Unit 2. The single-equation linear model

Block C. When is that meaning causal?

Instructor: Henry Redondo

Dpto. de Fundamentos del Análisis Económico
Universidad de Alicante

Econometrics II · Academic Year 2026/2027

Outline of Block C

1. C. When is that meaning causal?
    - Asymptotic properties of the OLS estimator
    - Finite-sample properties of the OLS estimator
    - Endogeneity
    - Omitted variables
    - Proxy variables
    - Measurement error
    - Simultaneity

Asymptotic properties of the OLS estimator

- In this section we are going to study the asymptotic properties of the OLS estimator.
- We first study under which assumptions the OLS estimator is consistent.
- Then, we will obtain the asymptotic distribution of the OLS estimator.
- Finally, we will find the estimated variance matrix of the OLS estimator.
- *Asymptotic* means that we ask what happens to $\widehat{\boldsymbol{\beta}}$ as the sample size N grows. The answers only need $OLS1$ and $OLS2$.
- After that we study the finite-sample properties: unbiasedness, the variance matrix and the Gauss–Markov theorem. They hold for any N, but they need stronger assumptions.
- Both parts answer the question of this block: when does $\widehat{\boldsymbol{\beta}}$ recover the $\boldsymbol{\beta}$ we want to interpret?

Unbiasedness and consistency

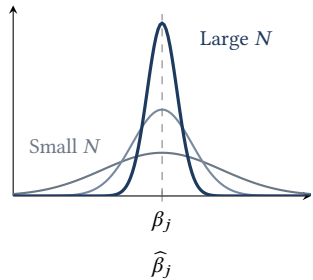
- $\widehat{\beta}_j$ is random: different samples produce different estimates.
- **Unbiasedness (fixed N):**

$$E(\widehat{\beta}_j) = \beta_j.$$

  Across repeated samples, estimates average to the true parameter.

- **Consistency (as $N \to \infty$):**

$$\widehat{\beta}_j \xrightarrow{p} \beta_j.$$



Sampling distributions of an unbiased, consistent estimator.

Consistency

Theorem (consistency). Under assumptions *OLS*1 and *OLS*2, the *OLS* estimator is a **consistent estimator** of $\boldsymbol{\beta}$.

- *Proof*
  - Substituting $y_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i$ in (7) we have that

$$\begin{aligned}\widehat{\boldsymbol{\beta}} &= \left(\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\left(\sum_{i=1}^{N}\mathbf{x}_i y_i\right) = \left(\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\left(\sum_{i=1}^{N}\mathbf{x}_i\left(\mathbf{x}_i'\boldsymbol{\beta}+u_i\right)\right)\\ &= \boldsymbol{\beta} + \left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i u_i\right) \qquad (21)\end{aligned}$$

  - Given that we have a random sample, $\mathbf{x}_i u_i$ are *iid*, using the law of large numbers we have that

$$\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i u_i \rightarrow_p E(\mathbf{x}u) \underset{\text{by } OLS1}{=} 0 \qquad (22)$$

Proof of consistency, continued

- *Proof (cont.)*
  - $\mathbf{x}_i\mathbf{x}_i'$ are also *iid*, and using the law of large numbers we have that

$$\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i' \rightarrow_p E(\mathbf{x}\mathbf{x}')$$

  - and since $E(\mathbf{x}\mathbf{x}')$ is non-singular (by *OLS2*), using the continuous function theorem, we have that

$$\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i'\right)^{-1} \rightarrow_p E(\mathbf{x}\mathbf{x}')^{-1} \tag{23}$$

  - Then, using (22), (23) and the continuous function theorem, we have that

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i u_i\right) \rightarrow_p \boldsymbol{\beta}$$

From consistency to the asymptotic distribution

- **What consistency gives:** $\widehat{\boldsymbol{\beta}} \to_p \boldsymbol{\beta}$ as $N \to \infty$.
- **What it does not give:** how far $\widehat{\boldsymbol{\beta}}$ can be from $\boldsymbol{\beta}$ in a sample of, say, $N = 1{,}000$. Without that, we have no confidence intervals and no tests.
- For inference we need the **sampling distribution** of $\widehat{\boldsymbol{\beta}}$:
  - *Exact*: requires an assumption on the distribution of u, e.g. normality (Block D).
  - *Asymptotic*: follows from the CLT, with no normality needed.
- **Next:** we state the CLT for a sample mean and apply it to $\frac{1}{N}\sum_i \mathbf{x}_i u_i$, the same average that appeared in the consistency proof.

Central Limit Theorem

Definition (Central Limit Theorem)

Let $\{X_i\}_{i=1}^n$ be independent and identically distributed random variables such that

$$E[X_i] = \mu, \quad Var(X_i) = \sigma^2 < \infty.$$

Define the sample mean

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Then, as $n \to \infty$,

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2).$$

Even if the X_i are not normally distributed, their average becomes approximately normal when n is large.

Asymptotic normality

Theorem (asymptotic normality). Under *OLS*1 and *OLS*2, if $E(u^2\mathbf{x}\mathbf{x}')$ exists, the *OLS* estimator is asymptotically normal and its asymptotic distribution is:

$$\sqrt{N}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to_d N(0, E(\mathbf{x}\mathbf{x}')^{-1}E(u^2\mathbf{x}\mathbf{x}')E(\mathbf{x}\mathbf{x}')^{-1})$$

- *Proof*
  - We have seen that
  $$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i u_i\right)$$
  - Multiplying by $\sqrt{N}$ we have
  $$\sqrt{N}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i u_i\right)$$
  - Given that we have a random sample, $\mathbf{x}_i u_i$ are iid, and since $E(\mathbf{x}_i u_i) = 0$ (by *OLS*1) and $E(u_i^2\mathbf{x}_i\mathbf{x}_i')$ exists, using the central limit theorem we have that
  $$\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i u_i\right) \to_d N(0, E(u^2\mathbf{x}\mathbf{x}')) \tag{24}$$

Proof of asymptotic normality, continued

- *Proof (cont.)*
  - $\mathbf{x}_i\mathbf{x}_i'$ are also *iid*, then, using the law of large numbers

$$\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i' \to_p E(\mathbf{x}\mathbf{x}') \tag{25}$$

  - and since $E(\mathbf{x}\mathbf{x}')$ is non-singular (by *OLS2*), using the continuous function theorem, we have that

$$\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i'\right)^{-1} \to_p E(\mathbf{x}\mathbf{x}')^{-1} \tag{26}$$

  - Then, using (24), (26) and Slutsky's theorem, we have that

$$\begin{aligned}\sqrt{N}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}) &= \left(\tfrac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\sqrt{N}\left(\tfrac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i u_i\right)\\ &\to_d N(0, E(\mathbf{x}\mathbf{x}')^{-1}E(u^2\mathbf{x}\mathbf{x}')E(\mathbf{x}\mathbf{x}')^{-1})\end{aligned}$$

No normal errors needed. Prices are skewed and heteroskedastic, and $\widehat{\boldsymbol{\beta}}$ is still approximately normal with $N = 1{,}000$ listings.

