Unit 2. The single-equation linear model

Block A. What does OLS compute?

Instructor: Henry Redondo

Dpto. de Fundamentos del Análisis Económico
Universidad de Alicante

Econometrics II · Academic Year 2026/2027

Motivation

The pricing team of a property portal asks:
how much is a terrace worth in Alicante?
What number do we give them?

- Automated valuation, mortgage appraisal and portal price suggestions are all **linear regressions** underneath.
- Every block answers one question about that number.

Five questions, five blocks.

A. **What does OLS compute?** Fit, residuals, R^2.

B. **What does a coefficient mean?** Logs, quadratics, dummies, interactions.

C. **When is that meaning causal?** Consistency, robust standard errors, unbiasedness, Gauss–Markov.

D. **How sure are we?** t, F, Wald and LM tests.

E. **Is OLS the best we can do?** Heteroskedasticity and GLS.

1,000 listings in Alicante, 2013–2026

Variable	Meaning
`precio`	asking price (€)
`m2`	floor area, land registry (m^2)
`m2_anuncio`	floor area *declared in the ad*
`terraza`, `m2_terraza`	has a terrace; its size
`anio_construccion`, `antig`	year built; age when listed
`bajo`	ground floor
`zona`	5 zones: Benalúa, Centro, Carolinas, San Blas, Playa de San Juan
`fecha`, `anio`	month and year of listing
`cert_energ`	energy certificate (1 = G … 7 = A)
`num_fotos`	number of photos in the ad

What the data look like

- Median price **131,750 €**; mean 158,666 €.
- Mean area 91.5 m^2; 43.5% have a terrace (13.4 m^2 on average).
- Mean price: 108,589 € in San Blas, 295,367 € in Playa de San Juan.

The rules of the game

These data are simulated. Inspired by the Alicante market, but not real listings from Idealista or any other portal.

Because then we **know the truth**. We will estimate the terrace premium seven different ways, and at the end we can check which one is right. With real data that check is impossible, which is exactly why applied work needs the theory.

The Limits of What We Observe. One variable, the unobserved **quality** of the flat (finishes, light, renovation), is kept in a separate file, `idealista_oraculo.dta`. An analyst never has it. We open it only to see what we are missing.

Outline of Block A

1. A. What does OLS compute?
   - The model and the estimator
   - Algebraic properties and goodness of fit

The single-equation linear model

- Consider a population model that is linear in parameters

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u \tag{1}$$

  - $y, x_1, .., x_k$ are observable random scalars.
  - y is called the explained variable, dependent variable or response variable.
  - $x_1, .., x_k$ are called explanatory variables, independent variables, regressors, covariates or control variables.
  - u is an unobservable random disturbance called the error term.
  - $\beta_0, \beta_1, .., \beta_k$ are unknown parameters that we would like to estimate.
- We will assume that

$$E(u) = 0 \quad \text{and} \quad cov(x_j, u) = 0, \quad \text{for all } j = 1, .., k \tag{2}$$

- The assumption that $E(u) = 0$ is guaranteed provided the model includes a constant since β_0 is defined as

$$\beta_0 = E(y) - \beta_1 E(x_1) - ... - \beta_k E(x_k)$$

Assumptions *OLS*1 and *OLS*2

- Let

$$y = \mathbf{x}'\boldsymbol{\beta} + u$$

be the population linear model where $\mathbf{x} = (x_1, .., x_k)'$ is a random vector and $x_1 \equiv 1$, u is the unobservable random disturbance and $\boldsymbol{\beta} = (\beta_1, .., \beta_k)'$ is the unknown vector of parameters that we would like to estimate.

- We assume that

$$E(\mathbf{x}u) = 0 \qquad OLS1 \tag{3}$$

and

$$\text{rank } E(\mathbf{x}\mathbf{x}') = k \qquad OLS2 \tag{4}$$

- Notice that, since the first element of $\mathbf{x}$ is 1, assumption *OLS*1 is equivalent to saying that u has zero mean and it is uncorrelated with each regressor (assumption (2)).

