Lecture 17. Heteroskedasticity

17.1. Introduction

In Lectures 14 and 15, we considered the estimation of reduced form equations by OLS. In Lecture 16, we considered cases in which the disturbances are autocorrelated ie. $E[u_t u_s] \neq 0$ for some $t \neq s$. ***Autocorrelation is a problem that is typically encountered when analysing time series data.***

In this lecture, we consider violations of the *homoskedasticity* (ie. 'constant-variance') assumption of the classical model, which says that $V[u_t] = E[u_t^2] = \sigma_u^2$ for all t (see page 6 of the handout for Lecture 14). When disturbances pertaining to different observations have different variances, the disturbances are said to be *heteroskedastic*. ***Heteroskedasticity is a problem that is typically encountered when analysing cross-section data.*** Hence, in this lecture, we will use the subscript 'i' rather than the subscript 't' to denote particular observations. The material is organised as follows:

Section 17.2. The causes of heteroskedasticity
Section 17.3. The consequences of heteroskedasticity for the OLS estimator
Section 17.4. Representations of heteroskedasticity
Section 17.5. Testing for heteroskedasticity
Section 17.6. Estimation in the presence of heteroskedasticity using the method of Weighted Least Squares (WLS)
Section 17.7. An interpretation of heteroskedasticity as 'equation misspecification' or 'variation in the coefficients'.

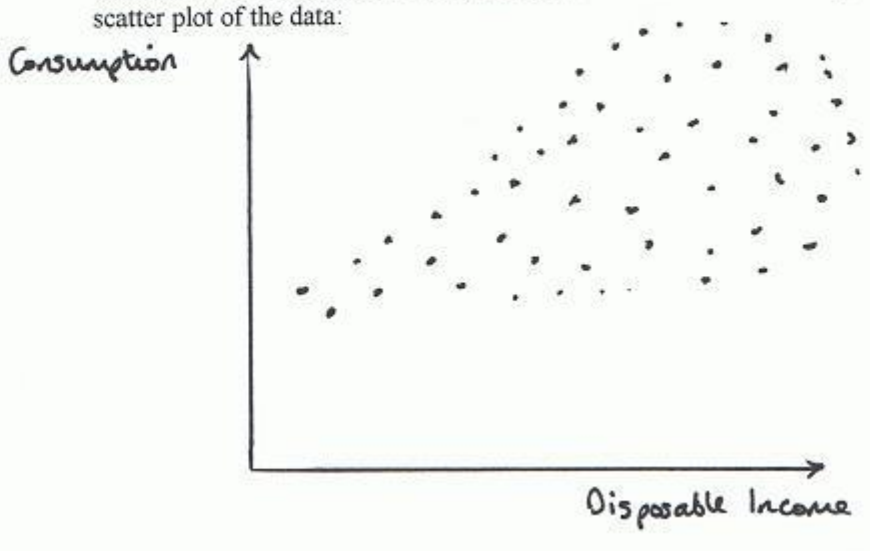
You should note that the term 'heteroskedasticity' is frequently written with a 'c' instead of a 'k' as *heteroscedasticity*.

17.2. The causes of heteroskedasticity

There are two main causes of heteroskedasticity which you should be aware of. Firstly, heteroskedasticity arises when the variance of the dependent variable (which is the same as the variance of the disturbance term) itself depends upon one or more independent variables. As a classic illustration of this, consider the following relationship between consumption and disposable income across a cross-section of families:

$$C_i = \alpha + \beta Y_i + u_i \qquad i = 1, \ldots, n$$

Those with higher incomes may have more flexibility over the choice of consumption level than those on lower incomes. Thus, we may find that the variance of consumption (or equivalently, the variance of the disturbance term) increases with income. This might give rise to the following scatter plot of the data:



Along the same lines, the variation in investment spending among large firms may be greater than among small firms, so the variances of the disturbances in a model of investment expenditure may themselves depend upon firm size. There are many other possible illustrations of a similar nature.

