DS-UA 201
Causal Inference
Instructor: Parijat Dube

Homework 5-100 points

Due: August 10, 2024

General Instructions

This homework must be turned in on Gradescope by August 10th 2024, 11:59pm. It must be your own work, and your own work only—you must not copy anyone's work, or allow anyone to copy yours. This extends to writing code. You may consult with others, but when you write up, you must do so alone. Your homework submission must be written and submitted using Rmarkdown. **No handwritten solutions will be accepted**. You should submit:

1. A compiled PDF file named yourNetID solutions.pdf containing your solutions to the problems.
2. A .Rmd file containing the code and text used to produce your compiled pdf named yourNetID solutions.Rmd. Note that math can be typeset in Rmarkdown in the same way as Latex.

Please make sure your answers are clearly structured in the Rmarkdown file:

1. Label each question part(e.g. 3.a).
2. Do not include written answers as code comments.
3. The code used to obtain the answer for each question part should accompany the written answer.

Problem 1 - (25 points)

Suppose you have a time-series dataset with repeated observations of the same n units over two time periods. No unit is treated at t = 0, and some units are treated at t = 1, each unit's treatment status at t = 1 is indicated by the random variable D_i. Suppose that exactly n_t units are treated, and n_c units are in the control group, so that $\Pr(D_i = 1) = \frac{n_t}{n}$. You know that observed outcomes take the following form:

$$Y_{i0} = \delta_0 + u_i + \epsilon_{i0},$$

$$Y_{i1} = \tau_i D_i + \delta_1 + u_i + \epsilon_{i1},$$

where δ_t , u_i , and τ_i are constants, and D_i and ϵ_{it} are random variables that satisfy: $E[\epsilon_{i1}|D_i = d] = E[\epsilon_{i0}|D_i = d] = \eta_i(d)$ for some constants $\eta_i(d)$ for both d = 1 and d = 0, and that $\eta_i(1) \neq \eta_i(0)$.

Part a (**5 points**)

In this setting, the time-dependent potential outcome under control is defined as:

$$Y_{it}(0) = \delta_t + u_i + \epsilon_{it},$$

for t = 0, 1. Show that the parallel trends assumption is satisfied in this setting by showing that $E[Y_{i1}(0) - Y_{i0}(0)|D_i = 1] = E[Y_{i1}(0) - Y_{i0}(0)|D_i = 0]$.

Homework 5-100 points

Part b (10 points)

Show that the estimator:

$$\hat{\tau} = \frac{1}{n_t}\sum_{i=1}^{n}(Y_{i1} - Y_{i0})D_i - \frac{1}{n_c}\sum_{i=1}^{n}(Y_{i1} - Y_{i0})(1 - D_i)$$

is unbiased for the ATE, i.e., show that $E[\hat{\tau}] = \frac{1}{n}\sum_{i=1}^{n}\tau_i$ in this setting.

Part c (10 points)

Suppose that we ignore the first time period, and we instead use the Neyman estimator with only the data for t = 1 for the ATE, that is, we use:

$$\hat{\tau}_{t=1} = \frac{1}{n_t}\sum_{i=1}^{n}Y_{i1}D_i - \frac{1}{n_c}\sum_{i=1}^{n}Y_{i1}(1 - D_i)$$

derive a formula for the bias of $\hat{\tau}_{t=1}$in terms of the constants $\eta_i(d)$; that is, derive a formula for the quantity:$E[\hat{\tau}_{t=1}] - \frac{1}{n}\sum_{i=1}^{n}\tau_i$.

Problem 2 - (15 points)

Suppose that you are in an instrumental variable setting with n units, a binary treatment D_i, and a binary instrument Z_i, where exactly n_t units are assigned to $Z_i = 1$ and n_c are assigned to $Z_i = 0$ uniformly at random, so that $\Pr(Z_i = 1) = \frac{n_t}{n}$. Suppose that you knew the **true** value of the probability of compliance: n $\Pr(D_i(1) > D_i(0)) = p$. Show that in this case the estimator:

$$\hat{\tau}^p_{Wald} = \frac{1}{p}(\frac{1}{n_t}\sum_{i=1}^{n}Y_iZ_i - \frac{1}{n_c}\sum_{i=1}^{n}Y_i(1 - Z_i))$$

is unbiased for the LATE, that is, show that $E[\hat{\tau}^p_{Wald}] = E[Y_i(1) - Y_i(0)|D_i(1) > D_i(0)]$.

Problem 3 - (35 points + 5 (bonus) points)

Despite heated political and media rhetoric, there are few causal estimates of the effect of expanded health-care insurance on healthcare outcomes. One landmark study, the [Oregon Health Insurance Experiment], covered new ground by utilizing a randomized control trial implemented by the state government of Oregon. To allocate a limited number of eligible coverage slots for the state's Medicaid expansion, about 30,000 low-income, uninsured adults (out of about 90,000 wait-list applicants) were randomly selected by lottery to be allowed to apply for Medicaid coverage. Researchers collected observable measures of health (blood pressure, cholesterol, and blood sugar levels), as well as hospital visitation and healthcare expenses for 6,387 selected adults and 5,842 not selected adults.

