Advanced Econometrics

Professor: Julie Shi

Timekeeper: Rui Zhou

Fall 2023

Contents

2 Properties of OLS Estimators

2.1 Finite Sample Properties of Least Squares Estimators

An “estimator” is a strategy, or formula for using the sample data that are drawn from a population. The “properties” of that estimator are a description of how that estimator can be expected to behave when it is applied to a sample of data. To consider an example, the concept of unbiasedness implies that “on average” an estimator (strategy) will correctly estimate the parameter in question; it will not be systematically too high or too low. It seems less than obvious how one could know this if they were only going to draw a single sample of data from the population and analyze that one sample. The argument adopted in classical econometrics is provided by the sampling properties of the estimation strategy. A conceptual experiment lies behind the description. One imagines “repeated sampling” from the population and characterizes the behavior of the “sample of samples”. The underlying statistical theory of the estimator provides the basis of the description.

2.1.1 Unbiasedness

The least squares estimator is unbiased in every sample.

$$\begin{aligned} b &= (X'X)^{-1}X'y = (X'X)^{-1}X'(X\beta+\varepsilon) = \beta + (X'X)^{-1}X'\varepsilon \\ \Longrightarrow &\mathrm{E}\left[b|X\right] = \beta + \mathrm{E}\left[(X'X)^{-1}X'\varepsilon|X\right] \end{aligned}$$

By assumption of exogeneity, the second term is **0**, so

$$\mathrm{E}\left[b|X\right] = \beta$$

Therefore,

$$\mathrm{E}\left[b\right] = \mathrm{E}_X\left[\mathrm{E}\left[b|X\right]\right] = \mathrm{E}_X\left[\beta\right] = \beta$$

The interpretation of this result is that, for any particular set of observations X, the least squares estimator has expectation β. Therefore, when we average this over the possible values of X, we find the unconditional mean is β as well.

2.1.1.1 Bias From Variable Omission From what we have discussed above, we can also see bias caused by omission of relevant variables. The most common one of specification errors are the omission of relevant variables and the inclusion of superfluous (irrelevant) variables. Suppose that a correctly specified regression model would be

$$y = X_1\beta_1 + X_2\beta_2 + \varepsilon$$

where the two parts of X have K_1 and K_2 columns, respectively. If we regress y on X_1, without including X_2, then the estimator is

$$b_1 = (X_1'X_1)^{-1}X_1'y = \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 + (X_1'X_1)^{-1}X_1'\varepsilon$$

Taking the expectation, we see that unless $X_1'X_2 = \mathbf{0}$ or $\beta_2 = \mathbf{0}$, b_1 is biased. The well-known result is the **omitted variable formula**:

$$\mathrm{E}\left[b_1|X\right] = \beta_1 + P_{1,2}\beta_2, \text{ where } P_{1,2} = (X_1'X_1)^{-1}X_1'X_2$$

Note that each column of the $K_1 \times K_2$ matrix $P_{1,2}$ is the column of slopes in the least squares regression of the corresponding column of X_2 on the columns of X_1.

Alternatively, we can view the omission of a set of variables as equivalent to imposing an incorrect restriction on the correct regression model. In particular, omitting X_2 is equivalent to *incorrectly* estimating the correct model subject to the restriction $\beta_2 = \mathbf{0}$. Incorrectly imposing a restriction produces a biased estimator. Another way to view this error is to note that it amounts to incorporating incorrect information in our estimation. Suppose, however, that our error is simply a failure to use some information that is correct. In this view, we can assert without formal proof that inclusion of irrelevant variables in the regression will not affect unbiasedness, but may face issues of overfitting and a larger covariance matrix for OLS estimators then.

2.1.1.2 Multicollinearity As a response to what appears to be a "multicollinearity problem," it is often difficult to resist the temptation to drop what appears to be an offending variable from the regression, if it seems to be the one causing the problem. This "strategy" creates a subtle dilemma for the analyst. Consider the partitioned multiple regression

$$y = X_1\beta_1 + X_2\beta_2 + \varepsilon$$

If we regress y only on X_1, the estimator is biased:

$$\begin{aligned} b_1 &= (X_1'X_1)^{-1}X_1'y \\ &= (X_1'X_1)^{-1}X_1'(X_1\beta_1 + X_2\beta_2 + \varepsilon) \\ &= \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 + (X_1'X_1)^{-1}X_1'\varepsilon \\ \Longrightarrow &\begin{cases} \mathrm{E}\left[b_1|X\right] = \beta_1 + P_{1,2}\beta_2 \\ b_1 - \mathrm{E}\left[b_1|X\right] = (X_1'X_1)^{-1}X_1'\varepsilon \end{cases} \end{aligned}$$

The covariance matrix of this estimator is

$$\mathrm{Var}\left[b_1|X\right] = \sigma^2(X_1'X_1)^{-1}$$

Note that the covariance matrix is around the $\mathrm{E}\left[b_1|X\right]$ instead of β_1 (Coincidentally in the unbiased estimator case, $\mathrm{E}\left[b_k\right] = b_k$). If β_2 is not actually zero, then in the multiple regression of y on (X_1, X_2), the variance of $b_{1.2}$ around its mean, β_1, would be

$$\begin{aligned} \mathrm{Var}\left[b_{1.2}|X\right] &= \sigma^2\left(X_1'M_2X_1\right)^{-1} \\ &= \sigma^2\left[X_1'X_1 - X_1X_2({X_2}'X_2)^{-1}{X_2}'X_1\right]^{-1} \end{aligned}$$

