Causal Inference
Problem Set 4

Due Monday May 14th

Problem 1

Consider the linear model $Y_i = \tau_X D_i + \beta X_i + \epsilon_i$; i.e., a model in which the treatment effect varies by strata of X (heterogenous treatment effects).

(a) Write down an expression for the $\hat{\tau}_{OLS}$ obtained by regressing Y on D and X.

(b) Write down an expression for the $\hat{\tau}_{OLS}$ obtained from the bivariate regression of Y on $\tilde{D}$, where $\tilde{D}$ is defined as $M_X D$ (i.e. the residuals from regressing D on X). Your expression should be a function only of Y and $\tilde{D}$.

(c) Show that the expressions from parts (a) and (b) are equivalent. (Hint: Recall the FWL theorem.)

(d) Beginning with the expression for $\hat{\tau}_{OLS}$ from part (b), derive an expression for $\hat{\tau}_{OLS}$ in terms of a weighted sum of the τ_X's (i.e. the stratum-specific treatment effects). If you have not yet done so, it will be helpful to write your expression from part (b) in non-matrix terms, which should be simple given its bivariate form. You will then need to algebraically manipulate this expression in order to obtain a weighted sum of the τ_X's. What are the weights?

(e) Recall the weights used in the subclassification estimator, which is unbiased for the true τ_{ATE}. Are these weights equal to or different than those implied by OLS? If so, how?

Problem 2

Use the data `dataQ2.csv` to estimate the ATE using:

i) Regression assuming the following model: $y_i = \alpha + \tau D_i + \epsilon_i$

ii) Regression assuming the following model: $y_i = \alpha + \tau D_i + \beta x_i + \epsilon_i$

iii) Subclassification

iv) Matching, using one match per treated unit (with replacement), setting `ties = TRUE`

v) Matching, using one match per treated unit (with replacement), setting `ties = FALSE`

In these data, selection on the observables is satisfied. With the insights from Problem 1, answer the following questions.

(a) Present your estimates of the ATE with standard errors for the five approaches.

(b) Why are the estimates different (or the same) across the approaches?

(c) Under what conditions would each approach be an unbiased estimate of the ATE? What method(s) deliver an unbiased estimate of the ATE in this case?

(d) Reproduce (exactly) the estimate of the ATE from the second regression by calculating and averaging (with appropriate weights) the within-stratum treatment effects.

Problem 3

This question will make use of two datasets, `simdata1.csv` and `simdata2.csv`. Each dataset is simulated and contains a treatment vector D, a response variable vector Y, and a covariate vector X. The assumption of selection on the observables is also met by both datasets, and the response variable was generated according to the following model: $y_i = \alpha + \tau D_i + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \epsilon_i$, with $\epsilon \sim \mathcal{N}(0, 0.25)$, $\alpha = 0$, $\tau = 10$, $\beta_1 = 5$, $\beta_2 = 1$, and $\beta_3 = -0.05$. However, the joint distribution of D and X is different in each dataset.

Because the data are simulated, we happen to know the true relationship between Y and D and X. However, in real-world situations, we will not know such true relationships. Even if we know that we must condition on X to achieve ignorability of the treatment, for instance, we will probably not know the functional form by which Y relates to X; you can think of the specification used for the simulated DGP in this problem as some arbitrary, complicated functional form that we are unlikely to figure out in the real world. We will explore the consequences of this for regression and matching estimators under different overlap conditions.

(a) Using `simdata1.csv`, report balance statistics for X across the treatment and control groups, and overlay its density for the treatment and control groups.

(b) Estimate τ using: (1) OLS regression with a bivariate model (i.e. $y_i = \alpha + \tau D_i + u_i$); (2) OLS regression with a model that controls for the covariate (i.e. $y_i = \alpha + \tau D_i + \beta_1 x_i + u_i$); (3) OLS regression that specifies the true model (i.e. $y_i = \alpha + \tau D_i + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + u_i$); and (4) a nearest-neighbor (with replacement) matching estimator. What do you notice?

(c) Repeat (a) and (b) with `simdata2.csv`, and comment on the differences. Can the matching approach balance treatment and control groups on x? How are the results affected for each estimation given these new data? Why?

