Lecture 8: Heteroskedasticity

Shuo Liu

UCLA Summer School Econ 103

July 24, 2017

Outline

1. The Nature of Heteroskedasticity
2. Detecting Heteroskedasticity
3. Heteroskedasticity-Consistent Standard Errors
4. Generalized Least Squares: Known Form of Variance
5. Generalized Least Squares: Unknown Form of Variance

The Nature of Heteroskedasticity

- Consider our basic linear function:

$$E(y_i) = \beta_1 + \beta_2 x_i \tag{1}$$

- As before, we define the random error term as:

$$e_i = y_i - E(y_i) = y_i - \beta_1 - \beta_2 x_i \tag{2}$$

- Equivalent model form is:

$$y_i = \beta_1 + \beta_2 x_i + e_i \tag{3}$$

- Homoskedasticity:

$$Var(e_i|x_i) = Var(e_i) = \sigma^2, \quad \forall i \tag{4}$$

- Heteroskedasticity:

$$Var(e_i|x_i) = \sigma_i^2, \quad \forall i \tag{5}$$

The Nature of Heteroskedasticity

- If random error e_i is heteroskedastic, by nonrandomness of x_i, y_i is also heteroskedastic

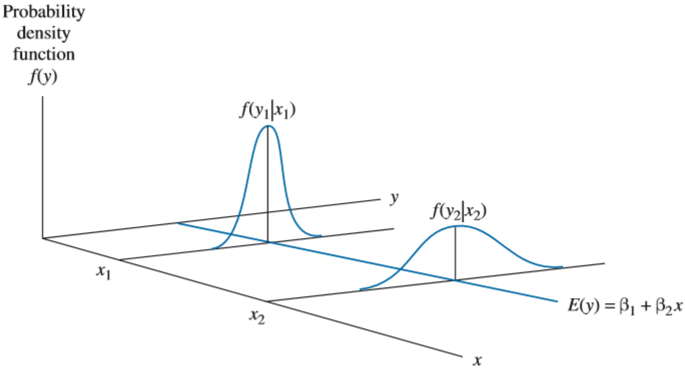


FIGURE 8.1 Heteroskedastic errors.

The Nature of Heteroskedasticity

- When there is heteroskedasticity, one of the least squares assumptions is violated. We still have that

$$E(e_i) = 0, \quad Cov(e_i, e_j) = 0 \tag{6}$$

- But now, the assumption that $Var(e_i|x_i) = \sigma^2$ is replaced by:

$$Var(e_i|x_i) = \sigma_i^2 = h(x_i), \quad \forall i \tag{7}$$

- Here $h(x_i)$ is a function of x_i.

The Nature of Heteroskedasticity

- Food expenditure example: $y = FOODEXP$, $x = INCOME$

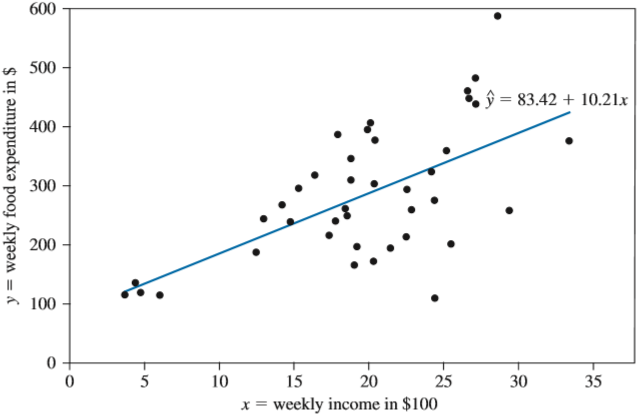
$$\hat{y} = 83.42 + 10.21x \tag{8}$$

- The residuals are defined as:

$$\hat{e}_i = y_i - \hat{y} = y_i - 83.42 - 10.21x \tag{9}$$

The Nature of Heteroskedasticity

- As the level of $INCOME$ increases, the variation (variance) in residuals $\hat{e}_i$ increases, so we guess there exists heteroskedasticity



There are two implications of heteroskedasticity:

- The least squares estimator is still a linear and unbiased estimator, but it is no longer the best estimator. In fact, there is another estimator with a smaller variance;
- The usual standard errors computed for the least squares estimator are incorrect. Thus, condence intervals and hypothesis tests that use these standard errors may be misleading.

