Unit 2. The single-equation linear model

Block B. What does a coefficient mean?

Instructor: Henry Redondo

Dpto. de Fundamentos del Análisis Económico
Universidad de Alicante

Econometrics II · Academic Year 2026/2027

Outline of Block B

1. B. What does a coefficient mean?
   - Levels, logs and quadratics
   - Dummy variables

A stronger assumption: zero conditional mean

- So far, we have used $OLS1$, $E(\mathbf{x}u) = 0$: the error is orthogonal to each regressor. A stronger assumption is

$$\boxed{E(u \mid \mathbf{x}) = 0} \qquad OLS1'$$

  The expected error is zero for every value of $\mathbf{x}$.
- $OLS1' \Rightarrow OLS1$, by the law of iterated expectations:

$$E(\mathbf{x}u) = E\big[\mathbf{x}E(u \mid \mathbf{x})\big] = 0.$$

- What each assumption implies:
  - Under $OLS1$, $\mathbf{x}'\boldsymbol{\beta}$ is the linear projection of y on $\mathbf{x}$: the best *linear* predictor.
  - Under $OLS1'$, $\mathbf{x}'\boldsymbol{\beta} = E(y \mid \mathbf{x})$: the linear model is the *conditional expectation*, and each β_j measures a partial effect on the conditional mean.

*OLS*1 and *OLS*1′ in words

- Both are assumptions about the unobserved **error** u: location, quality, views, everything that moves the price apart from the regressors. They are not statements about the residuals.
- *OLS*1′, $E(u \mid \mathbf{x}) = 0$: **at every value of x**, the omitted factors average zero. Among all the 60 m^2 flats they average zero, and among all the 120 m^2 flats too. In a scatter plot, every vertical slice of the cloud is centred on the regression line.
  - It implies that u is uncorrelated with *any* function of $\mathbf{x}$: x, x^2, $\log x$, …
  - It implies that OLS is unbiased (Block C).
- *OLS*1, $E(\mathbf{x}u) = 0$: **on average over the population**, the omitted factors do not move *linearly* with $\mathbf{x}$. Since $\mathbf{x}$ contains a constant, it means $E(u) = 0$ and $cov(x_j, u) = 0$ for each j. The line is not tilted relative to the cloud, but single slices can be off centre in a nonlinear pattern.
  - It is enough for $\mathbf{x}'\boldsymbol{\beta}$ to be the best linear predictor and for OLS to be consistent (Block C).
  - Its sample analogue, $\frac{1}{N}\sum_i \mathbf{x}_i \widehat{u}_i = 0$, is the first-order condition of OLS: method of moments.

*OLS*1 does not imply *OLS*1′

- **A textbook case.** Let $x \sim N(0,1)$ and $u = x^2 - 1$. Then $E(u) = 0$ and

$$E(xu) = E(x^3) - E(x) = 0, \qquad \text{but} \qquad E(u \mid x) = x^2 - 1 \neq 0.$$

- **In the listings.** Suppose the price is convex in `m2`, with small and very large flats both dearer per m^2, and we fit a straight line. The curvature goes into u:
  - $E(m2 \cdot u) = 0$ holds, because the best straight line is uncorrelated with what it misses;
  - but $E(u \mid m2) > 0$ for small and large flats and < 0 for medium ones.
- **When both fail.** Larger flats tend to be in better neighbourhoods, and neighbourhood quality is in u. Then $cov(m2, u) > 0$ and the 2,206 € of Analysis A1 mixes the value of space with the value of location. More data do not help: this is the omitted variable bias of Block C.

Remember. $\sum_i x_{ji}\widehat{u}_i = 0$ holds in every sample because OLS imposes it (property 1). $E(\mathbf{x}u) = 0$ is an assumption about the population, and the data alone cannot check it.

Interpreting the coefficients

- Under $OLS1'$, $E(u \mid \mathbf{x}) = 0$, the linear model is the conditional expectation

$$E(y|x_1, .., x_k) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$$

  and each β_j is the partial effect of x_j on $E(y|x_1, .., x_k)$.
- Then, the $\widehat{\beta}_j$ are interpreted as estimated partial effects (or estimated marginal effects). That is $\widehat{\beta}_j$ is the estimated change in $E(y|x_1, .., x_k)$ when x_j increases by one unit holding the other regressors fixed.
- Under $OLS1$ only, the $\widehat{\beta}_j$ still estimate the coefficients of the linear projection of y on $\mathbf{x}$: the best linear approximation to $E(y \mid \mathbf{x})$, which need not have these partial effects.

Analysis B1 · Coefficients in euros

The hedonic model in levels: price in euros on area, terrace, age, floor, zone and year. For now we read the two *continuous* regressors; the terrace and ground-floor dummies come with the dummy variables.

