CHAPTER 8

Heteroskedasticity

LEARNING OBJECTIVES

Based on the material in this chapter, you should be able to

1. Explain the meaning of heteroskedasticity and give examples of data sets likely to exhibit heteroskedasticity.
2. Explain how and why plots of least squares residuals can reveal heteroskedasticity.
3. Specify a variance function and use it to test for heteroskedasticity with (a) a Breusch–Pagan test and (b) a White test.
4. Test for heteroskedasticity using a Goldfeld–Quandt test applied to (a) two subsamples with potentially different variances and (b) a model where the variance is hypothesized to depend on an explanatory variable.
5. Describe and compare the properties of the least squares and generalized least squares estimators when heteroskedasticity exists.
6. Compute heteroskedasticity-consistent standard errors for least squares.
7. Describe how to transform a model to eliminate heteroskedasticity.
8. Compute generalized least squares estimates for heteroskedastic models where (a) the variance is known except for the proportionality constant σ^2, (b) the variance is a function of explanatory variables and unknown parameters, and (c) the sample is partitioned into two groups with different variances.
9. Explain why the linear probability model exhibits heteroskedasticity.
10. Compute generalized least squares estimates of the linear probability model.

KEYWORDS

Breusch–Pagan test
generalized least squares
Goldfeld–Quandt test
grouped heteroskadasticity
heteroskedasticity
heteroskedasticity-consistent standard errors
homoskedasticity
Lagrange multiplier test
linear probability model
regression function
residual plot
robust standard errors
skedastic function
transformed model
variance function
weighted least squares
White test

8.1 The Nature of Heteroskedasticity

In Chapter 2, we discussed the relationship between household food expenditure and household income. We proposed the simple population regression model

$$FOOD\_EXP_i = \beta_1 + \beta_2 INCOME_i + e_i \tag{8.1}$$

Given the parameter values, β_1 and β_2, we can predict food expenditures for households with any income. Income is an important factor in households' decisions about weekly food expenditure, but there are many other factors entering a particular household's decisions. The random error e_i represents the collection of all the factors other than income that affect household expenditure on food.

The assumption of **strict exogeneity** says that when using information on household income our best prediction of the random error is zero. If sample values are randomly selected, then the technical expression for this assumption is that given income the conditional expected value of the random error e_i is zero, $E(e_i|INCOME_i) = 0$. If the assumption of strict exogeneity holds then the regression function is

$$E(FOOD\_EXP_i|INCOME_i) = \beta_1 + \beta_2 INCOME_i$$

The slope parameter β_2 describes how expected (population mean, or average) household food expenditure changes when household income increases by \$100, holding all else constant. The intercept parameter β_1 measures average expenditure on food for a household with no income in a week.

The discussion above focuses on the level, or amount, of food expenditure. We now ask, "How much **variation** in household food expenditure is there at different levels of income?" The U.S. median household income is about \$1000 a week. For such a household, the expected weekly food expenditure is $E(FOOD\_EXP_i|INCOME = 10) = \beta_1 + \beta_2(10)$. If we observe many households with the median income, we would observe a wide range of actual weekly food expenditures. The variation arises because different households have differing tastes and preferences, and they have differing demographic characteristics, and life circumstances. Readers who are students, and living on typical student incomes, how much variation is there in your food expenditure from week to week? We suspect that regardless of your tastes and preferences you have calculated very carefully how much you can afford and stick closely to a spending plan each week. In general, households with low incomes have little scope for wide variations in food expenditures from week to week because of their income constraint. On the other hand, households with a large weekly income have more food choices. Some high-income households may choose champagne, caviar, and steaks, but others may choose beer, rice, pasta, and beans. We can expect to observe larger variations in weekly food expenditures by households with large incomes.

Holding income constant, and given our model, what is the source of the variation in household food expenditures? It must be from the random error, the collection of factors, other than income, that influence food expenditure. As we observe different households at a given level of income, there are variations in food expenditures because randomly sampled households have different tastes and preferences and differ in many other ways as well. Recall that the random error in the regression is the difference between any observation on the outcome variable and its conditional expectation, that is

$$e_i = FOOD\_EXP_i - E(FOOD\_EXP_i|INCOME_i) \tag{8.2}$$

If the assumption of strict exogeneity holds, then the population average value of the random errors is $E(e_i|INCOME_i) = E(e_i) = 0$. A positive random error corresponds to an observation in which food expenditure is greater than expected, while a negative random error corresponds to an observation in which food expenditure is less than expected.

Another way of describing the greater variation in food expenditures for high-income households is to say the probability of observing large positive or negative random errors is higher

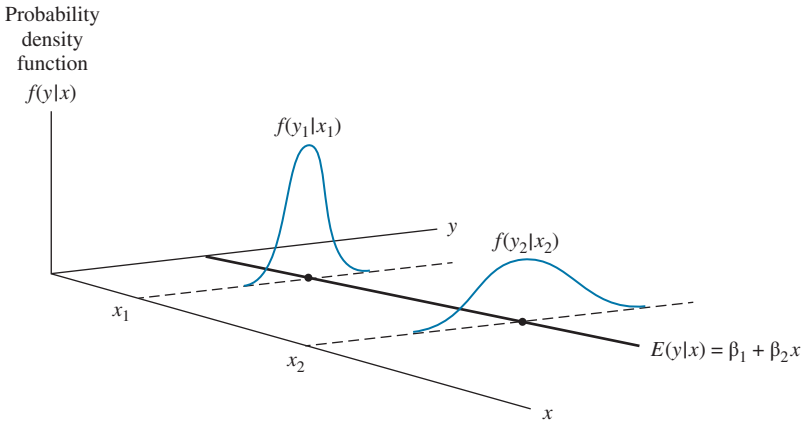

FIGURE 8.1 **Heteroskedastic errors.**

for high incomes than it is for low incomes. To illustrate this idea, examine Figure P.5 in the Probability Primer. First, suppose the probability distribution of the random errors is $N(0,1)$, the solid curve. What is the probability of observing a random value of e_i greater than two? Using Statistical Table 1, $P(e_i > 2) = P(Z > 2) = 0.0228$. Now, suppose the probability distribution of the random errors is $N(0, 4)$, the dot-dash curve. What is the probability of observing a random value of e_i greater than 2? Using Statistical Table 1, $P(e_i > 2) = P(Z > 1) = 0.1587$. The random error e_i has a higher probability of taking on a large value if its variance is large. In the context of the food expenditure example, we can capture the effect we are describing by assuming that $\text{var}(e_i|INCOME_i)$ increases as income increases. Food expenditure can deviate further from its mean, or expected value, when income is large.

In such a case, when the error variances for all observations are not the same, we say that **heteroskedasticity** exists. Alternatively, we say the random error e_i is **heteroskedastic**. Conversely, if all observations come from probability density functions with the same variance, we say that **homoskedasticity** exists, and e_i is **homoskedastic**. Heteroscedastic, homoscedastic, and heteroscedasticity are commonly used alternative spellings.

Figure 8.1 illustrates the heteroskedastic assumption. Let $y_i = FOOD\_EXP_i$ and $x_i = INCOME_i$. At x_1, the food expenditure probability density function $f(y_1|x_1)$ is such that y_1 will be close to $E(y_1|x_1)$ with high probability. When we move to the larger value x_2, the probability density function $f(y_2|x_2)$ is more spread out; we are less certain about where y_2 might fall, and much larger or smaller values than the average $E(y_2|x_2)$ are possible. When homoskedasticity exists, the probability density function for the errors does not change as x changes, as we illustrated in Figure 2.3.

8.2 Heteroskedasticity in the Multiple Regression Model

The existence of heteroskedasticity is a violation of one of our least squares assumptions listed in Section 5.1. For the multiple regression model $y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + e_i$, $i = 1, \ldots, N$, assumption MR3 is

$$\text{var}(e_i|\mathbf{X}) = \text{var}(y_i|\mathbf{X}) = \sigma^2$$

the conditional variance of the random error, and the dependent variable, is σ^2, a constant. Assumption MR3 is that the random error term is conditionally homoskedastic. The simplest

statement of the conditional heteroskedasticity assumption is

$$\text{var}(e_i|\mathbf{X}) = \text{var}(y_i|\mathbf{X}) = \sigma_i^2 \tag{8.3}$$

The change is very subtle, the error variance σ_i^2 now has a subscript, i, indicating that it is not always the same constant and may change from observation to observation, $i = 1, \ldots, N$. At the extreme, the error is heteroskedastic even if only one random error has a variance different than the other $N - 1$ random errors. Generally, however, we think of the problem as being more pervasive when it is present.

Assumptions MR1–MR5 apply to any type of regression, using time-series or cross-sectional data. Our notation $\mathbf{X}$ represents all N observations on $K - 1$ explanatory variables plus a constant term. Heteroskedasticity often arises 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*. The household data on income and food expenditure fall into this category. Other possible examples include data on costs, outputs, and inputs for a number of firms, and data on quantities purchased and prices for some commodity, or commodities, in a number of retail establishments. Cross-sectional data usually involve observations on economic units of varying sizes. For example, data on households will involve households with varying numbers of household members and different levels of household income. With data on a number of firms, we might measure the size of the firm by the quantity of output it produces. Frequently, the larger the firm, or the larger the household, the more difficult it is to explain the variation in some outcome variable y_i by the variation in a set of explanatory variables. Larger firms and households are likely to be more diverse and flexible with respect to the way in which values for y_i are determined. What this means for the linear regression model is that, as the size of the economic unit becomes larger, there is more uncertainty associated with the outcomes y_i. We model this greater uncertainty by specifying a conditional 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 an economic unit, such as a firm, a household, or even a whole economy, it is possible that the conditional error variance will change. This would be true if there was an external shock or change in circumstances that created more or less uncertainty about y.

For simplification, in the remainder of this chapter, we assume that the errors are uncorrelated and that heteroskedasticity is an observation-by-observation problem and that the conditional variance of the ith observation's random error e_i is unrelated to the jth observation. In the context of the cross-sectional data food expenditure example, we are ruling out the case in which the variability in the random error component for the ith household is connected to or explained by the characteristics of the jth household. In a time-series regression context, we are ruling out the case when the error variation at time t is related to conditions in the past, at time $t - s$. Can we always rule out these exceptions? No, we cannot. In the cross-sectional data context, we may find that households drawn from some geographical regions, or neighborhoods, are similar, so that the error variation for neighboring households might be similar, or connected. In the time-series context, we most certainly cannot rule out continuous periods of stability, perhaps many weeks at a time, and periods of instability that can similarly last many weeks or months, meaning that the error variation at time t is related to the error variation at times $t - 1$, $t - 2$, and so on. For now, however, we will rule out these interesting cases.

8.2.1 The Heteroskedastic Regression Model

The multiple regression model is $y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + e_i$. We assume we have a random sample so that the ith observation is statistically independent of the jth observation. Let $x_i = (1, x_{i2}, \ldots, x_{iK})$ denote the values of the K explanatory variables for the ith observation. The heteroskedasticity assumption in (8.3) becomes

$$\text{var}(y_i|\mathbf{x}_i) = \text{var}(e_i|\mathbf{x}_i) = \sigma^2 h(\mathbf{x}_i) = \sigma_i^2 \tag{8.4}$$

where $h(\mathbf{x}_i) > 0$ is a function of $\mathbf{x}_i$ that is sometimes called the **skedastic function**,[1] and $\sigma^2 > 0$ is a constant. If $h(\mathbf{x}_i) = 1$, then the conditional variance is homoskedastic. If $h(\mathbf{x}_i)$ is not constant, then the conditional variance is heteroskedastic. For example, when $h(\mathbf{x}_i) = x_{ik}$ the conditional variance becomes $\text{var}(e_i|\mathbf{x}_i) = \sigma^2 x_{ik}$, the error variance is proportional to the kth explanatory variable x_{ik}. Because variances must be positive, for the proportional heteroskedasticity model to work $h(\mathbf{x}_i) = x_{ik} > 0$. In (8.4) we assume the conditional variance depends on the values of some or all of the explanatory variables in the regression equation.

This chapter is concerned with the consequences of a variance assumption like (8.4). What are the consequences for the properties of least squares estimator? Is there a better estimation technique? How do we detect the existence of heteroskedasticity?

EXAMPLE 8.1 | Heteroskedasticity in the Food Expenditure Model

We can further illustrate the nature of heteroskedasticity and at the same time demonstrate an informal way of detecting heteroskedasticity using the food expenditure data. Using the $N = 40$ observations in the data file *food,* the OLS estimates are

$$\widehat{FOOD\_EXP}_i = 83.42 + 10.21\,INCOME_i$$

A graph of this fitted line, along with all the observed expenditure–income points, appears in Chapter 2, Figure 2.8. Notice that, as income grows, the prevalence of data points that deviate further from the estimated mean function increases. There are more points scattered further away from the line as income gets larger. Another way of describing this feature is to say that there is a tendency for the least squares residuals, defined by

$$\hat{e}_i = FOOD\_EXP_i - 83.42 - 10.21 INCOME_i$$

to increase in absolute value as income grows. The plot of the absolute value of the residuals, $|\hat{e}_i|$, versus income in Figure 8.2 shows this quite clearly. The plot of the calculated residuals, $\hat{e}_i$, versus income in Figure 8.3 shows the characteristic "spray" pattern shown in Chapter 4, Figure 4.7(b). Figure 4.7(a) shows the random scatter we anticipate if the errors are conditionally homoskedastic. Figures 4.7(b)–(d), spray, funnel, and bowtie, are patterns we might observe when the errors are conditionally heteroskedastic.

Since the observable least squares residuals $(\hat{e}_i)$ are the analogues of the unobservable errors (e_i), Figures 8.2 and 8.3 also suggest that the unobservable errors tend to increase in absolute value as income increases. That is, the variation of food expenditure around the conditional mean food expenditure $E(FOOD\_EXP_i|INCOME_i) = \beta_1 + \beta_2 INCOME_i$, and variation in the random error term, increase as income increases. The conditional variance $\text{var}(e_i|INCOME_i) = \sigma^2 h(INCOME_i)$, where $h(INCOME_i)$ is an increasing function of income. Possible variance functions include

$$\text{var}(e_i|INCOME_i) = \sigma^2 INCOME_i$$

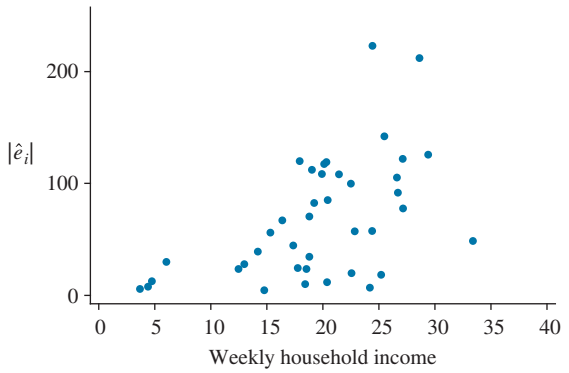

FIGURE 8.2 **Absolute value of food expenditure residuals vs. income.**

[1] See A. Colin Cameron and Pravin K. Trivedi (2010) *Microeconometrics Using Stata, Revised Edition*, Stata Press, p. 153.

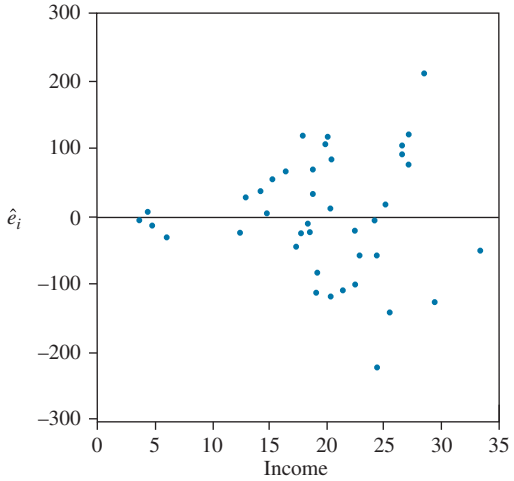

FIGURE 8.3 **Least squares food expenditure residuals plotted against income.**

or

$$\text{var}(e_i|INCOME_i) = \sigma^2 INCOME_i^2$$

These are consistent with the hypothesis that we posed earlier, namely, that the mean food expenditure function is better at explaining food expenditure for low-income households than it is for high-income households.

Plotting of least squares residuals is an informal way of detecting heteroskedasticity. Later in the chapter, in Section 8.6, we consider formal test procedures. First, however, we examine the consequences of heteroskedasticity for least squares estimation.

8.2.2 Heteroskedasticity Consequences for the OLS Estimator

Since the existence of heteroskedasticity violates the usual least squares assumption $\text{var}(e_i|\mathbf{x}_i) = \sigma^2$, we need to ask what consequences this violation has for our least squares estimator, and what we can do about it. There are two implications:

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.

Let's first consider the simple linear regression model with homoskedasticity

$$y_i = \beta_1 + \beta_2 x_i + e_i, \text{ with } \text{var}(e_i|\mathbf{x}) = \sigma^2 \tag{8.5}$$

We showed in Chapter 2 that the conditional variance of the least squares estimator for b_2 is

$$\text{var}(b_2|\mathbf{x}) = \sigma^2 \Big/ \sum_{i=1}^{N} (x_i - \bar{x})^2 \tag{8.6}$$

Now suppose the error variances for each observation are different and that we recognize this difference by putting a subscript i on σ^2, so that we have

$$y_i = \beta_1 + \beta_2 x_i + e_i, \text{ with } \text{var}(e_i|\mathbf{x}) = \sigma_i^2 \tag{8.7}$$

In Appendix 8A, we show that under the heteroskedastic specification in (8.7) the least squares estimator is unbiased with conditional variance

$$\text{var}(b_2|\mathbf{x}) = \left[\sum_{i=1}^{N}(x_i-\bar{x})^2\right]^{-1}\left[\sum_{i=1}^{N}(x_i-\bar{x})^2\sigma_i^2\right]\left[\sum_{i=1}^{N}(x_i-\bar{x})^2\right]^{-1} \tag{8.8}$$

If the errors are homoskedastic, then equation (8.8) reduces to the usual OLS estimator variance in equation (8.6). If the errors are heteroskedastic then (8.8) is correct and (8.6) is not. This is a practical problem because your computer software has programmed into it the estimated variances and covariance of the least squares estimator under homoskedasticity, given in Chapter 2, equations (2.20)–(2.22). This in turn means that if the errors are heteroskedastic, the usual standard errors in equations (2.23)–(2.24) are incorrect. Using incorrect standard errors in t-tests and confidence intervals may lead us to faulty conclusions. If we proceed to use the least squares estimator and its usual standard errors when $\text{var}(e_i) = \sigma_i^2$, we will be using an estimate of (8.6) to compute the standard error of b_2 when we should be using an estimate of (8.8).

8.3 Heteroskedasticity Robust Variance Estimator

Calculation of a correct estimate for the OLS variance (8.8) is astonishingly simple, although the theory leading to it is not. Simply replace σ_i^2 by $[N/(N-2)]\,\hat{e}_i^2$, the squared OLS residuals multiplied by an inflation factor.[2] The **White heteroskedasticity-consistent estimator (HCE)** that is valid in large samples for the simple regression model is

$$\widehat{\text{var}}(b_2) = \left[\sum(x_i-\bar{x})^2\right]^{-1}\left\{\sum\left[(x_i-\bar{x})^2\left(\frac{N}{N-2}\right)\hat{e}_i^2\right]\right\}\left[\sum(x_i-\bar{x})^2\right]^{-1} \tag{8.9}$$

where $\hat{e}_i$ is the least squares residual from the regression model, $y_i = \beta_1 + \beta_2 x_i + e_i$. The estimator is named after econometrician Halbert White who developed the idea. This variance estimator is **robust** because it is valid whether heteroskedasticity is present or not. Thus, if we are not sure whether the random errors are heteroskedastic or homoskedastic, then we can use a robust variance estimator and be confident that our standard errors, t-tests, and interval estimates are valid in large samples.

The formula in equation (8.9) has a lovely symmetry and is one illustration of a **variance sandwich**. Let $C = \left[\sum(x_i-\bar{x})^2\right]^{-1}$ be the "outside ***C***rust" and let $A = \left\{\sum\left[(x_i-\bar{x})^2\left(\frac{N}{N-2}\right)\hat{e}_i^2\right]\right\}$ be "***A***ny filling." Then our variance sandwich is *any filling* between two *crusts*, or $\widehat{\text{var}}(b_2) = CAC$. Modern Econometrics offers many such sandwiches. Equations (8.8) and (8.9) can be simplified, but we prefer to leave them as is to emphasize the "sandwich" form. Also the matrix approaches to multiple regression in your future econometrics courses will use the sandwich form.

EXAMPLE 8.2 | Robust Standard Errors in the Food Expenditure Model

Most regression packages include an option for calculating standard errors using White's estimator. If we do so for the food expenditure example, we obtain

$$\widehat{FOOD\_EXP} = 83.42 + 10.21\,INCOME$$
$$(27.46)\quad (1.81)\quad \text{(White robust se)}$$
$$(43.41)\quad (2.09)\quad \text{(incorrect OLS se)}$$

In this case, ignoring heteroskedasticity and using incorrect standard errors, based on the usual formula in (8.6), tends to understate the precision of estimation; we tend to get confidence intervals that are wider than they should be. Specifically, following the result in (3.6) in Chapter 3, we can construct corresponding 95% confidence intervals for β_2.

[2] See Appendix 8C for the logic of this inflation, and development of other versions of the robust variance.

White Robust se:

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

Incorrect OLS se:

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

If we ignore heteroskedasticity, we estimate that β_2 lies between 5.97 and 14.45. When we recognize the existence of heteroskedasticity, our information is judged more precise, and using the **robust standard error** we estimate that β_2 lies between 6.55 and 13.87. A caveat here is that the sample is small, which does mean that the robust standard error formula we have provided may not be as accurate as if the sample were large.

White's estimator for the standard errors helps us avoid computing incorrect interval estimates or incorrect values for test statistics in the presence of heteroskedasticity. However, it does not address the first implication of heteroskedasticity that we mentioned at the beginning of this section, that the least squares estimator is no longer best. However, failing to use the "best" estimator may not be too grave a sin if estimates are sufficiently precise for useful economic analysis. Many cross-sectional data sets have thousands of observations, resulting in robust standard errors that are small, making interval estimates narrow and *t*-tests powerful. Nothing further is required in these cases. If, however, your estimates are not sufficiently precise for economic analysis, then a better, more efficient, estimator is called for. In order to use such an estimator we must specify the **skedastic** function $h(\mathbf{x}_i) > 0$, a function of $\mathbf{x}_i$ and also perhaps other variables, that describes the pattern of conditional heteroskedasticity. In the next section, we describe an alternative estimator that has a smaller variance than the least squares estimator.

8.4 Generalized Least Squares: Known Form of Variance

To begin, consider the simple regression model $y_i = \beta_1 + \beta_2 x_i + e_i$. Let's assume the data are obtained by random sampling, so that the observations are statistically independent of one another, that $E(e_i|x_i) = 0$, and that the heteroskedasticity assumption is

$$\text{var}(e_i|x_i) = \sigma^2 h(x_i) = \sigma_i^2 \tag{8.10}$$

Although it is possible to obtain the White heteroskedasticity-consistent variance estimates by simply assuming the error variances σ_i^2 can be different for each observation, to develop an estimator that is better than the least squares estimator, we need to make a further assumption about how the variances σ_i^2 change with each observation. This means making an assumption about the skedastic function $h(x_i)$. The further assumption is necessary because the best linear unbiased estimator in the presence of heteroskedasticity, an estimator known as the **generalized least squares (GLS)** estimator, depends on the unknown σ_i^2. It is not practical to estimate N unknown variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_N^2$ with only N observations without making a restrictive assumption about how the σ_i^2 change. Thus, to make the GLS estimator operational some structure is imposed on σ_i^2. Alternative structures are considered in this and the following section. Details of the GLS estimator and the issues involved will become clear as we work our way through these sections.

