Econ 140 - Spring 2016
Section 12

GSI: Fenella Carpena

April 28, 2016

1 Experiments and Quasi-Experiments

Exercise 1.0. Consider the STAR Experiment discussed in lecture where students were randomly assigned to one of three groups: small class size, regular class size, or regular class size with a teacher's aide. Using data collected from the experiment, we estimate the regression:

$$Y_i = \beta_0 + \beta_1 SmallClass_i + \beta_2 RegAide_i + u_i$$

where Y_i is test score for student i, $SmallClass_i = 1$ if student i is in a small class and 0 otherwise, and $RegAide_i = 1$ if student i is in a regular class with an aide and 0 otherwise. How would you expect including $TeacherExp$ (teacher's years of experience) as an additional variable would alter the coefficient on $SmallClass$: would it increase, decrease, or stay the same? Explain.

> We would expect the coefficient on $SmallClass$ to stay the same because $SmallClass$ was randomly assigned. Hence, we would not expect the variable $TeacherExp$ to be correlated with $SmallClass$, so it would not be causing omitted variable bias.
>
> However, if it was the case that the more experienced teachers also taught the small classes (for example, because the principal set it up that way), then $cov(TeacherExp, SmallClass) > 0$. Combined with the fact that $TeacherExp$ probably has a positive effect on test scores, this would mean that omitting $TeacherExp$ from the regression results in an upward bias, and when it is added in the regression, we would expect the coefficient β_1 to decrease.

Exercise 1.1. (Stock & Watson, Review the Concepts 13.1) A researcher studying the effects of a new fertilizer on crop yields plans to carry out an experiment in which different amounts of fertilizer are applied to 100 different 1-acre parcels of land. There will be four treatment levels. Treatment level 1 has no fertilizer, treatment level 2 is 50% of the manufacturer's recommended amount of fertilizer, treatment level 3 is 100%, and treatment level 4 is 150%. The researcher plans to apply treatment 1 to the first 25 parcels of land, treatment level 2 to the second 25 parcels, and so forth. Can you suggest a better way to assign treatment levels? Why is your proposal better than the researcher's method?

> It would be better to assign the treatment level randomly to each parcel. The research plan outlined in the problem may be flawed because the different groups of parcels might differ systematically. For example, the first 25 parcels of land might have poorer drainage than the other parcels and this would lead to lower crop yields. The treatment assignment outlined in the problem would place these 25 parcels in the control group, thereby overestimating the effect of the fertilizer on crop yields. This problem is avoided with random assignment of treatments.

Exercise 1.2. (Stock & Watson, Review the Concepts 13.2) A clinical trial is carried out for new cholesterol-lowering drug. The drug is given to 500 patients, and a placebo is given to another 500 patients, using random assignment of the patients.

(a) How would you estimate the treatment effect of the drug?

(b) Suppose that you had data on the weight, age and gender of each patient. Could you use these data to improve your estimate? Explain.

(a) The treatment effect could be estimated as the difference in average cholesterol levels for the treated group and the untreated (control) group. We could also estimate a regression of cholesterol on a dummy variable for the treatment group, which will give us the same estimate.

(b) Data on the weight, age, and gender of each patient could be used to improve the estimate using the differences estimator with additional regressors. This regression may produce a more accurate estimate because it controls for these additional factors that may affect cholesterol.

Exercise 1.3. (Stock & Watson, Review the Concepts 13.3) Researchers studying the STAR data report anecdotal evidence that school principals were pressured by some parents to place their children in the small classes.

(a) Suppose that some principals succumbed to their children in the small classes. How would such transfers compromise the internal validity of the study?

(b) Suppose that you had data on the original random assignment of each student before the principal's intervention. How could you use this information to restore the internal validity of the study?

(a) If the students who were transferred to small classes differed systematically from the other students, then internal validity is compromised. For example, if the transferred students tended to have higher incomes and more learning opportunities outside of school, then they would tend to perform better on standardized tests. The experiment would incorrectly attribute this performance to the smaller class size.

(b) Information on original random assignment could be used as an instrument to restore internal validity. The original random assignment is a valid instrument because it is exogenous by virtue of random assignment (uncorrelated with the regression error) and is relevant (correlated with the actual assignment).

Exercise 1.4. (Adapted from Stock & Watson, Exercise 13.4) Going back to the Card and Krueger (1994) example, consider the difference-in-difference regression:

$$emp_{it} = \beta_0 + \beta_1 NJ_i + \beta_2 POST_t + \beta_3 NJ_i * POST_t + u_{it}$$

(a) In terms of coefficients β_0, β_1, β_2, β_3, what is the expected number of employees in:

(i) A New Jersey restaurant in 1991?

(ii) A New Jersey restaurant in 1993?

(iii) A Pennsylvania restaurant in 1991?