Remarks on asymptotic normality

- **Remarks**
  - Under the **homoskedasticity assumption**:

$$E(u^2\mathbf{x}\mathbf{x}') = \sigma^2 E(\mathbf{x}\mathbf{x}'), \text{ where } \sigma^2 = E(u^2) \qquad OLS3$$

    the asymptotic distribution of the *OLS* estimator is:

$$\sqrt{N}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \rightarrow_d N(0, \sigma^2 E(\mathbf{x}\mathbf{x}')^{-1})$$

  - When *OLS*3 is not satisfied the errors are said to be **heteroskedastic**.
  - Assumption *OLS*3 is the weakest homoskedasticity assumption. A sufficient condition for *OLS*3 is

$$E(u^2 \mid \mathbf{x}) = \sigma^2 \qquad OLS3'$$

    which is equivalent to $var(u \mid \mathbf{x}) = \sigma^2$ if $E(u \mid \mathbf{x}) = 0$.

Why we need to estimate the variance matrix

- The asymptotic variance

$$E(\mathbf{x}\mathbf{x}')^{-1}E(u^2\mathbf{x}\mathbf{x}')E(\mathbf{x}\mathbf{x}')^{-1}$$

  depends on population moments that we do not know and on the error u, which we do not observe.
- We estimate it with the same idea as OLS itself: replace population moments by sample averages (law of large numbers) and the errors u_i by the residuals $\widehat{u}_i$.
- There are two versions:
  - under homoskedasticity (*OLS*3), simpler, but valid only if *OLS*3 holds;
  - heteroskedasticity-robust, valid under *OLS*1 and *OLS*2 alone.
- The square roots of the diagonal are the **standard errors** that `lm` and Stata report, and that every t test of Block D uses.

Variance matrix under homoskedasticity

- Recall that under $OLS1$–$OLS2$,

$$\sqrt{N}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to_d N\left(\mathbf{0},\, \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}\right), \qquad \mathbf{A} = E(\mathbf{x}\mathbf{x}'), \quad \mathbf{B} = E(u^2\mathbf{x}\mathbf{x}')$$

- $OLS3$ (homoskedasticity) states that

$$\mathbf{B} = E(u^2\mathbf{x}\mathbf{x}') = \sigma^2 E(\mathbf{x}\mathbf{x}') = \sigma^2\mathbf{A}, \qquad \sigma^2 = E(u^2)$$

- The sandwich then collapses:

$$\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1} = \sigma^2\mathbf{A}^{-1} \quad \Longrightarrow \quad Avar(\widehat{\boldsymbol{\beta}}) = \frac{1}{N}\,\sigma^2\left[E(\mathbf{x}\mathbf{x}')\right]^{-1}$$

- This depends on two unknown population quantities, σ^2 and $E(\mathbf{x}\mathbf{x}')$, which we need to estimate.

Estimating σ^2 and $E(\mathbf{x}\mathbf{x}')$

- By the LLN,

$$\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i' \to_p E(\mathbf{x}\mathbf{x}') = \mathbf{A},$$

  and $\mathbf{A}$ is invertible by *OLS*2.
- Under *OLS*1–*OLS*3,

$$\widehat{\sigma}^2 = \frac{1}{N-k}\sum_{i=1}^{N}\widehat{u}_i^2 \to_p \sigma^2, \qquad \widehat{u}_i = y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}}$$

  - This holds because $\widehat{u}_i = u_i - \mathbf{x}_i'(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ and $\widehat{\boldsymbol{\beta}} \to_p \boldsymbol{\beta}$.
  - The $N - k$ is a degrees-of-freedom correction that does not matter asymptotically.
- By Slutsky's theorem,

$$\widehat{\sigma}^2\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i'\right)^{-1} \to_p \sigma^2\left[E(\mathbf{x}\mathbf{x}')\right]^{-1}$$

Estimated variance matrix and standard errors

- The estimated variance matrix of $\widehat{\boldsymbol{\beta}}$, valid under homoskedasticity, is

$$\widehat{var(\widehat{\boldsymbol{\beta}})} = \frac{1}{N}\widehat{\sigma}^2 \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i' \right)^{-1} = \widehat{\sigma}^2 \left(\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i' \right)^{-1} = \widehat{\sigma}^2 (\mathbf{X}'\mathbf{X})^{-1}$$

- The square roots of the elements on the main diagonal of $\widehat{var(\widehat{\boldsymbol{\beta}})}$ are the standard errors, denoted $SE(\widehat{\beta}_j)$.
- If $OLS3$ fails, then $\mathbf{B} \neq \sigma^2 \mathbf{A}$ and this estimator is inconsistent. In that case we need heteroskedasticity-robust standard errors.

Heteroskedasticity-robust variance matrix

- When $OLS3$ is not satisfied the estimated variance of $\widehat{\beta}$ given above is no longer valid.
- However, it can be shown that, under $OLS1$ and $OLS2$

$$\frac{1}{N}\sum_{i=1}^{N}\widehat{u}_i^2\mathbf{x}_i\mathbf{x}_i' \rightarrow_p E(u^2\mathbf{x}\mathbf{x}')$$

- Then

$$\left(\tfrac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\tfrac{1}{N}\sum_{i=1}^{N}\widehat{u}_i^2\mathbf{x}_i\mathbf{x}_i'\left(\tfrac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}$$
$$\rightarrow_p E(\mathbf{x}\mathbf{x}')^{-1}E(u^2\mathbf{x}\mathbf{x}')E(\mathbf{x}\mathbf{x}')^{-1}$$

- and the heteroskedasticity-robust estimator is given by

$$\widetilde{var(\widehat{\boldsymbol{\beta}})} = \left(\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\sum_{i=1}^{N}\widehat{u}_i^2\mathbf{x}_i\mathbf{x}_i'\left(\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}$$

- The square root of the elements in the main diagonal of $\widetilde{var(\widehat{\boldsymbol{\beta}})}$ are called the heteroskedasticity-robust standard errors and are denoted by $SE(\widehat{\beta}_j)$.