Levels: `lm(f_niv)`, robust s.e.

	Coef.	Robust s.e.
`m2`	1,615.76	71.93
`m2_terraza`	1,475.43	431.08

Exercise. How would you explain these two numbers to the pricing team?

Analysis B1 · Coefficients in euros

The hedonic model in levels: price in euros on area, terrace, age, floor, zone and year. For now we read the two *continuous* regressors; the terrace and ground-floor dummies come with the dummy variables.

Levels: `lm(f_niv)`, *robust s.e.*

	Coef.	Robust s.e.
`m2`	1,615.76	71.93
`m2_terraza`	1,475.43	431.08

Exercise. How would you explain these two numbers to the pricing team?

- Holding the other characteristics fixed, one more m^2 of floor area is associated with an asking price **1,615.76 €** higher on average.
- For a flat with a terrace, one more m^2 of terrace is associated with **1,475.43 €** more, holding the interior area and everything else fixed.

Quadratic terms

- The multiple regression model allows us to include in the model some functions of the same variable. For instance, a model can include a variable and its square

$$E(y|x) = \beta_0 + \beta_1 x + \beta_2 x^2$$

- What is the partial effect of x on $E(y|x)$ in this model?

Quadratic terms

- The multiple regression model allows us to include in the model some functions of the same variable. For instance, a model can include a variable and its square

$$E(y|x) = \beta_0 + \beta_1 x + \beta_2 x^2$$

- What is the partial effect of x on $E(y|x)$ in this model?
  - Clearly β_1 does not capture the partial effect, since β_1 would be the change in $E(y|x)$ given an increase of one unit in x, holding x^2 fixed, and obviously if we increase x by one unit we cannot hold constant its square.
  - In order to compute the partial effect we must compute the derivative

$$\frac{\partial E(y|x)}{\partial x} = \beta_1 + 2\beta_2 x$$

    and therefore, the partial effect (or marginal effect) of x on $E(y|x)$ is $\beta_1 + 2\beta_2 x$.
  - Note that in this model, the partial effect is not constant, it can be increasing or decreasing, depending on the sign of β_2. It changes sign at $x^* = -\beta_1/(2\beta_2)$.

A model with x^2 is still a linear model

- "Linear" means **linear in the parameters** $\boldsymbol{\beta}$, not in x. Take

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u, \qquad E(u \mid x) = 0.$$

- Conditioning on x, both x and x^2 are known numbers:

$$E(y \mid x) = \beta_0 + \beta_1 \underbrace{E(x \mid x)}_{=x} + \beta_2 \underbrace{E(x^2 \mid x)}_{=x^2} + \underbrace{E(u \mid x)}_{=0} = \beta_0 + \beta_1 x + \beta_2 x^2.$$

  If a flat has 80 m^2 we also know it has $80^2 = 6{,}400$: the square adds no new randomness once x is known. Conditioning on x is the same as conditioning on (x, x^2), so $E(u \mid x, x^2) = 0$ holds automatically.
- By the law of iterated expectations, with $\mathbf{z} = (1, x, x^2)'$,

$$E(\mathbf{z}u) = E\big[\mathbf{z}\,E(u \mid x)\big] = 0,$$

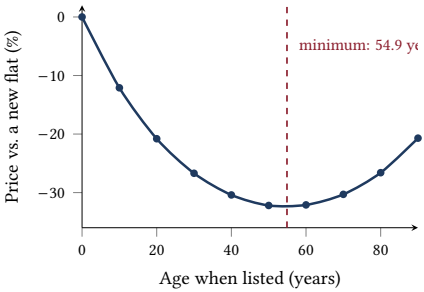
  so $y = \mathbf{z}'\boldsymbol{\beta} + u$ satisfies $OLS1$ with $\mathbf{z}$ in place of $\mathbf{x}$. The formula $\widehat{\boldsymbol{\beta}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}$, the algebraic properties and all the results of Block C apply unchanged. OLS does not know that one column is the square of another.

Linear and nonlinear in the parameters

Model	Linear in $\boldsymbol{\beta}$?	OLS?
$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u$	yes	yes
$\log y = \beta_0 + \beta_1 \log x + u$	yes	yes
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1 x_2 + u$	yes	yes
$y = \beta_0 e^{\beta_1 x} + u$	no	no, nonlinear least squares
$y = \beta_0 + x^{\beta_1} + u$	no	no

- The test: can the model be written as a sum of (known function of the data) $\times$ (parameter)? The first-order conditions $E[\mathbf{z}(y - \mathbf{z}'\boldsymbol{\beta})] = 0$ are then linear in $\boldsymbol{\beta}$ and are solved with one matrix inverse.
- x and x^2 are correlated, but x^2 is not an exact *linear* combination of 1 and x, so *OLS*2 holds (unless x takes only two values in the sample).
- A nonlinear relationship between y and x is a modelling choice inside the linear model. Adding x^2 moves the curvature from u into $\mathbf{z}'\boldsymbol{\beta}$, so $E(u \mid x) = 0$ can hold again.

