

Political Science 239

Problem Set 1

Due date: Wednesday, September 13, 2006

In this problem set, we'll analyze the impact of a house improvement program in Mexico on house infrastructure measures, mental health measures, and income. The program was implemented in many states of Mexico, and it gave poor households different materials (cement, bricks, etc.) to improve the quality of their houses. After receiving the materials, each household would provide its own labor to perform the improvements. There were no restrictions on the type of improvements that the household could make.

The dataset "ME_households.csv" contains a subset of a sample of Mexican households that received the program in the state of Coahuila (you are receiving only a subset of the original sample to preserve the original information). The treatments are households that applied and obtained the benefit in 2003. The controls are households from the same municipalities that applied for the program in 2004 but had not obtained the benefit at the moment of the survey. The file "ProblemSet1_Variables.xls" describes the variables in the dataset.

Exercise 1 *Identify the covariates that have missing values, and replace those missing values by the median of each covariate. If the median is not an integer and the variable takes only integer values, round the median to the largest integer not greater than the median.*

Exercise 2 *Report the minimum, maximum, mean and standard deviation of all covariates in the dataset, separated by treatments and controls. Are treatment and control similar in terms of these*

characteristics? What can you say about the demographic structure, economic status and sanitary conditions of both types of households? [Hint: use R function **apply**]

Exercise 3 For every outcome variable, report its mean separately for treatments and controls. Calculate the means excluding missing values. What differences do you see in these outcomes between treatment and controls? [Hint: use R function **apply**]

Exercise 4 Estimate a linear regression model for each house infrastructure, household income and mental health measure on the treatment dummy. For every outcome, run two regressions: 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, standard errors and p-values. Does the program have any impact on income? What about house infrastructure and mental health? Are these results similar to the ones you reported in Exercise 3? Why/Why not?

Exercise 5 In this question, we will analyze the residuals of the regression of the number of rooms. Use the output of the regression of the number of rooms on the treatment dummy and the covariates calculated in Exercise 4. First, plot an histogram of the residuals of this regression. Second, graph the residuals against the predicted values. Finally, plot a normal QQ-plot of the residuals, including a line that shows how this QQ-plot would look like if the residuals followed a normal distribution. From this graphical analysis, what can you conclude about the normality of the residuals of this regression? [Hint: use the R commands **hist**, **plot**, **qqplot** and **qqline**].