EC203: Applied Econometrics
Heteroskedasticity (and clustering) in the errors: consequences, tests and solutions

Dr. Tom Martin

University of Warwick

Illustrative reading:

- Wooldridge: Chapter 8
- Dougherty: Chapter 7
- Gujarati: Chapter 5

Heteroskedasticity in the error term

Suppose we have our usual CLRM:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \epsilon_i$$

Suppose all of our CLRM assumption hold except for the constant variance assumption. That is, the assumption $Var(\epsilon|X_1, ..., X_k) = \sigma^2$ does not hold.

In words, the constant variance (homoskedastic) assumption requires the variance of the error, conditional on the explanatory variables, to be constant. Homoskedasticity will fail whenever there is variance in the errors across different segments of the population. When this assumption fails the errors are said to exhibit heteroskedasticity.

Heteroskedasticity in the error term

If homoskedasticity fails the consequences for OLS estimators are:

- The estimators are still unbiased.
- However, the variance formula for the OLS estimators is wrong.
- Therefore, the standard errors used in hypothesis tests and confidence intervals are no longer valid: t and F-statistics are wrong.

Heteroskedasticity: consequences

Let us show this explicitly for the SLR model:[1]

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

If $E[\epsilon] = 0, E[\epsilon|X] = 0, Var(\epsilon|X) = \sigma^2$ and $E(\epsilon_i, \epsilon_j|X) = 0$, then, we know:

$$E(b) = \beta \tag{1}$$

and

$$V(b) = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \tag{2}$$

Now we show if the variance is non-constant:

1. The estimator is still unbiased, as presented in equation (1).
2. The estimator no longer has the usual variance equation presented in (2).

[1] The arguments pass over to the MLR model.

Heteroskedasticity: consequences

To illustrate unbiasedness and variance we need to express b in terms of β:[2]

$$b = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$b = \frac{\sum_{i=1}^{n}(X_i - \bar{X})Y_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$b = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(\alpha + \beta X_i + \epsilon_i)}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

[2]This is a repeat from the SLR lecture. But it is an exercise worth repeating.

Heteroskedasticity: consequences

$$b = \alpha \frac{\sum_{i=1}^{n}(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} + \beta \frac{\sum_{i=1}^{n}(X_i - \bar{X})X_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2} + \frac{\sum_{i=1}^{n}(X_i - \bar{X})\epsilon_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$b = \beta + \frac{\sum_{i=1}^{n}(X_i - \bar{X})\epsilon_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$b = \beta + \sum_{i=1}^{n} \lambda_i \epsilon_i$$

Where $\lambda_i = \frac{(X_i - \bar{X})}{\sum_{i}^{n}(X_i - \bar{X})^2}$

Heteroskedasticity: consequences

To show the estimator b is unbiased:[3]

$$b = \beta + \sum_{i=1}^{n} \lambda_i \epsilon_i$$

$$E[b] = E[\beta + \sum_{i=1}^{n} \lambda_i \epsilon_i]$$

$$E[b] = \beta + E[\lambda_1 \epsilon_1 + \lambda_2 \epsilon_2 + ... + \lambda_n \epsilon_n]$$

$$E[b] = \beta + \lambda_1 E[\epsilon_1] + \lambda_2 E[\epsilon_2] + ... + \lambda_n E[\epsilon_n]$$

$$E[b] = \beta + \lambda_1 E[\epsilon_1|X] + \lambda_2 E[\epsilon_2|X] + ... + \lambda_n E[\epsilon_n|X]$$

$$E[b] = \beta$$

[3] In the final line we make the conditioning on X explicit.

Heteroskedasticity: consequences

To illustrate the consequences of heteroskedasticity for variance, note:

$$b = \beta + \sum_{i=1}^{n} \lambda_i \epsilon_i$$

$$V(b) = E[(b - E[b])^2] = E[(\sum_{i=1}^{n} \lambda_i \epsilon_i)^2]$$

$$V(b) = E[(\lambda_1 \epsilon_1 + \lambda_2 \epsilon_2 + ... + \lambda_n \epsilon_n)^2]$$

$$\begin{aligned} V(b) = E[&\lambda_1^2 \epsilon_1^2 + \lambda_2^2 \epsilon_2^2 + ... + \lambda_n^2 \epsilon_n^2 ... \\ &...2\lambda_1 \lambda_2 \epsilon_1 \epsilon_2 + 2\lambda_1 \lambda_3 \epsilon_1 \epsilon_3 + ... + 2\lambda_1 \lambda_n \epsilon_1 \epsilon_n + ... \\ &...2\lambda_2 \lambda_3 \epsilon_2 \epsilon_3 + 2\lambda_2 \lambda_4 \epsilon_2 \epsilon_4 + ... + 2\lambda_2 \lambda_n \epsilon_2 \epsilon_n + ... \\ &...2\lambda_{n-1} \lambda_n \epsilon_{n-1} \epsilon_n] \end{aligned}$$