Identification

- Assumption *OLS*2 requires that there is no perfect collinearity among the regressors. Since the first element of $\mathbf{x}$ is 1, this assumption is equivalent to stating that the variance–covariance matrix of the $k-1$ non-constant regressors is positive definite (i.e., full rank). Assumption *OLS*2 fails only if one regressor can be expressed as an exact linear combination of the others, meaning the model is improperly specified—for example, if it includes a full set of dummy variables for a categorical attribute.
- Under assumptions *OLS*1 and *OLS*2 the vector of parameters $\boldsymbol{\beta}$ is identified and is defined as

$$\boldsymbol{\beta} = E(\mathbf{x}\mathbf{x}')^{-1}E(\mathbf{x}y) \tag{5}$$

  Thus, under assumptions *OLS*1 and *OLS*2, $\boldsymbol{\beta}$ is the vector of coefficients of the *LP* of y on $\mathbf{x}$.

OLS as a method-of-moments estimator

- We assume that, to estimate $\boldsymbol{\beta}$, we can obtain a random sample of size N from the population. Thus $\{(\mathbf{x}_i', y_i)' : i = 1, .., N\}$ are iid random variables and we can write the model for each element i of the sample as

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i, \qquad i = 1, 2, ..., N \tag{6}$$

- The *OLS* estimator is the sample counterpart of (5)

$$\widehat{\boldsymbol{\beta}} = \left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i y_i\right) = \left(\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\left(\sum_{i=1}^{N}\mathbf{x}_i y_i\right) \tag{7}$$

- Notice that $\widehat{\boldsymbol{\beta}}$ is the method of moments estimator of $\boldsymbol{\beta}$ since $\widehat{\boldsymbol{\beta}}$ is obtained by replacing the population expectations by the corresponding sample means.

Matrix notation

- We can also write (6) using a compact notation as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_N' \end{bmatrix} \quad \mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{bmatrix}$$

$\mathbf{X}$ is an $N \times k$ matrix and $\mathbf{y}$ and $\mathbf{u}$ are $N \times 1$ vectors.

- Then, we can write the *OLS* estimator using matrix notation as:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{8}$$

- We will use the vector notation in (7) to prove the asymptotic properties of the *OLS* estimator and the matrix notation in (8) to prove the finite sample properties of the *OLS* estimator.

The simple regression model

- If we consider the simple linear model

$$y = \beta_0 + \beta_1 x + u \tag{9}$$

where x is a scalar.

- The OLS estimators of β_0 and β_1 are given by

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})\,(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} = \frac{S_{xy}}{S_x^2} \tag{10}$$

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x} \tag{11}$$

where $\overline{x}$ and $\overline{y}$ are the sample means, S_{xy} is the sample covariance between x and y and S_x^2 is the sample variance of x.

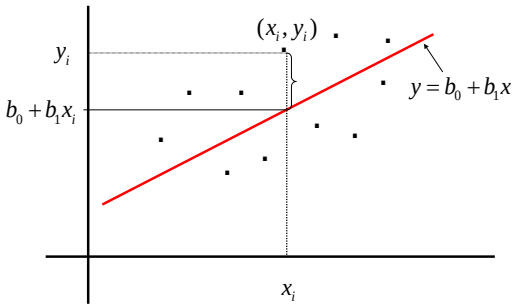
Why “least squares”?

- Now we are going to see the graphic interpretation of the OLS estimates of the simple regression model that justifies the name *least squares*.
- To do so, we draw a point cloud for a given sample of size N and some line

$$y = b_0 + b_1 x$$

- We will show that the OLS estimates defined in equations (11) and (10) are the "best" choice for the values of b_0 and b_1 if the goal is that the line is as "close" as possible to the points of the cloud according to a specified proximity criterion.
- Specifically, the proximity criterion which results in OLS estimates consists of minimizing the sum of the squares of the vertical distances of the points of the cloud to the regression line.

