Lecture 2: Properties of the OLS estimator

Economics 527: Econometric Methods

Vadim Marmer, UBC

The model and the assumptions

Where the first lecture ended

- The first lecture produced the estimator and stopped there:

$$\widehat{\beta} = (X^\top X)^{-1} X^\top Y = \left(\sum_{i=1}^n X_i X_i^\top \right)^{-1} \sum_{i=1}^n X_i Y_i,$$

  the unique minimiser of $\sum_{i=1}^n (Y_i - X_i^\top b)^2$.
- Nothing so far says whether it is a good estimator.
- $\widehat{\beta}$ is a function of the sample, so it is random, and in general it will not equal β.
- Three questions, taken in this order: whether $\widehat{\beta}$ is right on average, how far it moves from sample to sample, and whether a better estimator of the same kind exists.

The model in matrix form

- Stack the n observations, as in the first lecture:

$$\underset{n\times 1}{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \qquad \underset{n\times k}{X} = \begin{pmatrix} X_1^\top \\ \vdots \\ X_n^\top \end{pmatrix}, \qquad \underset{n\times 1}{U} = \begin{pmatrix} U_1 \\ \vdots \\ U_n \end{pmatrix}.$$

- The n scalar equations $Y_i = X_i^\top \beta + U_i$ become one:

$$\underset{n\times 1}{Y} = \underset{n\times k}{X}\ \underset{k\times 1}{\beta} + \underset{n\times 1}{U}.$$

- β is fixed and unknown. X, Y and U are random, and U is not observed.

The four assumptions

- **Assumption 1.** $Y = X\beta + U$ for some $\beta \in \mathbb{R}^k$.
- **Assumption 2.** $\mathrm{E}[U \mid X] = 0$.
- **Assumption 3.** $\mathrm{Var}(U \mid X) = \sigma^2 I_n$ for some $\sigma^2 > 0$.
- **Assumption 4.** $\mathrm{rank}(X) = k$, so $n \geq k$ and no regressor is an exact linear combination of the others.
- The numbering is the typed notes'. Assumption 4 is the sample rank condition, the sample counterpart of $\mathrm{rank}\left(\mathrm{E}\left[X_i X_i^\top\right]\right) = k$ from the first lecture.
- Assumption 3 is new, and nothing before the variance of $\widehat{\beta}$ uses it.

What unbiased means

- $\hat{\beta}$ is a function of the sample, hence a random vector:

$$\hat{\beta}(\omega) = \left(\sum_{i=1}^{n} X_i(\omega) X_i(\omega)^\top \right)^{-1} \sum_{i=1}^{n} X_i(\omega) Y_i(\omega).$$

- With a continuously distributed error, $\mathrm{P}(\hat{\beta} = \beta) = 0$. Hitting β exactly is not the standard.
- **Definition.** An estimator $\hat{\theta}$ of θ is **unbiased** if $\mathrm{E}\left[\hat{\theta}\right] = \theta$, and **conditionally unbiased given** X if

$$\mathrm{E}\left[\hat{\theta} \mid X\right] = \theta,$$

  whatever the true value of θ: the estimator is a rule that does not know θ, and the equality must hold for every value θ could take.
- The conditional statement is the stronger one, and it is the one this lecture proves.

Unbiasedness

The sampling error

- Substitute Assumption 1 into the estimator; Assumption 4 is what makes $(X^\top X)^{-1}$ exist:

$$\begin{aligned} \hat{\beta} &= (X^\top X)^{-1} X^\top Y \\ &= (X^\top X)^{-1} X^\top (X\beta + U) \\ &= \underbrace{(X^\top X)^{-1} X^\top X}_{= I_k} \beta + (X^\top X)^{-1} X^\top U. \end{aligned}$$

- The first term collapses, and what is left is the **sampling error**:

$$\boxed{\hat{\beta} = \underbrace{\beta}_{\text{signal}} + \underbrace{(X^\top X)^{-1} X^\top U}_{\text{noise}}}$$

- Everything in this section and the next is a statement about the noise term.

Taking expectations, first attempt

- Take expectations of the sampling error:

$$\mathrm{E}\left[\hat{\beta}\right] = \beta + \mathrm{E}\left[(X^\top X)^{-1} X^\top U\right].$$

- Suppose only that $\mathrm{E}\left[X_i U_i\right] = 0$ for every i. Then

$$\mathrm{E}\left[X^\top U\right] = \mathrm{E}\left[\sum_{i=1}^{n} X_i U_i\right] = \sum_{i=1}^{n} \mathrm{E}\left[X_i U_i\right] = 0.$$

- That is not enough, because in general

$$\mathrm{E}\left[(X^\top X)^{-1} X^\top U\right] \quad \text{need not equal} \quad \mathrm{E}\left[(X^\top X)^{-1}\right] \mathrm{E}\left[X^\top U\right].$$

- The two factors are built from the same X, and once $\mathrm{E}\left[U \mid X\right]$ is not zero, the expectation of their product need not factor.

The expectation of a ratio

- The one-regressor case makes the failure visible:

$$\mathrm{E}\left[\frac{\sum_i X_i U_i}{\sum_i X_i^2}\right] \quad \text{need not equal} \quad \frac{\mathrm{E}\left[\sum_i X_i U_i\right]}{\mathrm{E}\left[\sum_i X_i^2\right]} = 0.$$

- In general the expectation of a ratio is not the ratio of expectations. For any two scalar random variables Z and S,

$$\mathrm{E}\left[\frac{Z}{S}\right] = \iint \frac{z}{s}\, f_{Z,S}(z,s)\, dz\, ds \quad \text{need not equal} \quad \frac{\int z f_Z(z)\, dz}{\int s f_S(s)\, ds} = \frac{\mathrm{E}[Z]}{\mathrm{E}[S]}.$$

- The way out is to condition on X, which turns $(X^\top X)^{-1} X^\top$ into a known matrix.

