Unit 2. The single-equation linear model

Block E. Is OLS the best we can do?

Instructor: Henry Redondo

Dpto. de Fundamentos del Análisis Económico
Universidad de Alicante

Econometrics II · Academic Year 2026/2027

Outline of Block E

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

2. The answer for the pricing team

Testing for heteroskedasticity

- Under *OLS3′*, $E(u^2 \mid \mathbf{x}) = \sigma^2$ does not depend on $\mathbf{x}$. So we can test whether u^2 varies with the regressors.
- **Breusch–Pagan test.** Run the auxiliary regression

$$\widehat{u}_i^2 = \delta_0 + \delta_1 z_{i1} + \cdots + \delta_q z_{iq} + \text{error}$$

  where $z_{i1}, \ldots, z_{iq}$ are the regressors or functions of them (for example, the fitted values $\widehat{y}_i$), and test $H_0 : \delta_1 = \cdots = \delta_q = 0$.
- It is an *LM* test. Under H_0,

$$LM = N\, R^2_{\widehat{u}^2} \;\to_d\; \chi^2_q$$

  where $R^2_{\widehat{u}^2}$ is the R^2 of the auxiliary regression.
- If we reject, either keep OLS and use robust standard errors, or change the estimator and give noisy observations less weight (GLS).

Analysis D7 · Is there heteroskedasticity?

`bptest(m)`: $N \cdot R^2$ of $\widehat{u}^2$ on the fitted values

Model	$\chi^2(1)$	p
Levels (€)	122.80	10^{-28}
Logs	17.37	3×10^{-5}

Analysis D7 · Is there heteroskedasticity?

`bptest(m)`: $N \cdot R^2$ of $\widehat{u}^2$ on the fitted values

Model	$\chi^2(1)$	p
Levels (€)	122.80	10^{-28}
Logs	17.37	3×10^{-5}

- Logs remove most of it, not all: residual s.d. is **0.148** elsewhere and **0.201** at the beach. Beach prices are harder to predict.
- Two responses: keep OLS and use **robust** standard errors (what we did), or change the **estimator** and give noisy observations less weight. That is GLS.

The GLS estimator

- 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.

- Let $\{(\mathbf{x}_i', y_i)'\}$, $i = 1, .., N$, be a random sample of size N from the population. We can write the model for each element of the sample

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i \tag{54}$$

and we assume that the model satisfies $OLS1'$ and $OLS2$ but

$$E(u_i^2 \mid \mathbf{x}_i) = \sigma^2 w(\mathbf{x}_i) \tag{55}$$

- We know that the OLS estimator is unbiased. However, if $w(\mathbf{x}_i)$ is known, we can obtain a more efficient estimator of $\boldsymbol{\beta}$ by estimating a transformed model.

The transformed model

- If we divide model (54) by $\sqrt{w(\mathbf{x}_i)}$ we get

$$y_i^* = \mathbf{x}_i^{*\prime}\boldsymbol{\beta} + u_i^* \tag{56}$$

where

$$y_i^* = \frac{y_i}{\sqrt{w(\mathbf{x}_i)}} \quad \mathbf{x}_i^* = \frac{\mathbf{x}_i}{\sqrt{w(\mathbf{x}_i)}} = \begin{pmatrix} \frac{1}{\sqrt{w(\mathbf{x}_i)}} \\ \frac{x_{2i}}{\sqrt{w(\mathbf{x}_i)}} \\ \vdots \\ \frac{x_{ki}}{\sqrt{w(\mathbf{x}_i)}} \end{pmatrix} \quad u_i^* = \frac{u_i}{\sqrt{w(\mathbf{x}_i)}}$$

- We are going to show that this model satisfies $OLS1'$, $OLS2$ and $OLS3'$.

