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:

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

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*:

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

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,

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

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

(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)?

C8.9 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 $\check{u}_i = \hat{u}_i/\sqrt{\hat{h}_i}$ and $\check{y}_i = \hat{y}_i/\sqrt{\hat{h}_i}$ be the weighted quantities. Carry out the special case of the White test for heteroskedasticity by regressing $\check{u}_i^2$ on $\check{y}_i$, $\check{y}_i^2$, 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.

C8.10 Use the data set 401KSUBS.RAW 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?

C8.11 Use the data in 401KSUBS.RAW 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, $1/inc_i$, 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*.

C8.12 Use the data in MEAP00_01.RAW 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 *math4*, does OLS or WLS appear to be more precise?

KEY TERMS

Attenuation Bias
Average Partial Effect (APE)
Classical Errors-in-Variables (CEV)
Conditional Median
Davidson-MacKinnon Test
Endogenous Explanatory Variable
Endogenous Sample Selection
Exogenous Sample Selection
Functional Form Misspecification
Influential Observations
Lagged Dependent Variable
Least Absolute Deviations (LAD)
Measurement Error
Missing Data
Multiplicative Measurement Error
Nonnested Models
Nonrandom Sample
Outliers
Plug-In Solution to the Omitted Variables Problem
Proxy Variable
Random Coefficient (Slope) Model
Regression Specification Error Test (RESET)
Stratified Sampling
Studentized Residuals

PROBLEMS

9.1 In Problem 4.11, the R-squared from estimating the model

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + \beta_2 \log(mktval) + \beta_3 profmarg + \beta_4 ceoten + \beta_5 comten + u,$$

using the data in CEOSAL2.RAW, was $R^2 = .353$ ($n = 177$). When $ceoten^2$ and $comten^2$ are added, $R^2 = .375$. Is there evidence of functional form misspecification in this model?

9.2 Let us modify Computer Exercise C8.4 by using voting outcomes in 1990 for incumbents who were elected in 1988. Candidate A was elected in 1988 and was seeking reelection in 1990; *voteA90* is Candidate A's share of the two-party vote in 1990. The 1988 voting share of Candidate A is used as a proxy variable for quality of the candidate. All other variables are for the 1990 election. The following equations were estimated, using the data in VOTE2.RAW:

$$\widehat{voteA90} = 75.71 + .312\ prtystrA + 4.93\ democA$$
$$(9.25)\ \ (.046)\qquad\quad (1.01)$$
$$-.929 \log(expendA) - 1.950 \log(expendB)$$
$$(.684)\qquad\quad (.281)$$
$$n = 186, R^2 = .495, \bar{R}^2 = .483,$$

and

$$\widehat{voteA90} = 70.81 + .282\ prtystrA + 4.52\ democA$$
$$(10.01)\ \ (.052)\qquad\quad (1.06)$$
$$-.839 \log(expendA) - 1.846 \log(expendB) + .067\ voteA88$$
$$(.687)\qquad\quad (.292)\qquad\quad (.053)$$
$$n = 186, R^2 = .499, \bar{R}^2 = .485.$$

(i) Interpret the coefficient on *voteA88* and discuss its statistical significance.
(ii) Does adding *voteA88* have much effect on the other coefficients?

9.3 Let *math10* denote the percentage of students at a Michigan high school receiving a passing score on a standardized math test (see also Example 4.2). We are interested in estimating the effect of per student spending on math performance. A simple model is

$$math10 = \beta_0 + \beta_1 \log(expend) + \beta_2 \log(enroll) + \beta_3 poverty + u,$$

where *poverty* is the percentage of students living in poverty.

(i) The variable *lnchprg* is the percentage of students eligible for the federally funded school lunch program. Why is this a sensible proxy variable for *poverty*?

(ii) The table that follows contains OLS estimates, with and without *lnchprg* as an explanatory variable.

Dependent Variable: *math10*

Independent Variables	(1)	(2)
log(*expend*)	11.13 (3.30)	7.75 (3.04)
log(*enroll*)	.022 (.615)	−1.26 (.58)
lnchprg	—	−.324 (.036)
intercept	−69.24 (26.72)	−23.14 (24.99)
Observations *R*-squared	428 .0297	428 .1893

Explain why the effect of expenditures on *math10* is lower in column (2) than in column (1). Is the effect in column (2) still statistically greater than zero?