Conditional unbiasedness: the claim

- **Claim.** Under Assumptions 1, 2 and 4,

$$\mathrm{E}\left[\hat{\beta} \mid X\right] = \beta.$$

- Assumption 3 is not used. Unbiasedness does not depend on the variance of the errors.

Proof of conditional unbiasedness

- Condition the sampling error on X:

$$\begin{aligned}
\mathrm{E}\left[\hat{\beta} \mid X\right] &= \mathrm{E}\left[\beta + (X^\top X)^{-1} X^\top U \mid X\right] \\
&= \beta + \mathrm{E}\left[(X^\top X)^{-1} X^\top U \mid X\right] \\
&= \beta + (X^\top X)^{-1} X^\top \underbrace{\mathrm{E}[U \mid X]}_{=0 \text{ by Assumption 2}} \\
&= \beta.
\end{aligned}$$

- The third line takes out what is known, as in the first lecture. Given X, the matrix $(X^\top X)^{-1} X^\top$ is known and comes out of the expectation.
- This is exactly the step the unconditional version could not take.

From conditional to unconditional

- The law of iterated expectations, from the first lecture, finishes the proof whenever $\mathrm{E}\left[\hat{\beta}\right]$ exists:

$$\mathrm{E}\left[\hat{\beta}\right] = \mathrm{E}\left[\mathrm{E}\left[\hat{\beta} \mid X\right]\right] = \mathrm{E}[\beta] = \beta.$$

- Conditional unbiasedness implies unbiasedness whenever the unconditional mean exists. The converse does not hold.
- Existence is not automatic: with $n = k = 1$, a regressor uniform on $(0, 1)$ and a mean-zero error independent of it, the noise term U/X has no mean, although its conditional mean is zero.

The scalar case

- Take $k = 1$, so X is $n \times 1$ and

$$X^\top X = \sum_{i=1}^n X_i^2, \qquad X^\top Y = \sum_{i=1}^n X_i Y_i.$$

- The estimator and its sampling error:

$$\hat{\beta} = \frac{\sum_i X_i Y_i}{\sum_i X_i^2} = \beta + \frac{\sum_i X_i U_i}{\sum_i X_i^2}.$$

- Conditioning on $X_1, \ldots, X_n$ makes the denominator and every X_i in the numerator known:

$$\mathrm{E}\left[\hat{\beta} \mid X\right] = \beta + \frac{\sum_i X_i \overbrace{\mathrm{E}\left[U_i \mid X\right]}^{=0}}{\sum_i X_i^2} = \beta.$$

A weaker condition

- Replace Assumption 2 by

$$\mathrm{E}\left[X_i U_i\right] = 0 \quad \text{for all } i,$$

  which, together with $\mathrm{E}\left[U_i\right] = 0$, says that each regressor is uncorrelated with the error.
- The first lecture showed that β is still identified under this condition.
- It is not enough for unbiasedness:

$$\mathrm{E}\left[\hat{\beta}\right] = \beta + \underbrace{\mathrm{E}\left[(X^\top X)^{-1} X^\top U\right]}_{\text{need not be } 0}.$$

- So $\hat{\beta}$ need not be unbiased. The noise term is a nonlinear function of X and U, and once $\mathrm{E}\left[U \mid X\right]$ is not zero, nothing forces its mean to vanish.

Zero correlation does not pin down the conditional mean

- Take one regressor, so X_i is a scalar with density f_X, and write $g(x) = \mathrm{E}\left[U_i \mid X_i = x\right]$. By iterated expectations,

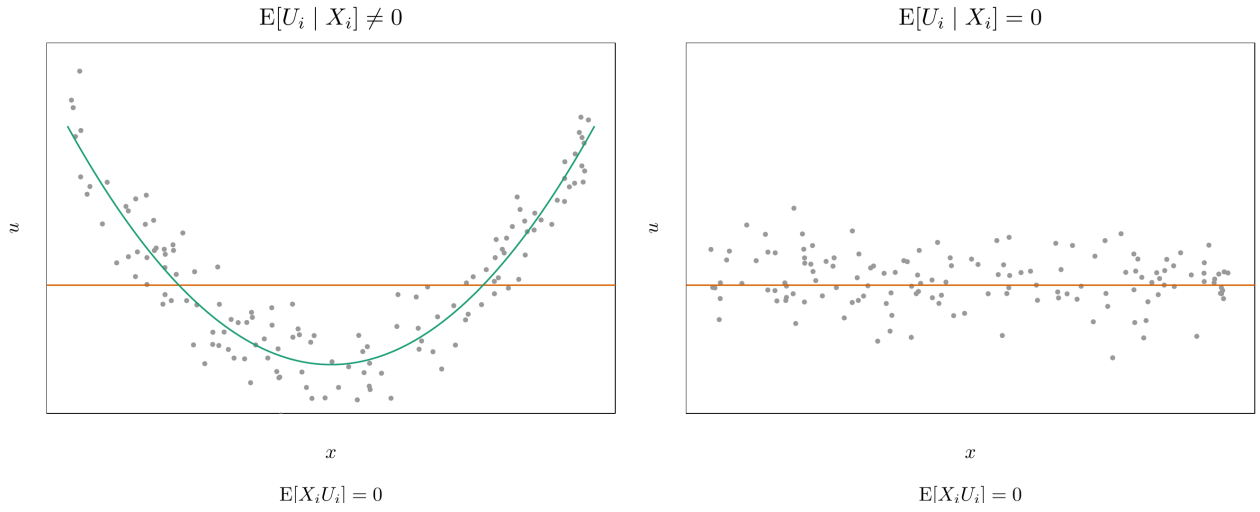
$$\mathrm{E}\left[X_i U_i\right] = \mathrm{E}\left[\mathrm{E}\left[X_i U_i \mid X_i\right]\right] = \mathrm{E}\left[X_i\, g(X_i)\right] = \int x\, g(x) f_X(x)\, dx.$$

