DS-UA 201
Causal Inference
Instructor: Parijat Dube
Final Exam-100 points Due: 11:59AM Aug 15, 2024

Instructions

You should submit your writeup (as a knitted `.pdf` along with the accompanying `.rmd` file) to the course website before ***11:59am EST on Wednesday, August 15th****. Please upload your solutions as a .pdf file saved as `Yourlastname_Yourfirstname_final.pdf`. In addition, an electronic copy of your `.rmd` file (saved as `Yourlastname_Yourfirstname_final.rmd`) should accompany this submission.*

Late finals will not be accepted, so start early and plan to finish early. Remember that exams often take longer to finish than you might expect.

This exam has ***5 questions*** *and is worth a total of* ***100 points****. Show your work in order to receive partial credit. Also, we will not accept un-compiled `.rmd` files.*

In general, you will receive points (partial credit is possible) when you demonstrate knowledge about the questions we have asked, you will not receive points when you demonstrate knowledge about questions we have not asked, and you will lose points when you make inaccurate statements (whether or not they relate to the question asked). Be careful, however, that you provide an answer to all parts of each question.

You may use your notes, books, and internet resources to answer the questions below. However, you are to work on the exam by yourself.

You are prohibited from corresponding with any human being or Generative AI (ChaptGPT) tools regarding the exam (unless following the procedures below). Detection of the use of Generative AI tools to answer exam questions will automatically results in a F grade for the course.

We will answer clarifying questions during the exam. We will not answer statistical or computational questions until after the exam is over. If you have a question, post it on **Campuswire** *as a* **private post***, visible only to the instructional staff. If your question is a clarifying one, we will reply.*

Problem 1 (20 points)

This problem will have you replicate and analyze the results from Moser and Voena's 2012 AER paper on the impact of the World War I *Trading with the Enemy Act* on U.S. domestic invention. The full citation is below

Moser, P., & Voena, A. (2012). *Compulsory licensing: Evidence from the trading with the enemy act.* American Economic Review, 102(1), 396-427.

The premise of the study is to evaluate the effect that "compulsory licensing" policy - that is, policies that permit domestic firms to violate foreign patents and produce foreign inventions without needing to obtain a license from the owner of the foreign patent - have on domestic invention. Does access to foreign inventions make domestic firms more innovative? The authors leverage an exogenous event in U.S. licensing policy that arose from World War I - the 1917 ”Trading with the Enemy Act” (TWEA) which permitted U.S. firms to violate patents owned by enemy-country firms. This had the consequence of effectively licensing all patents from German-owned firms to U.S. firms after 1918 (that is, from 1919 onward), allowing them to produce these inventions without paying for a license from the German-owned company.

The authors look specifically at domestic innovation and patent activity in the organic chemicals sector. They note that only some of the sub-classes of organic chemicals (as defined by the US Patent Office) received any compulsory licenses under the Trading with the Enemy Act while others did not. They leverage this variation in exposure to the ”treatment” of compulsory licensing to implement a differences-in-differences design looking at domestic firm patent activity in each of these sub-classes (comparing sub-classes that were exposed to compulsory licensing to those that were unexposed).

The unit of the dataset is the sub-class/year (471,120 observations) of 7248 US Patent and Trademark Office (USPTO) patent sub-classes over 65 years.

The dataset is `patents.csv` and the relevant variables are:

- `uspto_class` - USPTO Patent Sub-Class (unit)
- `grntyr` - Year of observation (year)
- `count_usa` - Count of patents granted to US-owned firms in the year
- `count_for` - Count of patents granted to foreign-owned (non-US) firms in the year
- `treat` - Treatment indicator - Whether the patent sub-class received any German patents under TWEA (after 1918 when the policy went into effect) (Note that this is not an indicator for the overall treatment group (whether the unit ever received treatment) - it is only 1 after 1918 for units that receive treatment but is still 0 for those ”treated” units prior to the initiation of treatment)

Question A (5 points)

If you try to use a two-way fixed effects estimator on the dataset as it is, it will likely freeze up your computer as this is a very large dataset. We'll instead first aggregate the data in a way that will let you use a simple difference-in-differences estimator to estimate the treatment effect.

Generate a point estimate for the average treatment effect of receiving treatment on the average annual count of US patents using a difference-in-differences estimator (using all post-treatment (1919-1939) and pretreatment (1875-1918) time periods. You should aggregate your data such that the outcome is the post-/pre-difference in the outcome (preferably using `tidyverse` functions like `group_by` and `summarize`) and each row is a USPTO patent sub-class (rather than a sub-class/year observation) and use a difference-in-means estimator with the differenced outcome. Again, if you use `lm_robust` or even `lm` with two-way fixed effects, your computer will likely freeze up as there are many FE parameters to estimate.