The transformed model satisfies *OLS*1′, *OLS*2, *OLS*3′

- *OLS*1′ :
  Given that u_i satisfies *OLS*1′

$$E[u_i^* \mid \mathbf{x}_i] = E\left[\frac{u_i}{\sqrt{w(\mathbf{x}_i)}} \mid \mathbf{x}_i\right] = \frac{1}{\sqrt{w(\mathbf{x}_i)}} E(u_i \mid \mathbf{x}_i) = 0$$

  Using the law of iterated expectations, since $\mathbf{x}_i^*$ is a function of $\mathbf{x}_i$

$$E[u_i^* \mid \mathbf{x}_i^*] = E\left[E[u_i^* \mid \mathbf{x}_i] \mid \mathbf{x}_i^*\right] = 0$$

- *OLS*2 : If $E(\mathbf{x}_i \mathbf{x}_i')$ has full rank $E(\mathbf{x}_i^* \mathbf{x}_i^{*\prime})$ has also full rank.
- *OLS*3′ :

$$E[(u_i^*)^2 \mid \mathbf{x}_i] = E\left[\frac{u_i^2}{w(\mathbf{x}_i)} \mid \mathbf{x}_i\right] = \frac{1}{w(\mathbf{x}_i)} E(u_i^2 \mid \mathbf{x}_i) \underset{\text{Using (55)}}{=} \sigma^2$$

  Using the law of iterated expectations, since $\mathbf{x}_i^*$ is a function of $\mathbf{x}_i$

$$E[(u_i^*)^2 \mid \mathbf{x}_i^*] = E\left[E[(u_i^*)^2 \mid \mathbf{x}_i] \mid \mathbf{x}_i^*\right] = \sigma^2$$

GLS is BLUE

- The *GLS* (Generalized Least Squares) estimator is the *OLS* estimator of (56)

$$\widehat{\boldsymbol{\beta}}_{GLS} = \left(\sum_{i=1}^{N} \mathbf{x}_i^* \mathbf{x}_i^{*\prime}\right)^{-1} \left(\sum_{i=1}^{N} \mathbf{x}_i^* y_i^*\right) = \left(\sum_{i=1}^{N} \frac{1}{w(\mathbf{x}_i)} \mathbf{x}_i \mathbf{x}_i'\right)^{-1} \left(\sum_{i=1}^{N} \frac{1}{w(\mathbf{x}_i)} \mathbf{x}_i y_i\right)$$

- Since the *GLS* estimator is an *OLS* estimator of a model that satisfies $OLS1'$, $OLS2$ and $OLS3'$, using Gauss-Markov's theorem, the *GLS* estimator is the Best Linear Unbiased Estimator and since *OLS* is linear and unbiased, we have that

$$Var(\widehat{\boldsymbol{\beta}} \mid \mathbf{X}) \geq Var(\widehat{\boldsymbol{\beta}}_{GLS} \mid \mathbf{X})$$

- The *GLS* estimator for correcting heteroskedasticity is also called Weighted Least Squares (WLS) estimator. This name comes from the fact that $\widehat{\boldsymbol{\beta}}_{GLS}$ minimizes the weighted sum of squared residuals, where each squared residual is weighted by $1/w(\mathbf{x}_i)$.

Feasible GLS

- The *GLS* estimator is very appealing from a theoretical point of view, however, in practice we rarely know $w(\mathbf{x}_i)$, and if we don't know $w(\mathbf{x}_i)$, the *GLS* estimator is not feasible.
- Sometimes it is possible to assume a certain functional form for $w(\mathbf{x}_i)$, like $w(\mathbf{x}_i) = \exp(\mathbf{x}_i'\delta)$. Then, although $w(\mathbf{x}_i)$ is unknown, the vector of parameters δ can be estimated using the *OLS* residuals.
- Then, using $\widehat{\delta}$ we can compute $\widehat{w(\mathbf{x}_i)} = \exp(\mathbf{x}_i'\widehat{\delta})$ and use these estimates to calculate the so called feasible *GLS* estimator:

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

- The feasible *GLS* estimator does not have the same good finite sample properties as the *GLS* estimator since it is not unbiased. However it can be shown that it is consistent and asymptotically more efficient than *OLS*.

When is $w(\mathbf{x})$ known?