Secondly, heteroskedasticity can arise as a result of misspecification of an equation. For example, the omission of a relevant explanatory variable will cause the errors to be heteroskedastic if that variable itself exhibits non-constant variance. In this case, the correct thing to do is to respecify the equation, rather than to try to account for heteroskedastic errors using the estimation procedures discussed below. As we shall see in Section 17.7, heteroskedasticity may also arise if one or more of the *coefficients* in an equation are random variables (rather than constants).

17.3. The consequences of heteroskedasticity for the OLS estimator

The consequences of heteroskedasticity are essentially the same as those of autocorrelation - they are both violations of the assumption that the variance-covariance matrix of the errors can be written as $E[\mathbf{uu}'] = \sigma_u^2\mathbf{I}$, where **I** is the (n×n) identity matrix (see Lecture 15, page 5). The OLS coefficient estimates are still both unbiased and consistent (the assumption of a constant error variance was not used in the proof of either - see Lecture 14). However, they are inefficient ie. they no longer have the property of minimum variance, so that it is possible to obtain more reliable estimates (we will see how later).

Finally, the OLS estimates of the variances of the coefficients will be biased, so the t-ratios which make use of these expressions will also be biased. To clarify the nature of the bias in the t-ratios, consider a cross-section bivariate regression model involving a single explanatory variable:

$$Y_i = \alpha + \beta X_i + u_i \qquad i = 1, \ldots, n$$

Suppose that the disturbances are heteroskedastic, so that the variance of each u_i depends upon i: $E[u_i^2] = \sigma_i^2$. If the derivation of the formula for $V[\hat{\beta}]$ in Lecture 14 (page 9) is carefully considered, it can be shown that the direction of the bias in the OLS estimate of the variance of $\hat{\beta}$ depends on the direction of the association between $(X_i - \overline{X})^2$ and σ_i^2. As an exercise for this lecture, you are asked to show that if $(X_i - \overline{X})^2$ and σ_i^2 are positively related, then the OLS estimate of the variance of $\hat{\beta}$ underestimates the true one. ***Hence, the calculated t-ratio will be an overestimate, and we may be misled into thinking that a variable is significant when in reality it is not.*** If, on the other hand, $(X_i - \overline{X})^2$ and σ_i^2 are negatively related, the OLS estimate of the variance of $\hat{\beta}$ overestimates the true one. ***Hence, the calculated t-ratio will be an underestimate, and we may be misled into thinking that a variable is insignificant when in reality it is significant.***

17.4. Representations of heteroskedasticity

When considering autocorrelation in Lecture 16, we had to assume that the errors were generated by an AR(1) or some other process in order to proceed with our analysis. We need to do the same sort of thing here. We have to assume some *form* for the way in which σ_i^2 varies with i in order to consider efficient estimates, tests against particular alternatives, etc. The most common approach is to formulate the possible heteroskedasticity as

$$\sigma_i^2 = f(\alpha_1 + \alpha_2 Z_{2i} + \alpha_3 Z_{3i} + \cdots + \alpha_m Z_{mi})$$

where f is some function which is assumed to be independent of i, the α's are unknown parameters, and the Z_j's are observed variables which may or may not include the explanatory variables in the original equation. As a simple example, consider the following specification, which we will use later to discuss estimation in the presence of heteroskedasticity by Weighted Least Squares:

$$\sigma_i^2 = \alpha_1 + \alpha_2 Z_i$$

A null hypothesis of homoskedastic errors can be specified in terms of parametric restrictions on these specifications eg. H_0: $\alpha_2 = 0$ in the context of the simple model $\sigma_i^2 = \alpha_1 + \alpha_2 Z_i$.