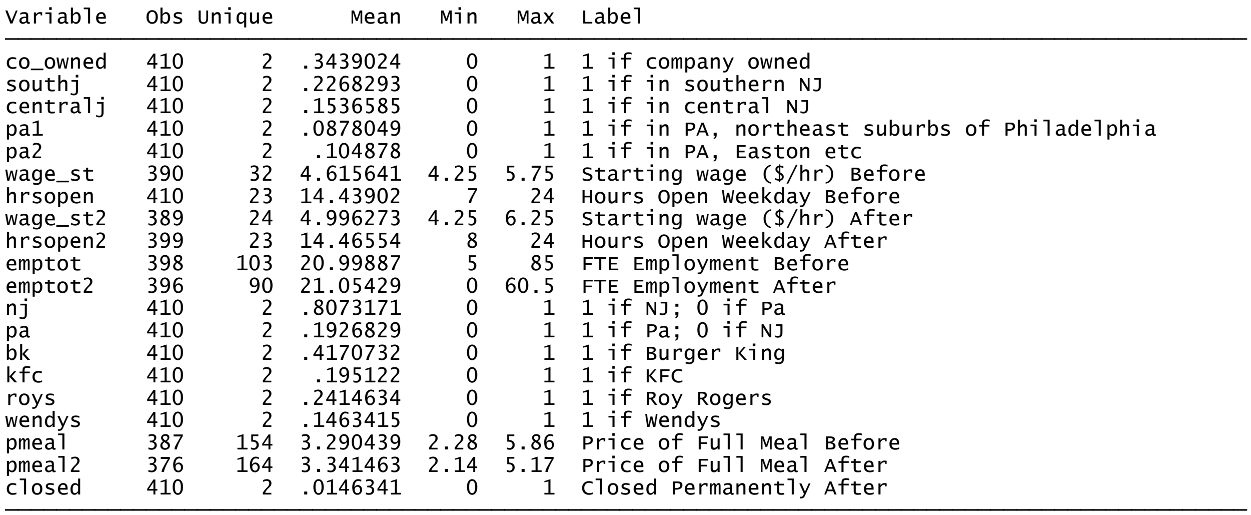
Problem 4

Now, you will replicate results in the following article:

> Card, D. and A. B. Krueger (1994), "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania," *American Economic Review*, vol. 84, 772-793.

You'll want to download the original paper and data (`card_krueger.dta`) from Piazza. Card and Krueger are interested in estimating the impact of minimum wage on teenage employment. Conventional economic wisdom states that raises in minimum wages hurt employment, especially teenage employment since teenage wages are often set at the minimum wage. Such is the main argument of those who oppose raising minimum wages. However, empirical analysis has failed to find evidence of employment responses to raises in minimum wages. In 1992, New Jersey's minimum wage increased from \$4.25 to \$5.05 while the minimum wage in Pennsylvania remained at \$4.25. The authors used data on employment at fast-food establishments in New Jersey and Pennsylvania before and after the increase in the minimum wage to measure the impact of the increase in minimum wage on teenage employment.

All variables in the dataset are listed below.



```
Variable   Obs Unique      Mean   Min   Max  Label
------------------------------------------------------------------------------------
co_owned   410      2  .3439024     0     1  1 if company owned
southj     410      2  .2268293     0     1  1 if in southern NJ
centralj   410      2  .1536585     0     1  1 if in central NJ
pa1        410      2  .0878049     0     1  1 if in PA, northeast suburbs of Philadelphia
pa2        410      2   .104878     0     1  1 if in PA, Easton etc
wage_st    390     32  4.615641  4.25  5.75  Starting wage ($/hr) Before
hrsopen    410     23  14.43902     7    24  Hours Open Weekday Before
wage_st2   389     24  4.996273  4.25  6.25  Starting wage ($/hr) After
hrsopen2   399     23  14.46554     8    24  Hours Open Weekday After
emptot     398    103  20.99887     5    85  FTE Employment Before
emptot2    396     90  21.05429     0  60.5  FTE Employment After
nj         410      2  .8073171     0     1  1 if NJ; 0 if Pa
pa         410      2  .1926829     0     1  1 if Pa; 0 if NJ
bk         410      2  .4170732     0     1  1 if Burger King
kfc        410      2   .195122     0     1  1 if KFC
roys       410      2  .2414634     0     1  1 if Roy Rogers
wendys     410      2  .1463415     0     1  1 if Wendys
pmeal      387    154  3.290439  2.28  5.86  Price of Full Meal Before
pmeal2     376    164  3.341463  2.14  5.17  Price of Full Meal After
closed     410      2  .0146341     0     1  Closed Permanently After
------------------------------------------------------------------------------------
```

