DS-UA 201
Causal Inference
Instructor: Parijat Dube

Homework 4-115 points

Due: July 27, 2024

General Instructions

This homework must be turned in on Gradescope by July 27, 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 - *Political Efficacy in China and Mexico* 25 points

In 2002, the World Health Organization conducted a survey of two provinces in China and three provinces in Mexico. [4] One issue of interest, which we analyze in this exercise, concerns political efficacy. First, the following self-assessment question was asked.

How much say do you have in getting the government to address issues that interest you?

(5) Unlimited say, (4) A lot of say, (3) Some say, (2) Little say, (1) No say at all.

After the self-assessment question, three vignette questions were asked.

[`Alison`] lacks clean drinking water. She and her neighbors are supporting an opposition candidate in the forthcoming elections that has promised to address the issue. It appears that so many people in her area feel the same way that the opposition candidate will defeat the incumbent representative.

[`Jane`] lacks clean drinking water because the government is pursuing an industrial development plan. In the campaign for an upcoming election, an opposition party has promised to address the issue, but she feels it would be futile to vote for the opposition since the government is certain to win.

[`Moses`] lacks clean drinking water. He would like to change this, but he can't vote, and feels that no one in the government cares about this issue. So he suffers in silence, hoping something will be done in the future.

The respondent was asked to assess each vignette in the same manner as the self assessment question. How much say does ["name"] have in getting the government to address issues that interest

Homework 4-115 points

Variable	Description
self	self-assessment response
alison	response to the Alison vignette
jane	response to the Jane vignette
moses	response to the Moses vignette
china	1 for China and 0 for Mexico
age	age of respondent in years

Table 1: Vignette Survey Data

[him/her]?

(5) Unlimited say, (4) A lot of say, (3) Some say, (2) Little say, (1) No say at all.

[`"name"`] is replaced by either Alison, Jane, or Moses.

The data set we analyze `vignettes.csv` contains the variables whose names and descriptions are given in Table 1. In the analysis that follows, we assume that these survey responses can be treated as numerical values. For example, "Unlimited say" = 5, and "Little say" = 2. This approach is not appropriate if, for example, the difference between "Unlimited say" and "A lot of say" is not the same as the difference between "Little say" and "No say at all." However, relaxing this assumption is beyond the scope of this chapter.

1. (**5 points**) We begin by analyzing the self-assessment question. Plot the distribution of responses separately for China and Mexico using bar plots, where the vertical axis is the proportion of respondents. In addition, compute the mean response for each country. According to this analysis, which country appears to have a higher degree of political efficacy? How does this evidence match with the fact that in the 2000 election, Mexican citizens voted out of office the ruling Institutional Revolutionary Party (PRI) who had governed the country for more than 80 years, while Chinese citizens have not been able to vote in a fair election to date?

2. (**5 points**) We examine the possibility that any difference in the levels of efficacy between Mexican and Chinese respondents is due to the difference in their age distributions. Create histograms for the age variable separately for Mexican and Chinese respondents. Add a vertical line representing the median age of the respondents for each country. In addition, use a quantile-quantile plot to compare the two age distributions. What differences in age distribution do you observe between the two countries? Answer this question by interpreting each plot.

3. (**5 points**) One problem with the self-assessment question is that survey respondents may interpret the question differently. For example, two respondents who choose the same answer may be facing quite different political situations and hence may interpret "A lot of say" differently. To address this problem, we rank a respondent's answer to the self-assessment question relative to the same respondent's answer to a vignette question. Compute the proportion of respondents, again separately for China and Mexico, who rank themselves (according to the self-assessment question) as having less say in the government's decisions than Moses (the last

Homework 4-115 points

vignette). How does the result of this analysis differ from that of the previous analysis? Give a brief interpretation of the result.

4. (**5 points**) We focus on survey respondents who ranked these three vignettes in the expected order (i.e., Alison ≥ Jane ≥ Moses). Create a variable that represents how respondents rank themselves relative to these vignettes. This variable should be equal to 1 if respondents rank themselves less than Moses, 2 if ranked the same as Moses or between Moses and Jane, 3 if ranked the same as Jane or between Jane and Alison, and 4 if ranked the same as Alison or higher. Create the bar plots of this new variable as done in question 1. The vertical axis should represent the proportion of respondents for each response category. Also, compute the mean value of this new variable separately for China and Mexico. Give a brief interpretation of the result by comparing these results with those obtained in question 1.

