Violations of OLS Assumptions: Heteroscedasticity

Econometric Methods, ECON 370

One of the critical assumptions in OLS is that we assume that the distribution of the error terms are the same (homoskedastic), and are independent (serially uncorrelated). We will deal with the latter later if necessary, but for now, we would like the ask the question what if the former is violated. What are its causes and consequences? How do we test for it? Is there a way we can accomodate it?

1 Heteroskedasticity

If you recall, the basic OLS model

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + ... + \beta_k x_{k,i} + \epsilon$$

The homoskedasticity assumption implies that

$$Var(\epsilon_i | x_{1,i}, x_{2,i}, ..., x_{k,i}) = \sigma^2, \forall i \in 1, 2, ..., n$$

Which essentially says that the error term for each observation are the same for all observations. When we are dealing with individuals in the general populace, this is hardly true, although on average it might be. When this assumption is violated, we have

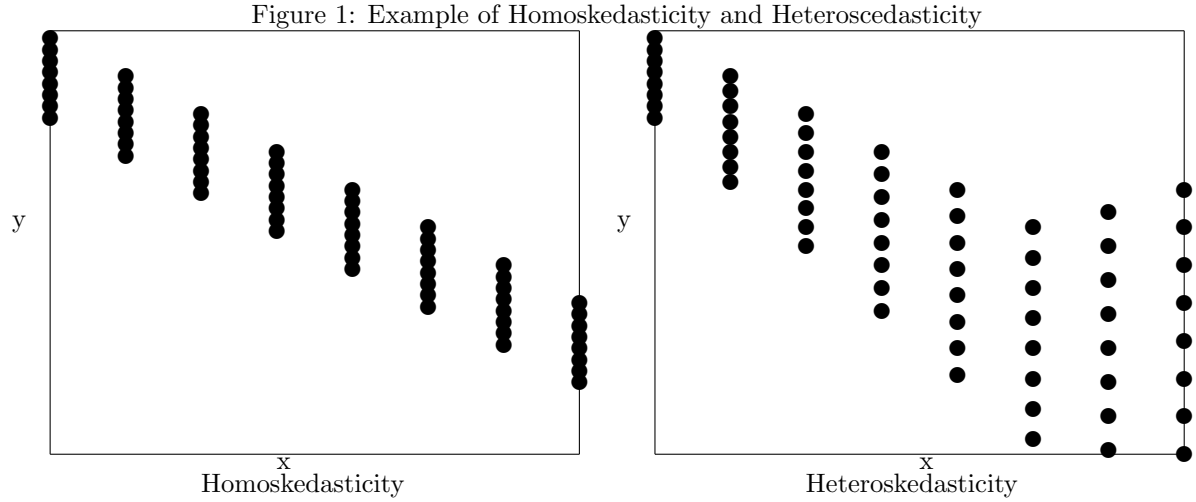
$$Var(\epsilon_i | x_{1,i}, x_{2,i}, ..., x_{k,i}) = \sigma_i^2, \forall i \in 1, 2, ..., n$$

which essentially says that the variance for each observation could be different. Visually, what this means is that if we were to plot the dependent and independent variables on the same diagram, the dispersion of the observations would not be the same across each value of the independent variable, such as depicted below. Although the diagram depicts the relationship as decreasing in the independent variable, and that the degree of dispersion is increasing with the value of the independent variable, this not the only type of case, since there is an infinite amount of permutation. The crux of the matter is that dispersion varies.

Why do they occur? When are they most common? Heteroskedasticity are most common in micro-econometrics where the observations are individuals or households. Principally, there is no reason to believe that families or individuals have the same background. If so, the behavioral inclinations are all alittle different. When performing regressions typically, what we are trying to do is to explain occurrences with one equation. If our observations are essentially different a priori, there is very little reason that they should behave the same way, and consequently heteroskedasticity.

