

Sampling Error of the OLS Coefficient estimates—recap

Suppose the assumptions of the classical linear model hold, we fix \mathbf{X} , *and* we know the true parameters of the model, $\boldsymbol{\beta}$ and σ .

Recall that the true model says that ε_i is drawn, iid, from the same distribution for every observation. So the batch of ε_i 's (or if you prefer, the vector $\boldsymbol{\varepsilon}$), will be different for every sample of fixed size.

Suppose we:

- drew an $\boldsymbol{\varepsilon}$ at random (given σ)
- then calculated $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ (*so we have generated some 'fake' data that we know follows the model's true data-generating process (DGP)*)
- then estimated \mathbf{b} by OLS on this fake data, i.e. on \mathbf{y} and \mathbf{X} .

Because $\boldsymbol{\varepsilon}$ (and therefore the \mathbf{y} vector it gives rise to) is random, the estimated \mathbf{b} will in general not equal $\boldsymbol{\beta}$ exactly. Instead, it will deviate from $\boldsymbol{\beta}$ according to the formula $\mathbf{b} - \boldsymbol{\beta} = \underbrace{(\mathbf{X}'\mathbf{X})^{-1}}_{K \times K} \underbrace{\mathbf{X}'\boldsymbol{\varepsilon}}_{K \times 1} \equiv \mathbf{A}\boldsymbol{\varepsilon}$. (notice—if by some bizarre chance, we happened to draw a vector of zeros for $\boldsymbol{\varepsilon}$, this means we'd get \mathbf{b} exactly right).

Lecture 3: Finite-Sample Properties of the OLS Estimator

(Hayashi, pp. 27-31)

A. Finite-Sample Properties: Definition

B The Properties [and the assumptions needed for each: see next page]:

1. Unbiasedness: $E(\mathbf{b}|\mathbf{X}) = \boldsymbol{\beta}$. [1-3]

2. Variance of \mathbf{b} : $\text{Var}(\mathbf{b}|\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. [1-4]

3. *Gauss-Markov Theorem*: The OLS estimator is the BLUE. [1-4]

4. Orthogonality of \mathbf{b} , \mathbf{e} : $\text{Cov}(\mathbf{b}, \mathbf{e} | \mathbf{X}) = \mathbf{0}$, where \mathbf{e} is the OLS residual. [1-4]

5. $s^2 = \frac{SSR}{n-K} = \frac{\mathbf{e}'\mathbf{e}}{n-K}$ is unbiased, i.e. $E(s^2) = \sigma^2$ [1-4]

Quick Recap of Assumptions:

1. **Linearity** y_i is a linear function of the x_i 's, plus an error term, ε_i , *i.e.*:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_{1k} x_{iK} + \varepsilon_i, \text{ or } y_i = \sum_{k=1}^K \beta_j x_{ik} + \varepsilon_i$$

2. **Strict Exogeneity**: The expected value of $\varepsilon_i | \mathbf{X}$ equals zero for every i .

3. **No Multicollinearity**: The columns of \mathbf{X} are linearly independent.

4. **Homoskedasticity**: The variance of ε_i equals a constant, σ^2 for every i ; the covariance of ε across observations is zero.

A. Finite Sample Properties

These are properties that are true for any given sample size n . For example, unbiasedness: this refers to what you would expect, in the limit, if you ran a regression of \mathbf{y} on \mathbf{X} over and over, each time using a sample of size n , and each time taking a new set of independent draws from the distribution of the error term, ε . If the mean of the estimated OLS coefficients converges to the true coefficients, $\boldsymbol{\beta}$, then the OLS coefficients are said to be unbiased.

These properties help us interpret what we get when we estimate an OLS regression on real data, if –when we run that regression– we believe that the data were generated by *one* such draw from the distribution of ε .

B. The Properties

1. Unbiasedness of \mathbf{b} : Under assumptions 1-3 (note we don't need homoskedasticity), $E(\mathbf{b}|\mathbf{X}) = \boldsymbol{\beta}$.