Analysis D1 · Classical vs. robust standard errors

Levels (€)

	Classical	Robust	Ratio
m2	46.12	71.93	1.56
m2_terraza	269.78	431.08	1.60
terraza	4,269.89	5,273.38	1.24
bajo	3,953.68	3,317.22	0.84

Logs

	Classical	Robust	Ratio
lm2	0.0178	0.0186	1.05
m2_terraza	0.0011	0.0014	1.26
terraza	0.0171	0.0189	1.11
bajo	0.0158	0.0157	0.99

Analysis D1 · Classical vs. robust standard errors

Levels (€)

	Classical	Robust	Ratio
`m2`	46.12	71.93	1.56
`m2_terraza`	269.78	431.08	1.60
`terraza`	4,269.89	5,273.38	1.24
`bajo`	3,953.68	3,317.22	0.84

Logs

	Classical	Robust	Ratio
`lm2`	0.0178	0.0186	1.05
`m2_terraza`	0.0011	0.0014	1.26
`terraza`	0.0171	0.0189	1.11
`bajo`	0.0158	0.0157	0.99

- **In euros**, a 400,000 € beach flat misses by far more euros than a 70,000 € flat in San Blas: strong heteroskedasticity. The classical s.e. of `m2` is 36% too small.
- **In logs**, errors are in per cent and are much closer to homoskedastic, so the two columns are closer, though not identical (`m2_terraza`: 1.26).
- Robust can also be *smaller* (`bajo` in levels: 0.84). It corrects in whichever direction the data require.

Rule. Report robust standard errors by default: `coeftest(m, vcovHC(m, "HC1"))`. If they barely differ from the classical ones, you have lost nothing.

Finite-sample properties of the OLS estimator

- In this section we are going to study the finite-sample properties of the OLS estimator.
- We first study under which assumptions the OLS estimator is unbiased.
- Then, we will obtain the variance matrix of the OLS estimator.
- Finally, we will prove that, under some assumptions, the OLS estimator is optimal.
- These results hold for *any* sample size, not only as $N \to \infty$, and they do not use the law of large numbers or the CLT.
- The price is stronger assumptions: $OLS1'$, $E(u \mid \mathbf{x}) = 0$, instead of $OLS1$, and, for the variance and Gauss–Markov, homoskedasticity $OLS3'$.

Why $E(\mathbf{x}u) = 0$ is not enough for unbiasedness

- From the proof of consistency, $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$. Unbiasedness requires

$$E\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\right] = 0.$$

- $(\mathbf{X}'\mathbf{X})^{-1}$ is a *nonlinear* function of all the regressors in the sample. Zero correlation between each $\mathbf{x}_i$ and u_i does not make this expectation zero.
- For consistency this did not matter: as N grows, $\frac{1}{N}\mathbf{X}'\mathbf{X}$ settles at $E(\mathbf{x}\mathbf{x}')$, a constant, and only $E(\mathbf{x}u) = 0$ is needed.
- In a finite sample, the solution is to condition on $\mathbf{X}$. Given $\mathbf{X}$, $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is a matrix of numbers, and what we need is $E(\mathbf{u} \mid \mathbf{X}) = 0$. That is $OLS1'$, the stronger assumption of Block B.

Unbiasedness

Theorem (unbiasedness). Under assumptions *OLS*1′ and *OLS*2, the *OLS* estimator is an **unbiased estimator** of $\boldsymbol{\beta}$.

Proof

- Under *OLS*1′, $E(u_i \mid \mathbf{x}_i) = 0$ for all $i = 1, .., N$, and given that the vectors $(\mathbf{x}_i', y_i)'$, $i = 1, .., N$ are independent, we have that

$$E(u_i \mid \mathbf{x}_i) = E(u_i \mid \mathbf{x}_1, \mathbf{x}_2, .., \mathbf{x}_N) = E(u_i \mid \mathbf{X}) = 0, \quad \text{for all } i = 1, .., N$$

- and using a compact notation we have that

$$E(\mathbf{u} \mid \mathbf{X}) = 0 \tag{27}$$

where now $\mathbf{u}$ is

$$\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{bmatrix}$$

Proof of unbiasedness, continued

Proof (cont.)

- Using the matrix definition of $\widehat{\boldsymbol{\beta}}$ in (8) and substituting $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ we have that

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\left(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}\right) = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$$

- Then, taking expectations conditional on $\mathbf{X}$ and using (27), we have that

$$E(\widehat{\boldsymbol{\beta}} \mid \mathbf{X}) = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{u} \mid \mathbf{X}) = \boldsymbol{\beta}$$

- and using the law of iterated expectations

$$E(\widehat{\boldsymbol{\beta}}) = E(E(\widehat{\boldsymbol{\beta}} \mid \mathbf{X})) = \boldsymbol{\beta}$$

Conditional variance matrix

Theorem (conditional variance). Under assumptions $OLS1'$, $OLS2$ and $OLS3'$

$$var(\widehat{\boldsymbol{\beta}} \mid \mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

Proof

- Using the definition of the conditional variance

$$var(\widehat{\boldsymbol{\beta}} \mid \mathbf{X}) = E[(\widehat{\boldsymbol{\beta}} - E(\widehat{\boldsymbol{\beta}} \mid \mathbf{X}))(\widehat{\boldsymbol{\beta}} - E(\widehat{\boldsymbol{\beta}} \mid \mathbf{X}))' \mid \mathbf{X}]$$

- Since $OLS1'$ and $OLS2$ are satisfied, we know that $E(\widehat{\boldsymbol{\beta}} \mid \mathbf{X}) = \boldsymbol{\beta}$, then

$$var(\widehat{\boldsymbol{\beta}} \mid \mathbf{X}) = E[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \mid \mathbf{X}]$$

- Now substituting $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$, we have that

$$\begin{aligned} var(\widehat{\boldsymbol{\beta}} \mid \mathbf{X}) &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\mathbf{u}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mid \mathbf{X}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{u}\mathbf{u}' \mid \mathbf{X})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \end{aligned} \tag{28}$$

Proof of the conditional variance matrix, continued

Proof (cont.)

- We know that

$$E(\mathbf{u}\mathbf{u}' \mid \mathbf{X}) = \begin{bmatrix} E(u_1^2 \mid \mathbf{X}) & E(u_1 u_2 \mid \mathbf{X}) & \cdots & E(u_1 u_N \mid \mathbf{X}) \\ E(u_2 u_1 \mid \mathbf{X}) & E(u_2^2 \mid \mathbf{X}) & \cdots & E(u_2 u_N \mid \mathbf{X}) \\ \vdots & \vdots & & \vdots \\ E(u_N u_1 \mid \mathbf{X}) & E(u_N u_2 \mid \mathbf{X}) & \cdots & E(u_N^2 \mid \mathbf{X}) \end{bmatrix}$$

- Since the observations are *iid* and $OLS3'$ is satisfied, we have that

$$\begin{aligned} E(u_i^2|\mathbf{X}) &= \sigma^2, \quad i = 1, 2, .., N \\ E(u_i u_j|\mathbf{X}) &= 0, \quad i, j = 1, 2, .., N, \ i \neq j \end{aligned}$$

- Therefore $E(\mathbf{u}\mathbf{u}' \mid \mathbf{X}) = \sigma^2 I_N$, and substituting in (28) and simplifying we have

$$var(\widehat{\boldsymbol{\beta}} \mid \mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' I_N \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

Is OLS the best we can do?

