}H' \quad \text{and} \\ S'S \cdot R'CR &= H\Lambda^{-1}H' \cdot H\Lambda H' = I_g. \end{aligned}$$

Further observe

$$\begin{aligned} \sqrt{N} \cdot R'(b_N - \beta) &\rightsquigarrow N(0, R'CR) \\ \left\{ \text{from } \sqrt{N}(b_N - \beta) &\rightsquigarrow N(0, C) \text{ "times } R'' \right\}, \\ \sqrt{N}SR'(b_N - \beta) &\rightsquigarrow N(0, I_g), \quad \text{since} \\ S \cdot R'CR \cdot S' &= H\Lambda^{-0.5}H' \cdot H\Lambda H' \cdot H\Lambda^{-0.5}H' = I_g, \\ N(R'b_N - R'\beta)'S'S(R'b_N - R'\beta) & \\ = N(R'b_N - R'\beta)'(R'CR)^{-1}(R'b - R'\beta) &\rightsquigarrow \chi_g^2, \end{aligned}$$

because $N(R'b_N - R'\beta)'S'S(R'b_N - R'\beta)$ is a sum of g -many squared, asymptotically uncorrelated $N(0, 1)$ random variables (rv). Replacing $R'\beta$ with c under $H_0 : R'\beta = c$, we get a *Wald test* statistic

$$N(R'b_N - c)'(R'C_N R)^{-1}(R'b_N - c) \rightsquigarrow \chi_g^2 \quad \text{where } C_N \rightarrow^p C.$$

The matrix $(R'C_N R)^{-1}$ in the middle standardizes the vector $\sqrt{N}(R'b_N - c)$.

3.2 Remarks

When b_N is the LSE of y on x , we get

$$\begin{aligned} C &\equiv E^{-1}(xx')E(xx'u^2)E^{-1}(xx'), \\ C_N &= \left(\frac{1}{N} \sum_i x_i x_i'\right)^{-1} \cdot \frac{1}{N} \sum_i x_i x_i' \hat{u}_i^2 \cdot \left(\frac{1}{N} \sum_i x_i x_i'\right)^{-1} \\ [&= N(X'X)^{-1}X'DX(X'X)^{-1}, \quad \text{in matrices where} \\ &D \equiv \text{diag}(\hat{u}_1^2, \dots, \hat{u}_N^2) \end{aligned}$$

If homoskedasticity holds, then instead of C and C_N , we can use C_o and C_{oN} where

$$C_{oN} \equiv s_N^2 \left(\frac{1}{N} \sum_i x_i x_i'\right)^{-1} = s_N^2 \left(\frac{X'X}{N}\right)^{-1} \rightarrow^p C_o \equiv \sigma^2 E^{-1}(xx').$$

To show $C_N \rightarrow^p C$, since $(N^{-1} \sum_i x_i x_i')^{-1} \rightarrow^p E^{-1}(xx')$ was noted already, we have to show

$$\frac{1}{N} \sum_i x_i x_i' \hat{u}_i^2 - E(xx'u^2) = o_p(1).$$

Here, we take the “working proposition” that, for the expected value $E(h(x, y, \beta))$ where $h(x, y, \beta)$ is a (matrix-valued) function of x , y , and β , it holds in general that

$$\frac{1}{N} \sum_i h(x_i, y_i, b_N) - E(h(x, y, \beta)) = o_p(1), \quad \text{if } b_N \rightarrow^p \beta.$$

Then, setting $h(x, y, b) = xx'(y - x'b)^2$ establishes $C_N \rightarrow^p C$. For the preceding display to hold, $h(\cdot, \cdot, b)$ should not be too variable as a function of b so that the LLN holds uniformly over b . In almost all cases we encounter, the preceding display holds.

Instead of C_N , MacKinnon and White (1985) suggested to use, for a better small sample performance,

$$\begin{aligned} \tilde{C}_N &\equiv (N-1)(X'X)^{-1} \left(X' \tilde{D} X - \frac{X' \tilde{r} \tilde{r}' X}{N} \right) (X'X)^{-1}, \quad \text{where} \\ \tilde{D} &\equiv \text{diag}(\tilde{r}_1^2, \dots, \tilde{r}_N^2), \quad \tilde{r}_i \equiv \frac{y_i - x_i' b_{lse}}{1 - d_{ii}}, \quad \tilde{r} \equiv (\tilde{r}_1, \dots, \tilde{r}_N)', \quad \text{and} \\ &d_{ii} \text{ is the } i\text{th diagonal element of the matrix } X(X'X)^{-1}X'. \end{aligned}$$

\tilde{C}_N and C_N are asymptotically equivalent as the term $X' \tilde{r} \tilde{r}' X / N$ in \tilde{C}_N is of smaller order than $X' \tilde{D} X$.

Although the two variance estimators C_N and C_{No} numerically differ in finite samples, we have $C_N - C_{No} = o_p(1)$ under homoskedasticity. As

already noted, too much difference between C_N and C_{No} would indicate the presence of heteroskedasticity, which is the basis for *White (1980) test for heteroskedasticity*. We will not, however, test for heteroskedasticity; instead, we will just allow it by using the heteroskedasticity-robust variance estimator C_N . There have been criticisms on the heteroskedasticity-robust variance estimator. For instance, Kauermann and Carroll (2001) showed that, when homoskedasticity holds, the heteroskedasticity-robust variance estimator has the higher variance than the variance estimator under homoskedasticity, and that confidence intervals based on the former have the coverage probability lower than the nominal value.

Suppose

$$y_i = x_i'\beta + d_i\beta_d + d_iw_i'\beta_{dw} + u_i$$

where d_i is a dummy variable of interest (e.g., a key policy variable on ($d = 1$) or off ($d = 0$)), and w_i consists of elements of x_i interacting with d_i . Here, the effect of d_i on y_i is $\beta_d + w_i'\beta_{dw}$ which varies across i ; i.e., we get N different individual effects. A way to summarize the N -many effects is using $\beta_d + E(w')\beta_{dw}$ (the effect evaluated at the “mean person”) or $\beta_d + Med(w')\beta_{dw}$ (the effect evaluated at the “median person”). Observe

$$\begin{aligned} E(\beta_d + w_i'\beta_{dw}) &= \beta_d + E(w')\beta_{dw} \quad \text{but} \\ Med(\beta_d + w_i'\beta_{dw}) &\neq \beta_d + Med(w')\beta_{dw}; \end{aligned}$$

$Med(z_1 + z_2) \neq Med(z_1) + Med(z_2)$ in general for two rv's z_1 and z_2 . The former is that the mean effect is also the effect at the mean person, whereas the latter is that the median effect is not the effect at the median person.

If we want to Wald-test “ $H_0 : \beta_d + E(w')\beta_{dw} = 0$,” then replace $E(w)$ with \bar{w} to set $c = 0$ and $R' = (0'_{k_x}, 1, \bar{w}')$ where 0_{k_x} is the $k_x \times 1$ vector of zero's. In this test, we may worry about the difference $\bar{w} - E(w)$ in replacing the unknown $(0'_{k_x}, 1, E(w'))$ with its estimator $(0'_{k_x}, 1, \bar{w}')$. But $\bar{w} - E(w)$ can be ignored, as we can just declare that we want to evaluate the effect at the sample mean \bar{w} . “ $H_0 : \beta_d + Med(w')\beta_{dw} = 0$ ” can be tested in the analogous way, replacing $Med(w)$ with the sample median.

3.3 Empirical Examples

EXAMPLE: HOUSE SALE (continued). The two variables BATH and ROOM looked insignificant. Since BATH and ROOM tend to be highly correlated, using the two individual t-values for BATH and ROOM for the two separate hypotheses $H_0 : \beta_{bath} = 0$ and $H_0 : \beta_{room} = 0$ may be different from testing the joint null hypothesis $H_0 : \beta_{bath} = \beta_{room} = 0$ with Wald test, because the latter involves the asymptotic covariance between $b_{N,bath}$ and $b_{N,room}$ that is not used for the two t-values. It does happen in practice that, when two regressors are highly correlated (“*multicollinearity*”), the two separate null hypotheses may not be rejected while the single joint null hypothesis is rejected. This is because either one of them has explanatory power, but

adding the other to the model when one is already included does not add any new explanatory power. With $k = 11$, $g = 2$, $c = (0, 0)'$, and

$$\begin{aligned} R'_{2 \times 11} &= \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 & ,... , & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & ,... , & 0 \end{bmatrix} \\ \beta_{11 \times 1} &= (\beta_{\ln(T)}, \beta_1, \beta_{bath}, \beta_{elec}, \beta_{room}, ..., \beta_{supply})', \end{aligned}$$

the Wald test statistic is 0.456 with the p-value $0.796 = P(\chi^2_2 > 0.456)$ for the model with $\ln(T)$ and C_N : the joint null hypothesis is not rejected. When C_{oN} is used instead of C_N , the Wald test statistic is 0.501 with the p-value 0.779—hardly any change. Although BATH and ROOM are important variables for house prices, they do not explain the discount % DISC. The t-values with C_N and \tilde{C}_N shown below for the 11 regressors are little different (tv- \tilde{C}_N was shown already) because $N = 467$ is not too small for the number of regressors:

x :	$\ln(T)$	1	BATH	ELEC	RM	TAX	YR	$\ln(LP)$	BIGS	RATE	SUPPLY
tv- C_N :	7.76	-0.23	0.18	2.46	-0.67	-1.28	-3.87	2.52	-2.56	-3.10	1.06
tv- \tilde{C}_N :	7.29	-0.22	0.17	2.38	-0.65	-1.22	-3.66	2.40	-2.45	-2.98	1.03

EXAMPLE: TRANSLOG PRODUCTION FUNCTION. Consider a “*translog production function*”:

$$\ln y = \beta_0 + \sum_{p=1}^m \beta_p \ln x_p + \sum_{p=1}^m \sum_{q=1}^m \beta_{pq} \frac{1}{2} \ln x_p \ln x_q + u \quad \text{where } \beta_{pq} = \beta_{qp}.$$

This becomes a Cobb-Douglas production function if $\beta_{pq} = 0 \forall p, q$. To see why the restriction $\beta_{pq} = \beta_{qp}$ appears, observe

$$\beta_{pq} \frac{1}{2} \ln x_p \ln x_q + \beta_{qp} \frac{1}{2} \ln x_q \ln x_p = \frac{\beta_{pq} + \beta_{qp}}{2} \ln x_p \ln x_q = \beta_{pq} \ln x_p \ln x_q :$$

we can only identify the average of β_{pq} and β_{qp} , and $\beta_{pq} = \beta_{qp}$ essentially redefines the average as β_{pq} .