- We have already mentioned that we rarely know $w(\mathbf{x}_i)$. However, there is one case where $w(\mathbf{x}_i)$ arises naturally from an underlying econometric model. This happens when, instead of using individual level data, we only have averages of data across some group or geographic region.
- Suppose we have per capita data at the city, county, state, or country level. If the individual-level equation satisfies the Gauss-Markov assumptions, then the error in the per capita equation has a variance proportional to one over the size of the population. Therefore, weighted least squares with weights equal to the population is appropriate.

Analysis D8 · Weighted least squares with aggregated data

A common situation in industry: you only receive **averages**, here the mean $\log(\text{£/m}^2)$ in each zone×year cell.

- 70 cells with between 5 and 25 listings (mean 14.3). A cell mean from 25 flats is far more precise: $\text{Var}(\bar{y}_c) = \sigma^2 / n_c$.
- GLS logic: weight each cell by n_c. `lm(lpm2 ~ zona + anio, weights = n)`

OLS (equal weights) vs. WLS (weights n_c)

	OLS		WLS	
	Coef.	s.e.	Coef.	s.e.
Centro	0.0347	0.0207	0.0419	0.0181
Playa de San Juan	0.5034	0.0207	0.5034	0.0196
2025	0.5376	0.0347	0.5447	0.0317

Analysis D8 · Weighted least squares with aggregated data

A common situation in industry: you only receive **averages**, here the mean $\log(\text{£/m}^2)$ in each zone×year cell.

- 70 cells with between 5 and 25 listings (mean 14.3). A cell mean from 25 flats is far more precise: $\text{Var}(\bar{y}_c) = \sigma^2 / n_c$.
- GLS logic: weight each cell by n_c. `lm(lpm2 ~ zona + anio, weights = n)`

OLS (equal weights) vs. WLS (weights n_c)

	OLS		WLS	
	Coef.	s.e.	Coef.	s.e.
Centro	0.0347	0.0207	0.0419	0.0181
Playa de San Juan	0.5034	0.0207	0.5034	0.0196
2025	0.5376	0.0347	0.5447	0.0317

The coefficients hardly move and the standard errors fall a little. The gain is modest because cell sizes are not very unequal. With cells of 5 and 5,000 observations (stores, regions, days), weighting matters a lot.

Blocks D and E in R

```
# D1: classical vs robust
mn <- lm(f_niv, d); ml <- lm(f_log, d)
sqrt(diag(vcov(mn)))
sqrt(diag(vcovHC(mn, "HC1")))
# D2-D3: t and F / Wald
coeftest(ml, vcovHC(ml, "HC1"))
zonas <- grep("^zona", names(coef(ml)), value = TRUE)
linearHypothesis(ml, paste(zonas, "= 0"),
                 vcov. = vcovHC(ml, "HC1"))
# D4-D5: the appraisers' rule
linearHypothesis(mn, "m2_terraza - 0.5*m2 = 0",
                 vcov. = vcovHC(mn, "HC1"))
# D6: LM
mr  <- lm(update(f_log, . ~ . - zona), d)
aux <- lm(update(f_log, resid(mr) ~ .), d)
nobs(aux) * summary(aux)$r.squared
# D7-D8: heteroskedasticity, WLS
bptest(ml, ~ fitted(ml))
lm(lpm2 ~ zona + anio, cells, weights = n)
```

1. Test the 50% rule in the *log* model. What exactly is the hypothesis there?
2. Is the beach terrace interaction (B7) significant with robust standard errors?
3. Compute LM and W for a *false* restriction that is close to true. Do they agree?

Exercise on heteroskedasticity

Exercise. Suppose $OLS1'$ and $OLS2$ hold but the errors are heteroskedastic. Which statements are true?

(a) OLS is inconsistent.

(b) OLS is biased.

(c) The usual standard errors $\sqrt{\widehat{\sigma}^2 a_{jj}}$ are no longer valid.

(d) OLS is no longer BLUE.

Exercise on heteroskedasticity

Exercise. Suppose $OLS1'$ and $OLS2$ hold but the errors are heteroskedastic. Which statements are true?

(a) OLS is inconsistent.

(b) OLS is biased.