Analysis B2 · Does an old building lose value forever?

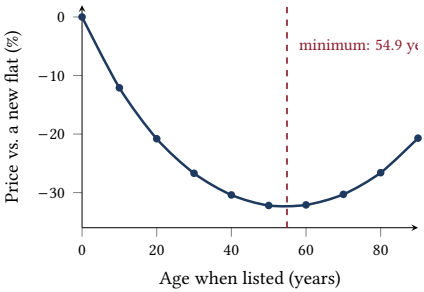


From the log model

$\widehat{\beta}_{antig} = -0.0142,\ \widehat{\beta}_{antig^2} = 0.0001297$

$$\frac{\partial E(\log precio \mid \mathbf{x})}{\partial\, antig} = \widehat{\beta}_{antig} + 2\widehat{\beta}_{antig^2}\, antig$$

Analysis B2 · Does an old building lose value forever?



From the log model

$\widehat{\beta}_{antig} = -0.0142,\ \widehat{\beta}_{antig^2} = 0.0001297$

$$\frac{\partial E(\log precio \mid \mathbf{x})}{\partial\, antig} = \widehat{\beta}_{antig} + 2\widehat{\beta}_{antig^2}\, antig$$

- Effect of one more year: −1.30% at 5 years, −0.78% at 25, −0.23% at 46, +0.391% at 70.
- The effect changes sign at $antig^* = -\widehat{\beta}_{antig}/(2\widehat{\beta}_{antig^2}) = 54.91$ years.
- Built around 1960–70, in the *desarrollismo* years: the least valued stock. Older buildings in Centro and Benalúa recover.

Is 54.9 years a fact about *age*, or about what comes with old buildings: poor insulation, worse finishes? Block C returns to this.

Logarithms

- Taking logs changes the units in which a coefficient is read: from units of y to percentages.
- **Log–log**, $\log y = \beta_0 + \beta_1 \log x + \ldots$: β_1 is an **elasticity**. A 1% increase in x changes y by about β_1%.
- **Log–level**, $\log y = \beta_0 + \beta_1 x + \ldots$: $100\beta_1$ is a **semi-elasticity**. One more unit of x changes y by about $100\beta_1$%.
- **Level–log**, $y = \beta_0 + \beta_1 \log x + \ldots$: a 1% increase in x changes y by about $\beta_1/100$ units.
- These approximations use $\log(1 + a) \approx a$, so they are good for small changes. In a log–level model the exact effect of a change Δx is

$$\boxed{100\left(e^{\beta_1 \Delta x} - 1\right)\%}$$

  which matters when $\beta_1 \Delta x$ is not small. We will need it for dummy variables, where Δx is the jump from 0 to 1.

Analysis B1 (cont.) · Coefficients in percentages

The same model with $\log(precio)$ on the left and $\log(m2)$ on the right:

Logs: `lm(f_log)`, *robust s.e.*

	Coef.	Robust s.e.
`lm2`	0.8997	0.0186
`m2_terraza`	0.0053	0.0014

Exercise. What does 0.8997 on `lm2` mean? And 0.0053 on `m2_terraza`?

Analysis B1 (cont.) · Coefficients in percentages

The same model with $\log(precio)$ on the left and $\log(m2)$ on the right:

Logs: `lm(f_log)`, *robust s.e.*

	Coef.	Robust s.e.
`lm2`	0.8997	0.0186
`m2_terraza`	0.0053	0.0014

Exercise. What does 0.8997 on `lm2` mean? And 0.0053 on `m2_terraza`?

- **Elasticity (log–log):** holding its other characteristics fixed, a 1% larger flat is listed about 0.90% higher; 10% more area $\Rightarrow 1.1^{0.8997} - 1 = 8.95\%$.
- **Semi-elasticity (log–level):** each extra m^2 of terrace raises the price by about $100 \times 0.0053 = 0.53\%$.

Analysis B1 (cont.) · Coefficients in percentages

The same model with $\log(precio)$ on the left and $\log(m2)$ on the right:

Logs: `lm(f_log)`, robust s.e.

	Coef.	Robust s.e.
`lm2`	0.8997	0.0186
`m2_terraza`	0.0053	0.0014

Exercise. What does 0.8997 on `lm2` mean? And 0.0053 on `m2_terraza`?

- **Elasticity (log–log):** holding its other characteristics fixed, a 1% larger flat is listed about 0.90% higher; 10% more area $\Rightarrow 1.1^{0.8997} - 1 = 8.95\%$.
- **Semi-elasticity (log–level):** each extra m^2 of terrace raises the price by about $100 \times 0.0053 = 0.53\%$.