8.4.1 Transforming the Model: Proportional Heteroskedasticity

Recall our earlier inspection of the least squares residuals for the food expenditure example. The variation in the OLS residuals increases as income increases, which suggests that the error

variance increases as income increases. One possible assumption for the variance σ_i^2 that has this characteristic is

$$\text{var}(e_i|x_i) = \sigma_i^2 = \sigma^2 h(x_i) = \sigma^2 x_i, \quad x_i > 0 \tag{8.11}$$

That is, we assume that the variance of the ith error term σ_i^2 is given by a positive unknown constant parameter σ^2 multiplied by the positive income variable x_i, so that $\text{var}(e_i|x_i)$ is **proportional** to income. We are assuming the skedastic function is $h(x_i) = x_i$. As explained earlier, in economic terms this assumption implies that, for low levels of income (x_i), food expenditure (y_i) will be clustered closer to the **regression function** $E(y_i|x_i) = \beta_1 + \beta_2 x_i$. Expenditure on food for low-income households will be largely explained by the level of income. At high levels of income, food expenditures can deviate more from the regression function. This means that there are likely to be many other factors, such as specific tastes and preferences, that reside in the error term, and that lead to a greater variation in food expenditure for high-income households.

The least squares estimator is **not** the best linear unbiased estimator when the errors are heteroskedastic. Is there a best linear unbiased estimator under these circumstances? Yes there is! The approach is to **transform the model** into one with homoskedastic errors. Leaving the basic structure of the model intact, we turn the heteroskedastic error model into a homoskedastic error model. After the transformation, applying OLS to the **transformed model** gives a best linear unbiased estimator. These steps define the new GLS estimator.

Given the model of proportional heteroskedasticity in equation (8.11), begin by dividing both sides of the original model in (8.7) by $\sqrt{x_i}$

$$\frac{y_i}{\sqrt{x_i}} = \beta_1\left(\frac{1}{\sqrt{x_i}}\right) + \beta_2\left(\frac{x_i}{\sqrt{x_i}}\right) + \frac{e_i}{\sqrt{x_i}} \tag{8.12}$$

Define the **transformed variables** and **transformed error** as

$$y_i^* = \frac{y_i}{\sqrt{x_i}}, \quad x_{i1}^* = \frac{1}{\sqrt{x_i}}, \quad x_{i2}^* = \frac{x_i}{\sqrt{x_i}} = \sqrt{x_i}, \quad e_i^* = \frac{e_i}{\sqrt{x_i}} \tag{8.13}$$

so that (8.12) can be rewritten as

$$y_i^* = \beta_1 x_{i1}^* + \beta_2 x_{i2}^* + e_i^* \tag{8.14}$$

The beauty of this transformed model is that the new transformed error term e_i^* is homoskedastic. To see this, recall equation (P.14) from the Probability Primer: If X is a random variable and a is a constant, then $\text{var}(aX) = a^2\text{var}(X)$. Applying that rule here we have

$$\text{var}(e_i^*|x_i) = \text{var}\left(\frac{e_i}{\sqrt{x_i}}\middle| x_i\right) = \frac{1}{x_i}\text{var}(e_i|x_i) = \frac{1}{x_i}\sigma^2 x_i = \sigma^2 \tag{8.15}$$

Using the rules of expected values, the transformed error term will retain a zero conditional mean $E(e_i^*|x_i) = 0$. As a consequence, we can apply OLS to the transformed variables, y_i^*, x_{i1}^*, and x_{i2}^* to obtain the best linear unbiased estimator for β_1 and β_2. Note that the transformed variables y_i^*, x_{i1}^*, and x_{i2}^* are easy to create. An important difference between the original and transformed models is that the transformed model no longer contains a constant term. In the original model, $x_{i1} = 1$. In the transformed model, the variable $x_{i1}^* = 1/\sqrt{x_i}$ is no longer constant. You will have to be careful to exclude the constant if your software automatically inserts one, but you can still proceed. The transformed model is linear in the unknown parameters β_1 and β_2. These are the original parameters that we are interested in estimating. They are unaffected by the transformation. In short, the transformed model is a linear model to which we can apply OLS estimation. The transformed model satisfies the conditions of the Gauss–Markov theorem, and the OLS estimators defined in terms of the transformed variables are BLUE.

To summarize, to obtain the best linear unbiased estimator for a model with heteroskedasticity of the type specified in equation (8.11), $\text{var}(e_i|x_i) = \sigma_i^2 = \sigma^2 h(x_i) = \sigma^2 x_i$:

1. Calculate the transformed variables given in (8.13).
2. Use OLS to estimate the transformed model given in (8.14), yielding estimates $\hat{\beta}_1$ and $\hat{\beta}_2$.

The estimates obtained in this way are the GLS estimates.

The GLS estimator is BLUE if the model assumption of proportional heteroskedasticity is correct. Of course, we never know if our assumed skedastic function is correct or not. It is likely that a thoughtfully chosen transformation will reduce the model heteroskedasticity. If, however, the chosen transformation does not completely eliminate the heteroskedasticity, the GLS estimator is linear and unbiased but not best, and the standard errors from the transformed model estimation are incorrect. What then? Easy. Use White robust standard errors with the transformed data model to obtain valid (in large samples) standard errors. Doing so we will have striven for a more efficient estimator, but been cautious to present valid standard errors, *t*-stats, and interval estimates. We illustrate this strategy in Example 8.3.

8.4.2 Weighted Least Squares: Proportional Heteroskedasticity

One way of viewing the GLS estimator is as a **weighted least squares (WLS)** estimator. Recall that the OLS estimates are those values of β_1 and β_2 that minimize the sum of squared errors

$$S(\beta_1, \beta_2|y_i, x_i) = \sum_{i=1}^{N} (y_i - \beta_1 - \beta_2 x_i)^2$$

The sum of squares function using the transformed data model (8.14) is

$$\begin{aligned} S(\beta_1, \beta_2|y_i, x_i) &= \sum_{i=1}^{N} (y_i^* - \beta_1 x_{i1}^* - \beta_2 x_{i2}^*)^2 = \sum_{i=1}^{N} \left(\frac{y_i}{\sqrt{x_i}} - \beta_1 \frac{1}{\sqrt{x_i}} - \beta_2 \frac{x_{i2}}{\sqrt{x_i}} \right)^2 \\ &= \sum_{i=1}^{N} \left[\frac{1}{\sqrt{x_i}} (y_i - \beta_1 - \beta_2 x_{i2}) \right]^2 \\ &= \sum_{i=1}^{N} \frac{(y_i - \beta_1 - \beta_2 x_{i2})^2}{x_i} \end{aligned} \tag{8.16}$$

The squared errors are *weighted* by $1/x_i$. Recall that our variance assumption is $\text{var}(e_i|x_i) = \sigma^2 x_i$. When x_i is smaller we are assuming the variance of the error is smaller and the data fall closer to the regression function. These data are *more informative* about the location of $E(y_i|x_i) = \beta_1 + \beta_2 x_i$. When x_i is larger we are assuming the variance of the error is larger, and the data may fall farther from the regression function. These data are *less informative* about the location of $E(y_i|x_i) = \beta_1 + \beta_2 x_i$. Intuitively, it makes sense to "down weight" observations with less information and weigh more heavily observations with more information. That is exactly what the weighted sum of squares function (8.16) achieves. When x_i is small, the data contain more information about the regression function and the observations are weighted heavily. When x_i is large, the data contain less information and the observations are weighted lightly. In this way, we take advantage of the heteroskedasticity to improve parameter estimation. On the other hand, OLS estimation treats all observations as equally informative and equally important, as it should under homoskedasticity.

Most software have a WLS or GLS option. If your software falls into this category, you do not have to transform the variables before estimation, nor do you have to worry about omitting the constant. The computer will do both the transforming and the estimating once you decipher the software command. If you do the transforming yourself, that is, you create y_i^*, x_{i1}^*, and x_{i2}^*,

and apply OLS, be careful not to include a constant in the regression. As noted before, there is no constant because $x_{i1}^* \neq 1$.

EXAMPLE 8.3 | Applying GLS/WLS to the Food Expenditure Data

In the food expenditure example, we assume $\text{var}(e_i|INCOME_i) = \sigma_i^2 = \sigma^2 INCOME_i$. Applying the generalized (weighted) least squares procedure to our household expenditure data yields the following GLS estimates:

$$\underset{\text{(se)}}{\widehat{FOOD\_EXP}_i} = \underset{(23.79)}{78.68} + \underset{(1.39)}{10.45} INCOME_i \tag{8.17}$$

That is, we estimate the intercept term as $\hat{\beta}_1 = 78.68$ and the slope coefficient that shows the response of food expenditure to a change in income as $\hat{\beta}_2 = 10.45$. These estimates are somewhat different from the least squares estimates $b_1 = 83.42$ and $b_2 = 10.21$ that did not allow for the existence of heteroskedasticity. It is important to recognize that the interpretations for β_1 and β_2 are the same in the transformed model in (8.14) as they are in the untransformed model in (8.7). Transformation of the variables is a technique for converting a heteroskedastic error model into a homoskedastic error model, *not* as something that changes the meaning of the coefficients.

The standard errors in (8.17), $\text{se}(\hat{\beta}_1) = 23.79$ and $\text{se}(\hat{\beta}_2) = 1.39$, are both lower than their least squares counterparts that were calculated from White's robust standard errors, namely, $\text{se}(b_1) = 27.46$ and $\text{se}(b_2) = 1.81$. Since GLS is a better estimation procedure than least squares, we expect the GLS standard errors to be lower. This statement needs to be qualified in two ways. First, remember that standard errors are square roots of *estimated* variances; in a single sample, the relative magnitudes of true variances may not always be reflected by their corresponding variance estimates. Second, the reduction in variance has come at the cost of making an additional assumption, namely, that the error variances have the structure given in (8.11).

The smaller standard errors have the advantage of producing narrower, more informative confidence intervals. For example, using the GLS results, a 95% confidence interval for β_2 is given by

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

The least squares confidence interval computed using White's standard errors was [6.55, 13.87].

In order to obtain the GLS estimates, we assumed the specific pattern of heteroskedasticity, namely $\text{var}(e_i|x_i) = \sigma_i^2 = \sigma^2 h(x_i) = \sigma^2 x_i$. We must ask ourselves whether this assumption adequately represents the pattern of heteroskedasticity in the data. If so, then the transformed model (8.14) should have homoskedastic errors. An informal check is to compute the residuals from the transformed model and plot them. That is, let $\hat{e}_i^* = y_i^* - \hat{\beta}_1 x_{i1}^* - \hat{\beta}_2 x_{i2}^*$. If you have used a WLS/GLS software, then the residuals it saves are, most likely, the GLS residuals $\hat{e}_{i,\text{WLS}} = y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{i2}$. In this case, $\hat{e}_i^* = \hat{e}_{i,\text{WLS}}/\sqrt{x_i}$. In Figure 8.4 we plot the residuals from the transformed model and the OLS residuals against household income.

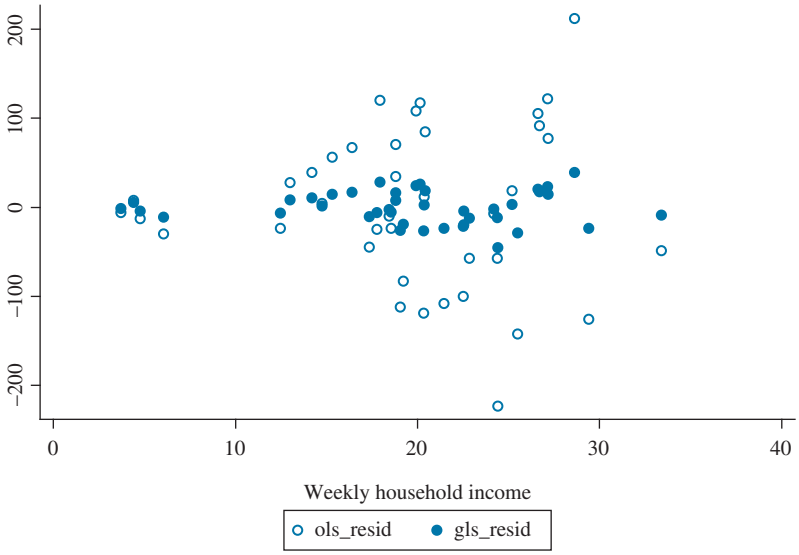

FIGURE 8.4 **OLS- and GLS-transformed residuals.**

It is evident that our transformation has substantially reduced the "spray" pattern indicating heteroskedasticity. If the transformation is a total success plotting the transformed residuals against *any* variable should reveal no pattern. If patterns remain, then you may try another skedastic function. Or, because it is visually clear that the transformation eliminated most, if not all, the heteroskedasticity, we can use a White heteroskedasticity robust standard error with the transformed model. In this way, we will have attempted to gain a more efficient estimator, but then protected ourselves against incorrect standard errors from any remaining heteroskedasticity. The GLS/WLS estimated model with robust standard errors is

$$\underset{\text{(robse)}}{\widehat{FOOD\_EXP}_i} = \underset{(12.04)}{78.68} + \underset{(1.17)}{10.45}INCOME_i$$

The 95% interval estimate of the slope is [8.07, 12.83].

8.5 Generalized Least Squares: Unknown Form of Variance

In the previous section, we assumed that heteroskedasticity could be described by the **variance function** $\text{var}(e_i|x_i) = \sigma^2 x_i$. This is convenient and simple in the food expenditure example because $x_i = INCOME_i > 0$ and intuitively reasonable. However, this is one possible choice of a skedasticity function $h(x_i)$. There are other alternatives such as $\text{var}(e_i|x_i) = \sigma^2 h(x_i) = \sigma^2 x_i^2$ and $\text{var}(e_i|x_i > 0) = \sigma^2 h(x_i) = \sigma^2 x_i^{1/2}$. Both have the property that the error variance increases as x_i increases. Why not choose one of these functions?

In a multiple regression $y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + e_i$ a heteroskedasticity pattern might be related to more than one of the explanatory variables, so that we might consider a skedastic function $h(x_{i2}, \ldots, x_{iK}) = h(\mathbf{x}_i)$. In fact, the heteroskedasticity pattern might be related to variables not even in the model! In order to deal with the more general specification that includes all these possibilities we need a model that is flexible, parsimonious, and for which $\sigma_i^2 > 0$. One specification that works well is

$$\begin{aligned}\sigma_i^2 &= \exp(\alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_S z_{iS}) \\ &= \exp(\alpha_1)\exp(\alpha_2 z_{i2} + \cdots + \alpha_S z_{iS}) \\ &= \sigma^2 h(z_{i2}, \ldots, z_{iS})\end{aligned} \tag{8.18}$$

The candidate variables $z_{i2}, \ldots, z_{iS}$ that are possibly associated with the heteroskedasticity may or may not be in $\mathbf{x}_i$. The exponential function is convenient because it ensures we will get positive values for the variances σ_i^2 for all possible values of the parameters $\alpha_1, \alpha_2, \ldots, \alpha_S$. Equation (8.18) is called the model of **multiplicative heteroskedasticity**. It includes homoskedasticity as a special case; when $\alpha_2 = \cdots = \alpha_S = 0$ the error variance is $\sigma_i^2 = \exp(\alpha_1) = \sigma^2$. It is called a multiplicative model because

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

Each candidate variable has a separate multiplicative effect. This model does introduce some new parameters, but as you have seen many times now, when there is an unknown parameter an econometrician will figure out how to estimate it. That is what we do.

This model is attractive because of the features mentioned above, it is flexible, parsimonious, and $\sigma_i^2 > 0$, and also because it has several special cases that are very useful.

Multiplicative Heteroskedasticity, Special Case 1: $\text{var}(e_i|x_i) = \sigma_i^2 = \sigma^2 x_i^{\alpha_2}$

As noted in the food expenditure example, three plausible variance functions are $\text{var}(e_i|x_i) = \sigma^2 x_i$, $\text{var}(e_i|x_i) = \sigma^2 h(x_i) = \sigma^2 x_i^2$, and $\text{var}(e_i|x_i > 0) = \sigma^2 h(x_i) = \sigma^2 x_i^{1/2}$. These are special cases of

$$\text{var}(e_i|x_i) = \sigma_i^2 = \sigma^2 x_i^{\alpha_2}$$

where α_2 is an unknown parameter. In the multiplicative model, let $S = 2$, $z_{i2} = \ln(x_i)$ and $h(z_{i2}) = \exp[\alpha_2 \ln(x_i)]$. Using the properties of logarithms and exponentials, we have

$$\begin{aligned}\sigma_i^2 &= \exp(\alpha_1 + \alpha_2 z_{i2}) \\ &= \exp(\alpha_1)\exp\left[\alpha_2 \ln(x_i)\right] = \exp(\alpha_1)\exp\left[\ln(x_i^{\alpha_2})\right] \\ &= \sigma^2 x_i^{\alpha_2}\end{aligned}$$

Multiplicative Heteroskedasticity, Special Case 2: Grouped Heteroskedasticity

Data partitions arise naturally in many economic examples. We might be estimating a wage equation with data on individuals from both urban and rural areas. It is likely that the labor market in the urban area is more diverse, leading to wage variations from one person to another that is greater than in a rural area. Or perhaps we are considering wages for individuals with different education levels, such as those with only primary school education, those with a high school education, and those with some postsecondary education. Or individuals in different industries, or countries, etc. It is possible that the same basic structure holds for each group, with perhaps intercept dummy variables, and an error variance that is different for one group versus another.

Suppose we are considering just two groups. Create an indicator variable $D_i = 1$ if an observation is in one group and $D_i = 0$ for observations in the other group. Then the variance function is

$$\text{var}(e_i|\mathbf{x}_i) = \exp(\alpha_1 + \alpha_2 D_i) = \begin{cases} \exp(\alpha_1) = \sigma^2 & D_i = 0 \\ \exp(\alpha_1 + \alpha_2) = \sigma^2 \exp(\alpha_2) & D_i = 1 \end{cases}$$

Using the multiplicative form $\sigma_i^2 = \exp(\alpha_1 + \alpha_2 D_i) = \exp(\alpha_1)\exp(\alpha_2 D_i) = \sigma^2 h(D_i)$, the skedastic function is $h(D_i) = \exp(\alpha_2 D_i)$. Note that if $\alpha_2 = 0$ the error variance is the same for the two groups, meaning that the assumption of homoskedasticity holds.

The same strategy works if there are more than two groups. Suppose there are $g = 1, 2, \ldots, G$ groups or data partitions. Create indicator variables for each group. Let $D_{ig} = 1$ if an observation is from group g, and otherwise $D_{ig} = 0$. If e_{ig} is the random error for the ith observation in group g, then a useful variance function is

$$\text{var}(e_{ig}|\mathbf{x}_{ig}) = \exp(\alpha_1 + \alpha_2 D_{i2} + \cdots + \alpha_G D_{iG}) = \begin{cases} \exp(\alpha_1) = \sigma^2 = \sigma_1^2 & g = 1; \text{ only } D_{i1} = 1 \\ \exp(\alpha_1 + \alpha_2) = \sigma_2^2 & g = 2; \text{ only } D_{i2} = 1 \\ \vdots & \\ \exp(\alpha_1 + \alpha_G) = \sigma_G^2 & g = G; \text{ only } D_{iG} = 1 \end{cases}$$

In this specification, we have chosen group 1 as the reference group and its indicator variable is omitted. This is similar to the indicator variable approach in Chapter 7. The variance of the reference group error can be denoted σ^2 or σ_1^2, to indicate that it is for group 1. For groups 2, … , G the skedastic function is $h(D_g) = \exp(\alpha_g D_g)$. Alternatively, let the variance function be $\text{var}(e_{ig}|\mathbf{x}_{ig}) = \exp(\alpha_1 D_{i1} + \alpha_2 D_{i2} + \cdots + \alpha_G D_{iG})$. Work out the variance for each group with this alteration. The end results using these two specifications are identical.

8.5.1 Estimating the Multiplicative Model

How do we proceed with estimation with an assumption like (8.18)? Our ultimate objective is to estimate the regression parameters $\beta_1, \beta_2, \ldots, \beta_K$. With the model of multiplicative heteroskedasticity, we use several estimation steps.

FEASIBLE GLS PROCEDURE

1. Estimate the original model $y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + e_i$ by OLS, saving the OLS residuals $\hat{e}_i$.
2. Use the least squares residuals and the variables $z_{i2}, \ldots, z_{iS}$ to estimate $\alpha_1, \alpha_2, \ldots, \alpha_S$.
3. Calculate the estimated skedastic function $\hat{h}(z_{i2}, \ldots, z_{iS})$.
4. Divide each observation by $\sqrt{\hat{h}(z_{i2}, \ldots, z_{iS})}$ and apply OLS to the transformed data, or use WLS regression with weighting factor $1/\hat{h}(z_{i2}, \ldots, z_{iS})$.

The resulting estimates, $\hat{\hat{\beta}}_1, \hat{\hat{\beta}}_2, \ldots, \hat{\hat{\beta}}_K$, are called **feasible generalized least squares (FGLS)** estimates or **estimated generalized least squares (EGLS)** estimates. If heteroskedasticity is present, the FGLS estimator is consistent and more efficient than OLS in large samples. We have placed a second "hat" on these estimates to differentiate them from the earlier GLS estimates and to remind us that these estimates depend on a first-stage estimation.

Step 2 in the procedure is accomplished through a very clever manipulation of the model of multiplicative heteroskedasticity. Taking logarithms of both sides of (8.18), we obtain

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

This looks like a regression model except for the fact that the left-hand side is unknown. Add the log of the squared least squares residuals to each side:

$$\ln(\sigma_i^2) + \ln(\hat{e}_i^2) = \alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_S z_{iS} + \ln(\hat{e}_i^2) \tag{8.19}$$

Rearrange and simplify equation (8.19):

$$\begin{aligned}
\ln(\hat{e}_i^2) &= \alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_S z_{iS} + \ln(\hat{e}_i^2) - \ln(\sigma_i^2) \\
&= \alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_S z_{iS} + \ln(\hat{e}_i^2/\sigma_i^2) \\
&= \alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_S z_{iS} + \ln\left[(\hat{e}_i/\sigma_i)^2\right] \\
&= \alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_S z_{iS} + v_i
\end{aligned}$$

We have taken the model of multiplicative heteroskedasticity and through some simple manipulations arrived to

$$\ln(\hat{e}_i^2) = \alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_S z_{iS} + v_i \tag{8.20}$$

Using this model we can estimate $\alpha_1, \alpha_2, \ldots, \alpha_S$ in (8.19) using OLS and continue with the steps of the procedure. Whether or not this procedure is a legitimate one depends on the properties of the new error term v_i that we introduced in (8.20). Does it have a zero mean? Is it homoskedastic? In small samples the answer to these questions is no. However, in large samples the answer is happier. It can be shown (see Appendix 8C.1) that $E(v_i|\mathbf{z}_i) \cong -1.2704$ and $\text{var}(v_i|\mathbf{z}_i) \cong 4.9348$, where $\mathbf{z}_i = (1, z_{i2}, \ldots, z_{iS})$, and if $e_i \sim N(0, \sigma_i^2)$. Because the regression error does not have conditional

mean zero, the estimated value of α_1 will be off by -1.2704. But $\hat{\alpha}_2, \ldots, \hat{\alpha}_S$ are consistent estimators, which for estimating the skedastic function $\hat{h}(z_{i2}, \ldots, z_{iS})$ is all that matters.

