Chapter **Eight**

Heteroskedasticity

The homoskedasticity assumption, introduced in Chapter 3 for multiple regression, states that the variance of the unobservable error, u, conditional on the explanatory variables, is constant. Homoskedasticity fails whenever the variance of the unobservables changes across different segments of the population, which are determined by the different values of the explanatory variables. For example, in a savings equation, heteroskedasticity is present if the variance of the unobserved factors affecting savings increases with income.

In Chapters 3 and 4, we saw that homoskedasticity is needed to justify the usual t tests, F tests, and confidence intervals for OLS estimation of the linear regression model, even with large sample sizes. In this chapter, we discuss the available remedies when heteroskedasticity occurs, and we also show how to test for its presence. We begin by briefly reviewing the consequences of heteroskedasticity for ordinary least squares estimation.

8.1 CONSEQUENCES OF HETEROSKEDASTICITY FOR OLS

Consider again the multiple linear regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u. \tag{8.1}$$

In Chapter 3, we proved unbiasedness of the OLS estimators $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$ under the first four Gauss-Markov assumptions, MLR.1 through MLR.4. In Chapter 5, we showed that the same four assumptions imply consistency of OLS. The homoskedasticity assumption MLR.5, stated in terms of the error variance as $\text{Var}(u|x_1,x_2,\ldots,x_k) = \sigma^2$, played no role in showing whether OLS was unbiased or consistent. It is important to remember that heteroskedasticity does not cause bias or inconsistency in the OLS estimators of the β_j, whereas something like omitting an important variable would have this effect.

If heteroskedasticity does not cause bias or inconsistency, why did we introduce it as one of the Gauss-Markov assumptions? Recall from Chapter 3 that the estimators of the *variances*, $\text{Var}(\hat{\beta}_j)$, are biased without the homoskedasticity assumption. Since the OLS standard errors are based directly on these variances, they are no longer valid for constructing confidence intervals and t statistics. The usual OLS t statistics do not have t distributions in the presence of heteroskedasticity, and the problem is not resolved by using large sample sizes. Similarly, *F* statistics are no longer *F* distributed, and the *LM* statistic no longer has an asymptotic chi-square distribution. In summary, the statistics we used to test hypotheses under the Gauss-Markov assumptions are not valid in the presence of heteroskedasticity.

We also know that the Gauss-Markov theorem, which says that OLS is best linear unbiased, relies crucially on the homoskedasticity assumption. If $\text{Var}(u|\boldsymbol{x})$ is not constant, OLS is no longer BLUE. In addition, OLS is no longer asymptotically efficient in the class of estimators described in Theorem 5.3. As we will see in Section 8.4, it is possible to find estimators that are more efficient than OLS in the presence of heteroskedasticity (although it requires knowing the form of the heteroskedasticity). With relatively large sample sizes, it might not be so important to obtain an efficient estimator. In the next section, we show how the usual OLS test statistics can be modified so that they are valid, at least asymptotically.

8.2 HETEROSKEDASTICITY-ROBUST INFERENCE AFTER OLS ESTIMATION

Since testing hypotheses is such an important component of any econometric analysis and the usual OLS inference is generally faulty in the presence of heteroskedasticity, we must decide if we should entirely abandon OLS. Fortunately, OLS is still useful. In the last two decades, econometricians have learned how to adjust standard errors, *t*, *F*, and *LM* statistics so that they are valid in the presence of **heteroskedasticity of unknown form**. This is very convenient because it means we can report new statistics that work, regardless of the kind of heteroskedasticity present in the population. The methods in this section are known as *heteroskedasticity-robust* procedures because they are valid—at least in large samples—whether or not the errors have constant variance, and we do not need to know which is the case.

We begin by sketching how the variances, $\text{Var}(\hat{\beta}_j)$, can be estimated in the presence of heteroskedasticity. A careful derivation of the theory is well-beyond the scope of this text, but the application of heteroskedasticity-robust methods is very easy now because many statistics and econometrics packages compute these statistics as an option.

First, consider the model with a single independent variable, where we include an *i* subscript for emphasis:

$$y_i = \beta_0 + \beta_1 x_i + u_i.$$

We assume throughout that the first four Gauss-Markov assumptions hold. If the errors contain heteroskedasticity, then

$$\text{Var}(u_i|x_i) = \sigma_i^2,$$

where we put an *i* subscript on σ^2 to indicate that the variance of the error depends upon the particular value of x_i.

Write the OLS estimator as

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n} (x_i - \bar{x})u_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

Under Assumptions MLR.1 through MLR.4 (that is, without the homoskedasticity assumption), and conditioning on the values x_i in the sample, we can use the same arguments from Chapter 2 to show that

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sigma_i^2}{\mathrm{SST}_x^2}, \tag{8.2}$$

where $\mathrm{SST}_x = \sum_{i=1}^{n} (x_i - \bar{x})^2$ is the total sum of squares of the x_i. When $\sigma_i^2 = \sigma^2$ for all i, this formula reduces to the usual form, σ^2/SST_x. Equation (8.2) explicitly shows that, for the simple regression case, the variance formula derived under homoskedasticity is no longer valid when heteroskedasticity is present.

Since the standard error of $\hat{\beta}_1$ is based directly on estimating $\mathrm{Var}(\hat{\beta}_1)$, we need a way to estimate equation (8.2) when heteroskedasticity is present. White (1980) showed how this can be done. Let $\hat{u}_i$ denote the OLS residuals from the initial regression of y on x. Then a valid estimator of $\mathrm{Var}(\hat{\beta}_1)$, for heteroskedasticity of *any* form (including homoskedasticity), is

$$\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2 \hat{u}_i^2}{\mathrm{SST}_x^2}, \tag{8.3}$$

which is easily computed from the data after the OLS regression.

In what sense is (8.3) a valid estimator of $\mathrm{Var}(\hat{\beta}_1)$? This is pretty subtle. Briefly, it can be shown that when equation (8.3) is multiplied by the sample size n, it converges in probability to $E[(x_i - \mu_x)^2 u_i^2]/(\sigma_x^2)^2$, which is the probability limit of n times (8.2). Ultimately, this is what is necessary for justifying the use of standard errors to construct confidence intervals and t statistics. The law of large numbers and the central limit theorem play key roles in establishing these convergences. You can refer to White's original paper for details, but that paper is quite technical. See also Wooldridge (1999, Chapter 4).

A similar formula works in the general multiple regression model

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u.$$

It can be shown that a valid estimator of $\mathrm{Var}(\hat{\beta}_j)$, under Assumptions MLR.1 through MLR.4, is

$$\hat{\mathrm{Var}}(\hat{\beta}_j) = \frac{\sum_{i=1}^{n} \hat{r}_{ij}^2 \hat{u}_i^2}{\mathrm{SST}_j^2}, \tag{8.4}$$

where $\hat{r}_{ij}$ denotes the i^{th} residual from regressing x_j on all other independent variables, and SSR_j is the sum of squared residuals from this regression (see Section 3.2 for the partialling out a representation of the OLS estimates). The square root of the quantity in (8.4) is called the **heteroskedasticity-robust standard error** for $\hat{\beta}_j$. In econometrics, these robust standard errors are usually attributed to White (1980). Earlier works in statistics, notably those by Eicker (1967) and Huber (1967), pointed to the possibility of obtaining such robust standard errors. In applied work, these are sometimes called *White*, *Huber*, or *Eicker standard errors* (or some hyphenated combination of these names). We will just refer to them as *heteroskedasticity-robust standard errors*, or even just *robust standard errors* when the context is clear.

Sometimes, as a degree of freedom correction, (8.4) is multiplied by $n/(n - k - 1)$ before taking the square root. The reasoning for this adjustment is that, if the squared OLS residuals $\hat{u}_i^2$ were the same for all observations i—the strongest possible form of homoskedasticity in a sample—we would get the usual OLS standard errors. Other modifications of (8.4) are studied in MacKinnon and White (1985). Since all forms have only asymptotic justification and they are asymptotically equivalent, no form is uniformly preferred above all others. Typically, we use whatever form is computed by the regression package at hand.

Once heteroskedasticity-robust standard errors are obtained, it is simple to construct a **heteroskedasticity-robust t statistic**. Recall that the general form of the t statistic is

$$t = \frac{\textit{estimate} - \textit{hypothesized value}}{\textit{standard error}}. \tag{8.5}$$

Since we are still using the OLS estimates and we have chosen the hypothesized value ahead of time, the only difference between the usual OLS t statistic and the heteroskedasticity-robust t statistic is in how the standard error is computed.

EXAMPLE 8.1

(Log Wage Equation with Heteroskedasticity-Robust Standard Errors)

We estimate the model in Example 7.6, but we report the heteroskedasticity-robust standard errors along with the usual OLS standard errors. Some of the estimates are reported to more digits so that we can compare the usual standard errors with the heteroskedasticity-robust standard errors:

$$\begin{array}{l}
\widehat{\log(wage)} = \underset{\substack{(.100)\\ [.109]}}{.321} + \underset{\substack{(.055)\\ [.057]}}{.213}\, marrmale - \underset{\substack{(.058)\\ [.058]}}{.198}\, marrfem - \underset{\substack{(.056)\\ [.057]}}{.110}\, singfem \\
\qquad + \underset{\substack{(.0067)\\ [.0074]}}{.0789}\, educ + \underset{\substack{(.0055)\\ [.0051]}}{.0268}\, exper - \underset{\substack{(.00011)\\ [.00011]}}{.00054}\, exper^2 \\
\qquad + \underset{\substack{(.0068)\\ [.0069]}}{.0291}\, tenure - \underset{\substack{(.00023)\\ [.00024]}}{.00053}\, tenure^2 \\
\qquad n = 526,\ R^2 = .461.
\end{array} \tag{8.6}$$

The usual OLS standard errors are in parentheses, ( ), below the corresponding OLS estimate, and the heteroskedasticity-robust standard errors are in brackets, [ ]. The numbers in brackets are the only new things, since the equation is still estimated by OLS.