Provide a 95% robust confidence interval and interpret your point estimate. Do we reject the null of no treatment effect at the $\alpha = .05$ level?

Question B (5 points)

A colleague suggests that you should instead just compare the average differences in the count of US patents in the post-1918 period between exposed and unexposed sub-classes to estimate the treatment effect. Based on what we observe in the pre-1919 period, is ignorability of the treatment likely to hold under this strategy? Discuss why or why not - what do you observe in the patent counts in the pre-treatment period between exposed and unexposed subclasses?

Question C (5 points)

We might be concerned that there are differential trends in pre-treatment patenting between those sub-classes exposed to the treatment and those exposed to control. Estimate the difference in the trend in US patents between exposed and unexposed sub-classes from 1918 to 1917, 1916, 1915, and 1914 (four estimates in total: 1918-1917, 1918-1916, 1918-1915, 1918-1914). Provide a 95% robust confidence interval for each of these estimates and interpret your results. Do we reject the null that any of these differ from 0 (at α = .05) ? If the outcome trends were evolving in parallel, what would we expect these estimates to be? What do your results suggest for the validity of the parallel trends assumption?

Question D (5 points)

The authors adjust for covariates out of concern for possible parallel trends violations. One possible confounder that might be driving a parallel trends violation is the overall amount of foreign patenting in the sub-class and its change over time - reflecting general technological differences that might differ between the patent sub-classes. Since the treatment does not affect the amount of foreign patenting, this is a valid control.

Create a variable for the change between the post- and pre-treatment count of foreign patents in the USPTO subclass. Bin this variable into six (6) roughly-equally sized strata and estimate the effect of the treatment on US patenting (again using the differenced outcome) using a stratified difference-in-means estimator. Provide a robust 95% confidence interval and interpret your results. Do we reject the null of no treatment effect at the α = .05 level? Compare your results to your estimate from Question A and discuss why they might differ.

Problem 2 (20 points)

In this problem you will be analyzing a dataset from a 2011 paper by Carpenter and Dobkin. The full citation for the paper is:

Carpenter, C., & Dobkin, C. (2011). *The minimum legal drinking age and public health.* Journal of Economic Perspectives, 25(2), 133-56.

This paper examines evidence linking the legal alcohol drinking age in the US (21) to increased likelihood of accidents, hospitalization, and health hazards in general. The main identification strategy employed by the authors is a sharp Regression Discontinuity Design (RDD), where age is the running variable, and 21 is the cutoff.

The dataset contains 80 observations, where each unit is an age group, and values are collected over 4 US states.

The dataset is `ER.csv` and it contains five variables:

- `age` - The age of the unit, where the decimal indicates month of the year
- `all` - The total number of ER admissions
- `injury` - The total number of ER admissions due to injury
- `illness` - The total number of ER admissions due to viral illness
- `alcohol` - An adjusted index of how many ER admissions were linked to alcohol consumption

Question A (8 points)

Estimate the effect of being legally able to purchase alcohol (age ≥ 21) on the all, injury, and alcohol variables using an RDD with bandwidth = 1. For each of the three outcomes report point estimates, standard errors, and 95% confidence intervals. Repeat the analysis for bandwidth = 0.5 years, and bandwidth = 2 years. Discuss and interpret your results. Which outcome variable seems to be associated with the largest effect? Does bandwidth selection influence results?

Question B (4 points)

Using the entire dataset, create and show RDD plots that visualize the discontinuity for each of the three outcome variables used in Question A. The plots should display both points and regression lines.

Question C (8 points)

Conduct a placebo RDD analysis using the illness variable as outcome: since viral illnesses are not caused by alcohol consumption, we have no reason to expect that being legally able to drink will have an effect on this variable. Report both RDD estimates with standard errors and 95% CIs, and make a RDD plot for this outcome variable. Is there a treatment effect and is it statistically significant? What does this suggest about the plausibility of the RDD assumptions?

Problem 3 (20 points)

In this problem, you will examine whether family income affects an individual's likelihood to enroll in college by analyzing a survey of approximately 4739 high school seniors that was conducted in 1980 with a follow-up survey taken in 1986.

This dataset is based on a dataset from

Rouse, Cecilia Elena. *Democratization or diversion? The effect of community colleges on educational attainment.* Journal of Business & Economic Statistics 13, no. 2 (1995): 217-224.

