Economics 241B
Large Sample Distribution of the OLS Estimator

The importance of the OLS procedure, originally developed for the classic regression model, is that it has good asymptotic properties for a family of models, distinct from the classic model, that are useful in economics. Of the models in the family, the model presented here has the widest range of application.

No specific distribution assumption for the error (such as the Gaussian assumption) is required to derive an asymptotic distribution for the OLSE. The requirement in finite sample theory that the regressors be strictly exogenous (or "fixed"), is replaced by a much weaker assumption that the regressors be predetermined.

The Model

We use the term data generating process (DGP) for the stochastic process that generated the finite sample (Y, X). Therefore, if we specify the DGP, then the finite sample distributions of (Y, X) can be determined. For finite sample theory, in which the number of distributions was fixed and finite, the model is defined as a set of finite sample distributions. In contrast, for asymptotic theory the model is stated as a set of DGP's. The model that we study is the set of DGP's that satisfy the following set of assumptions.

Assumption 2.1 (linearity):

$$Y_t = X_t'\beta + U_t \qquad (t = 1, \ldots, n,)$$

where X_t is a K-dimensional vector of regressors, β is a K-dimensional vector of coefficients and U_t is the latent error.

Assumption 2.2 (ergodic stationarity): *The $(K+1)$-dimensional vector stochastic process $\{Y_t, X_t\}$ is jointly stationary and ergodic.*

Assumption 2.3 (predetermined regressors): *All regressors are predetermined, in the sense that they are orthogonal to the contemporaneous error: $E(X_{tk}U_t) = 0$ for all t and $k(= 1, 2, \ldots, K)$. This can be written as*

$$E(g_t) = 0 \qquad \text{where } g_t \equiv X_t \cdot U_t.$$

Assumption 2.4 (rank condition): *The $K \times K$ matrix $E(X_t X_t')$ is nonsingular (and hence finite). We denote this matrix by Σ_{XX}.*

Assumption 2.5 (g_t is a martingale difference sequence with finite second moments): *$\{g_t\}$ is a martingale difference sequence (so by definition $E(g_t) = 0$). The $K \times K$ matrix of cross moments, $E(g_t g_t')$, is nonsingular. Let S denote $Avar(\bar{g})$ (the variance of the asymptotic distribution of $\sqrt{n}\bar{g}$, where $\bar{g} = \frac{1}{n}\sum_t g_t$). By Assumption 2.2 and the Ergodic Stationary Martingale Differences CLT, $S = E(g_t g_t')$.*

Assumption 2.1 follows directly from the classic regression model. The remaining assumptions require comment.

- (Ergodic stationarity) We in no way rule out cross-section data, as a trivial, but important special case is a random sample (that is, $\{Y_t, X_t\}$ is i.i.d.). Indeed, once independence has been assumed, large sample results can be derived for the case where $\{Y_t, X_t\}$ is independently but not identically distributed (i.n.i.d.) provided that some conditions on the higher moments of (U_t, X_t) are satisfied.

- The model also accommodates conditional heteroskedasticity. If $\{Y_t, X_t\}$ is stationary, then the error $U_t = Y_t - X_t'\beta$ is also stationary. Thus Assumption 2.2 implies that the unconditional second moment $E(U_t^2)$ - if it exists and is finite - is constant across t. That is, the error is unconditionally homoskedastic. As we have seen in our lecture on GLS, an error can be conditionally heteroskedastic and unconditionally homoskedastic.

- (Predetermined) The definition of predetermined used above is not universal. Another common definition of predetermined regressors is that all previous values of regressors are orthogonal to the error: $E(X_{t-j} \cdot U_t) = 0$ for all $j \geq 0$, not just for $j = 0$.

