Violations of Classical Assumptions IV: Heteroskedasticity

Marcio Santetti

EC 339 | Fall 2022

Table of contents

Introduction

In this lecture, we will study in more detail one of the *most common* violations of CLRM assumptions: **heteroskedasticity**. *Assumption V* states that observations of the error term are drawn from a distribution that has a *constant variance*. This is also known as *homoskedasticity*.[1] When the variance of the error term is no longer constant, we have *heteroskedasticity*.

[1] *homo* = equal; *skedasticity* = spread.

Such violation arises from the fact that the variance of the regression's residual term varies *as regressor values change*. In analytical terms, this means

$$Var(u_i|x_i) = \sigma_i^2 \quad \forall\, i = 1, 2, 3, ..., n$$

Translating the above statement, the *conditional variance* of the error term, by having the i subscript, is now dependent on observations of other regressors' values. It often occurs in data sets where there is a wide *disparity* between the largest and the smallest observed value of the dependent variable. The larger the disparity, the more likely u_i will be heteroskedastic. Such violation is very common in cross-sectional data, as well as in some kinds of time series, such as financial data.

Consider the following example. Assume we want to investigate the relationship between *test scores* and the *student-to-teacher ratio* for a given state. In Figure 1's left panel, notice that, as the student-to-teacher ratio increases, the *range* of test scores also increases. In other words, as this ratio increases, the *variance* of test scores also increases. If we fit a regression line to this panel, the *distance* between the data points and the OLS line will increase as we move from left to right.

In the right panel, we fit this OLS regression line (in red), and we add a box plot according to a few student-teacher ratios (10, 15, 20, and 25), helping us to visualize how the *spread* of test results increases. This is in accordance to reality, since it is straightforward to assume that, the more students the same

number of teachers must give attention to, the more it will reflect on their performance.

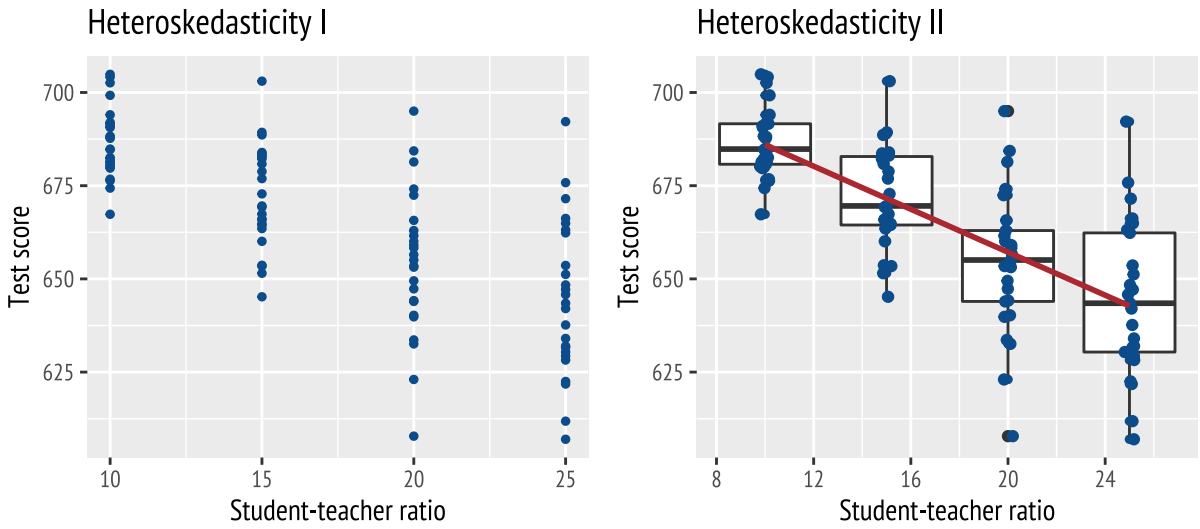

Figure 1: Heteroskedasticity in two ways.

Another example concerns the *wage-education* relationship. The next figure presents data and OLS fit for a simple regression model of hourly earnings and years of education for college-educated full-time US workers in 2004, according to the Current Population Survey (CPS). Notice that we have a similar behavior as the one observed in the previous example: as years of education increase, the range of hourly earnings also increases. Using the OLS regression line makes it easier to see how the distance between the data points and the red line increases, reflecting in a more likely non-constant variance of the error term.

A non-constant variance in the error term implies serious issues for OLS estimation, if not properly addressed. Before looking at ways to overcome this problem, let us study its consequences in more detail.

