Key Terms

Attenuation Bias
Average Marginal Effect
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

In Problem 11 in Chapter 4, 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?

Let us modify Computer Exercise C4 in Chapter 8 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:

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

and

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

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

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.

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

The following equation explains weekly hours a child plays video games, *gamehours*, in terms of child's age, mother's education, father's education, and number of siblings:

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

We are worried that *gamehours** is measured with error in our survey. The *gamehours* denote the reported hours of video game playing per week.
(i) Explain what the classical errors-in-variables (CEV) setup entails in this application.
(ii) Do you think the CEV assumptions are likely to hold? Explain.

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.

Computer Exercises

(i) Apply RESET from equation (9.3) to the model estimated in Computer Exercise C5 in Chapter 7. 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?

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?

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.

Use the data for the year 1990 in INFMRT.RAW for this exercise.
(i) Reestimate equation (9.37), 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 (9.38). What do you conclude about including a dummy variable for a single observation?

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?

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?

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 C8 in Chapter 7, 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?

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.

Use the data in 401KSUBS.RAW to answer this question, using only single people ($fsize = 1$). The equation we are interested in is

$$nettfa = \beta_0 + \beta_1 inc + \beta_2 (age - 25)^2 + \beta_4 male + \beta_5 e401k + u.$$

(i) Estimate the equation by OLS, report the results in the usual form, and interpret the coefficient on *e401k*.

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

(iii) Estimate the equation by LAD, and report the results in the same form as for OLS. Interpret the coefficient on *e401k*.

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

11. 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.