However, that is not to say that this only occurs in micro studies. Consider cross country, cross province, cross city, cross firm, cross industry and cross town comparisons. That is even when we are more concerned with a macro variable, it is still possible that homoskedasticity may be violated (and most likely is). In short it is most common in cross sectional data studies.

Figure 1: Example of Homoskedasticity and Heteroscedasticity



2 Consequences of Heteroskedasticity for OLS

The interesting and fortuitous thing is that even with heteroskedasticity, our estimators are unbiased. The unfortunate thing however is that our variance are incorrect, and we lose the B.L.U.E. feature. Consequently, our inference stand to be incorrect.

To see that the OLS estimator is unbiased, first recall that the formula for a simple OLS

$$\widehat{\beta}_1 = \sum(\frac{(x_i - \overline{x})y_i}{\sum(x_i - \overline{x})^2})$$

Next let us treat x_i as a random variable. Then taking expectations,

$$E(\widehat{\beta}_1) = E(\sum(\frac{(x_i - \overline{x})y_i}{\sum(x_i - \overline{x})^2}))$$

$$= \sum E(\frac{(x_i - \overline{x})(\beta_0 + \beta_1 x_i + \epsilon_i)}{\sum(x_i - \overline{x})^2})$$

$$= \beta_0 \sum E\frac{(x_i - \overline{x})}{\sum(x_i - \overline{x})^2} + \beta_1 \sum E\frac{x_i(x_i - \overline{x})}{\sum(x_i - \overline{x})^2} + \sum E\frac{\epsilon_i(x_i - \overline{x})}{\sum(x_i - \overline{x})^2}$$

$$= \beta_0 E\frac{\sum(x_i - \overline{x})}{\sum(x_i - \overline{x})^2} + \beta_1 E\frac{\sum x_i(x_i - \overline{x})}{\sum(x_i - \overline{x})^2} + E\frac{\sum \epsilon_i(x_i - \overline{x})}{\sum(x_i - \overline{x})^2}$$

$$= \beta_1 + \sum E\frac{\epsilon_i(x_i - \overline{x})}{\sum(x_i - \overline{x})^2} = \beta_1$$

You should recall that the first term on the right hand side is zero since:

$$\sum(x_i - \overline{x}) = \sum x_i - n\overline{x} = \sum x_i - \sum x_i = 0$$

While for the second term

$$\sum x_i(x_i - \overline{x}) = \sum x_i^2 - \overline{x}\sum x_i = \sum x_i^2 - n(\overline{x})^2 = \sum (x_i - \overline{x})^2$$

And for the final term,

$$\sum \epsilon_i x_i - \overline{x}\sum \epsilon_i$$

And since $E(\epsilon_i) = 0$, this then means that the second term is zero. Next $E(\epsilon_i x_i) = 0$. Then the expected value of the above term is zero, and our estimator is unbiased. Essentially, the OLS estimator is unbiased because the variance of the error term does not enter its calculations.

To show now that our estimator is no longer B.L.U.E. again note that;

$$Var(\widehat{\beta}_1) = Var(\sum(\frac{(x_i - \overline{x})y_i}{\sum(x_i - \overline{x})^2}))$$

$$\Rightarrow Var(\widehat{\beta}_1) = \sum \frac{(x_i - \overline{x})^2}{(\sum(x_i - \overline{x})^2)^2} Var(y_i) = \sum \frac{(x_i - \overline{x})^2}{(\sum(x_i - \overline{x})^2)^2}\sigma_i^2 \neq \frac{\sigma^2}{\sum(x_i - \overline{x})^2}$$

Note that in general, it is not possible to say whether the OLS estimated variance would be greater or less than the true parameter variance.

