Chapter 8
Heteroskedasticity

Walter R. Paczkowski
Rutgers University

Chapter Contents

- 8.1 The Nature of Heteroskedasticity
- 8.2 Detecting Heteroskedasticity
- 8.3 Heteroskedasticity-Consistent Standard Errors
- 8.4 Generalized Least Squares: Known Form of Variance
- 8.5 Generalized Least Squares: Unknown Form of Variance

8.1 The Nature of Heteroskedasticity

- Consider our basic linear function:

Eq. 8.1

$$E(y) = \beta_1 + \beta_2 x$$

  - To recognize that not all observations with the same *x* will have the same *y*, and in line with our general specification of the regression model, we let e_i be the difference between the ith observation y_i and mean for all observations with the same x_i.

Eq. 8.2

$$e_i = y_i - E(y_i) = y_i - \beta_1 - \beta_2 x_i$$

- Our model is then:

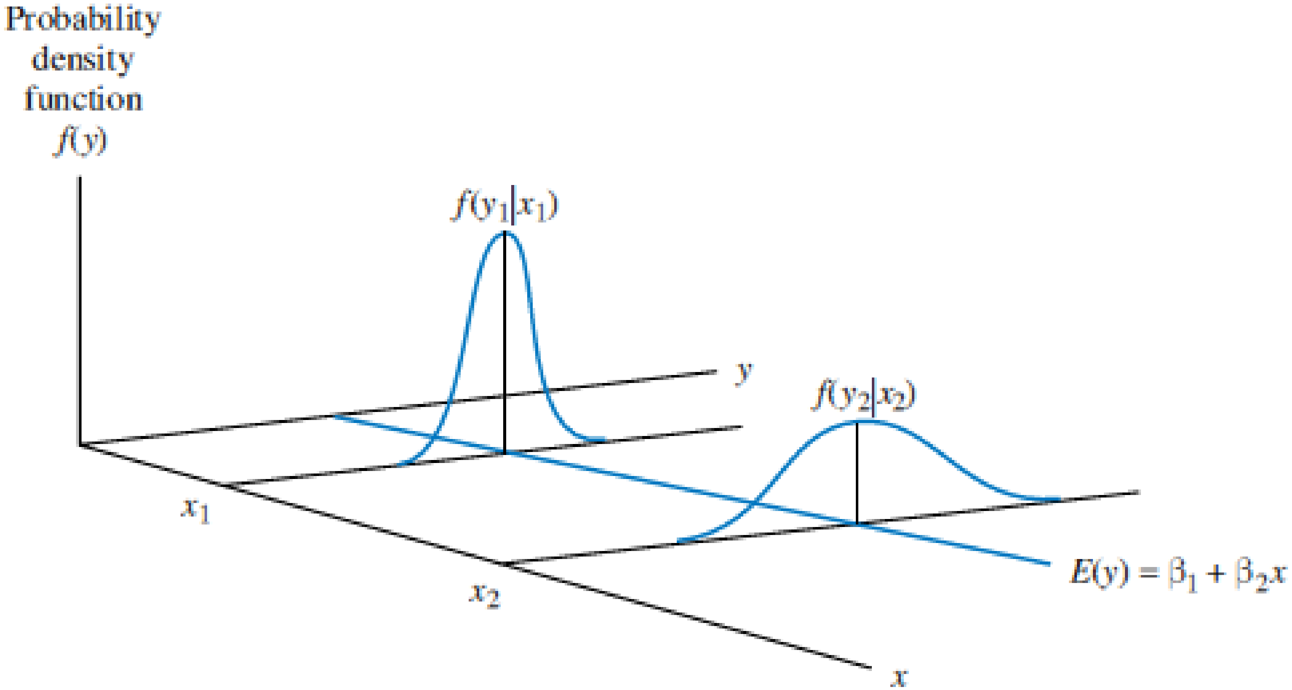
Eq. 8.3

$$y_i = \beta_1 + \beta_2 x_i + e_i$$

- The probability of getting large positive or negative values for *e* is higher for large (or small) values of *x* than it is for low (or high) values
  - A random variable, in this case *e*, has a higher probability of taking on large values if its variance is high.
  - We can capture this effect by having var(*e*) depend directly on *x*.
    - Or: var(e) increases as x increases