- Under *OLS*1′, *OLS*2 and *OLS*3′, OLS is unbiased with variance $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. It is not the only unbiased estimator. In the simple regression model, for example:
  - OLS on the first 500 listings only;
  - the slope between two listings, $(y_2 - y_1)/(x_2 - x_1)$.

  Both are linear in $\mathbf{y}$ and unbiased given $\mathbf{X}$, but they throw information away.
- Among all the estimators that are linear in $\mathbf{y}$ and unbiased, which one has the smallest variance?
- The Gauss–Markov theorem answers: OLS. "Smallest" is in the matrix sense: $var(\widetilde{\boldsymbol{\beta}} \mid \mathbf{X}) - var(\widehat{\boldsymbol{\beta}} \mid \mathbf{X})$ is positive semidefinite, so every linear combination $\mathbf{c}'\widehat{\boldsymbol{\beta}}$, each single coefficient included, has a smaller variance than $\mathbf{c}'\widetilde{\boldsymbol{\beta}}$.

Gauss–Markov theorem

Theorem (Gauss–Markov). Under assumptions $OLS1'$, $OLS2$ and $OLS3'$, the OLS estimator is the **Best Linear Unbiased Estimator (BLUE)** of $\boldsymbol{\beta}$.

Proof

- Let $\widetilde{\boldsymbol{\beta}}$ be another estimator of $\boldsymbol{\beta}$ that is linear in $\mathbf{y}$ and unbiased conditional on $\mathbf{X}$, i.e.

$$\widetilde{\boldsymbol{\beta}} = \mathbf{C}\mathbf{y} \tag{29}$$

  and

$$E(\widetilde{\boldsymbol{\beta}} \mid \mathbf{X}) = \boldsymbol{\beta} \tag{30}$$

- We have to prove that

$$var(\widetilde{\boldsymbol{\beta}} \mid \mathbf{X}) \geq var(\widehat{\boldsymbol{\beta}} \mid \mathbf{X})$$

  i.e. we have to show that $var(\widetilde{\boldsymbol{\beta}} \mid \mathbf{X}) - var(\widehat{\boldsymbol{\beta}} \mid \mathbf{X})$ is a positive semidefinite matrix.

Proof of the Gauss–Markov theorem, continued

Proof (cont.)

- Substituting $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ in (29), we can write $\widetilde{\boldsymbol{\beta}}$ as:

$$\widetilde{\boldsymbol{\beta}} = \mathbf{C}(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) = \mathbf{C}\mathbf{X}\boldsymbol{\beta} + \mathbf{C}\mathbf{u} \tag{31}$$

- Taking expectations conditional on $\mathbf{X}$, we have

$$E(\widetilde{\boldsymbol{\beta}} \,|\, \mathbf{X}) = E(\mathbf{C}\mathbf{X}\boldsymbol{\beta} + \mathbf{C}\mathbf{u} \,|\, \mathbf{X}) = \mathbf{C}\mathbf{X}\boldsymbol{\beta} + \mathbf{C}E(\mathbf{u} \,|\, \mathbf{X}) \underset{\text{by } OLS1'}{=} \mathbf{C}\mathbf{X}\boldsymbol{\beta}$$

- Then, using (30), we have that $\mathbf{C}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$, which implies that $\mathbf{C}\mathbf{X} = I_k$ and substituting in (31) we have that

$$\widetilde{\boldsymbol{\beta}} = \boldsymbol{\beta} + \mathbf{C}\mathbf{u}$$

- Moreover, under $OLS3'$

$$\begin{aligned} var(\widetilde{\boldsymbol{\beta}} \,|\, \mathbf{X}) &= E((\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})' \,|\, \mathbf{X}) \\ &= E(\mathbf{C}\mathbf{u}\mathbf{u}'\mathbf{C}' \,|\, \mathbf{X}) = \mathbf{C}E(\mathbf{u}\mathbf{u}' \,|\, \mathbf{X})\mathbf{C}' = \sigma^2\mathbf{C}\mathbf{C}' \end{aligned}$$

Proof of the Gauss–Markov theorem, continued

Proof (cont.)

- We know that under $OLS1'$, $OLS2$ and $OLS3'$

$$var(\widehat{\boldsymbol{\beta}} \mid \mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

- Then, to prove that $Var(\widetilde{\boldsymbol{\beta}} \mid \mathbf{X}) - Var(\widehat{\boldsymbol{\beta}} \mid \mathbf{X})$ is a positive semidefinite matrix, we have to prove that

$$\sigma^2\mathbf{C}\mathbf{C}' - \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \geq 0 \iff \mathbf{C}\mathbf{C}' - (\mathbf{X}'\mathbf{X})^{-1} \geq 0$$

- Let

$$\mathbf{D} = \mathbf{C} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

$$\begin{aligned}
\mathbf{D}\mathbf{D}' &= (\mathbf{C} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')(\mathbf{C} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')' \\
&= \mathbf{C}\mathbf{C}' - \mathbf{C}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}' + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \mathbf{C}\mathbf{C}' - (\mathbf{X}'\mathbf{X})^{-1}
\end{aligned}$$

$\mathbf{D}\mathbf{D}'$ is a positive semidefinite matrix and therefore
$Var(\widetilde{\boldsymbol{\beta}} \mid \mathbf{X}) - Var(\widehat{\boldsymbol{\beta}} \mid \mathbf{X})$ is a positive semidefinite matrix.

Remarks on the finite-sample results

Unbiasedness, the conditional variance $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ and Gauss–Markov hold for *any* sample size, which is elegant. But they rest on $E(u \mid \mathbf{x}) = 0$, the assumption that fails when quality is omitted (next subsections), and Gauss–Markov also needs homoskedasticity, which the listings reject (Analysis D1). In applied work, consistency and robust inference do the heavy lifting.

BLUE means best among *linear, unbiased* estimators. A biased estimator with a much smaller variance can predict better, which is the whole idea of ridge and lasso.

Exercise on Gauss–Markov

Exercise.

1. Does the Gauss–Markov theorem require normally distributed errors?
2. Does it say OLS has the smallest variance among *all* estimators?

Exercise on Gauss–Markov

Exercise.

1. Does the Gauss–Markov theorem require normally distributed errors?
2. Does it say OLS has the smallest variance among *all* estimators?

Answer.

1. No. It only uses $OLS1'$, $OLS2$ and $OLS3'$.
2. No. Only among estimators that are **linear** in $\mathbf{y}$ and **unbiased** conditional on $\mathbf{X}$.