In levels, one more m^2 is worth 1,615.76 € in every flat, at the beach or in San Blas. In logs it is worth about 0.9% of the price per 1% of area, so its value in euros grows with the price of the flat. The log model also fits better ($R^2 = 0.900$).

Reading coefficients

Dependent	Regressor	Δx_j	Change in $E(y \mid \mathbf{x})$
y	x_j	+1 unit	$+\beta_j$ units
$\log y$	$\log x_j$	+1%	$\approx +\beta_j\%$
$\log y$	x_j	+1 unit	$\approx +100\beta_j\%$
y	$\log x_j$	+1%	$\approx +\beta_j/100$ units
y	$x_j,\ x_j^2$	+1 unit	$\approx \beta_1 + 2\beta_2 x_j$ units

Dummy variables

- In many empirical studies we want to control for qualitative factors: the zone of the flat, whether it has a terrace, whether it is on the ground floor.
- Qualitative factors are represented by binary variables that take the values zero and one, called **dummy variables**.
- Consider the model

$$\log(precio) = \beta_0 + \delta_0\, bajo + \beta_1 \log(m2) + u, \qquad E(u \mid bajo, m2) = 0 \qquad (15)$$

where $bajo = 1$ if the flat is on the ground floor and 0 otherwise.

Dummy variables

- In many empirical studies we want to control for qualitative factors: the zone of the flat, whether it has a terrace, whether it is on the ground floor.
- Qualitative factors are represented by binary variables that take the values zero and one, called **dummy variables**.
- Consider the model

$$\log(precio) = \beta_0 + \delta_0\, bajo + \beta_1 \log(m2) + u, \qquad E(u \mid bajo, m2) = 0 \qquad (15)$$

  where $bajo = 1$ if the flat is on the ground floor and 0 otherwise.
- Then

$$E(\log(precio) \mid m2, bajo = 1) = \beta_0 + \delta_0 + \beta_1 \log(m2)$$
$$E(\log(precio) \mid m2, bajo = 0) = \beta_0 + \beta_1 \log(m2)$$

  so δ_0 is an **intercept shift**: the difference in expected log price between ground-floor flats and other flats with the same area.
- $100\delta_0$ is approximately the percentage difference; the exact one is $100\ (e^{\delta_0} - 1)\%$.

Which category gets the dummy?

- We can alternatively write the model with $alto = 1 - bajo$, equal to 1 if the flat is *not* on the ground floor:

$$\log(precio) = \gamma_0 + \gamma_1\, alto + \gamma_2 \log(m2) + u, \qquad E(u \mid alto, m2) = 0 \tag{16}$$

Which category gets the dummy?

- We can alternatively write the model with $alto = 1 - bajo$, equal to 1 if the flat is *not* on the ground floor:

$$\log(precio) = \gamma_0 + \gamma_1\, alto + \gamma_2 \log(m2) + u, \qquad E(u \mid alto, m2) = 0 \qquad (16)$$

- Then

$$E(\log(precio) \mid m2, alto = 0) = \gamma_0 + \gamma_2 \log(m2)$$
$$E(\log(precio) \mid m2, alto = 1) = \gamma_0 + \gamma_1 + \gamma_2 \log(m2)$$

  and $100\gamma_1$ is the approximate percentage difference between upper-floor and ground-floor flats with the same area.
- Models (15) and (16) are the same model with different parameters:

$$\begin{array}{l} \beta_0 = \gamma_0 + \gamma_1 \\ \beta_0 + \delta_0 = \gamma_0 \\ \beta_1 = \gamma_2 \end{array} \quad \Rightarrow \quad \begin{array}{l} \beta_0 = \gamma_0 + \gamma_1 \\ \delta_0 = -\gamma_1 \\ \beta_1 = \gamma_2 \end{array}$$

The dummy variable trap

- We can also include both dummies and drop the intercept:

$$\log(precio) = \gamma_0\, alto + \beta_0\, bajo + \beta_1 \log(m2) + u \tag{17}$$

The dummy variable trap

- We can also include both dummies and drop the intercept:

$$\log(precio) = \gamma_0\, alto + \beta_0\, bajo + \beta_1 \log(m2) + u \tag{17}$$

- Then

$$E(\log(precio) \mid m2, bajo = 0) = \gamma_0 + \beta_1 \log(m2)$$
$$E(\log(precio) \mid m2, bajo = 1) = \beta_0 + \beta_1 \log(m2)$$

  Each coefficient is that group's own intercept, and $100(\beta_0 - \gamma_0)$ is the approximate percentage difference between ground-floor and upper-floor flats with the same area.
- The OLS estimates and their standard errors satisfy the same equivalences as the parameters of the three models.
- Can we include $bajo$, $alto$ *and* an intercept?