What does it mean? First, fix consider a fixed set of conditioning variables, \mathbf{X} . Now consider a bunch of random draws of the true ($n \times 1$) vector of disturbances, $\boldsymbol{\varepsilon}$. For each such draw, the OLS estimator produces a ($K \times 1$) vector of estimated coefficients, \mathbf{b} . Property 1 says that the expected value of \mathbf{b} equals its true value. It follows that, if we took R draws ("replications") of $\boldsymbol{\varepsilon}$ and let R approach infinity, the mean of the estimated \mathbf{b} 's across these replications would approach the true value, $\boldsymbol{\beta}$.

Proof: First, rewrite the property as $E(\mathbf{b} - \boldsymbol{\beta} | \mathbf{X}) = \mathbf{0}$.

We have already shown (last class) that the sampling error, $\mathbf{b} - \boldsymbol{\beta} = \underbrace{(\mathbf{X}'\mathbf{X})^{-1}}_{K \times K} \underbrace{\mathbf{X}'\boldsymbol{\varepsilon}}_{K \times 1} \equiv \mathbf{A}\boldsymbol{\varepsilon}$.

(note, interestingly and importantly, that this doesn't depend on what the true $\boldsymbol{\beta}$ is)

So, $E(\mathbf{b} - \boldsymbol{\beta} | \mathbf{X}) = E(\mathbf{A}\boldsymbol{\varepsilon} | \mathbf{X}) = \mathbf{A}E(\boldsymbol{\varepsilon} | \mathbf{X})$.

But Assumption 2 (strict exogeneity) says $E(\boldsymbol{\varepsilon} | \mathbf{X}) = \mathbf{0}$. QED.

2. Variance of \mathbf{b} : Under assumptions 1-4, $\text{Var}(\mathbf{b}|\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

What does it mean? Again, fix \mathbf{X} and consider a bunch of realizations of the true disturbance vector, $\boldsymbol{\varepsilon}$. Each such realization produces a $(K \times 1)$ vector of coefficients, \mathbf{b} . If we were to do this over many realizations of $\boldsymbol{\varepsilon}$, the $(K \times K)$ variance-covariance matrix *among the elements of \mathbf{b}* would be given by the true σ^2 (a scalar), times $(\mathbf{X}'\mathbf{X})^{-1}$.

Proof. $\text{Var}(\mathbf{b} | \mathbf{X}) = \text{Var}(\mathbf{b} - \boldsymbol{\beta} | \mathbf{X})$ (because $\boldsymbol{\beta}$ is not random).

$$= \text{Var}(\mathbf{A}\boldsymbol{\varepsilon} | \mathbf{X}) \text{ (by the defn of } \mathbf{A})$$

$$= \mathbf{A}\text{Var}(\boldsymbol{\varepsilon} | \mathbf{X})\mathbf{A}' \text{ (b/c } \mathbf{A} \text{ is not random, given } \mathbf{X})$$

$$= \mathbf{A}E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' | \mathbf{X})\mathbf{A}' \text{ (because } E(\boldsymbol{\varepsilon}) = 0 \text{: strict exogeneity)}$$

$$= \mathbf{A}\sigma^2\mathbf{I}_n\mathbf{A}' \text{ (by independence and homoskedasticity)} = \sigma^2\mathbf{A}\mathbf{A}' .$$

Substituting the definition of $\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'$ and simplifying completes the proof. (note for future reference that this implies $\mathbf{A}\mathbf{A}' = (\mathbf{X}'\mathbf{X})^{-1}$) (*)

3. Gauss-Markov Theorem: Under assumptions 1-4, the OLS estimator is efficient in the class of linear unbiased estimators, i.e. it is the BLUE. Formally, for any unbiased estimator, $\hat{\boldsymbol{\beta}}$, that is linear in \mathbf{y} , $Var(\hat{\boldsymbol{\beta}} | \mathbf{X}) \geq Var(\mathbf{b} | \mathbf{X})$.

