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 *F* Statistic
Heteroskedasticity-Robust *LM* Statistic
Heteroskedasticity-Robust Standard Error
Heteroskedasticity-Robust *t* Statistic
Weighted Least Squares (WLS) Estimators
White Test for Heteroskedasticity

Problems

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.

2 Consider a linear model to explain monthly beer consumption:

$$beer = \beta_0 + \beta_1 inc + \beta_2 price + \beta_3 educ + \beta_4 female + u$$

$$\begin{aligned} \mathrm{E}(u|inc, price, educ, female) &= 0 \\ \mathrm{Var}(u|inc, price, educ, female) &= \sigma^2 inc^2. \end{aligned}$$

Write the transformed equation that has a homoskedastic error term.

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

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

$$\begin{array}{lllll} \widehat{trmgpa} = & -2.12 & + .900\ crsgpa & + .193\ cumgpa & + .0014\ tothrs \\ & (.55) & (.175) & (.064) & (.0012) \\ & [.55] & [.166] & [.074] & [.0012] \\ & + .0018\ sat & - .0039\ hsperc & + .351\ female & - .157\ season \\ & (.0002) & (.0018) & (.085) & (.098) \\ & [.0002] & [.0019] & [.079] & [.080] \\ n = 269, R^2 = .465. & & & & \end{array}$$

Here, *trmgpa* is term GPA, *crsgpa* is a weighted average of overall GPA in courses taken, *cumgpa* is GPA prior to the current semester, *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?

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

$$\begin{aligned}\widehat{smokes} = {} & \underset{[.856]}{\underset{(.855)}{.656}} - \underset{[.207]}{\underset{(.204)}{.069}} \log(cigpric) + \underset{[.026]}{\underset{(.026)}{.012}} \log(income) - \underset{[.006]}{\underset{(.006)}{.029}}\, educ \\ & + \underset{[.005]}{\underset{(.006)}{.020}}\, age - \underset{[.00006]}{\underset{(.00006)}{.00026}}\, age^2 - \underset{[.038]}{\underset{(.039)}{.101}}\, restaurn - \underset{[.050]}{\underset{(.052)}{.026}}\, white \\ n = {} & 807,\ R^2 = .062.\end{aligned}$$

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.

6 There are different ways to combine features of the Breusch-Pagan and White tests for heteroskedasticity. One possibility not covered in the text is to run the regression

$$\hat{u}_i^2 \text{ on } x_{i1}, x_{i2}, \ldots, x_{ik}, \hat{y}_i^2,\ i = 1, \ldots, n,$$

where the $\hat{u}_i$ are the OLS residuals and the $\hat{y}_i$ are the OLS fitted values. Then, we would test joint significance of $x_{i1}, x_{i2}, \ldots, x_{ik}$ and $\hat{y}_i^2$. (Of course, we always include an intercept in this regression.)

(i) What are the *df* associated with the proposed *F* test for heteroskedasticity?

(ii) Explain why the *R*-squared from the regression above will always be at least as large as the *R*-squareds for the BP regression and the special case of the White test.

(iii) Does part (ii) imply that the new test always delivers a smaller *p*-value than either the BP or special case of the White statistic? Explain.

(iv) Suppose someone suggests also adding $\hat{y}_i$ to the newly proposed test. What do you think of this idea?

7 Consider a model at the employee level,

$$y_{i,e} = \beta_0 + \beta_1 x_{i,e,1} + \beta_2 x_{i,e,2} + \ldots + \beta_k x_{i,e,k} + f_i + v_{i,e},$$

where the unobserved variable f_i is a "firm effect" to each employee at a given firm i. The error term $v_{i,e}$ is specific to employee e at firm i. The *composite error* is $u_{i,e} = f_i + v_{i,e}$, such as in equation (8.28).

(i) Assume that $\text{Var}(f_i) = \sigma_f^2$, $\text{Var}(v_{i,e}) = \sigma_v^2$, and f_i and $v_{i,e}$ are uncorrelated. Show that $\text{Var}(u_{i,e}) = \sigma_f^2 + \sigma_v^2$; call this σ^2.

(ii) Now suppose that for $e \neq g$, $v_{i,e}$ and $v_{i,g}$ are uncorrelated. Show that $\text{Cov}(u_{i,e}, u_{i,g}) = \sigma_f^2$

(iii) Let $\bar{u}_i = m_i^{-1}\sum_{e=1}^{mi} u_{i,e}$ be the average of the composite errors within a firm. Show that $\text{Var}(\bar{u}_i) = \sigma_f^2 + \sigma_v^2/m_i$.

(iv) Discuss the relevance of part (iii) for WLS estimation using data averaged at the firm level, where the weight used for observation i is the usual firm size.