- Assumption 2 says $g(x) = 0$ at every x, which makes the integral zero. The weaker condition says only that the integral is zero.
- An integral vanishes when the positive and negative parts cancel, so

$$\int x\, g(x) f_X(x)\, dx = 0 \;\text{ does not imply }\; g(x) = 0 \;\text{ for all } x.$$

The parabola

- Let the conditional mean of the error be a parabola in the regressor, centred at zero and shifted down so that $\mathrm{E}\left[U_i\right] = \mathrm{E}\left[g(X_i)\right] = 0$, and let the regressor be symmetric about zero. Then g is even, so $x\, g(x)$ is odd and $\mathrm{E}\left[X_i U_i\right] = 0$, while $\mathrm{E}\left[U_i \mid X_i\right] \neq 0$ at every value of X_i except the two zeros of g.
- The parabola shows that the conditional mean is not zero. It does not by itself bias $\hat{\beta}$: with g even, the conditional mean of the noise term, $\sum_i X_i g(X_i) / \sum_i X_i^2$, changes sign when every X_i does. The regressors are symmetric about zero, so the noise term has mean zero. Bias needs one of those two symmetries broken, as on the next slide but one.
- Read the horizontal axis as education and the vertical axis as ability: ability is low at middling education and high at both ends.



What survives under the weaker condition

- Under $\mathrm{E}[X_i U_i] = 0$ and $\mathrm{E}[U_i] = 0$ in place of Assumption 2: β is identified, and $\hat{\beta}$ is still the method-of-moments and least-squares estimator.
- What is lost: $X_i^\top \beta$ need not be the conditional mean any more, and $\hat{\beta}$ is no longer guaranteed to be unbiased at any sample size.

The variance of a random vector

The definition

- For a random vector V that is $n \times 1$ with finite second moments, and with $\mathrm{E}[V] = (\mathrm{E}[V_1], \ldots, \mathrm{E}[V_n])^\top$,

$$\underset{(n\times n)}{\mathrm{Var}(V)} = \mathrm{E}\left[\underset{n\times 1}{(V - \mathrm{E}[V])}\ \underset{1\times n}{(V - \mathrm{E}[V])^\top}\right]$$

- The product inside is an outer product, not an inner product, so the result is a matrix rather than a number.

The matrix entry by entry

- Entry (i, j) is $\mathrm{E}\left[(V_i - \mathrm{E}[V_i])(V_j - \mathrm{E}\left[V_j\right])\right]$, so

$$\mathrm{Var}(V) = \begin{pmatrix} \mathrm{Var}(V_1) & \mathrm{Cov}(V_1, V_2) & \cdots & \mathrm{Cov}(V_1, V_n) \\ \mathrm{Cov}(V_2, V_1) & \mathrm{Var}(V_2) & & \vdots \\ \vdots & & \ddots & \\ \mathrm{Cov}(V_n, V_1) & \cdots & & \mathrm{Var}(V_n) \end{pmatrix}.$$

- The variances sit on the diagonal and the covariances off it.

Symmetry

- $\mathrm{Var}(V)^\top = \mathrm{Var}(V)$, because $\mathrm{Cov}(V_i, V_j) = \mathrm{Cov}(V_j, V_i)$.
- The same statement from the definition, using that the transpose passes through the expectation entry by entry, and then $(AB)^\top = B^\top A^\top$:

$$\mathrm{Var}(V)^\top = \mathrm{E}\left[((V - \mathrm{E}[V])(V - \mathrm{E}[V])^\top)^\top\right] = \mathrm{E}\left[(V - \mathrm{E}[V])(V - \mathrm{E}[V])^\top\right] = \mathrm{Var}(V).$$

Positive semidefiniteness

- **Claim.** $\mathrm{Var}\,(V)$ is positive semidefinite.
- **Proof.** Take any non-random $a \in \mathbb{R}^n$ and set $W = a^\top (V - \mathrm{E}\,[V])$, a scalar. Then

$$\begin{aligned} a^\top \mathrm{Var}\,(V)\, a &= a^\top \mathrm{E}\,[(V - \mathrm{E}\,[V])(V - \mathrm{E}\,[V])^\top]\, a \\ &= \mathrm{E}\left[\underbrace{a^\top (V - \mathrm{E}\,[V])}_{W}\ \underbrace{(V - \mathrm{E}\,[V])^\top a}_{W} \right] \\ &= \mathrm{E}\,[W^2] \ \geq\ 0. \end{aligned}$$

- The constant a moves inside the expectation because expectation is linear.
- This is the argument the first lecture used for $\mathrm{E}\,[X_i X_i^\top]$, with $W_i = X_i^\top c$.

When it is positive definite

- For $a \neq 0$, with $W = a^\top (V - \mathrm{E}\,[V])$ as above, track the equality case:

$$a^\top \mathrm{Var}\,(V)\, a = 0 \iff \mathrm{E}\,[W^2] = 0 \iff a^\top (V - \mathrm{E}\,[V]) = 0,$$

  that is,

$$a_1 (V_1 - \mathrm{E}\,[V_1]) + \cdots + a_n (V_n - \mathrm{E}\,[V_n]) = 0.$$

- So $\mathrm{Var}\,(V)$ is positive definite exactly when no non-zero linear combination of $V_1, \ldots, V_n$ is constant, that is, when no exact linear relation holds among the centred variables $V_i - \mathrm{E}\,[V_i]$.

Variance under a linear map

- **Claim.** For a non-random $m \times n$ matrix Γ,

$$\boxed{\mathrm{Var}\,(\Gamma V) = \Gamma \mathrm{Var}\,(V)\, \Gamma^\top} \qquad (m \times m).$$

- **Proof.** $\mathrm{E}\,[\Gamma V] = \Gamma \mathrm{E}\,[V]$, so

$$\begin{aligned} \mathrm{Var}\,(\Gamma V) &= \mathrm{E}\left[(\Gamma V - \Gamma \mathrm{E}\,[V]\,)(\Gamma V - \Gamma \mathrm{E}\,[V]\,)^\top \right] \\ &= \mathrm{E}\,[\Gamma (V - \mathrm{E}\,[V])(V - \mathrm{E}\,[V])^\top \Gamma^\top] \\ &= \Gamma \mathrm{E}\,[(V - \mathrm{E}\,[V])(V - \mathrm{E}\,[V])^\top]\, \Gamma^\top = \Gamma \mathrm{Var}\,(V)\, \Gamma^\top, \end{aligned}$$

  using $(AB)^\top = B^\top A^\top$ on the second factor, and linearity to move the non-random Γ and $\Gamma^\top$ outside the expectation.
- Check against the scalar case: with $m = n = 1$ this is $\mathrm{Var}\,(\gamma V) = \gamma^2 \mathrm{Var}\,(V)$.