EXAMPLE 8.4 | Multiplicative Heteroskedasticity in the Food Expenditure Model

In the food expenditure example, with z_{i2} defined as $z_{i2} = \ln(INCOME_i)$, the least squares estimate of (8.19) is

$$\widehat{\ln(e_i^2)} = 0.9378 + 2.329 \ln(INCOME_i)$$

Notice that the estimate $\hat{\alpha}_2 = 2.329$ is more than twice the value of $\alpha_2 = 1$, which was an implicit assumption of the variance specification used in Example 8.3. This suggests the earlier transformation was not sufficiently aggressive. Following the steps to obtain FGLS estimates we transform the model by dividing both sides by $\sqrt{\hat{h}(z_{i2})}$, where $\hat{h}(z_{i2}) = \exp[\hat{\alpha}_2 \ln(INCOME_i)]$, then apply OLS to the transformed data, or use WLS with weight $1/\hat{h}(z_{i2})$. The resulting FGLS estimates for the food expenditure example are

$$\underset{\text{(se)}}{\widehat{FOOD\_EXP_i}} = \underset{(9.71)}{76.05} + \underset{(0.97)}{10.63} INCOME_I \qquad (8.21)$$

Compared to the GLS results for the variance specification $\sigma_i^2 = \sigma^2 INCOME_i$, the estimates for β_1 and β_2 have not changed a great deal, but there has been a considerable drop in the standard errors that, under the previous specification, were $\text{se}(\hat{\beta}_1) = 23.79$ and $\text{se}(\hat{\beta}_2) = 1.39$.

We must ask ourselves whether our FGLS transformation has been adequate; does the transformed model satisfy the homoskedasticity assumption? In Example 8.3, we computed the residuals from the transformed model $\hat{e}_i^* = y_i^* - \hat{\beta}_1 x_{i1}^* - \hat{\beta}_2 x_{i2}^*$. Similarly, let $\hat{\hat{e}}_i^* = y_i^{**} - \hat{\beta}_1 x_{i1}^{**} - \hat{\beta}_2 x_{i2}^{**}$, where $y_i^{**} = y_i / \sqrt{\hat{h}(z_{i2})}$, $x_{i1}^{**} = 1/\sqrt{\hat{h}(z_{i2})}$, and $x_{i2}^{**} = x_{i2}/\sqrt{\hat{h}(z_{i2})}$. In Figure 8.5, we plot $\hat{e}_i^*$ (empty circles) from the GLS-transformed model, and $\hat{\hat{e}}_i^*$ (solid dots), from the FGLS-transformed model, versus income. Note that the vertical axis scales in Figures 8.4 and 8.5 are different; so take that into account when comparing them. By "zooming in" on $\hat{e}_i^*$ (empty circles) from the GLS-transformed model, we see a fan-shaped pattern persisting, meaning that the GLS transformation did not completely eliminate heteroskedasticity. In Figure 8.4, we saw a great reduction in the "spray" pattern and in Figure 8.5 the FGLS-transformed model has yet smaller residuals and shows a further reduction in the "spray" pattern. Based on visual evidence, the FGLS model has done a better job at eliminating heteroskedasticity than the GLS model.

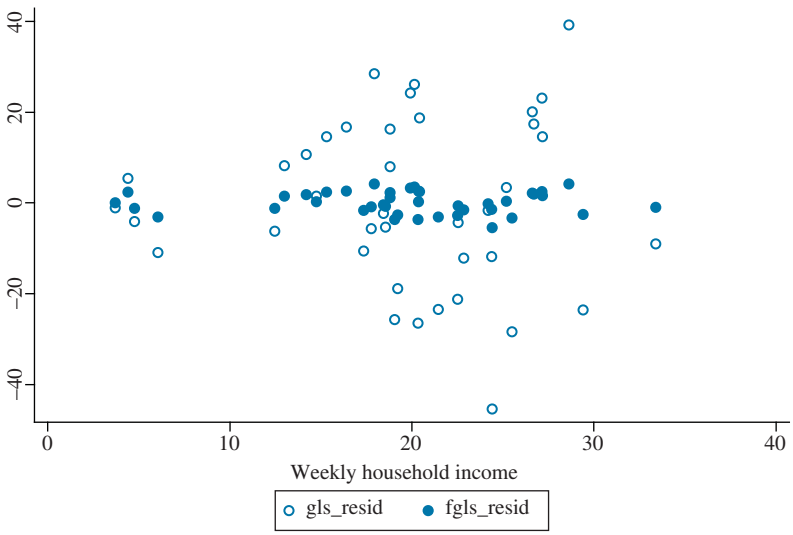

FIGURE 8.5 **GLS- and FGLS-transformed residuals.**

EXAMPLE 8.5 | A Heteroskedastic Partition

To illustrate the idea of a heteroskedastic partition we consider a simple wage equation in which a person's wage rate (*WAGE*) depends on their education (*EDUC*) and experience (*EXPER*). We also include an indicator variable for whether they live in a metropolitan, more urbanized, area or not. For convenience, think of the nonmetropolitan areas as "rural." That is

$$METRO = \begin{cases} 1 & \text{if person lives in a metropolitan area} \\ 0 & \text{if person lives in a rural area} \end{cases}$$

The wage equation is

$$WAGE_i = \beta_1 + \beta_2 EDUC_i + \beta_3 EXPER_i + \beta_4 METRO_i + e_i$$

The issue we address here is the possibility that the variance of the error term is different in metropolitan areas than in rural areas. That is, we suspect that

$$\text{var}(e_i|\mathbf{x}_i) = \begin{cases} \sigma_M^2 & \text{if } METRO = 1 \\ \sigma_R^2 & \text{if } METRO = 0 \end{cases}$$

For illustration, we use the data file *cps5_small* and restrict ourselves to observations from the Midwest region, $MIDWEST = 1$. First consider the summary statistics in Table 8.1 for metropolitan workers, $METRO = 1$, and rural workers, $METRO = 0$.

Observe that the average wage and the standard deviation of wage are higher in metropolitan areas than in rural areas. This is suggestive but not proof of heteroskedasticity. The standard deviation is an "unconditional" measure that does not depend on the regression model. Heteroskedasticity is a concern about the variation in the regression random errors holding other factors constant, in this case education and experience.

The OLS estimates with heteroskedasticity robust standard errors are

$$\underset{\text{(robse)}}{\widehat{WAGE}_i} = \underset{(4.023)}{-18.450} + \underset{(0.261)}{2.339}EDUC_i + \underset{(0.0478)}{0.189}EXPER_i + \underset{(1.159)}{4.991}METRO_i$$

We save the OLS residuals, $\hat{e}_i$, and estimate equation (8.20) using $z_{i2} = METRO_i$, $\ln\left(\hat{e}_i^2\right) = \alpha_1 + \alpha_2 METRO + v_i$, obtaining

$$\widehat{\ln\left(\hat{e}_i^2\right)} = 2.895 + 0.700METRO$$

The estimated skedastic function is

$$\hat{h}(z_{i2}) = \exp(\hat{\alpha}_2 METRO_i) = \exp(0.700METRO) = \begin{cases} 2.0147 & METRO = 1 \\ 1 & METRO = 0 \end{cases}$$

We estimate the conditional variance of the random error to be about twice as large for the metropolitan area as in the rural area. In the WLS regression, the observations in the metropolitan area will receive half the weight of the observations in the rural area. The feasible GLS estimates are

$$\underset{\text{(se)}}{\widehat{\widehat{WAGE}}_i} = \underset{(3.788)}{-16.968} + \underset{(0.239)}{2.258}EDUC_i + \underset{(0.0447)}{0.175}EXPER_i + \underset{(1.214)}{4.995}METRO_i$$

The FGLS coefficient estimates and standard errors for *EDUC* and *EXPER* are slightly smaller than in the OLS estimation.

TABLE 8.1 Summary Statistics, by *METRO*

	Variable	Obs	Mean	Std. Dev.
$METRO = 1$	*WAGE*	213	24.25	14.00
	EDUC	213	14.25	2.77
	EXPER	213	23.15	13.17
$METRO = 0$	*WAGE*	84	18.86	8.52
	EDUC	84	13.99	2.26
	EXPER	84	24.30	14.32

8.6 Detecting Heteroskedasticity

In our discussion of the food expenditure equation, we used the nature of the economic problem and data to argue why heteroskedasticity of a particular form might be present. However, in many applications, there is uncertainty about the presence, or absence, of heteroskedasticity. It is natural to ask: How do I know if heteroskedasticity is likely to be a problem for my model and my set of data? Is there a way of detecting heteroskedasticity so that I know whether to use GLS techniques? We consider three ways of investigating these questions. The first is the informal use of **residual plots**. The other two are more formal classes of statistical tests.

8.6.1 Residual Plots

One way of investigating the existence of heteroskedasticity is to estimate your model using least squares and to plot the least squares residuals. If the errors are homoskedastic, there should be no patterns of any sort in the residuals, as shown in Figure 4.7(a). If the errors are heteroskedastic, they may tend to exhibit greater, or less, variation in some systematic way, as in Figures 4.7(b)–(d). For example, for the household food expenditure data, we suspect that the variance increases as incomes increases. We illustrated the use of diagnostic residual plots in Examples 8.1–8.3. We discovered that the absolute values of the residuals do indeed tend to increase as income increases. This method of investigating heteroskedasticity can be followed for any simple regression.

When we have more than one explanatory variable, the estimated least squares function is not so easily depicted on a diagram. However, what we can do is plot the least squares residuals against each explanatory variable, or against the fitted values $\hat{y}_i$, to see if those residuals vary in a systematic way relative to the specified variable.

8.6.2 The Goldfeld–Quandt Test

The second test for heteroskedasticity that we consider is designed for the case where we have two subsamples with possibly different variances. The sub-samples might be based on an indicator variable. In Example 8.5, we considered metropolitan and rural sub-samples for estimating a wage equation. Alternatively, we might sort the data according to the magnitude of one continuous variable and then divide the data into subsamples, omitting a few central observations to create separation if possible. In either case, the **Goldfeld–Quandt** test uses the estimated error variances from separate sub-sample regressions as a basis for the test. The background for this test appears in Appendix C.7.3. The only difference is in the degrees of freedom. Let the first sub-sample contain N_1 observations and let the regression model in this partition have K_1 parameters, including the intercept. Let the true variance of the error in this sample be σ_1^2 with estimator $\hat{\sigma}_1^2 = SSE_1/(N_1 - K_1)$. Let the second sub-sample contain N_2 observations and let the regression model in this partition have K_2 parameters, including the intercept. Let the true variance of the error in this sample be σ_2^2 with estimator $\hat{\sigma}_2^2 = SSE_2/(N_2 - K_2)$. The test statistic is

$$GQ = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} \sim F_{(N_1-K_1,\, N_2-K_2)} \tag{8.22}$$

If the null hypothesis $H_0 : \sigma_1^2/\sigma_2^2 = 1$ is true, then the test statistic $GQ = \hat{\sigma}_1^2/\hat{\sigma}_2^2$ has an F-distribution with $(N_1 - K_1)$ numerator and $(N_2 - K_2)$ denominator degrees of freedom. If the alternative hypothesis is $H_1 : \sigma_1^2/\sigma_2^2 \neq 1$, then we carry out a two-tail test. If we choose level of significance $\alpha = 0.05$, then we reject the null hypothesis if $GQ \geq F_{(0.975,\, N_1-K_1,\, N_2-K_2)}$ or if $GQ \leq F_{(0.025,\, N_1-K_1,\, N_2-K_2)}$, where $F_{(\alpha,\, N_1-K_1,\, N_2-K_2)}$ denotes the 100α-percentile of the F-distribution with the specified degrees of freedom. If the alternative is one-sided, $H_1 : \sigma_1^2/\sigma_2^2 > 1$, then we reject the null hypothesis if $GQ \geq F_{(0.95,\, N_1-K_1,\, N_2-K_2)}$.

EXAMPLE 8.6 | The Goldfeld–Quandt Test with Partitioned Data

We illustrate the Goldfeld–Quandt test by continuing Example 8.5. The data partitions are based on the indicator variable

$$METRO = \begin{cases} 1 & \text{if person lives in a metropolitan area} \\ 0 & \text{if person lives in a rural area} \end{cases}$$

The issue we address here is the possibility that the variance of the error term is different in metropolitan areas than in rural areas. To test the homoskedasticity assumption, estimate the wage equation in each data partition:

$$WAGE_{Mi} = \beta_{M1} + \beta_{M2} EDUC_{Mi} + \beta_{M3} EXPER_{Mi} + e_{Mi}$$
$$WAGE_{Ri} = \beta_{R1} + \beta_{R2} EDUC_{Ri} + \beta_{R3} EXPER_{Ri} + e_{Ri}$$

Let $\text{var}(e_{Mi}|\mathbf{x}_{Mi}) = \sigma_M^2$ and $\text{var}(e_{Ri}|\mathbf{x}_{Ri}) = \sigma_R^2$. Our null hypothesis is $H_0 : \sigma_M^2/\sigma_R^2 = 1$. Let the alternative hypothesis

be $H_1: \sigma_M^2/\sigma_R^2 \neq 1$, so that we use a two-tail test. The metropolitan subsample has 213 observations and the rural subsample has 84. In this case, as in most, the number of parameters in each data-partition regression is the same, $K = K_1 = K_2 = 3$. The test critical values are $F_{(0.975,\,210,\,81)} = 1.4615$ and $F_{(0.025,\,210,\,81)} = 0.7049$. Using $\widehat{\text{var}}(e_{Mi}|\mathbf{x}_{Mi}) = \hat{\sigma}_M^2 = 147.62$ and $\widehat{\text{var}}(e_{Ri}|\mathbf{x}_{Ri}) = \hat{\sigma}_R^2 = 56.71$, the calculated value of the Goldfeld–Quandt test statistic is $GQ = 2.6033 > F_{(0.975,\,210,\,81)} = 1.4615$, so we reject the null hypothesis that the error variances in the two subsamples are equal.

EXAMPLE 8.7 | The Goldfeld–Quandt Test in the Food Expenditure Model

Although the Goldfeld–Quandt test is very convenient for instances where the sample divides naturally into two subsamples, it can also be used where, under H_1, the variance is a function of a single explanatory variable. In the food expenditure model, we suspect that the error variance increases as income increases. We order the observations according to the magnitude of income so that, if heteroskedasticity exists, the first half of the sample will correspond to observations with lower variances and the last half of the sample will correspond to observations with higher variances. Then, we split the sample into two approximately equal halves, carry out two separate least squares regressions that yield variance estimates, say $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$, and proceed with the test as described previously.

Following these steps for the food expenditure example, with the observations ordered according to income, we split the sample into two equal subsamples of 20 observations each. Because the sample is small, we do not omit any middle observations. Estimating the model on each subsample yields $\hat{\sigma}_1^2 = 3574.8$ and $\hat{\sigma}_2^2 = 12,921.9$, from which we obtain

$$F = \frac{\hat{\sigma}_2^2}{\hat{\sigma}_1^2} = \frac{12,921.9}{3574.8} = 3.61$$

Believing that the variances could increase, but not decrease with income, we use a one-tailed test with 5% level of significance critical value $F_{(0.95,\,18,\,18)} = 2.22$. Since $3.61 > 2.22$, a null hypothesis of homoskedasticity is rejected in favor of the alternative that the variance increases with income.

8.6.3 A General Test for Conditional Heteroskedasticity

In this section we consider a test for **conditional heteroskedasticity** that is related to some "explanatory" variables. Our equation of interest is the regression model

$$y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + e_i \tag{8.23}$$

Under assumptions MR1–MR5 the OLS estimator is the best linear unbiased estimator of the parameters $\beta_1, \beta_2, \ldots, \beta_K$. When conditional heteroskedasticity is a possibility, we hypothesize that the variance of the random error, e_i, depends on a set of explanatory variables $z_{i2}, z_{i3}, \ldots, z_{iS}$ that may include some or all of the explanatory variables $x_{i2}, \ldots, x_{iK}$. That is, assume a general expression for the conditional variance

$$\text{var}(e_i|\mathbf{z}_i) = \sigma_i^2 = E(e_i^2|\mathbf{z}_i) = h(\alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_S z_{iS}) \tag{8.24}$$

where $h(\cdot)$ is some smooth function and $\alpha_2, \alpha_3, \ldots, \alpha_S$ are **nuisance parameters**, meaning that we are not really interested in their values but must recognize that they are there. The beauty of the test we are about to present is that we do not have to actually know, or even guess, the function $h(\cdot)$. We will test for *any* relationship between the variance of the error term and *any* function of the selected variables. The function $h(\cdot)$ is similar to the skedastic function in equation (8.4), but here we have not factored out a constant σ^2, and unlike the feasible GLS estimation we do not have to choose an exponential form for $h(\cdot)$.

Notice what happens to the function $h(\cdot)$ when $\alpha_2 = \alpha_3 = \cdots = \alpha_S = 0$. It collapses to

$$h(\alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_S z_{iS}) = h(\alpha_1) \tag{8.25}$$

The term $h(\alpha_1)$, which we can define to be σ^2, is a constant, and $\text{var}(e_i|\mathbf{z}_i) = h(\alpha_1) = \sigma^2$. In other words, when $\alpha_2 = \alpha_3 = \cdots = \alpha_S = 0$ the random errors are homoskedastic. On the other hand, if *any* of the parameters $\alpha_2, \alpha_3, \ldots, \alpha_S$ are not zero, then heteroskedasticity is present. Consequently, the null and alternative hypotheses for a test for heteroskedasticity based on the variance function are

$$\begin{aligned} &\text{homoskedasticity} \leftrightarrow H_0 : \alpha_2 = \alpha_3 = \cdots = \alpha_S = 0 \\ &\text{heteroskedasticity} \leftrightarrow H_1 : \text{ not all the } \alpha_s \text{ in } H_0 \text{ are zero} \end{aligned} \tag{8.26}$$

The null and alternative hypotheses are the first components of a test. The next component is a test statistic. To obtain a test statistic, consider a *linear* conditional variance function

$$\sigma_i^2 = E(e_i^2|\mathbf{z}_i) = \alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_S z_{iS} \tag{8.27}$$

Despite using a linear conditional variance function the test is for the general heteroskedasticity pattern in (8.24). Let $v_i = e_i^2 - E(e_i^2|\mathbf{z}_i)$ be the difference between a squared error and its conditional mean. Then, from (8.27), we can write

$$e_i^2 = E(e_i^2|\mathbf{z}_i) + v_i = \alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_S z_{iS} + v_i \tag{8.28}$$

This looks very much like a linear regression model. The one problem is that the "dependent variable" e_i^2 is not observable. We overcome this problem by replacing e_i^2 with the squared OLS residuals $\hat{e}_i^2$. In large samples, this is valid because, as we show in Appendix 8B, the difference $e_i - \hat{e}_i$ goes to zero as $N \to \infty$. An operational version of (8.28) is

$$\hat{e}_i^2 = \alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_S z_{iS} + v_i \tag{8.29}$$

Strictly speaking, replacing e_i^2 by $\hat{e}_i^2$ also changes the definition of v_i, but we will retain the same notation to avoid unnecessary complication.

The test for heteroskedasticity is based on OLS estimation of (8.29). The question we ask is, do the variables $z_{i2}, z_{i3}, \ldots, z_{iS}$ help explain $\hat{e}_i^2$? Under homoskedasticity the variables $z_{i2}, z_{i3}, \ldots, z_{iS}$ should have no relation to $\hat{e}_i^2$. One alternative is to use an F-test for the null hypothesis. An asymptotically equivalent and convenient test is based on the R^2, goodness-of-fit statistic, from (8.29). If the null hypothesis is true, $\alpha_2 = \alpha_3 = \cdots = \alpha_S = 0$, then the R^2 should be small and close to zero. If R^2 is large, it is evidence against the assumption of homoskedasticity. How large does R^2 have to be for us to reject homoskedasticity? An answer requires a test statistic and a rejection region. It can be shown that if the random errors are homoskedastic, then the sample size multiplied by R^2, $N \times R^2$ or simply NR^2, has a chi-square (χ^2) distribution with $S - 1$ degrees of freedom in large samples. That is,

$$NR^2 \overset{a}{\sim} \chi^2_{(S-1)} \text{ if the null hypothesis of homoskedasticity is true} \tag{8.30}$$

Your exposure to the χ^2 distribution has been relatively limited. It is discussed in Appendix B.5.2. It was used for testing for normality in Section 4.3.4, and its relationship with the F-test was explored in Section 6.1.5. It is a distribution that is used for testing many different kinds of hypotheses. Like an F random variable, a χ^2 random variable only takes positive values. Critical values of the distribution appear in Statistical Table 3. Locate the test degrees of freedom in the left-hand column, and find the critical value from the columns, each of which corresponds to a percentile of the distribution. Because a large R^2 value is evidence against the null hypothesis of homoskedasticity (it suggests the z variables explain some changes in the variance), the rejection region for the statistic in (8.30) is in the right tail of the distribution. For an α-significance level test, we reject H_0 and conclude that heteroskedasticity exists when $NR^2 \geq \chi^2_{(1-\alpha,\, S-1)}$.

For example, if $\alpha = 0.01$ and $S = 2$, reject the hypothesis of homoskedasticity if $NR^2 \geq \chi^2_{(0.99,1)} = 6.635$. Your econometric software will have functions to calculate critical values, and p-values, for χ^2-tests.

There are several important features of this test:

1. It is a large sample test. The result in (8.30) holds approximately in large samples.
2. You will often see the test referred to as a **Lagrange multiplier test** (LM test) or a **Breusch–Pagan test** for heteroskedasticity. Breusch and Pagan used the LM principle (see Appendix C.8.4) to derive an earlier version of the test, which was later modified by other researchers to the form in (8.30). The test values for these and other slightly different versions of the test, one of which is the F-test, are automatically calculated by a number of software packages. The one provided by your software may or may not be exactly the same as the NR^2 version in (8.30). The relationships between the different versions of the test are described in Appendix 8B. As you proceed through the book and study more econometrics, you will find that many LM tests can be written in the form NR^2, where the R^2 comes from a convenient auxiliary regression related to the hypothesis being tested.
3. We motivated the test in terms of an alternative hypothesis with the very general conditional variance function $\sigma_i^2 = h(\alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_S z_{iS})$, yet we proceeded to carry out the test using the linear function $\hat{e}_i^2 = \alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_S z_{iS} + v_i$. One of the amazing features of the Breusch–Pagan/LM test is that 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 (8.24).
4. The Breusch–Pagan test is for conditional heteroskedasticity. **Unconditional heteroskedasticity** exists when the error term variance is completely random, changing from observation to observation but unrelated to any particular variable. The least squares estimator properties are unaffected by unconditional heteroskedasticity. We illustrate this point in Appendix 8D.

8.6.4 The White Test

One problem with the variance function test described so far is that it presupposes that we have knowledge of what variables will appear in the variance function if the alternative hypothesis of heteroskedasticity is true. In other words, it assumes we are able to specify $z_2, z_3, \ldots, z_S$. In reality, we may wish to test for heteroskedasticity without precise knowledge of the relevant variables. With this point in mind, White suggested defining the z's as equal to the x's, the squares of the x's, and their cross-products. Frequently, the variables that affect the variance are the same as those in the mean function. Also, by using a quadratic function we can approximate a number of other possible conditional variance functions. Suppose the regression model is

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i$$

The **White test** uses

$$z_2 = x_2 \quad z_3 = x_3 \quad z_4 = x_2^2 \quad z_5 = x_3^2 \text{ and } z_6 = x_2 x_3$$

If the regression model contains quadratic terms $(x_3 = x_2^2$ for example$)$, then some of the z's are redundant and are deleted. Also if x_3 is an indicator variable, taking the values 0 and 1, then $x_3^2 = x_3$ which is also redundant.

The White test is performed using the NR^2 test defined in (8.29), or an F-test (see Appendix 8B for details). One difficulty with the White test is that it can detect problems other than heteroskedasticity. Thus, while it is a useful diagnostic, be careful about interpreting the result of a significant White test. It may be that you have an incorrect functional form, or an omitted variable. In this sense, it is something like RESET, a specification error test discussed in Chapter 6.