17.5. <u>Testing for heteroskedasticity</u>

There are many tests available for examinination of the possibility of heteroskedastic errors, but we will only consider one called the *Breusch-Pagan* test, which is easy to carry out, and which encapsulates the basic ideas underlying most of the other tests. The Breusch-Pagan test formulates the possible heteroskedasticity as

$$\sigma_i^2 = f(\alpha_1 + \alpha_2 Z_{2i} + \alpha_3 Z_{3i} + \cdots + \alpha_m Z_{mi})$$

This is defined in the same way as in Section 17.4, except that the Z_j's are assumed to be <u>exogenous</u>, along with all the explanatory variables in the original equation (any of which can also appear as Z's). Note that the function f does not need to be specified. The null hypothesis of homoskedasticity can be written as

$$H_0: \alpha_2 = \alpha_3 = \cdots = \alpha_m = 0$$

since, in this case, $\sigma_i^2 = f(\alpha_1)$ which is constant. The test procedure is as follows:

(I). Estimate the original equation by OLS and calculate the residuals $e_1, e_2, \ldots, e_n$.

(II). Construct the variable g defined by

$$g_i = \frac{ne_i^2}{\sum_{i=1}^{n} e_i^2} \qquad i = 1, \ldots, n$$

(III). Regress g on the Z variables (including an intercept term).

(IV). Construct the test-tatistic

B = 0.5(the explained sum of squares from this regression).

<u>Under the null hypothesis</u> of homoskedastic errors, B has (asymptotically) a chi-square distribution with (m-1) degrees of freedom. Hence, to test the null hypothesis, we simply compare B with the appropriate critical point from the chi-square table.

Other tests, which you might like to read about for yourself, include the famous *Goldfeld-Quandt* test, and the *Glesjer* test.

17.6. <u>Estimation in the presence of heteroskedasticity using the method of Weighted Least Squares (WLS)</u>

Consider the following cross-section bivariate regression model:

$$Y_i = \alpha + \beta X_i + u_i \qquad i = 1, \ldots, n$$

The assumption of homoskedasticity in the context of this model can be written as $V[u_i] = E[u_i^2] = \sigma^2$ (constant for all i). Given any pair of observations in the data set, say (X_1, Y_1) and (X_2, Y_2), the assumption of homoskedasticity says that $V[u_1] = V[u_2]$, so that the two observations are *equally reliable* ie. each observation is just as likely as the other to be 'near' to the line $Y = \alpha + \beta X$. If it were the case that $V[u_1] < V[u_2]$, say, then the (X_1, Y_1) would be more likely to lie 'near' to the line $Y = \alpha + \beta X$ than (X_2, Y_2). ***Since (X_1, Y_1) is more reliable than (X_2, Y_2) in this case, we would want to put more 'emphasis' on (X_1, Y_1) in the estimation procedure.***

This intuitively appealing idea underlies the statistical method known as Weighted Least Squares (WLS), which can be used to estimate regression models in the presence of heteroskedasticity. Basically, WLS involves applying OLS to 'transformed' variables, where the transformation procedure attaches a greater weight to the more reliable observations, and a lower weight to the less reliable observations. Typically, the 'weight' used is an estimate of the inverse of the standard deviation.

To illustrate, suppose we want to estimate the parameters of the following linear equation:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i \qquad i = 1, \ldots, n$$

To keep things simple, we will assume that the heteroskedasticity can be characterised by the following relationship between the variance of u_i and a variable Z_i (recall that the formula for the variance of u_i is $E[u_i^2]$):

$$E[u_i^2] = \alpha_1 + \alpha_2 Z_i \qquad i = 1, \ldots, n$$

The variable Z may or may not be one of the explanatory variables in the original equation, but it must be exogenous. As we have already seen, the hypothesis of homoskedasticity can be written simply as H_0: $\alpha_2 = 0$. Given such a form for the mean of u_i^2, we might reasonably assume that

$$u_i^2 = \alpha_1 + \alpha_2 Z_i + v_i \qquad i = 1, \ldots, n$$

where v_i is an error term with $E[v_i] = 0$. ***If we replace the unobservable disturbances u_i by the residuals e_i, we can then estimate α_1 and α_2 by OLS. Hence, we can get estimates of the σ_i^2 and thus re-estimate the original equation taking account of the heteroskedasticity. This produces the WLS estimates. If the assumed form of the heteroskedasticity is correct, these estimates are efficient.*** The full procedure is as follows:

(I). Estimate the original equation by OLS, and calculate the residuals.