Adding a constant

- For a non-random α that is $m \times 1$,

$$\mathrm{Var}\,(\alpha + \Gamma V) = \mathrm{Var}\,(\Gamma V) = \Gamma \mathrm{Var}\,(V)\, \Gamma^\top.$$

- The constant shifts the mean, so the deviations from the mean are unchanged and it drops out.
- This is why β will drop out of the variance of the sampling error.

The conditional version

- The same definition with conditional expectations throughout:

$$\operatorname{Var}(V \mid X) = \mathrm{E}\left[(V - \mathrm{E}[V \mid X])(V - \mathrm{E}[V \mid X])^{\top} \,\middle|\, X\right].$$

- Everything above holds conditionally, with Γ allowed to be a function of X. Given X, such a Γ is known and comes out of the conditional expectation, by taking out what is known, as in the proof of unbiasedness.
- When the conditional mean is zero, the definition simplifies. Under Assumption 2,

$$\operatorname{Var}(U \mid X) = \mathrm{E}\left[UU^{\top} \mid X\right].$$

The variance of the OLS estimator

The variance matrix of $\hat{\beta}$

- $\hat{\beta}$ is $k \times 1$, so $\operatorname{Var}\left(\hat{\beta} \mid X\right)$ is $k \times k$: the conditional variance of each estimated coefficient on the diagonal, the conditional covariances between them off it.
- Conditioning on X again, for the same reason it worked for unbiasedness: it makes $(X^{\top}X)^{-1}X^{\top}$ a known matrix.

The sandwich formula

- Start from the sampling error and name the matrix that multiplies U:

$$\hat{\beta} = \beta + \underbrace{(X^{\top}X)^{-1}X^{\top}}_{\Gamma_X} U.$$

- β is a constant and drops out, and Γ_X is known given X, so the rule for a linear map applies:

$$\begin{aligned}\operatorname{Var}\left(\hat{\beta} \mid X\right) &= \operatorname{Var}(\beta + \Gamma_X U \mid X) \\ &= \operatorname{Var}(\Gamma_X U \mid X) \\ &= \Gamma_X \operatorname{Var}(U \mid X)\, \Gamma_X^{\top}.\end{aligned}$$

The sandwich formula written out

- Substituting $\Gamma_X = (X^{\top}X)^{-1}X^{\top}$, with $\Gamma_X^{\top} = X(X^{\top}X)^{-1}$ because $X^{\top}X$ and its inverse are symmetric:

$$\boxed{\operatorname{Var}\left(\hat{\beta} \mid X\right) = (X^{\top}X)^{-1}X^{\top}\ \operatorname{Var}(U \mid X)\ \ X(X^{\top}X)^{-1}}$$

- The box uses Assumptions 1 and 4, and needs the errors to have finite conditional second moments, so that $\operatorname{Var}(U \mid X)$ exists. Assumption 2 enters only to write the middle factor as $\mathrm{E}\left[UU^{\top} \mid X\right]$.
- Apart from requiring $\operatorname{Var}(U \mid X)$ to exist, nothing yet has been assumed about the errors' conditional variances or covariances.

Assumption 3: homoskedasticity and no autocorrelation

- **Assumption 3.** $\operatorname{Var}(U \mid X) = \sigma^2 I_n$ for some $\sigma^2 > 0$, that is,

$$\operatorname{Var}(U \mid X) = \begin{pmatrix} \sigma^2 & & 0 \\ & \ddots & \\ 0 & & \sigma^2 \end{pmatrix}.$$

- Read the diagonal: every error has the same conditional variance,

$$\operatorname{Var}(U_1 \mid X) = \cdots = \operatorname{Var}(U_n \mid X) = \sigma^2.$$

- This is **homoskedasticity**. It rules out errors that are more variable at some values of the regressors than at others.
- Read the off-diagonal entries: for $i \neq j$, with the first equality by Assumption 2,

$$\operatorname{Cov}\left(U_i, U_j \mid X\right) = \operatorname{E}\left[U_i U_j \mid X\right] = 0.$$

- The errors are conditionally uncorrelated across observations. This is **no autocorrelation**, also called no serial correlation.
- Assumption 3 is these two statements together, and nothing more. It does not say the errors are independent, and it does not say they are normal.

The variance of OLS under Assumption 3

- Put $\operatorname{Var}\left(U \mid X\right) = \sigma^2 I_n$ into the sandwich:

$$\begin{aligned}\operatorname{Var}\left(\hat{\beta} \mid X\right) &= (X^\top X)^{-1} X^\top \, \sigma^2 I_n \, X (X^\top X)^{-1} \\ &= \sigma^2 (X^\top X)^{-1} \underbrace{X^\top X (X^\top X)^{-1}}_{= I_k} \\ &= \sigma^2 (X^\top X)^{-1}.\end{aligned}$$

- Under Assumptions 1 to 4:

$$\boxed{\operatorname{E}\left[\hat{\beta} \mid X\right] = \beta, \qquad \operatorname{Var}\left(\hat{\beta} \mid X\right) = \sigma^2 (X^\top X)^{-1}}$$

- This pair is the central result of the lecture.