8.6.5 Model Specification and Heteroskedasticity

As hinted at the end of the previous section, heteroskedasticity can be present because of a model specification error. If data partitions are not recognized, or important variables omitted, or an incorrect functional form selected, then heteroskedasticity can appear to be present. Hence, one piece of advice is to "Trust no one." Don't necessarily believe that a significant heteroskedasticity test means that heteroskedasticity is the problem and that using robust standard errors will be an adequate fix. Critically examine the model from the point of view of economic reasoning and look for any specification problems.

One very common specification issue with economic data is the choice of functional form. In Section 4.3.2, we discussed a variety of model specifications that are useful when considering nonlinear, or curvilinear, relationships (see Figure 4.5). Many economic applications use "log-log" or "log-linear" models. Using a logarithmic transformation of the dependent variable has another feature, **variance stabilization**, that is useful in the context of heteroskedastic data.[3] Economic variables like wages, incomes, house prices, and expenditures are right-skewed, with a long tail to the right. The **log-normal** probability distribution is useful when modeling such variables. This idea was introduced first in the Probability Primer in Figure P.2, and we discuss the log-normal distribution in Appendix B.3.9. If the random variable y has a log-normal probability density function, then $\ln(y)$ has a normal distribution, which is symmetrical and bell-shaped, and not skewed. That is, $\ln(y) \sim N(\mu, \sigma^2)$. The feature of the log-normal random variable that we are now interested in is that its variance increases when its mean and median increase. This is illustrated in Appendix B.3.9, Figure B.10, and the surrounding discussion. In Figure 8.6 we modify Figure 4.5(e) for the log-linear model to show $E(y|x)$, the solid line, and include $E(y|x) \pm 2\sqrt{\text{var}(y|x)}$, the dashed lines. By choosing a log-linear or log-log model we are implicitly assuming a curvilinear and heteroskedastic relationship between the variables y and x. However, there is a linear and homoskedastic relation between $\ln(y)$ and x.

Let's look at an example.

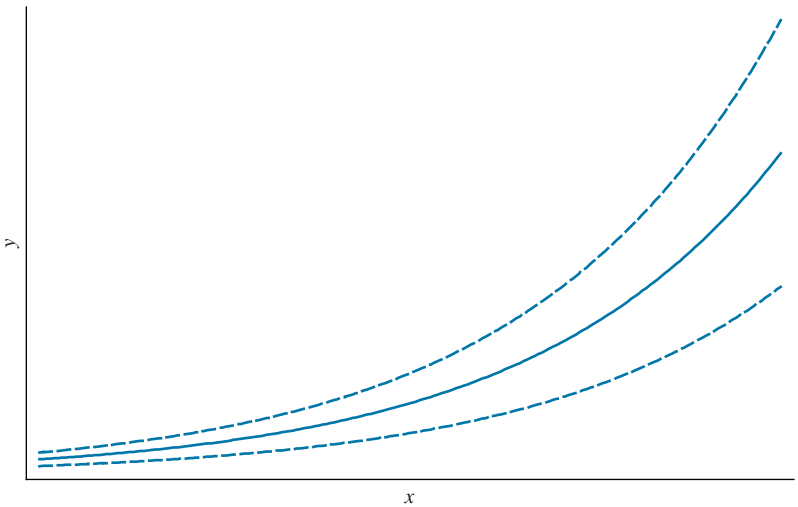

FIGURE 8.6 **A log-linear relationship.**

[3]The "Box-Cox Model" nests the linear and log-linear models in a more general nonlinear regression framework. See William Greene (2018) *Econometric Analysis, Eighth Edition*, 214–216.

EXAMPLE 8.8 | Variance Stabilizing Log-transformation

Consider the data file *cex5_small*. Figure 8.7(a) shows a histogram of household expenditures on entertainment per person, *ENTERT*, for those households who have positive spending, and Figure 8.7(b) is the histogram for ln(*ENTERT*).

Note the extremely skewed distribution of entertainment expenditures in Figure 8.7(a). Figure 8.7(b) shows the effect of the log-transformation. The distribution of ln(*ENTERT*) exhibits little skewness. Figure 8.8(a) shows the entertainment expenses plotted versus income and the least squares fitted line.

The variation in *ENTERT* about the fitted line increases as *INCOME* increases. Estimating the model $ENTERT = \beta_1 + \beta_2 INCOME + \beta_3 COLLEGE + \beta_4 ADVANCED + e$, we obtain the least squares residuals and then estimate by OLS the model $\hat{e}_i^2 = \alpha_1 + \alpha_2 INCOME_i + v_i$. From this regression, $NR^2 = 31.34$. The critical value for a 1% level of significance, heteroskedasticity test is 6.635, thus we conclude that heteroskedasticity is present. Figure 8.8(b) shows the log of entertainment expenses, ln(*ENTERT*), plotted versus income and the least squares fitted line. There is little if any visual evidence of heteroskedasticity and the value of the heteroskedasticity test statistic is $NR^2 = 0.36$, so we do not reject the null hypothesis of homoskedasticity. The log-transformation has "cured" the heteroskedasticity problem.

Among the 1200 households in the sample, 100 did not report any spending on entertainment. The log-transformation can only be used for positive values. We dropped the 100 with no spending, but that is not necessarily the best approach. In Section 16.7 we will discuss this type of data, which is called a **censored** sample.

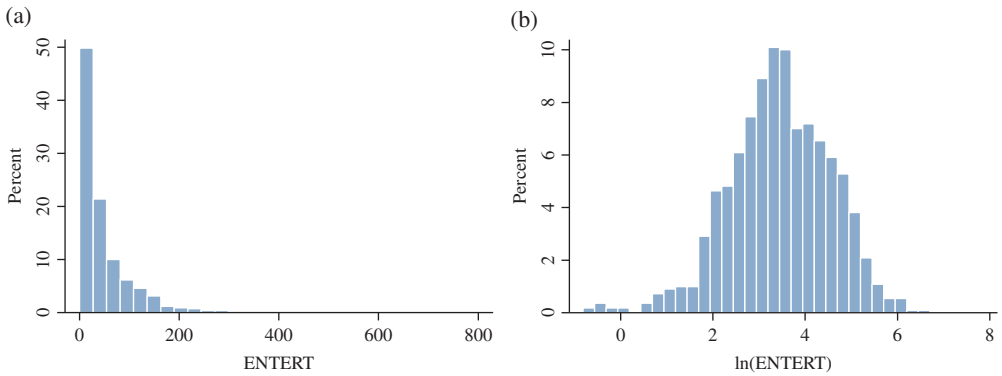

FIGURE 8.7 **Histograms of entertainment expenditures.**

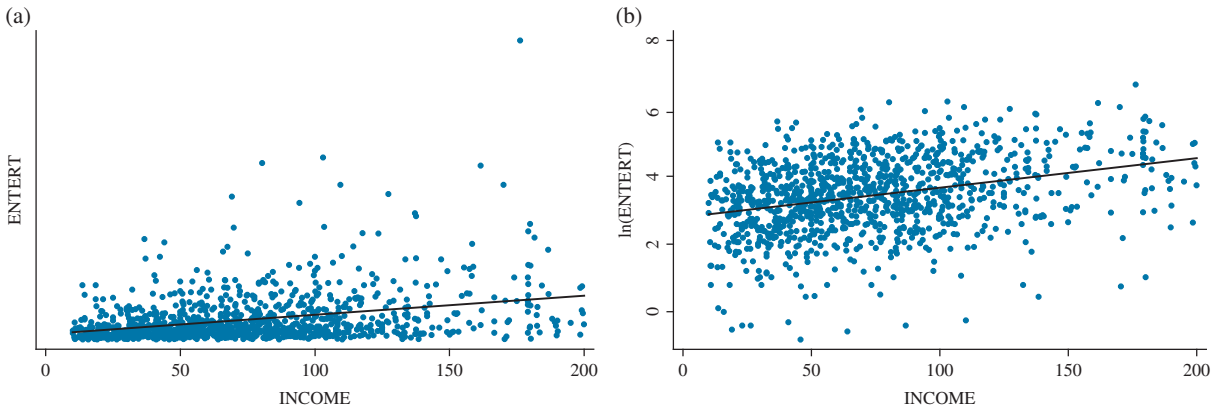

FIGURE 8.8 **Linear and log-linear models for entertainment expenditures.**

8.7 Heteroskedasticity in the Linear Probability Model

In Section 7.4 we introduced the **linear probability model** for explaining choice between two alternatives. We can represent this choice by an indicator variable y that takes the value one with probability p if the first alternative is chosen, and the value zero with probability $1 - p$ if the second alternative is chosen. An indicator variable with these properties is a Bernoulli random variable with mean $E(y) = p$ and variance $\text{var}(y) = p(1 - p)$. Interest centers on measuring the effect of explanatory variables $x_2, x_3, \ldots, x_k$ on the probability p. In the linear probability model the relationship between p and the explanatory variables is specified as the linear function

$$E(y_i|\mathbf{x}_i) = p = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK}$$

Defining the error e_i as the difference $y_i - E(y_i|\mathbf{x}_i)$ for the ith observation, we have the model

$$y_i = E(y_i|\mathbf{x}_i) + e_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + e_i \tag{8.31}$$

This model can be estimated with least squares—an example was given in Section 7.4—but it suffers from heteroskedasticity because

$$\begin{aligned} \text{var}(y_i|\mathbf{x}_i) &= \text{var}(e_i|\mathbf{x}_i) = p_i(1 - p_i) \\ &= (\beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK})(1 - \beta_1 - \beta_2 x_{i2} - \cdots - \beta_K x_{iK}) \end{aligned} \tag{8.32}$$

The error variance depends on the values of the explanatory variables. We can rectify this problem by applying the techniques described earlier in this chapter. Instead of using least squares standard errors, we can use heteroskedasticity-robust standard errors. Or, alternatively, we can apply a GLS procedure.

The first step toward obtaining GLS estimates is to estimate the variance in (8.32). An estimate of p_i can be obtained from the least squares predictions

$$\hat{p}_i = b_1 + b_2 x_{i2} + \cdots + b_K x_{iK} \tag{8.33}$$

giving an estimated variance of

$$\widehat{\text{var}}(e_i|\mathbf{x}) = \hat{p}_i(1 - \hat{p}_i) \tag{8.34}$$

A word of caution is required at this point. It is possible that some of the $\hat{p}_i$ obtained from (8.33) will not lie within the interval $0 < \hat{p}_i < 1$. If that happens, the corresponding variance estimate in (8.34) will be negative or zero, a nonsensical outcome. Thus, before proceeding to calculate the estimated variances from (8.34), it is necessary to check the estimated probabilities from (8.33) to ensure that they lie between zero and one. For those observations that violate this requirement, one possible solution is to set $\hat{p}_i$'s greater than 0.99 equal to 0.99, and $\hat{p}_i$'s less than 0.01 equal to 0.01. Another possible solution is to omit the offending observations. Neither of these solutions is totally satisfactory. Truncating at 0.99 or 0.01 is arbitrary, and the results could be sensitive to the truncation point. Omitting observations means that we are throwing away information. It might be preferable to use least squares with robust standard errors—that should, at least, be one of the options that is tried.

Once positive variance estimates have been obtained using (8.34), with adjustments where necessary, GLS estimates can be obtained by applying least squares to the transformed equation

$$\frac{y_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}} = \beta_1 \frac{1}{\sqrt{\hat{p}_i(1-\hat{p}_i)}} + \beta_2 \frac{x_{i2}}{\sqrt{\hat{p}_i(1-\hat{p}_i)}} + \cdots + \beta_K \frac{x_{iK}}{\sqrt{\hat{p}_i(1-\hat{p}_i)}} + \frac{e_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}}$$

EXAMPLE 8.9 | The Marketing Example Revisited

In Example 7.7 the choice of purchasing either Coke ($COKE = 1$) or Pepsi ($COKE = 0$) was modeled as depending on the relative price of Coke to Pepsi (*PRATIO*), and whether store displays for Coke and Pepsi were present (*DISP_COKE* = 1 if a Coke display was present, otherwise 0; *DISP_PEPSI* = 1 if a Pepsi display was present, otherwise 0). The data file *coke* contains 1140 observations on these variables. Table 8.2 contains the results for (1) least squares, (2) least squares with robust standard errors, (3) GLS with variances below 0.01 truncated to 0.01, and (4) GLS with observations not satisfying $0 < \hat{p}_i < 1$ omitted. For the GLS estimates there were no observations for which $\hat{p}_i > 0.99$ and there were 16 observations where $\hat{p}_i < 0.01$; for these latter cases it was also true that $\hat{p}_i < 0$.

Since the variance function in (8.32) contains the x's, their squares, and their cross products, a suitable test for heteroskedasticity is the White test described in Section 8.6.4. Applying this test to the residuals from the least squares estimated equation yields

$$\chi^2 = N \times R^2 = 25.817 \quad p\text{-value} = 0.0005$$

leading us to reject a null hypothesis of homoskedasticity at a 1% level of significance. Note that, when carrying out this test, your software will omit the squares of *DISP_COKE* and *DISP_PEPSI*. Because these variables are indicator variables, $DISP\_COKE^2 = DISP\_COKE$ and $DISP\_PEPSI^2 = DISP\_PEPSI$, leaving a χ^2-test with 7 degrees of freedom.

Examining the estimates in Table 8.2, we see there is little difference in the four sets of standard errors. In this particular case the use of least squares standard errors does not seem to matter. The four sets of coefficient estimates are also similar with the exception of those from GLS where the negative $\hat{p}$'s were truncated to 0.01. The weight on observations with variance $\text{var}(e_i) = 0.01(1 - 0.01) = 0.0099$ is a relatively large one. It appears that the large weights placed on those 16 observations are having a noticeable impact on the estimates. The signs are all as expected. Making Coke more expensive leads more people to purchase Pepsi. A Coke display encourages purchase of Coke, and a Pepsi display encourages purchase of Pepsi.

In Chapter 16 we study models which are specifically designed for modeling choice between two or more alternatives, and which do not suffer from the problems of the linear probability model.

TABLE 8.2 Linear Probability Model Estimates

	LS	LS-robust	GLS-trunc	GLS-omit
C	0.8902	0.8902	0.6505	0.8795
	(0.0655)	(0.0652)	(0.0568)	(0.0594)
PRATIO	−0.4009	−0.4009	−0.1652	−0.3859
	(0.0613)	(0.0603)	(0.0444)	(0.0527)
DISP	0.0772	0.0772	0.0940	0.0760
_COKE	(0.0344)	(0.0339)	(0.0399)	(0.0353)
DISP	−0.1657	−0.1657	−0.1314	−0.1587
_PEPSI	(0.0356)	(0.0343)	(0.0354)	(0.0360)

8.8 Exercises

8.8.1 Problems

8.1 For the simple regression model with heteroskedasticity, $y_i = \beta_1 + \beta_2 x_i + e_i$ and $\text{var}(e_i|\mathbf{x}_i) = \sigma_i^2$ show that the variance $\text{var}(b_2|\mathbf{x}_i) = \left[\sum_{i=1}^N (x_i - \bar{x})^2\right]^{-1} \left[\sum_{i=1}^N (x_i - \bar{x})^2 \sigma_i^2\right] \left[\sum_{i=1}^N (x_i - \bar{x})^2\right]^{-1}$ reduces to $\text{var}(b_2|\mathbf{x}) = \sigma^2 / \sum_{i=1}^N (x_i - \bar{x})^2$ under homoskedasticity.

8.2 Consider the regression model $y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + e_i$ with two explanatory variables, x_{i1} and x_{i2}, but no constant term.

a. The sum of squares function is $S(\beta_1, \beta_2 | \mathbf{x}_1, \mathbf{x}_2) = \sum_{i=1}^{N} (y_i - \beta_1 x_{i1} - \beta_2 x_{i2})^2$. Find the partial derivatives with respect to the parameters β_1 and β_2. Setting these derivatives to zero and solving, as in Appendix 2A, show that the least squares estimator of β_2 is

$$b_2 = \frac{\left(\sum x_{i1}^2\right)\left(\sum x_{i2} y_i\right) - \left(\sum x_{i1} x_{i2}\right)\left(\sum x_{i1} y_i\right)}{\left(\sum x_{i1}^2\right)\left(\sum x_{i2}^2\right) - \left(\sum x_{i1} x_{i2}\right)^2}$$

b. Let $x_{i1} = 1$ and show that the estimator in (a) reduces to

$$b_2 = \frac{\frac{\sum x_{i2} y_i}{N} - \frac{\sum x_{i2}}{N}\frac{\sum y_i}{N}}{\frac{\sum x_{i2}^2}{N} - \left(\frac{\sum x_{i2}}{N}\right)^2}$$

Compare this equation to equation (2A.5) and show that they are equivalent.

c. In the estimator in part (a), replace y_i, x_{i1} and x_{i2} by $y_i^* = y_i/\sqrt{h_i}$, $x_{i1}^* = x_{i1}/\sqrt{h_i}$ and $x_{i2}^* = x_{i2}/\sqrt{h_i}$. These are transformed variables for the heteroskedastic model $\sigma_i^2 = \sigma^2 h(\mathbf{z}_i) = \sigma^2 h_i$. Show that the resulting GLS estimator can be written as

$$\hat{\beta}_2 = \frac{\sum a_i x_{i2} y_i - \sum a_i x_{i2} \sum a_i y_i}{\sum a_i x_{i2}^2 - \left(\sum a_i x_{i2}\right)^2}$$

where $a_i = 1/(c h_i)$ and $c = \sum(1/h_i)$. Find $\sum_{i=1}^{N} a_i$.

d. Show that under homoskedasticity $\hat{\beta}_2 = b_2$.

e. Explain how $\hat{\beta}_2$ can be said to be constructed from "weighted data averages" while the usual least squares estimator b_2 is constructed from "arithmetic data averages." Relate your discussion to the difference between WLS and ordinary least squares.

8.3 Suppose that an outcome variable $y_{ij} = \beta_1 + \beta_2 x_{ij} + e_{ij}$, $i = 1, \ldots, N$; $j = 1, \ldots, N_i$. Assume $E(e_{ij}|\mathbf{X}) = 0$ and $\text{var}(e_{ij}|\mathbf{X}) = \sigma^2$. One illustration is y_{ij} = the ith farm's production on the jth acre of land, with each farm consisting of N_i acres. The variable x_{ij} is the amount of an input, labor or fertilizer, used by the ith farm on the jth acre.

a. Suppose that we do not have data on each individual acre, but only aggregate, farm-level data, $\sum_{j=1}^{N_i} y_{ij} = y_{Ai}$, $\sum_{j=1}^{N_i} x_{ij} = x_{Ai}$. If we specify the linear model $y_{Ai} = \beta_1 + \beta_2 x_{Ai} + e_{Ai}$, $i = 1, \ldots, N$, what is the conditional variance of the random error?

b. Suppose that we do not have data on each individual acre, but only average data for each farm, $\sum_{j=1}^{N_i} y_{ij}/N_i = \bar{y}_i$, $\sum_{j=1}^{N_i} x_{ij}/N_i = \bar{x}_i$. If we specify the linear model $\bar{y}_i = \beta_1 + \beta_2 \bar{x}_i + \bar{e}_i$, $i = 1, \ldots, N$, what is the conditional variance of the random error?

c. Suppose the outcome variable is binary. For example, suppose $y_{ij} = 1$ if a crop shows evidence of blight on the jth acre of the ith farm, and $y_{ij} = 0$ otherwise. In this case $\sum_{j=1}^{N_i} y_{ij}/N_i = p_i$, where p_i is the sample proportion of acres that show the blight on the ith farm. Suppose the probability of the ith farm showing blight on a particular acre is P_i. If we specify the linear model $\bar{y}_i = \beta_1 + \beta_2 \bar{x}_i + \bar{e}_i$, $i = 1, \ldots, N$, what is the conditional variance of the random error?

8.4 Consider the simple regression model $y_i = \beta_1 + \beta_2 x_i + e_i$ where we hypothesize heteroskedasticity of the form $\sigma_i^2 = \sigma^2 x_i^2$. We have $N = 4$ observations, with $x = (1\ \ 2\ \ 3\ \ 4)$ and $y = (3\ \ 4\ \ 3\ \ 5)$.

a. Use the formula for the least squares estimator in Exercise 8.2(b) to compute the OLS estimate of β_2. In this case $\sum x_{i2} y_i / N = 10$, $\sum x_{i2}^2 / N = 7$.

b. Referring to Exercise 8.2(c), what is the value $c = \sum(1/h_i)$?

c. Referring to Exercise 8.2(c), what are the values $a_i = 1/(c h_i)$, $i = 1, \ldots, 4$? What is $\sum_{i=1}^{4} a_i$?

d. Use the formula for the generalized least squares estimator in Exercise 8.2(c) to compute the GLS estimate of β_2.

e. Suppose that we know that $\sigma^2 = 0.2$. Calculate the true OLS variance given in equation (8.8). The values of $(x_i - \bar{x})^2$ are $(2.25,\ 0.25,\ 0.25,\ 2.25)$. What is the value of the incorrect variance in equation (8.6)?

8.5 Consider the simple regression model $y_i = \beta_1 + \beta_2 x_{i2} + e_i$. Suppose $N = 5$ and the values of x_{i2} are $(1, 2, 3, 4, 5)$. Let the true values of the parameters be $\beta_1 = 1$, $\beta_2 = 1$. Let the true random error values, which are never known in reality, be $e_i = (1,\ -1,\ 0,\ 6,\ -6)$.

a. Calculate the values of y_i.

b. The OLS estimates of the parameters are $b_1 = 3.1$ and $b_2 = 0.3$. Compute the least squares residual, $\hat{e}_1$, for the first observation, and $\hat{e}_4$, for the fourth observation. What is the sum of all the least squares residuals? In this example, what is the sum of the true random errors? Is the sum of the residuals always equal to the sum of the random errors? Explain.

c. It is hypothesized that the data are heteroskedastic with the variance of the first three random errors being σ_1^2, and the variance of the last two random errors being σ_2^2. We regress the squared residuals $\hat{e}_i^2$ on the indicator variable z_i, where $z_i = 0,\ i = 1, 2, 3$ and $z_i = 1,\ i = 4, 5$. The overall model F-statistic value is 12.86. Does this value provide evidence of heteroskedasticity at the 5% level of significance? What is the p-value for this F-value (requires computer)?

d. $R^2 = 0.8108$ from the regression in (c). Use this value to carry out the LM (Breusch–Pagan) test for heteroskedasticity at the 5% level of significance. What is the p-value for this test (requires computer)?

e. We now regress $\ln\left(\hat{e}_i^2\right)$ on z_i. The estimated coefficient of z_i is 3.81. We discover that the software reports using only $N = 4$ observations in this calculation. Why?

f. In order to carry out feasible generalized least squares using information from the regression in part (e), we first create the transformed variables $\left(y_i^*,\ x_{i1}^*,\ x_{i2}^*\right)$. List the values of the transformed observations for $i = 1$ and $i = 4$.