The dummy variable trap

- We can also include both dummies and drop the intercept:

$$\log(precio) = \gamma_0\, alto + \beta_0\, bajo + \beta_1 \log(m2) + u \tag{17}$$

- Then

$$E(\log(precio) \mid m2, bajo = 0) = \gamma_0 + \beta_1 \log(m2)$$
$$E(\log(precio) \mid m2, bajo = 1) = \beta_0 + \beta_1 \log(m2)$$

  Each coefficient is that group's own intercept, and $100(\beta_0 - \gamma_0)$ is the approximate percentage difference between ground-floor and upper-floor flats with the same area.
- The OLS estimates and their standard errors satisfy the same equivalences as the parameters of the three models.
- Can we include $bajo$, $alto$ *and* an intercept? No. $bajo + alto = 1$ for every flat: the two dummies add up to the constant. There is exact multicollinearity, *OLS*2 fails, $\mathbf{X'X}$ is singular and the OLS estimator does not exist. R silently drops one of them (its coefficient is reported as NA).

Analysis B3 · The ground-floor discount

Coefficient on `bajo` in the hedonic model, robust s.e.

	Coef.	Robust s.e.
Levels, `lm(f_niv)`	−22,357.73	3,317.22
Logs, `lm(f_log)`	−0.1210	0.0157

Exercise. How much cheaper is a ground-floor flat, in euros and in per cent? What would the coefficient be if we used *alto* instead of *bajo*?

Analysis B3 · The ground-floor discount

Coefficient on `bajo` *in the hedonic model, robust s.e.*

	Coef.	Robust s.e.
Levels, `lm(f_niv)`	−22,357.73	3,317.22
Logs, `lm(f_log)`	−0.1210	0.0157

Exercise. How much cheaper is a ground-floor flat, in euros and in per cent? What would the coefficient be if we used *alto* instead of *bajo*?

- **Levels:** a ground-floor flat is listed, on average, **22,357.73 €** below an otherwise identical flat on an upper floor.
- **Logs:** $100(e^{-0.1210} - 1) = -11.4\%$: a ground-floor flat is listed 11.4% below an otherwise identical flat upstairs: noise, security, light.
- With *alto* the coefficient is +0.1210, with the same standard error 0.0157. Upper floors are listed $100(e^{0.1210} - 1) = 12.9\%$ above ground floors: exact percentages are not symmetric.

Several categories

- Suppose we want to control for the zone. With $J = 5$ zones we define five dummies

$$Z_j = \begin{cases} 1 \text{ if the flat is in zone } j \\ 0 \text{ otherwise} \end{cases} \qquad j = 1, \ldots, 5$$

  (1 Benalúa, 2 Centro, 3 Carolinas, 4 San Blas, 5 Playa de San Juan).
- We include $J - 1$ of these dummies in the model:

$$\log(precio) = \beta_0 + \beta_2 Z_2 + \beta_3 Z_3 + \beta_4 Z_4 + \beta_5 Z_5 + \beta_6 \log(m2) + \cdots + u \qquad (18)$$

Several categories

- Suppose we want to control for the zone. With $J = 5$ zones we define five dummies

$$Z_j = \begin{cases} 1 \text{ if the flat is in zone } j \\ 0 \text{ otherwise} \end{cases} \qquad j = 1, \ldots, 5$$

  (1 Benalúa, 2 Centro, 3 Carolinas, 4 San Blas, 5 Playa de San Juan).
- We include $J - 1$ of these dummies in the model:

$$\log(precio) = \beta_0 + \beta_2 Z_2 + \beta_3 Z_3 + \beta_4 Z_4 + \beta_5 Z_5 + \beta_6 \log(m2) + \cdots + u \qquad (18)$$

- $100\beta_j$ measures, approximately, the percentage difference in price between a flat in zone j and an otherwise identical flat in the omitted zone, the **base group** (here Benalúa).
- We cannot include the five dummies and an intercept: $Z_1 + \cdots + Z_5 = 1$, exact multicollinearity again. `factor(zona)` creates the $J - 1$ dummies for you, leaving out the first level.

Analysis B4 · Five zones, base group Benalúa

Base group: Benalúa

Zone	Coef.	Robust s.e.	Exact effect
Centro	0.0559	0.0146	+5.8%
Carolinas	−0.1193	0.0145	−11.2%
San Blas	−0.1747	0.0162	−16.0%
Playa de San Juan	0.3172	0.0237	+37.3%

Exercise. Beach flats are listed at 295,367 € on average and flats in Benalúa at 142,884 €, more than double. Why is the estimated premium only 37.3%?