(iii) Does it appear that pass rates are lower at larger schools, other factors being equal? Explain.

(iv) Interpret the coefficient on *lnchprg* in column (2).

(v) What do you make of the substantial increase in R^2 from column (1) to column (2)?

9.4 The following equation explains weekly hours of television viewing by a child in terms of the child's age, mother's education, father's education, and number of siblings:

$$tvhours^* = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 motheduc + \beta_4 fatheduc + \beta_5 sibs + u.$$

We are worried that *tvhours** is measured with error in our survey. Let *tvhours* denote the reported hours of television viewing per week.

(i) What do the classical errors-in-variables (CEV) assumptions require in this application?

(ii) Do you think the CEV assumptions are likely to hold? Explain.

9.5 In Example 4.4, we estimated a model relating number of campus crimes to student enrollment for a sample of colleges. The sample we used was not a random sample of colleges in the United States, because many schools in 1992 did not report campus crimes. Do you think that college failure to report crimes can be viewed as exogenous sample selection? Explain.

9.6 In the model (9.17), show that OLS consistently estimates α and β if a_i is uncorrelated with x_i and b_i is uncorrelated with x_i and x_i^2, which are weaker assumptions than (9.19). [*Hint*: Write the equation as in (9.18) and recall from Chapter 5 that sufficient for consistency of OLS for the intercept and slope is $E(u_i) = 0$ and $Cov(x_i, u_i) = 0$.]

9.7 Consider the simple regression model with classical measurement error, $y = \beta_0 + \beta_1 x^* + u$, where we have m measures on x^*. Write these as $z_h = x^* + e_h$, $h = 1, \ldots, m$. Assume that x^* is uncorrelated with $u, e_1, \ldots, e_m$, that the measurement errors are pairwise uncorrelated, and have the same variance, σ_e^2. Let $w = (z_1 + \ldots + z_m)/m$ be the average of the measures on x^*, so that, for each observation i, $w_i = (z_{i1} + \ldots + z_{im})/m$ is the average of the m measures. Let $\bar{\beta}_1$ be the OLS estimator from the simple regression y_i on 1, w_i, $i = 1, \ldots, n$, using a random sample of data.

(i) Show that

$$\text{plim}(\bar{\beta}_1) = \beta_1 \left\{ \frac{\sigma_{x^*}^2}{[\sigma_{x^*}^2 + (\sigma_e^2/m)]} \right\}.$$

[*Hint*: The plim of $\bar{\beta}_1$ is $Cov(w, y)/Var(w)$.]

(ii) How does the inconsistency in $\bar{\beta}_1$ compare with that when only a single measure is available (that is, $m = 1$)? What happens as m grows? Comment.

COMPUTER EXERCISES

C9.1 (i) Apply RESET from equation (9.3) to the model estimated in Computer Exercise C7.5. Is there evidence of functional form misspecification in the equation?

(ii) Compute a heteroskedasticity-robust form of RESET. Does your conclusion from part (i) change?

C9.2 Use the data set WAGE2.RAW for this exercise.

(i) Use the variable *KWW* (the "knowledge of the world of work" test score) as a proxy for ability in place of *IQ* in Example 9.3. What is the estimated return to education in this case?

(ii) Now, use *IQ* and *KWW* together as proxy variables. What happens to the estimated return to education?

(iii) In part (ii), are *IQ* and *KWW* individually significant? Are they jointly significant?

C9.3 Use the data from JTRAIN.RAW for this exercise.

(i) Consider the simple regression model

$$\log(scrap) = \beta_0 + \beta_1 grant + u,$$

where *scrap* is the firm scrap rate and *grant* is a dummy variable indicating whether a firm received a job training grant. Can you think of some reasons why the unobserved factors in u might be correlated with *grant*?

(ii) Estimate the simple regression model using the data for 1988. (You should have 54 observations.) Does receiving a job training grant significantly lower a firm's scrap rate?