We compare the two covariance matrices. It is simpler to compare the inverses. Thus,

$$\{\mathrm{Var}\,[b_1|X]\}^{-1} - \{\mathrm{Var}\,[b_{1.2}|X]\}^{-1} = \frac{1}{\sigma^2} X_1' X_2 ({X_2}' X_2)^{-1} {X_2}' X_1$$

The difference matrix is a non-negative definite matrix. The implication is that, the variance of b_1 is not larger than the variance of $b_{1.2}$. It follows that although b_1 is biased, its variance is never larger than the variance of the unbiased estimator. In any realistic case, $X_1' X_2 \neq \mathbf{0}$, in fact it will be smaller.

2.1.2 Efficiency

If the regressors can be treated as non-stochastic, as they would be in an experimental situation in which the analyst chooses the values in X, then the sampling variance of the least squares estimator can be derived by treating X as a matrix of constants. Alternatively, we can allow X to be stochastic, do the analysis conditionally on the observed X, then consider averaging over X as we did in analysis of unbiasedness.

$$b = (X'X)^{-1}X'y = (X'X)^{-1}X'(X\beta + \varepsilon) = \beta + (X'X)^{-1}X'\varepsilon$$

From the result we can see, b is a linear function of the disturbances (which by definition we will see, makes it a linear estimator). As we have seen, the expected value of the second term is $\mathbf{0}$. Therefore, regardless of the distribution of ε, under our other assumptions, b is a linear, unbiased estimator of β. By assumption 4, $\mathrm{Var}\,[\varepsilon|X] = \sigma^2\mathbf{I}$. Thus, conditional covariance matrix of the least squares slope estimator is

$$\begin{aligned}
\mathrm{Var}\,[b|X] &= \mathrm{E}\left[(b-\beta)(b-\beta)'\,|X\right] \\
&= \mathrm{E}\left[(X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}|X\right] \\
&= (X'X)^{-1}X'\mathrm{E}\left[\varepsilon\varepsilon'|X\right]X(X'X)^{-1} \\
&= (X'X)^{-1}X'\left(\sigma^2\mathbf{I}\right)X(X'X)^{-1} \\
&= \sigma^2(X'X)^{-1}
\end{aligned}$$

Specifically, suppose that X contains only a constant term and a single regressor $\mathbf{x}$, the lower-right element of $\sigma^2(X'X)^{-1}$ is

$$Var[b|\mathbf{x}] = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Like what we have done in unbiasedness part, if we seek the unconditional covariance matrix:

$$\begin{aligned}
&\mathrm{Var}\,[b|X] = \sigma^2(X'X)^{-1} \\
\Longrightarrow &\mathrm{Var}\,[b] = \mathrm{E}_{\mathbf{X}}\,[\mathrm{Var}\,[b|X]] + \mathrm{Var}_{\mathbf{X}}\,[\mathrm{E}\,[b|X]] \\
\Longrightarrow &\mathrm{Var}\,[b] = \mathrm{E}_{\mathbf{X}}\left[\sigma^2(X'X)^{-1}\right] + \mathrm{Var}_{\mathbf{X}}\,[\beta] \\
\Longrightarrow &\mathrm{Var}\,[b] = \mathrm{E}\left[\sigma^2(X'X)^{-1}\right] = \sigma^2\mathrm{E}\left[(X'X)^{-1}\right]
\end{aligned}$$

Our earlier conclusion is altered slightly. We must replace $(X'X)^{-1}$ with its expected value to get the appropriate covariance matrix, which brings a subtle change in the interpretation of these results. The unconditional variance of b can only be described in terms of the average behavior of X, so to proceed further, it would be necessary to make some assumptions about the variances and covariances of the regressors. We will return to this subject later.

2.1.2.1 Gauss-Markov Theorem After obtaining the result of covariance matrix, we now obtain a general result for the class of linear unbiased estimator of β regarding its efficiency.

Gauss-Markov Theorem In the linear regression model with regressor matrix X, the least squares estimator b is the minimum variance linear unbiased estimator of β. For any vector of constants w, the minimum variance linear unbiased estimator of $w'\beta$ in the regression model is $w'b$, where b is the least squares estimator.

Proof A direct approach to proving this important theorem would be to *define the class of linear and unbiased estimators* ($b_L = Cy$ such that $\mathrm{E}\left[b_L|X\right] = \beta$) and then find the member of that class which has the smallest variance.