The Nature of Heteroskedasticity

- What happens to the standard errors?
- Consider the model form that we originally assumed:

$$y_i = \beta_1 + \beta_2 x_i + e_i, \quad Var(e_i) = \sigma^2 \tag{10}$$

- The variance of b_2 which is the least square estimator for β_2 is:

$$Var(b_2) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{11}$$

- Now let the variances of random errors differ. That is, consider the model:

$$y_i = \beta_1 + \beta_2 x_i + e_i, \quad Var(e_i) = \sigma_i^2 \tag{12}$$

- The variance of b_2 which is the least square estimator for β_2 is:

$$Var(b_2) = \sum_{i=1}^{n} w_i^2 \sigma_i^2 = \frac{\sum_{i=1}^{n}[(x_i - \bar{x})^2 \sigma_i^2]}{[\sum_{i=1}^{n}(x_i - \bar{x})^2]^2}, \quad w_i = \frac{x_i - \bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{13}$$

Detecting Heteroskedasticity

There are two methods we can use to detect heteroskedasticity:

- **Method 1**: An informal way using residual charts (as in Figure 8.2, 8.3);
- **Method 2**: A formal way using statistical tests

Method 1:

- If the errors are homoskedastic, there should be no patterns of any sort in the residuals;
- If the errors are heteroskedastic, they may tend to exhibit greater variation in some systematic way (this is just one specific case);
- This method of investigating heteroskedasticity can be followed for any simple regression (complex model still requires the use of method 2);
- In a regression with more than one explanatory variable we can **plot the residuals against each explanatory variable** $x_{ki}, i = 1, 2, \cdots, n, k = 2, 3, \cdots, K$, **or against** $\hat{y}_i$, to see if they vary in a systematic way

- As the level of the unique explanatory variable $INCOME$ increases, the variation (variance) in residuals $\hat{e}_i$ increases, so we guess there exists heteroskedasticity

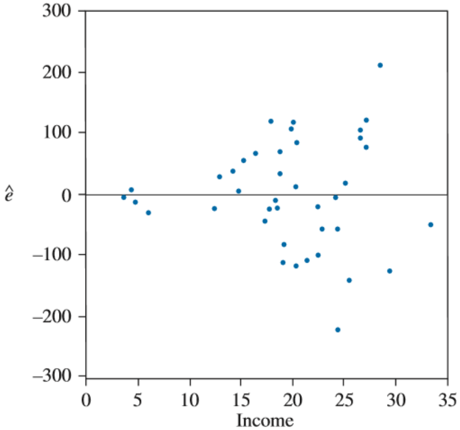


FIGURE 8.3 Least squares food expenditure residuals plotted against income.

Method 2:

- We need to have a test based on a variance function to detect heteroskedasticity;
- Consider the general multiple regression model:

$$E(y_i) = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots + \beta_K x_{Ki} \tag{14}$$

- **A general form for the variance function related to the multiple regression model above is:**

$$Var(y_i) = \sigma_i^2 = E(e_i^2) = h(\alpha_1 + \alpha_2 z_{2i} + \cdots + \alpha_S z_{Si}) \tag{15}$$

  where $z_{si}, s = 2, 3, \cdots, S$ are some other random variables used to explain the heteroskedastic σ_i^2 (z's may be correlated with the original explanatory variables x's)

Detecting Heteroskedasticity

- Possible functions for $h(x_i)$ are:
    1. Exponential function:

$$h(\alpha_1 + \alpha_2 z_{2i} + \cdots + \alpha_S z_{Si}) = exp(\alpha_1 + \alpha_2 z_{2i} + \cdots + \alpha_S z_{Si}) \quad (16)$$

    2. Linear function:

$$h(\alpha_1 + \alpha_2 z_{2i} + \cdots + \alpha_S z_{Si}) = \alpha_1 + \alpha_2 z_{2i} + \cdots + \alpha_S z_{Si} \quad (17)$$

- Note that in this latter case one must be careful to ensure that $h(x_i) > 0$.
- By the formula of $h(x_i)$, when will there be homoskedasticity?