Several things are apparent from equation (8.6). First, in this particular application, any variable that was statistically signficant using the usual *t* statistic is still statistically significant using the heteroskedasticity-robust *t* statistic. This is because the two sets of standard errors are not very different. (The associated *p*-values will differ slightly because the robust *t* statistics are not identical to the usual, nonrobust, *t* statistics.) The largest relative change in standard errors is for the coefficient on *educ*: the usual standard error is .0067, and the robust standard error is .0074. Still, the robust standard error implies a robust *t* statistic above 10.

Equation (8.6) also shows that the robust standard errors can be either larger or smaller than the usual standard errors. For example, the robust standard error on *exper* is .0051, whereas the usual standard error is .0055. We do not know which will be larger ahead of time. As an empirical matter, the robust standard errors are often found to be larger than the usual standard errors.

Before leaving this example, we must emphasize that we do not know, at this point, whether heteroskedasticity is even present in the population model underlying equation (8.6). All we have done is report, along with the usual standard errors, those that are valid (asymptotically) whether or not heteroskedasticity is present. We can see that no important conclusions are overturned by using the robust standard errors in this example. This often happens in applied work, but in other cases the differences between the usual and robust standard errors are much larger. As an example of where the differences are substantial, see Problem 8.7.

At this point, you may be asking the following question: If the heteroskedasticity-robust standard errors are valid more often than the usual OLS standard errors, why do we bother with the usual standard errors at all? This is a valid question. One reason they are still used in cross-sectional work is that, if the homoskedasticity assumption holds and the errors are normally distributed, then the usual *t* statistics have *exact t* distributions, regardless of the sample size (see Chapter 4). The robust standard errors and robust *t* statistics are justified only as the sample size becomes large. With small sample sizes, the robust *t* statistics can have distributions that are not very close to the *t* distribution, which would could throw off our inference.

In large sample sizes, we can make a case for always reporting only the heteroskedasticity-robust standard errors in cross-sectional applications, and this practice is being followed more and more in applied work. It is also common to report both standard errors, as in equation (8.6), so that a reader can determine whether any conclusions are sensitive to the standard error in use.

It is also possible to obtain *F* and *LM* statistics that are robust to heteroskedasticity of an unknown, arbitrary form. The **heteroskedasticity-robust *F* statistic** (or a simple transformation of it) is also called a *heteroskedasticity-robust Wald statistic*. A general treatment of this statistic is beyond the scope of this text. Nevertheless, since many statistics packages now compute these routinely, it is useful to know that heteroskedasticity-robust *F* and *LM* statistics are available. [See Wooldridge (1999) for details.]

EXAMPLE 8.2
(Heteroskedasticity-Robust *F* Statistic)

Using the data for the spring semester in GPA3.RAW, we estimate the following equation:

$$
\begin{array}{l}
\widehat{cumgpa} = \underset{\substack{(0.23)\\ [0.22]}}{1.47} + \underset{\substack{(.00018)\\ [.00019]}}{.00114}\, sat - \underset{\substack{(.00124)\\ [.00140]}}{.00857}\, hsperc + \underset{\substack{(.00073)\\ [.00073]}}{.00250}\, tothrs \\
\qquad + \underset{\substack{(.059)\\ [.059]}}{.303}\, female - \underset{\substack{(.147)\\ [.118]}}{.128}\, black - \underset{\substack{(.141)\\ [.110]}}{.059}\, white \\
\qquad n = 366,\ R^2 = .4006,\ \bar{R}^2 = .3905.
\end{array}
\qquad \textbf{(8.7)}
$$

Again, the differences between the usual standard errors and the heteroskedasticity-robust standard errors are not very big, and use of the robust *t* statistics does not change the statistical significance of any independent variable. Joint significance tests are not much affected either. Suppose we wish to test the null hypothesis that, after the other factors are controlled for, there are no differences in *cumgpa* by race. This is stated as H_0: $\beta_{black} = 0$, $\beta_{white} = 0$. The usual *F* statistic is easily obtained, once we have the *R*-squared from the restricted model; this turns out to be .3983. The *F* statistic is then $[(.4006 - .3983)/(1 - .4006)](359/2) \approx .69$. If heteroskedasticity is present, this version of the test is invalid. The heteroskedasticity-robust version has no simple form, but it can be computed using certain statistical packages. The value of the heteroskedasticity-robust *F* statistic turns out to be .75, which differs only slightly from the nonrobust version. The *p*-value for the robust test is .474, which is not close to standard significance levels. We fail to reject the null hypothesis using either test.

Computing Heteroskedasticity-Robust *LM* Tests

Not all regression packages compute *F* statistics that are robust to heteroskedasticity. Therefore, it is sometimes convenient to have a way of obtaining a test of multiple exclusion restrictions that is robust to heteroskedasticity and does not require a particular kind of econometric software. It turns out that a **heteroskedasticity-robust *LM* statistic** is easily obtained using virtually any regression package.

QUESTION 8.1

Evaluate the following statement: The heteroskedasticity-robust standard errors are always bigger than the usual standard errors.

To illustrate computation of the robust *LM* statistic, consider the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + u,$$

and suppose we would like to test H$_0$: $\beta_4 = 0$, $\beta_5 = 0$. To obtain the usual *LM* statistic, we would first estimate the restricted model (that is, the model without x_4 and x_5) to obtain the residuals, $\tilde{u}$. Then, we would regress $\tilde{u}$ on all of the independent variables and the $LM = n \cdot R^2_{\tilde{u}}$, where $R^2_{\tilde{u}}$ is the usual *R*-squared from this regression.

Obtaining a version that is robust to heteroskedasticity requires more work. One way to compute the statistic requires only OLS regressions. We need the residuals, say $\tilde{r}_1$, from the regression of x_4 on x_1, x_2, x_3. Also, we need the residuals, say $\tilde{r}_2$, from the regression of x_5 on x_1, x_2, x_3. Thus, we regress each of the independent variables excluded under the null on all of the included independent variables. We keep the residuals each time. The final step appears odd, but it is, after all, just a computational device. Run the regression of

$$1 \text{ on } \tilde{r}_1\tilde{u},\ \tilde{r}_2\tilde{u}, \tag{8.8}$$

without an intercept. Yes, we actually define a dependent variable equal to the value one for all observations. We regress this onto the products $\tilde{r}_1\tilde{u}$ and $\tilde{r}_2\tilde{u}$. The robust *LM* statistic turns out to be $n - \text{SSR}_1$, where SSR$_1$ is just the usual sum of squared residuals from regression (8.8).

The reason this works is somewhat technical. Basically, this is doing for the *LM* test what the robust standard errors do for the *t* test. [See Wooldridge (1991b) or Davidson and MacKinnon (1993) for a more detailed discussion.]

We now summarize the computation of the heteroskedasticity-robust *LM* statistic in the general case.

A HETEROSKEDASTICITY-ROBUST *LM* STATISTIC:

1. Obtain the residuals $\tilde{u}$ from the restricted model.
2. Regress each of the independent variables excluded under the null on all of the included independent variables; if there are q excluded variables, this leads to q sets of residuals ($\tilde{r}_1, \tilde{r}_2, \ldots, \tilde{r}_q$).
3. Find the products between each $\tilde{r}_j$ and $\tilde{u}$ (for all observations).
4. Run the regression of 1 on $\tilde{r}_1\tilde{u}$, $\tilde{r}_2\tilde{u}$, …, $\tilde{r}_q\tilde{u}$, without an intercept. The heteroskedasticity-robust *LM* statistic is $n - \text{SSR}_1$, where SSR$_1$ is just the usual sum of squared residuals from this final regression. Under H$_0$, *LM* is distributed approximately as χ^2_q.

Once the robust *LM* statistic is obtained, the rejection rule and computation of *p*-values is the same as for the usual *LM* statistic in Section 5.2.

EXAMPLE 8.3
(Heteroskedasticity-Robust *LM* Statistic)

We use the data in CRIME1.RAW to test whether the average sentence length served for past convictions affects the number of arrests in the current year (1986). The estimated model is

$$
\begin{array}{llllll}
\widehat{narr86} = & .567 & -\ .136\ pcnv & +\ .0178\ avgsen & -\ .00052\ avgsen^2 & \\
 & (.036) & (.040) & (.0097) & (.00030) & \\
 & [.040] & [.034] & [.0101] & [.00021] & \\
 & -\ .0394\ ptime86 & -\ .0505\ qemp86 & -\ .00148\ inc86 & & \textbf{(8.9)} \\
 & (.0087) & (.0144) & (.00034) & & \\
 & [.0062] & [.0142] & [.00023] & & \\
 & +\ .325\ black & +\ .193\ hispan & & & \\
 & (.045) & (.040) & & & \\
 & [.058] & [.040] & & & \\
 & n = 2{,}725,\ R^2 = .0728. & & & &
\end{array}
$$

In this example, there are more substantial differences between some of the usual standard errors and the robust standard errors. For example, the usual t statistic on $avgsen^2$ is about -1.73, while the robust t statistic is about -2.48. Thus, $avgsen^2$ is more significant using the robust standard error.

The effect of *avgsen* on *narr86* is somewhat difficult to reconcile. Since the relationship is quadratic, we can figure out where *avgsen* has a positive effect on *narr86* and where the effect becomes negative. The turning point is $.0178/[2(.00052)] \approx 17.12$; recall that this is measured in months. Literally, this means that *narr86* is positively related to *avgsen* when *avgsen* is less than 17 months; then *avgsen* has the expected deterrent effect after 17 months.

To see whether average sentence length has a statistically significant effect on *narr86*, we must test the joint hypothesis H_0: $\beta_{avgsen} = 0$, $\beta_{avgsen^2} = 0$. Using the usual *LM* statistic (see Section 5.2), we obtain $LM = 3.54$; in a chi-square distribution with two *df*, this yields a p-value $= .170$. Thus, we do not reject H_0 at even the 15% level. The heteroskedasticity-robust *LM* statistic is $LM = 4.00$ (rounding to two decimal places), with a p-value $= .135$. This is still not very strong evidence against H_0; *avgsen* does not appear to have a strong effect on *narr86*. [Incidentally, when *avgsen* appears alone in (8.9), that is, without the quadratic term, its usual t statistic is .658, and its robust t statistic is .592.]