- When the variances for all observations are not the same, we have **heteroskedasticity**
  - The random variable y and the random error e are **heteroskedastic**
  - Conversely, if all observations come from probability density functions with the same variance, **homoskedasticity** exists, and y and e are **homoskedastic**

FIGURE 8.1 Heteroskedastic errors


Probability density function $f(y)$
$f(y_1|x_1)$
y
$f(y_2|x_2)$
x_1
x_2
$E(y) = \beta_1 + \beta_2 x$
x

- When there is heteroskedasticity, one of the least squares assumptions is violated:

$$\mathrm{var}(e_i) = \sigma^2$$

  - Replace this with:

Eq. 8.4

$$\mathrm{var}(y_i) = \mathrm{var}(e_i) = h(x_i)$$

where $h(x_\mathrm{i})$ is a function of x_i that increases (or decreases) as x_i increases

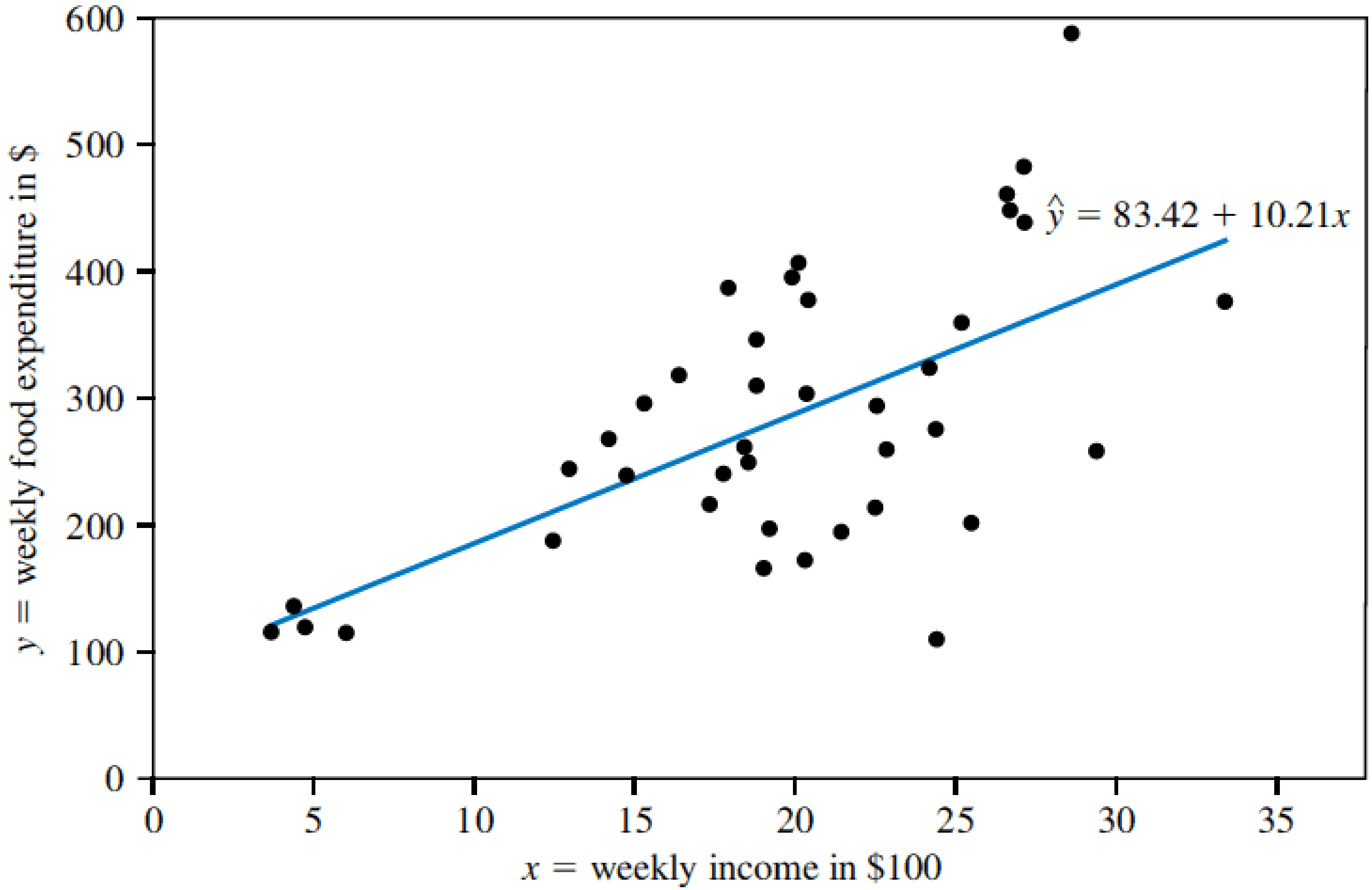
- Example from food data:

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

  - We can rewrite this as:

$$\hat{e}_i = y_i - 83.42 - 10.21x_i$$

FIGURE 8.2 Least squares estimated food expenditure function and observed data points



- Heteroskedasticity is often encountered when using cross-sectional data
  - The term **cross-sectional data** refers to having data on a number of economic units such as firms or households, *at a given point in time*
  - Cross-sectional data invariably involve observations on economic units of varying sizes

- This means that for the linear regression model, as the size of the economic unit becomes larger, there is more uncertainty associated with the outcomes *y*
  - This greater uncertainty is modeled by specifying an error variance that is larger, the larger the size of the economic unit

- Heteroskedasticity is not a property that is necessarily restricted to cross-sectional data
  - With time-series data, where we have data over time on one economic unit, such as a firm, a household, or even a whole economy, it is possible that the error variance will change

- There are two implications of heteroskedasticity:
  1. The least squares estimator is still a linear and unbiased estimator, but it is no longer best
     - There is another estimator with a smaller variance
  2. The standard errors usually computed for the least squares estimator are incorrect
     - Confidence intervals and hypothesis tests that use these standard errors may be misleading

- What happens to the standard errors?
  - Consider the model:

Eq. 8.5

$$y_i = \beta_1 + \beta_2 x_i + e_i \qquad \operatorname{var}(e_i) = \sigma^2$$

  - The variance of the least squares estimator for β_2 as:

Eq. 8.6

$$\operatorname{var}(b_2) = \frac{\sigma^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

- Now let the variances differ:
  - Consider the model:

Eq. 8.7

$$y_i = \beta_1 + \beta_2 x_i + e_i \qquad \operatorname{var}(e_i) = \sigma_i^2$$

  - The variance of the least squares estimator for β_2 is:

Eq. 8.8

$$\operatorname{var}(b_2) = \sum_{i=1}^{N} w_i^2 \sigma_i^2 = \frac{\sum_{i=1}^{N}\left[(x_i - \bar{x})^2 \sigma_i^2\right]}{\left[\sum_{i=1}^{N}(x_i - \bar{x})^2\right]^2}$$

- If we proceed to use the least squares estimator and its usual standard errors when:

$$\operatorname{var}\left(e_i\right) = \sigma_i^2$$

we will be using an estimate of Eq. 8.6 to compute the standard error of b_2 when we should be using an estimate of Eq. 8.8

– The least squares estimator, that it is no longer best in the sense that it is the minimum variance linear unbiased estimator