Exogeneity and endogeneity

- OLS is consistent under $OLS1$, $E(\mathbf{x}u) = \mathbf{0}$, and the rank condition $OLS2$. The stronger $OLS1'$, $E(u \mid \mathbf{x}) = 0$, also makes $\mathbf{x}'\boldsymbol{\beta}$ the conditional mean (Block B).
- x_j is **exogenous** in equation (1) if $cov(x_j, u) = 0$, and **endogenous** if $cov(x_j, u) \neq 0$. When assumption (2) holds, all explanatory variables in (1) are exogenous.
- If one regressor is endogenous, OLS still converges to the coefficients of the linear projection of y on $\mathbf{x}$, but these are no longer the structural β_j. In general *all* the $\widehat{\beta}_j$ are inconsistent, not only the coefficient of the endogenous regressor.

Where does endogeneity come from?

- In applied econometrics endogeneity arises mainly through three channels:

Omitted variables. We want to control for one or more additional variables but cannot include them because of data limitations.

e.g. unobserved ability in a wage equation.

Measurement error. We only have an imperfect measure of at least one of the explanatory variables.

e.g. self-reported income.

Simultaneity. At least one explanatory variable is determined simultaneously with y.

e.g. prices and quantities.

A structural model with an unobservable

- Consider the structural model

$$E(y \mid x_1, .., x_k, q) = \beta_0 + \beta_1 x_1 + .. + \beta_k x_k + \gamma q \tag{32}$$

  where q is unobservable. We are interested in estimating the β_j which are the partial effects of the observed explanatory variables holding the other explanatory variables (including the unobservable q) constant.
- Notice that in the context of the linear model there is no point in allowing for more than one unobservable factor, all the unobservables are included in q.
- *Our case*
  Let $y = \log(precio)$ and let q denote the unobserved quality of the flat (finishes, natural light, state of renovation). If $x_k = terraza$ is a dummy for having a terrace, then β_k in equation (32) is the terrace premium: the approximate percentage difference in price, between flats with and without a terrace that are otherwise identical in their observed characteristics and in q.

Putting the omitted variable in the error

- We can write (32) in error form as

$$y = \beta_0 + \beta_1 x_1 + .. + \beta_k x_k + \gamma q + v \quad \text{where } E(v \mid x_1, ..., x_k, q) = 0 \tag{33}$$

- We can include q in the error term and write the model as

$$y = \beta_0 + \beta_1 x_1 + .. + \beta_k x_k + u \quad \text{where } u = \gamma q + v \tag{34}$$

  Given that the model has an intercept we can assume that $E(u) = 0$, however if q is correlated with any of the regressors, then so is u, and we have an endogeneity problem, and, in general, the *OLS* estimator of (34) will not be a consistent estimator of any of the β_j.
- We know that $\widehat{\boldsymbol{\beta}}$ (the *OLS* estimator of (34)) is a consistent estimator of the coefficients of the *LP* of y on $1, x_1, ..x_k$, but the coefficients of this *LP* will not in general coincide with the β_j (unless q is uncorrelated with all the x_j).

The probability limit under omission

- Then, since $\widehat{\beta}$ is a consistent estimator of the coefficients of the LP of y on $1, x_1, ..x_k$, to compute the $plim$ of $\widehat{\beta}$ we have to find the coefficients of the LP of y on $1, x_1, ..x_k$.
- Using the law of iterated LP, we have that

$$L(y \mid 1, x_1, .., x_k) = L\left(L(y \mid 1, x_1, .., x_k, q) \mid 1, x_1, .., x_k\right) \tag{35}$$

- and since $E(y \mid x_1, ...x_k, q)$ is linear, we have that

$$L(y \mid 1, x_1, .., x_k, q) = E(y \mid x_1, ..., x_k, q) = \beta_0 + \beta_1 x_1 + .. + \beta_k x_k + \gamma q$$

- Then substituting in (35), we have that

$$L(y \mid 1, x_1, .., x_k) = \beta_0 + \beta_1 x_1 + .. + \beta_k x_k + \gamma L(q|1, x_1, .., x_k)$$

- Let $L(q \mid 1, x_1, .., x_k) = \delta_0 + \delta_1 x_1 + .. + \delta_k x_k$ be the LP of q on $1, x_1, ..x_k$.
- Then $$L(y \mid 1, x_1, .., x_k) = (\beta_0 + \gamma\delta_0) + (\beta_1 + \gamma\delta_1)x_1 + .. + (\beta_k + \gamma\delta_k)x_k$$
- and

$$plim\widehat{\beta}_j = \beta_j + \gamma\delta_j, \quad j = 0, .., k$$

The direction of omitted variable bias

- In the general case, when q might be correlated with all the x_j, it is very difficult to see which is the sign of the inconsistency bias.
- However, in the simple case when q is correlated with just one of the regressors, say x_k, it is often assumed that all δ_j except δ_0 and δ_k are zero.
- Then

$$plim\widehat{\beta}_j = \beta_j, \quad j = 1, .., k-1$$
$$plim\widehat{\beta}_k = \beta_k + \gamma \frac{cov(q,x_k)}{var(x_k)}$$

and the sign of the inconsistency bias is driven by the sign of $cov(\gamma q, x_k)$.

The direction of omitted variable bias

One omitted variable correlated with x_k only.

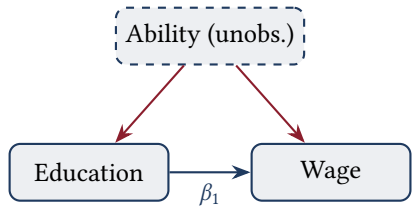
$$plim\,\widehat{\beta}_k - \beta_k = \gamma\,\frac{cov(q, x_k)}{var(x_k)}$$

	$cov(q, x_k) > 0$	$cov(q, x_k) < 0$
$\gamma > 0$	**upward** bias	**downward** bias
$\gamma < 0$	**downward** bias	**upward** bias

Omitted variables (education and wages)

Setup.

- Model: $Wage_i = \beta_0 + \beta_1 Education_i + \varepsilon_i$
- Omitted variable: ability or motivation.
- Ability affects both education and wages $\Rightarrow \text{Cov}(Education_i, \varepsilon_i) \neq 0$.
- OLS attributes the effect of ability to education.
- $\widehat{\beta_1}$ is **biased upward**.
- Solutions: a proxy for ability (next subsection) or instruments (Unit 3).



Signing the bias (training and productivity)

Exercise. We regress firm productivity on a training dummy, omitting management quality. Better-managed firms train more and are more productive. Sign of the bias?

Signing the bias (training and productivity)

Exercise. We regress firm productivity on a training dummy, omitting management quality. Better-managed firms train more and are more productive. Sign of the bias?

