Universidad Carlos III de Madrid

César Alonso

ECONOMETRICS

Topic 7: HETEROSKEDASTICITY[*]

Contents

[*] This material is based on previous work done by María Arrazola and José de Hevia.

1 Introduction

- For the appropriate analysis of most economic phenomena, the HOMOSKEDASTICITY assumption (constant conditional error variance), which is established in the classical regression model, must be relaxed.

- HETEROSKEDASTICITY is often found both with cross section and with time series data.

- The term **heteroskedasticity** means that situation where the variance of Y conditional to the variables in the model is not constant for the different X's values.

1.1 Examples

- Consider 100 students in a typewriting class. Some of them have already typewriting earlier, while some other have never done it. After the first class, some students typewrite badly while some others do quite well.

  If we depicted a figure showing the typewriting errors of students with respect to their cumulated typewriting hours, we would see that the dispersion of typewriting errors decreases as long as the explanatory variable (number of cumulated hours of typewriting) rises.

  Besides, the difference in the number of typewriting errors between the best one and the worst one in class will be possibly lower after the last class than after the first class.

- Consider household data on income and food expenditure. If we depicted a figure of food expenditure against income, we would possibly find heteroskedasticity.

It is very likely that, for different income levels, the dispersion in food expenditure to increase with income.
Poor households enkoy less flexibilility in their food expenditure, so we would expect low dispersion in food expenditure for them. On the contrary, among wealthy households, we will observe some spending a lot in food (e.g., eating caviar or expensive meals) while other ones with different preferences could spend much less in food, dedicating their income to other purposes.

2 The linear regression model with heteroskedasticity

- Without any loss of generality, consider the simple regresion case:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

where:

$$\begin{aligned} &E(\varepsilon|X) = 0 \\ &\qquad \Rightarrow E(Y|X) = \beta_0 + \beta_1 X = PLO(Y|X) \\ &\qquad \Rightarrow \ \beta_0 = E(Y) - \beta_1 E(X) \quad \text{y} \quad \beta_1 = \frac{C(Y,X)}{V(X)} \end{aligned}$$

and:

$$V(\varepsilon|X) = V(Y|X) = \sigma^2(X) \qquad \text{HETEROSKEDASTICITY}$$

- We will learn that:
  - OLS is unbiased and consistent, even in the absence of homoskedasticity.

- The standard errors of the estimates based on the conventional formula are biased and inconsistent in the presence of heteroskedasticity.

  (and therefore, they are not valid for inference).

2.1 Consequences on OLS estimation

- What are the properties of the OLS estimator in this context?
- let's focus on the estimation of β_1.
- If we have a random sample, tis properties are:
  - **Unbiasedness**

    We know that $\widehat{\beta}_1 = \sum_i c_i Y_i$, with $c_i = \dfrac{x_i}{\sum_i x_i^2}$.
    Since $\sum_i c_i = 0$ and $\sum_i c_i X_i = 1$, then:

    $$\begin{aligned} E(\widehat{\beta}_1|X_i) &= E(\textstyle\sum_i c_i Y_i|X_i) = \sum_i c_i E(Y_i|X_i) \\ &= \textstyle\sum_i c_i(\beta_0 + \beta_1 X_i) = \beta_1 \end{aligned}$$

    Thus,

    $$E(\widehat{\beta}_1) = \beta_1 \Rightarrow \quad \text{Unbiasedness}$$

  - **Consistency**

    Recall that

    $$\widehat{\beta}_1 = \frac{\sum_i x_i y_i}{\sum_i x_i^2} = \frac{\frac{1}{n}\sum_i x_i y_i}{\frac{1}{n}\sum_i x_i^2},$$

    so that

    $$p\lim \widehat{\beta}_1 = \frac{p\lim \frac{1}{n}\sum_i x_i y_i}{p\lim \frac{1}{n}\sum_i x_i^2} = \frac{C(X,Y)}{V(X)} = \beta_1 \Rightarrow \quad \text{Consistency}$$

- **Variance form**

$$\begin{aligned} V(\widehat{\beta}_1|X_i) = \sigma^2_{\widehat{\beta}_1} &= V(\textstyle\sum_i c_i Y_i|X_i) = \textstyle\sum_i c_i^2 V(Y_i|X_i) \\ &= \textstyle\sum_i c_i^2 \sigma_i^2 = \dfrac{\sum_i x_i^2 \sigma_i^2}{\left(\sum_i x_i^2\right)^2} \\ &\Rightarrow V(\widehat{\beta}_1) = E\left[\dfrac{\sum_i x_i^2 \sigma_i^2}{\left(\sum_i x_i^2\right)^2}\right] \end{aligned}$$

  Notice that this is no longer the usual expression.

  Hence, confidence intervals and hypotheses testing based on the conventional form (under homoskedasticity)

$$s^2_{\widehat{\beta}_1} = \frac{\sigma^2}{E\left(\sum_i x_i^2\right)}$$

  are not valid. The magnitude of the “bias” will depend on the magnitude of Heteroskedasticity.

- **Efficiency**

  OLS is **no longer**, in this context, the Best Linear Unbiased Estimation.

2.2 Consequences on the instrumental-variable estimators

- Likewise, the consistency property is not altered under heteroskedasticity, because variance assumptions are not needed to proof consistency.
- Nevertheless, the conventional variance expression (derived under homoskedasticity) is no longer valid, thus invalidating any inference based on the conventional variance expression.

3 Inference Robust to heteroskedasticity: White or Eiker-White proposal

- Note that, in general, the fuctional form for σ_i^2 is unknown.

- However, asymptotic results allow us to build and "approximate" measure of confidence intervals and hypotheses testing, by means of the proposal of White or Eiker-White for $\sigma^2_{\widehat{\beta}_1} = E\left[\frac{\sum_i x_i^2 \sigma_i^2}{\left(\sum_i x_i^2\right)^2}\right]$.

- This proposal is valid both for the OLS and the IV/2SLS estimators. However, formally we will focus on OLS.

- As its main advantage, the proposal does not require any knowledge of the pattern characterizing heteroskedasticity.

- The proposal consists on computing

  $$s^2_{\widehat{\beta}_1} = \frac{\sum_i x_i^2 \widehat{\varepsilon}_i^2}{\left(\sum_i x_i^2\right)^2},$$

  where $\widehat{\varepsilon}_i$ are the OLS residuals, i.e.:

  $$\widehat{\varepsilon}_i = Y_i - \widehat{Y}_i = Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 X_i)$$

- The proposed standard errors of the estimators are known as **robust standard errors**, or heteroskedasticity-consistent standard errors, or Eicker-White standard errors.

- It can be proved that:

  $$\frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\frac{\sum_i x_i^2 \widehat{\varepsilon}_i^2}{\left(\sum_i x_i^2\right)^2}}} \overset{\cdot}{\sim} N(0,1),$$

i.e.

$$\frac{\widehat{\beta}_1 - \beta_1}{s_{\widehat{\beta}_1}} \overset{a}{\sim} N(0,1),$$

This result can be applied to calculate confidence intervals and undertake tests of hypothesis as usual.

- In practice, most econometric packages provide the option of heteroskedasticity-robust standard errors, both with OLS and IV/2SLS estimation.

- **Important**: as the conditional variance of Y is not constant under heteroskedasticity, goodness-of-fit measures, like the standard error of the regression or the R^2, do not make any sense, as they are only meaningful under homoskedasticity.