Let $b_0 = Cy$ be another linear unbiased estimator of β, where C is a $K \times n$ matrix. If b_0 is unbiased, then

$$\begin{aligned}\mathrm{E}\left[b_0|X\right] &= \mathrm{E}\left[Cy|X\right] = \mathrm{E}\left[C\left(X\beta+\varepsilon\right)|X\right] = \mathrm{E}\left[CX\beta + C\varepsilon|X\right] \\ &= \mathrm{E}\left[CX\beta|X\right] = \beta\end{aligned}$$

which implies that $CX = \mathbf{I}$. There are many candidates. Now construct a "difference matrix" D, $D = C - (X'X)^{-1}X'$. (This idea is somewhat based on the conjecture that our OLS estimator $b = (X'X)^{-1}X'y$ is the best one so we are naturally interested in the difference.) So, $Dy = b_0 - b$. Then,

$$\begin{aligned}\mathrm{Var}\left[b_0|X\right] &= \mathrm{Var}\left[Dy + b|X\right] \\ &= \mathrm{Var}\left[D\left(X\beta+\varepsilon\right) + b|X\right] \\ &= \mathrm{Var}\left[DX\beta + D\varepsilon + b|X\right] \\ &= \mathrm{Var}\left[DX\beta + D\varepsilon + \beta + (X'X)^{-1}X'\varepsilon|X\right] \\ &= \mathrm{Var}\left[\left(DX + I\right)\beta + \left(D + (X'X)^{-1}X'\right)\varepsilon|X\right] \\ &= \mathrm{Var}\left[\left(D + (X'X)^{-1}X'\right)\varepsilon|X\right] \\ &= \left(D + (X'X)^{-1}X'\right)\mathrm{Var}\left[\varepsilon|X\right]\left(D + (X'X)^{-1}X'\right)' \\ &= \sigma^2\left(D + (X'X)^{-1}X'\right)\left(D + (X'X)^{-1}X'\right)' \\ &= \sigma^2(X'X)^{-1} + \sigma^2 DD'\end{aligned}$$

The last equality holds because since we know that $CX = \mathbf{I} = DX + (X'X)^{-1}(X'X)$, so DX must equal $\mathbf{0}$. This way is a bit tedious, another way to see the result is through the direct covariance matrix of b_0:

$$\begin{aligned}\text{Var}\left[b_0|X\right] &= \text{Var}\left[Cy|X\right] = \text{Var}\left[CX\beta + C\varepsilon|X\right] = \text{Var}\left[C\varepsilon|X\right] \\ &= C\text{Var}\left[\varepsilon|X\right]C' = \sigma^2 CC' \\ &= \sigma^2\left(D + (X'X)^{-1}X'\right)\left(D + (X'X)^{-1}X'\right)' \\ &= \sigma^2(X'X)^{-1} + \sigma^2 DD'\end{aligned}$$

Since a quadratic form in DD' is $q'DD'q = (D'q)'(D'q) = ||D'q||^2 \geq 0$, the conditional covariance matrix of b_0 equals that of b plus a non-negative definite matrix. Therefore, every quadratic form in $Var[b_0|X]$ is larger than the corresponding quadratic from in $Var[b|X]$, which establishes the first result.

The proof of the second statement follows from the previous derivation, since the varaince of $w'b$ is a quadratic form in $Var[b|X]$, and likewise for any b_0 and proves that each individual slope estimator b_k is the best linear unbiased estimator of β_k.

Remarks: Gauss-Markov theorem makes no use of assumption of normality of the distribution of the disturbances. Only the first four assumptions are necessary.

We have shown in preceding Gauss-Markov theorem that $\text{Var}\left[b|X\right] \leq \text{Var}\left[b_0|X\right]$ for any linear and unbiased $b_0 \neq b$ and for the specific X in our sample. But if this inequality holds for every particular X, together with unconditional covariance that $\text{Var}\left[b\right] = \sigma^2 \text{E}_{\mathbf{X}}\left[(X'X)^{-1}\right]$, then it must hold for $\text{Var}\left[b\right] = \text{E}_{\mathbf{X}}\left[\text{Var}\left[b|X\right]\right]$. That is, if it holds for every particular X, then it must hold over the average value(s) of X.

The conclusion, therefore, is that the important results we have obtained thus far for the least squares estimator, unbiasedness, and the Gauss–Markov theorem hold whether or not we condition on the particular sample in hand or consider, instead, sampling broadly from the population.

Gauss–Markov Theorem (Concluded) In the linear regression model, the least squares estimator b is the minimum variance linear unbiased estimator of β whether X is stochastic or nonstochastic, so long as the other assumptions of the model continue to hold.

From here on, we will be encountering many times the comparison between covariance matrices. So it shall be better to know something about matrices comparison.

Matrices Comparison

Derivations in econometrics often focus on whether one matrix is "larger" than another.We now consider how to make such a comparison. As a starting point, the two matrices must have the same dimensions. A useful comparison is based on

$$d = x'Ax - x'Bx = x'(A - B)x$$

If d is always positive for any nonzero vector x, then by this criterion, we can say that A is larger than B. The reverse would apply if d is always negative.

It follows from the definition that, if $d > 0$ for all nonzero $\mathbf{x}$, then $A - B$ is positive definite.

If d is only greater than or equal to zero, then $A - B$ is nonnegative definite. Notice that the ordering is not complete. For some pairs of matrices, d could have either sign, depending on $\mathbf{x}$. In this case, there is no simple comparison.

A particular case of the general result which we will encounter frequently is that, if A is positive definite and B is non-negative definite, then $A + B \geq A$. This is often used in variance matrix comparison and see which estimator is more efficient.

