Heteroskedasticity

Class 3

Wonmun Shin

(wonmun.shin@sejong.ac.kr)

Department of Economics, Sejong University

* This lecture note is written based on Professor Chang Sik Kim's lecture note.

What is Heteroskedasticity?

What is Heteroskedasticity?

- Heteroskedasticity: $\underbrace{\textit{hetero}}_{\text{different}} + \underbrace{\textit{skedasis}}_{\text{dispersion}}$

- Consider the following simple regression

$$Y_i = \beta_1 + \beta_2 X_i + e_i$$

  to explain household expenditure on food (Y_i) as a linear function of household income (X_i).

  - Then, we can consider the above linear relationship for two different groups: high-income group and low-income group.
  - Intuitively, income is less important as an explanatory variable for food expenditure of high-income group.
    - Food expenditure can be very different among high-income families due to their preferences.
  - Therefore, the variance of food expenditure (or, the variance of error term) is greater for high-income household.
  - This violates the homoskedasticity assumption in the classical assumptions.

What is Heteroskedasticity? [cont'd]

- Classical assumption (A3)

$$Var\left(Y_i\right) = Var\left(e_i\right) = \sigma^2 \quad for\ i = 1, \cdots, n$$

- **Heteroskedasticity**: To relax the above assumption, we allow for different variances for different observations.

$$Var\left(Y_i\right) = Var\left(e_i\right) = \sigma_i^2 \quad for\ i = 1, \cdots, n$$

  - Here, σ_i^2 are all different across i.

- The existence of different variances, or heteroskedasticity, is often encountered when using cross-sectional data.

Consequences of Heteroskedasticity

Consequences of Heteroskedasticity

$$Y_i = \beta_1 + \beta_2 X_i + e_i$$

- **Question 1** Can we obtain the OLS estimators under heteroskedasticity?
  - Note that we did not use (A3) to construct the OLS estimators.
  - $\therefore$ We can earn $\hat{\beta}_1$ and $\hat{\beta}_2$ through the ordinary least square method.

$$\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2}$$

Consequences of Heteroskedasticity [cont'd]

- **Question 2** Then, does still the properties of OLS estimator stand?
  - Unbiasedness

$$\hat{\beta}_2 - \beta_2 = \sum \omega_i e_i$$
$$\rightarrow E\left(\hat{\beta}_2 - \beta_2\right) = E\left(\sum \omega_i e_i\right) = \sum \omega_i E\left(e_i\right) = 0$$

    - So, $\hat{\beta}_2$ is a still unbiased (and linear) estimator.
  - Variance of OLS estimator

$$\begin{aligned} Var\left(\hat{\beta}_2\right) &= Var\left(\hat{\beta}_2 - \beta_2\right) = Var\left(\sum \omega_i e_i\right) \\ &= \omega_1^2 Var\left(e_1\right) + \omega_2^2 Var\left(e_2\right) + \cdots + \omega_n^2 Var\left(e_n\right) \\ &= \omega_1^2\sigma_1^2 + \omega_2^2\sigma_2^2 + \cdots + \omega_n^2\sigma_n^2 \\ &= \sum \omega_i^2\sigma_i^2 = \frac{\sum x_i^2\sigma_i^2}{\left[\sum x_i^2\right]^2} \end{aligned}$$

    - The usual formula for the variance of the OLS estimator is incorrect.

Consequences of Heteroskedasticity [cont'd]

- Consistency
  - Recall that the sufficient condition of consistency is $MSE \to 0$ as $n \to \infty$.
  - $Bias\left(\hat{\beta}_2\right) = 0$ and $Var\left(\hat{\beta}_2\right) \to 0$ as $n \to \infty$
  - $\therefore \hat{\beta}_2$ is a still consistent estimator.
- Recall, to do hypothesis testing, we needed the least squares standard error, which was (under homoskedasticity):

$$\sqrt{\widehat{Var\left(\hat{\beta}_2\right)}} = \sqrt{\frac{\hat{\sigma^2}}{\sum x_i^2}} \text{ where } \hat{\sigma^2} = \frac{\sum \hat{e}_i^2}{n-2}$$

  - However, as we have seen, the usual formula for $Var\left(\hat{\beta}_2\right)$ in incorrect when there exists heteroskedasticity.
  - Consequently, the usual least squares standard error is **inconsistent**.
  - Even if the OLS estimator of coefficients is consistent, the usual least squares standard error should not be used to test hypothesis.