Analysis B4 · Five zones, base group Benalúa

Base group: Benalúa

Zone	Coef.	Robust s.e.	Exact effect
Centro	0.0559	0.0146	+5.8%
Carolinas	−0.1193	0.0145	−11.2%
San Blas	−0.1747	0.0162	−16.0%
Playa de San Juan	0.3172	0.0237	+37.3%

Exercise. Beach flats are listed at 295,367 € on average and flats in Benalúa at 142,884 €, more than double. Why is the estimated premium only 37.3%?

"+37.3%" means *relative to an identical flat in Benalúa*: same area, age, floor, terrace and listing year. The difference between the average prices of the two zones also reflects that beach flats are bigger and newer.

Changing the base group

- Instead of the dummy for Benalúa we could exclude any other zone dummy, and the model would be equivalent to (18).
- If the excluded dummy is zone s, the coefficient of zone j (times 100) measures the percentage difference between zone j and zone s for otherwise identical flats.
- With any base, the difference between zones j and s is $\beta_j - \beta_s$. To obtain its standard error directly, re-estimate the model with s as the base group.

Changing the base group

- Instead of the dummy for Benalúa we could exclude any other zone dummy, and the model would be equivalent to (18).
- If the excluded dummy is zone s, the coefficient of zone j (times 100) measures the percentage difference between zone j and zone s for otherwise identical flats.
- With any base, the difference between zones j and s is $\beta_j - \beta_s$. To obtain its standard error directly, re-estimate the model with s as the base group.
- Changing the base changes every zone coefficient, but not the fitted values, the residuals, R^2, or any difference between two zones.

Practical rule. Choose as base the group your audience compares against: the largest zone, the reference market, or "the average flat". Then everyone reads the table in one pass.

Analysis B5 · Changing the base group

`relevel(zona, "Centro")`: base group Centro

Playa de San Juan vs. Centro: the coefficient of Playa de San Juan is now 0.26122.

Analysis B5 · Changing the base group

`relevel(zona, "Centro")`: *base group Centro*

Playa de San Juan vs. Centro: the coefficient of Playa de San Juan is now 0.26122.

$$\underbrace{0.26122}_{\text{base Centro}} = \underbrace{0.31716 - 0.05593}_{\text{base Benalúa}}$$

Same answer from the difference of the two coefficients in the Benalúa-base model, with the same robust s.e. 0.0240 ($t = 10.86$; tests come in Block D).

Two qualitative factors

- Consider now a model with two qualitative factors:

$$\log(precio) = \beta_0 + \delta_1\, terraza + \delta_2\, bajo + \beta_1 \log(m2) + u \tag{19}$$

- $100\delta_1$ is the approximate percentage difference between flats with and without a terrace, with the same area and on the same type of floor.
- $100\delta_2$ is the approximate percentage difference between ground-floor and upper-floor flats, with the same area and the same terrace status.

Two qualitative factors

- Consider now a model with two qualitative factors:

$$\log(precio) = \beta_0 + \delta_1\, terraza + \delta_2\, bajo + \beta_1 \log(m2) + u \tag{19}$$

- $100\delta_1$ is the approximate percentage difference between flats with and without a terrace, with the same area and on the same type of floor.
- $100\delta_2$ is the approximate percentage difference between ground-floor and upper-floor flats, with the same area and the same terrace status.
- Four groups, three parameters (intercepts, for a given area):

	$bajo = 0$	$bajo = 1$
$terraza = 0$	β_0	$\beta_0 + \delta_2$
$terraza = 1$	$\beta_0 + \delta_1$	$\beta_0 + \delta_1 + \delta_2$

- The model *imposes* that the terrace is worth the same on the ground floor as upstairs. That may be false: a ground-floor "terrace" is often a patio. To let it differ we need an **interaction**.

Analysis B6 · The terrace premium

The hedonic model contains both *terraza* (a dummy) and *m2_terraza* (the terrace size, zero without a terrace).

Terrace coefficients in the hedonic model, robust s.e.

	Levels, `lm(f_niv)`		Logs, `lm(f_log)`	
	Coef.	Robust s.e.	Coef.	Robust s.e.
`terraza`	9,008.58	5,273.38	0.1041	0.0189
`m2_terraza`	1,475.43	431.08	0.0053	0.0014

Exercise. How much is having a terrace worth, in euros and in per cent?

Analysis B6 · The terrace premium

The hedonic model contains both *terraza* (a dummy) and *m2_terraza* (the terrace size, zero without a terrace).

Terrace coefficients in the hedonic model, robust s.e.

	Levels, `lm(f_niv)`		Logs, `lm(f_log)`	
	Coef.	Robust s.e.	Coef.	Robust s.e.
`terraza`	9,008.58	5,273.38	0.1041	0.0189
`m2_terraza`	1,475.43	431.08	0.0053	0.0014

Exercise. How much is having a terrace worth, in euros and in per cent?