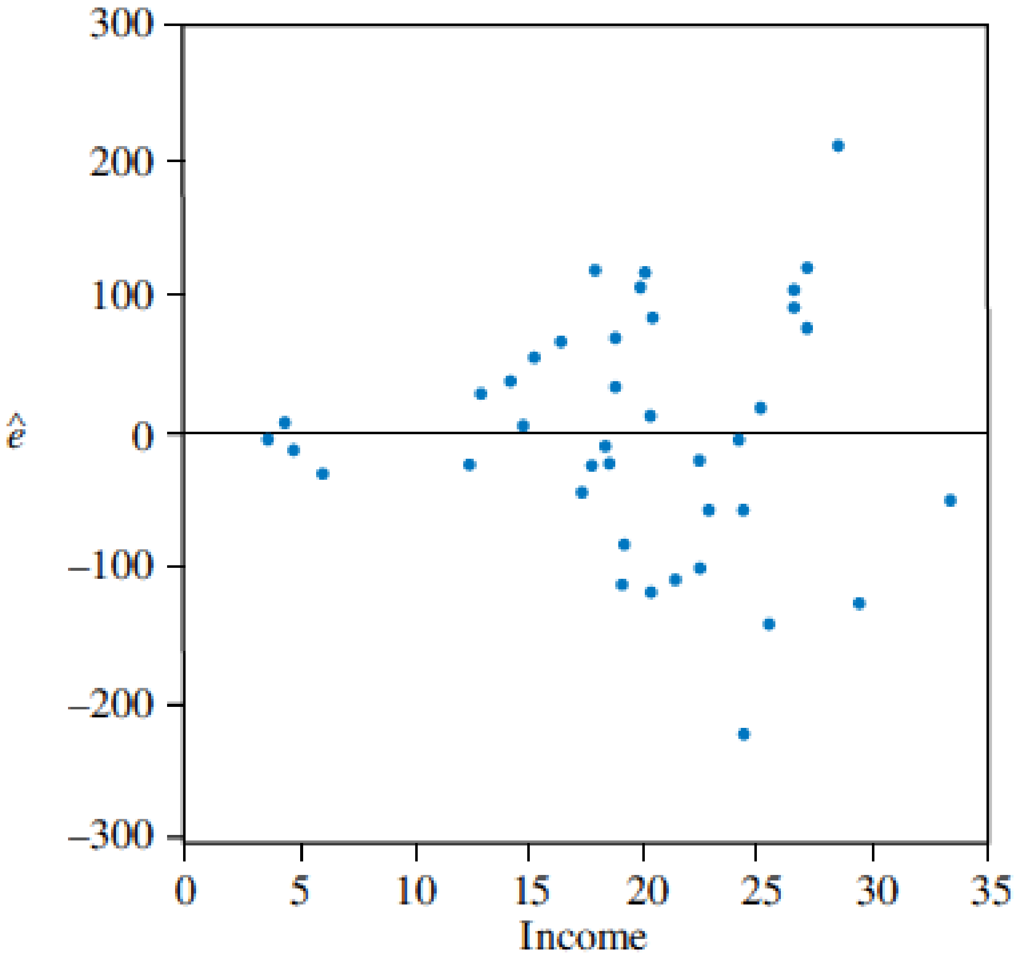
8.2
Detecting Heteroskedasticity

- There are two methods we can use to detect heteroskedasticity

1. An informal way using residual charts
2. A formal way using statistical tests

- 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 method of investigating heteroskedasticity can be followed for any simple regression
    - In a regression with more than one explanatory variable we can plot the least squares residuals against each explanatory variable, or against, $\hat{y}_i$, to see if they vary in a systematic way

FIGURE 8.3 Least squares food expenditure residuals plotted against income



- Let's develop a test based on a **variance function**
  - Consider the general multiple regression model:

Eq. 8.9

$$E(y_i) = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK}$$

  - A general form for the variance function related to Eq. 8.9 is:

Eq. 8.10

$$\text{var}(y_i) = \sigma_i^2 = E(e_i^2) = h(\alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_S z_{iS})$$

    - This is a general form because we have not been specific about the function $h(\bullet)$

- Two possible functions for $h(\bullet)$ are:
  - Exponential function:

Eq. 8.11

$$h(\alpha_1+\alpha_2 z_{i2}+\cdots+\alpha_S z_{iS})=\exp(\alpha_1+\alpha_2 z_{i2}+\cdots+\alpha_S z_{iS})$$