What does it mean?

a) “linearity” in the above expression means “linear in \mathbf{y} ”. The OLS estimator is linear in \mathbf{y} because it can be computed as

$$\underbrace{\mathbf{b}}_{K \times 1} = \underbrace{(\mathbf{X}'\mathbf{X})^{-1}}_{K \times K} \underbrace{\mathbf{X}'}_{K \times n} \underbrace{\mathbf{y}}_{n \times 1} = \mathbf{A}\mathbf{y}.$$

b) recall that the variance of the estimator refers to how the computed estimator (\mathbf{b} or $\hat{\boldsymbol{\beta}}$) varies across independent draws of the true error vector, $\boldsymbol{\varepsilon}$.

c) note that the inequality defining “best” is a matrix inequality, comparing two $K \times K$ matrices. This means that the difference between these matrices, $Var(\hat{\boldsymbol{\beta}} | \mathbf{X}) - Var(\mathbf{b} | \mathbf{X})$, itself a $K \times K$ matrix, is positive semidefinite. Among other things, this means that the diagonals of this matrix, $Var(\hat{\beta}_k | \mathbf{X}) - Var(b_k | \mathbf{X})$, are all positive.

Proof.

Write the alternative estimator, $\hat{\beta}$, as $\hat{\beta} = \mathbf{C}\mathbf{y}$, where \mathbf{C} can depend on \mathbf{X} (but not \mathbf{y}). (*This establishes the constraint that $\hat{\beta}$ must be linear*). Further, define $\mathbf{C} = \mathbf{D} + \mathbf{A}$, so that \mathbf{D} is in some sense the “divergence” between the matrix we use to calculate $\hat{\beta}$ and the matrix we use to calculate the OLS estimator, \mathbf{b} . (Recall that $\mathbf{A} \equiv (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, and $\mathbf{b} = \mathbf{A}\mathbf{y}$). So,

$$\hat{\beta} = (\mathbf{D} + \mathbf{A})\mathbf{y} = \mathbf{D}\mathbf{y} + \mathbf{A}\mathbf{y}, \text{ so}$$

$$\hat{\beta} = \mathbf{D}(\mathbf{X}\beta + \varepsilon) + \mathbf{b} = \mathbf{D}\mathbf{X}\beta + \mathbf{D}\varepsilon + \mathbf{b} \tag{1}$$

Now, to incorporate the constraint that $\hat{\beta}$ be unbiased, consider its expectation:

$$\begin{aligned} E(\hat{\beta} | \mathbf{X}) &= \mathbf{D}\mathbf{X}\beta + E(\mathbf{D}\varepsilon | \mathbf{X}) + E(\mathbf{b} | \mathbf{X}) \\ &= \mathbf{D}\mathbf{X}\beta + \beta \quad (\text{because } \mathbf{D}E(\varepsilon | \mathbf{X}) = \mathbf{0} \text{ (by strict exogeneity), and } E(\mathbf{b} | \mathbf{X}) = \beta \\ &\quad \text{(by unbiasedness of OLS, which, we have recently proved).} \end{aligned}$$

Unbiasedness of $\hat{\beta}$ therefore **requires $\mathbf{D}\mathbf{X}\beta = \mathbf{0}$** . Indeed, since we need $\hat{\beta}$ to be unbiased regardless of the true β , *unbiasedness of $\hat{\beta}$ thus requires $\mathbf{D}\mathbf{X} = \mathbf{0}$* .