3 Robust Inference after OLS

Given the above the next question is how can we estimate the variance of the parameters with heteroskedasticity? Let $\widehat{\epsilon_i}$ be the OLS residuals from the regression. We can estimate the new variance by the following formula,

$$\frac{\sum(x_i - \overline{x})^2\widehat{u}_i^2}{(\sum(x_i - \overline{x})^2)^2}$$

This is typically referred to as the the Robust Standard Errors, and would work even when we have homoskedasticity. Note that this is the formula for a simple regression. In the general case of a multiple variable regression, the robust standard errors for $\widehat{\beta}_j$ for $j \in 1, 2, ..., k$ in a k variable regression,

$$Var(\widehat{\beta}_j) = \frac{\sum \widehat{r}_{j,i}^2\widehat{\epsilon}_i^2}{(\sum(\widehat{r}_{j,i} - \overline{\widehat{r}}_j)^2)^2}$$

where you should recall that as in our discussion of the multiple variable regression $\widehat{r}_{j,i}$ is the i^th residual for the regression of x_j on all the other independent variables. The above is just the more general form of the **Heteroskedasticity-Robust Variance** for β_j. This is not the only form of a robust variance.

You may also correct for degrees of freedom by multiplying the above formula by $\frac{n}{n-(k+1)}$. The rationale for this correction is because if indeed the errors are homoskedastic, we will still get the homoskedastic variance and consequently the homoskedastic standard errors. However, as a matter of practice, we take what most statistical packages provide. Further, note that once the robust standard errors has been calculated, the manner in which we perform all inferences, such as the calculation of

the t statistic remains the same. As a final note, in applied work, especially where we're dealing with cross sectional data, when the sample size is large, economists typically report the robust standard errors given the above corrections.

4 Testing for Heteroskedasticity

To rationalize the test for heteroskedasticity we note first that the homoskedasticity assumption in OLS implies

$$Var(\epsilon|x_1, x_2, ..., x_k) = \sigma^2$$

In that case, if we want to test for heteroskedasticity, our maintained assumption is that the errors are actually homoskedastic, and we wish to examine if that is true. That is the null hypothesis is just the above,

$$H_0 : Var(\epsilon|x_1, x_2, ..., x_k) = \sigma^2$$

Next, note that in examining heteroskedasticity, the expected value of the errors being zero is still maintained. Which means that

$$Var(\epsilon|x_1, x_2, ..., x_k) = E(\epsilon^2|x_1, x_2, ..., x_k) = \sigma^2$$

So that we can rewrite the hypothesis being test as

$$H_0 : E(\epsilon^2|x_1, x_2, ..., x_k) = \sigma^2$$

So that if we assume a simple linear relationship between ϵ with respect to the dependent variables, we could then test the hypothesis. To see this, consider a general k variable regression where the dependent variable is ϵ^2. Let e be the error term in the linear relationship, and assume that it is normal distributed with mean 0 given the independent variables. That is,

$$\epsilon^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + ... + e$$

If homoskedasticity holds, then we would have

$$\delta_1 = \delta_2 = ... = \delta_k$$

Recall the first diagram illustrating the idea behind heteroskedasticity, and that it essentially implies a relationship between the error term vis-a-vis independent variables. Therefore if heteroskedasticity does not exists, the null hypothesis of homoskedasticity, can be written as,

$$H_0 : \delta_1 = \delta_2 = ... = \delta_k$$

This implies that we could test the hypothesis using the F statistic that is provided in standard statistical software (even if you write your own program, the calculation of the F statistic is not difficult given that we have already found the formula earlier in our discussion of OLS).

Of course we do not observe the true population error term, nor could we get a sample. However, we could use the residuals from the original OLS regression of y against x_1, x_2,..., x_k, call it $\widehat{e^2}$. That is we perform the following regression, and calculate the F statistic there after.

$$\widehat{\epsilon}^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + ... + e$$

The F statistic is dependent on the goodness of fit measure from the above regression. Let that be $R^2_{\widehat{\epsilon}^2}$, then the statistic is,

$$F = \frac{\frac{R^2_{\widehat{\epsilon}^2}}{k}}{\frac{1-R^2_{\widehat{\epsilon}^2}}{n-(k+1)}}$$

And the statistic is approximately distributed as a $F_{k,n-(k+1)}$ under the null hypothesis.