8.6 Consider the wage equation

$$WAGE_i = \beta_1 + \beta_2 EDUC_i + \beta_3 EXPER_i + \beta_4 METRO_i + e_i \qquad \text{(XR8.6a)}$$

where wage is measured in dollars per hour, education and experience are in years, and $METRO = 1$ if the person lives in a metropolitan area. We have $N = 1000$ observations from 2013.

a. We are curious whether holding education, experience, and *METRO* constant, there is the same amount of random variation in wages for males and females. Suppose $\text{var}\left(e_i|\mathbf{x}_i, FEMALE = 0\right) = \sigma_M^2$ and $\text{var}\left(e_i|\mathbf{x}_i, FEMALE = 1\right) = \sigma_F^2$. We specifically wish to test the null hypothesis $\sigma_M^2 = \sigma_F^2$ against $\sigma_M^2 \neq \sigma_F^2$. Using 577 observations on males, we obtain the sum of squared OLS residuals, $SSE_M = 97161.9174$. The regression using data on females yields $\hat{\sigma}_F = 12.024$. Test the null hypothesis at the 5% level of significance. Clearly state the value of the test statistic and the rejection region, along with your conclusion.

b. We hypothesize that married individuals, relying on spousal support, can seek wider employment types and hence holding all else equal should have more variable wages. Suppose $\text{var}\left(e_i|\mathbf{x}_i, MARRIED = 0\right) = \sigma_{SINGLE}^2$ and $\text{var}\left(e_i|\mathbf{x}_i, MARRIED = 1\right) = \sigma_{MARRIED}^2$. Specify the null hypothesis $\sigma_{SINGLE}^2 = \sigma_{MARRIED}^2$ versus the alternative hypothesis $\sigma_{MARRIED}^2 > \sigma_{SINGLE}^2$. We add *FEMALE* to the wage equation as an explanatory variable, so that

$$WAGE_i = \beta_1 + \beta_2 EDUC_i + \beta_3 EXPER_i + \beta_4 METRO_i + \beta_5 FEMALE + e_i \qquad \text{(XR8.6b)}$$

Using $N = 400$ observations on single individuals, OLS estimation of (XR8.6b) yields a sum of squared residuals is 56231.0382. For the 600 married individuals, the sum of squared errors is 100,703.0471. Test the null hypothesis at the 5% level of significance. Clearly state the value of the test statistic and the rejection region, along with your conclusion.

c. Following the regression in part (b), we carry out the NR^2 test using the right-hand-side variables in (XR8.6b) as candidates related to the heteroskedasticity. The value of this statistic is 59.03. What do we conclude about heteroskedasticity, at the 5% level? Does this provide evidence about the issue discussed in part (b), whether the error variation is different for married and unmarried individuals? Explain.

d. Following the regression in part (b) we carry out the White test for heteroskedasticity. The value of the test statistic is 78.82. What are the degrees of freedom of the test statistic? What is the 5% critical value for the test? What do you conclude?

e. The OLS fitted model from part (b), with usual and robust standard errors, is

$$\widehat{\text{WAGE}} = -17.77 + 2.50EDUC + 0.23EXPER + 3.23METRO - 4.20FEMALE$$
$$\begin{array}{lllll} \text{(se)} & (2.36)\ (0.14) & (0.031) & (1.05) & (0.81) \\ \text{(robse)} & (2.50)\ (0.16) & (0.029) & (0.84) & (0.80) \end{array}$$

For which coefficients have interval estimates gotten narrower? For which coefficients have interval estimates gotten wider? Is there an inconsistency in the results?

f. If we add *MARRIED* to the model in part (b), we find that its t-value using a White heteroskedasticity robust standard error is about 1.0. Does this conflict with, or is it compatible with, the result in (b) concerning heteroskedasticity? Explain.

8.7 Consider the simple treatment effect model $y_i = \beta_1 + \beta_2 d_i + e_i$. Suppose that $d_i = 1$ or $d_i = 0$ indicating that a treatment is given to randomly selected individuals or not. The dependent variable y_i is the outcome variable. See the discussion of the difference estimator in Section 7.5.1. Suppose that N_1 individuals are given the treatment and N_0 individual are in the control group, who are not given the treatment. Let $N = N_0 + N_1$ be the total number of observations.

a. Show that if $\text{var}(e_i|\mathbf{d}) = \sigma^2$ then the variance of the OLS estimator b_2 of β_2 is $\text{var}(b_2|\mathbf{d}) = N\sigma^2/(N_0N_1)$. [*Hint*: See Appendix 7B.]

b. Let $\bar{y}_0 = \sum_{i=1}^{N_0} y_i/N_0$ be the sample mean of the outcomes for the N_0 observations on the control group. Let $SST_0 = \sum_{i=1}^{N_0}(y_i - \bar{y}_0)^2$ be the sum of squares about the sample mean of the control group, where $d_i = 0$. Similarly, let $\bar{y}_1 = \sum_{i=1}^{N_1} y_i/N_1$ be the sample mean of the outcomes for the N_1 observations on the treated group, where $d_i = 1$. Let $SST_1 = \sum_{i=1}^{N_1}(y_i - \bar{y}_1)^2$ be the sum of squares about the sample mean of the treatment group. Show that $\hat{\sigma}^2 = \sum_{i=1}^{N} \hat{e}_i^2/(N-2) = (SST_0 + SST_1)/(N-2)$ and therefore that

$$\widehat{\text{var}}(b_2|\mathbf{d}) = N\hat{\sigma}^2/(N_0N_1) = \left(\frac{N}{N-2}\right)\left(\frac{SST_0 + SST_1}{N_0N_1}\right)$$

c. Using equation (2.14) find $\text{var}(b_1|\mathbf{d})$, where b_1 is the OLS estimator of the intercept parameter β_1. What is $\widehat{\text{var}}(b_1|\mathbf{d})$?

d. Suppose that the treatment and control groups have not only potentially different means but potentially different variances, so that $\text{var}(e_i|d_i = 1) = \sigma_1^2$ and $\text{var}(e_i|d_i = 0) = \sigma_0^2$. Find $\text{var}(b_2|\mathbf{d})$. What is the unbiased estimator for $\text{var}(b_2|\mathbf{d})$? [*Hint*: See Appendix C.4.1.]

e. Show that the White heteroskedasticity robust estimator in equation (8.9) reduces in this case to $\widehat{\text{var}}(b_2|\mathbf{d}) = \frac{N}{N-2}\left(\frac{SST_0}{N_0^2} + \frac{SST_1}{N_1^2}\right)$. Compare this estimator to the unbiased estimator in part (d).

f. What does the robust estimator become if we drop the degrees of freedom correction $N/(N-2)$ in the estimator proposed in part (e)? Compare this estimator to the unbiased estimator in part (d).

8.8 It can be shown that the theoretically useful form of the OLS estimator of β_1 in the simple linear regression model $y_i = \beta_1 + \beta_2 x_{i2} + e_i$ is $b_1 = \beta_1 + \sum(-\bar{x}w_i + N^{-1})e_i = \sum v_i e_i$, where $v_i = (-\bar{x}w_i + N^{-1})$ and $w_i = (x_i - \bar{x})/\sum(x_i - \bar{x})^2$. Using this formula consider the simple treatment effect model $y_i = \beta_1 + \beta_2 d_i + e_i$. Suppose that $d_i = 1$ or $d_i = 0$ indicating that a treatment is given to a randomly selected individual or not. The dependent variable y_i is the outcome variable. See the discussion of the difference estimator in Section 7.5.1. Suppose that N_1 individuals are given the treatment and N_0 individuals in the control group are not given the treatment. Let $N = N_0 + N_1$ be the total number of observations.

a. Show that when $d_i = 0$, $v_i = 1/N$ and that when $d_i = 1$, $v_i = 0$.

b. Derive $\text{var}(b_1|\mathbf{d})$ under the assumption of homoskedastic errors, $\text{var}(e_i|\mathbf{d}) = \sigma^2$. What is an unbiased estimator of $\text{var}(b_1|\mathbf{d})$ in this case?

c. Derive $\text{var}(b_1|\mathbf{d})$ under the assumption of heteroskedastic errors, $\text{var}(e_i|d_i = 1) = \sigma_1^2$ and $\text{var}(e_i|d_i = 0) = \sigma_0^2$. What is an unbiased estimator of $\text{var}(b_1|\mathbf{d})$ in this case?

8.9 We wish to estimate the hedonic regression model

$$PRICE_i = \beta_1 + \beta_2 SQFT_i + \beta_3 CLOSE_i + \beta_4 AGE_i + \beta_5 FIREPLACE_i + \beta_6 POOL_i + \beta_7 TWOSTORY_i + e_i$$

The variables are *PRICE* ($1000), *SQFT* (100s), *CLOSE* = 1 if located near a major university, 0 otherwise, *AGE* (years), *FIREPLACE*, *POOL*, *TWOSTORY* = 1 if present, 0 otherwise.

a. Using Table 8.3, comment on the sign, significance, and interpretation of the OLS coefficient estimate for the variable *CLOSE*.

b. Answer each of the following True or False. In a regression model with heteroskedasticity, (i) the OLS estimator is biased; (ii) the OLS estimator is inconsistent; (iii) the OLS estimator does not have an approximate normal distribution in large samples; (iv) the usual OLS standard error is too small; (v) the usual OLS estimator standard error is incorrect; (vi) the usual R^2 is no longer meaningful; (vii) the usual overall F-test is reliable in large samples.

c. Following the OLS regression, the residuals are saved as *EHAT*. In the regression labeled AUX in Table 8.3, the dependent variable is $EHAT^2$. Test for the presence of heteroskedasticity, using the 5% level of significance. State the test statistic, the test critical value, and your conclusion.

d. The model is reestimated by OLS using White heteroskedasticity-consistent standard errors. In what way are these standard errors robust? Are they valid when there is homoskedasticity, heteroskedasticity, in small samples and large? Which of the statistically significant coefficients has wider confidence intervals using the robust standard errors? Do any coefficients switch from being significant at 5% to not significant at 5%, or vice versa?

e. Our researcher estimates the equation after dividing each variable, and the constant term, by *SQFT* to obtain the GLS estimates. What assumption has been made about the form of heteroskedasticity in this estimation? Are the GLS estimates, shown in Table 8.3, noticeably different from the OLS estimates? Do any coefficients switch from being significant at 5% to not significant at 5%, or vice versa?

f. The residuals from the transformed regression in part (e) are called *ESTAR*. The researcher regresses $ESTAR^2$ on all the transformed variables and includes an intercept. The $R^2 = 0.0237$. Has the researcher eliminated heteroskedasticity?

g. The researcher estimates the model in (e) again but uses robust standard errors. These are reported in Table 8.3 as "Robust GLS." Do you consider this a prudent thing to do? Explain your reasoning.

TABLE 8.3 **Estimates for Exercise 8.9**

	OLS	AUX	Robust OLS	GLS	Robust GLS
C	−101.072***	−25561.243***	−101.072***	−4.764	−4.764
	(27.9055)	(5419.9443)	(34.9048)	(21.1357)	(35.8375)
SQFT	13.3417***	1366.8074***	13.3417***	7.5803***	7.5803***
	(0.5371)	(104.3092)	(1.1212)	(0.5201)	(0.9799)
CLOSE	26.6657***	1097.8933	26.6657***	39.1988***	39.1988***
	(9.8602)	(1915.0902)	(9.6876)	(7.0438)	(7.2205)
AGE	−2.7305	52.4499	−2.7305	1.4887	1.4887
	(2.7197)	(528.2353)	(3.2713)	(2.1034)	(2.5138)
FIREPLACE	−2.2585	−3005.1375	−2.2585	17.3827**	17.3827*
	(10.5672)	(2052.4109)	(10.6369)	(7.9023)	(9.3531)
POOL	0.3601	6878.0158*	0.3601	8.0265	8.0265
	(19.1855)	(3726.2941)	(27.2499)	(17.3198)	(15.6418)
TWOSTORY	5.8833	−7394.3869**	5.8833	26.7224*	26.7224*
	(14.8348)	(2881.2790)	(20.8733)	(13.7616)	(16.0651)
R^2	0.6472	0.3028	0.6472	0.4427	0.4427

Standard errors in parentheses
*$p < 0.10$
**$p < 0.05$
***$p < 0.01$

8.10 Does having more children drive parents to drink more alcohol? We have data on the following variables: *WALC* = budget share (percent of income spent) for alcohol expenditure; *INCOME* = total net household income (10,000 UK pounds); *AGE* = age of household head/10; *NK* = number of children (1 or 2). We are interested in the equation

$$\ln(WALC) = \beta_1 + \beta_2 INCOME + \beta_3 AGE + \beta_4 NK + e$$

a. The data we have is based on a survey. If we hope to establish a causal relationship between *NK* and the budget share spent on alcohol, what assumptions are sufficient to prove that the least squares estimator is BLUE?

b. Using 1278 observations on households with a positive budget share for alcohol, the OLS estimated equation, with conventional standard errors, is

$$\widehat{\ln(WALC)} = -1.956 + 0.837 INCOME - 0.228 AGE - 0.251 NK$$
$$\text{(se)} \qquad (0.166) \quad (0.516) \qquad (0.039) \qquad (0.058)$$

Test the null hypothesis that an increase in the number of children from one to two has no effect on the budget share of alcohol versus the alternative that an increase in the number of children increases the budget share of alcohol. Use the 5% level of significance.

c. We suspect that the regression error variance might be larger for households with two children rather than one. We estimate the budget share equation by least squares separately for households with one and two children. For the 489 households with one child, the sum of squared residuals is 465.83. For the 789 households with two children, the sum of squared residuals is 832.77. Test the null hypothesis that there is no difference between the regression error variances for these two groups, against the alternative that there is a difference. Use the Goldfeld–Quandt test at the 5% level of significance. Repeat the test using the alternative that the regression error variance for the subsample of households with two children is greater than the regression error variance for the subsample of households with one child. What do you conclude?

d. We save the least squares residuals from the estimation in part (b), calling them *EHAT*. We then obtain the second-stage regression results $EHAT^2 = 0.012 + 0.279 AGE + 0.025 NK$ with an $R^2 = 0.0208$. Is there evidence of heteroskedasticity? Set up the appropriate hypothesis and carry out the test at the 1% level of significance. What do you conclude?

e. We then carry out the regression $\widehat{\ln(EHAT^2)} = -2.088 + 0.291 AGE - 0.048 NK$. Holding *NK* constant, calculate the estimated variance ratio $\widehat{\text{var}}(e_i|AGE = 40)/\widehat{\text{var}}(e_i|AGE = 30)$. [*Hint*: Recall that *AGE* is measured in units of 10 years.] What is the estimated ratio $\widehat{\text{var}}(e_i|AGE = 60)/\widehat{\text{var}}(e_i|AGE = 30)$? Holding *AGE* constant, calculate the estimated variance ratio $\widehat{\text{var}}(e_i|NK = 2)/\widehat{\text{var}}(e_i|NK = 1)$.

f. Based on the results we have obtained so far, can we claim that the least squares estimator used in (b) is BLUE?

g. What model would we estimate by OLS to implement feasible generalized least squares estimation?

8.11 We are interested in the relationship between rice production, inputs of labor and fertilizer, and the area planted using data on $N = 44$ farms.

$$RICE_i = \beta_1 + \beta_2 LABOR_i + \beta_3 FERT_i + \beta_4 ACRES_i + e_i$$

a. We observe the least squares residuals, $\hat{e}_i$, increase in magnitude when plotted against *ACRES*. We regress $\hat{e}_i^2$ on *ACRES* and obtain a regression with $R^2 = 0.2068$. The estimated coefficient of *ACRES* is 2.024 with the standard error of 0.612. What can we conclude about heteroskedasticity based on these results? Explain your reasoning.

b. We instead estimate the model

$$RICE_i/ACRES_i = \alpha + \beta_1(1/ACRES_i) + \beta_2 LABOR_i/ACRES_i + \beta_3 FERT_i/ACRES_i + e_i$$

What is the implicit assumption about the heteroskedasticity pattern?

c. Many economists would omit $(1/ACRES_i)$ from the equation. What argument can you propose that would make this defensible?

d. Following the estimation of the model in (b) or (c), the squared residuals, $\tilde{e}_i^2$, are regressed on *ACRES*. The estimated coefficient is negative and significant at the 10% level. The regression

$R^2 = 0.0767$. What might you conclude about the models in (b) or (c)? That is, what could have led to such results?

e. In a further step, we estimate $\ln\left(\hat{e}_i^2\right) = -1.30 + 1.11\ln(ACRES)$ and $\ln\left(\tilde{e}_i^2\right) = -1.20 - 1.21\ln(ACRES)$. What evidence does this provide about the question in part (d)?

f. If we estimate the model in (c), omitting $\left(1/ACRES_i\right)$, would you advise using White heteroskedasticity robust standard errors? Explain why or why not.

8.12 An econometrician wishes to study the properties of an estimator using simulated data. Suppose the sample size N is set to be 100. The intercept and slope parameters are 100, and 10, respectively. The one explanatory variable, x, has a normal distribution with mean 10 and standard deviation 10. A standard normal random variable, z, independent of x, is created. The data generating process is $y_i = \beta_1 + \beta_2 x_i + e_i$, where

$$e_i = \begin{cases} z_i & \text{if } i \text{ is an odd number} \\ 2z_i & \text{if } i \text{ is an even number} \end{cases}$$

a. The OLS estimator is not the best linear unbiased estimator using the 100 data pairs $\left(y_i, x_i\right)$. True or false? Explain.

b. If we divide y and x for the even number observations by $\sqrt{2}$, leaving the odd number observations alone, and then run a least squares regression, the resulting estimator is BLUE. True or false? Explain.

c. Suppose you were assigned the task of showing that the heteroskedasticity in the data was "statistically significant." Using the 100 data pairs $\left(y_i, x_i\right)$, how exactly would you do it?

8.13 A researcher has 1100 observations on household expenditures on entertainment (per person in the previous quarter, \$) *ENTERT*. The researcher wants to explain these expenditures as a function of *INCOME* (monthly income during past year, \$100 units), whether the household lives in an *URBAN* area, and whether someone in the household has a *COLLEGE* degree (Bachelor's) or an *ADVANCED* degree (Master's or Ph.D.). *COLLEGE* and *ADVANCED* are indicator variables.

a. The OLS estimates and t-values are given in Table 8.4, on the next page. Taking the residuals from this regression, and regressing their squared values on all explanatory variables yields an $R^2 = 0.0344$. Such a small value implies there is no heteroskedasticity, correct? If that statement is not correct, then carry out the proper test. What do you conclude about the presence of heteroskedasticity?

b. To be safe the researcher uses White heteroskedasticity robust standard errors, given in Table 8.4. The researcher's paper has to do with the effect on entertainment expenditures of having someone with an advanced degree in the household. Compare the significance of *ADVANCED* in the two OLS regressions. What do you find? It is generally true that robust standard errors are larger than ones that are not robust. Is that true or false in this case?

c. Because of the importance of the variable *ADVANCED* in the model, the researcher takes some additional effort. Using the OLS residuals $\hat{e}_i$, the researcher obtains

$$\ln\left(\hat{e}_i^2\right) = 4.9904 + 0.0177INCOME_i + 0.2902ADVANCED_i$$
$$(t) \qquad (10.92) \qquad (1.80)$$

What evidence about heteroskedasticity is present in these results?

d. The researcher takes the results in (c) and then calculates

$$h_i = \exp\left(0.0177INCOME_i + 0.2902ADVANCED_i\right)$$

Each variable, including the intercept, is divided by $\sqrt{h_i}$ and the model reestimated to obtain the FGLS results in Table 8.4. Based on these results, how much of an effect on entertainment expenditures is there for households including someone with an advanced degree? Is this statistically significant? To which set of OLS results, can we make a valid comparison with the FGLS estimates? Have we improved the estimation of the effect of *ADVANCED* on entertainment by taking the steps in (c) and (d)? Provide a very careful answer to this question.

e. Looking for an easier way the researcher estimates a log-linear model shown in Table 8.4. Following this estimation, regressing the squared residuals on the explanatory variables, we find $NR^2 = 2.46$. Using White's test, including all the squares and cross-products of the explanatory variables, we obtain $NR^2 = 6.63$. What are the critical values for each of these test statistics? Using a test at the 5% level, do we reject homoskedasticity in the log-linear model or not?

f. Interpret the regression results in (e) from the point of view of the researcher interested in the effect of *ADVANCED* on entertainment expenditures. What exactly has happened by using the log-linear model? Provide an intuitive explanation. As a hint, Figure 8.9 shows entertainment expenditures for one range of income, between $7000/mo and $8000/mo.

TABLE 8.4 Estimates for Exercise 8.13

	OLS	Robust OLS	FGLS	Log-linear
C	20.5502	20.5502	18.5710	2.7600
	(3.19)	(3.30)	(4.16)	(25.79)
INCOME	0.5032	0.5032	0.4447	0.0080
	(10.17)	(6.45)	(8.75)	(9.77)
URBAN	−6.4629	−6.4629)	−0.8420	0.0145
	(−1.06)	(−0.81)	(−0.20)	(0.14)
COLLEGE	−0.7155	−0.7155	1.7388	0.0576
	(−0.16)	(−0.15)	(0.52)	(0.77)
ADVANCED	9.8173	9.8173	9.0123	0.2315
	(1.87)	(1.58)	(1.92)	(2.65)

t-values in parentheses

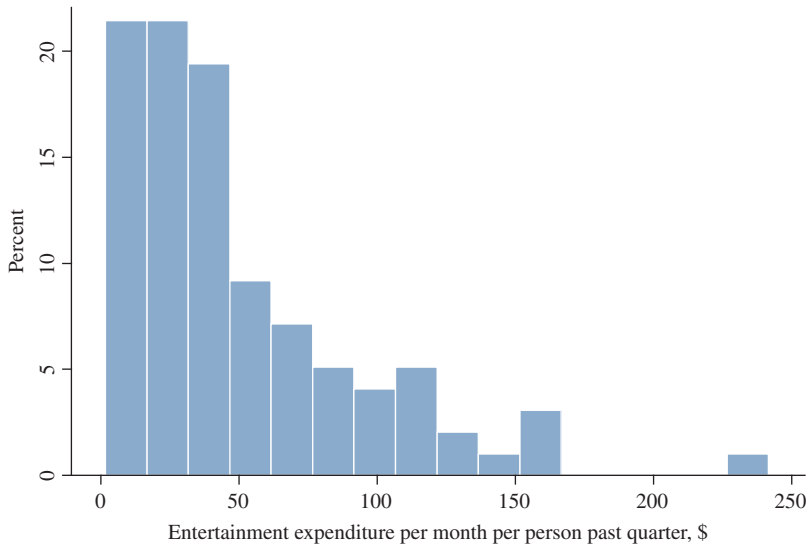

FIGURE 8.9 Histogram for entertainment expenditure.

8.14 Using data on 1000 home loan borrowers, we estimate the linear probability model

$$DEFAULT = \beta_1 + \beta_2 LTV + \beta_3 RATE + \beta_4 AMOUNT + \beta_5 FICO + e$$

where $DEFAULT = 1$ if the borrower has made a mortgage payment more than 90 days late, $LTV = 100$(loan amount/property value), *RATE* is the interest rate, *AMOUNT* ($10,000 units) of the loan, and *FICO* is the borrower's credit score.

a. Figure 8.10(a) is the histogram of the least squares residuals, $\hat{e}$. Explain the bimodal shape.

b. Figure 8.10(b) is the histogram of the least squares fitted values,

$$\widehat{DEFAULT} = 0.6887 + 0.0055LTV + 0.0482RATE - 0.0012AMOUNT - 0.0014FICO$$

Explain the interpretation of the fitted values. Do you find any unusual fitted values in the figure?

c. Let Y be a Bernoulli random variable, taking the values 1 and 0 with probabilities P and $1 - P$. Show that $\text{var}(Y) = P(1 - P)$.

d. Regressing $\hat{e}_i^2$ on the explanatory variables, we obtain $R^2 = 0.0206$ and the model F-statistic is 5.22. What does each of these values tell us about the null hypothesis of homoskedasticity in this model? Provide any relevant test statistics, and their 5% level of significance critical values. In light of part (c), are the results surprising?

e. Consider two hypothetical borrowers:

Borrower 1: $LTV = 85$, $RATE = 11$, $AMOUNT = 400$, $FICO = 500$

Borrower 2: $LTV = 50$, $RATE = 5$, $AMOUNT = 100$, $FICO = 700$

The 95% interval estimates, for the expected probability of default for the hypothetical borrowers using OLS, OLS with heteroskedasticity robust standard errors, and FGLS are given in Table 8.5. Discuss these interval estimates. If two such borrowers came for a loan, to whom would you offer one?

f. To obtain the FGLS estimates in (e), negative predicted values in nine observations are turned to positives by taking their absolute value. Why did we do that? What other alternatives did we have?