5. (**5 points**) Is the problem identified above more or less severe among older respondents when compared to younger ones? Answer the previous question separately for those who are 40 years or older and those who are younger than 40 years. Does your conclusion for the previous question differ between these two groups of respondents? Relate your discussion to your finding for question 2.

Problem 2 - *Election and Conditional Cash Transfer in Mexico* 30 points

In this exercise, we analyze the data from a study that seeks to estimate the electoral impact of Progresa, Mexico's conditional cash transfer program (CCT program). [7] The original study relied on a randomized evaluation of the CCT program in which eligible villages were randomly assigned to receive the program either 21 months (early Progresa) or 6 months (late Progresa) before the 2000 Mexican presidential election. The author of the original study hypothesized that the CCT program would mobilize voters, leading to an increase in turnout and support for the incumbent party (PRI, or Partido Revolucionario Institucional, in this case). The analysis was based on a sample of precincts that contain at most one participating village in the evaluation.

The data we analyze are available as the CSV file `progresa.csv`. Table in Figure 1 presents the names and descriptions of variables in the data set. Each observation in the data represents a precinct, and for each precinct the file contains information about its treatment status, the outcomes of interest, socioeconomic indicators, and other precinct characteristics.

1. (**5 points**) Estimate the impact of the CCT program on turnout and support for the incumbent party (PRI) by comparing the average electoral outcomes in the "treated" (early Progresa) precincts versus the ones observed in the "control" (late Progresa) precincts. Next, estimate these effects by regressing the outcome variable on the treatment variable. Interpret and compare the estimates under these approaches. Here, following the original analysis, use the turnout and support rates as shares of the eligible voting population (t2000 and pri 2000 s, respectively). Do the results support the hypothesis? Provide a brief interpretation.

2. (**5 points**) In the original analysis, the author fits a linear regression model that includes, as predictors, a set of pretreatment covariates as well as the treatment variable. Here, we

Homework 4-115 points

Variable	*Description*
`treatment`	whether an electoral precinct contains a village where households received early Progresa
`pri2000s`	PRI votes in the 2000 election as a share of precinct population above 18
`pri2000v`	official PRI vote share in the 2000 election
`t2000`	turnout in the 2000 election as a share of precinct population above 18
`t2000r`	official turnout in the 2000 election
`pri1994`	total PRI votes in the 1994 presidential election
`pan1994`	total PAN votes in the 1994 presidential election
`prd1994`	total PRD votes in the 1994 presidential election
`pri1994s`	total PRI votes in the 1994 election as a share of precinct population above 18
`pan1994s`	total PAN votes in the 1994 election as a share of precinct population above 18
`prd1994s`	total PRD votes in the 1994 election as a share of precinct population above 18
`pri1994v`	official PRI vote share in the 1994 election
`pan1994v`	official PAN vote share in the 1994 election
`prd1994v`	official PRD vote share in the 1994 election
`t1994`	turnout in the 1994 election as a share of precinct population above 18
`t1994r`	official turnout in the 1994 election
`votos1994`	total votes cast in the 1994 presidential election
`avgpoverty`	precinct average of village poverty index
`pobtot1994`	total population in the precinct
`villages`	number of villages in the precinct

Figure 1: Conditional Cash Transfer Program (Progresa) Data.

fit a similar model for each outcome that includes the average poverty level in a precinct (avgpoverty), the total precinct population in 1994 (pobtot1994), the total number of voters who turned out in the previous election (votos1994), and the total number of votes cast for each of the three main competing parties in the previous election (pri1994 for PRI, pan1994 for Partido Acción Nacional or PAN, and prd1994 for Partido de la Revolución Democrática or PRD). Use the same outcome variables as in the original analysis, which are based on the shares of the voting age population. According to this model, what are the estimated average effects of the program's availability on turnout and support for the incumbent party? Are these results different from those you obtained in the previous question?