Heteroskedasticity: consequences

continuing:[4]

$$
\begin{aligned}
V(b) =& \lambda_1^2 E[\epsilon_1^2] + \lambda_2^2 E[\epsilon_2^2] + ... + \lambda_n^2 E[\epsilon_n^2] ... \\
& ...2\lambda_1\lambda_2 E[\epsilon_1\epsilon_2] + 2\lambda_1\lambda_3 E[\epsilon_1\epsilon_3] + ... + 2\lambda_1\lambda_n E[\epsilon_1\epsilon_n] + ... \\
& ...2\lambda_2\lambda_3 E[\epsilon_2\epsilon_3] + 2\lambda_2\lambda_4 E[\epsilon_2\epsilon_4] + ... + 2\lambda_2\lambda_n E[\epsilon_2\epsilon_n] + ... \\
& ...2\lambda_{n-1}\lambda_n E[\epsilon_{n-1}\epsilon_n] \\
V(b) =& \lambda_1^2 E[\epsilon_1^2|X] + \lambda_2^2 E[\epsilon_2^2|X] + ... + \lambda_n^2 E[\epsilon_n^2|X] \\
V(b) =& \lambda_1^2\sigma^2 + \lambda_2^2\sigma^2 + ... + \lambda_n^2\sigma^2 \\
V(b) =& \sigma^2 \sum_{i=1}^{2} \lambda_i^2 \\
V(b) =& \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}
\end{aligned}
$$

[4] In the final line we make the conditioning on X explicit.

Heteroskedasticity: consequences

However, if $V(\epsilon) \neq \sigma^2$ then:

$$\begin{aligned} V(b) =& \lambda_1^2 E[\epsilon_1^2] + \lambda_2^2 E[\epsilon_2^2] + ... + \lambda_n^2 E[\epsilon_n^2] \\ =& \lambda_1^2 \sigma_1^2 + \lambda_2^2 \sigma_2^2 + ... + \lambda_n^2 \sigma_n^2 \\ =& \sum_{i=1} \lambda_i \sigma_i^2 \\ =& \frac{\sum_{i=1}^n \sigma_i^2 (X_i - \bar{X})^2}{[\sum_{i=1}^n (X_i - \bar{X})^2]^2} \end{aligned}$$

Thus, given:

$$V(b) = \frac{\sum_{i=1}^n \sigma_i^2 (X_i - \bar{X})^2}{[\sum_{i=1}^n (X_i - \bar{X})^2]^2} \neq \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

Our usual variance formula is incorrect, thus our usual hypothesis test are incorrect.

Heteroskedasticity: solutions

How can we construct an estimator for the variance that is robust to heteroskedasticity?

- We know $V(b) = \frac{\sum_{i=1}^{n} \sigma_i^2 (X_i - \bar{X})^2}{[\sum_{i=1}^{n} (X_i - \bar{X})^2]^2}$.
- The main problem is we don't know σ_i^2.
- One solution is to replace σ_i^2 by e_i^2: the squared residuals. Then use these in the above formula. Thus, (White's) heteroskedastic-robust variance is:
- $V(b) = \frac{\sum_{i=1}^{n} e_i^2 (X_i - \bar{X})^2}{[\sum_{i=1}^{n} (X_i - \bar{X})^2]^2}$.
- The square root of this is the heteroskedastic-robust standard error.

Heteroskedasticity: solutions

Notes on heteroskedastic-robust variance:

- The estimator is asymptotically valid under all forms of heteroskedasticity, including homoskedasticity.
- The arguments given above hold for the MLR model.
- Our estimated coefficient and the goodness of fit measures R^2 is unaffected by heteroskedasticity.