Finally, in comparing matrices, it may be more convenient to compare their inverses. The result analogous to a familiar result for scalars is:

$$A > B \Longrightarrow B^{-1} > A^{-1}$$

2.1.2.2 Variance Estimation If we wish to test hypotheses about β or to form confidence intervals, then we will require a sample estimate of the covariance matrix, $\operatorname{Var}[b|X] = \sigma^2(X'X)^{-1}$. The population parameter σ^2 remains to be estimated. Since σ^2 is the expected value of ε_i^2 and e_i is an estimate of ε_i, by analogy,

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2$$

would seem to be a natural estimator. But the least squares residuals are imperfect estimates of their population counterparts; $e_i = y_i - x_i'b = \varepsilon_i - x_i'(b - \beta)$. **The estimator is distorted (as might be expected) because β is not observed directly**. The expected square on the right-hand side involves a second term that might not have expected value zero.
The least squares residuals are

$$e = My = M\left(X\beta + \varepsilon\right) = M\varepsilon$$

As $MX = \mathbf{0}$. An estimator of σ^2 will be based on the sum of squared residuals:

$$e'e = \varepsilon' M\varepsilon$$

The expected value of this quadratic form is $\mathrm{E}\left[e'e|X\right] = \mathrm{E}\left[\varepsilon'M\varepsilon|X\right]$. Notice that the scalar $\varepsilon'M\varepsilon$ is a 1×1 matrix, namely a number, so it is equal to its trace. By using the result on cyclic permutations,

$$\begin{aligned}
\mathrm{E}\left[\operatorname{tr}\left(\varepsilon'M\varepsilon\right)|X\right] &= \mathrm{E}\left[\operatorname{tr}\left(M\varepsilon\varepsilon'\right)|X\right] \\
&= \operatorname{tr}\left(M\mathrm{E}\left[\varepsilon\varepsilon'|X\right]\right) \\
&= \operatorname{tr}\left(M\left(\sigma^2\mathbf{I}\right)\right) \\
&= \sigma^2\operatorname{tr}\left(M\right)
\end{aligned}$$

The trace of M is

$$\begin{aligned}
\operatorname{tr}(M) &= \operatorname{tr}\left(\mathbf{I}_n - X(X'X)^{-1}X'\right) \\
&= \operatorname{tr}\left(\mathbf{I}_n\right) - \operatorname{tr}\left(X(X'X)^{-1}X'\right) \\
&= \operatorname{tr}\left(\mathbf{I}_n\right) - \operatorname{tr}\left[(X'X)^{-1}(X'X)\right] \\
&= \operatorname{tr}\left(()\,\mathbf{I}_n\right) - \operatorname{tr}\left(\mathbf{I}_K\right) \\
&= n - K
\end{aligned}$$

Therefore,

$$\mathrm{E}\left[e'e|X\right] = (n-K)\sigma^2$$

An unbiased estimator of σ^2 is:

$$s^2 = \frac{e'e}{n-K}$$

So the natural estimator, $\frac{1}{n}\sum_{i=1}^{n} e_i^2$ is biased toward zero, although the bias becomes smaller as the sample size increases. Notice that the estimator s^2 is unbiased unconditionally as well, since $\mathrm{E}\left[s^2\right] = \mathrm{E}_{\mathbf{X}}\left[\mathrm{E}\left[s^2|X\right]\right] = \mathrm{E}_{\mathbf{X}}\left[\sigma^2\right] = \sigma^2$. The standard error of the regression is then s, the square root of s^2. With s^2, we can then compute

$$\mathrm{Est.Var}\left[b|X\right] = s^2(X'X)^{-1}$$

Henceforth, we shall use the notation $\mathrm{Est.Var}\left[\cdot\right]$ to indicate a sample estimate of the sampling variance of an estimator. The square root of the k-th diagonal element of this matrix, $\left\{s^2(X'X)^{-1}\right\}^{\frac{1}{2}}$, is the standard error of the estimator b_k, which is often denoted simply "the standard error of b_k".

2.1.3 Normality

To this point, our specification and analysis of the regression model are semiparametric. We have not used Assumption A6, normality of ε, in any of our results. The assumption is useful for constructing statistics for forming confidence intervals. As we have shown, b is a linear function of the disturbance vector ε. If we assume that ε has a multivariate normal distribution, then

$$\begin{aligned}
&b = \beta + (X'X)^{-1}X'\varepsilon \\
\Longrightarrow &b|X \sim N\left[\beta, \sigma^2(X'X)^{-1}\right]
\end{aligned}$$

This specifies a multivariate normal distribution, so each element of $b|X$ is normally distributed:

$$b_k|X \sim N\left[\beta_k, \sigma^2(X'X)^{-1}_{kk}\right]$$

The distribution of b is conditioned on X. The normal distribution of b in a finite sample is a consequence of our specific assumption of normally distributed disturbances.

Without this assumption, and without some alternative specific assumption about the distribution of ε, we will not be able to make any definite statement about the exact distribution of b, conditional or otherwise. Later, however, we will be able to obtain an approximate normal distribution for b, with or without assuming normally distributed disturbances and whether the regressors are stochastic or not.