(iv) A Pennsylvania restaurant in 1993?

(b) In terms of the coefficients β_0, β_1, β_2, β_3, what is the average causal effect of the minimum wage on employment?

(c) Explain why Card and Kruger used the difference-in-difference estimator of the causal effect instead of the "New Jersey after − New Jersey before" or the "1993 New Jersey − 1993 Pennsylvania" differences estimator.

1. New Jersey in 1991: $\beta_0 + \beta_1$, New Jersey in 1993: $\beta_0 + \beta_1 + \beta_2 + \beta_3$, Pennsylvania in 1991: β_0, Pennsylvania in 1993: $\beta_0 + \beta_2$.

2. We obtain the average causal effect from the difference-in-difference estimate, given by (New Jersey 1993 − New Jersey 1991) − (Pennsylvania 1993 − Pennsylvania 1991) = $(\beta_2 + \beta_3) - (\beta_2) = \beta_3$.
3. The estimators "New Jersey after − New Jersey before" and "1993 New Jersey − 1993 Pennsylvania" do not give us the effect of the minimum wage increase alone. "New Jersey after − New Jersey before" = $\beta_2 + \beta_3$, where β_2 is the time effect associated with changes in the economy between 1991 and 1993. "1993 New Jersey − 1993 Pennsylvania" = $\beta_1 + \beta_3$, where β_1 denotes the average difference in employment between New Jersey and Pennsylvania.

Exercise 1.5. (Stock and Watson, Exercise 13.3) Suppose that, in a randomized controlled experiment of the effect of an SAT preparatory course on SAT scores, the following results are reported:

	Treatment Group	Control Group
Average SAT Score ($\overline{X}$)	1241	1201
Standard deviation of SAT score (S_X)	93.2	97.1
Number of women	55	45
Number of men	45	55

(a) Estimate the average treatment effect on test scores.

(b) Is there evidence of non-random assignment? Explain.

(a) The estimated average treatment effect is $\overline{X}^{Treatment} - \overline{X}^{Control} = 1241 - 1201 = 40$ points.

(b) There would be nonrandom assignment if men and women had different probabilities of being assigned to the treatment and control groups. Let p_M denote the probability that a male is assigned to the treatment group, and let p_W denote the probability that a female is assigned to the treatment group. Random assignment means $p_M = p_W$ (i.e., probability of assignment does not depend on gender). Testing this null hypothesis results in a t-statistic of t-stat $= \frac{\widehat{p}_M - \widehat{p}_W}{\sqrt{\frac{\widehat{p}_M(1-\widehat{p}_M)}{n_M} + \frac{\widehat{p}_W(1-\widehat{p}_W)}{n_W}}} = \frac{0.55-0.45}{\sqrt{\frac{0.55\cdot 0.45}{100} + \frac{0.45\cdot 0.55}{100}}} = 1.42$, so that the null of random assignment cannot be rejected at all common significance levels.

2 Final Exam Review

Final Exam Spring 2014, Question 3. You are hired by the Government of Ghana to study the impact of income on the level of education. Using data on rural villages, you estimate the following population regression using OLS:

$$Educ_i = \beta_0 + \beta_1 Income_i + \beta_2 Pop_i + \beta_3 School_i + \beta_4 Age_i + u_i$$

where $Educ_i$ is average years of formal education in the village, $Income_i$ is average annual income per capita in the village, Pop_i is the number of village residents, $School_i$ is the number of schools in the village, and Age_i is the average age of the village population.

(a) (5 points) Explain what econometric problem is likely to arise that leads to biased and inconsistent estimates as a result of including $Income$ as a regressor in the education regression as is done above.

The variable is likely to be endogenous since not only does village income have an impact on education, but the average education may also cause income. If we ignore the endogeneity issue, using OLS will result in and biased and inconsistent estimate of β_1. Omitted variables such as religious or sectarian composition of the population that correlate with both income and education may be another source of endogeneity that biases the estimate of β_1.

You learn from Ghana's Minister of Agriculture that the country's citizens derive the bulk of their income from agriculture. As a result, you cleverly infer that average annual rainfall ($Rainfall$) may be a good instrument for income.

(b) (5 points) You recall from your econometrics course that an instrument can be used in a procedure called Two Stage Least Squares that is designed to solve this econometric problem. Describe carefully the first of the two stages and why TSLS will generate a consistent estimate of β_1.

Using TSLS, we would need to run the first stage regression of the endogenous regressor on the instruments and controls:

$$Income_i = \pi_0 + \pi_1 Rainfall_i + \pi_2 Pop_i + \pi_3 School_i + \pi_4 Age_i + v_i.$$

Using the OLSEs from this regression, compute the fitted values $\widehat{Income}$ from this regression. These fitted values are highly correlated with if the instrument is relevant, and if the instrument is exogenous then they should be uncorrelated with the population error term. More simply, the fitted values measure that portion of the endogenous regressor which is correlated with variable of interest and uncorrelated with the error term.