Breusch-Pagan Test There is another procedure that uses a rather easy statistic that is also dependent on the goodness of fit measure, call the (LM) **Lagrange Multiplier** statistic. This test based using the LM statistic is known as the Breusch-Pagan Test for Heteroskedasticity. The statistic is,

$$LM = n \times R^2_{\widehat{\epsilon}^2}$$

The LM statistic is distributed asymptotically as χ^2_k.

The standard version of the statistic when we assume homoskedasticity for calculating the LM statistic is in page 185-187 of your text. You should read it for your reference.

The procedure for the Breusch-Pagan Test for Heteroskedasticity is as follows,

1. Estimate the standard OLS of your desire,

   $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \epsilon$$

   and obtain $\widehat{\epsilon}^2$.

2. Estimate the following next,

   $$\widehat{\epsilon}^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + ... + e$$

   and keep $R^2_{\widehat{\epsilon}^2}$.

3. Calculate the LM statistic, and the p-value using the χ^2_k distribution. As usual, if the p-value is small, you reject H_0, or homoskedasticity.

Of course it need not be true that the errors are dependent on the full set of independent variables. If we suspect that it is dependent on only a subset of the independent variables, we can and should modify the Breusch-Pagan such that in the second stage, we run the regression on those suspected independent variables. Note that the degrees of freedom of the distribution has to change to the

number of independent variables used. And if the second regression performed has only one variable, it is equivalent to us simply checking for the significance of the coefficient for that variable, i.e. all we need to do is to do a t test.

As a final note, realize also that in fact the functional form need not be linear, that is if you believe that the heteroskedasticity has a particular pattern either by a priori expectations, or theory, you could and should alter the regression at the second step. An example say if you suspect that the relationship is concave in either one or a few of the independent variables.

White Test

The typical homoskedasticity assumption in truth may be replaced with a weaker assumption in the typical OLS, that the error term ϵ be uncorrelated with the independent variables, x_j, the squares of the independent variables, x_j^2, and the cross products of the indepedent variables, $x_j x_h$, where $j \neq h$, and j, $h \in \{1, 2, ..., k\}$. This led to the White Test by White (of course!) which suggested the inclusion of all the above as covariates in the second step regression. That is,

$$\widehat{\epsilon}^2 = \delta_0 + \delta_1 x_1 + ... + \delta_k x_k + \delta_{k+1} x_1^2 + \delta_{k+2} x_1 x_2 + ... + e$$

The White test for heteroskedasticity then test whether all the above parameters are all equal to 0. However, as is easy to see, the increase in the number of covariates is a weakness of the White Test since it uses too many degrees of freedom even at relatively small OLS models.

However, there is a method by which we could conserve the degrees of freedom, yet utilizing the idea of the White Test that we want to exhaust the various functional forms possible. The idea is derived by observing the original regression model. If we were to square the entire OLS regression model, we would get the cross products we had wanted. That is we could perform the regression in the second stage as,

$$\widehat{\epsilon}^2 = \delta_0 + \delta_1 \widehat{y} + \delta_2 \widehat{y}^2 + e$$

Where $\widehat{y}$ is just the predicted value of y, the dependent variable. Why? What are we including in the right hand side of the regression model above if we use y? We can again use either the F or LM statistic to test the following hypothesis for homoskedasticity,

$$H_0 : \delta_1 = \delta_2 = 0$$

Notice that this test greatly reduces the degrees of freedom burden of the original test. The suggested test is then a special case of the White Test.

Finally, it should be mentioned that the above tests work if sole problem in our regression model is derived from heteroskedasticity. If our model is of the wrong functional form, called misspecification error, our inference could be incorrect. That is, it is possible that under the wrong functional form, we actually have homoskedasticity, and yet, the tests would suggest that we should reject the null hypothesis. In general, because specification is a more important concern, we should perform a misspecification test first, before we test for heteroskedasticity.