TABLE 8.5 **Interval Estimates for Exercise 8.14(e)**

Borrower	Method	Lower Bound	$\widehat{DEFAULT}$	Upper Bound	Std. Err.
1	OLS	−0.202	0.527	1.257	0.372
1	OLS (robust)	−0.132	0.527	1.187	0.337
1	FGLS	−0.195	0.375	0.946	0.291
2	OLS	−0.043	0.116	0.275	0.082
2	OLS (robust)	−0.025	0.116	0.257	0.072
2	FGLS	−0.019	0.098	0.215	0.060

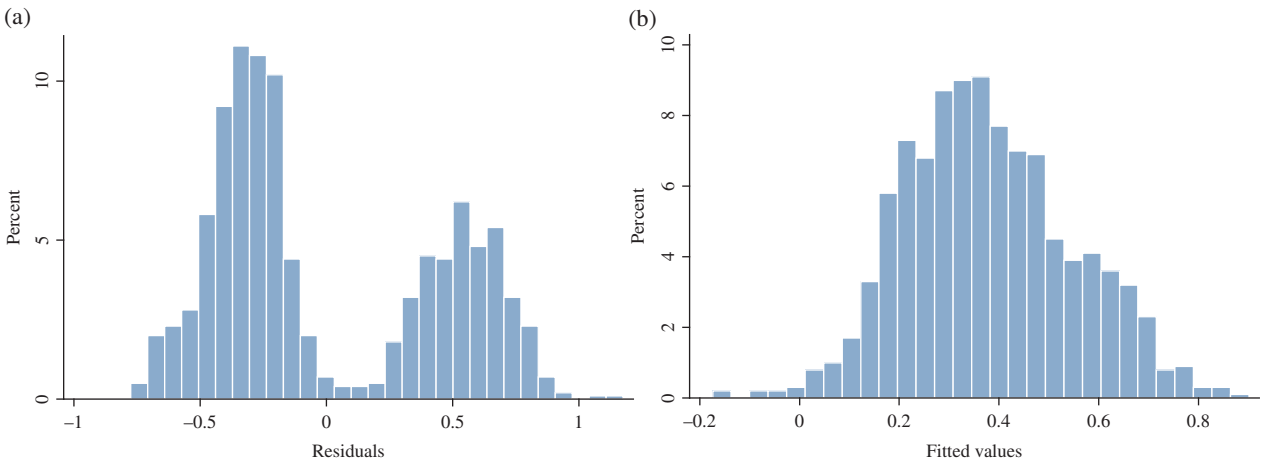

FIGURE 8.10 **Histograms for residuals and fitted values for Exercise 8.14.**

8.15 We have $N = 396$ observations on employment at fast-food restaurants in two neighboring states, New Jersey and Pennsylvania. In Pennsylvania, the control group $d_i = 0$, there is no minimum wage law. In New Jersey, the treatment group $d_i = 1$, there is a minimum wage law. Let the observed outcome variable be full-time employment FTE_i at comparable fast-food restaurants. Some sample summary statistics for FTE_i in the two states are in Table 8.6. For Pennsylvania, the sample size is $N_0 = 77$, the sample mean is $\overline{FTE}_0 = \sum_{i=1,d_i=0}^{N_0} FTE_i/N_0$, the sample variance is $s_0^2 = \sum_{i=1,d_i=0}^{N_0}\left(FTE_i - \overline{FTE}_0\right)^2/(N_0 - 1) = SST_0/(N_0 - 1)$, the sample standard deviation is $s_0 = \sqrt{s_0^2}$, and the standard error of mean is $se_0 = \sqrt{s_0^2/N_0} = s_0/\sqrt{N_0}$. For New Jersey, the definitions are comparable with subscripts "1".

TABLE 8.6 **Summary Statistics for Exercise 8.15**

State	d	N	Sample Mean	Sample Variance	Sample Standard Deviation	Standard Error of the Mean
Pennsylvania (control)	0	77	21.16558	68.50429	8.276732	0.9432212
New Jersey (treatment)	1	319	21.02743	86.36029	9.293024	0.5203094

a. Consider the regression model $FTE_i = \beta_1 + \beta_2 d_i + e_i$. The OLS estimates are given below, along with the usual standard errors (se), the White heteroskedasticity robust standard errors (robse), and an alternative robust standard error (rob2).

$$\begin{aligned} \widehat{FTE}_i &= 21.16558 - 0.1381549 d_i \\ \text{(se)} \quad & (1.037705) \quad (1.156182) \\ \text{(robse)} \quad & (0.9394517) \quad (1.074157) \\ \text{(rob2)} \quad & (0.9432212) \quad (1.077213) \end{aligned}$$

Show the relationship between the least squares estimates of the coefficients, the estimated slope and intercept, and the summary statistics in Table 8.6.

b. Calculate $\widehat{\text{var}}(b_2|\mathbf{d}) = N\hat{\sigma}^2/(N_0 N_1) = \left(\frac{N}{N-2}\right)\left(\frac{SST_0 + SST_1}{N_0 N_1}\right)$, derived in Exercise 8.7(b). Compare the standard error of the slope using this expression to the regression output in part (a).

c. Suppose that the treatment and control groups have not only potentially different means but potentially different variances, so that $\text{var}(e_i|d_i = 1) = \sigma_1^2$ and $\text{var}(e_i|d_i = 0) = \sigma_0^2$. Carry out the Goldfeld–Quandt test of the null hypothesis $\sigma_0^2 = \sigma_1^2$ at the 1% level of significance. [*Hint*: See Appendix C.7.3.]

d. In Exercise 8.7(e), we showed that the heteroskedasticity robust variance for the slope estimator is $\widehat{\text{var}}(b_2|\mathbf{d}) = \frac{N}{N-2}\left(\frac{SST_0}{N_0^2} + \frac{SST_1}{N_1^2}\right)$. Use the summary statistic data to calculate this quantity. Compare the heteroskedasticity robust standard error of the slope using this expression to those from the regression output. In Appendix 8D, we discuss several heteroskedasticity robust variance estimators. This one is most common and usually referred to as "HCE1," where HCE stands for "heteroskedasticity consistent estimator."

e. Show that the alternative robust standard error, rob2, can be computed from $\widehat{\text{var}}(b_2|\mathbf{d}) = \frac{SST_0}{N_0(N_0-1)} + \frac{SST_1}{N_1(N_1-1)}$. In Appendix 8D, this estimator is called "HCE2." Note that it can be written $\widehat{\text{var}}(b_2|\mathbf{d}) = \left(\hat{\sigma}_0^2/N_0\right) + \left(\hat{\sigma}_1^2/N_1\right)$, where $\hat{\sigma}_0^2 = SST_0/(N_0 - 1)$ and $\hat{\sigma}_1^2 = SST_1/(N_1 - 1)$. These estimators are unbiased and are discussed in Appendix C.4.1. Is the variance estimator unbiased if $\sigma_0^2 = \sigma_1^2$?

f. The estimator HCE1 is $\widehat{\text{var}}(b_2|\mathbf{d}) = \frac{N}{N-2}\left(\frac{SST_0}{N_0^2} + \frac{SST_1}{N_1^2}\right)$. Show that dropping the degrees of freedom correction $N/(N-2)$ it becomes HCE0, $\widehat{\text{var}}(b_2|\mathbf{d}) = \left(\tilde{\sigma}_0^2/N_0\right) + \left(\tilde{\sigma}_1^2/N_1\right)$, where $\tilde{\sigma}_0^2 = SST_0/N_0$ and $\tilde{\sigma}_1^2 = SST_1/N_1$ are biased but consistent estimators of the variances. See Appendix C.4.2. Calculate the standard error for b_2 using this alternative.

g. A third variant of a robust variance estimator, HCE3, is $\widehat{\text{var}}(b_2|\mathbf{d}) = \left(\frac{\hat{\sigma}_0^2}{N_0 - 1}\right) + \left(\frac{\hat{\sigma}_1^2}{N_1 - 1}\right)$, where $\hat{\sigma}_0^2 = SST_0/(N_0 - 1)$ and $\hat{\sigma}_1^2 = SST_1/(N_1 - 1)$. Calculate the robust standard error using HCE3 for this example. In this application, comparing HCE0 to HCE2 to HCE3, which is largest? Which is smallest?

8.8.2 Computer Exercises

8.16 A sample of 200 Chicago households was taken to investigate how far American households tend to travel when they take a vacation. Consider the model

$$MILES = \beta_1 + \beta_2 INCOME + \beta_3 AGE + \beta_4 KIDS + e$$

MILES is miles driven per year, *INCOME* is measured in \$1000 units, *AGE* is the average age of the adult members of the household, and *KIDS* is the number of children.

a. Use the data file *vacation* to estimate the model by OLS. Construct a 95% interval estimate for the effect of one more child on miles traveled, holding the two other variables constant.

b. Plot the OLS residuals versus *INCOME* and *AGE*. Do you observe any patterns suggesting that heteroskedasticity is present?

c. Sort the data according to increasing magnitude of income. Estimate the model using the first 90 observations and again using the last 90 observations. Carry out the Goldfeld–Quandt test for heteroskedastic errors at the 5% level. State the null and alternative hypotheses.

d. Estimate the model by OLS using heteroskedasticity robust standard errors. Construct a 95% interval estimate for the effect of one more child on miles traveled, holding the two other variables constant. How does this interval estimate compare to the one in (a)?

e. Obtain GLS estimates assuming $\sigma_i^2 = \sigma^2 INCOME_i^2$. Using both conventional GLS and robust GLS standard errors, construct a 95% interval estimate for the effect of one more child on miles traveled, holding the two other variables constant. How do these interval estimates compare to the ones in (a) and (d)?

8.17 In this exercise, we explore the relationship between total household expenditures and expenditures on clothing. Use the data file *malawi_small* (*malawi* has more observations) and observations for which *PCLOTHES* is positive. We consider three models:

$$PCLOTHES = \beta_1 + \beta_2 \ln(TOTEXP) + e \quad \text{(XR8.17a)}$$

$$\ln(CLOTHES) = \alpha_1 + \alpha_2 \ln(TOTEXP) + v \quad \text{(XR8.17b)}$$

$$CLOTHES = \gamma_1 + \gamma_2 TOTEXP + u \quad \text{(XR8.17c)}$$

a. Plot *PCLOTHES* versus ln(*TOTEXP*) and include the least squares fitted line. Calculate the point elasticity of clothing expenditures with respect to total expenditures at the means. See Exercise 4.12 for the elasticity in this model.

b. Calculate $CLOTHES = PCLOTHES \times TOTEXP$. Then plot ln(*CLOTHES*) versus ln(*TOTEXP*) and include the least squares fitted line. Calculate a 95% interval estimate of the elasticity of clothing expenditures with respect to total expenditures. Is the elasticity computed in part (a) within this interval?

c. Plot *CLOTHES* versus *TOTEXP* and include the least squares fitted line. Calculate a 95% interval estimate of the elasticity of clothing expenditures with respect to total expenditures at the means. Is the elasticity computed in part (a) within this interval?

d. Test for the presence of heteroskedasticity in each model in parts (a)–(c). Use the 1% level of significance. What are your conclusions? For which specification does heteroskedasticity seem less of an issue?

e. For the models in which heteroskedasticity was significant at the 1% level, use OLS with robust standard errors. Calculate a 95% interval estimate for the elasticity of clothing expenditures with respect to total expenditures at the means. How do the intervals compare to the ones based on conventional standard errors?

8.18 Consider the wage equation,

$$\ln(WAGE_i) = \beta_1 + \beta_2 EDUC_i + \beta_3 EXPER_i + \beta_4 EXPER_i^2 + \beta_5 FEMALE_i + \beta_6 BLACK + \beta_7 METRO_i + \beta_8 SOUTH_i + \beta_9 MIDWEST_i + \beta_{10} WEST + e_i$$

where *WAGE* is measured in dollars per hour, education and experience are in years, and *METRO* = 1 if the person lives in a metropolitan area. Use the data file *cps5* for the exercise.

a. We are curious whether holding education, experience, and *METRO* equal, there is the same amount of random variation in wages for males and females. Suppose $\text{var}(e_i|\mathbf{x}_i, FEMALE = 0) = \sigma_M^2$ and $\text{var}(e_i|\mathbf{x}_i, FEMALE = 1) = \sigma_F^2$. We specifically wish to test the null hypothesis $\sigma_M^2 = \sigma_F^2$ against $\sigma_M^2 \neq \sigma_F^2$. Carry out a Goldfeld–Quandt test of the null hypothesis at the 5% level of significance. Clearly state the value of the test statistic and the rejection region, along with your conclusion.

b. Estimate the model by OLS. Carry out the NR^2 test using the right-hand-side variables *METRO*, *FEMALE*, *BLACK* as candidates related to the heteroskedasticity. What do we conclude about heteroskedasticity, at the 1% level? Do these results support your conclusions in (a)? Repeat the test using all model explanatory variables as candidates related to the heteroskedasticity.

c. Carry out the White test for heteroskedasticity. What is the 5% critical value for the test? What do you conclude?

d. Estimate the model by OLS with White heteroskedasticity robust standard errors. Compared to OLS with conventional standard errors, for which coefficients have interval estimates gotten narrower? For which coefficients have interval estimates gotten wider? Is there an inconsistency in the results?

e. Obtain FGLS estimates using candidate variables *METRO* and *EXPER*. How do the interval estimates compare to OLS with robust standard errors, from part (d)?

f. Obtain FGLS estimates with robust standard errors using candidate variables *METRO* and *EXPER*. How do the interval estimates compare to those in part (e) and OLS with robust standard errors, from part (d)?

g. If reporting the results of this model in a research paper which one set of estimates would you present? Explain your choice.

8.19 In this exercise we explore the relationship between total household expenditures and expenditures on telephone services. Use the data file *malawi_small* (*malawi* has more observations).

a. Using observations for which $PTELEPHONE > 0$, create the variable $\ln(TELEPHONE) = \ln(PTELEPHONE \times TOTEXP)$. Plot $\ln(TELEPHONE)$ versus $\ln(TOTEXP)$ and include the least squares fitted line.

b. Based on the OLS regression of $\ln(TELEPHONE)$ on $\ln(TOTEXP)$ what is the estimated elasticity of telephone expenditures with respect to total expenditure. Compute a 95% interval estimate for the elasticity. Based on the estimates, would you classify telephone services as a necessity or a luxury?

c. Test for the presence of heteroskedasticity in the regression in part (b). What do you conclude?

d. Estimate the model $PTELEPHONE_i = \beta_1 + \beta_2 \ln(TOTEXP_i) + e_i$ by OLS. Test the null hypothesis that $\beta_2 \leq 0$ against $\beta_2 > 0$ using the 5% level of significance.

e. Calculate the elasticity of telephone expenditures with respect to total expenditure at the sample median of total expenditures. The expression for an elasticity in such a model was derived in Exercise 4.12. Use your software to compute a 95% interval estimate for the elasticity. Compare the estimated elasticity to that in (b).

f. Test for the presence of heteroskedasticity in the regression in part (d). What do you conclude?

g. Estimate the model in (d) using FGLS with $\ln(TOTEXP_i)$ being the variable that may be associated with the heteroskedasticity. Using the conventional FGLS standard errors, test the null hypothesis that $\beta_2 \leq 0$ against $\beta_2 > 0$ using the 5% level of significance.

h. Repeat part (g) but using FGLS with robust standard errors.

i. Summarize your findings about the elasticity of telephone services expenditure with respect to total expenditure.

8.20 The data file *br2* contains data on 1080 houses sold in Baton Rouge, Louisiana, during mid-2005. We will be concerned with the selling price (*PRICE*), the size of the house in square feet (*SQFT*), the age of the house in years (*AGE*), whether the house is on a waterfront (*WATERFRONT* = 1, 0), and if it is of a traditional style (*TRADITIONAL* = 1, 0).

a. Find OLS estimates of the following equation and save the residuals.

$$\ln(PRICE) = \beta_1 + \beta_2 \ln(SQFT) + \beta_3 AGE + \beta_4 AGE^2 + \beta_5 WATERFRONT + \beta_6 TRADITIONAL + e$$

At some point, is it possible that an old house will become "historic" with age increasing its value? Construct a 95% interval estimate for the age at which age begins to have a positive effect on price.

b. Use the NR^2 test for heteroskedasticity with the candidate variables AGE, AGE^2, $WATERFRONT$, and $TRADITIONAL$. Repeat the test dropping AGE, but keeping AGE^2. Plot the least residuals against AGE. Is there any visual evidence of heteroskedasticity?

c. Estimate the model in (a) by OLS with White heteroskedasticity robust standard errors. Construct a 95% interval estimate for the age at which age begins to have a positive effect on price. How does the interval compare to the one in (a)?

d. Assume $\sigma_i^2 = \sigma^2 \exp\left(\alpha_2 AGE_i^2 + \alpha_3 WATERFRONT_i + \alpha_4 TRADITIONAL\right)$. Obtain FGLS estimates of the model in (a) and compare the results to those in (c). Construct a 95% interval estimate for the age at which age begins to have a positive effect on price. How does the interval compare to the one in (c)?

e. Obtain the residuals from the transformed model based on the skedastic function in (d). Regress the squares of these residuals on AGE^2, $WATERFRONT$, $TRADITIONAL$, and a constant term. Using the NR^2, is there any evidence of remaining heteroskedasticity in the transformed model? Repeat the test using the transformed model version of the variables and a constant term. How do the results compare?

f. Modify the estimation in (d) to use FGLS with heteroskedasticity robust standard errors. Construct a 95% interval estimate for the age at which age begins to have a positive effect on price. How does the interval compare to the ones in (c) and (d)?

g. What do you conclude about the age at which historical value increases a house price?

8.21 In Example 8.9 we estimated the linear probability model

$$COKE = \beta_1 + \beta_2 PRATIO + \beta_3 DISP\_COKE + \beta_4 DISP\_PEPSI + e$$

where $COKE = 1$ if a shopper purchased Coke and $COKE = 0$ if a shopper purchased Pepsi. The variable $PRATIO$ was the relative price ratio of Coke to Pepsi and $DISP\_COKE$ and $DISP\_PEPSI$ were indicator variables equal to one if the relevant display was present. Suppose now that we have 1140 observations on randomly selected shoppers from 50 different grocery stores. Each grocery store has its own settings for $PRATIO$, $DISP\_COKE$ and $DISP\_PEPSI$. Let an (i, j) subscript denote the jth shopper at the ith store, so that we can write the model as

$$COKE_{ij} = \beta_1 + \beta_2 PRATIO_i + \beta_3 DISP\_COKE_i + \beta_4 DISP\_PEPSI_i + e_{ij}$$

Average this equation over all shoppers in the ith store so that we have

$$\overline{COKE}_i\cdot = \beta_1 + \beta_2 PRATIO_i + \beta_3 DISP\_COKE_i + \beta_4 DISP\_PEPSI_i + \bar{e}_i\cdot \qquad \text{(XR8.21)}$$

where

$$\bar{e}_i\cdot = \frac{1}{N_i}\sum_{j=1}^{N_i} e_{ij} \qquad \text{and} \qquad \overline{COKE}_i\cdot = \frac{1}{N_i}\sum_{j=1}^{N_i} COKE_{ij}$$

and N_i is the number of sampled shoppers in the ith store.

a. What is the interpretation of $\overline{COKE}_i\cdot$ for the ith store?

b. Assume that $E\left(COKE_{ij}|\mathbf{x}_{ij}\right) = P_i$ and $\text{var}\left(COKE_{ij}|\mathbf{x}_{ij}\right) = P_i\left(1 - P_i\right)$, show that $E\left(\overline{COKE}_i\cdot|\mathbf{X}\right) = P_i$ and $\text{var}\left(\overline{COKE}_i\cdot|\mathbf{X}\right) = P_i\left(1 - P_i\right)/N_i$.

c. Interpret P_i and express it in terms of $PRATIO_i$, $DISP\_COKE_i$, and $DISP\_PEPSI_i$.

d. Observations on the variables $\overline{COKE}_i\cdot$, $PRATIO_i$, $DISP\_COKE_i$, $DISP\_PEPSI_i$, and N_i appear in the data file *coke_grouped*. Obtain summary statistics for the data. Calculate the sample coefficient of variation, $\text{CV} = 100 s_x/\bar{x}$, for $\overline{COKE}_i\cdot$ and $PRATIO_i$. How much variation is there in these variables relative to their mean? Would we prefer larger or smaller coefficients of variation in these variables? Why? Construct histograms for $\overline{COKE}_i\cdot$ and $PRATIO_i$. What do you observe?

e. Find least squares estimates of equation (XR8.21) and use robust standard errors. Summarize the results. Test the null hypothesis $\beta_3 = -\beta_4$. Choose an appropriate alternative hypothesis and use the 5% level of significance. If the null hypothesis is true, what does it imply about the effect of store displays for *COKE* and *PEPSI*?

f. Create the variable $DISP = DISP\_COKE - DISP\_PEPSI$. Estimate the model $\overline{COKE}_i. = \beta_1 + \beta_2 PRATIO_i + \beta_3 DISP_i + \bar{e}_i.$ by OLS. Test for heteroskedasticity by applying the White test. Also carry out the NR^2 test for heteroskedasticity using the candidate variable N_i. What are your conclusions, at the 5% level?

g. Obtain the fitted values from (e), p_i, and estimate $\text{var}\left(\overline{COKE}_i.\right)$ for each of the stores. Report the mean, standard deviation, maximum and minimum values of the p_i.

h. Find generalized least squares estimates of the model in part (f). Comment on the results and compare them with those obtained in part (f). How might the results of part (d) help you?

8.22 Use data file *cps5* for this exercise.

a. Estimate the following wage equation by OLS and use heteroskedasticity robust standard errors:

$$\begin{aligned}\ln(WAGE) = {} & \beta_1 + \beta_2 EDUC + \beta_3 EXPER + \beta_4 EXPER^2 + \beta_5 (EXPER \times EDUC) \\ & + \beta_6 FEMALE + \beta_7 BLACK + \delta_1 SOUTH + \delta_2 MIDWEST + \delta_3 WEST + e\end{aligned} \tag{XR8.22}$$

Discuss the results.

b. Add *MARRIED* to the equation and reestimate. Holding education and experience constant, do white male married workers in the northeast get higher wages? Using a 5% significance level, test a null hypothesis that wages of married workers are less than or equal to those of unmarried workers against the alternative that wages of married workers are higher.

c. Examine the residuals from part (a) for the two values of *MARRIED*. Is there evidence of heteroskedasticity?

d. Estimate the model in part (a) twice—once using observations on only married workers and once using observations on only unmarried workers. Use the Goldfeld–Quandt test and a 5% significance level to test whether the error variances for married and unmarried workers are different.

e. Hypothesize that $\sigma_i^2 = \sigma^2 \exp(\alpha_2 MARRIED)$. Find generalized least squares of the model in part (a). Compare the estimates and standard errors with those obtained in part (a).

f. Find two 95% interval estimates for the marginal effect $\partial E(\ln(WAGE)) / \partial EDUC$ for a white male worker living in the northeast with 16 years of education and 10 years of experience. Use the results from part (a) for one interval and the results from part (e) for the other interval. Comment on any differences.