If we take the translog function as a second-order approximation to an underlying smooth function, say, $y = \exp\{f(x)\}$, then $\beta_{pq} = \beta_{qp}$ is a natural restriction from the symmetry of the second-order matrix. Specifically, observe

$$\ln y = f(x) \implies \ln y = f\{\exp(\ln x)\} = \tilde{f}(\ln x) \quad \text{where } \tilde{f}(t) \equiv f\{\exp(t)\}.$$

Now $\tilde{f}(\ln x)$ can be linearized around $x = 1$ (i.e., around $\ln x = 0$) with its second-order approximation where the β -parameters depend on the approximation point $x = 1$.

For a production function $y = f(x) + u$, it is “homogeneous of degree h ” if $t^h y = f(tx) + u \forall t$. To test for the h -homogeneity, apply $t^h y = f(tx) + u$ to the translog production function to get

$$\begin{aligned}
 h \ln t + \ln y &= \beta_0 + \sum_{p=1}^m \beta_p (\ln t + \ln x_p) \\
 &\quad + \sum_{p=1}^m \sum_{q=1}^m \beta_{pq} \frac{1}{2} (\ln t + \ln x_p) (\ln t + \ln x_q) + u \\
 \implies h \ln t + \ln y &= \beta_0 + \ln t \sum_{p=1}^m \beta_p + \sum_{p=1}^m \beta_p \ln x_p + \frac{(\ln t)^2}{2} \sum_{p=1}^m \sum_{q=1}^m \beta_{pq} \\
 &\quad + \frac{\ln t}{2} \left\{ \sum_{p=1}^m (\ln x_p \sum_{q=1}^m \beta_{pq}) + \sum_{q=1}^m (\ln x_q \sum_{p=1}^m \beta_{pq}) \right\} \\
 &\quad + \sum_{p=1}^m \sum_{q=1}^m \beta_{pq} \frac{1}{2} \ln x_p \ln x_q + u.
 \end{aligned}$$

For both sides to be equal for all t , it should hold that

$$\begin{aligned}
 \sum_{p=1}^m \beta_p &= h \quad \text{and} \quad \sum_{q=1}^m \beta_{pq} = 0 \quad \forall p \\
 \left(\iff \sum_{q=1}^m \beta_{qp} &= 0 \quad \forall p \iff \sum_{p=1}^m \beta_{pq} = 0 \quad \forall q \right).
 \end{aligned}$$

To be specific, for $m = 2$ and $h = 1$, there are six parameters $(\beta_0, \beta_1, \beta_2, \beta_{11}, \beta_{22}, \beta_{12})$ to estimate in

$$\ln y = \beta_0 + \beta_1 \ln x_1 + \beta_2 \ln x_2 + \beta_{11} \frac{(\ln x_1)^2}{2} + \beta_{22} \frac{(\ln x_2)^2}{2} + \beta_{12} \ln x_1 \ln x_2 + u.$$

Bear in mind $\beta_{12} = \beta_{21}$, and we use only β_{12} with the first subscript smaller than the second. The 1-homogeneity (i.e., constant returns to scale) restrictions are

$$H_0 : \beta_1 + \beta_2 = 1, \quad \beta_{11} + \beta_{12} = 0 \quad \text{and} \quad \beta_{12} + \beta_{22} = 0 \quad (\text{from } \beta_{21} + \beta_{22} = 0).$$

Clearly, we can estimate the model with LSE to test for this linear H_0 .

If H_0 is accepted, then one may want to impose the H_0 on the model using its equivalent form

$$\beta_2 = 1 - \beta_1, \quad \beta_{11} = -\beta_{12}, \quad \beta_{22} = -\beta_{12}.$$

That is, the H_0 -imposed model is

$$\begin{aligned}
 \ln y - \ln x_2 &= \beta_0 + \beta_1 (\ln x_1 - \ln x_2) + \beta_{12} \left\{ -\frac{(\ln x_1)^2}{2} - \frac{(\ln x_2)^2}{2} \right. \\
 &\quad \left. + \ln x_1 \ln x_2 \right\} + u.
 \end{aligned}$$

This can be estimated by the LSE of $\ln y - \ln x_2$ on the rhs regressors.

If $m = 3$ and $h = 1$, then there will be 10 parameters ($\beta_0, \beta_1, \beta_2, \beta_3, \beta_{11}, \beta_{22}, \beta_{33}, \beta_{12}, \beta_{13}, \beta_{23}$), and the 1-homogeneity restrictions are

$$\begin{aligned}\beta_1 + \beta_2 + \beta_3 &= 1, \quad \beta_{11} + \beta_{12} + \beta_{13} = 0, \\ \beta_{12} + \beta_{22} + \beta_{23} &= 0 \quad (\text{from } \beta_{21} + \beta_{22} + \beta_{23} = 0) \text{ and} \\ \beta_{13} + \beta_{23} + \beta_{33} &= 0 \quad (\text{from } \beta_{31} + \beta_{32} + \beta_{33} = 0).\end{aligned}$$

4 Instrumental Variable Estimator (IVE)

When $E(xu) \neq 0$, LSE becomes inconsistent. A solution is dropping (i.e., substituting out) the endogenous components of x from the model, but the ensuing LSE does not deliver what is desired: the “other-things-being-equal” effect. Another solution is to extract only the exogenous part of the endogenous regressors, which is the topic of this and the following sections.

4.1 IVE Basics

4.1.1 IVE in Narrow Sense

For the linear model $y = x'\beta + u$, suppose we have a $k \times 1$ moment condition

$$E(zu) = E(z(y - x'\beta)) = 0,$$

instead of $E(xu) = 0$, where z is a $k \times 1$ random vector such that $E(zx')$ is invertible. Solve the equation for β to get

$$\beta = E^{-1}(zx') \cdot E(zy).$$

The sample analog of this is the *instrumental variable estimator (IVE)*

$$b_{ive} = \left(\frac{1}{N} \sum_i z_i x'_i \right)^{-1} \frac{1}{N} \sum_i z_i y_i \quad \{ = (Z'X)^{-1} Z'Y \text{ in matrices} \}.$$

While IVE in its broad sense includes any estimator using instruments, here we define IVE in its narrow sense as the one taking this particular form. IVE includes LSE as a special case when $z = x$ (or $Z = X$ in matrices).

Substitute $y_i = x'_i \beta + u_i$ into the b_{ive} formula to get

$$\begin{aligned}b_{ive} &= \left(\frac{1}{N} \sum_i z_i x'_i \right)^{-1} \frac{1}{N} \sum_i z_i (x'_i \beta + u_i) = \beta + \left(\frac{1}{N} \sum_i z_i x'_i \right)^{-1} \\ &\quad \times \frac{1}{N} \sum_i z_i u_i.\end{aligned}$$

The consistency of the IVE follows simply by applying the LLN to the terms other than β in the last equation. As for the asymptotic distribution, observe

$$\sqrt{N}(b_{ive} - \beta) = \left(\frac{1}{N} \sum_i z_i x_i' \right)^{-1} \frac{1}{\sqrt{N}} \sum_i z_i u_i.$$

Applying the LLN to $N^{-1} \sum_i z_i x_i'$ and the CLT to $N^{-1/2} \sum_i z_i u_i$, it holds that

$$\sqrt{N}(b_{ive} - \beta) \rightsquigarrow N \{0, E^{-1}(zx') E(zz'u^2) E^{-1}(xz')\}.$$

This is informally stated as

$$b_{ive} \sim N \left\{ \beta, \frac{1}{N} E^{-1}(zx') E(zz'u^2) E^{-1}(xz') \right\}$$

the variance of which can be estimated with (defining $r_i \equiv y_i - x_i' b_{ive}$)

$$\left(\sum_i z_i x_i' \right)^{-1} \left(\sum_i z_i z_i' r_i^2 \right) \left(\sum_i x_i x_i' \right)^{-1} = (Z'X)^{-1} Z'DZ (X'Z)^{-1},$$

in matrices,

where $D \equiv \text{diag}(r_1^2, \dots, r_N^2)$ and $r_i = y_i - x_i' b_{ive}$, not $y_i - x_i' b_{ive}$.

4.1.2 Instrumental Variable (IV) qualifications

IVE is useful when LSE is not applicable because some regressors are endogenous in the sense $E(xu) \neq 0$. For instance, suppose $x_i = (1, x_{i2}, x_{i3}, x_{i4})'$ (thus $y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + u_i$) and

$$E(u) = E(x_2 u) = E(x_3 u) = 0, \text{ but } E(x_4 u) \neq 0;$$

x_2 and x_3 are *exogenous regressors* in the y -equation whereas x_4 is an *endogenous regressor*. If there is a rv w such that

$$(i) \text{ } COR(w, u) = 0 \quad (\iff E(wu) = 0)$$

$$(ii) \text{ } 0 \neq COR(w, x_4) \quad (\text{“inclusion restriction”})$$

$$(iii) \text{ } w \text{ does not appear in the } y \text{ equation} \quad (\text{“exclusion restriction”})$$

then w is a valid *instrumental variable (IV)*—or just *instrument*—for x_4 , and we can use $z_i = (1, x_{i2}, x_{i3}, w_i)'$. The reason why (ii) is called “inclusion restriction” is that w should be in the x_4 equation for (ii) to hold. Conditions (ii) and (iii) together are simply called “*inclusion/exclusion restrictions*.”