- Sometimes the orthogonality restriction $E(X_t \cdot U_t) = 0$ is replaced with the stronger assumption that $E(U_t|X_t) = 0$. This is stronger than the orthogonality assumption because it implies that for any (measurable) function f of X_t, $f(X_t)$ is orthogonal to U_t:

  $$E[f(X_t) U_t] = E[E(f(X_t) U_t|X_t)] = E[f(X_t) E(U_t|X_t)] = 0.$$

  The stronger conditional mean assumption identifies $X_t'\beta$ as the conditional mean of Y_t, which in essence implies that there are no other functions of the regressors that have been omitted from the model (and in that sense, the model is correctly specified). The weaker orthogonality equation makes no such claim, as X_t and X_t^2 can be uncorrelated, there is no sense with the weaker equation that the model is not omitting regressors. In rational expectations models the stronger assumption is satisfied, but to develop asymptotic theory we need only the weaker assumption. (On a minor, statistical note consider the example of bivariate binomial random variables: $(x = 1, u = 1)$ and $(x = -1, u = 2)$ each of which occurs with probability $\frac{1}{2}$. Orthogonality holds, but not zero conditional means.)

- The regressors are not required to be strictly exogenous. Recall that strict exogeneity requires that $E(X_{sk}U_t) = 0$ for all s and t, not just for $s = t$, which rules out the possibility that the error term U_t is correlated with future regressors X_{t+j} for $j > 0$. Assumption 2.3 ruling out only the contemporaneous correlation, does not restrict this possibility. For example, the AR(1) process, which does not satisfy the strict exogeneity assumption of the classic model, can be accommodated with the model presented here.

- (Rank condition implies no multicollinearity in the limit) Because $E(X_tX_t')$ is finite by Assumption 2.4, $\lim_{n\to\infty} S_{XX} = \Sigma_{XX}$ (where $S_{XX} = \frac{1}{n}\sum_t X_tX_t'$) with probability 1 by the Ergodic Theorem. So, for n sufficiently large the sample cross moment of the regressors S_{XX} is nonsingular by Assumptions 2.2 and 2.4. Because $S_{XX} = \frac{1}{n}X'X$ is nonsingular if and only if $rank(X) = K$, Assumption 1.3 (no multicollinearity) is satisfied with probability one for sufficiently large n. In the OLS formula $B = S_{XX}^{-1}S_{XY}$ (where $S_{XY} = \frac{1}{n}\sum_t X_tY_t$), S_{XX} needs to be inverted. If S_{XX} is singular in a finite sample (and so cannot be inverted), simply assign an arbitrary value to B so that the OLS estimator is well defined in any sample.

- (A sufficient condition for $\{g_t\}$ to be an m.d.s.) Because an m.d.s. is zero mean by definition, Assumption 2.5 is stronger than Assumption 2.3. We need Assumption 2.5 to prove the asymptotic normality of the OLS estimator. As the condition can be hard to interpret, consider a sufficient condition that is easier to interpret

$$E(U_t|U_{t-1}, U_{t-2}, \ldots, U_1, X_t, X_{t-1}, \ldots, X_1) = 0.$$

  We follow the analysis just mentioned in comparing the orthogonality condition with a conditional mean condition. We have that the current error term is uncorrelated with any function of the current and lagged values of the regressors. Moreover, the error term is serially uncorrelated, as the current error is uncorrelated with any function of lagged errors. To see that this condition is sufficient for the m.d.s. assumption, we have the following.

$$E\left(g_t | g_{t-1}, \ldots, g_1\right) = E\left[E\left(g_t | U_{t-1}, U_{t-2}, \ldots, U_1, X_t, X_{t-1}, \ldots, X_1\right) | g_{t-1}, \ldots, g_1\right],$$

which follows from the law of iterated expectations because the information set on the inside, $(U_{t-1}, U_{t-2}, \ldots, U_1, X_t, X_{t-1}, \ldots, X_1)$, contains more information than the information set on the outside $(g_{t-1}, \ldots, g_1)$. Therefore

$$\begin{aligned} E\left(g_t | g_{t-1}, \ldots, g_1\right) &= E\left[X_t E\left(U_t | U_{t-1}, U_{t-2}, \ldots, U_1, X_t, X_{t-1}, \ldots, X_1\right) | g_{t-1}, \ldots, g_1\right] \\ &= 0. \end{aligned}$$