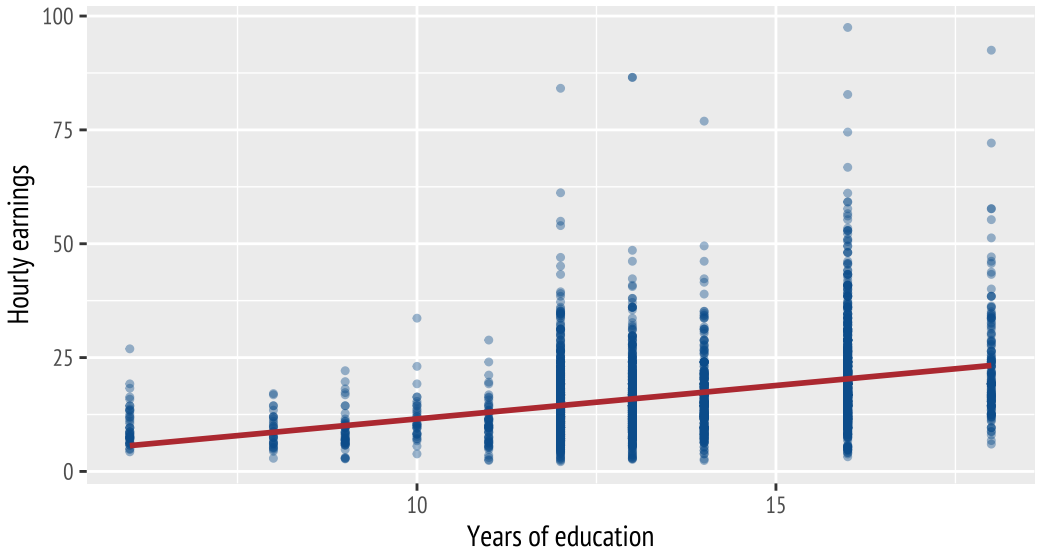

Figure 2: Hourly earnings vs. education.

Consequences of heteroskedasticity

First of all, heteroskedasticity **does not** cause bias in OLS estimates. As with mulicollinearity and serial correlation, heteroskedasticity affects the **standard errors** of OLS β estimates, undermining proper inference from our models.

As a consequence, under heteroskedasticity, OLS estimates will no longer be *BLUE*. Since the standard errors are affected, there is no way in which OLS will yield *minimum variance* coefficients. Lastly, *goodness-of-fit* measures (R^2 and $\bar{R}^2$) are not affected.

Fortunately, we can still make our models *robust* to heteroskedasticity within the OLS spectrum. We first look at a couple of statistical tests we can perform to detect heteroskedasticity, and later we will try to make up for this problem.

Testing for heteroskedasticity

The is no universally agreed upon method for testing for heteroskedasticity, given that there are several different techniques. As with previous violations of CLRM assumtpions, no statistical test has the power to *prove* the presence of heteroskedasticity in a model, but some can indicate its *likelihood*.

Before any test begins, a great starting point is trying to *visually* detect the presence of heteroskedasticity. In addition to the figures shown in the previous section, we can also plot the model's residuals, both in *levels* and in *squared* form. From the *earnings-education* model, we plot the error term in both forms in Figure 3. A dashed red line was added at the error's mean of zero (in accordance with CLRM Assumption II). Notice how most of the observations are concentrated around this mean, but several deviate from it, creating a non-constant variance across its entire support. The advantage of also visualizing *squared* residuals is that, in addition to removing negative signs, higher deviations from the mean will have greater magnitude, and thus will be more visually clear, as it is possible to see in the right panel.

The Breusch-Pagan test

In order to conduct a Breusch-Pagan (BP) test for heteroskedasticity, we start from the original regression model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + u_i$$

After the model is estimated, we store its *residuals*, $\hat{u}_i$, and estimate the following *auxiliary regression*, with its squared, $\hat{u}_i^2$, form as the dependent variable:

$$\hat{u}_i^2 = \delta_0 + \delta_1 x_{1i} + \delta_2 x_{2i} + ... + \delta_k x_{ki} + v_i$$