(iii) Now, add as an explanatory variable $\log(scrap_{87})$. How does this change the estimated effect of *grant*? Interpret the coefficient on *grant*. Is it statistically significant at the 5% level against the one-sided alternative $H_1: \beta_{grant} < 0$?

(iv) Test the null hypothesis that the parameter on $\log(scrap_{87})$ is one against the two-sided alternative. Report the *p*-value for the test.

(v) Repeat parts (iii) and (iv), using heteroskedasticity-robust standard errors, and briefly discuss any notable differences.

C9.4 Use the data for the year 1990 in INFMRT.RAW for this exercise.

(i) Reestimate equation (9.43), but now include a dummy variable for the observation on the District of Columbia (called *DC*). Interpret the coefficient on *DC* and comment on its size and significance.

(ii) Compare the estimates and standard errors from part (i) with those from equation (9.44). What do you conclude about including a dummy variable for a single observation?

C9.5 Use the data in RDCHEM.RAW to further examine the effects of outliers on OLS estimates and to see how LAD is less sensitive to outliers. The model is

$$rdintens = \beta_0 + \beta_1 sales + \beta_2 sales^2 + \beta_3 profmarg + u,$$

where you should first change *sales* to be in billions of dollars to make the estimates easier to interpret.

(i) Estimate the above equation by OLS, both with and without the firm having annual sales of almost $40 billion. Discuss any notable differences in the estimated coefficients.

(ii) Estimate the same equation by LAD, again with and without the largest firm. Discuss any important differences in estimated coefficients.

(iii) Based on your findings in (i) and (ii), would you say OLS or LAD is more resilient to outliers?

C9.6 Redo Example 4.10 by dropping schools where teacher benefits are less than 1% of salary.

(i) How many observations are lost?

(ii) Does dropping these observations have any important effects on the estimated tradeoff?

C9.7 Use the data in LOANAPP.RAW for this exercise.

(i) How many observations have $obrat > 40$, that is, other debt obligations more than 40% of total income?

(ii) Reestimate the model in part (iii) of Computer Exercise C7.8, excluding observations with $obrat > 40$. What happens to the estimate and *t* statistic on *white*?

(iii) Does it appear that the estimate of β_{white} is overly sensitive to the sample used?

C9.8 Use the data in TWOYEAR.RAW for this exercise.

(i) The variable *stotal* is a standardized test variable, which can act as a proxy variable for unobserved ability. Find the sample mean and standard deviation of *stotal*.

(ii) Run simple regressions of *jc* and *univ* on *stotal*. Are both college education variables statistically related to *stotal*? Explain.

(iii) Add *stotal* to equation (4.17) and test the hypothesis that the returns to two- and four-year colleges are the same against the alternative that the return to four-year colleges is greater. How do your findings compare with those from Section 4.4?

(iv) Add $stotal^2$ to the equation estimated in part (iii). Does a quadratic in the test score variable seem necessary?

(v) Add the interaction terms *stotal·jc* and *stotal·univ* to the equation from part (iii). Are these terms jointly significant?

(vi) What would be your final model that controls for ability through the use of *stotal*? Justify your answer.

C9.9 In this exercise, you are to compare OLS and LAD estimates of the effects of 401(k) plan eligibility on net financial assets. The model is

$$nettfa = \beta_0 + \beta_1 inc + \beta_2 inc^2 + \beta_3 age + \beta_4 age^2 + \beta_5 male + \beta_6 e401k + u.$$

(i) Use the data in 401KSUBS.RAW to estimate the equation by OLS and report the results in the usual form. Interpret the coefficient on *e401k*.

(ii) Use the OLS residuals to test for heteroskedasticity using the Breusch-Pagan test. Is u independent of the explanatory variables?

(iii) Estimate the equation by LAD and report the results in the same form as for OLS. Interpret the LAD estimate of β_6.

(iv) Reconcile your findings from parts (i) and (iii).

C9.10 You need to use two data sets for this exercise, JTRAIN2.RAW and JTRAIN3.RAW. The former is the outcome of a job training experiment. The file JTRAIN3.RAW contains observational data, where individuals themselves largely determine whether they participate in job training. The data sets cover the same time period.

