

Problem Set 6
(due at the beginning of class December 4, 2002)

Part I. For each of the following questions, choose the one best answer. Briefly explain your reasoning.

The first 5 questions are based on the following information: Suppose a researcher is interested in the effect of class attendance on college performance, and plans to estimate the following model: $colGPA = \beta_0 + \beta_1 hsGPA + \beta_2 ACT + \beta_3 skipped + u$, where $colGPA$ is current GPA, $hsGPA$ is high school GPA, ACT is score on a college entrance exam and $skipped$ is the average number of classes skipped per week. The researcher believes that a component of u is the student's inherent laziness.

1. OLS estimates of this model will most likely
 - a) be biased and inconsistent, because skipped is endogenous
 - b) be biased because skipped is endogenous, but still consistent
 - c) be unbiased and consistent, because it satisfies the Gauss-Markov assumptions
 - d) be unbiased, but possibly inconsistent

2. The researcher has information on the distance in miles students live from class ($dist$) and whether they have any classes at 8am ($early$), and regresses $skipped$ on $dist$, $early$, $hsGPA$, and ACT . He then saves the residuals, $uhat$, from this regression. If he is planning on doing IV, he should
 - a) test for the joint significance of all the exogenous variables
 - b) test for the joint significance of $dist$ and $early$
 - c) check the reported overall F-test for significance
 - d) regress $uhat$ on all the exogenous variables

3. The researcher next obtains the following estimates:

Source	SS	df	MS			
Model	4.53802712	4	1.13450678	Number of obs =	141	
Residual	14.8680723	136	.109324061	F(4, 136) =	10.38	
Total	19.4060994	140	.138614996	Prob > F =	0.0000	
				R-squared =	0.2338	
				Adj R-squared =	0.2113	
				Root MSE =	.33064	

colGPA	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
hsGPA	.4135611	.0943636	4.383	0.000	.2269514	.6001708
ACT	.0144984	.0106536	1.361	0.176	-.0065698	.0355666
skipped	-.0796302	.0308488	-2.581	0.011	-.1406356	-.0186249
uhat	-.0122316	.0578108	-0.212	0.833	-.1265559	.1020928
_cons	1.385228	.3333431	4.156	0.000	.7260219	2.044434

We can conclude that:

- a) all else equal, each skipped class reduces GPA by about .08 points
- b) the OLS estimates are consistent
- c) the IV estimates are not significantly different from the OLS estimates
- d) all of the above

4. The researcher estimates the model using IV, saves the residuals (*uhativ*) and then obtains:

Source	SS	df	MS			
Model	.09339178	4	.023347945	Number of obs = 141		
Residual	14.7815227	136	.108687667	F(4, 136) = 0.21		
Total	14.8749145	140	.106249389	Prob > F = 0.9298		
				R-squared = 0.0063		
				Adj R-squared = -0.0229		
				Root MSE = .32968		

<i>uhativ</i>	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
hsGPA	.003987	.0935295	0.043	0.966	-.1809732	.1889472
ACT	.0003169	.0104884	0.030	0.976	-.0204246	.0210584
dist	.0163524	.0185409	0.882	0.379	-.0203132	.0530181
early	-.111073	.1233414	-0.901	0.369	-.3549881	.1328422
_cons	-.0592899	.3391357	-0.175	0.861	-.7299514	.6113716

The above estimates would imply that

- a) we cannot reject the null of homoskedasticity in the errors
 - b) the instruments are endogenous
 - c) students probably can't completely choose where to live and whether to have 8am classes or not
 - d) none of the above
5. Turning to the IV estimates the researcher must have obtained in question 4, we can predict that
- a) the coefficient on *skipped* will not be systematically different from -.0796302
 - b) the coefficient on *skipped* will definitely be higher than -.0796302
 - c) the coefficient on *skipped* will definitely be lower than -.0796302
 - d) the coefficient on *skipped* will definitely be exactly -.0796302
6. Which of the following are true about time-series estimation?
- a) As with cross-section estimation, we can assume the observations are independent
 - b) A trending variable cannot be used as the dependent variable in multiple regression
 - c) Seasonality is not an issue when using annual time series observations
 - d) all of the above

Part II. Stata Problems. See "Important Things to Know about Stata" on our website for more about installing and using Stata. This PS uses the Stata commands **insheet**, **desc**, **sum**, **reg** (with the **robust** option and also with the **() syntax** for IV), **graph**, **gen**, **tsset**, **newey**, **prais** and **test**. All data sets are available through our web page.

1. This problem uses the spreadsheet **smoke.xls** which contains information on individual's age, race, education, income, and cigarette smoking, as well as information on cigarette prices in the individual's state and whether this state has restrictions on smoking in restaurants. You will need to save the spreadsheet as a text file and read it into Stata.

- a) A researcher has read the article "More Bad News for Smokers" and wants to use these data to see if he finds an effect of smoking on income. He thinks income is a function of not just cigarettes smoked, but also education and demographics like race and age. He also thinks any positive effect of age will decline over time and that it would be a good idea to use the log of income. Estimate this implied model, allowing for the possibility of heteroskedasticity.

b) The researcher thinks more about the problem and decides that he should really consider income and cigarette smoking to be jointly determined. That is, he thinks that cigarettes smoked are a function of $\ln(\text{income})$, as well as education, demographics, cigarette prices and whether there are restaurant smoking restrictions. Estimate the reduced form models for both cigarettes smoked and $\ln(\text{income})$.

c) Use IV to estimate the structural models (if possible). Discuss what variables are being used as instruments if you can estimate a model, or why you can't estimate a model, if it is impossible.

2. This problem uses the Stata data set **consump.dta** which contains annual macroeconomic data from 1959 to 1995. It assumes a researcher is interested in the relationship between aggregate income and consumption. Be sure to look carefully at the variable labels and means before starting.

a) Graph the logs of real per capita disposable income and real per capita consumption on a time trend. Do the series appear to be related? Use a simple regression to estimate the elasticity of consumption with respect to income.

b) Regress each of the logs of real per capita disposable income and real per capita consumption on a time trend. What can you conclude about the growth of each series? Use these regressions to produce detrended series. Graph the detrended series on a time trend. Do the series still appear to be related?

c) Use a simple regression with these detrended data to estimate the elasticity of consumption with respect to income. Use the data that has not been detrended to obtain this same elasticity estimate. How do these estimates compare with the elasticity estimated in a)?

d) Re-estimate the last model calculating Newey-West standard errors that are serial correlation-robust. (Note: Let g be equal to at least 3 or 4.) Compare this with using Cochrane-Orcutt estimation, and compare both with the OLS estimates from c). Are things as you would expect? Why or why not?

3. This problem uses the Stata data set **intdef.dta** that includes data on inflation, interest rates and the deficit. It assumes a researcher is interested in the effect of interest rates and deficits on inflation. Be sure to look carefully at the variable labels and means before starting.

a) Estimate a finite distributed lag model (of order 1) of the effect of inflation **and** the deficit on the interest rate. What is the impact propensity for inflation? What about for the deficit? Explain your answer.

b) What is the long-run propensity for inflation? What about for the deficit? In each case, test the null hypothesis for whether the long-run propensity is significantly different from zero. Discuss your results.