As an example, suppose that y is blood pressure, x_2 is age, x_3 is gender, x_4 is exercise, u includes health concern, and w is a randomized education dummy variable on health benefits of exercise (i.e., a coin is flipped to give person i the education if head comes up). Those who are health-conscious

may exercise more, which means $COR(x_4, u) \neq 0$. Checking out (i–iii) for w , first, w satisfies (i) because w is randomized. Second, those who received the health education are likely to exercise more, thus implying (ii). Third, receiving the education alone cannot affect blood pressure, and hence (iii) holds. (iii) does not mean that w should not influence y at all: (iii) is that w can affect y only indirectly through x_4 .

Condition (i) is natural in view of $E(zu) = 0$. Condition (ii) is necessary as w is used as a “proxy” for x_4 ; if $COR(w, x_4) = 0$, then w cannot represent x_4 —a rv from a coin toss is independent of x_4 and fails (ii) despite satisfying (i) and (iii). Condition (iii) is necessary to make $E(zx')$ invertible; an exogenous regressor x_2 (or x_3) already in the y -equation cannot be used as an instrument for x_4 despite it satisfies (i) and possibly (ii), because $E(zx')$ is not invertible if $z = (1, x_2, x_3, x_2)'$.

Recalling partial regression, only the part of x_4 not explained by the other regressors $(1, x_2, x_3)$ in the y equation contributes to explaining y . Among the part of x_4 , w picks only the part uncorrelated with u , because w is uncorrelated with u by condition (i). *The instrument w is said to extract the “exogenous variation” in x_4 .* In view of this, to be more precise, (ii) should be replaced with

$$\begin{aligned} (ii)' \quad & 0 \neq COR[w, \{\text{part of } x_4 \text{ unexplained by the other} \\ & \text{regressors}(1, x_2, x_3)\}] \\ \iff & 0 \neq COR[w, \{\text{residual of the linear projection of} \\ & x_4 \text{ on } (1, x_2, x_3)\}]. \end{aligned}$$

Condition (ii)' can be (and should be) verified by the LSE of x_4 on w and the other regressors: the slope coefficient of w should be non-zero in this LSE for w to be a valid instrument. But conditions (i) and (iii) cannot be checked out; they can be only “argued for.” In short, *an instrument should be excluded from the response equation and included in the endogenous regressor equation with zero correlation with the error term.*

There are a number of sources for the endogeneity of x_4 :

- First, a *simultaneous relation* when x_4 is affected by y (as well as affecting y). For example, $x_{i4} = q_i'\gamma + \alpha y_i + v_i$ where q_i are regressors and v_i is an error term. This implies that u is correlated with x_4 through y : $u \rightarrow y \rightarrow x_4$. If y is the work hours of a spouse and x_4 is the work hours of the other spouse in the same family, then the simultaneous relation may occur.
- Second, a *recursive relation with correlated errors*. For example, $x_{i4} = q_i'\gamma + v_i$ and $COR(v, u) \neq 0$ holds (no simultaneity). Here x_4 is correlated with u through v : $x_4 \leftarrow v \rightarrow u$. In the preceding family work hour case, if x_4 is for the “leader” (i.e., the dominant spouse), y is for the “follower,” and local labor-market-condition variables influencing both spouses are omitted, then these variables will lurk in u and v , leading to $COR(u, v) \neq 0$.

- Third, *errors-in-variables*. Here, x_4 is not observed, but its error-ridden version $x_{i4}^e = x_{i4} + e_i$ is. In this case, we can rewrite the y equation as

$$y_i = \dots, +\beta_4 x_{i4} + u_i = \dots, +\beta_4 (x_{i4}^e - e_i) + u_i = \dots, +\beta_4 x_{i4}^e + (u_i - \beta_4 e_i)$$

and use x_4^e as a regressor. But the new error $u - \beta_4 e$ is correlated with x_4^e through e .

4.1.3 Further Remarks

What if there is no variable available for instruments? In this case, it is tempting to use functions of exogenous regressors. Recall the above example:

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + u_i, \quad E(x_2 u) = E(x_3 u) = 0 \text{ but } E(x_4 u) \neq 0.$$

Functions of x_2 and x_3 (such as x_2^2 or $x_2 x_3$) qualify as instruments for x_4 , if we know a priori that those functions are excluded from the y equation. But “smooth” functions such as x_2^2 are typically not convincing instruments, because x_2^2 may very well be included in the y equation if x_2 is so. Instead of smooth functions, non-smooth functions of exogenous regressors may be used as instruments if there are due justifications that they appear in the x_4 equation, but not in the y equation. Such examples can be seen in relation to “*regression discontinuity design*” in the treatment-effect literature; see Lee (2005a) and the references there. Those discontinuous functions then serve as “local instruments” around the discontinuity points.

In case of no instrument, the endogenous regressors may be dropped and LSE may be applied. But this leads to an omitted variable bias as examined already. For instance, suppose $x_{i4} = \gamma_1 + \gamma_2 x_{i3} + v_i$. Substitute this into the y_i equation to get

$$\begin{aligned} y_i &= \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 (\gamma_1 + \gamma_2 x_{i3} + v_i) + u_i \\ &= (\beta_1 + \beta_4 \gamma_1) + \beta_2 x_{i2} + (\beta_3 + \beta_4 \gamma_2) x_{i3} + (u_i + \beta_4 v_i). \end{aligned}$$

When this is estimated by LSE, the slope estimator for x_3 is consistent for $\beta_3 + \beta_4 \gamma_2$, where $\beta_4 \gamma_2$ is nothing but the bias due to omitting x_4 in the LSE. The slope parameter $\beta_3 + \beta_4 \gamma_2$ of x_3 consists of two parts: the “direct effect” of x_3 on y , and the “indirect part” of x_3 on y through x_4 . If x_3 affects x_4 but not the other way around, then the indirect part can be interpreted as the “indirect effect” of x_3 on y through x_4 . So long as we are interested in the total effect $\beta_3 + \beta_4 \gamma_2$, the LSE is all right. But usually in economics, the desired effect is the “*ceteris paribus*” effect of changing x_3 while holding all the other variables (including x_4) constant.

The IVE can also be cast into a minimization problem. The sample analog of $E(zu)$ is $N^{-1} \sum_i z_i u_i$. Since u_i is unobservable, replace u_i by $y_i - x_i' b$

to get $N^{-1} \sum_i z_i(y_i - x'_i b)$. We can get the IVE by minimizing the deviation of $N^{-1} \sum_i z_i(y_i - x'_i b)$ from 0. Since $N^{-1} \sum_i z_i(y_i - x'_i b)$ is a $k \times 1$ vector, we need to choose how to measure the distance from 0. Adopting the squared Euclidean norm as usual and ignoring N^{-1} , we get

$$\begin{aligned} & \left\{ \sum_i z_i(y_i - x'_i b) \right\}' \cdot \sum_i z_i(y_i - x'_i b) = \{Z'(Y - X'b)\}' \cdot Z'(Y - X'b) \\ & = (Y - Xb)'ZZ'(Y - Xb) = Y'ZZ'Y - 2b'X'ZZ'Y + b'X'ZZ'Xb. \end{aligned}$$

The first-order condition of minimization is

$$0 = -2X'ZZ'Y + 2X'ZZ'Xb \implies b_{ive} = (Z'X)^{-1}Z'Y.$$

Although IVE can be cast into a minimization problem, it minimizes the distance of $N^{-1} \sum_i z_i(y_i - x'_i b)$ from 0. For LSE, we would be minimizing the distance of $N^{-1} \sum_i x_i(y_i - x'_i b)$ from 0, which is different from minimizing the scalar $N^{-1} \sum_i (y_i - x'_i b)^2$. This scalar minimand shows that LSE is a “prediction-error minimizing estimator” where y_i is the target and $x'_i b_{lse}$ is the predictor for the target. In minimizing $N^{-1} \sum_i (y_i - x'_i b)^2$, there is no concern for endogeneity: regardless of $E(xu) = 0$ holding or not, we can always minimize $N^{-1} \sum_i (y_i - x'_i b)^2$. The resulting estimator is, however, consistent for β in $y_i = x'_i \beta + u_i$ only if $E(xu) = 0$. The usual model fitness and R^2 are irrelevant for IVE, because, if they were, we would be using LSE, not IVE. Nevertheless, there is a pseudo R^2 to appear later that may be used for model selection with IVE, as the usual R^2 is used for the same purpose with LSE.

4.2 IVE Examples

Here we provide three empirical examples for IVE, using the same four regressor model as above with x_4 being endogenous. The reader will see that some instruments are more convincing than others. Among the three examples, the instruments in the first example will be the most convincing, followed by those in the second which are in turn more plausible than those in the third. More examples of instruments can be found in Angrist and Krueger (2001) and the references therein. It is not clear, however, who invented IVE. See Stock and Trebbi (2003) for some “detective work” on the origin of IVE.

EXAMPLE: FERTILITY EFFECT ON WORK. Understanding the relationship between fertility and female labor supply matters greatly in view of increasing labor market participation of women and declining fertility rates in many countries; the latter is also a long-term concern for pension systems. But finding a causal effect for either direction has proven difficult, as females are likely to decide on fertility and labor supply jointly, leading to a simultaneity problem. Angrist and Evans (1998) examined the effect of the number

of children (x_4) on labor market outcomes. Specifically, their x_4 is a dummy variable for more than two children.