  - Linear function:

Eq. 8.12

$$h(\alpha_1+\alpha_2 z_{i2}+\cdots+\alpha_S z_{iS})=\alpha_1+\alpha_2 z_{i2}+\cdots+\alpha_S z_{iS}$$

    - In this latter case one must be careful to ensure $h(\bullet)>0$

- Notice that when

$$\alpha_2 = \alpha_3 = \cdots = \alpha_S = 0$$

then:

$$h\left(\alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_S z_{iS}\right) = h\left(\alpha_1\right)$$

But $h(\alpha_1)$ is a constant

- So, when:

$$\alpha_2 = \alpha_3 = \cdots = \alpha_S = 0$$

heteroskedasticity is not present

- The null and alternative hypotheses are:

Eq. 8.13

$$H_0: \ \alpha_2 = \alpha_3 = \cdots = \alpha_S = 0$$
$$H_1: \ \text{not all the } \alpha_i \text{ in } H_0 \text{ are zero}$$

- For the test statistic, use Eq. 8.10 and 8.12 to get:

Eq. 8.14

$$\operatorname{var}(y_i) = \sigma_i^2 = E(e_i^2) = \alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_S z_{iS}$$

– Letting

$$v_i = e_i^2 - E(e_i^2)$$

we get

Eq. 8.15

$$e_i^2 = E(e_i^2) + v_i = \alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_S z_{iS} + v_i$$

- This is like the general regression model studied earlier:

Eq. 8.16

$$y_i = E(y_i) + e_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + e_i$$

– Substituting the least squares residuals $\hat{e}_i^2$ for e_i^2 we get:

Eq. 8.17

$$\hat{e}_i^2 = \alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_S z_{iS} + v_i$$

- Since the R^2 from Eq. 8.17 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 has a chi-square (χ^2) distribution with S - 1 degrees of freedom:

Eq. 8.18

$$\chi^2 = N \times R^2 \sim \chi^2_{(S-1)}$$

- Important features of this test:
  - It is a large sample test
  - You will often see the test referred to as a **Lagrange multiplier test** or a **Breusch-Pagan test** for heteroskedasticity
  - The value of the statistic computed from the linear function is valid for testing an alternative hypothesis of heteroskedasticity where the variance function can be of any form given by Eq. 8.10

- The previous test presupposes that we have knowledge of the variables appearing in the variance function if the alternative hypothesis of heteroskedasticity is true
  - We may wish to test for heteroskedasticity without precise knowledge of the relevant variables
  - Hal White suggested defining the *z*'s as equal to the *x*'s, the squares of the *x*'s, and possibly their cross-products

- Suppose:

$$E(y)=\beta_1+\beta_2x_2+\beta_3x_3$$

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

$$z_2 = x_2 \quad z_3 = x_3 \quad z_4 = x_2^2 \quad z_5 = x_3^2$$

  – Including interactions adds one further variable

$$z_5 = x_2x_3$$

- The White test is performed as an F-test or using:

$$\chi^2 = N \times R^2$$

- 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 $\hat{e}_i^2 = \alpha_1 + \alpha_2 x_i + v_i$ by least squares
  - Save the R^2 which is:

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

  - Calculate:

$$\chi^2 = N \times R^2 = 40 \times 0.1846 = 7.38$$

- Since there is only one parameter in the null hypothesis, the χ-test has one degree of freedom.
  - The 5% critical value is 3.84
  - Because 7.38 is greater than 3.84, we reject H_0 and conclude that the variance depends on income

- For the White version, estimate:

$$\hat{e}_i^2 = \alpha_1 + \alpha_2 x_i + \alpha_3 x_i^2 + v_i$$

  - Test H_0: $\alpha_2 = \alpha_3 = 0$ against H_1: $\alpha_2 \neq 0$ or $\alpha_3 \neq 0$
  - Calculate:

$$\chi^2 = N \times R^2 = 40 \times 0.18888 = 7.555 \quad p - \text{value} = 0.023$$