(c) The usual standard errors $\sqrt{\widehat{\sigma}^2 a_{jj}}$ are no longer valid.

(d) OLS is no longer BLUE.

Answer. Only (c) and (d). Consistency and unbiasedness do not use $OLS3$, so use **heteroskedasticity-robust** standard errors; a more efficient estimator (GLS) exists, as we have seen.

Which assumptions deliver what?

Result	$OLS1$	$OLS1'$	$OLS2$	$OLS3$	$OLS3'$	$OLS4$
Consistency	✓		✓			
Asymptotic normality[†]	✓		✓			
Avar $= \sigma^2 E(\mathbf{xx}')^{-1}$	✓		✓	✓		
Unbiasedness		✓	✓			
$var(\widehat{\boldsymbol{\beta}} \mid \mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, BLUE		✓	✓		✓	
Exact t and F tests			✓			✓

[†] provided $E(u^2\mathbf{xx}')$ exists.

How the assumptions nest. $OLS1' \Rightarrow OLS1$, $OLS3' \Rightarrow OLS3$, $OLS4 \Rightarrow OLS1'$ and $OLS3'$. Stronger assumptions buy stronger (finite-sample) results.

Summary of the OLS estimator

1. $\boldsymbol{\beta} = E(\mathbf{x}\mathbf{x}')^{-1}E(\mathbf{x}y)$; OLS is its sample analogue.
2. $OLS1 + OLS2$: consistent and asymptotically normal. Robust standard errors do not need homoskedasticity.
3. $OLS1' + OLS2$: unbiased. Adding $OLS3'$: variance $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ and BLUE.
4. Residuals are orthogonal to the regressors; $SST = SSE + SSR$; R^2 measures fit, not causality.
5. Coefficients are partial effects; logs give percentages; with quadratics the effect depends on x.
6. Exact t and F need normality; otherwise use Wald or LM tests in large samples.
7. Dummies shift intercepts, interactions shift slopes; always leave out one category.
8. GLS/WLS reweights observations when the form of heteroskedasticity is known.

Analysis E1 · How much is a terrace worth? Every version

Premium for a typical terrace (13.4 m^2), from Block B to Block E:

One question, five answers

Specification	Where	Premium	On a 130,000 € flat
No controls (`lm(lprecio ~ terraza)`)	–	46.0%	
Hedonic model	B	19.2%	24,935 €
+ proxy (number of photos)	C2	15.2%	
+ quality (the oracle)	C1	11.4%	14,810 €
Truth (simulation)		**13.0%**	**16,845 €**

Analysis E1 · How much is a terrace worth? Every version

Premium for a typical terrace (13.4 m^2), from Block B to Block E:

One question, five answers

Specification	Where	Premium	On a 130,000 € flat
No controls (`lm(lprecio ~ terraza)`)	–	46.0%	
Hedonic model	B	19.2%	24,935 €
+ proxy (number of photos)	C2	15.2%	
+ quality (the oracle)	C1	11.4%	14,810 €
Truth (simulation)		**13.0%**	**16,845 €**

- The naive number is three and a half times the truth. The "good" hedonic model still overstates it by almost half, around 8,000 € per flat.
- Every extra layer of econometrics moved the answer *towards* the truth. None of them showed up in R^2.

The memo

To: Pricing team From: Data science Re: terrace premium. A terrace adds about **11–15%** to the asking price of a comparable flat in Alicante, probably more on the beach front (Playa de San Juan) than inland. How a terrace square metre compares with an interior one is imprecisely estimated: we cannot reject the appraisers' 50% rule, but the data are also consistent with much higher values.

Do not use the raw difference (+46%) or the basic hedonic coefficient (+19%): both attribute to the terrace the higher overall quality of flats that have one. We correct for this with a quality proxy (number of photos); true quality is unobserved, so our figure is probably still slightly high.

Standard errors are robust to heteroskedasticity. Price levels are in the Alicante market of 2013–2026.

The estimate and its range. **What it means**, in the audience's units. **The key assumption**, stated plainly. **The main threat**, and which direction it pushes.