One instrument for x_4 is the dummy for the same sex children in a household: having only girls or boys in the first two births may result in more children than the couple planned otherwise. The random event (gender) for the first two children gives an exogenous variation to x_4 , and the dummy for the same sex children is to take advantage of this variation. Another instrument is the dummy for twin second birth: having a twin second birth means an exogenous increase to the third child. Part of Table 2 (using a 1990 data set in US for women aged 21–35 with two or more children) in Angrist and Evans (1998) shows descriptive statistics (SD is in (·)):

	#Children ever	More than two	First birth boy	Same sex	Twin second birth	First birth age
All women	2.50 (0.76)	0.375 (0.48)	0.512 (0.50)	0.505 (0.50)	0.012 (0.108)	21.8 (3.5)
Wives	2.48 (0.74)	0.367 (0.48)	0.514 (0.50)	0.503 (0.50)	0.011 (0.105)	22.4 (3.5)

The table shows that there is not much difference across all women data and wives only data, that the probability of boy is slightly higher than the probability of girl, and that the probability of twin birth is about 1%.

Part of Table 8 for the all women data in Angrist and Evans (1998) is:

y	Worked or not	Weeks worked	Hours per week	Labor income
LSE (SD)	−0.155 (0.002)	−8.71 (0.08)	−6.80 (0.07)	−3984 (44.2)
IVE (SD)	−0.092 (0.024)	−5.66 (1.16)	−4.08 (0.98)	−2100 (664.0)

For instance, having the third child decreases weeks worked by 6–9% and hours worked by 4–7 hours per week. Overall, IVE magnitudes are about 50–100% smaller than the LSE magnitudes.

EXAMPLE: POLICE IMPACTS ON CRIME. Whether the number of policemen (x_4) lowers crime rates (y) has been an important question in criminology. The main difficulty in assessing the effects has been the simultaneity problem between y and x_4 . Suppose that x_4 decreases y , and y increases x_4 (a higher crime rates leads to the more policemen):

$$\begin{aligned}
 y_i &= \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + u_i, & x_{i4} &= q'_i \gamma + \alpha y_i + v_i, \\
 &\text{(with } \beta_4 < 0 \text{ and } \alpha > 0) \\
 \implies y_i &= \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 (q'_i \gamma + \alpha y_i + v_i) + u_i, \\
 &\text{(substituting the } x_4 \text{ equation)} \\
 \implies y_i &= \frac{1}{1 - \beta_4 \alpha} (\beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 q'_i \gamma + \beta_4 v_i + u_i)
 \end{aligned}$$

which is the y “reduced form (RF).” Substituting the y RF into $x_{i4} = q'_i \gamma + \alpha y_i + v_i$, we also get the x_4 RF:

$$x_{i4} = q'_i \gamma + \frac{\alpha}{1 - \beta_4 \alpha} (\beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 q'_i \gamma + \beta_4 v_i + u_i) + v_i.$$

Judging from the u ’s slope $\alpha(1 - \beta_4 \alpha)^{-1} > 0$, we get $COR(x_4, u) > 0$. Suppose that LSE is run for the y equation ignoring the simultaneity. Then with $x_i = (1, x_{i2}, x_{i3}, x_{i4})'$, the LSE of y on x will be inconsistent by the magnitude

$$E^{-1}(xx')E(xu) = E^{-1}(xx')\{0, 0, 0, E(x_4 u)\}' :$$

the LSE for β_4 is upward biased and hence the LSE for β_4 can even be positive. Recalling the discussion on omitted variable bias, we can see that the bias is not restricted to β_4 if x_4 is correlated with x_2 or x_3 , because the last column of $E^{-1}(xx')$ can “spread” $E(x_4 u) \neq 0$ to all components of the LSE.

One way to overcome the simultaneity problem is to use data for short periods. For instance, if y is a monthly crime number for city i and x_4 is the number of policemen in the same month, then it is unlikely that y affects x_4 as it takes time to adjust x_4 , whereas x_4 can affect y almost instantly. Another way is to find an instrument. Levitt (1997) noted that the change in x_4 takes place almost always in election years, mayoral or gubernatorial. Thus he sets up a “panel (or longitudinal) data” model where y_{it} is a change in crime numbers for city i and year t , $x_{it,4}$ is a change in policemen, and $w_{it} = 1$ if year t is an election year at city i and 0 otherwise, because w_{it} is unlikely to be correlated with the error term in the crime number change equation. Levitt (1997) concluded that the police force size reduces (violent) crimes.

As McCrary (2002) noted, however, there was a small error in Levitt (1997). Levitt (2002) thus proposed the number of firefighters per capita as a new instrument for the number of policemen per capita. The panel data model used is

$$\Delta \ln(y_{it}) = \beta_p \ln(\text{police}_{i,t-1}) + x'_{it} \beta_x + \delta_i + \lambda_t + u_{it}$$

where i indexes large US cities with $N = 122$, t indexes years 1975–1995, $\text{police}_{i,t-1}$ instead of police_{it} is used to mitigate the endogeneity problem, and x_{it} are the regressors other than police; δ_i is for the “city effect” (estimated by city dummies) and λ_t is for the “year effect” (estimated year dummies).

Part of Table 3 in Levitt (2002) for police effect is shown below with SD in (\cdot), where “LSE without city dummies” means the LSE without city dummies but with year dummies. By not using city dummies, the parameters are identified mainly with cross-city variation because cross-city variation is much greater than over-time variation, and this LSE is thus similar to cross-section LSE pooling all panel data.

y	Violent crimes per capita	Property crimes per capita
LSE without city dummies	0.562 (0.056)	0.113 (0.038)
LSE with city/year dummies	-0.076 (0.061)	-0.218 (0.052)
IVE with city/year dummies	-0.435 (0.231)	-0.501 (0.235)

This table shows that the LSE’s are upward biased as analyzed above although the bias is smaller when the city dummies are used, and that police force expansion indeed reduces the number of crimes. The number of fire-fighters is an attractive instrument, but somewhat less convincing than the instruments in the fertility example.

EXAMPLE: ECONOMIC IMPACTS ON CRIME. In the preceding examples for IVE, the justification of the instruments was strong. Here is an IVE example with a weaker justification—this kind of cases are more common in practice.

Gould et al. (2002) analyzed the effect of local labor market conditions on crime rates in the US for 1979–1997. They set up a panel data model

$$y_{it} = x'_{it}\beta + \delta_i + u_{it}, \quad i = 1, \dots, N, \quad t = 1, \dots, T$$

where y_{it} is the number of various offenses per 100,000 people in county i at year t , x_{it} includes the mean log weekly wage of non-college educated men ($wage_{it}$), unemployment rate of non-college educated men (ur_{it}), and the mean log household income (inc_{it}), and time dummies, δ_i is a time-constant error and u_{it} is a time-variant error. Our presentation in the following is a rough simplification of their longer models.

Since δ_i represents each county’s unobserved long-standing culture and practice such as how extensively crimes are reported and so on, δ_i is likely to be correlated with x_{it} . They take the difference between 1979 and 1989 to remove δ_i and get (removing δ_i by differencing is a “standard” procedure in panel data)

$$\Delta y_i = \Delta x'_i \beta + \Delta u_i,$$

where $\Delta y_i \equiv y_{i,1989} - y_{i,1979}$, Δx_i and Δu_i are analogously defined, and $N = 564$. Their estimation results are as follows with SD in (\cdot):

$$\begin{aligned} LSE: \quad & b_{wage} = -1.13 (0.38), \quad b_{ur} = 2.35 (0.62), \quad b_{inc} = 0.71 (0.35) \\ & R^2 = 0.094, \\ IVE: \quad & b_{wage} = -1.06 (0.59), \quad b_{ur} = 2.71 (0.97), \quad b_{inc} = 0.093 (0.55); \end{aligned}$$

the instruments will be explained below. All three estimates in LSE are significant and show that low wage, high unemployment rate, and high household income increase the crime rate. The IVE is close to the LSE in $wage_{it}$ and ur_{it} , but much smaller for inc_{it} and insignificant. See Freeman (1999) for a survey on crime and economics.

It is possible that crime rates influence local labor market conditions, because firms may move out in response to high crime rates or firms may offer higher wages to compensate for high crime rates. This means that a simultaneous relation problem may occur between crime rates and labor market conditions. To avoid this problem, Gould et al. (2002) constructed a number of instruments. One of the instruments is

$$\sum_j (\text{employment share of industry } j \text{ in county } i \text{ in 1979}) \cdot (\text{national growth rate of industry } j \text{ for 1979–1989}).$$

The two conditions to check are $COR(\Delta u, z) = 0$ and $COR(\Delta x, z) \neq 0$. For $COR(\Delta u, z) = 0$, the primary reason to worry for endogeneity was the influence of the crime rate on the labor market conditions. But it is unlikely that a county's crime rates over 1979–1989 influenced the national industry growth rates over 1979–1989. Also the employment shares had been taken in 1979 before the crime rates were measured. These points support $COR(\Delta u, z) = 0$. For $COR(\Delta x, z) \neq 0$, consider a county in Michigan: if the national auto industry shrank during 1979–1989 and if the share of auto industry was large in the county in 1979, then the local labor market condition would have deteriorated.

4.3 IVE with More than Enough Instruments

4.3.1 IVE in Wide Sense

If a random variable w is independent u , then we get not just $COR(w, u) = 0$, but also $COR(w^2, u) = 0$. This means that if w is an instrument for an endogenous regressor x_4 , then we may use both w and w^2 as instruments for x_4 . In this case, $z_i = (1, x_{i2}, x_{i3}, w_i, w_i^2)'$, the dimension of which is bigger than the dimension of x_i : $E(zx')$ is not a square matrix as in the preceding subsection, and hence not invertible. There arises the question of selecting or combining more than enough instruments (i.e., more than enough moment conditions) for only k -many parameters. While a complete answer will be provided later, here we just provide one simple answer which also turns out to be optimal under homoskedasticity.

