[STAT-25] Heteroskedasticity

Miguel-Angel Canela

Associate Professor, IESE Business School

Heteroskedasticity

Homoskedasticity is one of the assumptions of the classical linear regression model. It means that the error variance is independent of the actual value of the X's. **Heteroskedasticity** is a generic term for the failure of homoskedasticity, used to indicate that the error variance depends on the X values. In most cases, we deal with heteroskedasticity of unknown form, that is, we have no previous information on the form of this dependence.

I discuss in this lecture two topics related to heteroskedasticity: (a) inference after OLS estimation in the presence of heteroskedasticity and (b) testing homoskedasticity. Heteroskedasticity already appeared in our discussion of the variants of the two-sample t test in lecture STAT-18. In the most popular version of the classical t test, subpopulation variances are assumed to be equal (this is homoskedasticity), but there is an alternative version in which this condition is relaxed. This alternative is a **heteroskedasticity-robust method**. There is a companion test on the equality of the subpopulation variances, the F test (there are, in fact, many tests available for this purpose), which may be seen as a test on homoskedasticity. Extensions of both methods to one-way ANOVA can be found in many textbooks oriented to applications in engineering, biology, etc, but, in Econometrics textbooks, these issues are always discussed in the linear regression context.

To avoid confusion about the homoskedasticity assumptions, some remarks may be worth:

- Homoskedasticity is not needed for the OLS estimators $\hat{\beta}_0$, $\hat{\beta}_1$, ..., $\hat{\beta}_k$ to be unbiased and consistent.
- In lecture STAT-19, I used the homoskedasticity assumption to derive the formula $\hat{\text{cov}}[\hat{\boldsymbol{\beta}}] = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}s^2$. Under heteroskedasticity, this formula is no longer valid. In particular, the standard errors reported by stat packages are not correct.
- In the presence of heteroskedasticity, the distribution of the t statistics associated to the regression coefficients are no longer Student's t. The same for the F statistics used in testing linear restrictions (lecture STAT-21).
- Without homoskedasticity, Gauss-Markov theorem (lecture STAT-19) fails. The OLS estimator is no longer the BLUE, and there are more efficient linear estimators. Nevertheless, this extra efficiency does not matter for big samples.

Heteroskedasticity-adjusted standard errors

Heteroskedasticity-robust estimators appear in Econometrics textbooks as White, Huber and/or Eicker estimators. Robust estimation does not affect the parameter estimates, only the standard errors. Standard errors obtained by robust estimation are frequently said to be "adjusted".

Some mathematical detail on the standard errors of regression coefficients can help to understand how robust estimation works. My presentation is restricted to simple regression. The extension to multiple regression involves some matrix algebra. The OLS estimator of the regression slope can be rewritten as

$$\hat{\beta}_1 = \beta_1 + \frac{\sum(x_i - \bar{x})e_i}{\sum(x_i - \bar{x})^2}\,.$$

By the linearity assumption of the classical linear regression model, the expected value of e_i is zero, hence this estimator is unbiased. Also, denoting the error variance at $X = x_i$ by σ_i^2, we get

$$\operatorname{var}\big[\hat{\beta}_1\big] = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{\big[\sum (x_j - \bar{x})^2\big]^2}\, \sigma_i^2,$$

which shows the variance of $\hat{\beta}_1$ as proportional to a weighted average of the σ_i^2. Putting $\sigma_i^2 = \sigma^2$ for all i, we get the formula of lecture STAT-19, valid under homoskedasticity. For a heteroskedasticity adjustment, we can use e_i^2 as an estimate of σ_i^2, which leads to the formula of the **White estimator**

$$\hat{\operatorname{var}}[\hat{\beta}_1] = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{\big[\sum (x_j - \bar{x})^2\big]^2}\, e_i^2.$$

This formula has many variants, not discussed here (see the example for one of them). A final word of caution. The jargon used in this context (e.g. "the standard errors are adjusted") may lead you to assume that robust estimation always results in higher standard errors for everything. It is not so. In regression analysis, we deal with a vector of parameters, and the loss efficiency of the OLS method refers to a covariance matrix, not to the standard errors of the individual coefficients. Although there is, on the average, a loss of significance when turning to robust estimation, a specific coefficient may improve its significance.