8.23 Using the data in *cps5* obtain OLS estimates of the wage equation

$$\begin{aligned}\ln(WAGE) = {} & \beta_1 + \beta_2 EDUC + \beta_3 EXPER + \beta_4 EXPER^2 + \beta_5 (EXPER \times EDUC) \\ & + \beta_6 FEMALE + \beta_7 BLACK + \beta_8 UNION + \beta_9 METRO \\ & + \delta_1 SOUTH + \delta_2 MIDWEST + \delta_3 WEST + e\end{aligned} \tag{XR8.23}$$

a. Interpret the coefficient of *UNION*. Test the null hypothesis that the coefficient of *UNION* is less than or equal to zero, against the alternative that is positive. What do you conclude?

b. Test for the presence of heteroskedasticity related to the variables *UNION* and *METRO* using the NR^2 test. What do you conclude at the 1% level of significance?

c. Regress the squared least squares residuals, $\hat{e}_i^2$, from (a) on *EDUC*, *UNION*, and *METRO*. Also regress $\ln\left(\hat{e}_i^2\right)$ on *EDUC*, *UNION*, and *METRO*. What do these results suggest about the effect of *UNION* membership on the variation in the random error? What do these results suggest about the effect of *METRO* on the variation in the random error?

d. Hypothesize that $\sigma_i^2 = \sigma^2 \exp(\alpha_2 EDUC + \alpha_3 UNION + \alpha_4 METRO)$. Find generalized least squares estimates of the wage equation. For the coefficient of *UNION*, compare the estimates and standard errors with those obtained from OLS estimation of (XR8.23) with heteroskedasticity robust standard errors.

8.24 In this exercise, we will explore some of the factors predicting costs at American universities using the data file *poolcoll2*. Let *TC* = the real (2008 dollars) total cost per student, *FTUG* = number of full-time undergraduate students, *FTGRAD* = number of full-time graduate students, *FTEF* = full-time faculty per 100 students, *CF* = number of contract faculty per 100 students, *FTENAP* = full

time nonacademic professionals per 100 students, *PRIVATE* = 1 if the school is private, and 0 if it is public.

a. Estimate the regression of ln(*TC*) on the remaining variables. What are the predicted effects of additional graduate students on total cost per student? What are the predicted effects of additional full-time faculty?

b. Include in the model not only *PRIVATE* but also *PRIVATE* × *FTEF*. Are these variables individually and jointly significant at the 5% level?

c. Use the NR^2 test for heteroskedasticity that is possibly related to *PRIVATE*. What do you conclude at the 1% level of significance?

d. Test the hypothesis in (b) using OLS estimates with robust standard errors.

e. Include in the model not only *PRIVATE* but also *PRIVATE* times all the other variables. Test the joint significance of *PRIVATE* and *PRIVATE* times all the other variables using an *F*-test. Use robust standard errors and carry out a robust *F*-test. Can we say "We reject the hypothesis that the models determining total cost per student are the same for public and private universities?"

f. Hypothesize $\sigma_i^2 = \exp(\alpha_1 + \alpha_2 PRIVATE)$. Obtain FGLS estimates of the model in (e) and carry out the *F*-test on *PRIVATE* and *PRIVATE* times all the other variables. What is the value of the *F*-test statistic? What is the 1% critical value?

8.25 What effect does having public health insurance have on the number of doctor visits a person has during a year? Using 1988 data, in the data file *rwm88_small*, from Germany, we will explore this question. The data file *rwm88* contains more observations.

a. Estimate the regression model with the dependent variable *DOCVIS* and the explanatory variables *PUBLIC*, *FEMALE*, *HHKIDS*, *MARRIED*, *SELF*, *EDUC2*, *HHNINC2*. Test the null hypothesis that the coefficient on *PUBLIC* is less than or equal to zero, versus the alternative that it is greater than zero at the 1% level of significance.

b. Test for the presence of heteroskedasticity. Obtain the squared least squares residuals from the regression in (a), regress them on all the explanatory variables, and carry out an *F*-test of their joint significance. What do we conclude about the presence of heteroskedasticity at the 1% level of significance?

c. Estimate the regression model with the dependent variable *DOCVIS* and the explanatory variables *FEMALE*, *HHKIDS*, *MARRIED*, *SELF*, *EDUC2*, *HHNINC2* separately for those with public insurance and those who do not have public insurance. Use equation (7.37) to obtain the estimate of the average treatment effect of public insurance.

d. Estimate the regression model with the dependent variable *DOCVIS* and the explanatory variables *PUBLIC*, *FEMALE*, *HHKIDS*, *MARRIED*, *SELF*, *EDUC2*, *HHNINC2* in "deviation from the mean" form. That is, for each variable x, create the variable $\tilde{x} = x - \bar{x}$, where $\bar{x}$ is the sample mean. Using robust standard errors, test the significance of the coefficient on *PUBLIC*.

e. Estimate the regression model with the dependent variable *DOCVIS* and the explanatory variables *FEMALE*, *HHKIDS*, *MARRIED*, *SELF*, *EDUC2*, *HHNINC2*, along with *PUBLIC* and *PUBLIC* times each of the variables in deviation about the mean form. What is the estimated average treatment effect? Using a robust standard error, is it statistically significant at the 5% level? [*Hint*: See equation (7.41) and the surrounding discussion.]

8.26 In the STAR experiment, Example 7.8, children were randomly assigned within schools into three types of classes: small classes with 13–17 students, regular-sized classes with 22–25 students, and regular–sized classes with a full-time teacher aide to assist the teacher. Student scores on achievement tests were recorded as well as some information about the students, teachers, and schools. Data for the kindergarten classes is contained in the data file *star5_small2*.

a. Regress *MATHSCORE* on *SMALL*, *AIDE*, *TCHEXPER*, *SCHRURAL*, *FREELUNCH*, and *BOY*. Test for heteroskedasticity related to *SMALL* and *AIDE* using the NR^2 test. What do you conclude at the 5% level?

b. Estimate the regression model in (a) by OLS including interactions between *FREELUNCH* and the other variables. Test for heteroskedasticity related to *SMALL* and *AIDE* using the NR^2 test. What do you conclude at the 5% level?

c. Using the model in (b), and both conventional and robust standard errors, test the joint significance of the interactions between *FREELUNCH* and *SMALL*, *AIDE*, and *TCHEXPER* at the 10% level in each regression. What do you conclude?

d. Estimate the model in (b) and include indicator variables for each school (*SCHOOLID*). Test for heteroskedasticity related to *SMALL* and *AIDE* using the NR^2 test. What do you conclude at the 5% level?

e. Using the model in (d), and both conventional and robust standard errors, test the joint significance of the interactions between *FREELUNCH* and *SMALL*, *AIDE*, and *TCHEXPER* at the 10% level in each regression. What do you conclude?

8.27 There were 64 countries who competed in the 1992 Olympics and won at least one medal. For each of these countries, let *MEDALTOT* be the total number of medals won, *POP* be population in millions, and *GDP* be GDP in billions of 1995 dollars.

a. Use the data file *olympics5*, excluding the United Kingdom, and use the $N = 63$ remaining observations. Estimate the model $MEDALTOT = \beta_1 + \beta_2 \ln(POP) + \beta_3 \ln(GDP) + e$ by OLS.

b. Calculate the squared least squares residuals $\hat{e}_i^2$ from the regression in (a). Regress $\hat{e}_i^2$ on $\ln(POP)$ and $\ln(GDP)$. Use the F-test from this regression to test for heteroskedasticity at the 5% level of significance. Use the R^2 from this regression to test for heteroskedasticity. What are the p-values of the two tests?

c. Reestimate the model in (a) but using heteroskedasticity robust standard errors. Using a 10% significance level, test the hypothesis that there is no relationship between the number of medals won and GDP against the alternative that there is a positive relationship. What happens if you change the significance level to 5%?

d. Using a 10% significance level, test the hypothesis that there is no relationship between the number of medals won and population against the alternative that there is a positive relationship. What happens if you change the significance level to 5%?

e. Use the model in (c) to find point and 95% interval estimates for the expected number of medals won by the United Kingdom whose population and GDP in 1992 were 58 million and $1010 billion, respectively.

f. The United Kingdom won 20 medals in 1992. Was the model successful in predicting the mean number of medals for the United Kingdom? Using the estimation in (c), with robust standard errors, what is the p-value for a test of $H_0 : \beta_1 + \ln(58) \times \beta_2 + \ln(1010) \times \beta_3 = 20$ versus $H_1 : \beta_1 + \ln(58) \times \beta_2 + \ln(1010) \times \beta_3 \neq 20$?

8.28 In this exercise you will create some simulated data and try out estimation and testing methods. Use your software to create a new data set, or "workfile," with $N = 100$ observations. All modern software has functions, called random number generators, to create uniformly distributed and normally distributed random values. Follow these steps.

1. Create $X2 = 1 + 5 \times U1$, where $U1$ is a random number between zero and one.
2. Create $X3 = 1 + 5 \times U2$, where $U2$ is another random number between zero and one.
3. Create $E = \sqrt{\exp(2 + 0.6X2)} \times Z$, where $Z \sim N(0, 1)$.
4. Create $Y = 5 + 4X2 + E$

You should now have 100 values for Y, $X2$, and $X3$. Note: Your results should be different from your classmates, and your results might change from one experiment to the next. To prevent this from happening, you can set the random number's "seed." See your software documentation for instructions.

a. Regress Y on $X2$ and $X3$ and obtain conventional OLS standard errors. Compare the estimated coefficients to the true values of the regression parameters, $\beta_1 = 5$, $\beta_2 = 4$, $\beta_3 = 0$. Do the t-values suggest that the coefficients are significantly different from 0 at the 5% level?

b. Calculate the least squares residuals $\hat{e}$ from the OLS estimation in (a) and regress $\hat{e}^2$ on $X2$ and $X3$. What evidence, if any, do you find for the presence of heteroskedasticity?

c. Regress Y on $X2$ and $X3$ and obtain robust standard errors. Compare these to the conventional standard errors in (a).

d. Assume the heteroskedasticity pattern is $\sigma^2 X2^2$. Obtain GLS estimates with conventional and robust standard errors. Are the GLS parameter estimates closer to the true parameter values or not? Which set of standard errors should be used?

e. Assume the multiplicative heteroskedasticity model $\exp(\alpha_1 + \alpha_2 X2 + \alpha_3 X3)$. Obtain FGLS estimates with conventional and robust standard errors. Are the FGLS estimates closer to the true parameter values than the GLS or OLS estimates? Which set of standard errors should be used?

8.29 The data file *mexican* contains data collected in 2001 from the transactions of 754 Mexican sex workers.

- **a.** Using OLS, estimate the hedonic log-linear model with *LNPRICE* as the dependent variable and independent variables *BAR*, *STREET*, *SCHOOL*, *AGE*, *RICH*, *ALCOHOL*, *ATTRACTIVE*. Interpret the estimated coefficients.
- **b.** Test for heteroskedasticity related to *ATTRACTIVE* using the NR^2 test at the 1% level of significance.
- **c.** Estimate the model separately by OLS for observations with *ATTRACTIVE* = 1 and *ATTRACTIVE* = 0. Using the results, carry out the Goldfeld–Quandt test for heteroskedasticity across the two regressions. Use a two-tailed test at the 5% level. Which regression has a larger estimated error variance?
- **d.** Compare the estimates from the two estimations in (c). Do they appear similar or dissimilar? Which coefficients are noticeably different? Use OLS to estimate the model that includes the original variables and interactions between *ATTRACTIVE* and the other explanatory variables. Test the joint significance of *ATTRACTIVE* and the interaction variables at the 1% level of significance. Is this a "valid" Chow test? Is homoskedasticity a necessary condition for this test? Recall that the test is described in Section 7.2.3.
- **e.** Using the estimation results in (d), test for heteroskedasticity related to *ATTRACTIVE* using the NR^2 test at the 1% level of significance.
- **f.** Use OLS with heteroskedasticity robust standard errors to estimate the model that includes the original variables and interactions between *ATTRACTIVE* and the other explanatory variables. Test the joint significance of *ATTRACTIVE* and the interaction variables at the 1% level of significance. Is this a "valid" Chow test?

8.30 The data file *grunfeld2* contains annual data on the gross investment, capital stock, and the value of the firm, measured by the value of common and preferred stock for General Electric and Westinghouse, during the period 1935–1954. These data have been used to train econometricians for almost 60 years, and still provide valuable lessons.

- **a.** Create an indicator variable $GE = 1$ for General Electric and $GE = 0$ for Westinghouse. Using the combined data on both firms, use OLS to estimate the model of investment, *INV*, as a function of the value of the firms, *V*, and capital stock, *K*, also the indicator variable *GE* and the interactions of *GE* with *V* and *K*. That is $INV = f(const, V, K, GE, GE \times V, GE \times K)$. Test the joint significance of the variables $GE, GE \times V, GE \times K$ at the 5% level. What does this test reveal about the two firms' investment characteristics?
- **b.** Obtain the OLS residuals from (a) and regress their squares on the indicator variable *GE*. Use the result of this regression to test for heteroskedasticity across the firms at the 1% level.
- **c.** Reestimate the model in (a) using OLS with heteroskedasticity robust standard errors. Test the joint significance of the variables $GE, GE \times V, GE \times K$ at the 5% level. Does your conclusion change?
- **d.** Estimate the investment model separately for General Electric and Westinghouse. Let the estimated error variances be $\hat{\sigma}^2_{GE}$ and $\hat{\sigma}^2_{WE}$. For which firm is the estimated error variance smaller?
- **e.** Create a variable *W* that takes the value $\hat{\sigma}^2_{GE}$ when $GE = 1$ and takes the value $\hat{\sigma}^2_{WE}$ when $GE = 0$. Estimate the model in (a) by FGLS with weighting variable *W*. Test the joint significance of the variables $GE, GE \times V, GE \times K$ at the 5% level. Does your conclusion change?

Appendix 8A Properties of the Least Squares Estimator

In Appendix 2D, we wrote the least squares estimator for β_2 in the simple regression model as $b_2 = \beta_2 + \sum w_i e_i$, where

$$w_i = \frac{x_i - \bar{x}}{\sum (x_i - \bar{x})^2}$$

This expression is a useful one for exploring the properties of the least squares estimator under heteroskedasticity. The first property that we establish is that of unbiasedness. This property was

derived under homoskedasticity in equation (2.13) of Chapter 2. The same proof holds under heteroskedasticity because the only error term assumption that was used is $E(e_i|\mathbf{x}) = 0$.

$$\begin{aligned} E(b_2|\mathbf{x}) &= E(\beta_2 + \sum w_i e_i|\mathbf{x}) = E(\beta_2 + w_1 e_1 + w_2 e_2 + \cdots + w_N e_N|\mathbf{x}) \\ &= E(\beta_2) + E(w_1 e_1|\mathbf{x}) + E(w_2 e_2|\mathbf{x}) + \cdots + E(w_N e_N|\mathbf{x}) \\ &= \beta_2 + \sum E(w_i e_i|\mathbf{x}) = \beta_2 + \sum w_i E(e_i|\mathbf{x}) = \beta_2 \end{aligned}$$

The least squares estimators are unbiased as long as $E(e_i|\mathbf{x}) = 0$, even if the errors are heteroskedastic. This is true in both the simple and multiple regression models.

The variance of the least squares estimator is

$$\begin{aligned} \text{var}(b_2|\mathbf{x}) &= \text{var}(\sum w_i e_i|\mathbf{x}) \\ &= \sum w_i^2 \text{var}(e_i|\mathbf{x}) + \sum_{i \neq j} \sum w_i w_j \text{cov}(e_i, e_j|\mathbf{x}) \\ &= \sum w_i^2 \sigma_i^2 \\ &= \sum \left\{ \frac{(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \right\}^2 \sigma_i^2 = \sum \left\{ \frac{(x_i - \bar{x})^2}{\left[\sum (x_i - \bar{x})^2\right]^2} \sigma_i^2 \right\} \\ &= \left[\sum (x_i - \bar{x})^2\right]^{-1} \sum \left[(x_i - \bar{x})^2 \sigma_i^2\right] \left[\sum (x_i - \bar{x})^2\right]^{-1} \end{aligned} \tag{8A.1}$$

Going from the second line to the third we used assumption MR4, conditionally uncorrelated errors, $\text{cov}(e_i, e_j|\mathbf{x}) = 0$. If the variances are all the same $(\sigma_i^2 = \sigma^2)$, then the third line becomes $\sigma^2 \sum w_i^2 = \text{var}(b_2|\mathbf{x}) = \sigma^2 / \sum (x_i - \bar{x})^2$, which is the usual OLS variance expression. This simplification is not possible under heteroskedasticity. The fourth and fifth lines are equivalent ways of writing the variance of the least squares estimator, equation (8.8), when the random errors are heteroskedastic.

Appendix 8B Lagrange Multiplier Tests for Heteroskedasticity

More insights into LM and other variance function tests can be developed by relating them to the F-test introduced in (6.8) for testing the significance of a mean function. To put that test in the context of a variance function, consider (8.15)

$$\hat{e}_i^2 = \alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_S z_{iS} + v_i \tag{8B.1}$$

and assume that our objective is to test $H_0: \alpha_2 = \alpha_3 = \cdots = \alpha_S = 0$ against the alternative that at least one α_s, for $s = 2, \ldots, S$, is nonzero. In Section 8.2.2 we considered a more general variance function than that in (8B.1), but we also pointed out that using the linear function in (8B.1) is valid for testing more general alternative hypotheses.

Adapting the F-value reported in (6.8) to test the overall significance of (8B.1), we have

$$F = \frac{(SST - SSE)/(S-1)}{SSE/(N-S)} \tag{8B.2}$$

where

$$SST = \sum_{i=1}^{N} \left[\hat{e}_i^2 - \overline{\hat{e}^2}\right]^2 \quad \text{and} \quad SSE = \sum_{i=1}^{N} \hat{v}_i^2$$

are the total sum of squares and sum of squared errors from estimating (8B.1). Note that $\overline{\hat{e}^2}$ is the mean of the dependent variable in (8B.1), or, equivalently, the average of the squares of the least squares residuals from the regression function. At a 5% significance level, a valid test is to reject H_0 if the F-value is greater than a critical value given by $F_{(0.95,\, S-1,\, N-S)}$.

Two further tests, the original Breusch–Pagan test and its $N \times R^2$ version, can be obtained by modifying (8B.2). Please be patient as we work through these modifications. We begin by rewriting (8B.2) as

$$\chi^2 = (S-1) \times F = \frac{SST - SSE}{SSE/(N-S)} \sim \chi^2_{(S-1)} \tag{8B.3}$$

The chi-square statistic $\chi^2 = (S-1) \times F$ has an approximate $\chi^2_{(S-1)}$ -distribution in large samples. That is, multiplying an F-statistic by its numerator degrees of freedom gives another statistic that follows a chi-square distribution. The degrees of freedom of the chi-square distribution are $S-1$, the same as that for the numerator of the F-distribution. The background for this result is given in Appendix 6A.

Next, note that

$$\widehat{\text{var}}\left(e_i^2\right) = \widehat{\text{var}}\left(v_i\right) = \frac{SSE}{N-S} \tag{8B.4}$$

That is, the variance of the dependent variable is the same as the variance of the error, which can be estimated from the sum of squared errors in (8B.1). Substituting (8B.4) into (8B.3) yields

$$\chi^2 = \frac{SST - SSE}{\widehat{\text{var}}\left(e_i^2\right)} \tag{8B.5}$$

This test statistic represents the basic form of the Breusch–Pagan statistic. Its two different versions occur because of the alternative estimators used to replace $\widehat{\text{var}}\left(e_i^2\right)$.

If it is assumed that e_i is normally distributed, it can be shown that $\text{var}\left(e_i^2\right) = 2\sigma_e^4$, and the statistic for the first version of the Breusch–Pagan test is

$$\chi^2 = \frac{SST - SSE}{2\hat{\sigma}_e^4} \tag{8B.6}$$

Note that $\sigma_e^4 = \left(\sigma_e^2\right)^2$ is the square of the error variance from the mean function; unlike *SST* and *SSE*, its estimate comes from estimating (8.16). The result $\text{var}\left(e_i^2\right) = 2\sigma_e^4$ might be unexpected—here is a little proof so that you know where it comes from. When $e_i \sim N\left(0, \sigma_e^2\right)$, then $\left(e_i/\sigma_e\right) \sim N(0,1)$, and $\left(e_i^2/\sigma_e^2\right) \sim \chi^2_{(1)}$. The variance of a $\chi^2_{(1)}$ random variable is 2. Thus,

$$\text{var}\left(\frac{e_i^2}{\sigma_e^2}\right) = 2 \Rightarrow \quad \frac{1}{\sigma_e^4}\text{var}\left(e_i^2\right) = 2 \Rightarrow \text{var}\left(e_i^2\right) = 2\sigma_e^4$$

Using (8B.6), we reject a null hypothesis of homoskedasticity when the χ^2-value is greater than a critical value from the $\chi^2_{(S-1)}$ distribution.

For the second version of (8B.5) the assumption of normally distributed errors is not necessary. Because this assumption is not used, it is often called the robust version of the Breusch–Pagan test. The sample variance of the squared least squares residuals, the $\hat{e}_i^2$, is used as an estimator for $\text{var}\left(e_i^2\right)$. Specifically, we set

$$\widehat{\text{var}}\left(e_i^2\right) = \frac{1}{N}\sum_{i=1}^{N}\left[\hat{e}_i^2 - \overline{\hat{e}^2}\right]^2 = \frac{SST}{N} \tag{8B.7}$$

This quantity is an estimator for $\text{var}\left(e_i^2\right)$ under the assumption that H_0 is true. It can also be written as the total sum of squares from estimating the variance function divided by the sample size.

Substituting (8B.7) into (8B.5) yields

$$\begin{aligned}\chi^2 &= \frac{SST - SSE}{SST/N} \\ &= N \times \left(1 - \frac{SSE}{SST}\right) \\ &= N \times R^2\end{aligned} \tag{8B.8}$$

where R^2 is the R^2 goodness-of-fit statistic from estimating the variance function. At a 5% significance level, a null hypothesis of homoskedasticity is rejected when $\chi^2 = N \times R^2$ exceeds the critical value $\chi^2_{(0.95,S-1)}$.

Software often reports the outcome of the White test described in Section 8.6.4 a as an F-value or a χ^2-value. The F-value is from the statistic in (8B.4), with the z's chosen as the x's and their squares and possibly cross-products. The χ^2-value is from the statistic in (8B.8), with the z's chosen as the x's and their squares and possibly cross-products.

Appendix 8C Properties of the Least Squares Residuals

The least squares residuals are $\hat{e}_i = y_i - \hat{y}_i$. Substituting in the fitted value $\hat{y}_i = b_1 + b_2 x_i$ we obtain for the simple regression model

$$\begin{aligned}\hat{e}_i &= y_i - \hat{y}_i = \beta_1 + \beta_2 x_i + e_i - (b_1 + b_2 x_i) \\ &= (\beta_1 - b_1) + (\beta_2 - b_2) x_i + e_i \\ &= e_i - (b_1 - \beta_1) - (b_2 - \beta_2) x_i\end{aligned}$$

Using the last line we find

$$E(\hat{e}_i|\mathbf{x}) = E(e_i|\mathbf{x}) - E(b_1 - \beta_1|\mathbf{x}) - E(b_2 - \beta_2|\mathbf{x}) x_i = 0$$

The expected value of the least squares residual is zero under assumptions SR1–SR5. Also, note what happens if we consider large samples, with $N \to \infty$. The least squares estimators b_1 and b_2 are unbiased, and recall from Section 2.4.4 that their variances get smaller and smaller as N gets larger. This means that in large samples $(b_1 - \beta_1)$ and $(b_2 - \beta_2)$ are close to zero, so that in large samples the difference $\hat{e}_i - e_i$ is close to zero. In econometric terms, the *probability limit* of $\hat{e}_i - e_i$ is zero, that is, $\text{plim}(\hat{e}_i - e_i) = 0$. The two random variables become essentially the same and thus have the same probability distribution. This means, that in large samples, if $e_i \sim N(0, \sigma^2)$ then $\hat{e}_i \overset{a}{\sim} N(0, \sigma^2)$, where "$\overset{a}{\sim}$" means **approximately distributed**, or **asymptotically** (in large samples) **distributed**. Learning asymptotic analysis is an important feature of econometrics. See Section 5.7 for further discussion.