Suppose $E(zu) = 0$, where z is $p \times 1$ with $p \geq k$, the rank of $E(xz') = k$, and $E^{-1}(zz')$ exists. Observe

$$\begin{aligned} E\{z(y - x'\beta)\} &= 0 \iff E(zy) = E(xz')\beta \\ \implies E(xz')E^{-1}(zz') \cdot E(zy) &= E(xz')E^{-1}(zz') \cdot E(xz')\beta \\ \implies \beta &= \{E(xz')E^{-1}(zz')E(xz')\}^{-1} \cdot E(xz')E^{-1}(zz')E(zy). \end{aligned}$$

For the product AB of two matrices A and B where B^{-1} exists, $\text{rank}(AB) = \text{rank}(A)$; i.e., multiplication by a non-singular matrix B does not alter the rank of A . This fact implies that $E(xz')E^{-1}(zz')E(zx')$ has rank k and thus is invertible. If $E(zx')$ is invertible, then β in the last display becomes $E^{-1}(zx') \cdot E(zy)$ and the resulting b_{ive} is the IVE when the number of instruments is the same as the number of parameters.

The sample analog for β is the following *instrumental variable estimator*

$$b_{ive} = \left\{ \sum_i x_i z'_i \left(\sum_i z_i z'_i \right)^{-1} \sum_i z_i x'_i \right\}^{-1} \cdot \sum_i x_i z'_i \left(\sum_i z_i z'_i \right)^{-1} \sum_i z_i y_i$$

where many N^{-1} 's are ignored that cancel one another out. The consistency is obvious, and the asymptotic distribution of $\sqrt{N}(b_{ive} - \beta)$ is

$$N(0, G \cdot E(zz'u^2) \cdot G'), \quad \text{where } G \equiv \{E(xz')E^{-1}(zz')E(zx')\}^{-1}E(xz')E^{-1}(zz').$$

A consistent estimator for the asymptotic variance is

$$G_N \cdot \frac{1}{N} \sum_i z_i z'_i r_i^2 \cdot G'_N,$$

where $r_i \equiv y_i - x'_i b_{ive}$, and

$$G_N \equiv \left\{ \frac{1}{N} \sum_i x_i z'_i \left(\frac{1}{N} \sum_i z_i z'_i \right)^{-1} \frac{1}{N} \sum_i z_i x'_i \right\}^{-1} \cdot \frac{1}{N} \sum_i x_i z'_i \left(\frac{1}{N} \sum_i z_i z'_i \right)^{-1}.$$

4.3.2 Various Interpretations of IVE

It is informative to see the IVE in matrices:

$$\begin{aligned} b_{ive} &= \{X'Z(Z'Z)^{-1}Z'X\}^{-1}X'Z(Z'Z)^{-1}Z'Y \\ &= [\{Z(Z'Z)^{-1}Z'X\}'X]^{-1}\{Z(Z'Z)^{-1}Z'X\}'Y \\ &= (\hat{X}'X)^{-1}\hat{X}'Y, \quad \text{where } \hat{X} \equiv Z(Z'Z)^{-1}Z'X; \end{aligned}$$

\hat{X} that has dimension $N \times k$ is “ x fitted by z ,” or the part of x explained by z ; $(Z'Z)^{-1}Z'X$ is the LSE of z on x . \hat{X} combines more than k instruments into just k many.

Using $P_Z \equiv Z(Z'Z)^{-1}Z'$, b_{ive} can also be written as (recall that P_Z is idempotent)

$$b_{ive} = \{(P_Z X)'P_Z X\}^{-1}(P_Z X)'P_Z Y = (\hat{X}'\hat{X})^{-1}\hat{X}'Y$$

as if b_{ive} were the LSE for the equation $Y = \hat{X}\beta + \text{error}$ where the “error” is $Y - \hat{X}\beta$. This rewriting accords an interesting interpretation of b_{ive} . As X is

endogenous, we can decompose X as $\hat{X} + (X - \hat{X})$ where the first component is exogenous, “sifted” from X using P_Z , and the second component is the remaining endogenous component. Then

$$Y = X\beta + U = \{\hat{X} + (X - \hat{X})\}\beta + U = \hat{X}\beta + \{(X - \hat{X})\beta + U\}.$$

Hence, β can be estimated by the LSE of Y on \hat{X} so long as \hat{X} is “asymptotically uncorrelated” with the error term $(X - \hat{X})\beta + U$, which is shown in the following.

Recall

“ $A_N = o_p(1)$ ” means $A_N \rightarrow^p 0$ (and $B_N = C_N + o_p(1)$ means $B_N - C_N \rightarrow^p 0$).

The error vector $(X - \hat{X})\beta + U$ satisfies $N^{-1}\hat{X}'\{(X - \hat{X})\beta + U\} = o_p(1)$ because

$$\begin{aligned} \frac{1}{N}\hat{X}'\{(X - \hat{X})\beta + U\} &= \frac{1}{N}\left(\hat{X}'X\beta - \hat{X}'\hat{X}\beta + \hat{X}'U\right) = \frac{1}{N}\hat{X}'U, \\ &\text{for } \hat{X}'X = \hat{X}'\hat{X} \\ &= \frac{1}{N}\{Z(Z'Z)^{-1}Z'X\}'U \\ &= \frac{1}{N}X'Z \cdot \left(\frac{1}{N}Z'Z\right)^{-1} \cdot \frac{1}{N}ZU = o_p(1). \end{aligned}$$

The expression $(\hat{X}'\hat{X})^{-1}\hat{X}'Y$ also demonstrates that the so-called “*two-stage LSE (2SLSE)*” for simultaneous equations is nothing but IVE. For simplification, consider two simultaneous equations with two endogenous variables y_1 and y_2 :

$$\begin{aligned} y_1 &= \alpha_1 y_2 + x_1' \beta_1 + u_1, & y_2 &= \alpha_2 y_1 + x_2' \beta_2 + u_2, \\ \text{COR}(x_j, u_{j'}) &= 0, \quad j, j' = 0, 1 & \text{and} & \quad x_1 \neq x_2. \end{aligned}$$

Let z denote the system exogenous regressors (i.e., the collection of the elements in x_1 and x_2). Denoting the regressors for the y_1 equation as $x \equiv (y_2, x_1)'$, the first step of 2SLSE for (α_1, β_1) is the LSE of y_2 on z to obtain the fitted value \hat{y}_2 of y_2 , and the second step is the LSE of y_1 on (\hat{y}_2, x_1) . This 2SLSE is nothing but the IVE where the first step is $P_Z X$ to obtain the LSE fitted value of x on z —the LSE fitted value of x_1 on z is simply x_1 —and the second step is the LSE of y_1 on $P_Z X$.

4.3.3 Further Remarks

We already mentioned that the usual R^2 is irrelevant for IVE. Despite this, sometimes

$$1 - \frac{\sum_i (y_i - x_i' b_{ive})^2}{\sum_i (y_i - \bar{y})^2}$$

is reported in practice as a measure of model fitness. Pesaran and Smith (1994) showed, however, that this should not be used as a model selection criterion. Instead, they propose the following *pseudo R^2 for IVE*:

$$R_{ive}^2 = 1 - \frac{\sum_i (y_i - \hat{x}'_i b_{ive})^2}{\sum_i (y_i - \bar{y})^2}$$

where \hat{x}_i is the i th row of \hat{X} . R_{ive}^2 satisfies $0 \leq R_{ive}^2 \leq 1$ and takes 1 if $y = \hat{x}'_i b_{ive}$ and 0 if all slope components of b_{ive} are zero. The intuition for R_{ive}^2 was given ahead already: $Y = \hat{X}\beta + \text{error}$ with the error term asymptotically orthogonal to \hat{X} .

Observe

$$\begin{aligned} E(u^2|z) &= \sigma^2 \text{ (homoskedasticity) implies } G \cdot E(zz'u^2) \cdot G' \\ &= \sigma^2 \{E(xz')E^{-1}(zz')E(zx')\}^{-1}; \end{aligned}$$

here, homoskedasticity is wrt z , not x . To compare this to the LSE asymptotic variance $\sigma^2 E^{-1}(xx')$ under homoskedasticity, observe

$$\begin{aligned} E(xx') &= E(xz')E^{-1}(zz')E(zx') + E\{(x - \gamma'z)(x - \gamma'z)'\} \quad \text{where} \\ \gamma &\equiv E^{-1}(zz')E(zx'). \end{aligned}$$

This is a decomposition of $E(xx')$ into two parts, one explained by z and the other unexplained by z ; $x - \gamma'z$ is the “residual” (compared with the linear projection, $E(x|z)$ is often called the *projection of x on z*).

From the decomposition, we get

$$\begin{aligned} E(xx') &\geq E(xz')E^{-1}(zz')E(zx') \iff E^{-1}(xx') \\ &\leq \{E(xz')E^{-1}(zz')E(zx')\}^{-1}; \end{aligned}$$

the former is called *generalized Cauchy-Schwarz inequality*. This shows that the “explained variation” $E(xz')E^{-1}(zz')E(zx')$ is not greater than the “total variation” $E(xx')$. Hence, under homoskedasticity, LSE is more efficient than IVE; under homoskedasticity, there is no reason to use IVE unless $E(xu) \neq 0$. Under heteroskedasticity, however, the asymptotic variances of LSE and IVE are difficult to compare, because the comparison depends on the functional forms of $V(u|x)$ and $V(u|z)$.