8.3 TESTING FOR HETEROSKEDASTICITY

The heteroskedasticity-robust standard errors provide a simple method for computing t statistics that are asymptotically t distributed whether or not heteroskedasticity is present. We have also seen that heteroskedasticity-robust F and LM statistics are available. Implementing these tests does not require knowing whether or not heteroskedasticity is present. Nevertheless, there are still some good reasons for having simple tests that can detect its presence. First, as we mentioned in the previous section, the usual t statistics have exact t distributions under the classical linear model assumptions. For this reason, many economists still prefer to see the usual OLS standard errors and test statistics reported, unless there is evidence of heteroskedasticity. Second, if heteroskedasticity is present, the OLS estimator is no longer the best linear unbiased estimator. As we will see in Section 8.4, it is possible to obtain a better estimator than OLS when the form of heteroskedasticity is known.

Many tests for heteroskedasticity have been suggested over the years. Some of them, while having the ability to detect heteroskedasticity, do not directly test the assumption that the variance of the error does not depend upon the independent variables. We will restrict ourselves to more modern tests, which detect the kind of heteroskedasticity that invalidates the usual OLS statistics. This also has the benefit of putting all tests in the same framework.

As usual, we start with the linear model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u, \tag{8.10}$$

where Assumptions MLR.1 through MLR.4 are maintained in this section. In particular, we assume that $\mathrm{E}(u|x_1,x_2,\ldots,x_k) = 0$, so that OLS is unbiased and consistent.

We take the null hypothesis to be that Assumption MLR.5 is true:

$$\mathrm{H}_0\colon \mathrm{Var}(u|x_1,x_2,\ldots,x_k) = \sigma^2. \tag{8.11}$$

That is, we assume that the ideal assumption of homoskedasticity holds, and we require the data to tell us otherwise. If we cannot reject (8.11) at a sufficiently small significance level, we usually conclude that heteroskedasticity is not a problem. However, remember that we never accept H_0; we simply fail to reject it.

Because we are assuming that u has a zero conditional expectation, $\mathrm{Var}(u|\boldsymbol{x}) = \mathrm{E}(u^2|\boldsymbol{x})$, and so the null hypothesis of homoskedasticity is equivalent to

$$\mathrm{H}_0\colon \mathrm{E}(u^2|x_1,x_2,\ldots,x_k) = \mathrm{E}(u^2) = \sigma^2.$$

This shows that, in order to test for violation of the homoskedasticity assumption, we want to test whether u^2 is related (in expected value) to one or more of the explanatory variables. If H_0 is false, the expected value of u^2, given the independent variables, can be any function of the x_j. A simple approach is to assume a linear function:

$$u^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \ldots + \delta_k x_k + v, \tag{8.12}$$

where v is an error term with mean zero given the x_j. Pay close attention to the dependent variable in this equation: it is the *square* of the error in the original regression equation, (8.10). The null hypothesis of homoskedasticity is

$$\mathrm{H}_0\colon \delta_1 = \delta_2 = \ldots = \delta_k = 0. \tag{8.13}$$

Under the null hypothesis, it is often reasonable to assume that the error in (8.12), v, is independent of $x_1, x_2,\ldots,x_k$. Then, we know from Section 5.2 that either the F or LM statistics for the overall significance of the independent variables in explaining u^2 can be used to test (8.13). Both statistics would have asymptotic justification, even though u^2 cannot be normally distributed. (For example, if u is normally distributed, then u^2/σ^2 is distributed as χ_1^2.) If we could observe the u^2 in the sample, then we could easily compute this statistic by running the OLS regression of u^2 on $x_1, x_2,\ldots,x_k$, using all n observations.

As we have emphasized before, we never know the actual errors in the population model, but we do have estimates of them: the OLS residual, $\hat{u}_i$, is an estimate of the error u_i for observation i. Thus, we can estimate the equation

$$\hat{u}^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \ldots + \delta_k x_k + error \tag{8.14}$$

and compute the F or LM statistics for the joint significance of $x_1, \ldots, x_k$. It turns out that using the OLS residuals in place of the errors does not affect the large sample distribution of the F or LM statistics, although showing this is pretty complicated.

The F and LM statistics both depend on the R-squared from regression (8.14); call this $R^2_{\hat{u}2}$ to distinguish it from the R-squared in estimating equation (8.10). Then, the F statistic is

$$F = \frac{R^2_{\hat{u}2}/k}{(1 - R^2_{\hat{u}2})/(n - k - 1)}, \tag{8.15}$$

where k is the number of regressors in (8.14); this is the same number of independent variables in (8.10). Computing (8.15) by hand is rarely necessary, since most regression packages automatically compute the F statistic for overall significance of a regression. This F statistic has (approximately) an $F_{k,n-k-1}$ distribution under the null hypothesis of homoskedasticity.

The LM statistic for heteroskedasticity is just the sample size times the R-squared from (8.14):

$$LM = n \cdot R^2_{\hat{u}2}. \tag{8.16}$$

Under the null hypothesis, LM is distributed asymptotically as χ^2_k. This is also very easy to obtain after running regression (8.14).

The LM version of the test is typically called the **Breusch-Pagan test for heteroskedasticity** (**BP test**). Breusch and Pagan (1980) suggested a different form of the test that assumes the errors are normally distributed. Koenker (1983) suggested the form of the LM statistic in (8.16), and it is generally preferred due to its greater applicability.

We summarize the steps for testing for heteroskedasticity using the BP test:

THE BREUSCH-PAGAN TEST FOR HETEROSKEDASTICITY.

1. Estimate the model (8.10) by OLS, as usual. Obtain the squared OLS residuals, $\hat{u}^2$ (one for each observation).
2. Run the regression in (8.14). Keep the R-squared from this regression, $R^2_{\hat{u}2}$.
3. Form either the F statistic or the LM statistic and compute the p-value (using the $F_{k,n-k-1}$ distribution in the former case and the χ^2_k distribution in the latter case). If the p-value is sufficiently small, that is, below the chosen significance level, then we reject the null hypothesis of homoskedasticity.

If the BP test results in a small enough p-value, some corrective measure should be taken. One possibility is to just use the heteroskedasticity-robust standard errors and test statistics discussed in the previous section. Another possibility is discussed in Section 8.4.

EXAMPLE 8.4
(Heteroskedasticity in Housing Price Equations)

We use the data in HPRICE1.RAW to test for heteroskedasticity in a simple housing price equation. The estimated equation using the levels of all variables is

$$\begin{aligned}\widehat{price} = &-21.77 + .00207\ lotsize + .123\ sqrft + 13.85\ bdrms \\ &(29.48)\quad (.00064)\qquad\quad (.013)\qquad\quad (9.01) \\ &n = 88,\ R^2 = .672.\end{aligned} \tag{8.17}$$

This equation tells us *nothing* about whether the error in the population model is heteroskedastic. We need to regress the squared OLS residuals on the independent variables. The *R*-squared from the regression of $\hat{u}^2$ on *lotsize*, *sqrft*, and *bdrms* is $R^2_{\hat{u}^2} = .1601$. With $n = 88$ and $k = 3$, this produces an *F* statistic for significance of the independent variables of $F = [.1601/(1 - .1601)](84/3) \approx 5.34$. The associated *p*-value is .002, which is strong evidence against the null. The *LM* statistic is $88(.1601) \approx 14.09$; this gives a *p*-value $\approx$ .0028 (using the χ^2_3 distribution), giving essentially the same conclusion as the *F* statistic. This means that the usual standard errors reported in (8.17) are not reliable.

In Chapter 6, we mentioned that one benefit of using the logarithmic functional form for the dependent variable is that heteroskedasticity is often reduced. In the current application, let us put *price*, *lotsize*, and *sqrft* in logarithmic form, so that the elasticities of *price*, with respect to *lotsize* and *sqrft*, are constant. The estimated equation is

$$\begin{aligned}\log(\widehat{price}) = &\ 5.61 + .168\log(lotsize) + .700\log(sqrft) + .037\ bdrms \\ &(.65)\quad (.038)\qquad\qquad (.093)\qquad\qquad (.028) \\ &n = 88,\ R^2 = .643.\end{aligned} \tag{8.18}$$

Regressing the squared OLS residuals from this regression on log(*lotsize*), log(*sqrft*), and *bdrms* gives $R^2_{\hat{u}^2} = .0480$. Thus, $F = 1.41$ (*p*-value = .245), and $LM = 4.22$ (*p*-value = .239). Therefore, we fail to reject the null hypothesis of homoskedasticity in the model with the logarithmic functional forms. The occurrence of less heteroskedasticity with the dependent variable in logarithmic form has been noticed in many empirical applications.

QUESTION 8.2

Consider wage equation (7.11), where you think that the conditional variance of log(*wage*) does not depend on *educ*, *exper*, or *tenure*. However, you are worried that the variance of log(*wage*) differs across the four demographic groups of married males, married females, single males, and single females. What regression would you run to test for heteroskedasticity? What are the degrees of freedom in the *F* test?

If we suspect that heteroskedasticity depends only upon certain independent variables, we can easily modify the Breusch-Pagan test: we simply regress $\hat{u}^2$ on whatever independent variables we choose and carry out the appropriate *F* or *LM* test. Remember that the appropriate degrees of freedom depends upon the number of independent variables in the regression with $\hat{u}^2$ as the dependent variable; the number of independent variables showing up in equation (8.10) is irrelevant.

If the squared residuals are regressed on only a single independent variable, the test for heteroskedasticity is just the usual t statistic on the variable. A significant t statistic suggests that heteroskedasticity is a problem.