(II). Regress e_i^2 on Z_i to obtain OLS estimates of α_1 and α_2 and from these calculate

$$\hat{\sigma}_i = \sqrt{\hat{\alpha}_1 + \hat{\alpha}_2 Z_i} \qquad i = 1, \ldots, n$$

(III). Divide through the original equation by $\hat{\sigma}_i$, and estimate the transformed equation by OLS:

$$(Y_i/\hat{\sigma}_i) = \beta_1 + \beta_2(X_{2i}/\hat{\sigma}_i) + \cdots + \beta_k(X_{ki}/\hat{\sigma}_i) + (u_i/\hat{\sigma}_i) \qquad i = 1, \ldots, n$$

In this way, observations which are associated with a higher variance (and which are therefore less reliable) are given a lower weight, and observations which are associated with a lower variance are given a higher weight. Note that there are different procedures for different assumptions about the precise form of the heteroskedasticity, but the basic intuition is the same.

17.7. An interpretation of heteroskedasticity as 'equation misspecification' or 'variation in the coefficients'

It was pointed out in Section 17.2 that observed heteroskedasticity in the calculated residuals can be the result of misspecification of the equation and/or variation in the coefficients, as well as of heteroskedasticity in the unobservable disturbance term of the 'true' equation. To conclude this lecture, we will now look at this more closely.

Consider first the possibility that the equation is misspecified to the extent that a relevant explanatory variable has been left out. Call it W_t. Thus, the 'true' relationship might be given by

$$Y_t = \beta_1 + \beta_2 X_{2t} + \cdots + \beta_k X_{kt} + \gamma W_t + u_t$$

whilst we are estimating the equation

$$Y_t = \beta_1 + \beta_2 X_{2t} + \cdots + \beta_k X_{kt} + u_t^*$$

where $u_t^* = \gamma W_t + u_t$. Assuming that W_t is exogenous, we have $V[u_t^*] = \gamma^2 V[W_t] + \sigma_u^2$, which will in general not be constant. (Note: I have used the rule that, for any two random variables X and Y, $V[X + Y] = V[X] + V[Y] + 2Cov[X, Y]$. When W_t is exogenous, $Cov[W_t, u_t] = 0$, so we have $V[\gamma W_t + u_t] = V[\gamma W_t] + V[u_t] = \gamma^2 V[W_t] + \sigma_u^2$ as stated above). Thus, for example, if one of the relevant Z's in the Breusch-Pagan test is from outside the equation, one possible interpretation of a significant test statistic is that the variable should have been *in* the equation in the first place!

Consider next the possibility that one of the *coefficients* in the equation is not constant. Suppose, for example, that the 'true' relationship is given by

$$Y_t = \beta_1 + \beta_2 X_{2t} + \cdots + \beta_{kt} X_{kt} + u_t$$

where β_{kt} is composed of a fixed and a random component as follows:

$$\beta_{kt} = \beta_k + v_t$$

It is assumed that $E[v_t] = 0$. If we ignored the random variation in the coefficient of X_{kt}, we would effectively be estimating the equation

$$Y_t = \beta_1 + \beta_2 X_{2t} + \cdots + \beta_k X_{kt} + u_t^*$$

with $u_t^* = v_t X_{kt} + u_t$. Assuming that X_{kt} is exogenous and that u and v are not correlated, we would have $V[u_t^*] = V[v_t X_{kt} + u_t] = X_{kt}^2 \sigma_v^2 + \sigma_u^2$. This will clearly not be constant in general. Hence, when significant heteroskedasticity is discovered in a Breusch-Pagan test based on one of the explanatory variables, one possible explanation is that there is variation in the corresponding coefficient.