5 Generalized Method-of-Moment Estimator (GMM)

In IVE, we saw an answer to the question of how to combine more than enough moment conditions. There, the idea was to multiply the more than enough p -many equations $E(zy) = E(zx')\beta$ from $E(zu) = 0$ with a $k \times p$ matrix. But, there are many candidate $k \times p$ matrices, with $E(xz')E^{-1}(zz')$

for IVE being just one of them. If we use $E(xz')W^{-1}$ where W is a $p \times p$ p.d. matrix, we will get

$$\begin{aligned} E(xz')W^{-1}E(zy) &= E(xz')W^{-1}E(zx')\beta \\ \implies \beta &= \{E(xz')W^{-1}E(zx')\}^{-1}E(xz')W^{-1}E(zy). \end{aligned}$$

As it turns out, $W = E(zz'u^2)$ is optimal for iid samples, which is the theme of this section.

5.1 GMM Basics

Suppose there are $p(\geq k)$ population moment conditions

$$E\psi(y, x, z, \beta) = 0$$

which may be nonlinear in β ; we will often write $\psi(y, x, z, \beta)$ simply as $\psi(\beta)$. The *generalized method-of-moment (GMM)* estimator is a class of estimators indexed by W that is obtained by minimizing the following wrt b :

$$\frac{1}{\sqrt{N}} \sum_i \psi(b)' \cdot W^{-1} \cdot \frac{1}{\sqrt{N}} \sum_i \psi(b).$$

The question in GMM is which W to use. Hansen (1982) showed that the W yielding the smallest variance for the class of GMM estimator is

$$V \left\{ \frac{1}{\sqrt{N}} \sum_i \psi(\beta) \right\} \quad [= E\{\psi(\beta)\psi(\beta)'\} \text{ for iid samples}];$$

this becomes $E(zz'u^2)$ when $\psi(\beta) = z(y - x'\beta)$.

The intuition for $W = V\{N^{-1/2} \sum_i \psi(\beta)\}$ is that, in the minimization, it is better to standardize $N^{-1} \sum_i \psi(b)$; otherwise one component with a high variance can unduly dominate the minimand. The optimal GMM is simply called (the) GMM. The GMM with $W = I_p$ is sometimes called the “unweighted (or equally weighted) GMM”; the name “equally weighted GMM,” however, can be misleading, for the optimal GMM has this interpretation. It may seem that we may be able to do better than GMM by using a distance other than the quadratic distance. But Chamberlain (1987) showed that the GMM is the efficient estimator under the given moment condition $E\psi(\beta) = 0$. In statistics, $\psi(y, x, z, \beta) = 0$ is called “estimating functions” (Godambe, 1960) and $E\psi(y, x, z, \beta) = 0$ “*estimating equations*”; see Owen (2001) and the references therein.

While GMM with nonlinear models will be examined in another chapter in detail, for the linear model, we have

$$E\psi(\beta) = E\{z(y - x'\beta)\} = E(zu) = 0.$$

In matrices, the GMM minimand with W is

$$\begin{aligned} & \{Z'(Y - Xb)\}'W^{-1}\{Z'(Y - Xb)\} \\ &= (Y'ZW^{-1} - b'X'ZW^{-1}) \cdot (Z'Y - Z'Xb) \\ &= Y'ZW^{-1}Z'Y - 2b'X'ZW^{-1}Z'Y + b'X'ZW^{-1}Z'Xb. \end{aligned}$$

From the first-order condition of minimization, we get $X'ZW^{-1}Z'Y = X'ZW^{-1}Z'Xb$. Solve this to obtain

$$\begin{aligned} b_W &= (X'ZW^{-1}Z'X)^{-1} \cdot (X'ZW^{-1}Z'Y) \\ &= \left(\sum_i x_i z_i' W^{-1} \sum_i z_i x_i' \right)^{-1} \cdot \sum_i x_i z_i' W^{-1} \sum_i z_i y_i, \quad \text{in vectors.} \end{aligned}$$

Clearly b_W is consistent for β , and its asymptotic distribution is

$$\begin{aligned} \sqrt{N}(b_W - \beta) &\rightsquigarrow N(0, C_W), \quad \text{where} \\ C_W &\equiv \{E(xz')W^{-1}E(zx')\}^{-1}E(xz')W^{-1}E(zz'u^2)W^{-1}E(zx') \\ &\quad \{E(xz')W^{-1}E(zx')\}^{-1}. \end{aligned}$$

With $W = E(zz'u^2)$, this matrix becomes $\{E(xz')E^{-1}(zz'u^2)E(zx')\}^{-1}$, and we get the GMM with

$$\sqrt{N}(b_{gmm} - \beta) \rightsquigarrow N(0, \{E(xz')E^{-1}(zz'u^2)E(zx')\}^{-1}).$$

Since $W = E(zz'u^2)$ can be estimated consistently with

$$\frac{1}{N} \sum_i z_i z_i' r_i^2 = \frac{1}{N} Z'DZ,$$

where $r_i = y_i - x_i' b_{ive}$ and $D = \text{diag}(r_1^2, \dots, r_N^2)$, we get

$$\begin{aligned} b_{gmm} &= \left\{ \sum_i x_i z_i' \left(\sum_i z_i z_i' r_i^2 \right)^{-1} \sum_i z_i x_i' \right\}^{-1} \cdot \\ &\quad \sum_i x_i z_i' \left(\sum_i z_i z_i' r_i^2 \right)^{-1} \sum_i z_i y_i \\ &= (X'Z(Z'DZ)^{-1}Z'X)^{-1} (X'Z(Z'DZ)^{-1}Z'Y) \quad \text{in matrices.} \end{aligned}$$

Differently from IVE, $Z(ZDZ')^{-1}Z'$ is no longer the linear projection matrix of Z . A consistent estimator for the GMM asymptotic variance $\{E(xz')E^{-1}(zz'u^2)E(zx')\}^{-1}$ is easily obtained: it is simply the first part $(X'Z(Z'DZ)^{-1}Z'X)^{-1}$ of b_{gmm} times N .

5.2 GMM Remarks

A nice feature of GMM is that it also provides a specification test, called “*GMM over-identification test*”: with $u_{Ni} \equiv y_i - x_i' b_{gmm}$,

$$\frac{1}{\sqrt{N}} \sum_i z_i' u_{Ni} \cdot \left(\frac{1}{N} \sum_i z_i z_i' u_{Ni}^2 \right)^{-1} \cdot \frac{1}{\sqrt{N}} \sum_i z_i u_{Ni} \rightsquigarrow \chi_{p-k}^2.$$

Too big a value, greater than an upper quantile of χ^2_{p-k} , indicates that some moment conditions do not hold (or some other assumptions of the model may be violated). The reader may wonder how we can test for the very moment conditions that were used to get the GMM. If there are only k moment conditions, this concern is valid. But when there are more than k moment conditions (p -many), essentially only k of them get to be used in obtaining the GMM. The GMM over-identification test checks if the remaining $p - k$ moment conditions are satisfied by the GMM, as can be seen in the degrees of freedom (“dof”) of the test.

The test statistics may be viewed as

$$\sum_i \left[\left\{ \frac{u_{Ni} z'_i}{\sqrt{N}} \left(\sum_i \frac{z_i u_{Ni}}{\sqrt{N}} \frac{z'_i u_{Ni}}{\sqrt{N}} \right)^{-1} \sum_i \frac{z_i u_{Ni}}{\sqrt{N}} 1 \right\}' \cdot \left\{ \frac{u_{Ni} z'_i}{\sqrt{N}} \left(\sum_i \frac{z_i u_{Ni}}{\sqrt{N}} \frac{z'_i u_{Ni}}{\sqrt{N}} \right)^{-1} \sum_i \frac{z_i u_{Ni}}{\sqrt{N}} 1 \right\} \right].$$

Defining the matrix version for $z_i u_{Ni} / \sqrt{N}$ as G —i.e., the i th row of G is $z'_i u_{Ni} / \sqrt{N}$ —this display can be written as

$$\{G(G'G)^{-1}G'1_N\}' \{G(G'G)^{-1}G'1_N\} = 1'_N G (G'G)^{-1} G' 1_N$$

The inner-product form shows that the test statistic is non-negative at least.

Using the GMM over-identification test, a natural thing to do is to use only those moment conditions that are not rejected by the test. This can be done in practice by doing GMM on various subsets of the moment conditions, which would be ad hoc, however. Andrews (1999) and Hall and Peixe (2003) provided a formal discussion on this issue of selecting valid moment conditions, although how popular these suggestions will be in practice remains to be seen.

Under the homoskedasticity $E(u^2|z) = \sigma^2$, $W = \sigma^2 E(zz') = \sigma^2 Z'Z/N + o_p(1)$. But any multiplicative scalar in W is irrelevant for the minimization. Hence setting $W = Z'Z$ is enough, and b_{gmm} becomes b_{ive} under homoskedasticity; the aforementioned optimality of b_{ive} comes from the GMM optimality under homoskedasticity. Under homoskedasticity, we do not need an initial estimator to get the residuals r_i ’s. But when we do not know whether homoskedasticity holds or not, GMM is obtained in two stages: first apply IVE to get the r_i ’s, then use $\sum_i z_i z'_i r_i^2$ to get the GMM. For this reason, the GMM is sometimes called a “two-stage IVE.”