- Inclusion of a constant in the regressors allows us to recast Assumptions 2.3 and 2.5. With a constant in the regression we typically set $X_{t1} = 1$ for all t. Now Assumption 2.3 implies that the mean of the error is zero (which is implied by $E\left(X_{t1} U_t\right) = 0$) and that the contemporaneous correlation between the regressors and error is zero (which is implied by $E\left(X_{tk} U_t\right) = 0$ for $k \neq 1$ and $E\left(U_t\right) = 0$). Hence orthogonality and zero correlation are not quite the same statement, as inclusion of a constant in the regressors is needed for orthogonality to imply zero correlation. Also, because the first element of g_t $(\equiv X_t \cdot U_t)$ is U_t, Assumption 2.5 implies

  $$E\left(U_t | g_{t-1}, g_{t-2}, \ldots, g_1\right) = 0.$$

  Then, by the law of iterated expectations, $\{U_t\}$ is a scalar m.d.s.:

  $$E\left(U_t | U_{t-1}, U_{t-2}, \ldots, U_1\right) = 0.$$

  Therefore, Assumption 2.5 implies that the error by itself is an m.d.s. and hence is serially uncorrelated.

- (Assumptions on $\{g_t\}$ are linked to assumptions on $\{X_t\}$ and $\{U_t\}$) To understand the point, we focus on the i.i.d. case. The assumption that $\{g_t\}$ is an i.i.d. sequence is distinct from the assumptions that $\{X_t\}$ and $\{U_t\}$ are each sequences of i.i.d. random variables. The latter does not imply the former, because X_t could depend on U_{t-1}, in which case g_t $(= X_t U_t)$ is dependent on g_{t-1} $(= X_{t-1} U_{t-1})$. A sufficient condition is that $\{X_t\}$ be independent of $\{U_t\}$, which implies that all collections of $\{X_t\}$

are independent of all collections of $\{U_t\}$. (For example, (X_{t-1}, X_t) is independent of (U_t) and (X_{t-1}, X_t) is independent of (U_t, U_{t-1}).) As g_t contains only X_t we do not need independence of all collections of $\{X_t\}$ from $\{U_t\}$, we merely need each element of $\{X_t\}$ to be independent of $\{U_t\}$. As the same logic holds for $\{U_t\}$, we need only assume that each element of $\{X_t\}$ is independent of every element of $\{U_t\}$.

- (S is a matrix of fourth moments) Because $g_t \equiv X_t U_t$, the quantity S in Assumption 2.5 is $E\left(U_t^2 X_t X_t'\right)$. The (k, j) element of S is $E\left(U_t^2 X_{tk} X_{tj}\right)$, which is a fourth moment (the expectation of products of four random variables). Consistent estimation of S will require a further assumption.

- (S will take a different expression without Assumption 2.5) Thanks to the assumption that $\{g_t\}$ is an m.d.s., S $(= Avar\,(\bar{g}))$ is equal to $E\left(g_t g_t'\right)$. Without the assumption, the expression for S is more complicated and involves the autocovariances of g_t.

Asymptotic Distribution of the OLS Estimator

We now prove that the OLS estimator is consistent and asymptotically normal. The OLS estimator depends on the sample size n, although we do not explicitly note this dependence. The number of regressors is held fixed at K as we track the OLS estimator for different values of n. At this point we presume there is a consistent estimator of S, denoted $\hat{S}$.