It can be shown that the conditional variance of the least squares residual is

$$\text{var}(\hat{e}_i|\mathbf{x}) = E\left(\hat{e}_i^2|\mathbf{x}\right) = \sigma^2 \left\{1 - \frac{1}{N} - \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right\} = \sigma^2 (1 - h_i) \tag{8C.1}$$

where h_i is the **leverage** of the ith observation, a term we introduced in Section 4.3.6. Note that:

- **i.** The conditional variance of the least squares residual is not constant even if the random error is homoskedastic.
- **ii.** Because $0 \le h_i \le 1$ and $0 \le (1 - h_i) \le 1$, $\text{var}(\hat{e}_i|\mathbf{x}) < \text{var}(e_i|\mathbf{x}) = \sigma^2$. The variation in the least squares residual is less than the variance of the true random error.
- **iii.** The variance of the least squares residual is closest to $\text{var}(e_i|\mathbf{x}) = \sigma^2$ when $x_i = \bar{x}$, reflecting the fact that the fitted value $\hat{y}_i$ has the least prediction error at that point.
- **iv.** The expression (8C.1) is valid in both simple and multiple regression, with h_i redefined in multiple regression.

v. The sum of the leverage values is K, $\sum h_i = K$. As a check, verify that for the simple regression model $\sum h_i = 2$.

vi. $\sum_{i=1}^{N} \text{var}(\hat{e}_i|\mathbf{x}) = \sum_{i=1}^{N} E\left(\hat{e}_i^2|\mathbf{x}\right) = \sigma^2(N-K)$ while $\sum_{i=1}^{N} \text{var}(e_i|\mathbf{x}) = \sum_{i=1}^{N} E(e_i^2|\mathbf{x}) = N\sigma^2$

8C.1 Details of Multiplicative Heteroskedasticity Model

We showed that the least squares residuals and the true random error have the same probability distribution in large samples. If $e_i \sim N(0, \sigma_i^2)$ then in large samples the least squares residual $\hat{e}_i \overset{a}{\sim} N(0, \sigma_i^2)$. In large samples, then $(\hat{e}_i/\sigma_i) \overset{a}{\sim} N(0,1)$ and $(\hat{e}_i/\sigma_i)^2 \overset{a}{\sim} [N(0,1)]^2 \sim \chi^2_{(1)}$. Thus,

$$\ln\left[(\hat{e}_i/\sigma_i)^2\right] = v_i \overset{a}{\sim} \ln\left[\chi^2_{(1)}\right]$$

Statisticians have studied this random variable and found that $E\left\{\ln\left[\chi^2_{(1)}\right]\right\} = -1.2704$ and $\text{var}\left\{\ln\left[\chi^2_{(1)}\right]\right\} = 4.9348$.

Appendix 8D Alternative Robust Sandwich Estimators

The robust variance estimators carry over to the multiple regression model $y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + e_i$ quite easily. Recall from Appendix 6B that we can express the least squares estimator b_2 as

$$b_2 = \frac{\sum(x_{i2} - \tilde{x}_{i2})\, y_i}{\sum(x_{i2} - \tilde{x}_{i2})^2}$$

where $\tilde{x}_{i2}$ is the fitted value from the auxiliary regression of x_2 on all the other explanatory variables, $x_{i2} = c_1 + c_3 x_{i3} + \cdots + c_K x_{iK} + r_{i2}$. Substituting for y_i and simplifying leads us to

$$b_2 = \beta_2 + \frac{\sum(x_{i2} - \tilde{x}_{i2})\, e_i}{\sum(x_{i2} - \tilde{x}_{i2})^2}$$

If the errors are heteroskedastic and serially uncorrelated, then the conditional variance of b_2 is

$$\begin{aligned} \text{var}(b_2|\mathbf{X}) &= \text{var}\left[\left.\frac{\sum(x_{i2} - \tilde{x}_{i2})\, e_i}{\sum(x_{i2} - \tilde{x}_{i2})^2}\right| \mathbf{X}\right] = \frac{\sum(x_{i2} - \tilde{x}_{i2})^2 \text{var}(e_i|\mathbf{X})}{\left[\sum(x_{i2} - \tilde{x}_{i2})^2\right]^2} \\ &= \frac{\sum(x_{i2} - \tilde{x}_{i2})^2 \sigma_i^2}{\left[\sum(x_{i2} - \tilde{x}_{i2})^2\right]^2} \\ &= \left[\sum(x_{i2} - \tilde{x}_{i2})^2\right]^{-1} \left\{\sum(x_{i2} - \tilde{x}_{i2})^2 \sigma_i^2\right\} \left[\sum(x_{i2} - \tilde{x}_{i2})^2\right]^{-1} \end{aligned} \tag{8D.1}$$

The **original** White heteroskedasticity corrected variance estimator replaces σ_i^2 by the squared OLS residuals

$$\widehat{\text{var}}(b_2) = \left[\sum(x_{i2} - \tilde{x}_{i2})^2\right]^{-1} \left\{\sum(x_{i2} - \tilde{x}_{i2})^2 \hat{e}_i^2\right\} \left[\sum(x_{i2} - \tilde{x}_{i2})^2\right]^{-1} = \text{HCE0} \tag{8D.2}$$

The version in equation (8D.2) is valid in large samples. In practice, some alternatives are used that are designed to work better in smaller samples. These alternatives account for the fact that the least squares residuals are on average a little smaller than the true random errors. As noted in

Appendix 8C, in the simple regression model, with assumptions SR1–SR5 holding, the variance of the least squares residual is

$$\text{var}(\hat{e}_i|\mathbf{X}) = E\left(\hat{e}_i^2|\mathbf{X}\right) = \sigma^2\left\{1 - \frac{1}{N} - \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right\} = \sigma^2(1 - h_i) \tag{8D.3}$$

where h_i is the **leverage** of the ith observation, a term we introduced in Section 4.3.6. In the simple regression model

$$h_i = \frac{1}{N} + \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}$$

The expression

$$\text{var}(\hat{e}_i|\mathbf{X}) = \sigma^2(1 - h_i) \tag{8D.4}$$

is valid in both simple and multiple regression, with h_i redefined when $K > 2$. For both simple and multiple regression, $0 \le h_i \le 1$ and $0 \le (1 - h_i) \le 1$.

The first modification of HCE0 is based on the observation that the expected value of the squared least squares residual is smaller than the expected value of the squared random errors.

$$\text{var}(\hat{e}_i|\mathbf{X}) = E\left(\hat{e}_i^2|\mathbf{X}\right) = \sigma^2(1 - h_i) < \text{var}(e_i|\mathbf{X}) = E(e_i^2|\mathbf{X}) = \sigma^2$$

The average value of $E\left(\hat{e}_i^2|\mathbf{X}\right)$ is $[(N - K)/N]\sigma^2$ while the average value of $E(e_i^2|\mathbf{X}) = \sigma^2$. To adjust for the size difference of the least squares residuals, multiply $\hat{e}_i^2$ in HCE0 by $N/(N - K)$. That is,

$$\begin{aligned}\widehat{\text{var}}(b_2) &= \left[\sum(x_{i2} - \tilde{x}_{i2})^2\right]^{-1}\left\{\sum\left[(x_{i2} - \tilde{x}_{i2})^2\left(\frac{N}{N - K}\right)\hat{e}_i^2\right]\right\}\left[\sum(x_{i2} - \tilde{x}_{i2})^2\right]^{-1} \\ &= \text{HCE1}\end{aligned} \tag{8D.5}$$

This correction will have little effect if the sample is large, but it may have an effect when the number of explanatory variables in the model, $K - 1$, is large.

A second modification adjusts the squared least squares residual to have the same conditional expectation as the random error. That is,

$$E\left(\left.\frac{\hat{e}_i^2}{1 - h_i}\right|\mathbf{X}\right) = \sigma^2 = E(e_i^2|\mathbf{X})$$

Then, HCE2 is

$$\begin{aligned}\widehat{\text{var}}(b_2) &= \left[\sum(x_{i2} - \tilde{x}_{i2})^2\right]^{-1}\left\{\sum\left[(x_{i2} - \tilde{x}_{i2})^2\frac{\hat{e}_i^2}{(1 - h_i)}\right]\right\}\left[\sum(x_{i2} - \tilde{x}_{i2})^2\right]^{-1} \\ &= \text{HCE2}\end{aligned} \tag{8D.6}$$

In large samples HCE0, HCE1, and HCE2 are equivalent, but in samples that are not very large, the adjustments make useful differences. In econometric software, the "default" robust variance estimator is HCE0 or HCE1. If the random errors are actually homoskedastic, using HCE2 seems appropriate. Recall that part of the genius of the White heteroskedasticity robust variance estimators is that in large samples they can be applied whether the random errors are heteroskedastic or not. The modification introduced in HCE2 "tweaks" the robust estimator in such a way that it works when the errors are heteroskedastic and a little better than HCE0 and HCE1 when errors are homoskedastic.

Recall that $0 \le (1 - h_i) \le 1$ so HCE2 inflates the least squares residuals and the larger the leverage, h_i, the larger the adjustment becomes. Observations with high leverage, ones that have

a larger impact on regression estimates and predictions, are also the observations for which the least squares residual is much too small, thus the third modification inflates the residual again, using

$$\frac{\hat{e}_i^2/(1-h_i)}{(1-h_i)} = \frac{\hat{e}_i^2}{(1-h_i)^2}$$

Then

$$\widehat{\text{var}}(b_2) = \left[\sum(x_{i2}-\tilde{x}_{i2})^2\right]^{-1}\left\{\sum\left[(x_{i2}-\tilde{x}_{i2})^2\frac{\hat{e}_i^2}{(1-h_i)^2}\right]\right\}\left[\sum(x_{i2}-\tilde{x}_{i2})^2\right]^{-1}$$
$$= \text{HCE3} \tag{8D.7}$$

Some research shows that if heteroskedasticity is present in the data, then HCE3 is a good choice.

To summarize, replacing σ_i^2 in (8D.1) by $\hat{e}_i^2$, $[N/(N-K)]\hat{e}_i^2$, $\hat{e}_i^2/(1-h_i)$, or $\hat{e}_i^2/(1-h_i)^2$ leads to the robust sandwich variance estimators HCE0, HCE1, HCE2, or HCE3. These robust sandwich variance estimators are equivalent in large samples but may yield different results in small samples. "Robust" means that the variance estimates, and standard errors, are valid whether heteroskedasticity is present or not. When a priori reasoning *does not* lead you to suspect heteroskedasticity, but you are suspicious and/or risk averse, and if your sample is not small, then using the robust sandwich variance estimator HCE2 may be the best choice. When a priori reasoning *does* lead you to suspect heteroskedasticity, and if your sample is not small, then using the robust sandwich variance estimator HCE3 may be the better choice. Because the calculations are complex, it is best to use proper econometric software for robust variances.

EXAMPLE 8.10 | Alternative Robust Standard Errors in the Food Expenditure Model

Most regression packages include an option for calculating standard errors using White's estimator. If we do so for the food expenditure example, we obtain

$$\widehat{FOOD\_EXP} = 83.42 + 10.21\,INCOME$$
$$(27.46)\quad (1.81)\quad \text{(White robust se-HCE1)}$$
$$(27.69)\quad (1.82)\quad \text{(White robust se-HCE2)}$$
$$(28.65)\quad (1.89)\quad \text{(White robust se-HCE3)}$$
$$(43.41)\quad (2.09)\quad \text{(incorrect OLS se)}$$

In this case, ignoring heteroskedasticity and using incorrect standard errors, based on the usual formula in (8.6), tends to understate the precision of estimation; we tend to get confidence intervals that are wider than they should be. Specifically, following the result in (3.6) in Chapter 3, we can construct four corresponding 95% confidence intervals for β_2.

White HCE1: $b_2 \pm t_c\text{se}(b_2) = 10.21 \pm 2.024 \times 1.81 = [6.55,\ 13.87]$

White HCE2: $b_2 \pm t_c\text{se}(b_2) = 10.21 \pm 2.024 \times 1.82 = [6.52,\ 13.90]$

White HCE3: $b_2 \pm t_c\text{se}(b_2) = 10.21 \pm 2.024 \times 1.89 = [6.39,\ 14.03]$

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

If we ignore heteroskedasticity, we estimate that β_2 lies between 5.97 and 14.45. When we recognize the existence of heteroskedasticity, our information is more precise, and using HCE3 we estimate that β_2 lies between 6.39 and 14.03. Why HCE3? Because a priori we could reason that heteroskedasticity should be present. A caveat here is that the sample is small, which does mean that the robust standard error formulas we have provided may not be as accurate as if the sample were large.

Appendix 8E Monte Carlo Evidence: OLS, GLS, and FGLS

White's estimator for the standard errors helps us avoid computing incorrect interval estimates or incorrect values for test statistics in the presence of heteroskedasticity. The least squares estimator is no longer best, but failing to use the "best" estimator may not be too grave a sin if estimates are sufficiently precise for useful economic analysis. Many cross-sectional data sets have thousands of observations, resulting in robust standard errors that are small, making interval estimates narrow and *t*-tests powerful. Nothing further is required in these cases. If, however, your estimates are not sufficiently precise for economic analysis, then a better, more efficient estimator is called for. In order to use such an estimator, we must specify the **skedastic** function $h(\mathbf{x}_i) > 0$, a function of $\mathbf{x}_i$ and also perhaps other variables, that describe the pattern of conditional heteroskedasticity. In this appendix, we use a Monte Carlo study to illustrate an alternative estimator, feasible generalized least squares, that has a smaller variance than the least squares estimator in large samples.

Using Monte Carlo experiments, we illustrate the properties of the OLS estimator, the correct FGLS estimator and an incorrect GLS estimator. The data generating process[4] is based on the population model

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i = 5 + x_{i2} + 0x_{i3} + e_i$$

The variables x_2 and x_3 are statistically independent uniform (Appendix B.3.4) random variables over the interval (1, 5). They vary randomly with all values being equally likely in the interval. The random error is $e_i = h(\mathbf{x}_i) z_i$, where $z_i \sim N(0, 1)$. The skedasticity function $h(\mathbf{x}_i)$ is

$$h(\mathbf{x}_i) = 3\exp(1 + \alpha_2 x_{i2} + 0x_{i3}) / \overline{h}$$

The value of α_2 changes from $\alpha_2 = 0$, homoskedasticity, to $\alpha_2 = 0.3$, strong heteroskedasticity, to $\alpha_2 = 0.5$, very strong heteroskedasticity. The scalar $\overline{h}$ is a constant such that $\sum_{i=1}^{N} h(\mathbf{x}_i)/N \cong 3$ so that $\sum_{i=1}^{N} \text{var}(e_i|\mathbf{x}_i)/N \cong 9$. We use two sample sizes, $N = 100$, a moderate sample size, and $N = 5000$, a large sample. We use $M = 1000$ Monte Carlo replications and do not hold x_2 and x_3 constant across these experiments.

In Table 8E.1 we report the results of the experiments. The FGLS procedure follows the description in Section 8.5.1, with equation (8.20) being $\ln(\hat{e}_i^2) = \alpha_1 + \alpha_2 x_{i2} + \alpha_3 x_{i3} + v_i$. The GLS estimation incorrectly assumes $\text{var}(e_i|\mathbf{x}_i) = \sigma^2 x_{i2}$. This is the proportional heteroskedasticity assumption illustrated in Section 8.4.1. In the first row of Table 8E.1 is the sample size, N, and in the second row is the value of α_2. First, the results of experiments (1)–(4):

1. Let the OLS estimator of β_2 be b_2. The OLS estimator is unbiased in the presence of heteroskedasticity, which is revealed by the Monte Carlo average across 1000 samples $\overline{b}_2$ in row (3) that is close to the true value $\beta_2 = 1$. The averages of the (correct) FGLS estimates, $\overline{\hat{\beta}}_2$, in row (8) and the (incorrect) GLS estimates, $\overline{\tilde{\beta}}_2$, in row (13) are also close to the true parameter value.
2. The sample standard deviation of the 1000 Monte Carlo OLS estimates is $\text{sd}(b_2)$ in row (4). It measures the actual amount of sampling variation of the OLS estimator—how much it varies from sample to sample due solely to randomness inherent in sampling from a population. Compare to it the sample average of the 1000 Monte Carlo calculated values of the

[4]This design is adapted from James G. MacKinnon (2013) "Thirty Years of Heteroskedasticity-Robust Inference," in *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis: Essays in Honor of Halbert L. White Jr*, editors Xiaohong Chen and R. Norman Swanson, New York: Springer, 437–461.

usual, or nominal, OLS standard error of the estimator b_2, $\overline{\text{se}}(b_2)$ in row (5). Note that when $N = 100$ and $\alpha_2 \neq 0$ the average standard error is less than the standard deviation, meaning that the OLS standard error is too small on average. When $N = 5000$ both of these values are dramatically reduced, but the OLS standard error is still on average too small. Now compare the average of the White robust standard errors HCE1 with the simple inflation factor $N/(N-3)$ as described in Appendix 8C, $\overline{\text{robse}}(b_2)$ in row (6). The average of these standard errors is very close to the actual variation measured by $\text{sd}(b_2)$. That means that the robust standard error correction for the OLS estimator is doing its job, on average, in measuring actual sampling variation.

3. When heteroskedasticity is present, the actual variation in the FGLS estimates, $\text{sd}(\hat{\hat{\beta}}_2)$ in row (9) is less than the actual variation in the OLS estimates, $\text{sd}(b_2)$. The ratio $\text{sd}(\hat{\hat{\beta}}_2)/\text{sd}(b_2)$ in row (10) shows the improvement obtained by using FGLS. By using FGLS, we have obtained estimates that are more precise than the OLS estimates, as we should have. The sample average of the standard error estimates $\overline{\text{se}}(\hat{\hat{\beta}}_2)$, row (11), is slightly smaller than $\text{sd}(\hat{\hat{\beta}}_2)$ when $N = 100$. In this sample size, the FGLS standard errors are a little too small. When $N = 5000$ this is no longer the case. We are reminded that the properties of the FGLS estimator are valid in large samples. We used the correct model for the heteroskedasticity in the FGLS calculations; hence, there is no need to compute FGLS

TABLE 8E.1 **Monte Carlo Simulation Results**

		Experiment				
Result	**Item**	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**
1	N	100	100	100	5000	5000
2	α_2	0	0.3	0.5	0.5	NA
3	$\bar{b}_2$	1.0058	1.0044	1.0033	0.9996	1.0007
4	$\text{sd}(b_2)$	0.2657	0.3032	0.3574	0.0496	0.0414
5	$\overline{\text{se}}(b_2)$	0.2626	0.2831	0.3081	0.0423	0.0406
6	$\overline{\text{robse}}(b_2)$	0.2614	0.3035	0.3586	0.0498	0.0406
7	$\text{rej}(NR^2)$	0.0570	0.9620	1.0000	1.0000	0.0420
8	$\bar{\hat{\beta}}_2$	1.0070	1.0114	1.0116	1.0000	1.0013
9	$\text{sd}(\hat{\hat{\beta}}_2)$	0.2746	0.2731	0.2522	0.0312	0.0452
10	$\text{sd}(\hat{\hat{\beta}}_2)/\text{sd}(b_2)$	1.0338	0.9007	0.7058	0.6299	1.0920
11	$\overline{\text{se}}(\hat{\hat{\beta}}_2)$	0.2608	0.2555	0.2351	0.0323	0.0415
12	$\overline{\text{robse}}(\hat{\hat{\beta}}_2)$	0.2610	0.2565	0.2371	0.0323	0.0442
13	$\bar{\hat{\beta}}_2$	1.0124	1.0092	1.0073	0.9996	1.0007
14	$\text{sd}(\hat{\beta}_2)$	0.2924	0.2680	0.2894	0.0392	0.0414
15	$\text{sd}(\hat{\beta}_2)/\text{sd}(b_2)$	1.1009	0.8839	0.8099	0.7900	0.0406
16	$\overline{\text{se}}(\hat{\beta}_2)$	0.2677	0.2512	0.2561	0.0349	0.0406
17	$\overline{\text{robse}}(\hat{\beta}_2)$	0.2794	0.2645	0.2888	0.0395	0.0420

with robust standard errors, but we report these values for reference in row (12), $\overline{\text{robse}}\left(\hat{\hat{\beta}}_2\right)$. The averages are not much different from $\overline{\text{se}}\left(\hat{\hat{\beta}}_2\right)$, as we would have guessed.

4. The sampling variation of the GLS estimator, $\text{sd}\left(\hat{\beta}_2\right)$, is in row (14). The average of the usual, or nominal, GLS standard errors, $\overline{\text{se}}\left(\hat{\beta}_2\right)$, in row (16) is a bit too small. On average the usual GLS standard error understates the true sampling variation of the GLS estimator. However, using the heteroskedasticity robust standard error, HCE1, in row (17), on average closely measures the actual variation $\text{sd}\left(\hat{\beta}_2\right)$.
5. How well does the incorrect GLS estimator do relative to OLS and the correct FGLS estimator? When the random errors are homoskedastic, $\alpha_2 = 0$, the standard deviation of the GLS estimator is larger than that of the OLS estimator. Using GLS when OLS is appropriate is not a good idea. Note that FGLS does almost as well as OLS in this case, so there is not as much of a penalty when the pattern of heteroskedasticity is estimated. When heteroskedasticity is present the incorrect, but reasonable, GLS transformation yields estimates that are more precise than the OLS estimates. In row (15) we see that the ratio $\text{sd}\left(\hat{\beta}_2\right)/\text{sd}\left(b_2\right) < 1$ when $\alpha_2 \neq 0$. Partially curing the heteroskedasticity has produced an improvement. However, the GLS estimator improvement is not as great as for the FGLS estimator when heteroskedasticity is severe, $\alpha_2 = 0.5$.
6. How well does the NR^2 test do in detecting heteroskedasticity? Using the OLS residuals, the rejection rates of the test are $\text{rej}\left(NR^2\right)$ in row (7). When errors are homoskedastic, $\alpha_2 = 0$, the test rejects about 5% of the time as desired. When heteroskedasticity is present the test rejects homoskedasticity a very large percentage of the time, which is also desirable.
7. Finally, compare experiment (4) to experiment (3). These experiments have the same data generating process, except in experiment (3) we have 100 observations in a sample and in experiment (4) we have 5000 observations per sample. With 100 observations the standard deviation of the OLS estimates, which is the true sampling variation, is about 0.36. Using a two standard deviation rule, would being within ± 0.72 of the true parameter value $\beta_2 = 1.0$ be adequately informative for your work? If not, then the sampling variation can be reduced using FGLS, in this case so that the margin of error is ± 0.50. If that is not adequate you will need to build a better model or obtain more sample data. With 5000 observations the two standard deviation margin of error of the OLS estimates is about ± 0.10. Would that be adequate for your work? If so then nothing beyond OLS estimation with robust standard errors is needed. If not, then pursuing FGLS can reduce the margin of error to about ± 0.06. Having more good data facilitates statistical inference.

Experiment (5) is based on a different skedasticity function, $h\left(\mathbf{x}_i\right) = 3u_i/\bar{h}$, where $u_i \sim uniform(1, 11)$ is a uniform random variable, varying over the range (1,11). In this case $\text{var}\left(e_i\right) = h\left(\mathbf{x}_i\right) z_i = \sigma_i^2$ is different for each observation, heteroskedasticity is present, but the variance changes randomly from one observation to the next with no pattern and no relationship to the model explanatory variables or any other variables. This is **unconditional heteroskedasticity** and it has no effect on the properties of the OLS estimator and OLS is the best linear unbiased estimator. The NR^2 test has no ability to detect this type of heteroskedasticity.