We can summarize our analysis for the linear model under $p \times 1$ moment condition $E(zu) = 0$ and heteroskedasticity of unknown form as follows. First, the efficient estimator when $p \geq k$ is

$$b_{gmm} = \{X'Z(Z'DZ)^{-1}Z'X\}^{-1}X'Z(Z'DZ)^{-1}Z'Y.$$

If homoskedasticity prevails,

$$b_{ive} = \{X'Z(Z'Z)^{-1}Z'X\}^{-1}X'Z(Z'Z)^{-1}Z'Y$$

is the efficient estimator to which b_{gmm} becomes asymptotically equivalent. If $p = k$ and $(N^{-1} \sum_i z_i x_i')^{-1}$ exists, then

$$b_{gmm} = \{X'Z(Z'DZ)^{-1}Z'X\}^{-1}X'Z(Z'DZ)^{-1}Z'Y = (Z'X)^{-1}(Z'Y);$$

i.e., $b_{gmm} = b_{ive}$. Furthermore, if $z = x$, then $b_{gmm} = b_{ive} = b_{lse}$. Since GMM is efficient under $E(zu) = 0$, IVE is also efficient under homoskedasticity; if $(\sum_i z_i x_i')^{-1}$ exists, IVE inherits the efficiency from GMM because $b_{gmm} = b_{ive}$, homoskedasticity or not; furthermore, LSE is efficient when $z = x$, because $b_{gmm} = b_{lse}$. This way of characterizing the LSE efficiency is more relevant to economic data than using the conventional Gauss–Markov theorem in many econometric textbooks that requires non-random regressors.

There have been some further developments in linear-model IVE/GMM. The main issues there are weak instruments (i.e., small correlations between instruments and endogenous variables), small sample performance of IVE/GMM (e.g., small sample bias), small sample distribution (i.e., non-normal contrary to the asymptotic normality), and estimation of the variance (e.g., under-estimation of the variance). Just to name a few studies for readers interested in these topics, Donald and Newey (2001) showed how to choose the number of instruments by minimizing a mean-squared error criterion for IVE and other estimators, Stock et al. (2002) provided a survey on the literature, and Windmeijer (2005) suggested a correction to avoid the variance under-estimation problem. See also Hall (2005) for an extensive review on GMM.

5.3 GMM Examples

As an example of GMM moment conditions, consider a “rational expectation” model:

$$\begin{aligned} y_t &= \rho \cdot E(y_{t+1}|I_t) + x_t'\beta + \varepsilon_t, \quad t = 1, \dots, T, \\ E(\varepsilon_t x_{t-j}) &= E(\varepsilon_t y_{t-j}) = 0 \quad \forall j = 1, \dots, t \end{aligned}$$

where I_t is the information available up to period t including $x_t, y_{t-1}, x_{t-1}, \dots$. Here ρ captures the effect of the expectation of y_{t+1} on y_t . One way to estimate ρ and β is to replace $E(y_{t+1}|I_t)$ by y_{t+1} :

$$\begin{aligned} y_t &= \rho y_{t+1} + x_t'\beta + \varepsilon_t + \rho\{E(y_{t+1}|I_t) - y_{t+1}\} \\ &\equiv \rho y_{t+1} + x_t'\beta + u_t, \quad t = 1, \dots, T-1, \quad \text{where } u_t \\ &\equiv \varepsilon_t + \rho\{E(y_{t+1}|I_t) - y_{t+1}\}. \end{aligned}$$

Then $y_{t-1}, x_{t-1}, y_{t-2}, x_{t-2}, \dots$ are all valid instruments because the error term $E(y_{t+1}|I_t) - y_{t+1}$ is uncorrelated with all available information up to t . If $E(x_t \varepsilon_t) = 0$, then x_t is also a good instrument.

EXAMPLE: HOUSE SALE (continued). In estimating the DISC equation, $\ln(T)$ is a possibly endogenous variable as already noted. The endogeneity

can be dealt with IVE and GMM. We will use L1, L2, and L3 as instruments. An argument for the instruments would be that, while the market conditions and the characteristics of the house and realtor may influence DISC directly, it is unlikely that DISC is affected directly by when to list the house in the market. The reader may object to this argument, in which case the following should be taken just as an illustration.

Although we cannot test for the exclusion restriction, we can at least check whether the three variables have explanatory power for the potentially endogenous regressor $\ln(T)$. For this, the LSE of $\ln(T)$ on the instruments and the exogenous regressors was done to yield (heteroskedasticity-robust variance used):

$$\ln(T_i) = -1.352 + \dots - 0.294 \cdot L_1 - 0.269 \cdot L_2 - 0.169 \cdot L_3, R^2 = 0.098, \\ (t\text{-value}) \quad (-0.73) \quad (-2.77) \quad (-2.21) \quad (-1.22)$$

which shows that indeed L1, L2, and L3 have explanatory power for $\ln(T)$. Table 2 shows the LSE, IVE, and GMM results (LSE is provided here again for the sake of comparison). The pseudo R^2 for the IVE is 0.144; compare this to the $R^2 = 0.34$ of the LSE. The GMM over-identification test statistic value and its p -value are, respectively, 2.548 and 0.280, not rejecting the moment conditions.

Table 2: LSE, IVE, and GMM for House Sale Discount %

	LSE			IVE			GMM	
	b_{lse}	tv-ho	tv-het	b_{ive}	tv-ho	tv-het	b_{gmm}	tv
Ln(T)	4.60	12.23	7.76	10.57	2.66	2.84	10.35	2.78
1	-2.46	-0.24	-0.23	3.51	0.26	0.21	1.65	0.10
BATH	0.11	0.17	0.18	-0.15	-0.19	-0.19	-0.43	-0.54
ELEC	1.77	2.60	2.46	0.75	0.69	0.68	0.87	0.80
RM	-0.18	-0.71	-0.67	0.08	0.22	0.20	0.14	0.36
TAX	-1.74	-1.65	-1.28	-1.27	-0.94	-0.86	-1.05	-0.71
YR	-0.15	-5.96	-3.87	-0.15	-4.93	-3.71	-0.15	-3.75
Ln(LP)	6.07	3.73	2.52	3.15	1.12	0.97	2.79	0.87
BIGS	-2.15	-3.10	-2.56	-1.57	-1.66	-1.56	-1.47	-1.46
RATE	-2.99	-3.25	-3.10	-5.52	-2.73	-2.51	-5.05	-2.39
SUPPLY	1.54	1.02	1.06	1.96	1.03	1.14	2.11	1.23

In the IVE, the tv-ho's are little different from the tv-het's other than for YR. The difference between the tv for GMM and the tv-het for IVE is also negligible. The LSE have far more significant variables than the IVE and GMM which are close to each other. $\ln(T)$'s estimate is about 50% smaller in LSE than in IVE and GMM. ELEC and $\ln(LP)$ lose its significance in the IVE and GMM and the estimate sizes are also halved. YR has almost the same estimates and t-values across the three estimators. BIGS have similar estimates across the three estimators, but not significant in the IVE and GMM. RATE is significant for all three estimators, and its value changes from -3 in LSE to -5 in the IVE and GMM. Overall, the signs of the significant estimates are the same across all estimators, and most earlier remarks made for LSE apply to IVE and GMM.

6 Generalized Least Squares Estimator (GLS)

In WLS, we assumed that u_1, \dots, u_N are independent and that $E(u_i^2|x_i) = \omega(x_i, \theta)$ is a parametric function of x_i with some unknown parameter vector θ . In this section, we generalize WLS further by allowing u_1, \dots, u_N to be correlated and $u_i u_j$ to be heteroskedastic, which leads to “Generalized LSE (GLS).” Although GLS is not essential for our main theme of using “simple” moment conditions, understanding GLS is helpful in understanding GMM and its efficiency issue. GLS will appear later in other contexts as well.

6.1 GLS Basics

Suppose

$$E(u_i u_j | x_1, \dots, x_N) = \omega(x_1, \dots, x_N, \theta) \quad \forall i, j$$

for some parametric function ω . The product $u_i u_j$ may depend only on x_i and x_j , but for more generality, we put all x_1, \dots, x_N in the conditioning set. For example, if the data come from a small town, then $u_i u_j$ may depend on all x_1, \dots, x_N . If we set $E(u_i u_j | x_1, \dots, x_N) = \sigma$ which is a non-zero constant for all $i \neq j$, then we are allowing for dependence between u_i and u_j while ruling out heteroskedasticity. Recall that the consistency of LSE requires only $E(xu) = 0$; there is no restriction on the dependence among u_1, \dots, u_N nor on the form of heteroskedasticity. Hence, correlations among u_1, \dots, u_N or unknown forms of heteroskedasticity do not make LSE inconsistent; they may make either the LSE asymptotic variance matrix formula invalid or the LSE inefficient.

Writing the assumption succinctly using matrix notations, we have

$$E(UU'|X) = \Omega(X; \theta);$$

denote $\Omega(X; \theta)$ just as Ω to simplify notation, and pretend that θ is known for a while. As we transformed the original equation in WLS so that the resulting error term variance matrix becomes homoskedastic with unit variance, multiply $Y = X\beta + U$ by $\Omega^{-1/2} = H\Lambda^{-0.5}H'$ where $\Omega = H\Lambda H'$, Λ is the diagonal matrix of the eigenvalue of Ω , and H is a matrix whose columns are orthonormal eigenvectors) to get

$$\begin{aligned} \Omega^{-1/2}Y &= \Omega^{-1/2}X\beta + U^* \quad \text{where } U^* \equiv \Omega^{-1/2}U \\ \implies E(U^*U^{*'}|X) &= E(\Omega^{-1/2}UU'\Omega^{-1/2}|X) = \Omega^{-1/2}\Omega\Omega^{-1/2} \\ &= H\Lambda^{-0.5}H'H\Lambda H'H\Lambda^{-0.5}H' = I_N. \end{aligned}$$

Define

$$X^* \equiv \Omega^{-1/2}X \quad \text{and} \quad Y^* \equiv \Omega^{-1/2}Y,$$

and apply LSE to get the *Generalized LSE (GLS)*

$$\begin{aligned} b_{glS} &= (X^{*'}X^*)^{-1}(X^{*'}Y^*) \\ &= (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y = \beta + (X'\Omega^{-1}X)^{-1} \cdot X'\Omega^{-1}U. \end{aligned}$$