What σ^2 is

- From Assumptions 2 and 3 together,

$$\sigma^2 = \operatorname{Var}\left(U_i \mid X\right) = \operatorname{E}\left[U_i^2 \mid X\right],$$

  and the law of iterated expectations gives

$$\sigma^2 = \operatorname{E}\left[U_i^2\right] = \operatorname{Var}\left(U_i\right),$$

  the last step because $\operatorname{E}\left[U_i\right] = \operatorname{E}\left[\operatorname{E}\left[U_i \mid X\right]\right] = 0$ by Assumption 2.
- σ^2 is one unknown number, the same for every observation. Estimating it is a question for the next lecture.

One regressor

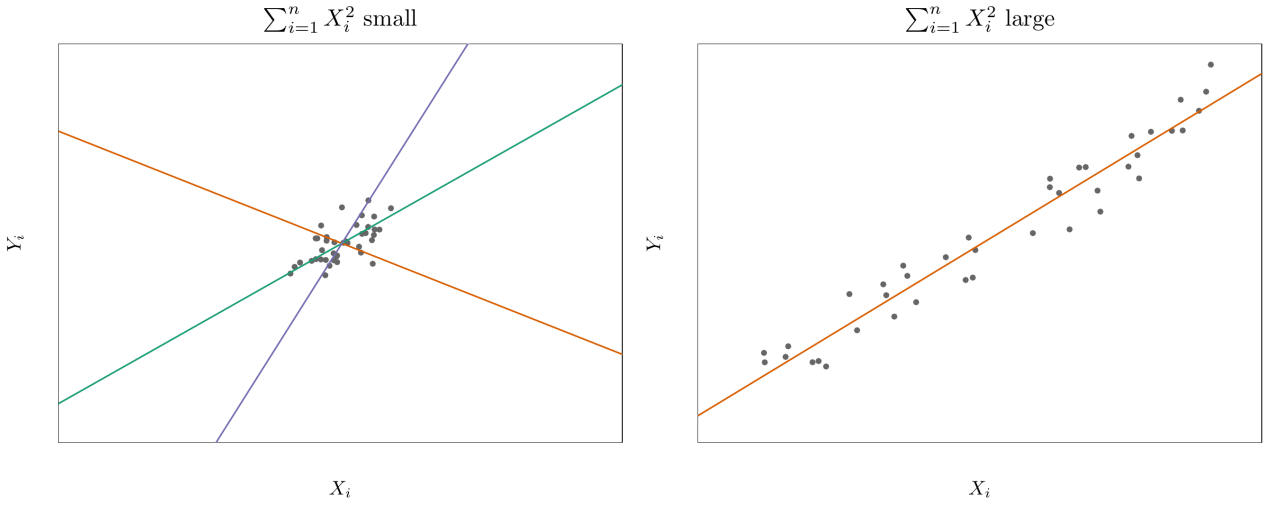
- Take $k = 1$ and no intercept, so $Y_i = \beta X_i + U_i$ and X is $n \times 1$. Then $X^\top X = \sum_i X_i^2$ and

$$\operatorname{Var}\left(\hat{\beta} \mid X\right) = \frac{\sigma^2}{\sum_{i=1}^n X_i^2}.$$

- Two things move the variance. Noisier errors raise it. A larger $\sum_i X_i^2$ lowers it: with no intercept this is the sum of squares of the regressor about zero. With an intercept the same formula gives the variance of the slope, with $\sum_i (X_i - \bar{X})^2$, the spread of the regressor about its mean, in place of $\sum_i X_i^2$.

Why spread in the regressor helps

- With the regressor bunched together, many quite different slopes fit the data about equally well, so the estimate moves a lot from sample to sample.
- With the regressor spread out, the slope is pinned down.



Without homoskedasticity

- If the errors are heteroskedastic but still conditionally uncorrelated across observations, so that $\mathrm{Var}\,(U \mid X)$ is diagonal with entries that may depend on X,

$$\mathrm{Var}\,(U \mid X) = \Omega = \begin{pmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_n^2 \end{pmatrix},$$

the sandwich does not collapse. The middle factor becomes

$$X^\top \Omega X = \begin{pmatrix} X_1 & \cdots & X_n \end{pmatrix} \Omega \begin{pmatrix} X_1^\top \\ \vdots \\ X_n^\top \end{pmatrix} = \sum_{i=1}^n \sigma_i^2 X_i X_i^\top,$$

so

$$\mathrm{Var}\left(\hat{\beta} \mid X\right) = (X^\top X)^{-1} \left(\sum_{i=1}^n \sigma_i^2 X_i X_i^\top \right) (X^\top X)^{-1}.$$

- Unbiasedness is untouched. Only the variance formula changes.

Where each assumption was used

Assumptions	What they deliver
1, 2, 4	$\mathrm{E}\left[\hat{\beta} \mid X\right] = \beta$, and $\mathrm{E}\left[\hat{\beta}\right] = \beta$ whenever that mean exists
1, 4, finite conditional second moments	$\mathrm{Var}\left(\hat{\beta} \mid X\right) = \Gamma_X \mathrm{Var}\,(U \mid X)\,\Gamma_X^\top$, with $\Gamma_X = (X^\top X)^{-1} X^\top$
1, 3, 4	$\mathrm{Var}\left(\hat{\beta} \mid X\right) = \sigma^2 (X^\top X)^{-1}$

The Gauss-Markov theorem

Linear estimators

- **Definition.** An estimator of β is **linear** if it can be written

  $$\tilde{\beta} = AY$$

  for a matrix A that is $k \times n$ and depends on X alone, not on Y.
- OLS is linear, with $A = (X^\top X)^{-1} X^\top$.
- Given X, each entry of a linear estimator is a fixed linear combination of the observations on the dependent variable.

