Department of Economics | Economics 120B
Winter 2008 | Prof. Berman

Problem Set #4
due Thursday, March 13 at the beginning of class

Note: This problem set should help you study for the final.
Your problem set grade for the course will be the average of the top three problem set grades, so you may choose to skip this one.

1. A few years ago UCSD economics student Lisa Ton was interested in estimating the effects of class attendance on midterm test scores. She surveyed students and estimated the following equation using 120 observations from an Economics 120B class.

Test_score = $\beta_0 + \beta_1$num_classes + u

test_score	Coefficient	Robust Std. Err.	t
num_classes	1.452607	.5980065	2.43
Constant	64.09062	4.580781	13.99

a) What omitted variables might bias the estimate of coefficient β_1 , if we want to think of it as the causal effect of number of classes attended on test scores?

b) Explain why distance lived from classroom is likely to be correlated with num_classes.

c) Would that make distance a good instrumental variable for num_classes in Lisa's regression? (Discuss relevance and validity.)

d) Imagine that dorms were randomly assigned to UCSD students. Would that change your answer to question c)?
If so, how?

2. Alan Krueger and Orley Ashenfelter of Princeton collected data on twin pairs. The data contain information on the log of hourly earnings, and educational attainment. Krueger and Ashenfelter collected both the respondents' and the twins' report on his educational attainment. By comparing these two measures of the same quantity, we can get some idea of how important measurement error in education is. A subset of the data, including 199 twin pairs, has been kindly provided by Krueger and Ashenfelter and is available on the course site as twins.dta .

Variables on the file, in order, are: w - hourly wages, id - twin pair identification number, ed - years of education (self reported), ed2 - years of education as reported by sibling,

a) Construct a variable containing the natural logarithm of wages. [gen lw = log(w)]

b) Regress log wages on education for each of the 398 individuals. Report the coefficient and standard error.

c) Now regress log wages on education, but to eliminate measurement error bias, instrument the twins report of his educational attainment (ed) with the report of the same quantity by the twin's sibling (ed2). Report the coefficient and standard error.

What must you assume for the IV estimate to yield a consistent estimate of the effect of education on wages?

Under those assumptions, what is the ratio of measurement error variance to total variance in (self reported) educational attainment implied by the comparison of your results in parts (b) and (c)?

d) In order to deal with omitted variable bias due to twin effects, perhaps due to unobserved ability), difference the data within pairs. [To difference the data, sort so that twin pairs are adjacent and create a variable dx= x - x[_n-1] if id= = id[_n-1], and construct dlw, ded, ded2. You should now have 199 observations representing twin pairs.]

Using the differenced data, regress dlw on ded. Report the coefficient and the standard error.

What must you assume about measurement error and about unobserved ability in order to conclude that this coefficient is a consistent estimate of the effect of education on wages?

e) Now regress dlw on ded, but instrument with ded2 in order to undo possible measurement error. State the necessary assumptions about unobserved ability and about measurement error necessary to conclude that the coefficient is a consistent estimate of the effect of education on wages.

Under those assumptions, is the ratio of measurement error variance to total variance larger in the differenced variable educational attainment variable (ded) or in the original variable (ed)?

Does the evidence suggest that differencing exacerbated (i.e., worstened) measurement error bias?

(Bonus) f) What, if anything, have you learned about the causal effect of education on wages from this exercise? (Answer on the reverse side).