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

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:

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

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?

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*:

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

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.

8.6 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?

8.7 Consider a model at the employee level,

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

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 $\bar{u}_i = m_i^{-1}\sum_{e=1}^{m_i} u_{i,e}$ 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.

COMPUTER EXERCISES

C8.1 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.RAW 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?

C8.2 (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?

C8.3 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?

C8.4 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?

C8.5 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?

C8.6 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.

C8.7 Use the data in LOANAPP.RAW for this exercise.

(i) Estimate the equation in part (iii) of Computer Exercise C7.8, 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?

C8.8 Use the data set GPA1.RAW 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 $\hat{h}_i$.