(i) In the data set JTRAIN2.RAW, what fraction of the men received job training? What is the fraction in JTRAIN3.RAW? Why do you think there is such a big difference?

(ii) Using JTRAIN2.RAW, run a simple regression of *re78* on *train*. What is the estimated effect of participating in job training on real earnings?

(iii) Now add as controls to the regression in part (ii) the variables *re74*, *re75*, *educ*, *age*, *black*, and *hisp*. Does the estimated effect of job training on *re78* change much? How come? (*Hint:* Remember that these are experimental data.)

(iv) Do the regressions in parts (ii) and (iii) using the data in JTRAIN3.RAW, reporting only the estimated coefficients on *train*, along with their t statistics. What is the effect now of controlling for the extra factors, and why?

(v) Define $avgre = (re74 + re75)/2$. Find the sample averages, standard deviations, and minimum and maximum values in the two data sets. Are these data sets representative of the same populations in 1978?

(vi) Almost 96% of men in the data set JTRAIN2.RAW have *avgre* less than \$10,000. Using only these men, run the regression

re78 on *train,re74,re75,educ,age,black,hisp*

and report the training estimate and its t statistic. Run the same regression for JTRAIN3.RAW, using only men with $avgre \leq 10$. For the subsample of low-income men, how do the estimated training effects compare across the experimental and nonexperimental data sets?

(vii) Now use each data set to run the simple regression *re78* on *train*, but only for men who were unemployed in 1974 and 1975. How do the training estimates compare now?

(viii) Using your findings from the previous regressions, discuss the potential importance of having comparable populations underlying comparisons of experimental and nonexperimental estimates.

C9.11 Use the data for the year 1993 for this question, although you will need to first obtain the lagged murder rate, say $mrdrte_{-1}$.

(i) Run the regression of *mrdrte* on *exec*, *unem*. What are the coefficient and *t* statistic on *exec*? Does this regression provide any evidence for a deterrent effect of capital punishment?

(ii) How many executions are reported for Texas during 1993? (Actually, this is the sum of executions for the current and past two years.) How does this compare with the other states? Add a dummy variable for Texas to the regression in part (i). Is its *t* statistic unusually large? From this, does it appear Texas is an "outlier"?

(iii) To the regression in part (i) add the lagged murder rate. What happens to $\hat{\beta}_{exec}$ and its statistical significance?

(iv) For the regression in part (iii), does it appear Texas is an outlier? What is the effect on $\hat{\beta}_{exec}$ from dropping Texas from the regression?

C9.12 Use the data in ELEM94_95 to answer this question. See also Computer Exercise C4.10.

(i) Using all of the data, run the regression *lavgsal* on *bs*, *lenrol*, *lstaff*, and *lunch*. Report the coefficient on *bs* along with its usual and heteroskedasticity-robust standard errors. What do you conclude about the economic and statistical significance of $\hat{\beta}_{bs}$?

(ii) Now drop the four observations with $bs > .5$, that is, where average benefits are (supposedly) more than 50% of average salary. What is the coefficient on *bs*? Is it statistically significant using the heteroskedasticity-robust standard error?

(iii) Verify that the four observations with $bs > .5$ are 68, 1,127, 1,508, and 1,670. Define four dummy variables for each of these observations. (You might call them *d*68, *d*1127, *d*1508, and *d*1670.) Add these to the regression from part (i), and verify that the OLS coefficients and standard errors on the other variables are identical to those in part (ii). Which of the four dummies has a *t* statistic statistically different from zero at the 5% level?

(iv) Verify that, in this data set, the data point with the largest studentized residual (largest *t* statistic on the dummy variable) in part (iii) has a large influence on the OLS estimates. (That is, run OLS using all observations except the one with the large studentized residual.) Does dropping, in turn, each of the other observations with $bs > .5$ have important effects?

(v) What do you conclude about the sensitivity of OLS to a single observation, even with a large sample size?

(vi) Verify that the LAD estimator is not sensitive to the inclusion of the observation identified in part (iii).