Detecting Heteroskedasticity

- When

$$\alpha_2 = \alpha_3 = \cdots = \alpha_S = 0 \tag{18}$$

we have

$$h(\alpha_1 + \alpha_2 z_{2i} + \cdots + \alpha_S z_{Si}) = h(\alpha_1) \tag{19}$$

where $h(\alpha_1)$ is a constant.

- So when $\alpha_2 = \alpha_3 = \cdots = \alpha_S = 0$, heteroskedasticity is not present;
- Use the joint hypothesis to test whether there exists heteroskedasticity

$$H_0 : \alpha_2 = \alpha_3 = \cdots = \alpha_S = 0 \tag{20}$$

$$H_1 : \text{At least one } \alpha_s \neq 0, s = 2, 3, \cdots, S \tag{21}$$

Detecting Heteroskedasticity

- Suppose we use the specific case in equation (17),

$$Var(y_i) = \sigma_i^2 = E(e_i^2) = \alpha_1 + \alpha_2 z_{2i} + \cdots + \alpha_S z_{Si} \tag{22}$$

- For the last equality, we can define a new multiple regression model:

$$e_i^2 = E(e_i^2) + \nu_i = \alpha_1 + \alpha_2 z_{2i} + \cdots + \alpha_S z_{Si} + \nu_i \tag{23}$$

- We use the squares of residuals $\{\hat{e}_i^2\}_{i=1}^n$ as dependent variable to regress on z's:

$$\hat{e}_i^2 = \alpha_1 + \alpha_2 z_{2i} + \cdots + \alpha_S z_{Si} + \nu_i \tag{24}$$

- If the multiple regression model fit the data well, which means there exists significant relationship between $\hat{e}_i^2$ and $z_2, z_3, \cdots, z_S$ (usually functions of $x_2, x_3, \cdots, x_K$), what does it imply?

Detecting Heteroskedasticity

- Since the R^2 from the new multiple regression above measures the proportion of variation in $\hat{e}_i^2$ explained by the z's, it is a natural candidate for a test statistic;
- It can be shown that when H_0 is true, the sample size multiplied by R^2 follows a χ^2 distribution with $S-1$ degrees of freedom

$$n \times R^2 \sim \chi^2_{(S-1)} \tag{25}$$

- It is important to note that **the test is a large sample test**, that is, it applies only when n is large;
- **Note** that this method presupposes that we have knowledge of the variables appearing in the variance function (z's) if heteroskedasticity were true.

How to set the z's (One option is White test):

- Define the variables z's as equal to the x's, the squares of the x's, and possibly their cross-products;
- Consider the model:

$$E(y) = \beta_1 + \beta_2 x_2 + \beta_3 x_3 \tag{26}$$

- The White test **without** cross-product terms (interactions) specifies:

$$z_2 = x_2, z_3 = x_3, z_4 = x_2^2, z_5 = x_3^2 \tag{27}$$

- Of course, we can further add one more interaction term:

$$z_6 = x_2 x_3 \tag{28}$$

- The White test is performed using:

$$n \times R^2 \sim \chi_{(S-1)} \tag{29}$$

Example:

- We test $H_0 : \alpha_2 = 0$ against $H_1 : \alpha_2 \neq 0$ in the variance function $\sigma_i^2 = h(\alpha_1 + \alpha_2 x_i)$;
- First estimate $\hat{e}_i^2 = \alpha_1 + \alpha_2 x_i + \nu_i$ by OLS method;
- Calculate measure of goodness-of-fit:

$$R^2 = 1 - \frac{SSE}{SST} = 0.1846 \tag{30}$$

- Suppose sample size $n = 40$, construct test statistic:

$$\chi^2_{(1)} = n \times R^2 = 40 \times 0.1846 = 7.38 \tag{31}$$

- χ^2 test is always one-tail (right-tail) test: in this case, the 5% critical value is 3.84, so since $7.38 > 3.84$, we reject H_0 and conclude that the variance depends on income, that is, **there exists heteroskedasticity**.