Proposition 2.1 (asymptotic distribution of the OLS estimator):
(a) (Consistency of B for β) Under Assumptions 2.1-2.4, $B \rightarrow_p \beta$ as $n \rightarrow \infty$. (So Assumption 2.5 is not needed for consistency.)
(b) (Asymptotic normality of B) If Assumption 2.3 is strengthened by Assumption 2.5, then

$$\sqrt{n}\,(B - \beta) \underset{d}{\rightarrow} N\left(0, Avar\,(B)\right) \quad as n \rightarrow \infty,$$

where

$$Avar\,(B) = \Sigma_{XX}^{-1}\; S\; \Sigma_{XX}^{-1}.$$

(Recall: $\Sigma_{XX} = E\left(X_t X_t'\right)$, $S = E\left(g_t g_t'\right)$, $g_t \equiv X_t \cdot U_t$.)
(c) (Consistent estimator of $Avar\,(B)$) Suppose there is available a consistent estimator $\hat{S}$, of the $K \times K$ matrix S. Then under Assumption 2.2, $Avar\,(B)$ is

consistently estimated by

$$\widehat{Avar\,(B)} = S_{XX}^{-1}\;\hat{S}\;S_{XX}^{-1},$$

where S_{XX} *is the sample mean of* $X_t X_t'$*:*

$$\underset{(K\times K)}{S_{XX}} \equiv \frac{1}{n}\sum_{t=1}^{n} X_t X_t'.$$

Proof of this proposition introduces the standard tricks of asymptotic analysis. To prove (a) and (b), three tricks are employed: (1) write the object in question in terms of sample moments, (2) apply the relevant LLN (the Ergodic Theorem in the present context) and CLT (the ergodic stationary Martingale Differences CLT) to sample moments, and (3) use Lemma 2.4(c) (the Slutzky Theorem) to derive the asymptotic distribution. Proof of (c) is not detailed, as it is an immediate consequence of ergodic stationarity.

Proof (Parts (a) and (b)).

(1) We first write the sampling error $B-\beta$ in terms of sample means.

$$\begin{aligned} B-\beta &= \left(\frac{1}{n}\sum_{t=1}^{n} X_t X_t'\right)^{-1}\left(\frac{1}{n}\sum_{t=1}^{n} X_t U_t\right) \\ &= S_{XX}^{-1}\;\bar{g}. \end{aligned}$$

The sample moments S_{XX} and $\bar{g}$ depend upon the sample size n, although the notation does not make the dependence explicit.

(2) (Consistency) Because $\{X_t X_t'\}$ is ergodic stationary under Assumption 2.2, $S_{XX} \rightarrow_p \Sigma_{XX}$. (The convergence is almost surely, but almost sure convergence implies convergence in probability.) Because Σ_{XX} is invertible under Assumption 2.4, $S_{XX}^{-1} \rightarrow_p \Sigma_{XX}^{-1}$ by Lemma 2.3(a). By the ergodic theorem, $\bar{g} \rightarrow_p E\,(g_t)$, which equals 0 by Assumption 2.3. So by Lemma 2.3(a), $S_{XX}^{-1}\;\bar{g} \rightarrow_p \Sigma_{XX}^{-1}\;0 = 0$. Therefore $p\lim_{n\to\infty}(B-\beta) = 0$, which implies $B \rightarrow_p \beta$ as $n \to \infty$.

(3) (Asymptotic normality) Rewrite the expression for $B-\beta$ as

$$\sqrt{n}\,(B-\beta) = S_{XX}^{-1}\;\left(\sqrt{n}\bar{g}\right).$$

From the ergodic stationary Martingale Differences CLT, $\sqrt{n}\bar{g} \rightarrow_d N\,(0,S)$. So by Lemma 2.4(c), $\sqrt{n}\,(B-\beta)$ converges in distribution to a normal distribution with mean 0 and variance $\Sigma_{XX}^{-1} S \left(\Sigma_{XX}^{-1}\right)'$. Because Σ_{XX} is symmetric, this expression equals the expression for $Avar\,(B)$ given in Proposition 2.1.

Q.E.D.

The results says that the distribution of $\sqrt{n}$ times the sampling error is approximated arbitrarily well by a normal distribution when the sample size is sufficiently large. Although the result holds for all DGP's that satisfy the assumption, the adequacy of the approximation for a given sample size varies over DGP's within this class. The adequacy of approximations can be understood through several means, one of which is Monte Carlo experiments.

s^2 is Consistent

The OLS estimator, s^2, is a consistent estimator of the error variance.