2.2 Large Sample Properties of Least Squares Estimators

Unbiasedness is a useful starting point for assessing the virtues of an estimator. It assures the analyst that their estimator will not persistently miss its target, either systematically too high or too low. However, as a guide to estimation strategy, it has two shortcomings. First, save for the least squares slope estimator we are discussing in this chapter, it is relatively rare for an econometric estimator to be unbiased. In nearly all cases beyond the multiple regression model, the best one can hope for is that the estimator improves in the sense suggested by unbiasedness as more information (data) is brought to bear on the study. As such, we will need a broader set of tools to guide the econometric inquiry. Second, the property of unbiasedness does not, in fact, imply that more information is better than less in terms of estimation of parameters. The sample means of random samples of 100 and $10,000$ are all unbiased estimators of a population mean—by this criterion all are equally desirable. Logically, one would hope that a larger sample is better than a smaller one in some sense that we are about to define (and, by extension, an extremely large sample should be much better, or even perfect). The property of consistency improves on unbiasedness in both of these directions.

2.2.1 Consistency

To begin, we leave the data generating mechanism for X unspecified—X may be any mixture of constants and random variables generated independently of the process that generates ε. We do make two crucial assumptions. The first is a modification of Assumption A5:

$(\mathbf{x}_i, \varepsilon_i), i = 1, \cdots, n$ is a sequence of *independent* observations.

The second concerns the behavior of the data in large samples:

$$\text{plim}\ \frac{X'X}{n} = Q,\ \text{ a positive definite matrix.}$$

The least squares estimator may be written as:

$$\begin{aligned} b &= \beta + (X'X)^{-1}X'\varepsilon \\ &= \beta + \left(\frac{X'X}{n}\right)^{-1}\left(\frac{X'\varepsilon}{n}\right) \end{aligned}$$

If Q^{-1} exists, then

$$\text{plim } b = \beta + Q^{-1}\text{plim } \left(\frac{X'\varepsilon}{n}\right)$$

This stands because the inverse is a continuous function of the original matrix. We require the probability limit of the last term. Let

$$\frac{X'\varepsilon}{n} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\varepsilon_i = \frac{1}{n}\sum_{i=1}^{n}\mathbf{w}_i = \bar{\mathbf{w}}$$

Then,

$$\text{plim } b = \beta + Q^{-1}\text{plim } \bar{\mathbf{w}}$$

From the exogeneity assumption, we have

$$\begin{aligned} \mathrm{E}\left[\mathbf{w}_i\right] &= \mathrm{E}_{\mathbf{x}}\left[\mathrm{E}\left[\mathbf{w}_i|\mathbf{x}_i\right]\right] \\ &= \mathrm{E}_{\mathbf{x}}\left[\mathbf{x}_i\mathrm{E}\left[\varepsilon_i|\mathbf{x}_i\right]\right] \\ &= \mathbf{0} \\ \Longrightarrow \mathrm{E}\left[\bar{\mathbf{w}}\right] &= \mathbf{0} \end{aligned}$$

For any element in $\mathbf{x}_i$ that is non-stochastic, the zero expectations follow from the marginal distribution of ε_i. We now consider the variance:

$$\begin{aligned} \mathrm{Var}\left[\bar{\mathbf{w}}\right] &= \mathrm{E}\left[\mathrm{Var}\left[\bar{\mathbf{w}}|X\right]\right] + \mathrm{Var}\left[\mathrm{E}\left[\bar{\mathbf{w}}|X\right]\right] \\ &= \mathrm{E}\left[\mathrm{Var}\left[\bar{\mathbf{w}}|X\right]\right] + \mathrm{Var}\left[\mathbf{0}\right] \\ &= \mathrm{E}\left[\mathrm{Var}\left[\bar{\mathbf{w}}|X\right]\right] \end{aligned}$$

To obtain the first term, $\mathrm{E}\left[\mathrm{Var}\left[\bar{\mathbf{w}}|X\right]\right]$, w use $\mathrm{E}\left[\varepsilon\varepsilon'|X\right] = \sigma^2\mathbf{I}$, so

$$\begin{aligned} \mathrm{Var}\left[\bar{\mathbf{w}}|X\right] &= \mathrm{E}\left[\bar{\mathbf{w}}\bar{\mathbf{w}}'|X\right] \\ &= \mathrm{E}\left[\left(\frac{X'\varepsilon}{n}\right)\left(\frac{X'\varepsilon}{n}\right)'\middle|X\right] \\ &= \frac{1}{n}X'\mathrm{E}\left[\varepsilon\varepsilon'|X\right]X\frac{1}{n} \\ &= \left(\frac{\sigma^2}{n}\right)\left(\frac{X'X}{n}\right) \end{aligned}$$

Therefore,

$$\mathrm{Var}\left[\bar{\mathbf{w}}\right] = \left(\frac{\sigma^2}{n}\right)\mathrm{E}\left[\frac{X'X}{n}\right]$$

The variance will collapse to zero if the expectation in parentheses is (or converges to) a constant matrix, so that the leading scalar will dominate the product as n increases. It then follows that

$$\lim_{n\to\infty} \operatorname{Var}[\bar{\mathbf{w}}] = 0 \cdot \frac{X'X}{n} = \mathbf{0}$$

Since the mean of $\bar{\mathbf{w}}$ is identically zero and its variance converges to zero, $\bar{\mathbf{w}}$ converges in mean square to zero, so we establish:

$$\text{plim } \bar{\mathbf{w}} = \mathbf{0} \iff \text{plim } \frac{X'\varepsilon}{n} = \mathbf{0}$$

So

$$\text{plim } b = \beta + Q^{-1} \cdot \mathbf{0} = \beta$$

The result establishes that, under the first 4 assumptions and the additional assumption of plim $\frac{X'X}{n} = Q$, b is a consistent estimator of β in the linear regression model.