Consequences of Heteroskedasticity [cont'd]

- **Question 3** We have different formula for the variance of the OLS estimator. Then, the OLS estimator is still BLUE?
  - Gauss-Markov theorem does not apply any more, so OLS estimators are *no longer BLUE*.
  - There exists another linear unbiased estimator of β_2 which has a smaller variance than $\hat{\beta}_2$ when the errors are heteroskedastic, namely *Generalized Least Squares (GLS)* estimators.
  - Nevertheless, it doesn't mean we cannot use the OLS estimator any more.
    - But we should sacrifice the accuracy of estimator owing to the larger variance.
    - Also, the probability of rejecting the null falls.
    - In fact, the usual least squares standard errors are inconsistent, which implies the test results will be wrong if we do not consider the existence of heteroskedasticity.

Detecting Heteroskedasticity

Detecting Heteroskedasticity

- How does one know that heteroskedasticity is present in a specific situation?
  - There are no golden rule for detecting heteroskedasticity.
  - In most cases, heteroskedasticity may be a matter of intuition, educated guesswork, and prior empirical experience.
- There are some informal or formal methods of detecting heteroskedasticity.
  - Most of methods are based on the examination of the OLS residuals ($\hat{e}_i$) since they are the ones we can observe, and not the error terms (e_i).
  - **Informal methods**
    - Nature of problem: Sometimes, the nature of the problem suggests whether heteroskedasticity is likely to be encountered. (ex. regression of consumption on income, regression of investment on sales)
    - Graphical method: Plot of residual squared ($\hat{e}_i^2$)
  - **Formal methods**: Park test, Goldfeld-Quandt (GQ) Test, White test, Glejser test, Spearman's rank correlation test, Breusch-Pagan LM test, Koenker-Bassett test, $\cdots\cdots\cdots$

Park Test

1. Park test

- Park suggests that σ_i^2 is some function of the explanatory variable X_i.
- The functional form he suggests is

$$\sigma_i^2 = \sigma^2 X_i^{\gamma} e^{\nu_i}$$
$$\rightarrow \ \ln \sigma_i^2 = \ln \sigma^2 + \gamma \ln X_i + \nu_i$$

where ν_i is white noise, $i.e.$ $\nu_i \sim iid \left(0, \sigma_\nu^2\right)$

- Since σ_i^2 is generally not known, Park suggests using $\hat{e}_i^2$ as a proxy.

$$\ln \hat{e}_i^2 = constant + \gamma \ln X_i + \nu_i$$

- We can obtain $\hat{\gamma}$ and test $H_0 : \gamma = 0$.
    - If we reject the null, $\exists$ heteroskedasticity.
- Caveats
    - Assumption would not be correct.
    - ν_i is white noise? It might be heteroskedastic.

Goldfeld-Quandt Test

2. Goldfeld-Quandt Test (GQ Test)

- Assumption: σ_i^2 is positively related to X_i
- Test

$$\begin{cases} H_0 : \textit{Homoskedasticity} \\ H_1 : \sigma_i^2 \approx \textit{monotonically related to } X_i \end{cases}$$

- Reorder the observations according to the values of X_i (beginning with the lowest X value)
- After omitting c central observations, divide the remaining $(n - c)$ observations into two groups: Group 1 (high X_i) and Group 2 (low X_i)
- Compute RSS_1 and RSS_2

$$F\text{-statistic} = \frac{RSS_1 / df_1}{RSS_2 / df_2} \sim F(df_1, df_2)$$

- If we reject the null (F-statistic $>$ critical value), $\exists$ heteroskedasticity.
- Caveats
  - Assumption would not be correct.
  - Choice of c is arbitrary.

White Test

3. White Test

- It is called *White's General Heteroskedasticity Test*.
- Test

$$\begin{cases} H_0: \textit{Homoskedasticity} \\ H_1: \textit{Not homoskedasticity} \end{cases}$$

- Auxiliary regression

$$\hat{e}_i^2 = \alpha_1 + \alpha_2 X_i + \alpha_3 X_i^2 + u_i$$

- We can compute R^2 $\left(= 1 - \frac{RSS}{TSS}\right)$ which represents a measure of goodness of fit.

$$n \cdot R^2 \overset{asy}{\sim} \chi^2_{df}$$

where the degree of freedom is (# of regressors - # of constant).

- If we reject the null ($n \cdot R^2 >$ critical value), $\exists$ heteroskedasticity.
- Caveats H_1 is too general $\rightarrow$ low power of test (especially, in small sample)

Solutions for Heteroskedasticity

1. Take Logarithms

- Suppose that we recognize the existence of heteroskedasticity but do not know any other things.
- We want to reduce the variance even a little bit.
- Log transformation

$$\begin{aligned} Y_i &= \beta_1 + \beta_2 X_i + e_i \\ \rightarrow \ \ln Y_i &= \beta_1 + \beta_2 \ln X_i + e_i \end{aligned}$$

- Caution: The interpretation of $\hat{\beta}_2$ becomes different.