Incorporating the requirement for unbiasedness ($\mathbf{DX}=\mathbf{0}$) into (1) now yields:

$\hat{\boldsymbol{\beta}} = \mathbf{D}\boldsymbol{\varepsilon} + \mathbf{b}$. Subtracting $\boldsymbol{\beta}$ from both sides yields an expression for the sampling error of this alternative, linear unbiased estimator $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$, in terms of the OLS sampling error ($\mathbf{b} - \boldsymbol{\beta}$):

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \mathbf{D}\boldsymbol{\varepsilon} + (\mathbf{b} - \boldsymbol{\beta}).$$

Using our expression for the OLS sampling error, $\mathbf{b} - \boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\boldsymbol{\varepsilon} = \mathbf{A}\boldsymbol{\varepsilon}$, yields:

$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \mathbf{D}\boldsymbol{\varepsilon} + \mathbf{A}\boldsymbol{\varepsilon} = (\mathbf{D} + \mathbf{A})\boldsymbol{\varepsilon}$. So the sampling error of *any* linear, unbiased estimator is related to the OLS sampling error, $\mathbf{A}\boldsymbol{\varepsilon}$, in this simple way.

Our next step is to calculate the variance of $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ above so we can compare it to the variance of the OLS estimator.

$$\begin{aligned}
\text{Var}(\hat{\boldsymbol{\beta}} | \mathbf{X}) &= \text{Var}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} | \mathbf{X}) = \text{Var}[(\mathbf{D} + \mathbf{A})\boldsymbol{\varepsilon} | \mathbf{X}] \\
&= \underbrace{(\mathbf{D} + \mathbf{A})}_{K \times n} \underbrace{\text{Var}(\boldsymbol{\varepsilon})}_{n \times n} \underbrace{(\mathbf{D} + \mathbf{A})'}_{n \times K} \quad (\text{since } \mathbf{D} \text{ and } \mathbf{A} \text{ both depend only on } \mathbf{X}) \\
&= \sigma^2 (\mathbf{D} + \mathbf{A})(\mathbf{D} + \mathbf{A})' \quad (\text{since } \text{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}_n) \\
&= \sigma^2 (\mathbf{D}\mathbf{D}' + \mathbf{A}\mathbf{D}' + \mathbf{D}\mathbf{A}' + \mathbf{A}\mathbf{A}')
\end{aligned}$$

Now, look at these expressions, term by term:

$$\mathbf{D}\mathbf{A}' = \mathbf{D}[(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}']' = \mathbf{D}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \mathbf{0}, \text{ since } \mathbf{D}\mathbf{X} = \mathbf{0} \text{ (by unbiasedness of } \hat{\boldsymbol{\beta}})$$

(note the above uses symmetry of $\mathbf{X}'\mathbf{X}$)

By the same token, $\mathbf{A}\mathbf{D}' = \mathbf{0}$,

and from our proof of Property 2 (Variance of \mathbf{b} : eqn (*)), $\mathbf{A}\mathbf{A}' = (\mathbf{X}'\mathbf{X})^{-1}$.

So, $\text{Var}(\hat{\boldsymbol{\beta}} | \mathbf{X}) = \sigma^2 (\mathbf{D}\mathbf{D}' + (\mathbf{X}'\mathbf{X})^{-1}) \geq \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} = \text{Var}(\mathbf{b} | \mathbf{X})$, where the inequality follows from positive semidefiniteness of $\mathbf{D}\mathbf{D}'$. QED.

Comments:

1. How surprising is this, really? The 'loss function' minimized by OLS is a sum of squared error terms. The variance of \mathbf{b} is also a sum of squares. So it's not totally surprising that minimizing a sum of squares (rather than absolute deviations, or some other function of the errors) minimizes the variance of \mathbf{b} .
2. What are the alternative, unbiased linear estimators that OLS dominates? It's hard to think of promising examples. (WLS?)

4. Orthogonality of \mathbf{b}, \mathbf{e} : Under Assumptions 1-4, $\text{Cov}(\mathbf{b}, \mathbf{e} \mid \mathbf{X}) = 0$, where \mathbf{e} is the OLS residual.

What does this mean? First, fix \mathbf{X} . Now consider a bunch of random draws of the true $(n \times 1)$ vector of disturbances, $\boldsymbol{\varepsilon}$. For each such draw, the OLS estimator produces a $(K \times 1)$ vector of estimated coefficients, \mathbf{b} . It also produces an $(n \times 1)$ vector of estimated residuals, \mathbf{e} . Now consider the $(n \times K)$ covariance matrix between these two vectors across realizations of $\boldsymbol{\varepsilon}$ (an element of this matrix is the Cov between b_k and e_i across draws of $\boldsymbol{\varepsilon}$). **All** the elements of this matrix are zero. i.e. the *constructed* \mathbf{b} 's and \mathbf{e} 's are uncorrelated with each other.