Least squares, graphically



The least squares criterion

- Graphically, we can see that the vertical distance from the point (x_i, y_i) to the line $y = b_0 + b_1 x$ is given by

$$y_i - b_0 - b_1 x_i$$

and therefore the objective function to be minimized is

$$s(b_0, b_1) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 \tag{12}$$

- The partial derivatives are:

$$\begin{aligned} \frac{\partial s(b_0, b_1)}{\partial b_0} &= -2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) \\ \frac{\partial s(b_0, b_1)}{\partial b_1} &= -2 \sum_{i=1}^{n} x_i (y_i - b_0 - b_1 x_i) \end{aligned}$$

First-order conditions

- The estimated coefficients are obtained by equating to zero the partial derivatives of the objective function

$$\sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) = 0$$

$$\sum_{i=1}^{n} x_i(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) = 0$$

- These two equations are called **first order conditions** for OLS estimates.
- The solution of this system of equations coincides with the OLS estimates of the simple linear model given in (10) and (11).

Fitted values and residuals

- The **fitted value** is defined for y when $x = x_i$ as

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$$

  This is the value we predict for y when $x = x_i$. Note that there is a fitted value for each observation in the sample.
  - In the multivariate regression model, fitted values are defined analogously

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{1i} + \cdots + \widehat{\beta}_k x_{ki}$$

- The **residual** for each observation in the sample is defined as the difference between the observed value y_i and the fitted value $\widehat{y}_i$

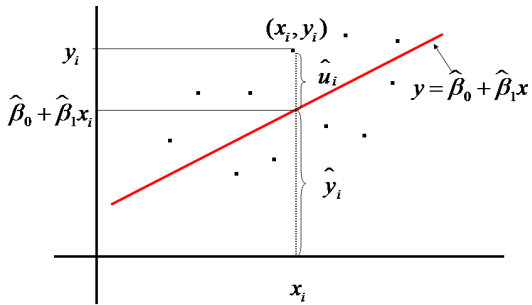
$$\widehat{u}_i = y_i - \widehat{y}_i$$

  and we have a residual for each of the observations in the sample.

Residuals as vertical distances

- Note that the residual of each observation is the vertical distance (with its corresponding sign) of the point to the regression line $y = \widehat{\beta}_0 + \widehat{\beta}_1 x$. Therefore, the OLS criterion is to minimize the sum of squared residuals. If the point lies above the regression line, the residual is positive; and if it lies below the line, the residual is negative.
  - In the multivariate regression model the OLS criterion is also to minimize the sum of squared residuals, and the residual of each observation is the vertical distance (with its corresponding sign) of the point to the regression hyperplane $y = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \cdots + \widehat{\beta}_k x_k$.

Residuals, graphically



Analysis A1 · Euros per square meter

```
lm(precio ~ m2)
```

	Coef.	Std. err.	t
m2	2,206.21	74.97	29.43
_cons	−43,151.50	7,210.13	−5.98
$N = 1{,}000$	$R^2 = 0.465$		

- **Slope:** one more square meter is associated with 2,206 € more on the asking price.
- **Intercept:** a flat with zero square meters would cost −43,151 €. Meaningless, because nobody lists a flat of 0 m^2. It only positions the line.
- Floor area alone accounts for 46% of the variation in prices. What is the other 54%?

Analysis A2 · The first-order conditions hold in the data

OLS chooses $\widehat{\boldsymbol{\beta}}$ so that

$$\sum_i \widehat{u}_i = 0, \qquad \sum_i x_{ij}\,\widehat{u}_i = 0.$$

After `lm(precio ~ m2)`

$\sum_i \widehat{u}_i$	2.9×10^{-8}
$\sum_i m2_i\,\widehat{u}_i$	2.2×10^{-6}
$\mathrm{corr}(\widehat{u}, m2)$	$\approx 10^{-16}$
mean of $\widehat{y}$	158,666.30
mean of y	158,666.30

These are zeros up to rounding error in the computer. Nothing here is *estimated*: they hold by construction in *every* sample, true model or not.