Time-series settings that involve time trends, polynomial time series, and trending variables often pose cases in which the preceding assumptions are too restrictive. A somewhat weaker set of assumptions about X that is broad enough to include most of these is the **Grenander conditions**. The conditions ensure that the data matrix is "well behaved" in large samples. The assumptions are very weak and likely to be satisfied by almost any data set encountered in practice.

2.2.2 Asymptotic Normality

As a guide to estimation, consistency is an improvement over unbiasedness. Since we are in the process of relaxing the more restrictive assumptions of the model, including normality of the disturbances, we will also lose the normal distribution of the estimator that will enable us to form confidence intervals (covered later). It seems that the more general model we have built here has come at a cost. In this section, we will find that normality of the disturbances is not necessary for establishing the distributional results we need to allow statistical inference including confidence intervals and testing hypotheses. Under generally reasonable assumptions about the process that generates the sample data, large sample distributions will provide a reliable foundation for statistical inference in the regression model.

To derive the asymptotic distribution of the least squares estimator, we need to make use of some basic central limit theorems. So in addition to assumption of exogeneity, we will assume that observations are *independent.* First rewrite the b as

$$b = (X'X)^{-1}X'y = \beta + (X'X)^{-1}X'\varepsilon$$

$$\iff \sqrt{n}\,(b - \beta) = \left(\frac{X'X}{n}\right)^{-1} \left(\frac{1}{\sqrt{n}}\right) X'\varepsilon$$

Since the inverse matrix is a continuous function of the original matrix, plim $\left(\frac{X'X}{n}\right)^{-1} = Q^{-1}$. Therefore, if the limiting distribution of the random vector $\left(\frac{X'X}{n}\right)^{-1}\left(\frac{1}{\sqrt{n}}\right)X'\varepsilon$ exists, then that limiting distribution is the same as that of

$$\left[\text{plim}\ \left(\frac{X'X}{n}\right)^{-1}\right]\left(\frac{1}{\sqrt{n}}\right)X'\varepsilon = Q^{-1}\left(\frac{1}{\sqrt{n}}\right)X'\varepsilon$$

Thus, we must establish the limiting distribution of

$$\left(\frac{1}{\sqrt{n}}\right)X'\varepsilon = \sqrt{n}\left(\bar{\mathbf{w}} - \mathrm{E}\left[\bar{\mathbf{w}}\right]\right)$$

where $\mathrm{E}\left[\bar{\mathbf{w}}\right] = \mathbf{0}$. We can use the multivariate Lindeberg-Feller version of the central limit theorem to obtain the limiting distribution of $\sqrt{n}\bar{\mathbf{w}}$. By definition of $\bar{\mathbf{w}}$, $\bar{\mathbf{w}}$ is the average of n independent random vectors $\mathbf{w}_i = \mathbf{x}_i\varepsilon_i$, with means $\mathbf{0}$ and variances

$$\mathrm{Var}\left[\mathbf{w}_i\right] = \mathrm{Var}\left[\mathbf{x}_i\varepsilon_i\right] = \sigma^2\mathrm{E}\left[\mathbf{x}_i\mathbf{x}_i'\right] = \sigma^2 Q_i$$

The variance of $\sqrt{n}\bar{\mathbf{w}}$ is

$$\sigma^2\bar{Q}_n = \sigma^2\left(\frac{1}{n}\right)\left[Q_1 + Q_2 + \cdots + Q_n\right]$$

As long as the sum is not dominated by any particular term and the regressors are well behaved, which in this case means plim $\frac{X'X}{n} = Q$ holds,

$$\lim_{n\to\infty}\sigma^2\bar{Q}_n = \sigma^2 Q$$

Therefore, we may apply the Lindeberg–Feller central limit theorem to the vector $\sqrt{n}\bar{\mathbf{w}}$. We now have the elements we need for a formal result. If $[\mathbf{x}_i\varepsilon_i], i = 1, 2, \cdots, n$ are independent vectors distributed with mean $\mathbf{0}$ and variance $\sigma^2 Q_i$, and if plim $\frac{X'X}{n} = Q$ holds, then

$$\left(\frac{1}{\sqrt{n}}\right)X'\varepsilon \overset{\mathrm{d}}{\to} N\left[\mathbf{0}, \sigma^2 Q\right]$$

It then follows that

$$\begin{aligned}
&Q^{-1}\left(\frac{1}{\sqrt{n}}\right)X'\varepsilon \overset{\mathrm{d}}{\to} N\left[Q^{-1}\mathbf{0}, Q^{-1}\left(\sigma^2 Q\right)Q^{-1}\right] \\
\Longleftrightarrow\ &Q^{-1}\left(\frac{1}{\sqrt{n}}\right)X'\varepsilon \overset{\mathrm{d}}{\to} N\left[\mathbf{0}, \sigma^2 Q^{-1}\right] \\
\Longleftrightarrow\ &\sqrt{n}\left(b - \beta\right) \overset{\mathrm{d}}{\to} N\left[\mathbf{0}, \sigma^2 Q^{-1}\right]
\end{aligned}$$

Thus, we obtain the asymptotic distribution of b.