(End of Lecture 17)

A problem for Lecture 17. Heteroskedasticity

It was asserted in the text that if the disturbances are heteroskedastic in the context of a bivariate regression model of the form

$$Y_i = \alpha + \beta X_i + u_i \qquad i = 1, \ldots, n$$

then the direction of the bias in the OLS estimate of the variance of $\hat{\beta}$ depends on the direction of the association between $(X_i - \overline{X})^2$ and σ_i^2. If $(X_i - \overline{X})^2$ and σ_i^2 are positively related, then the OLS estimate of the variance of $\hat{\beta}$ underestimates the true one. If, on the other hand, $(X_i - \overline{X})^2$ and σ_i^2 are negatively related, the OLS estimate of the variance of $\hat{\beta}$ overestimates the true one. Prove this.

(Hint: It can be shown that if the disturbances are heteroskedastic, and the standard error of $\hat{\beta}$ *is computed as* $\mathrm{SE}(\hat{\beta}) = \sqrt{\dfrac{\mathrm{SSR}}{(n-2)\sum_{t=1}^{n}(X_t - \overline{X})^2}}$*, then*

$$E[\mathrm{SE}(\hat{\beta})^2] = \frac{\overline{\sigma}^2}{\sum_{t=1}^{n}(X_t - \overline{X})^2} - \frac{\sum_{t=1}^{n}(X_t - \overline{X})^2\theta_i}{(n-2)\left(\sum_{t=1}^{n}(X_t - \overline{X})^2\right)^2}$$

where $\overline{\sigma}^2 \equiv \frac{1}{n}\sum_{i=1}^{n}\sigma_i^2$ *and* $\theta_i = \sigma_i^2 - \overline{\sigma}^2$*. Compare this expected value with the expression you get for* $V[\hat{\beta}]$ *when you alter the derivation on page 9 in the handout for Lecture 14 by assuming that the disturbances are heteroskedastic).*

Solution to the problem for Lecture 17

In the derivation on page 9 of the handout for Lecture 14, we replace the subscript 't' by the subscript 'i', and assume that $E[u_i^2] = \sigma_i^2$. All the other assumptions stay the same. Following precisely the same steps, we obtain

$$V[\hat\beta] = \Sigma_{i=1}^{n} E[(w_i u_i)^2]$$

$$= \Sigma_{i=1}^{n} w_i^2 E[u_i^2]$$

$$= \Sigma_{i=1}^{n} w_i^2 \sigma_i^2$$

$$= \frac{\Sigma_{i=1}^{n} (x_i - \bar{x})^2 \sigma_i^2}{\left\{\Sigma_{i=1}^{n} (x_i - \bar{x})^2\right\}^2} = \frac{\bar\sigma^2}{\Sigma_{i=1}^{n} (x_i - \bar{x})^2} + \frac{\Sigma_{i=1}^{n} (x_i - \bar{x})^2 \theta_i}{\left\{\Sigma_{i=1}^{n} (x_i - \bar{x})^2\right\}^2}$$

where $\bar\sigma^2 \equiv \frac{1}{n} \Sigma_{i=1}^{n} \sigma_i^2$ and $\theta_i = \sigma_i^2 - \bar\sigma^2$. The result follows immediately by comparing this expression for $V[\hat\beta]$ with the expression for $E[SE(\hat\beta)^2]$ given in the Hint. We see that

$$E[SE(\hat\beta)^2] \neq V[\hat\beta]$$

Furthermore, the expression in the numerator of the second term of the right hand sides is $(n-1) \cdot Cov[(x_i - \bar{x})^2, \sigma_i^2]$. Therefore if $(x_i - \bar{x})^2$ and σ_i^2 are positively related, $E[SE(\hat\beta)^2] < V[\hat\beta]$. Conversely, if $(x_i - \bar{x})^2$ are negatively related, $E[SE(\hat\beta)^2] > V[\hat\beta]$.

QED.