Testing homoskedasticity

I introduce in this section a method for testing homoskedasticity. My presentation is limited to the **Breusch-Pagan test**, as presented in Wooldridge textbook. Let us suppose that the error variance is a function of $X_1, \ldots, X_k$. All heteroskedasticity tests assume that this function has some particular form. The Breusch-Pagan test takes the simplest approach, assuming a linear form for that function, performing a regression of the squared residuals on the X's and testing the coefficients.

Since the expected value of ϵ^2, given the X's, is the error variance, the test assume a linear model

$$\mathrm{E}[\epsilon^2\big] = \alpha_0 + \alpha_1 X_1 + \cdots + \alpha_k X_k,$$

and the homoskedasticity null is rewritten as $H_0 : \alpha_1 = \cdots = \alpha_k = 0$. A rough approximation to the results of this test is given by the statistics that come with the regression of the squared residuals on the independent variables.

Example 1. The `hprice1` data set, taken from Wooldridge's data collection, uses cross-sectional data, with sample size 88. The objective of the analysis is to develop a housing price equation that allows a rough prediction of the price from the other three variables:

- `price`, house price, in thousands US dollars.
- `bdrms`, number of bedrooms.
- `lotsize`, size of the lot, in square feet.
- `sqrft`, size of the house, in square feet.

It would not be surprising to find bigger prediction errors in big (and hence expensive) houses. So, we have a case of potential heteroskedasticity. Since I have a small sample, the significance of the effects estimated from these data is a sensible issue.

I regress the price on the three explanatory variables, using first the default method (Table 1) and then recalculating the standard errors (Table 2), iusing one of the many available methods. More specifically, I use here the default of the option `robust` of the Stata command `regress`, which

is the one that is more likely to be found in papers. In `R`, this variant is available through the combination of the packages `sandwich` and `lmtest`.

TABLE 1. Linear regression results (Example 1)

Coefficient	Estimate	Std. error	t value	p-value
Intercept	–21.77	29.48	–0.74	0.462
bdrms	13.85	9.01	1.54	0.128
lotsize	0.0021	0.0006	3.22	0.002
sqrft	0.123	0.013	9.28	0.000

The coefficient estimates and the R-squared statistic ($R^2 = 0.672$) are the same. The changes in the regression output are that there is no ANOVA table and that there is, overall, a loss of significance, due to lower efficiency in the estimation. This is seen in the F statistic. Nevertheless, it is not true, in general, that all the standard errors of the coefficients increase.

TABLE 2. Linear regression results (heteroskedasticity-adjusted)

Coefficient	Estimate	Std. error	t value	p-value
Intercept	–21.77	37.14	–0.59	0.559
bdrms	13.85	8.48	1.63	0.106
lotsize	0.0021	0.0012	1.65	0.102
sqrft	0.123	0.018	6.93	0.000

Heteroskedasticity can be explored directly, using the squared residuals as a proxy of the conditional variance. Then, the correlation of the squared residuals with the regressors, or better,the regression of the squared residuals on the regressors, can be informative (Table 3). Here, the lotsize seems to play an important role.

TABLE 3. Regression results (squared residuals)

Coefficient	Estimate	Std. error	t value	p-value
Intercept	–5522.8	3259.5	–1.69	0.094
bdrms	1041.76	996.38	1.05	0.299
lotsize	0.2015	0.0710	2.84	0.006
sqrft	1.691	1.464	1.16	0.251

Hetereroskedasticity can be confirmed with a number of tests. In practical terms, the null in these tests is that the error variance is the same irrespective of the value of the X's. The most popular version of the Breusch-Pagan test (see Wooldridge's textbook, chapter 8) uses a chi square statistic, which in this case gives us BP = 14.09. The number of degrees of freedom is the number of regressors (3). This gives $P = 0.003$, supporting my guess about heteroskedasticity.