1. Replicate Table 2 of Card and Krueger (1994) using t-tests, assuming unequal variance. Can you successfully replicate it? If not, explain why you think you cannot.
2. Based on the results you report in your replicated Table 2, what is the major finding with regard to FTE before and after the rise of the minimum wage?
3. Replicate Figure 1. Explain what is going on after the minimum wage legislation.
4. Replicate Table 4 by carefully generating relevant variables and running regressions. Interpret your coefficient estimates for the variables of "New Jersey dummy" and "Initial wage gap." Be

sure to interpret the statistical and substantive significance.

5. Explain the DID identification strategy in the context of this example. Do you think it is credible? Why or why not? Be sure to think hard about potential violations, and then decide what you think about how much you'd trust these results. How could you improve the study?

Problem 5

We observe random samples of campaign contributions from two groups of voters (call them red and blue) in two different periods. We don't necessarily observe any given person in both periods (we can think of this as the repeated cross-section case discussed in class). We sample and observe $N_{g,t}$ individuals from group $g \in \{red, blue\}$ at time $t \in \{0, 1\}$. Let G_i be the group individual i in the sample belongs to, with the convention that $G_i = 1$ means the individual is in the red group, and T_i be the period in which we observe i's contributions.

As in a difference-in-difference scenario, assume that in the second period, the red group received a treatment (for example, a change in contribution rules for the red party). In class, we assumed a very particular model of the dependence between outcomes, time periods, treatment status, and unobservables *in the absence of treatment.* Instead, suppose we posit a more general model. Individual i in the sample has an unobservable determinant of contributions in the absence of treatment. Call this unobservable determinant ϵ_i, and think of it as the strength of partisan sentiment. In keeping with the potential outcomes framework, we let the relationship between partisan sentiment, time period, and potential outcomes (contributions) *in the absence of treatment* be as general as possible:

$$Y_{0i} = h_g(\epsilon_i, T_i) \tag{1}$$

That is to say, each group of voters $g \in \{red, blue\}$ might have a different function linking time and sentiment to contributions in the absence of treatment, but we are willing to assume that the function is the same for everyone in a given group (for instance, you might suppose that blue workers' contributions increase less rapidly as their partisan sentiment rises than contributions for red workers). Since, in the DID setup, people were not randomized into the two groups, we have to presume that the distribution of the unobservables differs across groups. In other words, the distribution of ability ϵ_i in group g at time t has a cdf $F_{g,t}(\epsilon_i)$ and density $f_{g,t}(\epsilon_i)$.

We leave the outcome under treatment (which is only observable for the red group in $t = 1$) unrestricted; that is, $Y_{1i}(t)$ is the period t outcome after having been treated, but we don't impose any conditions on it, other that its being independent across i's. In other words, we don't write down a function for the Y_{1i}s the way we did for the Y_{0i}s in equation (1).

(a) Recall that the ATT in the DID framework is expressed as

$$ATT = E[Y_{1i}(1) - Y_{0i}(1) \big| D_i = 1] = E[Y_{1i}(1) \big| D_i = 1] - E[Y_{0i}(1) \big| D_i = 1]$$

Using the notation described above, express the ATT terms of the h's, f's, ϵ's, and counterfactual outcome under treatment. (Hint: recall the definition of expected value: $E(X) = \int X f(x) dx$. This problem simply asks you to rewrite the second expectation in the ATT expression as an integral. The first expectation in the ATT expression can be kept as is.)