Heteroskedasticity: solutions

Notes on heteroskedastic-robust variance:

- Practically speaking, I will not ask you to calculate robust standard errors manually, however, they are easily implementable in Stata, just specify the robust option:
- reg Y X1 X2 X3, robust
- You can then perform t-tests as usual in Stata. Where $t = (b - \beta_0)/se(b)$, where $se(b)$ is the robust standard error.
- You can also perform F-tests as usual in Stata. However, you will not be able to manually calculate the results, since the equations for robust F-statistics are slightly different.

Heteroskedasticity: tests

We can formally test for heteroskedasticity. Consider again our standard multivariate population model:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \epsilon_i$$

Further, suppose all the CLRM assumptions hold, except for possibly homoskedasticity. Suppose the form of the heteroskedasticity can be represented as:

$$V(\epsilon_i | X_1, .., X_k) = E[\epsilon_i^2 | X_1, ..., X_k] = \sigma_i^2 = \gamma_0 + \gamma_1 Z_{1i} + ... + \gamma_p Z_{pi}$$

Where the Z random variables can represent any combination of our independent variables $(X_1, ..., X_k)$. Therefore,

$$\epsilon_i^2 = \gamma_0 + \gamma_1 Z_{1i} + ... + \gamma_p Z_{pi} + \xi_i$$

Heteroskedasticity: tests

The null of homoskedasticity is then:

$$H_0 : \gamma_1 = \gamma_2 = .. = \gamma_p = 0$$
$$H_1 : \text{ at least one } \gamma_j \neq 0$$

If we fail to reject the null we have the result that $\epsilon_i^2 = \gamma_0 + \xi_i$, showing the variance only depends on a constant. If we reject H_0 in favour of the alternative then we have heteroskedasticity. The test required is the usual F-statistic. Note however, we don't know ϵ_i so we have to use its sample counterpart e_i.

$$e_i^2 = \gamma_0 + \gamma_1 Z_{1i} + ... + \gamma_p Z_{ip} + \xi_i$$

Heteroskedasticity: tests

Further, we don't know the exact specification the heteroskedasticity takes.

1. A common procedure is to assume it is just a linear function of the independent variables:

$$\epsilon_i^2 = \gamma_0 + \gamma_1 X_{1i} + ... + \gamma_k X_{ki} + \xi_i$$

where the null will be $H_0 : \gamma_1 = ... = \gamma_k = 0$.

2. Another common specification is a linear function and their respective squares (White's test):

$$\epsilon_i^2 = \gamma_0 + \gamma_1 X_{1i} + ... + \gamma_k X_{ki} + \gamma_{k+1} X_{1i}^2 + ... + \gamma_{k+k} X_{ki}^2 + \xi_i$$

where the null will be
$H_0 : \gamma_1 = ... = \gamma_k = \gamma_{k+1} = ... = \gamma_{k+k} = 0$.

Heteroskedasticity: tests

To summarise, each procedure can be carried out manually. Let's carry out 1. as an example:

- Run the regression: $Y_i = \alpha + \beta_1 X_{1i} + ... + \beta_k X_{ki} + \epsilon_i$
- Predict the residuals for each observation: $e_i = Y_i - \hat{Y}_i$, then take their square to get e_i^2.
- Estimate an auxiliary regression: $e_i^2 = \gamma_0 + \gamma_1 X_{1i} + ... + \gamma_k X_{ki} + \xi_i$
- Save the the R^2 from the auxiliary regression.
- Test $H_0 : \gamma_1 = ... = \gamma_k = 0$ using the F-stat: $F = \frac{R^2/k}{(1-R^2)/(n-k-1)}$
- Apply the usual rejection rules.
- Alternatively in Stata:
- reg y x_1 ... x_k
- estat hettest x_1 ... x_k, fstat

Heteroskedasticity: a note of caution

If you detect heteroskedasicity, it may be a symptom of an omitted relevant variable.[5] To see this suppose the true population model is:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$$

But instead you estimate the model:

$$Y_i = \delta_0 + \delta_1 X_{1i} + u_i$$

Thus, $u_i = \beta X_{2i} + \epsilon_i$, and $Var(u_i) = E[u_i^2] = E[(\beta_2 X_{2i} + \epsilon_i)^2] = \beta_2^2 E[X_{2i}^2] + 2\beta_2 E[X_{2i}\epsilon_i] + E[\epsilon_i^2]$, which assuming ϵ_i satisfies the CLRM assumptions simplifies to: $E[u_i^2] = \beta_2^2 E[X_{2i}^2] + \sigma^2$. Thus the apparent heteroskedasticity in our wrongly specified model is driven by an omitted relevant variable and not actual heteroskedasticity.