8 The following equations were estimated using the data in ECONMATH. The first equation is for men and the second is for women. The third and fourth equations combine men and women.

$$\widehat{score} = \underset{(3.72)}{20.52} + \underset{(0.94)}{13.60}\, colgpa + \underset{(0.150)}{0.670}\, act$$

$$n = 406.\ R^2 = .4025,\ \text{SSR} = 38{,}781.38.$$

$$\widehat{score} = \underset{(4.11)}{13.79} + \underset{(1.09)}{11.89}\, colgpa + \underset{(0.18)}{1.03}\, act$$

$$n = 408,\ R^2 = .3666,\ \text{SSR} = 48{,}029.82.$$

$$\widehat{score} = \underset{(2.80)}{15.60} + \underset{(0.73)}{3.17}\, male + \underset{(0.72)}{12.82}\, colgpa + \underset{(0.116)}{0.838}\, act$$

$$n = 814,\ R^2 = .3946,\ \text{SSR} = 87{,}128.96.$$

$$\widehat{score} = \underset{(3.91)}{13.79} + \underset{(5.55)}{6.73}\, male + \underset{(1.04)}{11.89}\, colgpa + \underset{(0.17)}{1.03}\, act + \underset{(1.44)}{1.72}\, male \cdot colgpa - \underset{(0.232)}{0.364}\, male \cdot act$$

$$n = 814,\ R^2 = .3968,\ \text{SSR} = 86{,}811.20.$$

(i) Compute the usual Chow statistic for testing the null hypothesis that the regression equations are the same for men and women. Find the p-value of the test.

(ii) Compute the usual Chow statistic for testing the null hypothesis that the slope coefficients are the same for men and women, and report the p-value.

(iii) Do you have enough information to compute heteroskedasticity-robust versions of the tests in (ii) and (iii)? Explain.

Computer Exercises

C1 Consider the following model to explain sleeping behavior:

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

(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) Use the data in SLEEP75 to estimate the parameters of the model for heteroskedasticity. (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?

C2 (i) Use the data in HPRICE1 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?

C3 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?

C4 Use VOTE1 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?

C5 Use the data in PNTSPRD 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?

C6 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) Using the data in CRIME1, 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.

C7 Use the data in LOANAPP for this exercise.

(i) Estimate the equation in part (iii) of Computer Exercise C8 in Chapter 7, 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?

C8 Use the data set GPA1 for this exercise.

(i) Use OLS to estimate a model relating *colGPA* to *hsGPA*, *ACT*, *skipped*, and *PC*. Obtain the OLS residuals.

(ii) Compute the special case of the White test for heteroskedasticity. In the regression of $\hat{u}_i^2$ on $\widehat{colGPA}_i$, $\widehat{colGPA}_i^2$, obtain the fitted values, say $\hat{h}_i$.

(iii) Verify that the fitted values from part (ii) are all strictly positive. Then, obtain the weighted least squares estimates using weights $1/\hat{h}_i$. Compare the weighted least squares estimates for the effect of skipping lectures and the effect of PC ownership with the corresponding OLS estimates. What about their statistical significance?

(iv) In the WLS estimation from part (iii), obtain heteroskedasticity-robust standard errors. In other words, allow for the fact that the variance function estimated in part (ii) might be misspecified. (See Question 8.4.) Do the standard errors change much from part (iii)?

C9 In Example 8.7, we computed the OLS and a set of WLS estimates in a cigarette demand equation.

(i) Obtain the OLS estimates in equation (8.35).

(ii) Obtain the $\hat{h}_i$ used in the WLS estimation of equation (8.36) and reproduce equation (8.36). From this equation, obtain the *unweighted* residuals and fitted values; call these $\hat{u}_i$ and $\hat{y}_i$, respectively. (For example, in Stata, the unweighted residuals and fitted values are given by default.)

(iii) Let $\breve{u}_i = \hat{u}_i/\sqrt{\hat{h}_i}$ and $\breve{y}_i = \hat{y}_i/\sqrt{\hat{h}_i}$ be the weighted quantities. Carry out the special case of the White test for heteroskedasticity by regressing $\breve{u}_i^2$ on $\breve{y}_i$, $\breve{y}_i^2$, being sure to include an intercept, as always. Do you find heteroskedasticity in the weighted residuals?

(iv) What does the finding from part (iii) imply about the proposed form of heteroskedasticity used in obtaining (8.36)?

(v) Obtain valid standard errors for the WLS estimates that allow the variance function to be misspecified.

C10 Use the data set 401KSUBS for this exercise.