Answer. $\gamma > 0$ and $cov(q, x_k) > 0$: the effect of training is **overstated**.

Is the terrace premium causal?

- Block B gave us a number: a terrace is associated with **+11.0%**. The pricing team will use it as "add a terrace and the price goes up 11.0%". Is that a partial effect *holding everything else fixed*?

Four ways the listing data could fool us.

1. **Omitted variable.** Flats with terraces tend to be better overall: finishes, light, renovation. We do not observe *quality*.
2. **Proxy.** We do observe something correlated with quality: the number of photos in the ad.
3. **Bad control.** We could "control for" the energy certificate. Should we?
4. **Measurement error.** The m^2 in the ad are declared by the seller, not measured.

We have the oracle. For each problem we can compare what an analyst would estimate with what is actually true, and check the theory's predictions *numerically*.

Analysis C1 · Opening the oracle on omitted quality

Short vs. long regression

	Short (analyst)	Long (+ `calidad`)
`terraza`	0.1041	0.0329
`calidad`	–	0.0975
R^2	0.900	0.933

Auxiliary regression

`lm(update(f_log, calidad ~ .))`: coefficient on `terraza`

$$\widehat{\delta} = 0.7301$$

$\text{corr}(terraza, calidad) = 0.41$

Analysis C1 · Opening the oracle on omitted quality

Short vs. long regression

	Short (analyst)	Long (+ `calidad`)
`terraza`	0.1041	0.0329
`calidad`	–	0.0975
R^2	0.900	0.933

Auxiliary regression

`lm(update(f_log, calidad ~ .))`: coefficient on `terraza`

$$\widehat{\delta} = 0.7301$$

$\text{corr}(terraza, calidad) = 0.41$

The omitted-variable formula, checked.

$$\underbrace{0.1041 - 0.0329}_{\text{short}-\text{long} = 0.0712} \;=\; \underbrace{0.0975 \times 0.7301}_{\widehat{\gamma} \times \widehat{\delta} = 0.0712}$$

In the sample, the formula $plim\,\widehat{\beta}_j = \beta_j + \gamma\delta_j$ holds **exactly**, as an identity. Better flats have terraces ($\delta > 0$) and quality raises prices ($\gamma > 0$), so the analyst's terrace premium is about three times the premium holding quality fixed.

What is a proxy variable?

- Omitted variables bias can be eliminated, or at least mitigated, if there is a proxy variable for the unobservable q.
- z is said to be **a proxy** for q if the *OLS* estimator of the regression of y on $1, x_1, ..x_k, z$ consistently estimates the structural parameters $\beta_1, .., \beta_k$.
- There are two formal requirements for a proxy variable z :
  - **1.** z is irrelevant in explaining y once $\mathbf{x}$ and q have been controlled for. i.e.

$$E(y \mid \mathbf{x}, q, z) = E(y \mid \mathbf{x}, q) \tag{36}$$

    where $\mathbf{x} = (x_1, .., x_k)'$.
    - In our case z could be the number of photos in the ad. Requirement (36) says that, if we could control for the flat's true quality, the number of photos would not affect its price.
  - **2.** The correlation between q and each x_j is zero once we partial out z. i.e.

$$L(q \mid 1, x_1, .., x_k, z) = L(q \mid 1, z) \tag{37}$$

Why a proxy works

- We are going to show that, if requirements (36) and (37) are satisfied, the *OLS* estimators of the coefficients of $x_1, .., x_k$ in the regression of y on $1, x_1, .., x_k, z$ are consistent estimators of the structural parameters $\beta_1, .., \beta_k$.
- We know that the *OLS* estimators of the regression of y on $1, x_1, .., x_k, z$ consistently estimates the parameters of the *LP* of y on $1, x_1, .., x_k, z$. Thus, we are going to prove that the coefficients of $x_1, .., x_k$ in this *LP* are $\beta_1, .., \beta_k$.
- Using property 5 of the linear projection we have that

$$L(y \mid 1, \mathbf{x}, z) = L(E(y \mid \mathbf{x}, z, q) \mid 1, \mathbf{x}, z) \underset{\text{by (36)}}{=} L(E(y \mid \mathbf{x}, q) \mid 1, \mathbf{x}, z) \tag{38}$$

- and since

$$E(y \mid \mathbf{x}, q) = \beta_0 + \beta_1 x_1 + .. + \beta_k x_k + \gamma q$$

Why a proxy works (cont.)

- substituting in (38) we have that

$$L(y \mid 1, \mathbf{x}, z) = \beta_0 + \beta_1 x_1 + .. + \beta_k x_k + \gamma L(q \mid 1, \mathbf{x}, z)$$

- and using (37) we have that

$$\begin{aligned} L(y \mid 1, \mathbf{x}, z) &= \beta_0 + \beta_1 x_1 + .. + \beta_k x_k + \gamma L(q \mid 1, z) \\ &= \beta_0 + \beta_1 x_1 + .. + \beta_k x_k + \gamma (\theta_0 + \theta_1 z) \\ &= \beta_0 + \gamma\theta_0 + \beta_1 x_1 + .. + \beta_k x_k + \gamma\theta_1 z \end{aligned}$$

- and therefore, the *OLS* estimator of the regression of y on $1, x_1, .., x_k, z$ is a consistent estimator of $\beta_0 + \gamma\theta_0, \beta_1, .., \beta_k, \gamma\theta_1$.

Imperfect proxies

- When requirement (37) is not satisfied we say the z is an imperfect proxy.
- Then the LP of q on $1, x_1, ..x_k, z$ will be

$$q = \theta_0 + \rho_1 x_1 + .. + \rho_k x_k + \theta_1 z + r$$

  where not all the ρ_j are zero.
- In that case, the OLS estimator of the regression of y on $1, x_1, ..x_k, z$ will not produce consistent estimators of the β_j but

$$plim\widehat{\beta_j} = \beta_j + \gamma\rho_j$$

- The hope is that the bias is smaller than the bias of the OLS estimator of the regression of y on $1, x_1, ..x_k$ and this is what is assumed to be the case if z is a reasonable proxy.

Analysis C2 · The number of photos as a proxy for quality

lm(f_log + num_fotos), robust s.e.

	Coef.	Robust s.e.	
terraza	0.0705	0.0169	short 0.1041 → long 0.0329
num_fotos	0.0115	0.0008	corr($fotos, calidad$) = 0.74

Analysis C2 · The number of photos as a proxy for quality

`lm(f_log + num_fotos)`, robust s.e.

	Coef.	Robust s.e.	
`terraza`	0.0705	0.0169	short 0.1041 → long 0.0329
`num_fotos`	0.0115	0.0008	corr($fotos, calidad$) = 0.74

- The proxy removes $(0.1041 - 0.0705)/(0.1041 - 0.0329) = 47\%$ of the bias. Useful, not perfect: photos are an *imperfect* proxy (previous slides), so part of the bias survives.