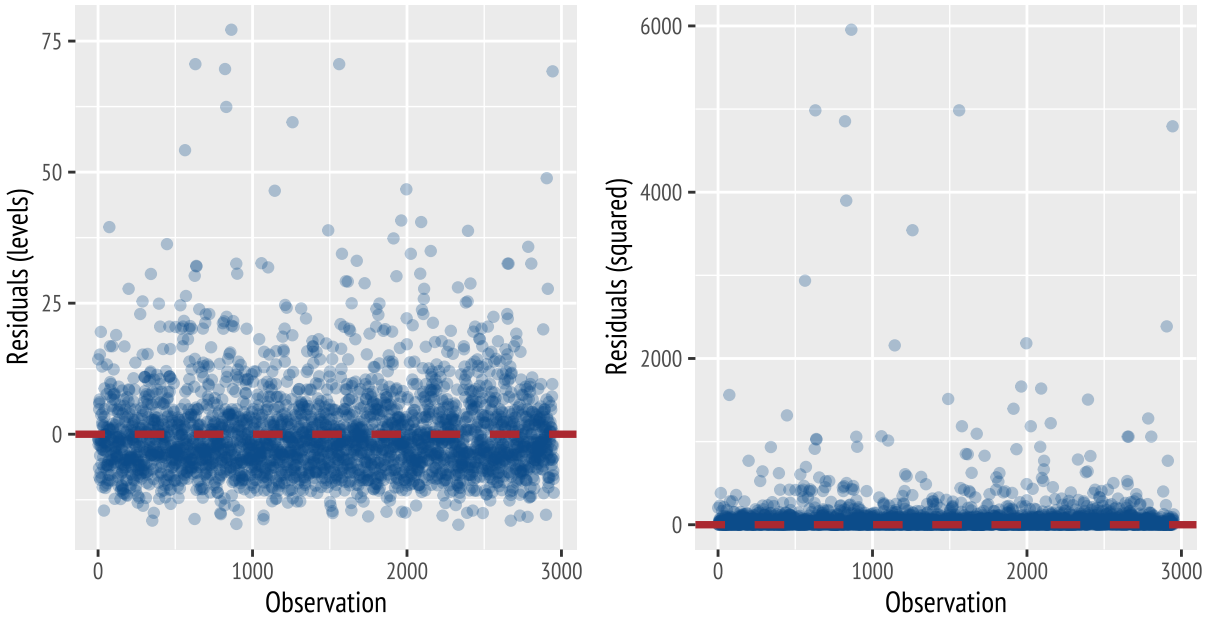

Figure 3: Regression residuals: in levels (left), squared (right).

That is, we regress the estimated squared residuals on *all* the original model's independent variables.

The Breusch-Pagan test's *null hypothesis* is that CLRM Assumption V is *true*, that is, the model does not suffer from heteroskedasticity. Translating this assumption to an *F-test* procedure, we have:

- $H_0 : \delta_1 = \delta_2 = ... = \delta_k = 0$ (*homoskedasticity*)
- $H_1 : H_0$ is not true (*heteroskedasticity*)

Then, we can either calculate *F* or *LM* (Lagrange-multiplier) statistics for this null hypothesis:

$$F = \frac{R^2_{\hat{u}^2}/k}{1 - R^2_{\hat{u}^2}/(n-k-1)} \qquad \text{or} \qquad LM = n \cdot R^2_{\hat{u}^2}$$

where $R^2_{\hat{u}^2}$ is the R-squared coefficient from the auxiliary regression. The BP test is evaluated through the LM test statistic,

which is Chi-squared distributed with k degrees-of-freedom, where k is the number of slope coefficients in the auxiliary regression.

In case we *reject* the null hypothesis, CLRM Assumption V is *violated* and we have evidence of heteroskedasticity in our regression model.

The White test

White (1980)[2] proposed a more general form of the Breusch-Pagan test, allowing for $\hat{u}^2$ to be correlated with squares, cubes, interaction terms, among other functional forms of all independent variables.

[2] White, H. (1980). *A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity*. **Econometrica**, 48(4): pp. 817–838.

Following the same reasoning we adopted when studying the *RESET* test for functional form misspecification, including all these terms in a regression model would consume several degrees-of-freedom. To circumvent it, we use instead functional forms of the *estimated dependent variable*, such as $\hat{y}^2$, $\hat{y}^3$, etc. Usually, including squares of the dependent variable will suffice.

The basic difference regarding the White test, relative to the BP procedure, is to estimate the *auxiliary regression* by adding these functional forms of $\hat{y}$:

$$\hat{u}_i^2 = \delta_0 + \delta_1 \hat{y}_i + \delta_2 \hat{y}_i^2 + \nu_i$$

In this specification $\hat{y}_i$ captures the entire right-hand side of the BP test's auxiliary regression, and we augment this equation by adding higher powers of the estimated dependent variable according to our needs. In the above regression, we have just added its squared form. Then, the null and alternative hypotheses become:

- $H_0 : \delta_1 = \delta_2 = 0$ (*homoskedasticity*)
- $H_1 : H_0$ is not true (*heteroskedasticity*)

In case we *reject* the White test's null hypothesis, we violate CLRM Assumption V and there is evidence of heteroskedasticity.

Dealing with heteroskedasticity: Robust standard errors

Since heteroskedasticity causes problems to our coefficients' *standard errors*, we must be able to improve their estimation, while maintaining the unbiasedness of our $\hat{\beta}$s. Given that inference is a crucial component of any econometric estimation, we would like to have more reliable standard errors in the presence of heteroskedasticity.

For a given multiple regression model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + u_i$$

We compute the standard errors of a $\hat{\beta}_i$ coefficient in the following way:

$$SE(\hat{\beta}_j) = \frac{RSS/(n-k-1)}{\sqrt{TSS_j(1-R_j^2)}}$$

where *RSS* is the residual sum of squares of the original regression model, and TSS_j and R_j^2 are the total sum of squares and coefficient of determination from a regression of x_j on all the other independent variables.