- **Levels:** a terrace adds 9,008.58 €, plus 1,475.43 € for each square metre of terrace.
- **Logs:** having a terrace raises the price by $e^{0.1041} - 1 = 11.0\%$, plus about 0.53% per terrace m^2.

Analysis B6 · The terrace premium

The hedonic model contains both *terraza* (a dummy) and *m2_terraza* (the terrace size, zero without a terrace).

Terrace coefficients in the hedonic model, robust s.e.

	Levels, `lm(f_niv)`		Logs, `lm(f_log)`	
	Coef.	Robust s.e.	Coef.	Robust s.e.
`terraza`	9,008.58	5,273.38	0.1041	0.0189
`m2_terraza`	1,475.43	431.08	0.0053	0.0014

Exercise. How much is having a terrace worth, in euros and in per cent?

- **Levels:** a terrace adds 9,008.58 €, plus 1,475.43 € for each square metre of terrace.
- **Logs:** having a terrace raises the price by $e^{0.1041} - 1 = 11.0\%$, plus about 0.53% per terrace m^2.

Percentages. "A terrace adds 9,009 €" is wrong for a 400,000 € flat on the beach and for a 70,000 € flat in San Blas; "it adds about 11%" travels much better.

Analysis B6 (cont.) · Terrace and ground floor together

Two dummies, no interaction, logs

Relative to “no terrace, not ground floor”:

- terrace: +0.1041
- ground floor: −0.1210
- terrace *and* ground floor:

Analysis B6 (cont.) · Terrace and ground floor together

Two dummies, no interaction, logs

Relative to "no terrace, not ground floor":

- terrace: +0.1041
- ground floor: −0.1210
- terrace *and* ground floor: $\widehat{\delta_1} + \widehat{\delta_2} = -0.0168$

A ground-floor flat with a terrace is listed about as high as an upper-floor flat without one: the terrace roughly offsets the ground-floor discount.

Interactions between two dummies

- To let the terrace premium differ between the beach and the rest of the city, include the interaction of $terraza$ with $costa$ (= 1 in Playa de San Juan):

$$\log(precio) = \beta_0 + \delta_1\, terraza + \gamma_1\, costa + \delta_3\, (terraza \times costa) + \beta_1 \log(m2) + u \quad (20)$$

$terraza \times costa$ is equal to one only for beach flats with a terrace.

Interactions between two dummies

- To let the terrace premium differ between the beach and the rest of the city, include the interaction of $terraza$ with $costa$ (= 1 in Playa de San Juan):

$$\log(precio) = \beta_0 + \delta_1\, terraza + \gamma_1\, costa + \delta_3\, (terraza \times costa) + \beta_1 \log(m2) + u \quad (20)$$

  $terraza \times costa$ is equal to one only for beach flats with a terrace.
- Expected log prices of the four groups, for a given area:

$$\begin{aligned}
E(\log(precio) \mid terraza = 0, costa = 0) &= \beta_0 + \beta_1 \log(m2) \\
E(\log(precio) \mid terraza = 1, costa = 0) &= \beta_0 + \delta_1 + \beta_1 \log(m2) \\
E(\log(precio) \mid terraza = 0, costa = 1) &= \beta_0 + \gamma_1 + \beta_1 \log(m2) \\
E(\log(precio) \mid terraza = 1, costa = 1) &= \beta_0 + \delta_1 + \gamma_1 + \delta_3 + \beta_1 \log(m2)
\end{aligned}$$

- The terrace premium is $100\delta_1$ away from the beach and $100(\delta_1 + \delta_3)$ at the beach: δ_3 *is* the difference. Whether it is the same in both places is the test $H_0 : \delta_3 = 0$ (Block D).

In the hedonic model the zone dummies already contain $costa$ (it is Z_5), so only the interaction is added.

Analysis B7 · Is a terrace worth more at the beach?

```
update(f_log, .  ~ .  + terraza:costa)
```

	Coef.	Robust s.e.
terraza	0.1026	0.0188
terraza × costa	0.0502	0.0418

Analysis B7 · Is a terrace worth more at the beach?

update(f_log, . ~ . + terraza:costa)

	Coef.	Robust s.e.
terraza	0.1026	0.0188
terraza × costa	0.0502	0.0418

- Premium: **10.3%** elsewhere, **15.28%** at Playa de San Juan.

A beach terrace is plausibly worth more: sea air, views, outdoor living.