Two more linear estimators

- **The two-point slope.** With $k = 1$, $Y_i = \beta X_i + U_i$ and $X_1 \neq X_n$,

  $$\tilde{\beta} = \frac{Y_n - Y_1}{X_n - X_1} = -\frac{1}{X_n - X_1} Y_1 + \frac{1}{X_n - X_1} Y_n.$$

  It uses two observations and throws the rest away.
- **Weighted least squares.** For a known $n \times n$ matrix W that is symmetric and positive definite, and that may depend on X but not on Y,

  $$\tilde{\beta} = (X^\top W X)^{-1} X^\top W Y.$$

  OLS is the case $W = I_n$. With W diagonal, this estimator weights observation i by the ith diagonal entry w_i, as in the first lecture's slide on other linear estimators. The inverse $(X^\top W X)^{-1}$ exists: for $c \neq 0$, $Xc \neq 0$ by Assumption 4, so $c^\top X^\top W X c = (Xc)^\top W (Xc) > 0$.

Both are unbiased

- The two-point slope, substituting $Y_i = \beta X_i + U_i$ and using Assumption 2:

  $$\tilde{\beta} = \beta + \frac{U_n - U_1}{X_n - X_1}, \qquad \mathrm{E}\left[\tilde{\beta} \mid X\right] = \beta.$$

- Weighted least squares, substituting $Y = X\beta + U$ and using Assumption 2:

  $$\tilde{\beta} = \beta + (X^\top W X)^{-1} X^\top W U, \qquad \mathrm{E}\left[\tilde{\beta} \mid X\right] = \beta.$$

- Being unbiased does not single OLS out. When $n > k$, there are many conditionally unbiased linear estimators.

The variance of the weighted estimator

- The route that gave the sandwich formula applies, with $A = (X^\top W X)^{-1} X^\top W$ in place of Γ_X. Since $W^\top = W$, the matrix $X^\top W X$ and its inverse are symmetric too, so $A^\top = W X (X^\top W X)^{-1}$. Under Assumptions 1, 3 and 4,

  $$\mathrm{Var}\left(\tilde{\beta} \mid X\right) = \sigma^2 (X^\top W X)^{-1} X^\top W W X (X^\top W X)^{-1},$$

  which reduces to $\sigma^2 (X^\top X)^{-1}$ when $W = I_n$.
- Which of these matrices is smallest, and what smallest means for matrices, are the next questions.

The comparison to settle

- For every linear conditionally unbiased $\tilde{\beta}$, compare the two conditional variance matrices:

  $$\underset{k \times k}{\mathrm{Var}\left(\tilde{\beta} \mid X\right)} - \underset{k \times k}{\mathrm{Var}\left(\hat{\beta} \mid X\right)} \overset{?}{\geq} 0.$$

- Here ≥ 0 means positive semidefinite.

The Gauss-Markov theorem

- **Theorem.** Under Assumptions 1 to 4, let $\tilde{\beta}$ satisfy
  1. **linearity**: $\tilde{\beta} = AY$, with A a $k \times n$ matrix depending on X alone;
  2. **conditional unbiasedness**: $\mathrm{E}\left[\tilde{\beta} \mid X\right] = \beta$ for every value of β.
- Then

$$\mathrm{Var}\left(\tilde{\beta} \mid X\right) - \mathrm{Var}\left(\hat{\beta} \mid X\right) \quad \text{is positive semidefinite.}$$

- OLS is the **best linear unbiased estimator** (BLUE): best in the sense above, linear in Y, conditionally unbiased, and an estimator of β.
- The theorem is about conditional unbiasedness, because it compares conditional variances.

Reading the matrix ordering

- If a $k \times k$ matrix M is positive semidefinite, then $c^\top M c \geq 0$ for **every** $c \in \mathbb{R}^k$.
- Take c to be the jth coordinate vector e_j, the vector with a 1 in position j and a 0 everywhere else:

$$e_j = (0, \ldots, 0, 1, 0, \ldots, 0)^\top, \qquad \text{the 1 sitting in position } j.$$

- Writing the quadratic form out entry by entry, and using $(e_j)_l = 0$ for every $l \neq j$ and $(e_j)_j = 1$,

$$e_j^\top M e_j = \sum_{l=1}^{k} \sum_{m=1}^{k} (e_j)_l \, M_{lm} \, (e_j)_m = M_{jj}.$$

- So every diagonal entry of a positive semidefinite matrix is non-negative.
- Apply this to $M = \mathrm{Var}\left(\tilde{\beta} \mid X\right) - \mathrm{Var}\left(\hat{\beta} \mid X\right)$, whose jth diagonal entry is the difference of the two variances of the jth coefficient:

$$\mathrm{Var}\left(\tilde{\beta}_j \mid X\right) \;\geq\; \mathrm{Var}\left(\hat{\beta}_j \mid X\right), \qquad j = 1, \ldots, k.$$

- A general c says the same about every linear combination $c^\top \beta$ of the coefficients, since $\mathrm{Var}\left(c^\top V \mid X\right) = c^\top \mathrm{Var}\left(V \mid X\right) c$:

$$\mathrm{Var}\left(c^\top \tilde{\beta} \mid X\right) \;\geq\; \mathrm{Var}\left(c^\top \hat{\beta} \mid X\right) \qquad \text{for every } c \in \mathbb{R}^k.$$

Proof, step 1: linearity and unbiasedness force $AX = I_k$

- Take conditional expectations of $\tilde{\beta} = AY$:

$$\begin{aligned} \beta = \mathrm{E}\left[\tilde{\beta} \mid X\right] &= \mathrm{E}\left[A(X\beta + U) \mid X\right] \\ &= AX\beta + A \underbrace{\mathrm{E}\left[U \mid X\right]}_{=0 \text{ by Assumption 2}} \\ &= AX\beta. \end{aligned}$$

- This must hold for every value of β, which is what unbiasedness means, so

$$\boxed{AX = I_k}$$

- Linearity plus conditional unbiasedness is exactly this one condition.