Asymptotic Distribution of b with Independent Observations If $\{\varepsilon_i\}$ are independently distributed with mean zero and finite variance σ^2, and x_{ik} is such that the Grenander conditions are met, then

$$b \overset{\text{a}}{\sim} N\left[\beta, \frac{\sigma^2}{n} Q^{-1}\right]$$

In practice, it is necessary to estimate $\dfrac{1}{n}Q^{-1}$ (recall that plim $\frac{X'X}{n} = Q$) with $(X'X)^{-1}$ and σ^2 with $\dfrac{e'e}{n-K}$.

Remarks: If ε is normally distributed, then normality of $b|X$ holds in *every* sample, so it holds asymptotically as well. The important implication of this derivation is that if the regressors are well behaved and observations are independent, then the asymptotic normality of the least squares estimator does not depend on normality of the disturbances; it is a consequence of the central limit theorem.

To complete the derivation of the asymptotic properties of b, we require an estimator of Asy.Var $[b] = \frac{\sigma^2}{n} Q^{-1}$. So we still need to assess the consistency of s^2 as an estimator of σ^2.

$$\begin{aligned} s^2 &= \frac{e'e}{n-K} = \frac{\varepsilon' M \varepsilon}{n-K} \\ &= \frac{1}{n-K} \cdot \left[\varepsilon'\varepsilon - \varepsilon X(X'X)^{-1}X'\varepsilon\right] \\ &= \frac{n}{n-K} \cdot \left[\frac{\varepsilon'\varepsilon}{n} - \left(\frac{\varepsilon' X}{n}\right)\left(\frac{X'X}{n}\right)^{-1}\left(\frac{X'\varepsilon}{n}\right)\right] \end{aligned}$$

The leading constant clearly converges to 1. As we have shown, plim $\left(\frac{X'X}{n}\right)^{-1} = Q^{-1}$, plim $\left(\frac{X'\varepsilon}{n}\right) = 0$. Using the product rule for probability limits, we assert that the second term in the brackets converges to 0. That leaves

$$\bar{\varepsilon}^2 = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2$$

This is a narrow case in which the random variables ε_i^2 are independent with the same finite mean σ^2, so not much is required to get the mean to converge almost surely to $\sigma^2 = \text{E}\left[\varepsilon_i^2\right]$. Only weak conditions about ε_i^2 are needed (skipped for discussions). This gives our result:

$$\text{plim } s^2 = \sigma^2$$

So the appropriate estimator of the asymptotic covariance matrix of b is

$$\text{Est.Asy.Var}\,[b] = s^2(X'X)^{-1}$$

2.3 Interval Estimation

2.3.1 Confidence Interval Construction

The objective of interval estimation is to present the best estimate of a parameter with an explicit expression of the uncertainty attached to that estimate. A general approach, for estimation of a parameter θ, would be

$$\widehat{\theta} \pm \text{ sampling variability}$$

(We are assuming that the interval of interest would be symmetric around $\widehat{\theta}$.) Following the logic that the range of the sampling variability should convey the degree of (un)certainty, we consider the logical extremes. We can be absolutely (100 percent) certain that the true value of the parameter we are estimating lies in the range $\widehat{\theta} \pm \infty$. Of course, this is not particularly informative. At the other extreme, we should place no certainty (0 percent) on the range $\widehat{\theta} \pm 0$. The probability that our estimate precisely hits the true parameter value should be considered zero. The point is to choose a value of α – 0.05 or 0.01 is conventional—such that we can attach the desired confidence (prob- ability), $100(1-\alpha)$ percent to the interval. We consider how to find that range and then apply the procedure to three familiar problems, *interval estimation for one of the regression parameters*, *estimating a function of the parameters* and *predicting the value of the dependent variable in the regression using a specific setting of the independent variables.* For this purpose, we depart from Assumption A6 that the disturbances are normally distributed. We will then relax that assumption and rely instead on the asymptotic normality of the estimator.

Under assumption 6 that the disturbances are normally and independently distributed,

$$\begin{aligned} b_k &\sim N\left[\beta_k, \sigma^2 S^{kk}\right] \\ \Longrightarrow z_k &= \frac{b_k - \beta_k}{\sqrt{\sigma^2 S^{kk}}} \sim N\,[0,1] \end{aligned}$$

Note that z_k, which is a function of b_k, β_k, σ^2 and S^{kk}, nonetheless has a distribution that involves none of the model parameters or the data; z_k is a pivotal statistic.

Using our conventional α (say $\alpha = 95\%$) percent confidence level, we know from distribution of z_k that

$$\Pr\left[-z_{\alpha/2} \leq z_k \leq z_{\alpha/2}\right] = \alpha$$

where $\alpha = 0.05$. By a simple manipulation, we find that

$$\Pr\left[b_k - z_{1-\alpha/2} \cdot \sqrt{\sigma^2 S^{kk}} \leq z_k \leq b_k + z_{1-\alpha/2} \cdot \sqrt{\sigma^2 S^{kk}}\right] = 1 - \alpha$$

Remarks: This is a statement about the probability that the random interval contains β_k, not the probability that β_k lies in the specified interval.

We would have our desired confidence interval as above, save for the complication that σ^2 is unknown, so the interval is not operational. It would seem natural to use s^2 from the regression. This is, indeed, an appropriate approach, but later we will see it follows a different distribution, t distribution with $(n-K)$ degrees of freedom.