Example: for the White test

- We estimate:

$$\hat{e}_i^2 = \alpha_1 + \alpha_2 x_i + \alpha_3 x_i^2 + \nu_i \tag{32}$$

- Then $S = 3$, $n = 40$, and we test $H_0 : \alpha_2 = \alpha_3 = 0$ against $H_1 : \alpha_2 \neq 0$ and/or $\alpha_3 \neq 0$.
- **Although it is joint hypothesis, since it is to detect heteroskedasticity, we still just need to use χ^2 test**:

$$\chi^2_{(2)} = n \times R^2 = 40 \times 0.1888 = 7.555 \tag{33}$$

- Given significance level $\alpha = 0.05$, either by critical value $\chi_{(0.95,2)} = 5.99 < 7.555$ or by the calculated p-value $0.023 < 0.05$, we will reject H_0.
- We conclude there exists heteroskedasticity.

Heteroskedasticity-Consistent Standard Errors

- Recall that there are two problems with using the least squares estimator in the presence of heteroskedasticity:
    1. The least squares estimator, although still being unbiased, is no longer the best;
    2. The usual least squares standard errors are incorrect, which invalidates interval estimates and, more generally, hypothesis tests.
- There is a way of correcting the standard errors so that our interval estimates and hypothesis tests are still valid.

Heteroskedasticity-Consistent Standard Errors

- Under heteroskedasticity:

$$Var(b_2) = \frac{\sum_{i=1}^{n}[(x_i - \bar{x})^2\sigma_i^2]}{[\sum_{i=1}^{n}(x_i - \bar{x})^2]^2} \tag{34}$$

- A consistent estimator for this variance has been developed and is known as the **Whites heteroskedasticity-consistent standard errors**;
- In STATA it is called robust standard errors.
- What is the straight forward way to construct such consistent estimator?

Heteroskedasticity-Consistent Standard Errors

- If the number of explanatory variables in the original model is K, we have:

$$\widehat{Var}(b_2) = \frac{n}{n-K} \frac{\sum_{i=1}^{n}[(x_i - \bar{x})^2 \hat{\sigma}_i^2]}{[\sum_{i=1}^{n}(x_i - \bar{x})^2]^2} \tag{35}$$

- Food expenditure example:

$$\begin{array}{rl} \hat{y} & = \underset{(27.46)}{83.42} + \underset{(1.81)}{10.21}x \\ \text{White se} & \\ \text{Incorrect se} & \quad (43.41) \quad\;\; (2.09) \end{array} \tag{36}$$

- The two corresponding 95% confidence intervals for β_2 are:
  1. White:

$$b_2 \pm t_c se(b_2) = 10.21 \pm 2.204 \times 1.81 = [6.55, 13.87] \tag{37}$$

  2. Incorrect:

$$b_2 \pm t_c se(b_2) = 10.21 \pm 2.204 \times 2.09 = [5.97, 14.45] \tag{38}$$

Generalized Least Squares: Known Form of Variance

- Recall the food expenditure example with heteroskedasticity:

$$y_i = \beta_1 + \beta_2 x_i + e_i \tag{39}$$

$$E(e_i) = 0, Var(e_i) = \sigma_i^2, Cov(e_i, e_j) = 0$$

- Now OLS estimator is no longer the best one, to develop an estimator that is better than the OLS estimator, we need to make a further assumption about σ_i^2;
- An estimator known as the **generalized least squares (GLS) estimator**, depends on the unknown σ_i^2;
- We impose some structure on σ_i^2: $Var(e_i) = \sigma_i^2 = \sigma^2 x_i$.

Generalized Least Squares: Known Form of Variance

- By assuming this structure, we can **transform the model with heteroskedastic errors into one with homoskedastic errors**:

$$\frac{y_i}{\sqrt{x_i}} = \beta_1 \left(\frac{1}{\sqrt{x_i}} \right) + \beta_2 \left(\frac{x_i}{\sqrt{x_i}} \right) + \frac{e_i}{\sqrt{x_i}} \tag{40}$$

- Define the following transformed variables:

$$y_i^* = \frac{y_i}{\sqrt{x_i}}, x_{1i}^* = \frac{1}{\sqrt{x_i}}, x_{2i}^* = \frac{x_i}{\sqrt{x_i}}, e_i^* = \frac{e_i}{\sqrt{x_i}} \tag{41}$$