DS-UA 201
Causal Inference
Instructor: Parijat Dube

Final Exam-100 points

Due: 11:59AM Aug 15, 2024

The dataset is `college.csv` and it contains the following variables:

- `college` Indicator for whether an individual attended college. (Outcome)
- `income` Is the family income above USD 25,000 per year (Treatment)
- `distance` distance from 4-year college (in 10s of miles).
- `score` These are achievement tests given to high school seniors in the sample in 1980.
- `fcollege` Is the father a college graduate?
- `tuition` Average state 4-year college tuition (in 1000 USD).
- `wage` State hourly wage in manufacturing in 1980.
- `urban` Does the family live in an urban area?

Question A (5 points)

Draw a DAG of the variables included in the dataset, and explain why you think arrows between variables are present or absent. You can use any tool you want to create an image of your DAG, but make sure you embed it in your compiled .pdf file. Assuming that there are no unobserved confounders, what variables should you condition on in order to estimate the effect of the treatment on the outcome, according to the DAG you drew?

Question B (5 points)

Choose one among the methodologies for ATE estimation under conditional ignorability that we have covered in class to apply to this dataset. Explain why you made your choice, and discuss the assumptions that are needed to apply your method of choice to this dataset. State if and why you think these assumptions hold in this dataset. In addition, choose a method to compute variance estimates for the estimator you chose, and discuss the reasons behind your choice in the context of this dataset.

Question C (10 points)

Using the methodology you chose in Question B to control for the confounders you have selected in Question A, as well as the relevant R packages, provide your estimate of the ATE of the treatment on the outcome. Using your variance estimator of choice, report standard errors and 95% confidence intervals around your estimates. Interpret your results and discuss both their statistical significance and their substantive implications.

Variable	Description
`shareself`	proportion of self-employed potential voters
`shareblue`	proportion of blue-collar potential voters
`sharewhite`	proportion of white-collar potential voters
`sharedomestic`	proportion of domestically employed potential voters
`shareunemployed`	proportion of unemployed potential voters
`nvoter`	number of eligible voters
`nazivote`	number of votes for Nazis

Table 1: 1932 German Election Data.

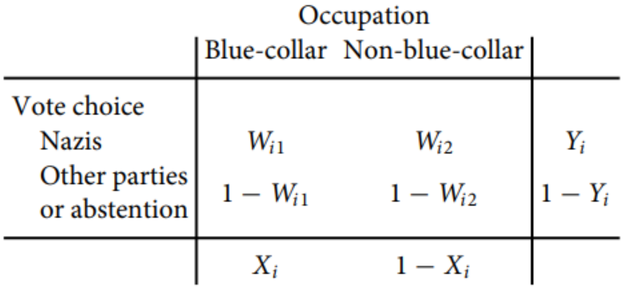
Problem 4 (20 points)

Who voted for the Nazis? Researchers attempted to answer this question by analyzing aggregate election data from the 1932 German election during the Weimar Republic. This question is based on the following article:
G. King, O. Rosen, M. Tanner, A. F. Wagner(2008). *Ordinary economic voting behavior in the extraordinary election of Adolf Hitler.* Journal of Economic History, vol. 68, pp. 951-996. 2008

We analyze a simplified version of the election outcome data, which records, for each precinct, the number of eligible voters as well as the number of votes for the Nazi party. In addition, the data set contains the aggregate occupation statistics for each precinct. Table 1 presents the variable names and descriptions of the CSV data file `nazis.csv`. Each observation represents a German precinct.

The goal of the analysis is to investigate which types of voters (based on their occupation category) cast ballots for the Nazi party in 1932. One hypothesis says that the Nazis received much support from blue-collar workers. Since the data do not directly tell us how many blue-collar workers voted for the Nazis, we must infer this information using a statistical analysis with certain assumptions. Such an analysis, where researchers try to infer individual behaviors from aggregate data, is called ecological inference.

To think about ecological inference more carefully in this context, consider the following simplified table for each precinct i.