The White Test for Heteroskedasticity

In Chapter 5, we showed that the usual OLS standard errors and test statistics are asymptotically valid, provided all of the Gauss-Markov assumptions hold. It turns out that the homoskedasticity assumption, $\text{Var}(u_1|x_1,\ldots,x_k) = \sigma^2$, can be replaced with the weaker assumption that the squared error, u^2, is *uncorrelated* with all the independent variables (x_j), the squares of the independent variables (x_j^2), and all the cross products (x_jx_h for $j \neq h$). This observation motivated White (1980) to propose a test for heteroskedasticity that adds the squares and cross products of all of the independent variables to equation (8.14). The test is explicitly intended to test for forms of heteroskedasticity that invalidate the usual OLS standard errors and test statistics.

When the model contains $k = 3$ independent variables, the White test is based on an estimation of

$$\begin{aligned}\hat{u}^2 = \delta_0 + \delta_1x_1 + \delta_2x_2 + \delta_3x_3 + \delta_4x_1^2 + \delta_5x_2^2 + \delta_6x_3^2 \\ + \delta_7x_1x_2 + \delta_8x_1x_3 + \delta_9x_2x_3 + error.\end{aligned} \tag{8.19}$$

Compared with the Breusch-Pagan test, this equation has six more regressors. The **White test for heteroskedasticity** is the LM statistic for testing that all of the δ_j in equation (8.19) are zero, except for the intercept. Thus, nine restrictions are being tested in this case. We can also use an F test of this hypothesis; both tests have asymptotic justification.

With only three independent variables in the original model, equation (8.19) has nine independent variables. With six independent variables in the original model, the White regression would generally involve 27 regressors (unless some are redundant). This abundance of regressors is a weakness in the pure form of the White test: it uses many degrees of freedom for models with just a moderate number of independent variables.

It is possible to obtain a test that is easier to implement than the White test and more conserving on degrees of freedom. To create the test, recall that the difference between the White and Breusch-Pagan tests is that the former includes the squares and cross products of the independent variables. We can achieve the same thing by using fewer functions of the independent variables. One suggestion is to use the OLS fitted values in a test for heteroskedasticity. Remember that the fitted values are defined, for each observation i, by

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1x_{i1} + \hat{\beta}_2x_{i2} + \ldots + \hat{\beta}_kx_{ik}.$$

These are just linear functions of the independent variables. If we square the fitted values, we get a particular function of all the squares and cross products of the independent variables. This suggests testing for heteroskedasticity by estimating the equation

$$\hat{u}^2 = \delta_0 + \delta_1\hat{y} + \delta_2\hat{y}^2 + error, \tag{8.20}$$

where $\hat{y}$ stands for the fitted values. It is important not to confuse $\hat{y}$ and y in this equation. We use the fitted values because they are functions of the independent variables (and the estimated parameters); using y in (8.20) does not produce a valid test for heteroskedasticity.

We can use the F or LM statistic for the null hypothesis H_0: $\delta_1 = 0$, $\delta_2 = 0$ in equation (8.20). This results in two restrictions in testing the null of homoskedasticity, regardless of the number of independent variables in the original model. Conserving on degrees of freedom in this way is often a good idea, and it also makes the test easy to implement.

Since $\hat{y}$ is an estimate of the expected value of y, given the x_j, using (8.20) to test for heteroskedasticity is useful in cases where the variance is thought to change with the level of the expected value, $E(y|\boldsymbol{x})$. The test from (8.20) can be viewed as a special case of the White test, since equation (8.20) can be shown to impose restrictions on the parameters in equation (8.19).

A SPECIAL CASE OF THE WHITE TEST FOR HETEROSKEDASTICITY:

1. Estimate the model (8.10) by OLS, as usual. Obtain the OLS residuals $\hat{u}$ and the fitted values $\hat{y}$. Compute the squared OLS residuals $\hat{u}^2$ and the squared fitted values $\hat{y}^2$.
2. Run the regression in equation (8.20). Keep the R-squared from this regression, $R^2_{\hat{u}^2}$.
3. Form either the F or LM statistic and compute the p-value (using the $F_{2,n-3}$ distribution in the former case and the χ^2_2 distribution in the latter case).

EXAMPLE 8.5

(Special Form of the White Test in the Log Housing Price Equation)

We apply the special case of the White test to equation (8.18), where we use the *LM* form of the statistic. The important thing to remember is that the chi-square distribution always has two *df*. The regression of $\hat{u}^2$ on $\widehat{lprice}$, $(\widehat{lprice})^2$, where $\widehat{lprice}$ denotes the fitted values from (8.18), produces $R^2_{\hat{u}^2} = .0392$; thus, $LM = 88(.0392) \approx 3.45$, and the p-value $= .178$. This is stronger evidence of heteroskedasticity than is provided by the Breusch-Pagan test, but we still fail to reject homoskedasticity at even the 15% level.

Before leaving this section, we should discuss one important caveat. We have interpreted a rejection using one of the heteroskedasticity tests as evidence of heteroskedasticity. This is appropriate provided we maintain Assumptions MLR.1 through MLR.4. But, if MLR.3 is violated—in particular, if the functional form of $E(y|\boldsymbol{x})$ is misspecified—then a test for heteroskedastcity can reject H_0, even if $\text{Var}(y|\boldsymbol{x})$ is constant. For example, if we omit one or more quadratic terms in a regression model or use the level model when we should use the log, a test for heteroskedasticity can be significant. This has led some economists to view tests for heteroskedasticity as general misspecification tests. However, there are better, more direct tests for functional form misspecification, and we will cover some of them in Section 9.1. It is better to use explicit tests for functional form first, since functional form misspecification is more important than heteroskedasticity. Then, once we are satisfied with the functional form, we can test for heteroskedasticity.

8.4 WEIGHTED LEAST SQUARES ESTIMATION

If heteroskedasticity is detected using one of the tests in Section 8.3, we know from Section 8.2 that one possible response is to use heteroskedasticity-robust statistics after estimation by OLS. Before the development of heteroskedasticity-robust statistics, the response to a finding of heteroskedasticity was to model and estimate its specific form. As we will see, this leads to a more efficient estimator than OLS, and it produces *t* and *F* statistics that have *t* and *F* distributions. While this seems attractive, it actually requires more work on our part because we must be very specific about the nature of any heteroskedasticity.

The Heteroskedasticity Is Known up to a Multiplicative Constant

Let $\boldsymbol{x}$ denote all the explanatory variables in equation (8.10) and assume that

$$\mathrm{Var}(u|\boldsymbol{x}) = \sigma^2 h(\boldsymbol{x}), \tag{8.21}$$

where $h(\boldsymbol{x})$ is some function of the explanatory variables that determines the heteroskedasticity. Since variances must be positive, $h(\boldsymbol{x}) > 0$ for all possible values of the independent variables. We assume in this subsection that the function $h(\boldsymbol{x})$ is known. The population parameter σ^2 is unknown, but we will be able to estimate it from a data sample.

For a random drawing from the population, we can write $\sigma_i^2 = \mathrm{Var}(u_i|\boldsymbol{x}_i) = \sigma^2 h(\boldsymbol{x}_i) = \sigma^2 h_i$, where we again use the notation $\boldsymbol{x}_i$ to denote all independent variables for observation *i*, and h_i changes with each observation because the independent variables change across observations. For example, consider the simple savings function

$$sav_i = \beta_0 + \beta_1 inc_i + u_i \tag{8.22}$$

$$\mathrm{Var}(u_i|inc_i) = \sigma^2 inc_i. \tag{8.23}$$

Here, $h(inc) = inc$: the variance of the error is proportional to the level of income. This means that, as income increases, the variability in savings increases. (If $\beta_1 > 0$, the expected value of savings also increases with income.) Because *inc* is always positive, the variance in equation (8.23) is always guaranteed to be positive. The standard deviation of u_i, conditional on inc_i, is $\sigma\sqrt{inc_i}$.

How can we use the information in equation (8.21) to estimate the β_j? Essentially, we take the original equation,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + u_i, \tag{8.24}$$

which contains heteroskedastic errors, and transform it into an equation that has homoskedastic errors (and satisfies the other Gauss-Markov assumptions). Since h_i is just a function of $\boldsymbol{x}_i$, $u_i/\sqrt{h_i}$ has a zero expected value conditional on $\boldsymbol{x}_i$. Further, since $\text{Var}(u_i|\boldsymbol{x}_i) = \text{E}(u_i^2|\boldsymbol{x}_i) = \sigma^2 h_i$, the variance of $u_i/\sqrt{h_i}$ (conditional on $\boldsymbol{x}_i$) is σ^2:

$$\text{E}\left((u_i/\sqrt{h_i})^2\right) = \text{E}(u_i^2)/h_i = (\sigma^2 h_i)/h_i = \sigma^2,$$

where we have suppressed the conditioning on $\boldsymbol{x}_i$ for simplicity. We can divide equation (8.24) by $\sqrt{h_i}$ to get

$$\begin{aligned} y_i/\sqrt{h_i} = \beta_0/\sqrt{h_i} &+ \beta_1(x_{i1}/\sqrt{h_i}) + \beta_2(x_{i2}/\sqrt{h_i}) + \ldots \\ &+ \beta_k(x_{ik}/\sqrt{h_i}) + (u_i/\sqrt{h_i}) \end{aligned} \tag{8.25}$$

or

$$y_i^* = \beta_0 x_{i0}^* + \beta_1 x_{i1}^* + \ldots + \beta_k x_{ik}^* + u_i^*, \tag{8.26}$$

where $x_{i0}^* = 1/\sqrt{h_i}$ and the other starred variables denote the corresponding original variables divided by $\sqrt{h_i}$.

Equation (8.26) looks a little peculiar, but the important thing to remember is that we derived it so we could obtain estimators of the β_j that have better efficiency properties than OLS. The intercept β_0 in the original equation (8.24) is now multiplying the variable $x_{i0}^* = 1/\sqrt{h_i}$. Each slope parameter in β_j multiplies a new variable that rarely has a useful interpretation. This should not cause problems if we recall that, for interpreting the parameters and the model, we always want to return to the original equation (8.24).

In the preceding savings example, the transformed equation looks like

$$sav_i/\sqrt{inc_i} = \beta_0(1/\sqrt{inc_i}) + \beta_1\sqrt{inc_i} + u_i^*,$$

where we use the fact that $inc_i/\sqrt{inc_i} = \sqrt{inc_i}$. Nevertheless, β_1 is the marginal propensity to save out of income, an interpretation we obtain from equation (8.22).