2. White Correction

- Given that the conventional lest squares are incorrect under the heteroskedasticity, the **White correction** gives you a **consistent** estimator for the variance of OLS estimator.

- Recall

$$Var\left(\hat{\beta}_2\right) = \frac{\sum x_i^2 \sigma_i^2}{\left[\sum x_i^2\right]^2}$$

- **White estimator** (of standard error)

$$\widehat{Var\left(\hat{\beta}_2\right)} = \frac{\sum x_i^2 \hat{e}_i^2}{\left[\sum x_i^2\right]^2}$$

Solutions for Heteroskedasticity: White Correction [cont'd]

- If we use $\widehat{Var\left(\hat{\beta}_2\right)}$, we can correct standard errors and t-statistics for OLS estimators.
- The squared residuals are used to approximate the variances, the White estimator is appropriate in large samples.
- However, OLS estimator $\hat{\beta}_2$ is still inefficient (since Gauss-Markov theorem no longer holds).
- One advantage for this procedure is that you need not know the form of the heteroskedasticity.

3. Generalized Least Squares (GLS)

- Under heteroskedasticity, OLS is not BLUE.
- But an estimation method known as **Generalized Least Squares (GLS)** which takes the heteroskedasticity into account explicitly obtains the minimum variance within the class of linear unbiased estimators.
- Assume that the heteroskedastic variances σ_i^2 are known.
- **Idea:** let's transform the regression model to make the error variances homoskedastic.

$$Y_i = \beta_1 + \beta_2 X_i + e_i \quad \text{under } Var\left(e_i\right) = \sigma_i^2$$

  - Divide the regression by σ_i as

$$\left(\frac{Y_i}{\sigma_i}\right) = \beta_1 \left(\frac{1}{\sigma_i}\right) + \beta_2 \left(\frac{X_i}{\sigma_i}\right) + \left(\frac{e_i}{\sigma_i}\right)$$

- Why?

$$Var\left(\frac{e_i}{\sigma_i}\right) = E\left[\left(\frac{e_i}{\sigma_i}\right)^2\right] - \left[E\left(\frac{e_i}{\sigma_i}\right)\right]^2$$
$$= \frac{1}{\sigma_i^2}E\left(e_i^2\right)$$
$$= \frac{1}{\sigma_i^2}\sigma_i^2 = 1$$

  - That is, the error term divided by σ_i does not have heteroskedasticity problem any more.

- Define

$$Y_i^* = \frac{Y_i}{\sigma_i},\ X_{1i}^* = \frac{1}{\sigma_i},\ X_{2i}^* = \frac{X_i}{\sigma_i},\ e_i^* = \frac{e_i}{\sigma_i}$$

Solutions for Heteroskedasticity: GLS [cont'd]

- Then, we obtain the new regression model

$$Y_i^* = \beta_1 X_{1i}^* + \beta_2 X_{2i}^* + e_i^* \quad \text{under } Var\left(e_i^*\right) = 1$$

  - Note that the transformed regression satisfies the classical assumptions.

- Let $\hat{\beta}_1^*$ and $\hat{\beta}_2^*$ be the OLS estimators in the transformed regression: **GLS estimator**

- OLS estimators in the transformed model (which is the GLS estimators in the original model) is the **BLUE** (according to Gauss-Markov Theorem)

Solutions for Heteroskedasticity: GLS [cont'd]

- In the transformed model, $\hat{\beta}_1^*$, $\hat{\beta}_2^*$ minimizes the following criterion function:

$$\sum \hat{e}_i^{*2} = \sum \left(\frac{\hat{e}_i^2}{\sigma_i^2} \right) = \sum \frac{1}{\sigma_i^2} \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i \right)^2$$

  - $\therefore$ In the GLS estimation, we minimize a *weighted* sum of squares of residuals.
  - GLS estimator under heteroskedasticity = **WLS (Weighted Least Squares)** estimator
  - Weight: $\frac{1}{\sigma_i^2}$ $\rightarrow$ light weight on less informative ones, heavy weight on more informative ones

- **However!** We have to know the values (or structure) of σ_i^2 which are generally not known $\rightarrow$ ***Infeasible GLS***

Solutions for Heteroskedasticity: FGLS

- **Feasible Generalized Least Squares (FGLS)**
  - We can use estimated value $\hat{\sigma_i^2}$ for the true parameters σ_i^2.
  - We call the GLS estimator based on $\hat{\sigma_i^2}$ as *feasible GLS*.
  - Since we use $\hat{\sigma_i^2}$ instead of σ_i^2, FGLS estimator may not be more efficient than OLS estimator with White correction (particularly in small sample).
  - Furthermore, misspecification of heteroskedasticity may lead to an inconsistent estimation.