(b) The usual DID estimator is

$$\hat{\beta} = \{\bar{Y}_{red,1} - \bar{Y}_{red,0}\} - \{\bar{Y}_{blue,1} - \bar{Y}_{blue,0}\}$$

We could obtain this directly from the means themselves, or by running a regression of Y_i on a constant, T_i, G_i and $T_i \times G_i$, as we saw in lecture. Express the estimator in terms of the functions, sample sizes, and random variables given—i.e., rewrite the latter three $\bar{Y}$s in terms of equation (1), but keep the first $\bar{Y}$ (outcome under treatment) unrestricted.

(c) What is the probability limit of this estimator? Your answer should in terms of the h's and $f_{g,t}$'s for all terms referring to outcomes in the absence of treatment. What do you have to assume about the sample sizes $N_{g,t}$ to derive this plim?

(d) Does this equal the expression for ATT derived in (a)?

(e) Without assuming anything else about the f's and h's, state a necessary and sufficient condition such that plim($\hat{\beta}$) =ATT. (Hint: this is not a difficult part of the question; your assumption should simply equate a term from the ATT expression in (b) with terms from the plim expression in (c).)

(f) What did we call this condition (assumption) in class? In this general case in which we don't restrict the form of the h function, does this condition have a behavioral (structural) interpretation? If so, what is it?

(g) Suppose you don't want to assume the condition above directly because you are not clear about what it implies about the h's and f's. Rather, you want to understand what would have to be true about the h's and f's in order for the condition to be satisfied, so that you can estimate the ATT with the DID estimator. You start by making the plausibly uncontroversial assumption that the functions $h_1(\cdot)$ and $h_0(\cdot)$ are the same—that is, that the mapping between partisan sentiment, time, and contributions (in the absence of treatment) is the same across groups. Call this assumption (A1). Can you now conclude that plim($\hat{\beta}$) = ATT? In other words, is the condition from (d) now satisfied? If so, prove your result. If not, state why this new assumption does not suffice.

(h) Suppose, instead, that you choose to make simplifying assumptions about the f's (distributions of unobservable ability in the samples), without restricting the h's. You start by assuming that $f_{g,1}(\epsilon) = f_{g,0}(\epsilon)$ for all g. (For notational convenience, we can now rewrite the fs with only one subscript, for group.) Call this assumption (A2). Explain what this assumption means substantively for the example of red and blue workers. Using this assumption, can you now conclude that $\text{plim}(\hat{\beta}) = \text{ATT}$? Why or why not?

(i) Now consider the special case in which $h_g(\epsilon_j, 1) = h_g(\epsilon_j, 0) + k$ for each group. What does this condition mean substantively? Under this condition together with (A2), does our condition hold? Can we use the DID estimator to get at the ATT?

(j) What if you combined (A2)—that the distribution of (unobserved) partisan sentiment strength is constant within group over time—with (A1), that the functions $h_1(\cdot)$ and $h_0(\cdot)$ are the same? Can you conclude, with (A1) and (A2) together (but without the assumption in (i)) that the necessary and sufficient condition for $\text{plim}(\hat{\beta}) = \text{ATT}$ holds? If so, prove your result. If not, state why not, and give one additional assumption about the f's that is needed to be able to conclude that $\text{plim}(\hat{\beta}) = \text{ATT}$. Does this additional assumption strike you as reasonable?

(k) A researcher suggests that the assumption $f_{red,1}(\epsilon) = f_{blue,1}(\epsilon)$ (call it (A3)), added to (A1), would let you estimate ATT consistently with $\{\bar{Y}_{red,1} - \bar{Y}_{blue,1}\}$. Show this.

(l) Explain, in social scientific terms, the meaning of assumption (A3). Also explain why (A3) is a much more problematic assumption, in behavioral terms, than (A2).

(m) Taking a slightly different approach, suppose we assume that 1) the h's are the same for both groups and 2) h is separable in time effects and the strength of partisan sentiment:

$$h(\epsilon_i, t) = h(t) + g(\epsilon_i)$$

Call this assumption (A1′) (a stronger version of (A1)). Would (A1′) and (A2) together make the DID estimator consistent? What other assumption would you need?

(n) Suppose, in addition to (A1′), we assumed (A4), $g(\epsilon) = \epsilon$, i.e., sentiment comes in linearly into the wage function in the absence of treatment. What minimal condition could you assume about the distributions of the ϵ's for (A1′) and (A4) together to justify DID? Is this weaker or stronger than (A2)? Than (A3)?