

**Problem Set 4**

(due at the beginning of class November 13, 2002)

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

**1, 2, 3 and 4.** Consider the following population model for household consumption, which has been estimated below:

$$\text{consumption} = \beta_0 + \beta_1 \text{income} + \beta_2 \text{education} + \beta_3 \text{household size} + u$$

cons	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
inc	.8918447	.0651406	13.691	0.000	.7625415 1.021148
educ	-143.3155	105.3268	-1.361	0.177	-352.3876 65.7565
size	-65.86678	216.6922	-0.304	0.762	-495.9974 364.2638
_cons	1438.803	1477.885	0.974	0.333	-1494.775 4372.38

1. Suppose that our variable for consumption is measured with error, so  $\text{cons} = \text{consumption} + e_0$ , where  $e_0$  is uncorrelated with  $\text{inc}$ ,  $\text{educ}$  and  $\text{size}$ . We would expect that:

- our estimates of  $\beta_0$ ,  $\beta_1$ ,  $\beta_2$ , and  $\beta_3$  will all definitely be biased
- our estimates of  $\beta_0$ ,  $\beta_1$ ,  $\beta_2$ , and  $\beta_3$  will all definitely be unbiased
- all our standard errors will be bigger than they would be without the measurement error
- both b) and c)

2. Now suppose that consumption is fine, but that our variable for income is measured with error, so  $\text{inc} = \text{income} + e_1$ , where  $e_1$  is uncorrelated with **true** income,  $\text{educ}$  and  $\text{size}$ . Assuming that **true** income is uncorrelated with  $\text{educ}$  and  $\text{size}$ , we would expect that:

- our estimate of  $\beta_1$  will likely be less than .89
- our estimates of  $\beta_2$  and  $\beta_3$  will likely be around -143 and -66 respectively
- all our standard error will be bigger than they would be without measurement error
- all of the above

3. Suppose the researcher gets the predicted value,  $\hat{y}$ , and then runs a regression of  $\hat{y}$  on  $\text{inc}$ ,  $\text{educ}$  and  $\text{size}$ . We can

- be absolutely certain that the  $R^2$  is equal to 1
- be absolutely certain that the  $R^2$  is equal to 0
- use the overall F-test to determine if our original model had the correct specification
- none of the above

4. Suppose the data was collected through a telephone survey, and the last 4 digits of the household's telephone number was accidentally named  $\text{assets}$  and included in the regression. We would expect

- our estimates of  $\beta_1$ ,  $\beta_2$ , and  $\beta_3$  to stay about the same, but with larger standard errors
- our estimates of  $\beta_1$ ,  $\beta_2$ , and  $\beta_3$  and their standard errors to stay about the same
- our estimates of  $\beta_1$ ,  $\beta_2$ , and  $\beta_3$  to move closer to zero, with larger standard errors
- our estimates of  $\beta_1$ ,  $\beta_2$ , and  $\beta_3$  to move closer to zero, but the standard errors will be about the same

5. Suppose the following population model meets the Gauss-Markov assumptions:

$\text{pct of students passing a reading exam} = \beta_0 + \beta_1 \ln(\text{expenditures}) + \beta_2 \ln(\text{enrollment}) + \beta_3 \text{ pct of students living in poverty} + u$ , but we do not have data on the percentage of students living in poverty and plan to use the percentage of students eligible for the federally funded school lunch program as a proxy. Our estimates of  $\beta_1$  and  $\beta_2$  will be consistent if

- pct of students living in poverty* is closely related to *pct of students eligible for the lunch program*
- $u$  is uncorrelated with *pct of students eligible for the lunch program*
- the difference between *pct of students living in poverty* and *pct of students eligible for the lunch program* is unrelated to  $\ln(\text{expenditures})$ ,  $\ln(\text{enrollment})$  and *pct of students living in poverty*
- all of the above must be true

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 **desc**, **sum**, **gen**, **reg** (and **reg** with the **robust** option), and **dprobit**. All data sets are available through our web page.

1. This problem uses *youth.dta*, which is a sample of high school students. It assumes that a researcher is concerned about youth suicide, in particular the relationship with drug use and sexual activity. Take a look at what variables are in the data set and what their means are before getting started.

a) Estimate a linear probability model for **considering** suicide as a function of basic demographics (i.e. age, race, sex), sexual partners and drug use (consider alcohol and tobacco to be drugs). Be sure to allow for heteroskedasticity. Considering all else equal: Are males or females most likely to consider suicide? Older or younger students? What racial/ethnic group is least likely to consider suicide? What is the effect of each additional sexual partner? Discuss the role of drug use. For each observation, predict the probability of having considered suicide. Are there any problems with these predictions?

b) Re-estimate the model from part a) using a probit, focusing on the marginal effects. Are there any meaningful differences in the implications of this model compared to that from part a)? If so, discuss these differences. For each observation, predict the probability of having considered suicide. How do these predictions differ, at all, from those of part a)?

c) Add *grade* and *partners*<sup>2</sup> to the model from b). How would you interpret the coefficients on *grade* and *ageyrs* now? How would you describe the effect of additional sexual partners?

2. This problem uses *vote1.dta*, which is a sample of congressional election outcomes. It assumes that with the election over (and campaign ads thankfully off of the airwaves!), a researcher wants to study whether candidates really can “buy votes.” Take a look at what variables are in the data set and what their means are before getting started.

a) The researcher is certain that he wants to model the percent of the vote received by candidate A as a function of which party the candidate belongs to, party strength, and spending. He cannot decide whether spending is best represented by including the log of each candidates spending separately, or simply including candidate A’s share of total spending. Estimate each of these implied models.

b) Can a Davidson-MacKinnon test reject either model? If so, which one? If not, or if both are rejected, how else might you decide between the two specifications?