So "the residuals are uncorrelated with $m2$" tells you nothing about whether $m2$ is exogenous. It is an algebraic fact, not evidence. That question belongs to Block C.

Algebraic property 1 (residuals and regressors)

Property 1.
a. The sum, and therefore the sample average of the OLS residuals, is zero.
b. The sample covariance between the residuals and each of the regressors is zero.
Proof

- We have to show that

$$\sum_{i=1}^{N} \widehat{u}_i = 0 \quad and \quad \sum_{i=1}^{N} x_{ji}\widehat{u}_i = 0, \quad j = 2, .., k \tag{13}$$

which can be written in compact notation as:

$$\sum_{i=1}^{N} \mathbf{x}_i \widehat{u}_i = 0 \tag{14}$$

where $\mathbf{x}_i = (1, x_{2i}, .., x_{ki})'$

Proof of algebraic property 1, continued

Proof (cont.)

- Then, using the definition of $\widehat{u}_i$ and substituting $\widehat{\boldsymbol{\beta}}$ by (7)

$$\begin{aligned}\sum_{i=1}^{N} \mathbf{x}_i \widehat{u}_i &= \sum_{i=1}^{N} \mathbf{x}_i (y_i - \mathbf{x}_i' \widehat{\boldsymbol{\beta}}) = \sum_{i=1}^{N} \mathbf{x}_i y_i - \sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i' \widehat{\boldsymbol{\beta}} \\ &= \sum_{i=1}^{N} \mathbf{x}_i y_i - \sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i' \left(\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \sum_{i=1}^{N} \mathbf{x}_i y_i = 0\end{aligned}$$

Remark If the model does not include a constant, it is not guaranteed that the sum of the residuals is zero. Moreover, although $\sum_{i=1}^{N} x_{ji} \widehat{u}_i = 0, \quad j = 1, 2, .., k$, holds, since $\sum_{i=1}^{N} \widehat{u}_i$ can be different from zero, the sample covariance between the regressors and the residuals can also be different from zero.

Why property 1 holds

- OLS leaves no pattern in the residuals that it could still use to lower the sum of squared residuals.
- **Sum equal to zero.** Suppose the residuals of `precio` on `m2` averaged +5,000 €: the line would sit too low on average. Shifting it up by 5,000 € would lower the sum of squared residuals. OLS has already made that shift, so at the optimum the errors cancel out.
- **Zero covariance with each regressor.** Suppose large flats had mostly positive residuals and small flats mostly negative ones. The line would be too flat, and tilting it would lower the errors. OLS stops tilting exactly when no linear relationship between `m2` and the residuals is left.
- The residuals keep everything floor area cannot explain *linearly*.
- **Without a constant** the line cannot move up or down, so nothing forces the residuals to sum to zero.

Algebraic property 2 (fitted values)

- **Property 2.**
  If the model includes a constant

$$\sum_{i=1}^{N} y_i = \sum_{i=1}^{N} \widehat{y}_i$$

  where $\widehat{y}_i = \mathbf{x}_i'\widehat{\boldsymbol{\beta}}$, $i = 1, 2, , .., N$, are the fitted values.

- *Proof*
  - Since the residuals are defined $\widehat{u}_i = y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}}$, we have that $y_i = \widehat{y}_i + \widehat{u}_i$, and using property 1

$$\sum_{i=1}^{N} y_i = \sum_{i=1}^{N} \widehat{y}_i + \sum_{i=1}^{N} \widehat{u}_i = \sum_{i=1}^{N} \widehat{y}_i$$

What property 2 says

- Dividing by N, $\overline{y} = \overline{\widehat{y}}$: on average the fitted values are exactly right. In Analysis A2 the mean asking price and the mean fitted price are both 158,666.30 €.
- In the simple regression model it also means that the regression line passes through the point of means $(\overline{x}, \overline{y})$, since $\overline{y} = \widehat{\beta}_0 + \widehat{\beta}_1 \overline{x}$.