  - The 5% critical value is $\chi_{(0.95,\, 2)} = 5.99$
    - Again, we conclude that heteroskedasticity exists

- The **Goldfeld-Quandt test** is designed to test for this form of heteroskedasticity, where the sample can be partitioned into two groups and we suspect the variance could be different in the two groups
  - Write the equations for the two groups as:

Eq. 8.20a

Eq. 8.20b

$$y_{Mi} = \beta_{M1} + \beta_{M2} x_{1Mi} + \ldots + \beta_{MK} x_{K\,Mi} + e_{Mi} \quad i = 1,\ 2,\ \ldots,\ N_M$$

$$y_{Ri} = \beta_{R1} + \beta_{R2} x_{2Ri} + \ldots + \beta_{RK} x_{K\,Ri} + e_{Ri} \quad i = 1,\ 2,\ \ldots,\ N_R$$

  - Test the null hypothesis:

$$\sigma_M^2 = \sigma_R^2$$

8.2.3
The Goldfeld-
Quandt Test

- The test statistic is:

Eq. 8.21

$$F = \frac{\hat{\sigma}_M^2 / \sigma_M^2}{\hat{\sigma}_R^2 / \sigma_R^2} \sim F_{(N_M - K_M, N_R - K_R)}$$

  – Usually $K_M = K_R = K$

  – Suppose we want to test:

Eq. 8.22

$$H_0 : \sigma_M^2 = \sigma_R^2 \text{ against } H_1 : \sigma_M^2 \neq \sigma_R^2$$

  – When H_0 is true, we have:

Eq. 8.23

$$F = \hat{\sigma}_M^2 / \hat{\sigma}_R^2$$

  – This is a two-tail test. We reject the null if

$F < F_{Lc} = F_{(a/2, N_M - K, N_R - K)}$ or $F > F_{Uc} = F_{(1-a/2, N_M - K, N_R - K)}$

8.3 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 best
  2. The usual least squares standard errors are incorrect, which invalidates interval estimates and hypothesis tests
     - There is a way of correcting the standard errors so that our interval estimates and hypothesis tests are valid

- Under heteroskedasticity:

Eq. 8.24

$$\operatorname{var}(b_2) = \frac{\sum_{i=1}^{N}\left[(x_i - \bar{x})^2 \sigma_i^2\right]}{\left[\sum_{i=1}^{N}(x_i - \bar{x})^2\right]^2}$$

- A consistent estimator for this variance has been developed and is known as:
  - White’s heteroskedasticity-consistent standard errors, or
  - Heteroskedasticity robust standard errors, or
  - Robust standard errors
    - The term “robust” is used because they are valid in large samples for both heteroskedastic and homoskedastic errors

- For K = 2, the White variance estimator is:

Eq. 8.25

$$\widehat{\operatorname{var}}\left(b_2\right)=\frac{N}{N-2}\frac{\sum_{i=1}^{N}\left[\left(x_i-\bar{x}\right)^2\hat{e}_i^2\right]}{\left[\sum_{i=1}^{N}\left(x_i-\bar{x}\right)^2\right]^2}$$

- For the food expenditure example:

$$\begin{aligned}\hat{y} &= 83.42 + 10.21x \\ &\quad (27.46)\ (1.81) \quad (\text{White se}) \\ &\quad (43.41)\ (2.09) \quad (\text{incorrect se})\end{aligned}$$

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

White: $b_2 \pm t_c \text{se}\left(b_2\right) = 10.21 \pm 2.024 \times 1.81 = \left[6.55, 13.87\right]$

Incorrect: $b_2 \pm t_c \text{se}\left(b_2\right) = 10.21 \pm 2.024 \times 2.09 = \left[5.97, 14.45\right]$