"Each extra photo raises the price by 1.2%" is the wrong reading. The photos do not cause anything: the coefficient is there to *absorb* quality. A proxy earns its place by what it does to the *other* coefficients, and its own coefficient has no causal meaning.

Text length, photo count and "professional photography" flags are cheap proxies for unobserved quality. Machine-learning features often do this job without anyone naming it.

Summary of omitted variables

1. Omitting q puts it in the error term: OLS estimates the *LP* of y on the observed regressors, so $plim\,\widehat{\beta}_j = \beta_j + \gamma\delta_j$.
2. The bias disappears only if q is uncorrelated with the regressors (all $\delta_j = 0$).
3. With one correlated regressor, the sign of the bias is the sign of $\gamma \cdot cov(q, x_k)$.
4. A proxy z works if it is redundant given $(\mathbf{x}, q)$ and absorbs all the correlation between q and $\mathbf{x}$.
5. Imperfect proxies leave bias $\gamma\rho_j$, hopefully smaller than without them.

Measurement error vs. omitted variables

- Measurement errors and omitted variables problems are very similar from the statistical point of view (in both cases there is a variable that we do not observe) but are conceptually very different.
- In the omitted variable case, the variable that we cannot observe is in many cases vaguely defined.
- In the measurement error case, the variable that we cannot observe is perfectly well defined but for some reason we observe it with some error.

Error in the dependent variable

- Consider the model
$$y^* = \beta_0 + \beta_1 x_1 + .. + \beta_k x_k + v \tag{39}$$
that satisfies $OLS1$ and $OLS2$.
- Suppose that we do not observe y^*, we observe an imperfect measure y. The population measurement error is defined as
$$e_0 = y - y^*$$
- Then we can write model (39) as
$$y = \beta_0 + \beta_1 x_1 + .. + \beta_k x_k + u, \quad \text{where } u = v + e_0 \tag{40}$$
- If the measurement error has zero mean and it is uncorrelated with all the explanatory variables, as is generally assumed, model (40) will satisfy $OLS1$ and $OLS2$, and therefore, the OLS estimator will be a consistent estimator of the β_j.
- The only cost is precision: in general the variance of u is larger than the variance of v, so the asymptotic variance of the OLS estimator is larger than if we could observe y^*. A nonzero mean of e_0 only affects the constant term β_0.

Error in an explanatory variable

- Consider the model with one explanatory variable measured with error

$$y = \beta_0 + \beta_1 x_1 + .. + \beta_{k-1} x_{k-1} + \beta_k x_k^* + v \tag{41}$$

  where x_k^* is not observable.
- We assume model (41) satisfies $OLS1$ and $OLS2$, so that, if x_k^* were observed, the OLS estimator of (41) will be a consistent estimator of the β_j.
- We observe x_k which is an imperfect measure of x_k^*. The measurement error in the population is defined as:

$$e_k = x_k - x_k^*$$

- We assume that x_k is uncorrelated with v (x_k is redundant in the LP of y on $1, x_1, ..x_{k-1}, x_k^*$) which implies that e_k is uncorrelated with v.

Two polar assumptions on the measurement error

- The question now is which are the properties of the *OLS* estimator obtained by replacing x_k^* by x_k?. The answer to this question depends crucially on the assumptions about the measurement error.
- Generally it is assumed that the measurement error is uncorrelated with the other explanatory variables, i.e.

$$cov(x_j, e_k) = 0, \quad j = 1, .., k-1$$

- The key assumption is about the relationship between the measurement error and x_k and x_k^*.
- There are two alternative polar assumptions that have been the focus of the econometric literature:

$$cov(x_k, e_k) = 0 \quad \Leftrightarrow \quad \begin{array}{l} e_k \text{ is uncorrelated with} \\ \text{the observed measure } x_k \end{array} \tag{42}$$

$$cov(x_k^*, e_k) = 0 \quad \Leftrightarrow \quad \begin{array}{l} e_k \text{ is uncorrelated with} \\ \text{the unobserved measure } x_k^* \end{array} \tag{43}$$

When is measurement error harmless?

- We can write model (41) as

$$y = \beta_0 + \beta_1 x_1 + .. + \beta_k x_k + u, \quad \text{where } u = v - \beta_k e_k \tag{44}$$

- Under assumption (42), the composite error term u is uncorrelated with the x_j and the *OLS* estimator of (44) will be a consistent estimator of the β_j.
- The only consequence is that the variance of the error term u is larger than the variance of v, and therefore, the asymptotic variance of the *OLS* estimator is larger than in the case that we could observe x_k^*.
- Under (42) the *OLS* estimator still has all the nice properties, then, measurement error + assumption (42) is not really a problem.
- When econometricians talk about the problem of measurement errors in the explanatory variables, the underlying assumption is assumption (43), which is called the **classical errors-in-variables** assumption.

Attenuation bias under classical errors-in-variables

- Under assumption (43)

$$cov(x_k, u) = cov(x_k^* + e_k, v - \beta_k e_k) = -\beta_k \sigma_{e_k}^2 \neq 0$$

  and the *OLS* estimator of the regression of y on $1, x_1, ..x_{k-1}, x_k$ is inconsistent.
- The $plim$ of $\widehat{\beta}_j,\ j = 1, .., k-1$ is difficult to characterize in the general case. However, we can characterize the $plim$ of $\widehat{\beta}_k$
- It can be shown that

$$\widehat{\beta}_k \rightarrow_p \beta_k \frac{\sigma_{r_k^*}^2}{\sigma_{r_k^*}^2 + \sigma_{e_k}^2} \tag{45}$$

  where r_k^* is the *LP* error in

$$x_k^* = \delta_0 + \delta_1 x_1 + .. + \delta_{k-1} x_{k-1} + r_k^* \tag{46}$$

- Notice that $\frac{\sigma_{r_k^*}^2}{\sigma_{r_k^*}^2 + \sigma_{e_k}^2}$ is between 0 and 1 and hence $\left| plim \widehat{\beta}_k \right| \leq |\beta_k|$. This is called the attenuation bias due to classical errors-in-variables.

Attenuation and the other regressors

- In the case of just one explanatory variable ($k = 1$) measured with error

$$plim\widehat{\beta_1} = \beta_1 \frac{\sigma^2_{x_1^*}}{\sigma^2_{x_1^*} + \sigma^2_{e_1}}$$

- The general bias in (45) shows that when there are additional explanatory variables (apart from the one measured with error) what matters is not $var(x_k^*)$ but the variance of x_k^* after netting out the other explanatory variables. The more collinear is x_k^* with other explanatory variables the worse the attenuation bias.