Algebraic property 3 (the residual-maker matrix M)

- **Property 3.** We can write the vector of residuals $\widehat{\mathbf{u}} = (\widehat{u}_1, \widehat{u}_2, .., \widehat{u}_N)'$ as

$$\widehat{\mathbf{u}} = \mathbf{M}\mathbf{y} = \mathbf{M}\mathbf{u}$$

where

$$\mathbf{M} = \mathbf{I}_N - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

and $\mathbf{M}$ has the following properties

- $\mathbf{M}$ is a symmetric idempotent matrix
- $\mathbf{M}\mathbf{X} = 0$
- $trace(\mathbf{M}) = N - k$.

We will use this property to find the distribution of the exact test statistics under normality.

Proof of algebraic property 3

- *Proof*
  - Since $\widehat{u}_i = y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}}$, we can write the vector of residuals as $\widehat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$, and substituting $\widehat{\boldsymbol{\beta}}$ we have that

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

    where

$$\mathbf{M} = \mathbf{I}_N - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

  - We will now show that $\mathbf{M}$ satisfies the three properties above
  -

$$\begin{aligned}\mathbf{M}\mathbf{M} &= \left(\mathbf{I}_N - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)\left(\mathbf{I}_N - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)\\ &= \mathbf{I}_N - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\\ &= \mathbf{I}_N - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{M}\end{aligned}$$

    and $\mathbf{M}$ is an idempotent matrix.

Proof of algebraic property 3, continued

- *Proof (cont.)*
  - $$\mathbf{MX} = \left(\mathbf{I}_N - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)\mathbf{X} = \mathbf{X} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} = 0$$
  - $$\begin{aligned} trace(\mathbf{M}) &= trace\left(\mathbf{I}_N - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right) \\ &= trace\left(\mathbf{I}_N\right) - trace\left(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right) \\ &= N - trace\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\right) = N - trace(\mathbf{I}_k) = N - k \end{aligned}$$
  - Finally, substituting $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ and using that $\mathbf{MX} = 0$, we have that

$$\widehat{\mathbf{u}} = \mathbf{My} = \mathbf{M}(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) = \mathbf{Mu}$$

Algebraic property 4 ($SST = SSE + SSR$)

- **Property 4**

$$\sum_{i=1}^{N} (y_i - \overline{y})^2 = \sum_{i=1}^{N} \left(\widehat{y}_i - \overline{\widehat{y}}\right)^2 + \sum_{i=1}^{N} \widehat{u}_i^2$$

  where
  - $\sum_{i=1}^{N} (y_i - \overline{y})^2$ is called the **total sum of squares** (SST)
  - $\sum_{i=1}^{N} \left(\widehat{y}_i - \overline{\widehat{y}}\right)^2$ is called the **explained sum of squares** (SSE)
  - $\sum_{i=1}^{N} \widehat{u}_i^2$ is called the **residual sum of squares** (SSR).
- *Proof*
  - Using the definition of the residuals and the fitted values we have that

$$y_i = \widehat{y}_i + \widehat{u}_i, \quad i = 1, 2, .., N$$

  - From property 2, $\overline{y} = \overline{\widehat{y}}$, and

$$y_i - \overline{y} = \widehat{y}_i - \overline{\widehat{y}} + \widehat{u}_i, \quad i = 1, 2, .., N$$

Proof of algebraic property 4, continued

- *Proof (cont.)*
  - Then

$$\begin{aligned}\sum_{i=1}^{N}(y_i-\overline{y})^2 &= \sum_{i=1}^{N}\left(\widehat{y}_i-\overline{\widehat{y}}+\widehat{u}_i\right)^2\\ &= \sum_{i=1}^{N}\left(\widehat{y}_i-\overline{\widehat{y}}\right)^2+\sum_{i=1}^{N}\widehat{u}_i^2+2\sum_{i=1}^{N}\left(\widehat{y}_i-\overline{\widehat{y}}\right)\widehat{u}_i\end{aligned}$$