[5] We return to the important problem of omitted variables shortly.

A very brief but important note on clustering

While on the topic of standard errors the CLRM assumptions includes the following:

$$Cov(\epsilon_i, \epsilon_j | X_1, ..., X_k) = 0 \text{ for all } i \neq j$$

- Under **random sampling** the data are independent: each observation is treated as a random draw from the same population, uncorrelated with the observation before or after. In such a situation zero correlation between the errors can be justified.

A very brief but important note on clustering

However, **random sampling** designs are often unrealistic. A random sample implies every individual in the population of interest (sampling frame) has an equal probability of selection. Many surveys that you may work with in applied projects (BHPS, Understanding Society, ... etc) will have non-random samples, they will have **complex sampling** designs. For example, complex sampling designs may place a higher probability of selection on households. This may be for cost reasons (collecting information from households in the same area is less costly). This leads to systematic correlation between (unobserved and observed) factors within groups of the population.[6]

[6] It is interesting to note that the opposite to clustering is stratification. Stratification ensures that certain parts of the population are included in the survey (minority groups, for example). This actually has the effect of increasing the precision of the estimator.

A very brief but important note on clustering

Another example, includes the level of randomisation in randomised control trials. For example, it may be that NGOs randomise programs across villages, not households. Or as in the following example, class size is randomised across classes not individuals. This leads to common (observed and unobserved) factors within classes.

In time-series econometrics ($Y_t = \alpha + \beta X_t + \epsilon_t$, $t = 1, ..., T$) it is common to assume errors are serially correlated across time (events in the previous time period, are hihgly likely to affect events in the current period). In cross-sectional data (and panel data) econometricians worry about correlation between observations in groups.

A brief but important note on clustering

The following example - the STAR exepriment - highlights dependence in the errors due to the group structure of the data.[7] Suppose we are interested in the following regression:

$$Y_{ig} = \alpha + \beta X_g + \epsilon_{ig} \text{ for } \; i = 1, ..., n; \; g = 1, ...G$$

Where:

- Where Y_{ig} is the exam score for student i in class g.
- The independent variable, class size, X_g ONLY varies at the group level.
- The error ϵ_{ig} varies at the individual level.

[7] The example is taken from Mostly Harmless Econometrics chapter 8. Note, in the STAR experiment, class size was randomised across classes. Therefore, in this example we are not concerned with bias, we are solely concerned with constructing correct standard errors.

A brief but important note on clustering

Given the structure of the data it is informative to split the error term (ϵ_{ig}):

$$\epsilon_{ig} = \nu_g + \eta_{ig}$$

Where:

- ϵ_{ig} is the error term from the original model and is often referred to as a composite error term.
- ν_g represents group level errors. When the group is a classroom this may include: environment, teacher quality, peer group quality and material.
- η_{ig} represent individual level errors, which may include: latent ability, motivation and effort.

A brief but important note on clustering

Given the structure of the data it is informative to split the error term:

$$\epsilon_{ig} = \nu_g + \eta_{ig}$$

The group error ν_g represents group level errors which are likely to cause correlation between the errors such that:

- $E[\epsilon_{ig}, \epsilon_{jg}|X] \neq 0$ for $i \neq j$, implying $cov(\epsilon_{ig}, \epsilon_{jg}|X) \neq 0$.
- Further, it is typically the case that $E[\epsilon_{ig}, \epsilon_{jg}|X] > 0$, since it is likely children in the class (or school) tend to have test scores that are positively correlated due to the same environment, teaching, and material provisions.
- If this group correlation structure is not taken account of unadjusted standard errors can be greatly under-estimated.[8]

[8] Typically correlation between the errors can be more problematic for bias estimation of the standard errors than can heteroskedasticity.

A brief but important note on clustering

Solutions:

- There are a number of solutions to the clustering problem. They typically involve either: i) modelling the group structure of the errors, or ii) making standard errors robust to the presence of the group structure.
- As with the case of heteroskedasticity we follow the later route. Suppose we have data on the STAR experiment where class size was randomised (across classes). In Stata we would run the following:
  - reg score class_size, cluster(class_id).
- A very conservative alternative is to run the regression at the group (rather than individual) level: $Y_g = \alpha + \beta X_g + \epsilon_g$. Where Y_g is the average score in class g. Condensing the information in this manner removes the problem of within class correlation. However, you may run into small sample problems.