Step 1 checked on the examples

- OLS: $A = (X^\top X)^{-1} X^\top$, so $AX = (X^\top X)^{-1} X^\top X = I_k$.
- Weighted least squares: $A = (X^\top W X)^{-1} X^\top W$, so $AX = (X^\top W X)^{-1} X^\top W X = I_k$.
- The two-point slope, with $k = 1$: A has $-1/(X_n - X_1)$ in position 1 and $1/(X_n - X_1)$ in position n, so

$$AX = \frac{X_n - X_1}{X_n - X_1} = 1.$$

Proof, step 2: the two sampling errors

- Using $AX = I_k$ for the general estimator, and the sampling error for OLS,

$$\hat{\beta} = \beta + (X^\top X)^{-1} X^\top U, \qquad \tilde{\beta} = AX\beta + AU = \beta + AU.$$

- Both estimators differ from β only through a known matrix times U. Everything that follows is an expectation of $UU^\top$.

The transpose rule for covariance matrices

- For a $k \times 1$ random vector V and an $\ell \times 1$ random vector Z,

$$\operatorname{Cov}(V, Z) = \mathrm{E}\left[(V - \mathrm{E}[V])(Z - \mathrm{E}[Z])^\top\right] \qquad (k \times \ell),$$

and transposing gives

$$\operatorname{Cov}(V, Z)^\top = \mathrm{E}\left[(Z - \mathrm{E}[Z])(V - \mathrm{E}[V])^\top\right] = \operatorname{Cov}(Z, V).$$

- The same definition and the same rule hold conditionally, with $\mathrm{E}[\cdot]$ replaced by $\mathrm{E}[\cdot \mid X]$ throughout. Step 4 needs this, because both orders appear there.

Proof, step 3: the covariance equals the variance of OLS

- Both estimators are conditionally unbiased, so their deviations from their conditional means are the two noise terms of step 2:

$$\begin{aligned}
\operatorname{Cov}\left(\tilde{\beta}, \hat{\beta} \mid X\right) &= \mathrm{E}\left[(\tilde{\beta} - \beta)(\hat{\beta} - \beta)^\top \,\middle|\, X\right] \\
&= \mathrm{E}\left[AU\, U^\top X (X^\top X)^{-1} \,\middle|\, X\right] \\
&= A \underbrace{\mathrm{E}\left[UU^\top \mid X\right]}_{= \sigma^2 I_n \text{ by Assumptions 2 and 3}} X (X^\top X)^{-1} \\
&= \sigma^2 \underbrace{AX}_{= I_k} (X^\top X)^{-1} \\
&= \sigma^2 (X^\top X)^{-1}.
\end{aligned}$$

- The right-hand side is $\operatorname{Var}\left(\hat{\beta} \mid X\right)$, so

$$\boxed{\operatorname{Cov}\left(\tilde{\beta}, \hat{\beta} \mid X\right) = \operatorname{Var}\left(\hat{\beta} \mid X\right)}$$

- Every linear conditionally unbiased estimator has the same covariance with OLS, namely the variance of OLS itself. This is the one surprising step in the proof.

Proof, step 4: the difference of the variances

- As for a scalar, expand the variance of the difference using the definition; the cross terms are the two covariance matrices:

$$\operatorname{Var}\left(\tilde{\beta}-\hat{\beta} \mid X\right)=\operatorname{Var}\left(\tilde{\beta} \mid X\right)+\operatorname{Var}\left(\hat{\beta} \mid X\right) \\ -\operatorname{Cov}\left(\tilde{\beta}, \hat{\beta} \mid X\right)-\operatorname{Cov}\left(\hat{\beta}, \tilde{\beta} \mid X\right).$$

- Step 3 gives the first covariance term, the second is its transpose by the rule above, and $\operatorname{Var}\left(\hat{\beta} \mid X\right)$ is symmetric. All three terms in blue equal $\operatorname{Var}\left(\hat{\beta} \mid X\right)$, and they combine into one with a minus sign:

$$\operatorname{Var}\left(\tilde{\beta}-\hat{\beta} \mid X\right)=\operatorname{Var}\left(\tilde{\beta} \mid X\right)-\operatorname{Var}\left(\hat{\beta} \mid X\right).$$

- The left-hand side is a conditional variance matrix, hence positive semidefinite. Therefore

$$\operatorname{Var}\left(\tilde{\beta} \mid X\right)-\operatorname{Var}\left(\hat{\beta} \mid X\right) \geq 0,$$

  which is the theorem.

Uniqueness

- Suppose some linear conditionally unbiased $\tilde{\beta}$ attains the same variance matrix. By the identity of step 4,

$$\operatorname{Var}\left(\tilde{\beta} \mid X\right)=\operatorname{Var}\left(\hat{\beta} \mid X\right) \quad \Longrightarrow \quad \operatorname{Var}\left(\tilde{\beta}-\hat{\beta} \mid X\right)=0.$$

- A random vector with zero conditional variance equals its conditional mean (the equality case of positive semidefiniteness, one coordinate at a time), so

$$\tilde{\beta}-\hat{\beta}=d(X)$$

  for some function of X alone.

- Both are conditionally unbiased, so taking conditional expectations,

$$\underbrace{\mathrm{E}\left[\tilde{\beta} \mid X\right]}_{=\beta}=\underbrace{\mathrm{E}\left[\hat{\beta} \mid X\right]}_{=\beta}+d(X) \quad \Longrightarrow \quad d(X)=0.$$

- Therefore $\tilde{\beta}=\hat{\beta}$. A linear conditionally unbiased competitor that is also best attains the same variance matrix: both $M=\operatorname{Var}\left(\tilde{\beta} \mid X\right)-\operatorname{Var}\left(\hat{\beta} \mid X\right)$ and $-M$ are then positive semidefinite, so $c^{\top} M c=0$ for every c, and a symmetric matrix with that property is zero. So OLS is not just a best linear unbiased estimator; it is the only one.

Where Assumption 3 was used

- Assumption 3 entered in step 3, where $\mathrm{E}\left[UU^{\top} \mid X\right]$ was replaced by $\sigma^2 I_n$, and again when $\sigma^2(X^{\top}X)^{-1}$ was recognised as $\operatorname{Var}\left(\hat{\beta} \mid X\right)$. Without it the cancellation that produced $\sigma^2 AX$ fails, and the covariance need not equal the variance of OLS.
- Assumptions 1, 2 and 4 give unbiasedness, and Assumptions 1 and 4 with finite conditional second moments give the sandwich formula. Assumption 3 is what makes OLS efficient.
- If Assumption 3 fails, OLS is still unbiased but need not be best: Assumption 3 is sufficient for efficiency, not necessary.