  - and since

$$\begin{aligned}\sum_{i=1}^{N}\left(\widehat{y}_i-\overline{\widehat{y}}\right)\widehat{u}_i &= \sum_{i=1}^{N}\widehat{y}_i\widehat{u}_i-\sum_{i=1}^{N}\overline{\widehat{y}}\widehat{u}_i\\ &= \sum_{i=1}^{N}\widehat{\beta}'x_i\widehat{u}_i-\overline{\widehat{y}}\sum_{i=1}^{N}\widehat{u}_i \underset{\text{by prop 1}}{=} 0\end{aligned}$$

  - We have that

$$\sum_{i=1}^{N}(y_i-\overline{y})^2=\sum_{i=1}^{N}\left(\widehat{y}_i-\overline{\widehat{y}}\right)^2+\sum_{i=1}^{N}\widehat{u}_i^2$$

Goodness of fit and R^2

- Using property 4 we can define a goodness of fit measure of the regression that is called $R-$**squared** and is defined as

$$R^2 = \frac{\sum_{i=1}^{N} (\widehat{y}_i - \overline{\widehat{y}})^2}{\sum_{i=1}^{N} (y_i - \overline{y})^2} = 1 - \frac{\sum_{i=1}^{N} \widehat{u}_i^2}{\sum_{i=1}^{N} (y_i - \overline{y})^2}$$

- The $R-$squared measures the proportion of the variability of y that is explained by the regressors.

Properties of R^2 and adjusted R^2

- $0 \leq R^2 \leq 1$
- If $R^2 = 1 \Rightarrow y_i = \mathbf{x}_i'\widehat{\boldsymbol{\beta}}, \quad i = 1, 2, .., N$. The fit is perfect.
- If $R^2 = 0 \Rightarrow \widehat{y}_i = \overline{y}, \quad i = 1, 2, .., N$. None of the variation of y is accounted for by the variation in the xs.
- R^2 has no meaning if the regression does not include a constant.
- R^2 always increases when we include additional explanatory variables. To account for this the **adjusted** $R-$**squared**, $\overline{R}^2$, is defined as

$$\overline{R}^2 = 1 - \frac{\frac{\sum_{i=1}^{N} \widehat{u}_i^2}{N-k}}{\frac{\sum_{i=1}^{N} (y_i - \overline{y})^2}{N-1}}$$

Why R^2 never falls when we add a regressor

- **Intuition.** With the new regressor, OLS can always choose to ignore it by setting its coefficient to zero. That reproduces the smaller model, so the best fit of the larger model can only be as good or better.
- **Proof.** Compare $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{u}$ (restricted, R) with $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{x}_2\beta_2 + \mathbf{u}$ (unrestricted, U), both with a constant:

$$SSR_U = \min_{\mathbf{b}_1,\, b_2} \sum_{i=1}^{N} (y_i - \mathbf{x}_{1i}'\mathbf{b}_1 - x_{2i}b_2)^2 \;\le\; \sum_{i=1}^{N} (y_i - \mathbf{x}_{1i}'\widehat{\boldsymbol{\beta}}_{1R} - x_{2i}\cdot 0)^2 = SSR_R$$

  because $(\widehat{\boldsymbol{\beta}}_{1R}, 0)$ is one of the choices available in the minimisation. SST depends only on y, so

$$R^2_U = 1 - \frac{SSR_U}{SST} \;\ge\; 1 - \frac{SSR_R}{SST} = R^2_R.$$

- The result compares a model with its own extension. It says nothing about two models with *different* regressors.

Analysis A3 · $SST = SSE + SSR$ with the full hedonic model

`lm(f_niv)`: price in euros on the full set of controls

SSE (explained, `e(mss)`)	7.6798×10^{12}
SSR (residual, `e(rss)`)	1.5507×10^{12}
$SST = SSE + SSR$	9.2305×10^{12}
$R^2 = SSE/SST$	0.8320
$\overline{R}^2$	0.8280
Parameters k	24

- Adding terrace, age, floor, zone and year raises R^2 from 0.46 to **0.83**.
- A valuation tool would be delighted. Is the *terrace coefficient* now right? R^2 cannot tell you. Hold that thought until Block C.