Equation (8.26) is linear in its parameters (so it satisfies MLR.1), and the random sampling assumption has not changed. Further, u_i^* has a zero mean and a constant variance (σ^2), conditional on $\boldsymbol{x}_i^*$. This means that if the original equation satisfies the first four Gauss-Markov assumptions, then the transformed equation (8.26) satisfies all five Gauss-Markov assumptions. Also, if u_i has a normal distribution, then u_i^* has a normal distribution with variance σ^2. Therefore, the transformed equation satisfies the classical linear model assumptions (MLR.1 through MLR.6), if the original model does so, except for the homoskedasticity assumption.

Since we know that OLS has appealing properties (is BLUE, for example) under the Gauss-Markov assumptions, the discussion in the previous paragraph suggests estimating the parameters in equation (8.26) by ordinary least squares. These estimators, β_0^*, β_1^*, …, β_k^*, will be different from the OLS estimators in the original equation. The β_j^* are examples of **generalized least squares** (**GLS**) **estimators**. In this case, the GLS estimators are used to account for heteroskedasticity in the errors. We will encounter other GLS estimators in Chapter 12.

Since equation (8.26) satisfies all of the ideal assumptions, standard errors, *t* statistics, and *F* statistics can all be obtained from regressions using the transformed variables. The sum of squared residuals from (8.26) divided by the degrees of freedom is an unbiased estimator of σ^2. Further, the GLS estimators, because they are the best linear unbiased estimators of the β_j, are necessarily more efficient than the OLS estimators $\hat{\beta}_j$ obtained from the untransformed equation. Essentially, after we have transformed the variables, we simply use standard OLS analysis. But we must remember to interpret the estimates in light of the original equation.

The *R*-squared that is obtained from estimating (8.26), while useful for computing *F* statistics, is not especially informative as a goodness-of-fit measure: it tells us how much variation in y^* is explained by the x_j^*, and this is seldom very meaningful.

The GLS estimators for correcting heteroskedasticity are called **weighted least squares** (**WLS**) **estimators**. This name comes from the fact that the β_j^* minimize the *weighted* sum of squared residuals, where each squared residual is weighted by $1/h_i$. The idea is that less weight is given to observations with a higher error variance; OLS gives each observation the same weight because it is best when the error variance is identical for all partitions of the population. Mathematically, the WLS estimators are the values of the b_j that make

$$\sum_{i=1}^{n} (y_i - b_0 - b_1x_{i1} - b_2x_{i2} - \ldots - b_kx_{ik})^2/h_i \qquad \textbf{(8.27)}$$

as small as possible. Bringing the square root of $1/h_i$ inside the squared residual shows that the weighted sum of squared residuals is identical to the sum of squared residuals in the transformed variables:

$$\sum_{i=1}^{n} (y_i^* - b_0x_{i0}^* - b_1x_{i1}^* - b_2x_{i2}^* - \ldots - b_kx_{ik}^*)^2.$$

It follows that the WLS estimators that minimize (8.27) are simply the OLS estimators from (8.26).

A weighted least squares estimator can be defined for any set of positive weights. OLS is the special case that gives equal weight to all observations. The efficient procedure, GLS, weights each squared residual by the *inverse* of the conditional variance of u_i given $\boldsymbol{x}_i$.

Obtaining the transformed variables in order to perform weighted least squares can be tedious, and the chance of making mistakes is nontrivial. Fortunately, most modern regression packages have a feature for doing weighted least squares. Typically, along with the dependent and independent variables in the original model, we just specify the weighting function. In addition to making mistakes less likely, this forces us to interpret weighted least squares estimates in the original model. In fact, we can write out the estimated equation in the usual way. The estimates and standard errors will be different from OLS, but the way we *interpret* those estimates, standard errors, and test statistics is the same.

EXAMPLE 8.6
(Family Saving Equation)

Table 8.1 contains estimates of saving functions from the data set SAVING.RAW (on 100 families from 1970). We estimate the simple regression model (8.22) by OLS and by weighted least squares, assuming in the latter case that the variance is given by (8.23). We then add variables for family size, age of the household head, years of education for the household head, and a dummy variable indicating whether the household head is black.

In the simple regression model, the OLS estimate of the marginal propensity to save (MPS) is .147, with a *t* statistic of 2.53. (The standard errors in Table 8.1 for OLS are the nonrobust standard errors. If we really thought heteroskedasticity was a problem, we would probably compute the heteroskedasticity-robust standard errors as well; we will not do that here.) The WLS estimate of the MPS is somewhat higher: .172, with $t = 3.02$. The standard errors of the OLS and WLS estimates are very similar for this coefficient. The intercept estimates are very different for OLS and WLS, but this should cause no concern since the *t* statistics are both very small. Finding fairly large changes in coefficients that are insignificant is not uncommon when comparing OLS and WLS estimates. The *R*-squareds in columns (1) and (2) are not comparable.

Table 8.1

Dependent Variable: *sav*

Independent Variables	(1) OLS	(2) WLS	(3) OLS	(4) WLS
inc	.147 (.058)	.172 (.057)	.109 (.071)	.101 (.077)
size	———	———	67.66 (222.96)	−6.87 (168.43)
educ	———	———	151.82 (117.25)	139.48 (100.54)
age	———	———	.286 (50.031)	21.75 (41.31)
black	———	———	518.39 (1,308.06)	137.28 (844.59)
intercept	124.84 (655.39)	−124.95 (480.86)	−1,605.42 (2,830.71)	−1,854.81 (2,351.80)
Observations *R*-Squared	100 .0621	100 .0853	100 .0828	100 .1042

Adding demographic variables reduces the MPS whether OLS or WLS is used; the standard errors also increase by a fair amount (due to multicollinearity that is induced by adding these additional variables). It is easy to see, using either the OLS or WLS estimates, that none of the additional variables is individually significant. Are they jointly significant? The *F* test based on the OLS estimates uses the *R*-squareds from columns (1) and (3). With 94 *df* in the unrestricted model and four restrictions, the *F* statistic is $F = [(.0828 - .0621)/(1 - .0828)](94/4) \approx .53$ and *p*-value = .715. The *F* test, using the WLS estimates, uses the *R*-squareds from columns (2) and (4): $F \approx .50$ and *p*-value = .739. Thus, using either OLS or WLS, the demographic variables are jointly insignificant. This suggests that the simple regression model relating savings to income is sufficient.

What should we choose as our best estimate of the marginal propensity to save? In this case, it does not matter much whether we use the OLS estimate of .147 or the WLS estimate of .172. Remember, both are just estimates from a relatively small sample, and the OLS 95% confidence interval contains the WLS estimate, and vice versa.

In practice, we rarely know how the variance depends on a particular independent variable in a simple form. For example, in the savings equation that includes all demographic variables, how do we know that the variance of *sav* does not change with age or education levels? In most applications, we are unsure about $\text{Var}(y|x_1,x_2\ldots,x_k)$.

QUESTION 8.3

Using the OLS residuals obtained from the OLS regression reported in column (1) of Table 8.1, the regression of $\hat{u}^2$ on *inc* yields a *t* statistic on *inc* of .96. Is there any need to use weighted least squares in Example 8.6?

There is one case where the weights needed for WLS arise naturally from an underlying econometric model. This happens when, instead of using individual level data, we only have averages of data across some group or geographic region. For example, suppose we are interested in determining the relationship between the amount a worker contributes to his or her 401(k) pension plan as a function of the plan generosity. Let *i* denote a particular firm and let *e* denote an employee within the firm. A simple model is

$$contrib_{i,e} = \beta_0 + \beta_1 earns_{i,e} + \beta_2 age_{i,e} + \beta_3 mrate_i + u_{i,e}, \tag{8.28}$$

where $contrib_{i,e}$ is the annual contribution by employee *e* who works for firm *i*, $earns_{i,e}$ is annual earnings for this person, and $age_{i,e}$ is the person's age. The variable $mrate_i$ is the amount the firm puts into an employee's account for every dollar the employee contributes.

If (8.28) satisfies the Gauss-Markov assumptions, then we could estimate it, given a sample on individuals across various employers. Suppose, however, that we only have *average* values of contributions, earnings, and age by employer. In other words, individual-level data are not available. Thus, let $\overline{contrib}_i$ denote average contribution for people at firm *i*, and similarly for $\overline{earns}_i$ and $\overline{age}_i$. Let m_i denote the number of employees at each firm; we assume that this is a known quantity. Then, if we average equation (8.28) across all employees at firm *i*, we obtain the firm-level equation

$$\overline{contrib}_i = \beta_0 + \beta_1 \overline{earns}_i + \beta_2 \overline{age}_i + \beta_3 mrate_i + \bar{u}_i, \tag{8.29}$$

where $\bar{u}_i = m_i^{-1}\sum_{e=1}^{m_i} u_{i,e}$ is the average error across all employees in firm i. If we have n firms in our sample, then (8.29) is just a standard multiple linear regression model that can be estimated by OLS. The estimators are unbiased if the original model (8.28) satisfies the Gauss-Markov assumptions and the individual errors $u_{i,e}$ are independent of the firm's size, m_i (because then the expected value of $\bar{u}_i$, given the explanatory variables in (8.29), is zero).

If the equation at the individual level satisfies the homoskedasticity assumption, then the firm-level equation (8.29) must have heteroskedasticity. In fact, if $\text{Var}(u_{i,e}) = \sigma^2$ for all i and e, then $\text{Var}(\bar{u}_i) = \sigma^2/m_i$. In other words, for larger firms, the variance of the error term $\bar{u}_i$ decreases with firm size. In this case, $h_i = 1/m_i$, and so the most efficient procedure is weighted least squares, with weights equal to the number of employees at the firm ($1/h_i = m_i$). This ensures that larger firms receive more weight. This gives us an efficient way of estimating the parameters in the individual-level model when we only have averages at the firm level.