Proof—see Hayashi.

5. Unbiasedness of s^2 : Under assumptions 1-4, $s^2 \equiv \frac{SSR}{n-K} = \frac{\mathbf{e}'\mathbf{e}}{n-K}$ is also unbiased, i.e. $E(s^2|\mathbf{X}) = \sigma^2$.

Proof:

First, note that $s^2 = \frac{\mathbf{e}'\mathbf{e}}{n-K}$ is unbiased iff $E\left(\frac{\mathbf{e}'\mathbf{e}}{n-K} \mid \mathbf{X}\right) = \sigma^2$, or:

$E(\mathbf{e}'\mathbf{e} \mid \mathbf{X}) = \sigma^2(n-K)$, or: $E(\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon} \mid \mathbf{X}) = \sigma^2(n-K)$, where

$\mathbf{M} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the $n \times n$ annihilator matrix that calculates \mathbf{e} from $\boldsymbol{\varepsilon}$.

We now show, in turn, that $E(\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon} \mid \mathbf{X}) = \sigma^2 \cdot \text{trace}(\mathbf{M})$, and that $\text{trace}(\mathbf{M}) = n - K$.

Part 1: $E(\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon} \mid \mathbf{X}) = \sum_{i=1}^n \sum_{j=1}^n m_{ij} E(\varepsilon_i \varepsilon_j \mid \mathbf{X})$ (this just writes out the quadratic form, and uses the fact that \mathbf{M} depends only on \mathbf{X})

$$= \sum_{i=1}^n m_{ij} \sigma^2 \quad (\text{since all the cross terms equal zero, by no autocorrelation})$$

$$= \sigma^2 \sum_{i=1}^n m_{ij} = \sigma^2 \text{trace}(\mathbf{M}).$$

Part 2: $\text{trace}(\mathbf{M}) = \text{trace}[\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] = \text{trace}[\mathbf{I}_n] - \text{trace}[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']$

$$= n - \text{trace}[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']$$

And,

$$\text{trace}[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] = \text{trace}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}] \quad (\text{because } \text{trace}(\mathbf{AB}) = \text{trace}(\mathbf{BA}))$$

$$= \text{trace}(\mathbf{I}_K) = K.$$

QED.

Corollary: the variance of the estimated residuals, $\frac{SSR}{n} = \frac{\mathbf{e}'\mathbf{e}}{n}$, i.e. the *sample analog* of σ^2 , is a biased estimator of σ^2 . It is consistent, however (since $n/(n-K)$ approaches one in large samples).

Intuition for the $n-K$ “degrees of freedom correction”: By assumption, the n errors (ε 's) are each independently drawn from the same distribution, and generate n independent y 's. From these y 's the OLS procedure then calculates n residuals, $e_i = y_i - \mathbf{x}_i \mathbf{b}$. But, by construction, these n generated residuals have to satisfy K equality constraints, given by the OLS normal equations. So, in some sense, they *can't* vary as much as the ε 's. We adjust for this by inflating the variance of the e 's by the factor $n/(n-K)$, which exceeds one.

Additional Intuition: Consider the extreme case where the number of observations = the number of variables ($n=K$). In this case OLS yields a perfect fit. That does not, of course, mean that the true σ^2 is zero.

Final note: given property 5 plus property 2 --the formula for $Var(\mathbf{b}|\mathbf{X})$ -- a natural estimator of $Var(\mathbf{b}|\mathbf{X})$ is

$$\mathit{Varhat}(\mathbf{b} | \mathbf{X}) = s^2 (\mathbf{X}'\mathbf{X})^{-1} .$$

We'll talk about its properties in the next lesson, on hypothesis testing.