Exercise on goodness of fit

Exercise. A colleague adds 20 irrelevant regressors and the R^2 rises from 0.30 to 0.33. Is the new model better?

Exercise on goodness of fit

Exercise. A colleague adds 20 irrelevant regressors and the R^2 rises from 0.30 to 0.33. Is the new model better?

Answer. Not necessarily. R^2 can never fall when regressors are added. Look at $\overline{R}^2$, which penalizes extra regressors, and remember that neither measure tells us whether the coefficients have a causal interpretation.

R^2 out of sample

- R^2 measures how well the regression fits the flats in the estimation sample. What matters to a property portal, however, is how accurately the model predicts prices for flats it has not yet observed.
- To test how well the model predicts new observations, estimate it using only part of the sample:
  1. split the sample; estimate $\widehat{\boldsymbol{\beta}}_{\text{train}}$ on the *training* half only;
  2. predict each flat j in the *test* half: $\widehat{y}_j = \mathbf{x}_j' \widehat{\boldsymbol{\beta}}_{\text{train}}$;
  3. compute the same ratio on the test flats:

$$R^2_{\text{out}} = 1 - \frac{\sum_{j \in \text{test}} \left(y_j - \mathbf{x}_j' \widehat{\boldsymbol{\beta}}_{\text{train}} \right)^2}{\sum_{j \in \text{test}} \left(y_j - \overline{y}_{\text{test}} \right)^2}$$

- $SST = SSE + SSR$ no longer holds.

Analysis A4 · More parameters, worse predictions

Train on the odd `ids` (500 flats), predict the even ones (500 flats), $y = \log(precio)$:

In sample vs. out of sample

Model	k	R^2	$\overline{R}^2$	R^2 out of sample
Hedonic model	24	0.904	0.899	**0.890**
+ zone×year, slopes by zone and year	93	0.917	0.899	**0.865**

- 69 more parameters: R^2 up 1.3 points…
- …R^2 on new flats down 2.5 points.
- $\overline{R}^2$ barely moves: its penalty is mild.

Adjusted R^2, AIC, BIC and cross-validation are four versions of one idea: **penalise complexity**. A portal's price-suggestion model lives or dies on the out-of-sample number.

Block A in R

```
d <- read.csv("data/idealista_alicante_clase.csv")
d$zona <- factor(d$zona); d$anio <- factor(d$anio)
d$lprecio <- log(d$precio); d$lm2 <- log(d$m2)
f_niv <- precio ~ m2 + m2_terraza + terraza +
  antig + I(antig^2) + bajo + zona + anio
f_log <- lprecio ~ lm2 + terraza + m2_terraza +
  antig + I(antig^2) + bajo + zona + anio

# A1: euros per square metre
m1 <- lm(precio ~ m2, data = d)

# A2: the first-order conditions
e <- resid(m1); cor(e, d$m2)

# A3: SST = SSE + SSR
m3 <- lm(f_niv, data = d)
sum(resid(m3)^2) +
  sum((fitted(m3) - mean(d$precio))^2)

# A4: in and out of sample
lm(f_log, data = d[d$id %% 2 == 1, ])
```

1. Do the two sums of squares really add up to *SST* (the sum of squared deviations of `precio` from its mean)?
2. Regress `precio` on `m2` *without* a constant, `lm(precio ~ 0 + m2)`. Do the residuals still sum to zero? Why not?
3. Add `factor(mes)` to the model. What happens to R^2 in and out of sample?
4. Add a column of pure noise, `d$ruido <- rnorm(nrow(d))`, to `lm(precio ~ m2)`. Does R^2 rise? Does $\overline{R}^2$?

Full script: `01_analisis_clase.R`.