Beyond the classical assumption

- If measurement error is present in more than one explanatory variable, deriving the inconsistency in the OLS estimators under extensions of the classical errors-in-variables assumption is complicated and does not lead to very useful results.
- There are some cases where it is clear that the classical errors-in-variables assumption cannot hold
  - For example, suppose that frequency of marijuana usage is used as an explanatory variable in a wage equation.
  - Let $smoked^*$ be the number of days, out of the last 30, that a worker has smoked marijuana, and the variable *smoked* is the self-reported number of days.
  - Suppose that people try to report the truth when asked about marijuana consumption. Then, non-smokers will always report zero consumption whereas smokers may miscount the number of days.
  - In this case, the measurement error is correlated with $smoked^*$.

Where is the error?

Exercise. In a wage equation, compare two problems:

(a) wages are misreported (error in y);

(b) years of schooling are misreported (classical error in x).

Which one biases the return to education?

Where is the error?

Exercise. In a wage equation, compare two problems:

(a) wages are misreported (error in y);

(b) years of schooling are misreported (classical error in x).

Which one biases the return to education?

Answer. Only (b). An error in y uncorrelated with the regressors just adds noise: OLS stays consistent but less precise. A classical error in x makes the regressor correlated with the composite error and **attenuates** the coefficient toward zero.

Analysis C4 · The square metres in the ad

Land-registry area vs. area declared in the ad

$\mathrm{sd}(m2_anuncio - m2) = 11.6\ \mathrm{m}^2$ $\quad$ $\mathrm{corr}(m2, m2_anuncio) = 0.93$

Regressor	Elasticity of price w.r.t. area
$\log(m2)$, land registry	0.900
$\log(m2_anuncio)$, declared	0.712
Ratio	0.791
Theoretical attenuation factor	0.779

Analysis C4 · The square metres in the ad

Land-registry area vs. area declared in the ad
$\mathrm{sd}(m2_anuncio - m2) = 11.6\ \mathrm{m}^2$ $\quad \mathrm{corr}(m2, m2_anuncio) = 0.93$

Regressor	Elasticity of price w.r.t. area
$\log(m2)$, land registry	0.900
$\log(m2_anuncio)$, declared	0.712
Ratio	0.791
Theoretical attenuation factor	0.779

- Classical error in a regressor biases its coefficient **towards zero**. The theory predicts the size of the shrinkage (the residual variance ratio, given the other regressors), and it is close.
- A correlation of 0.93 sounds like "basically the same variable". It still costs 21% of the elasticity.

The noisier a feature, the smaller its coefficient, so a ranking of "feature importance" is partly a ranking of *measurement quality*. Before trusting it, ask how each variable was measured.

Summary of measurement error

1. Error in y uncorrelated with the regressors: OLS remains consistent, only less precise.
2. Error in x_k uncorrelated with the *observed* x_k: harmless for consistency.
3. Classical errors-in-variables (error uncorrelated with x_k^*): **attenuation bias**, $|plim\,\widehat{\beta}_k| \leq |\beta_k|$.
4. Attenuation is worse the more x_k^* is explained by the other regressors.
5. When the error is related to the true value (e.g. non-smokers always report zero), the classical results do not apply.

Simultaneity

- A regressor is **simultaneous** with y when both are determined in the same equilibrium, so the regressor responds to the shocks that move y.
- Supply and demand (intercepts omitted, u_1 and u_2 uncorrelated):

$$q = \alpha_1 p + u_1 \ \text{(supply)}, \qquad q = \beta_1 p + u_2 \ \text{(demand)} \quad \Rightarrow \quad p = \frac{u_2 - u_1}{\alpha_1 - \beta_1}.$$

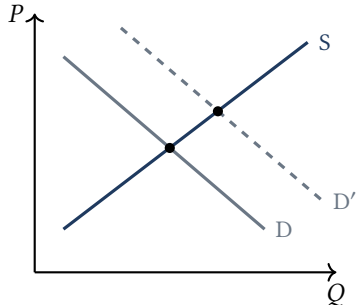
  The equilibrium price depends on both shocks, so $cov(p, u_1) \neq 0$ and $cov(p, u_2) \neq 0$: p is endogenous in *both* equations, and OLS of q on p recovers neither slope.
- Nothing is omitted and nothing is mismeasured, so neither a proxy nor extra controls help. Identifying the supply curve needs a variable that shifts demand but not supply: an **instrument**.
- That is why simultaneity is treated in **Unit 3**. Instrumental variables address all three sources of endogeneity: omitted variables, measurement error and simultaneity.

Simultaneity (price and quantity)

Setup.

- Model: $Q_i = \beta_0 + \beta_1 Price_i + \varepsilon_i$
- Price and quantity are determined together: when demand rises, both increase.
- $\Rightarrow \text{Cov}(Price_i, \varepsilon_i) \neq 0$

- OLS cannot separate supply from demand.
- $\widehat{\beta_1}$ is **biased and inconsistent**.
- Solutions: instrumental variables or structural modelling.



we only observe equilibrium points

Which channel?

Exercise. For each case, which source of endogeneity is most likely?

1. Regressing city crime rates on the number of police officers.
2. Using self-reported weekly hours worked as a regressor.
3. Evaluating a training programme in which the most motivated workers enrol.

Which channel?

Exercise. For each case, which source of endogeneity is most likely?

1. Regressing city crime rates on the number of police officers.
2. Using self-reported weekly hours worked as a regressor.
3. Evaluating a training programme in which the most motivated workers enrol.

Answer.

1. **Simultaneity**: cities with more crime hire more police.
2. **Measurement error**: hours are misreported.
3. **Omitted variable**: motivation affects both enrolment and outcomes.

Block C in R

```
# The oracle (never available in real life)
o <- read.csv("data/idealista_oraculo_clase.csv")
d <- merge(d, o, by = "id")

# C1: short, long and auxiliary regressions
lm(f_log, d)
lm(update(f_log, . ~ . + calidad), d)
lm(update(f_log, calidad ~ .), d)

# C2: proxy
lm(update(f_log, . ~ . + num_fotos), d)

# C3: bad control
mc <- lm(update(f_log, . ~ . + cert_energ), d)
deltaMethod(mc, "40*antig + 2400*`I(antig^2)`")

# C4: measurement error
d$lm2_anuncio <- log(d$m2_anuncio)
lm(update(f_log, . ~ . - lm2 + lm2_anuncio), d)
```

1. Verify $\widehat{\gamma\delta}$ = short − long to the fourth decimal.
2. Which way is the bias on `m2_terraza`? Explain it with the formula.
3. Use *both* `num_fotos` and `cert_energ`. Which question does that model answer?
4. Error in the *price* instead: what would it do?