3. (**5 points**) Next, we consider an alternative, and more natural, model specification. We will use the original outcome variables as in the previous question. However, our model should include the previous election outcome variables measured as shares of the voting age population (as done for the outcome variables t1994, pri1994s, pan1994s, and prd1994s) instead of those measured in counts. In addition, we apply the natural logarithmic transformation to the precinct population variable when including it as a predictor. As in the original model, our model includes the average poverty index as an additional predictor. Are the results based on these new model specifications different from those we obtained in the previous question? If the results are different, which model fits the data better?

4. (**5 points**) We examine the balance of some pretreatment variables used in the previous analyses. Using box plots, compare the distributions of the precinct population (on the original scale), average poverty index, previous turnout rate (as a share of the voting age population), and previous PRI support rate (as a share of the voting age population) between the treatment and control groups. Comment on the patterns you observe.

5. (**5 points**) We next use the official turnout rate t2000r (as a share of the registered voters) as the outcome variable rather than the turnout rate used in the original analysis (as a share of the voting age population). Similarly, we use the official PRI's vote share pri2000v (as a share of all votes cast) rather than the PRI's support rate (as a share of the voting age population). Compute the average treatment effect of the CCT program using a linear regression with the average poverty index, the log-transformed precinct population, and the previous official election outcome variables (t1994r for the previous turnout; pri1994v, pan1994v, and pra1994v for the previous PRI, PAN, and PRD vote shares). Briefly interpret the results.

6. (**5 points**) So far we have focused on estimating the average treatment effects of the CCT program. However, these effects may vary from one precinct to another. One important dimension to consider is poverty. We may hypothesize that since individuals in precincts with higher levels of poverty are more receptive to cash transfers, they are more likely to turn out in the election and support the incumbent party when receiving the CCT program. Assess this possibility by examining how the average treatment effect of the policy varies by different levels of poverty for precincts. To do so, fit a linear regression with the following predictors: the treatment variable, the log-transformed precinct population, the average poverty index and its square, the interaction between the treatment and the poverty index, and the interaction between the treatment and the squared poverty index. Estimate the average effects for unique observed values and plot them as a function of the average poverty level. Comment on the

resulting plot.

Problem 3 - *Data Generating Process* 25 points

Consider a setting in which you have a lot of contextual knowledge about your data, and you know that it is generated according to the following Data Generating Process (DGP):

$$Y_i|X_i, D_i = X_i\beta + \tau D_i + \epsilon_i, \tag{1}$$

$$D_i|X_i = \mathbb{I}(X_i\gamma + \nu_i > 0), \tag{2}$$

$$\epsilon_i \sim \mathcal{N}(0,1), \tag{3}$$

$$\nu_i \sim \mathcal{N}(0,1), \tag{4}$$

$$X_i \sim Bernoulli(\pi) \tag{5}$$

Let's parse this: starting from the bottom, Equation (5) tells us that our covariate, X, is a **binary** random variable that takes value 1 with probability π. Equations (4) and (3) tell us that the statistical noise terms in our treatment and outcome models follow a standard normal distribution. Equation (2) tells us that the treatment indicator is a binary random variable that takes value 1 whenever the condition $X_i\gamma + \nu_i > 0$ is met. Finally, equation (1) tells us that our outcome can be described in the real world with a linear combination of the covariates, the treatment, and some normally distributed statistical noise. The fixed (i.e., non-random), but unknown parameters in this DGP are β, τ, γ and π.

1. (**4 points**) Throughout the rest of the question it will be important to keep in mind several known properties of the normal distribution that we covered in our statistics review. Let's practice here. First, using what is given in Equations (1) and (3), together with the properties of the normal distribution, write down the distribution of $Y_i|X_i, D_i$, together with its mean and variance. Second, recall that any normal random variable can be *standardized* by subtracting its mean from it, and dividing it by its standard deviation. That is, if $A \sim N(\mu, \sigma^2)$, then $Z = \frac{A-\mu}{\sigma} \sim \mathcal{N}(0,1)$. Using $\Phi(a) = Pr(Z \leq a)$ to denote the CDF of a standard normal r.v evaluated at a point $z \in \mathbb{R}$ and $\phi(z)$ to denote the PDF of a standard normal r.v evaluated at the same point, write down an expression for the PDF and CDF of $Y_i|X_i, D_i$ evaluated at a point $y \in \mathbb{R}$.