- Our model can be written now as:

$$y_i^* = \beta_1 x_{1i}^* + \beta_2 x_{2i}^* + e_i^* \tag{42}$$

Generalized Least Squares: Known Form of Variance

- The new transformed error term is homoskedastic:

$$Var(e_i^*) = Var(\frac{e_i}{\sqrt{x_i}}) = \frac{1}{x_i}Var(e_i) = \frac{1}{x_i}\sigma^2 x_i = \sigma^2 \tag{43}$$

- The transformed error term will maintain the properties of zero mean and zero correlation between different observations;
- **To obtain the best linear unbiased estimator for a model with heteroskedasticity**:
    1. Calculate the transformed variables $y_i^*, x_{1i}^*, x_{2i}^*$;
    2. Use OLS method to estimate the transformed model.
- The estimator obtained in this way is called a generalized least squares (GLS) estimator.

Generalized Least Squares: Known Form of Variance

- One way of viewing the generalized least squares estimator is as a **weighted-least-square** estimator;
- The difference now is: minimizing the sum of squared transformed errors

$$\sum_{i=1}^{n} {e_i^*}^2 = \sum_{i=1}^{n} \frac{e_i^2}{x_i} = \sum_{i=1}^{n} \left(x_i^{-1/2} e_i \right)^2 \tag{44}$$

- That is, the errors are weighted by $x_i^{-1/2}$.

Generalized Least Squares: Unknown Form of Variance

- Consider a more general specification of the error variance:

$$Var(e_i) = \sigma_i^2 = \sigma^2 x_i^\gamma \tag{45}$$

where γ is an unknown parameter.

- **When you have unknown power, most time you need to take** ln **on both sides**:

$$ln(\sigma_i^2) = ln(\sigma^2) + \gamma ln(x_i) \tag{46}$$

where by assumption $ln(\sigma^2)$ is constant, can be denoted as α_1; γ is constant, can be denoted as α_2.

- Now we have the variance function as a log-linear function:

$$ln(\sigma_i^2) = \alpha_1 + \alpha_2 z_i = \alpha_1 + \alpha_2 ln(x_i) \tag{47}$$

- Then we use residuals from the OLS estimation of the original model, we estimate α_1 and α_2:

$$ln(\hat{e}_i^2) = \alpha_1 + \alpha_2 z_i + \nu_i, \quad z_i = ln(x_i) \tag{48}$$

Generalized Least Squares: Unknown Form of Variance

- For the food expenditure data, we have:

$$\widehat{ln(\sigma_i^2)} = \widehat{ln(\hat{e}_i^2)} = 0.9378 + 2.329 z_i + \nu_i, \quad z_i = ln(x_i) \tag{49}$$

- We can obtain estimator of variance:

$$\widehat{\sigma_i^2} = exp(\hat{\alpha_1} + \hat{\alpha_2} z_i) \tag{50}$$

then **transform the original model by dividing both sides by** $\widehat{\sigma_i}$:

$$\frac{y_i}{\widehat{\sigma_i}} = \beta_1 \left(\frac{1}{\widehat{\sigma_i}} \right) + \beta_2 \left(\frac{x_i}{\widehat{\sigma_i}} \right) + \frac{e_i}{\widehat{\sigma_i}} \tag{51}$$

- Theoretically the transformed error term is homoskedastic (since $\widehat{\sigma_i}^2$ is unbiased estimator of σ_i^2):

$$Var(\frac{e_i}{\sigma_i}) = \frac{1}{\sigma_i^2} Var(e_i) = \frac{1}{\sigma_i^2} \sigma_i^2 = 1 \tag{52}$$

Generalized Least Squares: Unknown Form of Variance

- To obtain a generalized least squares estimator for β_1 and β_2, define the transformed variables:

$$y_i^* = \frac{y_i}{\widehat{\sigma}_i}, x_{1i}^* = \frac{1}{\widehat{\sigma}_i}, x_{2i}^* = \frac{x_i}{\widehat{\sigma}_i} \tag{53}$$

- Use OLS method to estimate the transformed model:

$$y_i^* = \beta_1 x_{1i}^* + \beta_2 x_{2i}^* + e_i^* \tag{54}$$