

Political Science 239

Problem Set 2

Due date: Wednesday, September 20, 2006

In this problem set, we'll analyze both theoretical and applied questions. For the last two questions, use the house improvement data ("ME_households.csv") that we used in Problem Set 1.

Exercise 1 *20% of people in a town usually vote for the Green Party. 30% of the people in town are unhappy with the current government. But 80% of the people who usually vote for the Green Party are unhappy with the current Government. If you know that your neighbor is not happy with the current government, what is the probability that she usually votes for the Green Party?*

Exercise 2 *Suppose there are two types of social scientists who use quantitative methods. Type A understand the techniques they use and type B do not. Assume that 5 percent of social scientists who use quantitative methods are type A and 95 percent of them are type B (there is no overlap). Also assume that the mean social scientist classifies 10 percent of social scientists who are actually type A as type B. Also assume that the mean social scientist classifies 80 percent of social scientists who are actually type B as type A.*

(2.1) *What is the probability that the mean social scientist will classify a randomly chosen social scientist who uses quantitative methods as type A?*

(2.2) *What is the probability that the mean social scientist will classify a randomly chosen social scientist who uses quantitative methods as type B?*

Exercise 3 *A researcher wants to figure out once and for all if giving students high school vouchers (for private schools) would significantly improve the educational outcomes (as measured by test scores) of the students. The researcher, with the help of the government, is able to conduct an almost perfect randomized experiment. This researcher's experiment does not suffer from any of the standard problems that plague this field (including the experiments we discussed in class). In particular, there are no compliance issues—e.g., everyone who gets a voucher goes to private school and everyone who does not get a voucher does not go to private school. The researcher then estimates the following model:*

$$Y = \alpha + \beta_1 \text{voucher} + \beta_2 \text{race} + \beta_3 \text{parents.education} + \beta_4 \text{parents.income} + \varepsilon$$

where Y is the test score a student has at the end of the experiment, and "voucher" is a dummy variable for whether a student received a voucher or not. The other variables are "controls", which measure other things which are related to test scores in particular and to educational outcomes in general. The researcher finds that β_1 is significant and sends her work in for publication. A reviewer really doesn't like our researcher's model. The reviewer argues: since the experiment has no compliance problems, and because it is an experiment, the researcher should only estimate the following model:

$$Y = \alpha + \beta_1 \text{voucher} + \varepsilon$$

Our researcher estimates the model that the reviewer has recommended, but finds that the estimate for β_1 is no longer significant (but has the same sign as before). Answer the following questions:

(3.1) *What is the reviewer's rationale?*

(3.2) *Why do the results differ between the two models? Hint: this is an experiment so some of the problems that observational studies face are not relevant. Prove your answer.*

(3.3) *Which model should we prefer? Why?*

Exercise 4 Using the dataset "ME_households.csv", compare the means of the log of household consumption, the log of household assets, and the title dummy for those households whose head has less than 8 years of education with the means of the same variables for those households whose head has 8 or more years of education. Do you see any differences? Now repeat the same analysis, but this time separating both education groups by treatments and controls. Are these characteristics similar between treatment and controls across education groups?

Exercise 5 In this exercise, we will estimate a linear regression model for each house infrastructure outcome and each mental health outcome on the treatment dummy. In order to run these regressions, you are asked to create your own regression function. This means that you should not use R regression functions such as "lm()" or "glm()" to perform your estimations. However, you may use other R functions such as "matrix()", "t()", "solve()", etc., within the regression function that you create. Similarly to Problem Set 1, run two regressions for each outcome variable: one with no covariates and the other with household head's age and education, spouse's age and education, household size, log of consumption, log of assets and the title dummy as covariates. Include an intercept in all your regressions. Report estimated coefficients and standard errors. Are your estimates similar to the ones you obtained in Problem Set 1? [HINT: See accompanying R code and accompanying OLS hints on the website. Also, see "R code 2" on the website for hints about functions that perform matrix operations]