For this problem, you will need the `OHIE.dta` file. The variables you will need are:

- `treatment` - Selected in the lottery

Homework 5-100 points

- `ohp_all_ever_admin` - Ever enrolled in Medicaid from matched notification date to September 30, 2009 (actually "took" the treatment)
- `tab2bp_hyper` - Outcome: Binary indicator for elevated blood pressure (defined a systolic pressure of 140mm Hg or more and a diastolic pressure of 90mm Hg or more)
- `tab2phqtot_high` - Outcome: Binary indicator for a positive screening result for depression (defined as a score of 10 or higher on the Patient Health Questionnaire)
- `tab4_catastrophic_exp_inp` - Outcome: Indicator for catastrophic medical expenditure (total out-of-pocket medical expenses ≥ 30 percent of household income)
- `tab5_needmet_med_inp` - Outcome: Participant feels that they received all needed medical care in past 12 months (binary indicator)

1. (**10 points**) Estimate the intent-to-treat effects of assignment to treatment (being eligible to apply) on each of the four outcomes (elevated blood pressure, depression, catastrophic medical expenditure, and whether respondents had their health care needs met). Provide 95% confidence intervals for each estimate and interpret your results.
2. (**10 points**) Suppose that researchers actually wanted to estimate the effect of Medicaid enrollment on each of the four outcomes. Suppose they first used a naive regression of each of the the outcomes on the indicator of Medicaid enrollment. Report a 95% confidence interval for each of your estimates and interpret your results. Why might these be biased estimates for the causal effect of Medicaid enrollment?
3. (**5 points**) Suppose we were to use assignment to treatment as an instrument for actually receiving Medicaid cover-age. Consider that not everyone who was selected to apply for Medicaid actually ended up applying and receiving coverage. Likewise, some applicants who were not selected to receive the treatment nevertheless were eventually covered. What were the compliance rates (the level of Medicaid enrollment) for subjects who were selected and subjects who were not selected? Use a "first stage" regression to estimate the effect of being selected on Medicaid enrollment to estimate the compliance rates. Is the instrument of assignment-to-treatment a strong instrument for actual Medicaid enrollment?
4. (**10 points**) Now estimate the effect of Medicaid enrollment on each of the four outcomes using an instrumental variables strategy. Report a 95% confidence interval for your estimates and interpret your results. Compare the estimates to those you obtained in Question 3.
5. (**Bonus question - 5 points**) What additional assumptions do you have to make in order to interpret your estimates from Question 4 as an Average Treatment Effect for the entire sample?

Problem 4 - (25 points)

Does US military assistance strengthen or further weaken fragile and conflict-affected foreign governments? Aid may bolster state capacity and suppress violence from nonstate actors such as

Homework 5-100 points

paramilitary groups. On the other hand, aid may be diverted to those same violent groups. To answer the question, [Dube and Naidu 2015] leverage changes in the allocation of US military aid to Colombian military bases. They test whether Colombian muncipalities in which military bases are located have more or less paramilitary violence when the level of U.S. military aid increases, relative to Colombian municipalities in which military bases are not located.

For this problem, you will need the `bases_replication_final.dta` file. The variables you will need are:

- `parattq` - DV here is paramilitary attacks
- `bases6` - indicator variable whether or not there is a base in the municipality
- `lrmilnar_col` - (logged) U.S. military and narcotics aid to Colombia
- `bases6xlrmilnar_col` - the treatment i.e., the interaction between the level of U.S. military and narcotics aid and whether or not there is a base in the municipality
- `lnnewpop` - is log of population

1. (**10 points**) The treatment in this case is a continuous 'intensity' variable that changes over time. The authors use the interaction between the level of U.S. military and narcotics aid and whether a base exists in a municipality. How many units are in the 'control' group (no bases)? Does the bases variable change over time or is it a unit-constant factor? How about the logged military aid variable, does it change across units for a given year? What do the authors seem to be assuming about how military aid is allocated?

2. (**5 points**) The authors use a common empirical strategy called *two-way fixed effects* to estimate the average treatment effect of military aid. The model they estimate includes fixed effects for both time periods and units (and includes logged population as an additional covariate):

$$Y_{it} = \gamma_t + \alpha_i + \tau D_{it} + \beta X_{it} + \epsilon_{it}$$

   What assumptions are the authors making in order to identify the treatment effect of military aid?

3. (**10 points**) Using the two-way fixed effects estimator, estimate the effect of U.S. military and narcotics aid on the number of paramilitary attacks, including log of population as a covariate. The two sets of fixed effects are for municipality (municipality) and year (year). Cluster your standard errors at the unit level (see the cluster argument in lmrobust. Report a 95% confidence interval for your estimate and interpret your results.