As in WLS, we need to replace θ with a first-stage \sqrt{N} -consistent estimator, say $\hat{\theta}$; call the GLS with $\hat{\theta}$ the “feasible GLS” and the GLS with θ the “infeasible GLS.” Whether the feasible GLS is consistent with the same asymptotic distribution as the infeasible GLS follows depends on the form of $\Omega(X; \theta)$, but in all cases we will consider GLS for, this will be the case as to be shown in a later chapter. In the transformed equation, the error terms are iid and homoskedastic with unit variance. Thus we get

$$\sqrt{N}(b_{GLS} - \beta) \rightsquigarrow N(0, E^{-1}(x^* x^{*'})).$$

The variance matrix $E(x^* x^{*'})$ can be estimated consistently with

$$\frac{1}{N} \sum_i x_i^* x_i^{*'} = \frac{1}{N} X^{*'} X^* = \frac{1}{N} X' \Omega(X; \hat{\theta})^{-1} X.$$

6.2 GLS Remarks

If we define $Z \equiv \Omega^{-1} X$, then $b_{gls} = (Z' X)^{-1} Z' Y$, which is reminiscent of IVE. But, differently from that IVE was motivated to avoid inconsistency of LSE, the main motivation for GLS is a more efficient estimation than GMM. In GMM, the functional form $E(UU'|X)$ is not specified; rather, GMM just allows $E(UU'|X)$ to be an arbitrary unknown function of X . In contrast, GLS specifies fully the functional form $E(UU'|X)$. Hence, GLS makes use of more assumptions than GMM, and as a consequence, GLS is more efficient than GMM—more on this later. But the obvious disadvantage of GLS is that the functional form assumption on $E(UU'|X)$ can be wrong, which then nullifies the advantage and makes the GLS asymptotic variance formula invalid.

Recall that, when Ω is diagonal, we have two ways to proceed: one is doing LSE with an asymptotic variance estimator allowing for an unknown form of heteroskedasticity, and the other is specifying the form of Ω to do WLS; the latter is more efficient if the specified form is correct. When Ω is not diagonal, an example of which is provided below, again we can think of two ways to proceed, one of which is specifying the form of Ω to do GLS. The other way would be doing LSE with an asymptotic variance estimator allowing for an unknown form of Ω . When nonlinear GMM is discussed later, we will see asymptotic variance matrix estimators allowing for an unknown form of heteroskedasticity and correlations among u_1, \dots, u_N .

To see an example of GLS with a specified non-diagonal Ω , consider the following model with dependent error terms (the so-called “*auto-regressive errors of order one*”):

$$\begin{aligned} y_t &= x_t' \beta + u_t, & u_t &= \rho u_{t-1} + v_t, & |\rho| &< 1, \\ & & u_0 &= 0, & t &= 1, \dots, T. \end{aligned}$$

$$\{v_t\} \text{ are iid with } E(v) = 0 \text{ and } E(v^2) \equiv \sigma_v^2 < \infty, \text{ and independent of } x_1, \dots, x_T.$$

By substituting u_{t-1}, u_{t-2}, \dots successively, we get

$$u_t = \rho u_{t-1} + v_t = \rho^2 u_{t-2} + v_t + \rho v_{t-1} = \rho^3 u_{t-3} + v_t + \rho v_{t-1} + \rho^2 v_{t-2} = \dots,$$

from which $E(u_t^2) \rightarrow \sigma_u^2 \equiv \sigma_v^2 / (1 - \rho^2)$ as $t \rightarrow \infty$. Also observe

$$\begin{aligned} E(u_t u_{t-1}) &= E\{(\rho u_{t-1} + v_t) u_{t-1}\} = \rho E(u_{t-1}^2) \simeq \rho \sigma_u^2, \\ E(u_t u_{t-2}) &= E\{(\rho^2 u_{t-2} + v_t + \rho v_{t-1}) u_{t-2}\} = \rho^2 E(u_{t-2}^2) \simeq \rho^2 \sigma_u^2. \end{aligned}$$

As $\{v_t\}$ are independent of x_1, \dots, x_T and u_t consists of $\{v_t\}$, u_t is independent of x_1, \dots, x_T . Hence, $E(UU'|X) = E(UU')$ and

$$\begin{aligned} \Omega &= \begin{bmatrix} E(u_1 u_1) & \cdots & E(u_1 u_N) \\ \vdots & & \vdots \\ E(u_N u_1) & \cdots & E(u_N u_N) \end{bmatrix} \\ &\simeq \sigma_u^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \cdots & \rho^{N-1} \\ \rho & 1 & \rho & \rho^2 & \cdots & \rho^{N-2} \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ \rho^{N-1} & \rho^{N-2} & \rho^{N-3} & \cdots & \rho & 1 \end{bmatrix}. \end{aligned}$$

To implement the GLS, first do the LSE of y_t on x_t to get the residual \hat{u}_t . Second, replace ρ with the LSE estimator $\hat{\rho}$ of \hat{u}_t on \hat{u}_{t-1} ; σ_u^2 can be replaced by 1 because any scale factor in Ω is canceled in the GLS formula. Third, transform the equation with $\hat{\Omega}^{-1/2}$ and carry out the final LSE on the transformed equation.

6.3 Efficiency of LSE, GLS, and GMM

One may ask why we use LSE instead of some other estimators. For instance, minimizing $N^{-1} \sum_i |y_i - x_i' b|$ may be more natural than LSE. The usual answer found in many econometric textbooks is that LSE has the smallest variance among the “unbiased linear estimators” where a linear estimator a_N should be written as $A \cdot Y$ for some $N \times N$ constant matrix A , and a_N is said to be unbiased for β if $E(a_N) = \beta$. However, this answer is not satisfactory, for unbiasedness is hard to establish for nonlinear estimators. Also, focusing on the linear estimators is too narrow. In the following, we provide a modern answer which shows an optimality of LSE and efficiency comparison of LSE and GLS. The optimality of LSE is implied by the optimality of GMM.

Chamberlain (1987) showed the smallest possible variance (or, the “efficiency bound”) under a general moment condition. His results are valid for nonlinear as well as linear models. For the linear model under the iid assumption on observations, suppose we have a moment condition

$$E(zu) = E\{z(y - x'\beta)\} = 0$$

where $y = x'\beta + u$ and z has at least k components; z may include x and u may be a vector. Using the moment condition only, the smallest possible variance for (“regular”) estimators for β is

$$\{E(xz') \cdot E^{-1}(zuuz') \cdot E(zx')\}^{-1}.$$

This is the asymptotic variance of GMM, which means that GMM is efficient under $E(zu) = 0$, $y = x'\beta + u$, and the iid assumption. When $z = x$ and u is a scalar, the efficiency bound becomes

$$E^{-1}(xx') \cdot E(xx'u^2) \cdot E^{-1}(xx')$$

which is the asymptotic variance of LSE. Thus LSE is the most efficient under the moment condition $E(xu) = 0$, the linear model, and the iid assumption.

Chamberlain (1987) also showed that if

$$E(u|z) = E(y - x'\beta|z) = 0$$

then the smallest possible variance (or the efficiency bound) is

$$E_z^{-1} \left\{ E \left(\frac{\partial (y - x'\beta)}{\partial \beta} | z \right) \cdot E^{-1}(uu'|z) \cdot E \left(\frac{\partial (y - x'\beta)}{\partial \beta'} | z \right) \right\}.$$

If z includes x , then this becomes

$$E_z^{-1} \{ -x \cdot E^{-1}(uu'|z) \cdot (-x') \}.$$

If $z = x$ and u is a scalar, then the bound becomes the asymptotic variance of GLS

$$E^{-1} \left\{ \frac{xx'}{V(u|x)} \right\}.$$

Interestingly, if the error term is homoskedastic, then the two bounds under $E(xu) = 0$ and $E(u|x) = 0$ agree:

$$\sigma^2 E^{-1}(xx').$$

This observation might be, however, misleading, because the homoskedasticity condition is an extra information which could change the efficiency bound.

Observe $E\{xx'/V(u|x)\} = E[\{x/SD(u|x)\}\{x'/SD(u|x)\}]$ which is the “variation” of $x/SD(u|x)$. Also observe

$$\begin{aligned} E(xx')E^{-1}(xx'u^2)E(xx') &= E(xx')E_x^{-1} \{xx'E_{u|x}(u^2)\} E(xx') \\ &= E \left\{ \frac{x}{SD(u|x)} x' SD(u|x) \right\} E^{-1} [\{x SD(u|x)\} \{x' SD(u|x)\}] E \left\{ x SD(u|x) \frac{x'}{SD(u|x)} \right\} \end{aligned}$$

which is the variation of the projection of $x/SD(u|x)$ on $x \cdot SD(u|x)$. This shows that

$$\begin{aligned} E \left\{ \frac{xx'}{V(u|x)} \right\} &\geq E(xx')E^{-1}(xx'u^2)E(xx') \\ \iff E^{-1} \left\{ \frac{xx'}{V(u|x)} \right\} &\leq E^{-1}(xx')E(xx'u^2)E^{-1}(xx') : \end{aligned}$$

GLS is more efficient than LSE.

The condition $E(u|z) = 0$ is stronger than $E(zu) = 0$, because $E(u|z) = 0$ implies $E\{g(z)u\} = E\{g(z)E(u|z)\} = 0$ for any square-integrable function $g(z)$ (i.e., $E\{g(z)^2\} < \infty$). Under the stronger moment condition, the efficiency bound becomes smaller, which is attained by GLS. But this comes at the price that GLS should specify correctly the form of heteroskedasticity. Also, the known parametric form of heteroskedasticity is an extra information which may change the efficiency bound.

Micro-Econometrics

Methods of Moments and Limited Dependent Variables

Lee, M.-j.

2010, XXVII, 770 p., Hardcover

ISBN: 978-0-387-95376-2