Interactions with a continuous variable

- In all the models so far the area elasticity is the same for every flat: on the beach and elsewhere, with or without a terrace.
- To let a slope vary with a qualitative factor, interact the regressor with the dummy:

$$\log(precio) = \beta_0 + \gamma_1\, costa + \beta_1 \log(m2) + \beta_2\, (\log(m2) \times costa) + u$$

Interactions with a continuous variable

- In all the models so far the area elasticity is the same for every flat: on the beach and elsewhere, with or without a terrace.
- To let a slope vary with a qualitative factor, interact the regressor with the dummy:

$$\log(precio) = \beta_0 + \gamma_1\, costa + \beta_1 \log(m2) + \beta_2\, (\log(m2) \times costa) + u$$

- Then

$$E(\log(precio) \mid m2, costa = 0) = \beta_0 + \beta_1 \log(m2)$$
$$E(\log(precio) \mid m2, costa = 1) = \beta_0 + \gamma_1 + (\beta_1 + \beta_2) \log(m2)$$

  and

$$\frac{\partial E(\log(precio) \mid m2, costa = 0)}{\partial \log(m2)} = \beta_1, \qquad \frac{\partial E(\log(precio) \mid m2, costa = 1)}{\partial \log(m2)} = \beta_1 + \beta_2$$

- β_2 is the difference in the area elasticity between the beach and the rest. On its own a dummy shifts the **intercept**; interacted with a regressor it shifts the **slope**.

Analysis B8 · Does the area elasticity differ at the beach?

```
update(f_log, .  ~ .  + lm2:costa)
```

$\widehat{\beta}_1 = 0.9081$ (0.0191), $\widehat{\beta}_2 = -0.0438$ (0.0563), $t = -0.78$.

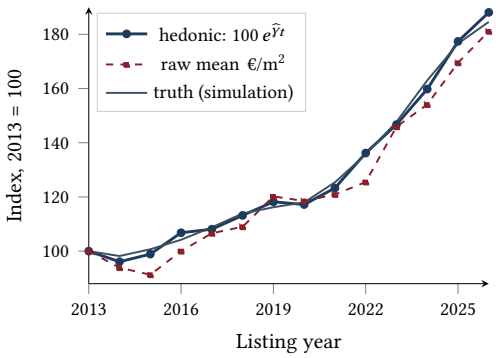
Analysis B8 · Does the area elasticity differ at the beach?

`update(f_log, . ~ . + lm2:costa)`

$\widehat{\beta}_1 = 0.9081$ (0.0191), $\widehat{\beta}_2 = -0.0438$ (0.0563), $t = -0.78$.

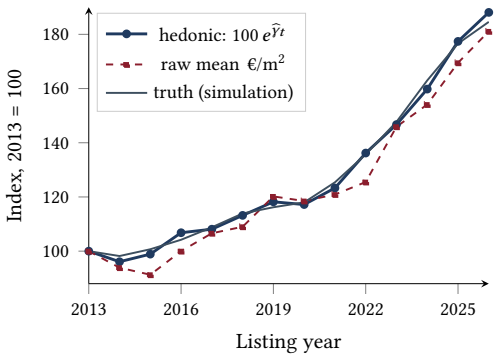
The interaction is small and less than one standard error away from zero (Block D turns this into a formal test). No evidence that the price of an extra square meter scales differently at the beach. Report it anyway: “we checked, and it is not there” is information too, and it protects against a colleague’s hunch becoming a pricing rule.

Analysis B9 · Year dummies are a house-price index



- `factor(anio)`: each year's coefficient is the price change of an *identical* flat relative to 2013, the base year.

Analysis B9 · Year dummies are a house-price index



- `factor(anio)`: each year's coefficient is the price change of an *identical* flat relative to 2013, the base year.
- The raw €/m² index mixes price changes with changes in *which* flats get listed. In 2022 it says +26%; the hedonic index says +36%.
- Average miss against the truth: 1.7 points (hedonic) vs. 5.0 (raw).

Portals and statistical agencies build house-price indices this way: a regression with time dummies, holding quality fixed.

Block B in R

```
library(sandwich); library(lmtest); library(car)
M <- lm(f_log, data = d); b <- coef(M)
coeftest(M, vcovHC(M, "HC1"))     # robust s.e.

# B2: turning point of the age profile
deltaMethod(M, "-antig / (2 * `I(antig^2)`)")
# B5: change the base group
lm(f_log, transform(d, zona = relevel(zona, "Centro")))
# B6: exact percentage for a dummy
100 * (exp(b["terraza"]) - 1)
# B7: terrace x beach
m7 <- lm(update(f_log, . ~ . + terraza:costa), d)
deltaMethod(m7, "terraza + `terraza:costa`")
# B9: hedonic index
100 * exp(b["anio2025"])
```

1. Use `relevel(zona, "Playa San Juan")`. Which coefficients change? Does R^2?
2. Put in *all five* zone dummies by hand plus a constant. What does R do?
3. Add `terraza:bajo`. Is a ground-floor terrace worth less?
4. Build the index by month with `factor(fecha)`. Is it too noisy?

The slides use `HC1` robust s.e. (the usual finite-sample correction); the default of `vcovHC` is `HC3`, which differs in the third decimal.