A similar weighting arises when we are using per capita data at the city, county, state, or country level. If the individual-level equation satisfies the Gauss-Markov assumptions, then the error in the per capita equation has a variance proportional to one over the size of the population. Therefore, weighted least squares with weights equal to the population is appropriate. For example, suppose we have city-level data on per capita beer consumption (in ounces), the percentage of people in the population over 21 years old, average adult education levels, average income levels, and the city price of beer. Then the city-level model

$$beerpc = \beta_0 + \beta_1 perc21 + \beta_2 avgeduc + \beta_2 incpc + \beta_2 price + u$$

can be estimated by weighted least squares, with the weights being the city population.

The advantage of weighting by firm size, city population, and so on relies on the underlying individual equation being homoskedastic. If heteroskedasticity exists at the individual level, then the proper weighting depends on the form of the heteroskedasticity. This is one reason why more and more researchers simply compute robust standard errors and test statistics when estimating models using per capita data. An alternative is to weight by population but to report the heteroskedasticity-robust statistics in the WLS estimation. This ensures that, while the estimation is efficient if the individual-level model satisfies the Gauss-Markov assumptions, any heteroskedasticity at the individual level is accounted for through robust inference.

The Heteroskedasticity Function Must Be Estimated: Feasible GLS

In the previous subsection, we saw some examples of where the heteroskedasticity is known up to a multiplicative form. In most cases, the exact form of heteroskedasticity is not obvious. In other words, it is difficult to find the function $h(\boldsymbol{x}_i)$ of the previous section. Nevertheless, in many cases we can model the function h and use the data to estimate the unknown parameters in this model. This results in an estimate of each h_i, denoted as $\hat{h}_i$. Using $\hat{h}_i$ instead of h_i in the GLS transformation yields an estimator called the **feasible GLS (FGLS) estimator**. Feasible GLS is sometimes called *estimated GLS*, or EGLS.

There are many ways to model heteroskedasticity, but we will study one particular, fairly flexible approach. Assume that

$$\mathrm{Var}(u|\boldsymbol{x}) = \sigma^2\exp(\delta_0 + \delta_1 x_1 + \delta_2 x_2 + \ldots + \delta_k x_k), \tag{8.30}$$

where $x_1, x_2, \ldots, x_k$ are the independent variables appearing in the regression model [see equation (8.1)], and the δ_j are unknown parameters. Other functions of the x_j can appear, but we will focus primarily on (8.30). In the notation of the previous subsection, $h(\boldsymbol{x}) = \exp(\delta_0 + \delta_1 x_1 + \delta_2 x_2 + \ldots + \delta_k x_k)$.

You may wonder why we have used the exponential function in (8.30). After all, when *testing* for heteroskedasticity using the Breusch-Pagan test, we assumed that heteroskedasticity was a linear function of the x_j. Linear alternatives such as (8.12) are fine when testing for heteroskedasticity, but they can be problematic when correcting for heteroskedasticity using weighted least squares. We have encountered the reason for this problem before: linear models do not ensure that predicted values are positive, and our estimated variances must be positive in order to perform WLS.

If the parameters δ_j were known, then we would just apply WLS, as in the previous subsection. This is not very realistic. It is better to use the data to estimate these parameters, and then to use these estimates to construct weights. How can we estimate the δ_j? Essentially, we will transform this equation into a linear form that, with slight modification, can be estimated by OLS.

Under assumption (8.30), we can write

$$u^2 = \sigma^2\exp(\delta_0 + \delta_1 x_1 + \delta_2 x_2 + \ldots + \delta_k x_k)v,$$

where v has a mean equal to unity, conditional on $\boldsymbol{x} = (x_1, x_2, \ldots, x_k)$. If we assume that v is actually independent of $\boldsymbol{x}$, we can write

$$\log(u^2) = \alpha_0 + \delta_1 x_1 + \delta_2 x_2 + \ldots + \delta_k x_k + e, \tag{8.31}$$

where e has a zero mean and is independent of $\boldsymbol{x}$; the intercept in this equation is different from δ_0, but this is not important. The dependent variable is the log of the squared error. Since (8.31) satisfies the Gauss-Markov assumptions, we can get unbiased estimators of the δ_j by using OLS.

As usual, we must replace the unobserved u with the OLS residuals. Therefore, we run the regression of

$$\log(\hat{u}^2) \text{ on } x_1, x_2, \ldots, x_k. \tag{8.32}$$

Actually, what we need from this regression are the fitted values; call these $\hat{g}_i$. Then, the estimates of h_i are simply

$$\hat{h}_i = \exp(\hat{g}_i). \tag{8.33}$$

We now use WLS with weights $1/\hat{h}_i$. We summarize the steps.

A FEASIBLE GLS PROCEDURE TO CORRECT FOR HETEROSKEDASTICITY:

1. Run the regression of y on x_1, x_2, ..., x_k and obtain the residuals, $\hat{u}$.
2. Create $\log(\hat{u}^2)$ by first squaring the OLS residuals and then taking the natural log.
3. Run the regression in equation (8.32) and obtain the fitted values, $\hat{g}$.
4. Exponentiate the fitted values from (8.32): $\hat{h} = \exp(\hat{g})$.
5. Estimate the equation

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + u$$

by WLS, using weights $1/\hat{h}$.

If we could use h_i rather than $\hat{h}_i$ in the WLS procedure, we know that our estimators would be unbiased; in fact, they would be the best linear unbiased estimators, assuming that we have properly modeled the heteroskedasticity. Having to estimate h_i using the same data means that the FGLS estimator is no longer unbiased (so it cannot be BLUE, either). Nevertheless, the FGLS estimator is consistent and *asymptotically* more efficient than OLS. This is difficult to show because of estimation of the variance parameters. But if we ignore this—as it turns out we may—the proof is similar to showing that OLS is efficient in the class of estimators in Theorem 5.3. At any rate, for large sample sizes, FGLS is an attractive alternative to OLS when there is evidence of heteroskedasticity that inflates the standard errors of the OLS estimates.

We must remember that the FGLS estimators are estimators of the parameters in the equation

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + u.$$

Just as the OLS estimates measure the marginal impact of each x_j on y, so do the FGLS estimates. We use the FGLS estimates in place of the OLS estimates because they are more efficient and have associated test statistics with the usual t and F distributions, at least in large samples. If we have some doubt about the variance specified in equation (8.30), we can use heteroskedasticity-robust standard errors and test statistics in the transformed equation.

Another useful alternative for estimating h_i is to replace the independent variables in regression (8.32) with the OLS fitted values and their squares. In other words, obtain the $\hat{g}_i$ as the fitted values from the regression of

$$\log(\hat{u}^2) \textit{ on } \hat{y}, \hat{y}^2 \qquad \textbf{(8.34)}$$

and then obtain the $\hat{h}_i$ exactly as in equation (8.33). This changes only step (3) in the previous procedure.

If we use regression (8.32) to estimate the variance function, you may be wondering if we can simply test for heteroskedasticity using this same regression (an F or LM test can be used). In fact, Park (1966) suggested this. Unfortunately, when compared with the tests discussed in Section 8.3, the Park test has some problems. First, the null hypothesis must be something stronger than homoskedasticity: effectively, u and $\boldsymbol{x}$ must be independent. This is not required in the Breusch-Pagan or White tests. Second, using the OLS residuals $\hat{u}$ in place of u in (8.32) can cause the F statistic to deviate from the

F distribution, even in large sample sizes. This is not an issue in the other tests we have covered. For these reasons, the Park test is not recommended when testing for heteroskedasticity. The reason that regression (8.32) works well for weighted least squares is that we only need consistent estimators of the δ_j, and regression (8.32) certainly delivers those.

EXAMPLE 8.7

(Demand for Cigarettes)

We use the data in SMOKE.RAW to estimate a demand function for daily cigarette consumption. Since most people do not smoke, the dependent variable, *cigs*, is zero for most observations. A linear model is not ideal because it can result in negative predicted values. Nevertheless, we can still learn something about the determinants of cigarette smoking by using a linear model.

The equation estimated by ordinary least squares, with the usual OLS standard errors in parentheses, is

$$
\begin{array}{l}
\widehat{cigs} = \underset{(24.08)}{-3.64} + \underset{(.728)}{.880}\ \log(income) - \underset{(5.773)}{.751}\ \log(cigpric) \\
\quad - \underset{(.167)}{.501}\ educ + \underset{(.160)}{.771}\ age - \underset{(.0017)}{.0090}\ age^2 - \underset{(1.11)}{2.83}\ restaurn \\
\qquad n = 807,\ R^2 = .0526,
\end{array}
\tag{8.35}
$$

where *cigs* is number of cigarettes smoked per day, *income* is annual income, *cigpric* is the per pack price of cigarettes (in cents), *educ* is years of schooling, *age* is measured in years, and *restaurn* is a binary indicator equal to unity if the person resides in a state with restaurant smoking restrictions. Since we are also going to do weighted least squares, we do not report the heteroskedasticity-robust standard errors for OLS. (Incidentally, 13 out of the 807 fitted values are less than zero; this is less than 2% of the sample and is not a major cause for concern.)

Neither income nor cigarette price is statistically significant in (8.35), and their effects are not practically large. For example, if income increases by 10%, *cigs* is predicted to increase by (.880/100)(10) = .088, or less than one-tenth of a cigarette per day. The magnitude of the price effect is similar.

Each year of education reduces the average cigarettes smoked per day by one-half, and the effect is statistically significant. Cigarette smoking is also related to age, in a quadratic fashion. Smoking increases with age up until $age = .771/[2(.009)] \approx 42.83$, and then smoking decreases with age. Both terms in the quadratic are statistically significant. The presence of a restriction on smoking in restaurants decreases cigarette smoking by almost three cigarettes per day, on average.