Beyond its ugliness, the main problem with this estimator is that it does not work properly when the model presents heteroskedasticity. Thus, we need *heteroskedasticity-robust* procedures to perform proper inference when Assumption V is violated. For the same regression model above, if we calculate the *variance* of $\hat{\beta}_j$ in the following way,

$$Var(\hat{\beta}_i) = \sum_{i=1}^{n} \hat{r}_{ij}^2 \hat{u}_i^2 \Big/ RSS_j^2$$

its *square root* is called the **heteroskedasticity-robust standard error for** $\hat{\beta}_j$. From the formula above, $\hat{r}_{ij}$ is the i^{th} residual from regressing x_j on all the other independent variables, RSS_j is the residual sum of squares from this regression, and $\hat{u}_i$ is the estimated residual from the original regression model.

You don't need to wrap your head too much around these formulas, since we will not compute these values manually. What is important here is to get the idea behind this procedure: since we cannot trust in the standard errors when a regression model presents heteroskedasticity, we address this issue by using *robust standard errors*. The SEs derived from the last formula are known as the **Eicker-Huber-White standard errors**. A correction for degrees-of-freedom was later suggested to these standard errors by MacKinnon and White (1985)[3]. Since both of these procedures are consistent, in the sense to have adequate large-sample properties, no form is preferred over the other.

[3] MacKinnon, J. G., and White, H. (1985). *Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties*. **Journal of Econometrics**, 29(3), 305–325.

Let us apply these procedures to an actual example. Consider the following estimated model for wages (in logs), controlling for education, experience, tenure, marital status, and gender:

$$\widehat{\log(wage)}_i = .321 + .213\ married_i - 1.98\ married_i \cdot female_i - .11\ female_i + .0789\ educ_i + 0269\ exper_i - .00054 exper_i^2 + 0.291\ tenure_i - .00053 tenure_i^2$$

The next table summarizes the 3 kinds of standard errors for each slope coefficient. Compare and contrast the ones from the original regression, along with Eicker-Huber-White and MacKinnon-White standard errors. Notice that both robust procedures produce similar values. Furthermore, the largest relative change occurs to the SE on $educ_i$: from .0067 in the original to .0074 with MacKinnon-White SEs. This model,

however, does not have heteroskedasticity, according to the BP and White tests. Let us next look at an example with non-constant residual variance.

Variable	Original SEs	Eicker-Huber-White	MacKinnon-White
married	.055	.057	.057
$\texttt{married} \cdot \texttt{female}$	.058	.072	.058
female	.056	.057	.057
educ	.0067	.007	.0074
exper	.0055	.0051	.0051
$\texttt{exper}^2$	.00011	.0001	.00011
tenure	.0068	.0069	.0069
$\texttt{tenure}^2$	.00023	.00024	.00024

Recall the *earnings-education* model analyzed before. Let us estimate it via OLS:

$$\widehat{\texttt{earnings}}_i = \underset{(0.959)}{-3.13} + \underset{(0.069)}{1.47}\ \texttt{education}_i$$

$$n = 2,950 \qquad \bar{R}^2 = .13$$

When we plotted the two variables together, we had good reasons to believe that heteroskedasticity is present. Indeed, both Breusch-Pagan and White tests *reject* the null hypothesis of homoskedasticity, thus violating Assumption V.

Let us next re-estimate this model, this time using MacKinnon-White robust standard errors:

$$\widehat{\texttt{earnings}}_i = \underset{(0.926)}{-3.13} + \underset{(0.072)}{1.47}\ \texttt{education}_i$$

Notice how the original model was *underestimating* the slope coefficients' standard errors. In other words, it was 0.069 in the original, and it becomes 0.072 when correcting for heteroskedastic residuals. Now, we can properly perform inference for this

model. The coefficient on *education* is statistically significant at $\alpha = 1\%$, which we could not confirm with more certainty without robust standard errors.

What happens if we *log-transform* the dependent variable from this model? Let's see:

$$\widehat{\log(earnings)}_i = \underset{(0.051)}{1.48} + \underset{(0.0037)}{0.088}\ education_i$$

Are our standard errors reliable for inference? BP and White tests for heteroskedasticity on this model *do not reject* the null hypothesis of homoskedasticity. But what just happened here? The main reason for this model not having heteroskedasticity is that log-transforming a dependent variable tends to *decrease its variance*, thus leading to a decreased distance between the data points and the regression line.

Therefore, *when the case allows* for it, log-transforming the dependent variable not only brings interpretation benefits, but also decreases the likelihood of heteroskedasticity.

We will apply these procedures addressing heteroskedasticity in our applied lecture. After that, dealing with this violation will be made very accessible.