2. (**6 points**) Using the properties of the normal distribution, indicator random variables, and standardization, derive expressions for $E[D_i|X_i = 1]$, and $E[D_i|X_i = 0]$.

3. (**5 points**) For two random variables, A, B, Bayes' rule states that $Pr(A = a|B = b) = \frac{Pr(B=b|A=a)Pr(A=a)}{\sum_a Pr(B=b|A=a)(Pr(A=a)}$Using this rule and your previous answers, derive expressions for $Pr(X_i = 1|D_i = 1)$ and $Pr(X_i = 1|D_i = 0)$.

4. (**10 points**) Suppose that we estimate the associational difference:

$$\eta = E[Y_i|D_i = 1] - E[Y_i|D_i = 0],$$

using your previous answers, derive a formula for the bias $\eta - \tau$ in the context of the DGP in Equations (1)-(5) that only involves numbers, the unknown parameters in the dgp, and the function Φ

Problem 4 - *Propensity Score Regression and IPW* 20 points

Using the same DGP as in question 1, suppose that we predict the propensity score for X = 1 by running a **simple linear regression** of D_i on X_i, that is, we estimate γ_i in the model: $D_i = X_i\gamma_i + \delta_i$, where $E[\delta_i] = 0$ using OLS regression, then we predict $\hat{e}(1)$ using the estimated coefficent. We then repeat the process for X = 0 by estimating γ_0 in the model: $D_i = (1 - X_i)\gamma_0 + \delta_i$ and predict $\hat{e}(0)$ in a similar way.

1. (**5 points**) Using the fact that, in the DGP above: $(X'X)^{-1} = \frac{1}{\sum_{i=1}^n X_i}$ and $X'D = \sum_{i=1}^n X_i D_i$, give a formula for when the predicted propensity scores: $\hat{e}(0)$ (the predicted pscore for X = 0 and $\hat{e}(1)$ (the predicted pscore when X = 1).

2. (**10 points**) Using your answers to part a, show that the stratified estimator on X and the IPW estimator with the pscore estimated in the way described above are exactly the same in the context of the DGP in (1)-(5), i.e.,show that:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i D_i}{\hat{e}(X_i)} - \frac{Y_i(1 - D_i)}{(1 - \hat{e}(X_i))} = \sum_{x=0}^{1}\frac{N_{X=x}}{n}\left(\frac{1}{N_{X_i=x,D_i=1}}\sum_{i:X_i=x} Y_i D_i - \frac{1}{N_{X_i=x,D_i=0}}\sum_{i:X_i=x} Y_i(1 - D_i)\right) \quad (6)$$

3. (**5 points**) Equation (2) tells us that the "true" model for $D_i|X_i$ is $D_i|X_i = \mathbb{I}(X_i\gamma + \nu_i > 0)$, however we have estimated the propensity score by assuming that $D_i|X_i = \alpha + X_i\gamma + \delta_i$. Will the estimates of the propensity score that we have obtained in this way be biased? (both explaining in words or proving mathematically are valid answers)

Problem 5 - *Matching Approach using TRC* 15 points

Use the same data set as in Question 4 in Homework 2.

Part a (3 points)

Estimate the ATE of TRC exposure on respondents' racial attitudes using the Matching approach. You can use the Match function from Matching package in R. And implement the Mahalanobis Distance matching to get the matching data. Report the 95% confidence interval of your estimate.

Homework 4-115 points

Part b (3 points)

Now estimate the ATE by matching to 3 observations instead of 1 (which is the default in the Match function). Did the Standard Error change compared to part A?

Part c (3 points)

Now adjust for the bias due to inexact matching and compute the ATE estimate. Did the standard error decrease compared to the previous two parts?

Part d (6 points)

Compute the matching weights from part c for the observations and create the balance table using the match- ing weights computed without using the inbuilt "MatchBalance" function in the Matching package.