	Occupation		
	Blue-collar	Non-blue-collar	
Vote choice			
Nazis	W_{i1}	W_{i2}	Y_i
Other parties or abstention	$1 - W_{i1}$	$1 - W_{i2}$	$1 - Y_i$
	X_i	$1 - X_i$	

The data at hand tells us only the proportion of blue-collar voters X_i and the vote share for the Nazis Y_i in each precinct, but we would like to know the Nazi vote share among the blue-collar voters W_{i1} and among the non-blue-collar voters W_{i2}. Then, there is a deterministic relationship between X, Y, and $\{W_1, W_2\}$. Indeed, for each precinct i, we can express the overall Nazi vote share as the weighted average of the Nazi vote share of each occupation:

$$Y_i = X_i W_{i1} + (1 - X_i) W_{i2} \tag{1}$$

Question A (4 points)

We exploit the linear relationship between the Nazi vote share Y_i and the proportion of blue-collar voters X_i given in equation (1) by regressing the former on the latter. That is, fit the following linear regression model:

$$\mathbb{E}(Y_i \mid X_i) = \alpha + \beta X_i \tag{2}$$

Compute the estimated slope coefficient, its standard error, and the 95% confidence interval. Give a substantive interpretation of each quantity.

Question B (4 points)

Based on the fitted regression model from the previous question, predict the average Nazi vote share Y_i given various proportions of blue-collar voters X_i. Specifically, plot the predicted value of Y_i (the vertical axis) against various values of X_i within its observed range (the horizontal axis) as a solid line. Add 95% confidence intervals as dashed lines. Give a substantive interpretation of the plot.

Question C (4 points)

Fit the following alternative linear regression model:

$$\mathbb{E}(Y_i \mid X_i) = \alpha^* X_i + (1 - X_i) \beta^*. \tag{3}$$

Note that this model does not have an intercept. How should one interpret α^* and β^* ? How are these parameters related to the linear regression model given in equation (2)?

Question D (4 points)

Fit a linear regression model where the overall Nazi vote share is regressed on the proportion of each occupation. The model should contain no intercept and 5 predictors, each representing the proportion of a certain occupation type. Interpret the estimate of each coefficient and its 95% confidence interval. What assumption is necessary to permit your interpretation?

Question E (4 points)

Finally, we consider a model-free approach to ecological inference. That is, we ask how much we can learn from the data alone without making an additional modeling assumption. Given the relationship in equation (1), for each precinct, obtain the smallest value that is logically possible for

W_{i1} by considering the scenario in which all non-blue-collar voters in precinct i vote for the Nazis. Express this value as a function of X_i and Y_i. Similarly, what is the largest possible value for W_{i1} ? Calculate these bounds, keeping in mind that the value for W_{i1} cannot be negative or greater than 1. Finally, compute the bounds for the nationwide proportion of blue-collar voters who voted for the Nazis (i.e., combining the blue-collar voters from all precincts by computing their weighted average based on the number of blue-collar voters). Give a brief substantive interpretation of the results.

Problem 5 (20 points)

In this problem we are interested in the question of whether an extra year of education causes increased wages. We will use `wage2.csv` data used in the following paper:

M. Blackburn and D. Neumark (1992), *Unobserved Ability, Efficiency Wages, and Interindustry Wage Differentials,* Quarterly Journal of Economics 107, 1421-1436. https://doi.org/10.3386/w3857

This dataset includes a bunch of different variables. These are the variables we care about for this question:

Variable name	Description
`wage`	Monthly wage (1980 dollars)
`educ`	Years of education
`feduc`	Years of education for father
`meduc`	Years of education for mother

Question A (4 points)

1. **(2 points)** First create a naive model of the relationship between years of education and wage, by treating wage as a dependent variable. What is the effect of education on wage as predicted by the model?

2. **(2 points)** Does this naive model correctly estimates the effect of education on wages? Why or why not?

Question B (6 points)

We next want to apply 2SLS (two-stage least squares) approach to account for unobservable confounders. For this we need to identify an instrument variable (IV). Remember, for an instrument to be valid, it should meet these criteria:

1. *Relevance:* Instrument is correlated with policy variable

2. *Exclusion:* Instrument is correlated with outcome only through the policy variable

3. *Exogeneity:* Instrument isn't correlated with anything else in the model (i.e. omitted variables)

We have two choices for the instrument variable: years of education of father and years of education of mother. Select one of them and show using the data that it satisfies all the three criteria to be an instrument variable.

Question C (5 points)

Using the IV identified in Question B, you will next perform 2SLS manually.

1. **(2 points)** In the first stage, predict education using the IV.
2. **(3 points)** In the second stage, use the predicted education to estimate the exogenous effect of education on wages. What is the causal effect of education on wage using this 2SLS analysis?

Question D (5 points)

Now using `iv_robust()` perform 2SLS in a single step.

1. **(2 points)** Compare the exogenous effect of education on wages predicted by the manual 2SLS from Question C with the effect obtained using `iv_robust()`. Are they same or different?
2. **(3 points)** Next compare the standard errors of the effect estimate from the two approaches. Which approach correctly estimates the standard error and why?