Do the errors underlying equation (8.35) contain heteroskedasticity? The Breusch-Pagan regression of the squared OLS residuals on the independent variables in (8.35) [see equation (8.14)] produces $R^2_{\hat{u}^2} = .040$. This small *R*-squared may seem to indicate no heteroskedasticity, but we must remember to compute either the *F* or *LM* statistic. If the sample size is large, a seemingly small $R^2_{\hat{u}^2}$ can result in a very strong rejection of homoskedasticity. The *LM* statistic is $LM = 807(.040) = 32.28$, and this is the outcome of a χ^2_6 random variable. The *p*-value is less than .000015, which is very strong evidence of heteroskedasticity.

Therefore, we estimate the equation using the previous feasible GLS procedure. The estimated equation is

$$\begin{aligned}\widehat{cigs} = &\ \underset{(17.80)}{5.64} + \underset{(.44)}{1.30}\log(income) - \underset{(4.46)}{2.94}\log(cigpric) \\ &- \underset{(.120)}{.463}\, educ + \underset{(.097)}{.482}\, age - \underset{(.0009)}{.0056}\, age^2 - \underset{(.80)}{3.46}\, restaurn \\ &n = 807,\ R^2 = .1134.\end{aligned} \tag{8.36}$$

The income effect is now statistically significant and larger in magnitude. The price effect is also notably bigger, but it is still statistically insignificant. (One reason for this is that *cigpric* varies only across states in the sample, and so there is much less variation in log(*cigpric*) than in log(*income*), *educ*, and *age*.)

The estimates on the other variables have, naturally, changed somewhat, but the basic story is still the same. Cigarette smoking is negatively related to schooling, has a quadratic relationship with *age*, and is negatively affected by restaurant smoking restrictions.

We must be a little careful in computing *F* statistics for testing multiple hypotheses after estimation by WLS. (This is true whether the sum of squared residuals or *R*-squared form of the *F* statistic is used.) It is important that the same weights be used to estimate the unrestricted and restricted models. We should first estimate the unrestricted model by OLS. Once we have obtained the weights, we can use them to estimate the restricted model as well. The *F* statistic can be computed as usual. Fortunately, many regression packages have a simple command for testing joint restrictions after WLS estimation, so we need not perform the restricted regression ourselves.

Example 8.7 hints at an issue that sometimes arises in applications of weighted least squares: the OLS and WLS estimates can be substantially different. This is not such a big problem in the demand for cigarettes equation because all the coefficients maintain the same signs, and the biggest changes are on variables that were statistically insignificant when the equation was estimated by OLS. The OLS and WLS estimates will always differ due to sampling error. The issue is whether their difference is enough to change important conclusions.

QUESTION 8.4

Suppose that the model for heteroskedasticity in equation (8.30) is not correct, but we use the feasible GLS procedure based on this variance. WLS is still consistent, but the usual standard errors, *t* statistics, and so on will not be valid, even asymptotically. What can we do instead? [*Hint*: See equation (8.26), where u_i^* contains heteroskedasticity if $\text{Var}(u|\boldsymbol{x}) \neq \sigma^2 h(\boldsymbol{x})$.]

If OLS and WLS produce statistically significant estimates that differ in sign—for example, the OLS price elasticity is positive and significant, while the WLS price elasticity is negative and signficant—or the difference in magnitudes of the estimates is practically large, we should be suspicious. Typically, this indicates that one of the *other*

Gauss-Markov assumptions is false, particularly the zero conditional mean assumption on the error (MLR.3). Correlation between u and any independent variable causes bias and inconsistency in OLS *and* WLS, and the biases will usually be different. The *Hausman test* [Hausman (1978)] can be used to formally compare the OLS and WLS estimates to see if they differ by more than the sampling error suggests. This test is beyond the scope of this text. In many cases, an informal "eyeballing" of the estimates is sufficient to detect a problem.

8.5 THE LINEAR PROBABILITY MODEL REVISITED

As we saw in Section 7.6, when the dependent variable y is a binary variable, the model must contain heteroskedasticity, unless all of the slope parameters are zero. We are now in a position to deal with this problem.

The simplest way to deal with heteroskedasticity in the linear probability model is to continue to use OLS estimation, but to also compute robust standard errors in test statistics. This ignores the fact that we actually know the form of heteroskedasticity for the LPM. Nevertheless, OLS estimates of the LPM is simple and often produces satisfactory results.

EXAMPLE 8.8
(Labor Force Participation of Married Women)

In the labor force participation example in Section 7.6 [see equation (7.29)], we reported the usual OLS standard errors. Now we compute the heteroskedasticity-robust standard errors as well. These are reported in brackets below the usual standard errors:

$$
\begin{array}{l}
\widehat{inlf} = \underset{\substack{(.154)\\ [.151]}}{.586} - \underset{\substack{(.0014)\\ [.0015]}}{.0034}\, nwifeinc + \underset{\substack{(.007)\\ [.007]}}{.038}\, educ + \underset{\substack{(.006)\\ [.006]}}{.039}\, exper \\
\quad - \underset{\substack{(.00018)\\ [.00019]}}{.00060}\, exper^2 - \underset{\substack{(.002)\\ [.002]}}{.016}\, age - \underset{\substack{(.034)\\ [.032]}}{.262}\, kidslt6 + \underset{\substack{(.0132)\\ [.0135]}}{.0130}\, kidsge6 \\
\qquad n = 753,\ R^2 = .264.
\end{array}
\tag{8.37}
$$

Several of the robust and OLS standard errors are the same to the reported degree of precision; in all cases the differences are practically very small. Therefore, while heteroskedasticity is a problem in theory, it is not in practice, at least not for this example. It often turns out that the usual OLS standard errors and test statistics are similar to their heteroskedasticity-robust counterparts. Furthermore, it requires a minimal effort to compute both.

Generally, the OLS esimators are inefficient in the LPM. Recall that the conditional variance of y in the LPM is

$$\mathrm{Var}(y|\boldsymbol{x}) = p(\boldsymbol{x})[1 - p(\boldsymbol{x})], \tag{8.38}$$

where

$$p(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k \tag{8.39}$$

is the response probability (probability of success, $y = 1$). It seems natural to use weighted least squares, but there are a couple of hitches. The probability $p(\boldsymbol{x})$ clearly depends on the unknown population parameters, β_j. Nevertheless, we do have unbiased estimators of these parameters, namely the OLS estimators. When the OLS estimators are plugged into equation (8.39), we obtain the OLS fitted values. Thus, for each observation i, $\mathrm{Var}(y_i|\boldsymbol{x}_i)$ is estimated by

$$\hat{h}_i = \hat{y}_i(1 - \hat{y}_i), \tag{8.40}$$

where $\hat{y}_i$ is the OLS fitted value for observation i. Now we apply feasible GLS, just as in Section 8.4.

Unfortunately, being able to estimate h_i for each i does not mean that we can proceed directly with WLS estimation. The problem is one that we briefly discussed in Section 7.6: the fitted values $\hat{y}_i$ need not fall in the unit interval. If either $\hat{y}_i < 0$ or $\hat{y}_i > 1$, equation (8.40) shows that $\hat{h}_i$ will be negative. Since WLS proceeds by multiplying observation i by $1/\sqrt{\hat{h}_i}$, the method will fail if $\hat{h}_i$ is negative (or zero) for any observation. In other words, all of the weights for WLS must be positive.

In some cases, $0 < \hat{y}_i < 1$ for all i, in which case WLS can be used to estimate the LPM. In cases with many observations and small probabilities of success or failure, it is very common to find some fitted values outside the unit interval. If this happens, as it does in the labor force participation example in equation (8.37), it is easiest to abandon WLS and to report the heteroskedasticity-robust statistics. An alternative is to adjust those fitted values that are less than zero or greater than unity, and then to apply WLS. One suggestion is to set $\hat{y}_i = .01$ if $\hat{y}_i < 0$ and $\hat{y}_i = .99$ if $\hat{y}_i > 1$. Unfortunately, this requires an arbitrary choice on the part of the researcher—for example, why not use .001 and .999 as the adjusted values? If many fitted values are outside the unit interval, the adjustment to the fitted values can affect the results; in this situation, it is probably best to just use OLS.

ESTIMATING THE LINEAR PROBABILITY MODEL BY WEIGHTED LEAST SQUARES:

1. Estimate the model by OLS and obtain the fitted values, $\hat{y}$.
2. Determine whether all of the fitted values are inside the unit interval. If so, proceed to step (3). If not, some adjustment is needed to bring all fitted values into the unit interval.
3. Construct the estimated variances in equation (8.40).
4. Estimate the equation

   $$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u$$

   by WLS, using weights $1/\hat{h}$.

EXAMPLE 8.9

(Determinants of Personal Computer Ownership)

We use the data in GPA1.RAW to estimate the probability of owning a computer. Let *PC* denote a binary indicator equal to unity if the student owns a computer, and zero otherwise. The variable *hsGPA* is high school GPA, *ACT* is achievement test score, and *parcoll* is a binary indicator equal to unity if at least one parent attended college. (Separate college indicators for the mother and the father do not yield individually significant results, as these are pretty highly correlated.)

The equation estimated by OLS is

$$\begin{array}{lllll} \widehat{PC} = & -.0004 & + .065\, hsGPA & + .0006\, ACT & + .221\, parcoll \\ & (.4905) & (.137) & (.0155) & (.093) \\ & [.4888] & [.139] & [.0158] & [.087] \\ \multicolumn{5}{c}{n = 141,\ R^2 = .0415.} \end{array} \tag{8.41}$$

Just as with Example 8.8, there are no striking differences between the usual and robust standard errors. Nevertheless, we also estimate the model by WLS. Because all of the OLS fitted values are inside the unit interval, no adjustments are needed:

$$\begin{array}{lllll} \widehat{PC} = & .026 & + .033\, hsGPA & + .0043\, ACT & + .215\, parcoll \\ & (.477) & (.130) & (.0155) & (.086) \\ \multicolumn{5}{c}{n = 141,\ R^2 = .0464.} \end{array} \tag{8.42}$$

There are no important differences in the OLS and WLS estimates. The only significant explanatory variable is *parcoll*, and in both cases we estimate that the probability of PC ownership is about .22 higher, if at least one parent attended college.