To start with our adjustments of σ^2 by s^2, the quantity

$$\frac{(n-K)\,s^2}{\sigma^2} = \frac{e'e}{\sigma^2} = \frac{\varepsilon' M \varepsilon}{\sigma^2} = \left(\frac{\varepsilon}{\sigma}\right)' M \left(\frac{\varepsilon}{\sigma}\right)$$

Notice this quantity is an idempotent quadratic form in a standard normal vector, $\left(\frac{\varepsilon}{\sigma}\right)$. Therefore, it has a chi-squared distribution with degrees of freedom equal to the $\text{rank}(M) = \text{tr}(M) = n - K$.

Till now, to construct the t statistic, we still need the independency of $\left(\frac{\varepsilon}{\sigma}\right)' M \left(\frac{\varepsilon}{\sigma}\right)$ and $\left(\frac{b-\beta}{\sigma}\right)$. To prove this, it suffices to show that $\left(\frac{b-\beta}{\sigma}\right) = (X'X)^{-1}X'\left(\frac{\varepsilon}{\sigma}\right)$ is independent of $\left(\frac{\varepsilon}{\sigma}\right)' M \left(\frac{\varepsilon}{\sigma}\right)$. We use the following result to finish the proof.

Independence of b and s^2 If ε is normally distributed, then the least squares coefficient estimator b is statistically independent of the residual vector e and therefore, all functions of e, including s^2.

Proof A sufficient condition for the independence of a linear form $L\mathbf{x}$ and an idempotent quadratic form $\mathbf{x}'A\mathbf{x}$ in a standard normal vector $\mathbf{x}$ is that, $LA = \mathbf{0}$. Letting $\left(\frac{\varepsilon}{\sigma}\right)$ be the $\mathbf{x}$. The requirement here would be $(X'X)^{-1}X'M = \mathbf{0}$. Notice that $MX = \mathbf{0}$, so the requirement is met. So we have proved the independency of $\left(\frac{\varepsilon}{\sigma}\right)' M \left(\frac{\varepsilon}{\sigma}\right)$ and $\left(\frac{b-\beta}{\sigma}\right)$. Since $\left(\frac{b-\beta}{\sigma}\right)$ is just a function of b, and $s^2 = \left(\frac{\varepsilon}{\sigma}\right)' M \left(\frac{\varepsilon}{\sigma}\right)$, so we can end our proof to the general result.

Therefore, the ratio

$$t_k = \frac{(b_k - \beta_k)/\sqrt{\sigma^2 S^{kk}}}{\sqrt{[(n-K)s^2/\sigma^2]/(n-K)}} = \frac{b_k - \beta_k}{\sqrt{s^2 S^{kk}}}$$

has a t distribution with $(n-K)$ degrees of freedom. We can use t_k to test hypothesis or form confidence intervals about the individual elements of β.

The result of t_k differs from z_k in the use of s^2 instead of σ^2, and in the pivotal distribution, t with $(n-K)$ degrees of freedom, rather than standard normal. It follows that a confidence interval for β_k can be formed using

$$\left[b_k - t_{1-\alpha/2,n-K} \cdot \sqrt{s^2 S^{kk}} \leq \beta_k \leq b_k + t_{1-\alpha/2,n-K} \cdot \sqrt{s^2 S^{kk}}\right] = 1-\alpha$$

Notice here, the distribution of the pivotal statistic depends on the sample size through $(n-K)$, but once again, not on the parameters or the data. The practical advantage of t_k is that it does not involve any unknown parameters.

2.3.2 Prediction

Suppose that we wish to predict the value of y_0 associated with a regressor vector $\mathbf{x}_0$. The actual value would be

$$y_0 = \mathbf{x}_0'\beta + \varepsilon_0$$

It follows from the Gauss-Markov theorem that $\hat{y}_0 = \mathbf{x}_0'b$ is the minimum variance linear unbiased estimator of $\mathrm{E}\left[y_0|\mathbf{x}_0\right] = \mathbf{x}_0'\beta$. The prediction error is

$$e_0 = \hat{y}_0 - y_0 = \mathbf{x}_0'(b-\beta) + \varepsilon_0$$

The prediction variance of this estimator is

$$\begin{aligned} \mathrm{Var}\left[e_0|X,\mathbf{x}_0\right] &= \sigma^2 + \mathrm{Var}\left[\mathbf{x}_0'(b-\beta)|X,\mathbf{x}_0\right] \\ &= \sigma^2 + \mathrm{Var}\left[\mathbf{x}_0'(X'X)^{-1}X'\varepsilon|X,\mathbf{x}_0\right] \\ &= \sigma^2 + \mathbf{x}_0'\left[\sigma^2(X'X)^{-1}\right]\mathbf{x}_0 \\ &= \mathrm{Var}\left[b|X\right] + \mathbf{x}_0'\mathrm{Var}\left[b|X\right]\mathbf{x}_0 \end{aligned}$$

The prediction variance can be estimated by using s^2 in place of σ^2. A confidence (prediction) interval for y_0 would then be formed using

$$\hat{y}_0 \pm t_{1-\alpha/2,n-K} \cdot \mathrm{se}\left(e_0\right)$$