You want to check the Minister's suggestion that rainfall has an impact on incomes in Ghana. You have information on average annual incomes in 1996 and 1997 for two regions: the "coastal region," which had the same precipitation level in both years, and the "hill region," which experienced a 30% increase in rainfall. Comparing 1996 and 1997, income in the coastal region fell from 124 to 104, while income in the hill region fell from 98 to 96. You also recall from your econometrics course that this situation might represent a "natural" or "quasi" experiment, allowing you to estimate the "treatment effect" of rainfall.

(c) (8 points) Perform a difference in differences analysis of the effect of rainfall on average income. Summarize the analysis in a table.

Formally, the D-in-D impact of a 30% increase in rainfall is $\beta = [Income(Hill, 1997) - Income(Hill, 1996)] - [Income(Coast, 1997) - Income(Coast, 1996)] = (96-98)-(104-124) = -2+20 = 18$. Income raised by 18 units due to the 30% increase in rainfall.

Region	Rainfall	1996	1997
Coast (control)	No change	124	104
Hill (treatment)	+30%	98	96

(d) (6 points) Describe a multivariate regression that when estimated using OLS will generate exactly the same estimate of the effect of rainfall on income as was generated by the analysis in part (c).

Let $G_i = 1$ if village is in Hills and $G_i = 0$ if village is on the Coast; $D_t = 1$ if year is 1997 and $D_t = 0$ if year is 1996. Consider the OLS regression: $I_i = \beta_0 + \beta_1 G_i + \beta_2 D_t + \beta_3 G_i \times D_t + u_i$ where I_i is income of village i. The differences in differences estimate of the rainfall effect is the OLSE of β_3.

(e) (6 points) Describe in detail one threat to the internal validity of the OLS estimates when treating these data as a quasi-experiment, and how it would bias the coefficient estimate.

Failure of randomization: the villages may not have been randomly chosen in the two regions; in fact, rainfall is likely not uniform throughout a region, so e.g. a dry part of the Hill region could be no different than the coast. Failure of compliance: should not be an issue here since cannot easily control rainfall. Attrition: movement of people especially between two regions would affect results. Hawthorne effect: depends on whether villages informed about researcher collecting data on income, education and other information.

Final Exam Spring 2011, Question 5. In 1980, due to a temporary easing of Cuban emigration rules, there was a huge influx of Cuban immigrants into the state of Florida. As a result of this so-called "Mariel boatlift," the low-skilled labor force of Miami increased by 7%. David Card compared the average hourly wages in Miami and comparison cities (Atlanta, Houston, Los Angeles, and Tampa-St. Petersburg). The average hourly wages expressed in logarithms are given in the below table:

		Cities	
		Miami	Comparison
Year	1979	1.85	1.93
	1981	1.85	1.91

(a) (10 points) Calculate the percentage change in average hourly wages in the treatment group and in the control group, and uses those changes to the differences-in-differences ("DiD") estimate. Is the sign of the DID estimate what would be predicted by economic theory? Explain.

> Change in the treatment group: 0%, change in the control group: -2% (keeping in mind that a 1% change in wage is equivalent to a 0.01 change in the log of wage). Effect of the increase in labor supply on average hourly wages is equal to +2%. Standard economic theory suggests a negative, not positive, change.

(b) (10 points) Give an example of a relevant variable that is omitted from the DiD estimation, and predict the likely bias it would cause.

> Cuban emigrants likely come to Miami where other Cubans have settled in earlier years who now offer them job opportunities that are not available in the other cities This would bias DiD upward.

(c) (12 points) To accommodate other determinants of metropolitan wage rates, you suggest including a measure of the size of the metropolitan manufacturing sector M_i since it might reflect ability to absorb low-skilled workers. Write down a linear regression that generates a DiD estimate while incorporating this control variable. Why would you believe that this regression approach would change your estimate of the effect of the Mariel boatlift from (a)?

> Two possible specifications, one with differences $\Delta Y_i = Y_i^{after} - Y_i^{before} = \beta_0 + \beta_1 X_i + \beta_2 M_i$, and a second as $Y_i = \beta_0 + \beta_1 X_i + \beta_2 G_i + \beta_3 D_i + \beta_4 M_i$ with usual definitions of the dummies. It is different because it stems from using the multiple regression model rather than the regression with a single regressor. In that case, β_1 is consistent (as long as we have conditional mean independence). Intuitively, by including the additional controls, the differences estimator controls for the fact that the treatment probability can depend on their values. The inclusion of the characteristics also allows for testing for random receipt of treatment and random assignment using the usual F-statistic in auxiliary regressions.