- White’s estimator for the standard errors helps avoid computing incorrect interval estimates or incorrect values for test statistics in the presence of heteroskedasticity
  - It does not address the other implication of heteroskedasticity: the least squares estimator is no longer best
    - Failing to address this issue may not be that serious
    - With a large sample size, the variance of the least squares estimator may still be sufficiently small to get precise estimates
      - To find an alternative estimator with a lower variance it is necessary to specify a suitable variance function
      - Using least squares with robust standard errors avoids the need to specify a suitable variance function

8.4 Generalized Least Squares: Known Form of Variance

8.4.1
Variance Proportional to x

- Recall the food expenditure example with heteroskedasticity:

Eq. 8.26

$$y_i = \beta_1 + \beta_2 x_i + e_i$$

$$E(e_i) = 0, \quad \text{var}(e_i) = \sigma_i^2, \quad \text{cov}(e_i, e_j) = 0 \quad (i \neq j)$$

  – To develop an estimator that is better than the least squares estimator we need to make a further assumption about how the variances $\sigma^2{}_{\text{i}}$ change with each observation

- An estimator known as the **generalized least squares estimator**, depends on the unknown σ^2_i
  - To make the generalized least squares estimator operational, some structure is imposed on σ^2_i
  - One possibility:

Eq. 8.27

$$\text{var}\left(e_i\right) = \sigma_i^2 = \sigma^2 x_i$$

8.4.1a
Transforming the Model

- We change or transform the model into one with homoskedastic errors:

Eq. 8.28

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

- Define the following transformed variables:

Eq. 8.29

$$y_i^* = \frac{y_i}{\sqrt{x_i}}, \quad x_{i1}^* = \frac{1}{\sqrt{x_i}}, \quad x_{i2}^* = \frac{x_i}{\sqrt{x_i}} = \sqrt{x_i}, \quad e_i^* = \frac{e_i}{\sqrt{x_i}}$$

– Our model is now:

Eq. 8.30

$$y_i^* = \beta_1 x_{i1}^* + \beta_2 x_{i2}^* + e_i^*$$

- The new transformed error term is homoskedastic:

Eq. 8.31

$$\operatorname{var}\left(e_i^*\right) = \operatorname{var}\left(\frac{e_i}{\sqrt{x_i}}\right) = \frac{1}{x_i}\operatorname{var}\left(e_i\right) = \frac{1}{x_i}\sigma^2 x_i = \sigma^2$$

  – The transformed error term will retain the properties of zero mean and zero correlation between different observations

- To obtain the best linear unbiased estimator for a model with heteroskedasticity of the type specified in Eq. 8.27:
  1. Calculate the transformed variables given in Eq. 8.29
  2. Use least squares to estimate the transformed model given in Eq. 8.30
- The estimator obtained in this way is called a generalized least squares estimator

- One way of viewing the generalized least squares estimator is as a **weighted least squares** estimator
  - Minimizing the sum of squared transformed errors:

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

  - The errors are weighted by $x_i^{-1/2}$

- Applying the generalized (weighted) least squares procedure to our food expenditure problem:

Eq. 8.32

$$\begin{aligned} \hat{y}_i &= 78.68 + 10.45 x_i \\ (se) &\ \ (23.79) \quad (1.39) \end{aligned}$$

  – A 95% confidence interval for β_2 is given by:

$$\hat{\beta}_2 \pm t_c \mathrm{se}\left(\hat{\beta}_2\right) = 10.451 \pm 2.024 \times 1.386 = [7.65, 13.26]$$

8.5 Generalized Least Squares: Unknown Form of Variance

- Consider a more general specification of the error variance:

Eq. 8.37

$$\operatorname{var}(e_i) = \sigma_i^2 = \sigma^2 x_i^\gamma$$

where γ is an unknown parameter

- To handle this, take logs

$$\ln(\sigma_i^2) = \ln(\sigma^2) + \gamma \ln(x_i)$$

and then take the exponential:

Eq. 8.38

$$\sigma_i^2 = \exp\left(\ln(\sigma^2) + \gamma \ln(x_i)\right)$$

$$= \exp(\alpha_1 + \alpha_2 z_i)$$

- We can extend this function to:

Eq. 8.39

$$\sigma_i^2 = \exp(\alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_S z_{iS})$$

Eq. 8.40

- Now write Eq. 8.38 as:

$$\ln\left(\sigma_i^2\right) = \alpha_1 + \alpha_2 z_i$$

  - To estimate α_1 and α_2 we recall our basic model:

$$y_i = E\left(y_i\right) + e_i = \beta_1 + \beta_2 x_i + e_i$$

  - that we estimate it using least squares and keep the residuals $\hat{e}_i$

- Apply the least squares strategy to Eq. 8.40 using $\hat{e}_i^2$:

Eq. 8.41

$$\ln(\hat{e}_i^2) = \ln(\sigma_i^2) + v_i = \alpha_1 + \alpha_2 z_i + v_i$$

- For the food expenditure data, we have:

$$\widehat{\ln(\sigma_i^2)} = 0.9378 + 2.329 z_i$$

- Again, given the estimates $\hat{\sigma}_i$ then we can divide both sides of the regression model by $\hat{\sigma}_i$ to obtain the generalized least squares estimators.

- This works because dividing Eq. 8.26 by σ_i yields:

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

  - The error term is homoskedastic:

Eq. 8.42

$$\operatorname{var}\left(\frac{e_i}{\sigma_i}\right)=\left(\frac{1}{\sigma_i^2}\right)\operatorname{var}(e_i)=\left(\frac{1}{\sigma_i^2}\right)\sigma_i^2=1$$

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

Eq. 8.43

$$y_i^* = \left(\frac{y_i}{\hat{\sigma}_i}\right) \qquad x_{i1}^* = \left(\frac{1}{\hat{\sigma}_i}\right) \qquad x_{i2}^* = \left(\frac{x_i}{\hat{\sigma}_i}\right)$$

and apply least squares to:

Eq. 8.44

$$y_i^* = \beta_1 x_{i1}^* + \beta_2 x_{i2}^* + e_i^*$$

- To summarize for the general case, suppose our model is:

Eq. 8.45

$$y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_k x_{iK} + e_i$$

where:

Eq. 8.46

$$\text{var}(e_i) = \sigma_i^2 = \exp(\alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_s z_{iS})$$

- The steps for obtaining a generalized least squares estimator are:
  1. Estimate Eq. 8.45 by least squares and compute the squares of the least squares residuals $\hat{e}_i^2$
  2. Estimate $\alpha_1, \alpha_2, \ldots, \alpha_S$ by applying least squares to the equation $\ln \hat{e}_i^2 = \alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_S z_{iS} + v_i$
  3. Compute variance estimates $\hat{\sigma}_i^2 = \exp(\hat{\alpha}_1 + \hat{\alpha}_2 z_{i2} + \cdots + \hat{\alpha}_S z_{iS})$
  4. Compute the transformed observations defined by Eq. 8.43, including $x_{i3}^*, \ldots, x_{iK}^*$ if $K > 2$
  5. Apply least squares to Eq. 8.44, or to an extended version of (8.44), if $K > 2$

Eq. 8.47

- Following these steps for our food expenditure problem:

$$\hat{y}_i = 76.05 + 10.63x$$
$$\text{(se)} \quad (9.71) \; (0.97)$$

  - The estimates for β_1 and β_2 have not changed much
  - There has been a considerable drop in the standard errors that under the previous specification were

$$\text{se}\left(\hat{\beta}_1\right) = 23.79 \text{ and } \text{se}\left(\hat{\beta}_2\right) = 1.39$$

- Robust standard errors can be used not only to guard against the possible presence of heteroskedasticity when using least squares, they can be used to guard against the possible misspecification of a variance function when using generalized least squares

Key Words

Keywords

- Breusch–Pagan test
- generalized least squares
- Goldfeld–Quandt test
- Heteroskedasticity
- Heteroskedasticity-consistent standard errors
- homoskedasticity
- Lagrange multiplier test
- mean function
- residual plot
- transformed model
- variance function
- weighted least squares
- White test