An alternative approach is to try a transformation that waters down the heteroskedasticity. In this example, log transforming the dependent variable is enough to fix the problem. Nevertheless, the regression coefficients have to be interpreted then in percentage terms.

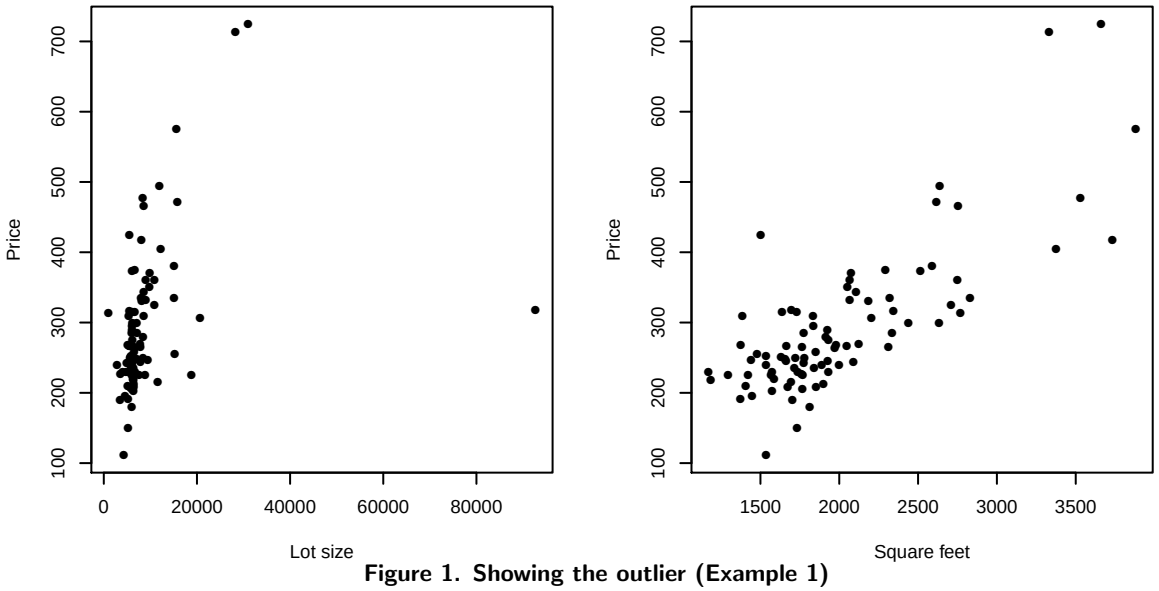

Figure 1. Showing the outlier (Example 1)

TABLE 4. Regression results (log scale)

Coefficient	Estimate	Std. error	t value	p-value
Intercept	4.759	0.0935	50.9	0.000
bdrms	0.025	0.029	0.88	0.380
lotsize	5.60e-06	2.04e-06	2.75	0.007
sqrft	0.0004	0.00004	8.67	0.000

Now the Breusch-Pagan statistic is BP = 0.79 ($P = 0.373$). A third approach to the analysis shows that the heterogeneity test just detected something in the data which does not agree with the assumptions of the model. Indeed, Figure 1 shows that there is one house with an enormous lot but a moderate price. If you drop this observation, the heteroskedasticity issue disappears, and the lot size term becomes significant.

Homework

A. Over the last decade, several large-scale cross-cultural studies have focused on well-being in a wide range of nations and cultures, but, in general, Latin countries have only been sporadically represented in these studies. The data for this example ($n = 819$) were collected on a sample of managers following a part-time MBA program at business schools in nine Latin countries. Test the differences between the nine countries, taking care of the uniformity of the variance.

¶ Source: S Poelmans & MA Canela, Statistical analysis of the results of a nine-country study of Latin managers, XIth European Congress on Work and Organizational Psychology (Lisboa, 2003).