Two extensions, and the summary

Adding normality

- **Assumption 5.** $U \mid X \sim N(0, \sigma^2 I_n)$. Assumptions 1 to 5 define the classical normal regression model.
- Given X, $\hat{\beta} = \beta + (X^\top X)^{-1} X^\top U$ is a constant plus a linear function of U, and Assumption 5 fixes the conditional distribution of U. For non-random α and Γ, such a function of a normal vector is normal:

$$V \sim N(\mu, \Sigma) \quad \Longrightarrow \quad \alpha + \Gamma V \sim N(\alpha + \Gamma\mu,\ \Gamma\Sigma\Gamma^\top).$$

- Under Assumptions 1, 4 and 5, since Assumption 5 implies Assumptions 2 and 3, the conditional mean and variance are the ones already computed:

$$\boxed{\hat{\beta} \mid X \ \sim \ N(\beta,\ \sigma^2 (X^\top X)^{-1})}$$

- Nothing before this slide needed normality. It is what confidence intervals and tests will be built on.

If Assumption 3 fails: generalised least squares

- Suppose now that $\mathrm{Var}\,(U \mid X) = \Omega$ for a known positive definite Ω, not necessarily diagonal, with $\Omega \neq \sigma^2 I_n$ for every $\sigma^2 > 0$. Let $\Omega^{-1/2}$ be the symmetric positive definite matrix with $\Omega^{-1/2}\Omega^{-1/2} = \Omega^{-1}$, built from the eigenvalue decomposition $\Omega = C\Lambda C^\top$, and write $\Omega^{1/2}$ for the inverse of $\Omega^{-1/2}$.
- Multiply the model through by $\Omega^{-1/2}$:

$$\underbrace{\Omega^{-1/2} Y}_{Y^*} = \underbrace{\Omega^{-1/2} X}_{X^*}\,\beta + \underbrace{\Omega^{-1/2} U}_{U^*}, \qquad \mathrm{E}\,[U^* \mid X] = 0, \qquad \mathrm{Var}\,(U^* \mid X) = \Omega^{-1/2}\Omega\Omega^{-1/2} = I_n,$$

  by taking out what is known and by the rule for a linear map.
- The transformed model satisfies all four assumptions (X^* has rank k because $\Omega^{-1/2}$ is nonsingular), so Gauss-Markov applies to it. Since $A \mapsto A\,\Omega^{1/2}$ matches the linear conditionally unbiased estimators of the two models one to one, OLS on the transformed model is also best among estimators linear in Y and conditionally unbiased, and it is

$$\hat{\beta}^* = (X^{*\top} X^*)^{-1} X^{*\top} Y^* = (X^\top \Omega^{-1} X)^{-1} X^\top \Omega^{-1} Y,$$

  which is weighted least squares with $W = \Omega^{-1}$.
- In practice Ω is not known. Later lectures return to this.

Summary: unbiasedness and variance

- Substituting the model into the estimator gives the sampling error $\hat{\beta} = \beta + (X^\top X)^{-1} X^\top U$. Everything follows from it.
- **Unbiasedness.** Under Assumptions 1, 2 and 4, conditioning on X makes $(X^\top X)^{-1} X^\top$ known, so $\mathrm{E}\left[\hat{\beta} \mid X\right] = \beta$, and the law of iterated expectations gives $\mathrm{E}\left[\hat{\beta}\right] = \beta$ whenever that mean exists. The weaker condition $\mathrm{E}\,[X_i U_i] = 0$ is not enough, because once $\mathrm{E}\,[U \mid X]$ is not zero, the expectation of the noise term need not vanish, and the parabola shows that zero correlation does not make the conditional mean zero.
- **Variance.** Under Assumptions 1 and 4, given finite conditional second moments,

$$\mathrm{Var}\left(\hat{\beta} \mid X\right) = (X^\top X)^{-1} X^\top \mathrm{Var}\,(U \mid X)\, X (X^\top X)^{-1},$$

  and adding Assumption 3, that $\mathrm{Var}\,(U \mid X) = \sigma^2 I_n$, gives

$$\mathrm{Var}\left(\hat{\beta} \mid X\right) = \sigma^2 (X^\top X)^{-1}.$$

  With one regressor and no intercept this is $\sigma^2 / \sum_i X_i^2$: noisier errors raise it; a larger $\sum_i X_i^2$ lowers it.

Summary: efficiency

- **Gauss-Markov.** Under Assumptions 1 to 4, OLS has the smallest conditional variance among all estimators that are linear in Y and conditionally unbiased, whatever the value of β. For every such $\tilde{\beta}$, the difference of the two conditional variance matrices is positive semidefinite:

  $$\mathrm{Var}\left(\tilde{\beta} \mid X\right) - \mathrm{Var}\left(\hat{\beta} \mid X\right) \;\geq\; 0,$$

  so in particular $\mathrm{Var}\left(\tilde{\beta}_j \mid X\right) \geq \mathrm{Var}\left(\hat{\beta}_j \mid X\right)$ for every coefficient.

- **The proof.** Linearity and conditional unbiasedness force $AX = I_k$. Assumption 3 then makes $\mathrm{Cov}\left(\tilde{\beta}, \hat{\beta} \mid X\right) = \mathrm{Var}\left(\hat{\beta} \mid X\right)$. Expanding $\mathrm{Var}\left(\tilde{\beta} - \hat{\beta} \mid X\right)$, which is positive semidefinite, leaves exactly the difference of the two variances.
- **Uniqueness.** Equality of the variances forces $\tilde{\beta} = \hat{\beta}$.
- Assumption 3 is what buys efficiency. Assumptions 1, 2 and 4 give unbiasedness on their own.