(i) Using OLS, estimate a linear probability model for *e401k*, using as explanatory variables *inc, inc*^{2}, *age, age*^{2}, and *male*. Obtain both the usual OLS standard errors and the heteroskedasticity-robust versions. Are there any important differences?

(ii) In the special case of the White test for heteroskedasticity, where we regress the squared OLS residuals on a quadratic in the OLS fitted values, $\hat{u}_i^2$ on $\hat{y}_i$, $\hat{y}_i^2$, $i = 1, \ldots, n$, argue that the probability limit of the coefficient on $\hat{y}_i$ should be one, the probability limit of the coefficient on $\hat{y}_i^2$ should be -1, and the probability limit of the intercept should be zero. {*Hint:* Remember that $\text{Var}(y|x_1, \ldots, x_k) = p(\mathbf{x})[1 - p(\mathbf{x})]$, where $p(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$.}

(iii) For the model estimated from part (i), obtain the White test and see if the coefficient estimates roughly correspond to the theoretical values described in part (ii).

(iv) After verifying that the fitted values from part (i) are all between zero and one, obtain the weighted least squares estimates of the linear probability model. Do they differ in important ways from the OLS estimates?

C11 Use the data in 401KSUBS for this question, restricting the sample to $fsize = 1$.

(i) To the model estimated in Table 8.1, add the interaction term, $e401k \cdot inc$. Estimate the equation by OLS and obtain the usual and robust standard errors. What do you conclude about the statistical significance of the interaction term?

(ii) Now estimate the more general model by WLS using the same weights, $1/inc_{\text{i}}$, as in Table 8.1. Compute the usual and robust standard error for the WLS estimator. Is the interaction term statistically significant using the robust standard error?

(iii) Discuss the WLS coefficient on *e401k* in the more general model. Is it of much interest by itself? Explain.

(iv) Reestimate the model by WLS but use the interaction term $e401k \cdot (inc - 30)$; the average income in the sample is about 29.44. Now interpret the coefficient on *e401k*.

C12 Use the data in MEAP00 to answer this question.

(i) Estimate the model

$$math4 = \beta_0 + \beta_1 lunch + \beta_2 \log(enroll) + \beta_3 \log(exppp) + u$$

by OLS and obtain the usual standard errors and the fully robust standard errors. How do they generally compare?

(ii) Apply the special case of the White test for heteroskedasticity. What is the value of the *F* test? What do you conclude?

(iii) Obtain $\hat{g}_i$ as the fitted values from the regression $\log(\hat{u}_i^2)$ on $\widehat{math4}_i$, $\widehat{math4}_i^2$, where $\widehat{math4}_i$ are the OLS fitted values and the $\hat{u}_i$ are the OLS residuals. Let $\hat{h}_i = \exp(\hat{g}_i)$. Use the $\hat{h}_i$ to obtain WLS estimates. Are there big differences with the OLS coefficients?

(iv) Obtain the standard errors for WLS that allow misspecification of the variance function. Do these differ much from the usual WLS standard errors?

(v) For estimating the effect of spending on *math*4, does OLS or WLS appear to be more precise?

C13 Use the data in FERTIL2 to answer this question.

(i) Estimate the model

$$children = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 educ + \beta_4 electric + \beta_5 urban + u$$

and report the usual and heteroskedasticity-robust standard errors. Are the robust standard errors always bigger than the nonrobust ones?

(ii) Add the three religious dummy variables and test whether they are jointly significant. What are the *p*-values for the nonrobust and robust tests?

(iii) From the regression in part (ii), obtain the fitted values $\hat{y}$ and the residuals, $\hat{u}$. Regress $\hat{u}^2$ on $\hat{y}$, $\hat{y}_2$ and test the joint significance of the two regressors. Conclude that heteroskedasticity is present in the equation for *children*.

(iv) Would you say the heteroskedasticity you found in part (iii) is practically important?

C14 Use the data in BEAUTY for this question.

(i) Using the data pooled for men and women, estimate the equation

$$lwage = \beta_0 + \beta_1 belavg + \beta_2 abvavg + \beta_3 female + \beta_4 educ + \beta_5 exper + \beta_5 exper^2 + u,$$

and report the results using heteroskedasticity-robust standard errors below coefficients. Are any of the coefficients surprising in either their signs or magnitudes? Is the coefficient on *female* practically large and statistically significant?

(ii) Add interactions of *female* with all other explanatory variables in the equation from part (i) (five interactions in all). Compute the usual *F* test of joint significance of the five interactions and a heteroskedasticity-robust version. Does using the heteroskedasticity-robust version change the outcome in any important way?

(iii) In the full model with interactions, determine whether those involving the looks variables—female • *belavg* and *female* • *abvavg*—are jointly significant. Are their coefficients practically small?