Proposition 2.2 (consistent estimation of error variance): *Let $\hat{U}_t = Y_t - X_t'B$ be the OLS residual for observation t. Under Assumptions 2.1-2.4,*

$$s^2 \equiv \frac{1}{n-K}\sum_{t=1}^{n}\hat{U}_t^2 \underset{p}{\rightarrow} E\left(U_t^2\right),$$

provided $E\left(U_t^2\right)$ exists and is finite.

If we could observe the error term, the obvious estimator would be the sample mean of U_t^2. It is consistent by ergodic stationarity. The message of Proposition 2.2 is that substitution of the OLS residual $\hat{U}_t$ for the population error U_t does not impair consistency. We sketch the proof, because understanding how to handle the discrepancy between U_t and its estimate $\hat{U}_t$ is useful in other contexts as well. Because

$$s^2 = \frac{n}{n-K}\left(\frac{1}{n}\sum_{t=1}^{n}\hat{U}_t^2\right),$$

it suffices to prove that the sample mean of $\hat{U}_t^2$, $\frac{1}{n}\sum_t \hat{U}_t^2$, converges in probability to $E\left(U_t^2\right)$. The relation between $\hat{U}_t$ and U_t is given by

$$\begin{aligned}\hat{U}_t &= Y_t - X_t'B \\ &= Y_t - X_t'\beta - X_t'\left(B-\beta\right) \\ &= U_t - X_t'\left(B-\beta\right),\end{aligned}$$

so that

$$\hat{U}_t^2 = U_t^2 - 2\left(B-\beta\right)'X_t \cdot U_t + \left(B-\beta\right)'X_tX_t'\left(B-\beta\right).$$

Summing over t we obtain

$$\begin{aligned}
\frac{1}{n}\sum_{t=1}^{n}\hat{U}_t^2 &= \frac{1}{n}\sum_{t=1}^{n}U_t^2 - 2\left(B-\beta\right)'\frac{1}{n}\sum_{t=1}^{n}X_t\cdot U_t + \left(B-\beta\right)'\left(\frac{1}{n}\sum_{t=1}^{n}X_tX_t'\right)\left(B-\beta\right) \\
&= \frac{1}{n}\sum_{t=1}^{n}U_t^2 - 2\left(B-\beta\right)'\bar{g} + \left(B-\beta\right)'S_{XX}\left(B-\beta\right).
\end{aligned}$$

The last two terms converge in probability to zero, so $plim\frac{1}{n}\sum_{t=1}^{n}U_t^2$ equals $plim\frac{1}{n}\sum_{t=1}^{n}\hat{U}_t^2$. Note, all that is needed for the proof is a consistent estimator of β. If the OLS estimator is replaced with another consistent estimator in forming the residuals, then the error variance estimator remains consistent for $E\left(U_t^2\right)$.

One can also derive the limit distribution of S^2:[1]

$$\begin{aligned}
n^{\frac{1}{2}}\left(S^2-\sigma^2\right) &= n^{\frac{1}{2}}\frac{1}{n}\sum_{t=1}^{n}\left(\hat{U}_t\right)^2 - n^{\frac{1}{2}}\sigma^2 \\
&= n^{\frac{1}{2}}\frac{1}{n}\sum_{t=1}^{n}U_t^2 - n^{\frac{1}{2}}\sigma^2 + o_P\left(1\right) \\
&= \frac{1}{n^{\frac{1}{2}}}\sum_{t=1}^{n}\left(U_t^2-\sigma^2\right) \xrightarrow{D} N\left(0, EU_t^4-\sigma^4\right),
\end{aligned}$$

where convergence in distribution is implied by the ergodic stationary Martingale Differences CLT.

[1] A sequence of random variables z_n is $o_P\left(n^{\delta}\right)$ if $n^{-\delta}z_n \xrightarrow{P} 0$.