SUMMARY

We began by reviewing the properties of ordinary least squares in the presence of heteroskedasticity. Heteroskedasticity does not cause bias or inconsistency in the OLS estimators, but the usual standard errors and test statistics are no longer valid. We showed how to compute heteroskedasticity-robust standard errors and t statistics, something that is routinely done by many regression packages. Most regression packages also compute a heteroskedasticity-robust, F-type statistic.

We discussed two common ways to test for heteroskedasticity: the Breusch-Pagan test and a special case of the White test. Both of these statistics involve regressing the *squared* OLS residuals on either the independent variables (BP) or the fitted and squared fitted values (White). A simple F test is asymptotically valid; there are also Lagrange multiplier versions of the tests.

OLS is no longer the best linear unbiased estimator in the presence of heteroskedasticity. When the form of heteroskedasticity is known, generalized least squares (GLS) estimation can be used. This leads to weighted least squares as a means of obtaining the BLUE estimator. The test statistics from the WLS estimation are either exactly valid when the error term is normally distributed or asymptotically valid under nonnormality. This assumes, of course, that we have the proper model of heteroskedasticity.

More commonly, we must estimate a model for the heteroskedasticity before applying WLS. The resulting *feasible* GLS estimator is no longer unbiased, but it is consistent and asymptotically efficient. The usual statistics from the WLS regression are asymptotically valid. We discussed a method to ensure that the estimated variances are strictly positive for all observations, something needed to apply WLS.

As we discussed in Chapter 7, the linear probability model for a binary dependent variable necessarily has a heteroskedastic error term. A simple way to deal with this problem is to compute heteroskedasticity-robust statistics. Alternatively, if all the fitted values (that is, the estimated probabilities) are strictly between zero and one, weighted least squares can be used to obtain asymptotically efficient estimators.

KEY TERMS

Breusch-Pagan Test for Heteroskedasticity (BP Test)
Feasible GLS (FGLS) Estimator
Generalized Least Squares (GLS) Estimators
Heteroskedasticity of Unknown Form
Heteroskedasticity-Robust Standard Error
Heteroskedasticity-Robust *F* Statistic
Heteroskedasticity-Robust *LM* Statistic
Heteroskedasticity-Robust *t* Statistic
Weighted Least Squares (WLS) Estimators
White Test for Heteroskedasticity

PROBLEMS

8.1 Which of the following are consequences of heteroskedasticity?

(i) The OLS estimators, $\hat{\beta}_j$, are inconsistent.

(ii) The usual *F* statistic no longer has an *F* distribution.

(iii) The OLS estimators are no longer BLUE.

8.2 Consider a linear model to explain monthly beer consumption:

$$beer = \beta_0 + \beta_1 inc + \beta_2 price + \beta_3 educ + \beta_4 female + u$$
$$\mathrm{E}(u|inc,price,educ,female) = 0$$
$$\mathrm{Var}(u|inc,price,educ,female) = \sigma^2 inc^2.$$

Write the transformed equation that has a homoskedastic error term.

8.3 True or False: WLS is preferred to OLS, when an important variable has been omitted from the model.

8.4 Using the data in GPA3.RAW, the following equation was estimated for the fall and second semester students:

$$
\begin{array}{lllll}
\widehat{trmgpa} = & -2.12 & +\ .900\ crsgpa & +\ .193\ cumgpa & +\ .0014\ tothrs \\
 & (.55) & (.175) & (.064) & (.0012) \\
 & [.55] & [.166] & [.074] & [.0012]
\end{array}
$$

$$
\begin{array}{llll}
+\ .0018\ sat & -\ .0039\ hsperc & +\ .351\ female & -\ .157\ season \\
(.0002) & (.0018) & (.085) & (.098) \\
[.0002] & [.0019] & [.079] & [.080]
\end{array}
$$

$$n = 269,\ R^2 = .465.$$

Here, *trmgpa* is term GPA, *crsgpa* is a weighted average of overall GPA in courses taken, *tothrs* is total credit hours prior to the semester, *sat* is SAT score, *hsperc* is graduating percentile in high school class, *female* is a gender dummy, and *season* is a dummy variable equal to unity if the student's sport is in season during the fall. The usual and heteroskedasticity-robust standard errors are reported in parentheses and brackets, respectively.

(i) Do the variables *crsgpa*, *cumgpa*, and *tothrs* have the expected estimated effects? Which of these variables are statistically significant at the 5% level? Does it matter which standard errors are used?

(ii) Why does the hypothesis H_0: $\beta_{crsgpa} = 1$ make sense? Test this hypothesis against the two-sided alternative at the 5% level, using both standard errors. Describe your conclusions.

(iii) Test whether there is an in-season effect on term GPA, using both standard errors. Does the significance level at which the null can be rejected depend on the standard error used?

8.5 The variable *smokes* is a binary variable equal to one if a person smokes, and zero otherwise. Using the data in SMOKE.RAW, we estimate a linear probability model for *smokes*:

$$
\begin{array}{lllll}
\widehat{smokes} = & .656 & -\ .069\ \log(cigpric) & +\ .012\ \log(income) & -\ .029\ educ \\
 & (.855) & (.204) & (.026) & (.006) \\
 & [.856] & [.207] & [.026] & [.006]
\end{array}
$$

$$
\begin{array}{llll}
+\ .020\ age & -\ .00026\ age^2 & -\ .101\ restaurn & -\ .026\ white \\
(.006) & (.00006) & (.039) & (.052) \\
[.005] & [.00006] & [.038] & [.050]
\end{array}
$$

$$n = 807,\ R^2 = .062.$$

The variable *white* equals one if the respondent is white, and zero otherwise; the other independent variables are defined in Example 8.7. Both the usual and heteroskedasticity-robust standard errors are reported.

(i) Are there any important differences between the two sets of standard errors?

(ii) Holding other factors fixed, if education increases by four years, what happens to the estimated probability of smoking?

(iii) At what point does another year of age reduce the probability of smoking?

(iv) Interpret the coefficient on the binary variable *restaurn* (a dummy variable equal to one if the person lives in a state with restaurant smoking restrictions).

(v) Person number 206 in the data set has the following characteristics: $cigpric = 67.44$, $income = 6{,}500$, $educ = 16$, $age = 77$, $restaurn = 0$, $white = 0$, and $smokes = 0$. Compute the predicted probability of smoking for this person and comment on the result.

COMPUTER EXERCISES

8.6 Use the data in SLEEP75.RAW to estimate the following sleep equation:

$$sleep = \beta_0 + \beta_1 totwrk + \beta_2 educ + \beta_3 age + \beta_4 age^2 + \beta_5 yngkid + \beta_6 male + u.$$

(i) Write down a model that allows the variance of u to differ between men and women. The variance should not depend on other factors.

(ii) Estimate the parameters of the model for heteroskedasticty. (You have to estimate the *sleep* equation by OLS, first, to obtain the OLS residuals.) Is the estimated variance of u higher for men or for women?

(iii) Is the variance of u statistically different for men and for women?

8.7 (i) Use the data in HPRICE1.RAW to obtain the heteroskedasticity-robust standard errors for equation (8.17). Discuss any important differences with the usual standard errors.

(ii) Repeat part (i) for equation (8.18).

(iii) What does this example suggest about heteroskedasticity and the transformation used for the dependent variable?

8.8 Apply the full White test for heteroskedasticity [see equation (8.19)] to equation (8.18). Using the chi-square form of the statistic, obtain the p-value. What do you conclude?

8.9 Use VOTE1.RAW for this exercise.

(i) Estimate a model with *voteA* as the dependent variable and *prtystrA*, *democA*, log(*expendA*), and log(*expendB*) as independent variables. Obtain the OLS residuals, $\hat{u}_i$, and regress these on all of the independent variables. Explain why you obtain $R^2 = 0$.

(ii) Now compute the Breusch-Pagan test for heteroskedasticity. Use the F statistic version and report the p-value.

(iii) Compute the special case of the White test for heteroskedasticity, again using the F statistic form. How strong is the evidence for heteroskedasticity now?

8.10 Use the data in PNTSPRD.RAW for this exercise.

(i) The variable *sprdcvr* is a binary variable equal to one if the Las Vegas point spread for a college basketball game was covered. The expected value of *sprdcvr*, say μ, is the probability that the spread is covered in a randomly selected game. Test H_0: $\mu = .5$ against H_1: $\mu \neq .5$ at the 10% significance level and discuss your findings. (*Hint*: This is easily done using a t test by regressing *sprdcvr* on an intercept only.)

(ii) How many games in the sample of 553 were played on a neutral court?

(iii) Estimate the linear probability model

$$sprdcvr = \beta_0 + \beta_1 favhome + \beta_2 neutral + \beta_3 fav25 + \beta_4 und25 + u$$

and report the results in the usual form. (Report the usual OLS standard errors and the heteroskedasticity-robust standard errors.) Which variable is most significant, both practically and statistically?

(iv) Explain why, under the null hypothesis H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$, there is no heteroskedasticity in the model.

(v) Use the usual *F* statistic to test the hypothesis in part (iv). What do you conclude?

(vi) Given the previous analysis, would you say that it is possible to systematically predict whether the Las Vegas spread will be covered using information available prior to the game?

8.11 In Example 7.12, we estimated a linear probability model for whether a young man was arrested during 1986:

$$arr86 = \beta_0 + \beta_1 pcnv + \beta_2 avgsen + \beta_3 tottime + \beta_4 ptime86 + \beta_5 qemp86 + u.$$

(i) Estimate this model by OLS and verify that all fitted values are strictly between zero and one. What are the smallest and largest fitted values?

(ii) Estimate the equation by weighted least squares, as discussed in Section 8.5.

(iii) Use the WLS estimates to determine whether *avgsen* and *tottime* are jointly significant at the 5% level.

8.12 Use the data in LOANAPP.RAW for this exercise.

(i) Estimate the equation in part (iii) of Problem 7.16, computing the heteroskedasticity-robust standard errors. Compare the 95% confidence interval on β_{white} with the nonrobust confidence interval.

(ii) Obtain the fitted values from the regression in part (i). Are any of them less than zero? Are any of them greater than one? What does this mean about applying weighted least squares?