ECON 31720: Problem Set 1

Steven Buschbach

February 6, 2023

Problem 1

Suppose that $D \in \{0, 1\}$ is a binary treatment variable and $Y(0), Y(1)$ are potential outcomes for an outcome variable, Y. Assume that selection on observables holds, so that $Y(0), Y(1)$ is independent of D, conditional on X. Let $p(x) \equiv \mathbb{P}[D = 1 \mid X = x]$ and $P \equiv p(X)$. Let

$$W \equiv D + (1 - D)\left(\frac{P}{1 - P}\right)$$

Consider the weighted linear regression of Y on D and a constant, with weights W. That is,

$$(\beta_0, \beta_1) = \underset{b_0, b_1}{\operatorname{argmin}} \mathbb{E}\left[W\left(Y - b_0 - b_1 D\right)^2\right].$$

True, False, or Uncertain: β_1 is equal to the average treatment on the treated (ATT).

Solution
This is **true**. To see why, first take the first order conditions of the minimization problem:

$$\begin{aligned}
[\beta_0] : 0 &= \mathbb{E}[-2W(Y - \beta_0 - \beta_1 D)] \\
0 &= \mathbb{E}[WY] - \beta_0 \mathbb{E}[W] - \beta_1 \mathbb{E}[WD] \\
[\beta_1] : 0 &= \mathbb{E}[-2WD(Y - \beta_0 - \beta_1 D)] \\
0 &= \mathbb{E}[WYD] - \beta_0 \mathbb{E}[WD] - \beta_1 \mathbb{E}[WD^2]
\end{aligned}$$

Since D is binary, then $D^2 = 1$. And from the definition of W, we have

$$WD = D^2 + (1 - D)D\left(\frac{P}{1 - P}\right) = D$$

Using this fact to rewrite the first order conditions:

$$\begin{aligned}
\mathbb{E}[WY] &= \beta_0 \mathbb{E}[W] + \beta_1 \mathbb{E}[D] \\
\mathbb{E}[DY] &= \beta_0 \mathbb{E}[D] + \beta_1 \mathbb{E}[D]
\end{aligned}$$

Solving this linear system gives

$$\begin{aligned}
\beta_0 &= \frac{\mathbb{E}[Y(W - D)]}{\mathbb{E}[W - D]} \\
\beta_1 &= \frac{\mathbb{E}[DY]}{\mathbb{E}[D]} - \beta_0 \\
&= \mathbb{E}[Y(1) | D = 1] - \beta_0
\end{aligned}$$

Since ATT is defined as $\mathbb{E}[Y(1)|D=1] - \mathbb{E}[Y(0)|D=1]$, it remains to show that $\beta_0 = \mathbb{E}[Y(0)|D=1]$. To do so, we use the law of iterated expectations to rewrite our result for β_0 as

$$\begin{aligned}\beta_0 &= \frac{\mathbb{E}_X\big[\mathbb{E}[Y(W-D)|X]\big]}{\mathbb{E}_X\big[\mathbb{E}[W-D|X]\big]} \\ &= \frac{\mathbb{E}_X\big[\mathbb{E}[Y(1)(W-D)|D=1,X]\mathbb{E}[D|X] + \mathbb{E}[Y(0)(W-D)|D=0,X](1-\mathbb{E}[D|X])\big]}{\mathbb{E}_X\big[\mathbb{E}[W-D|D=1,X]\mathbb{E}[D|X] + \mathbb{E}[W-D|D=0,X](1-\mathbb{E}[D|X])\big]}\end{aligned}$$

Note that from the definition of W, we have

$$\begin{aligned}(W-D|D=1,X) &= 0 \\ (W-D|D=0,X) &= \frac{p(X)}{1-p(X)}\end{aligned}$$

Plugging this in and using the selection on observables assumption and the fact that $p(X) = \mathbb{E}[D|X]$, we get

$$\begin{aligned}\beta_0 &= \frac{\mathbb{E}_X\big[\frac{p(X)}{1-p(X)}\mathbb{E}[Y(0)|D=0,X](1-\mathbb{E}[D|X])\big]}{\mathbb{E}_X\big[\frac{p(X)}{1-p(X)}(1-\mathbb{E}[D|X])\big]} \\ &= \frac{\mathbb{E}_X\big[p(X)\mathbb{E}[Y(0)|X]\big]}{\mathbb{E}[D]}\end{aligned}$$

Note that this implicitly required an overlap assumption that $p(X) \in (0,1)$. Now we use Bayes' rule to simplify:

$$\begin{aligned}\beta_0 &= \frac{\int_X p(x)\mathbb{E}[Y(0)|X=x]dF_X(x)}{\mathbb{E}[D]} \\ &= \frac{\int_X p(x)\mathbb{E}[Y(0)|X=x]\frac{\mathbb{E}[D]dF_{X|D=1}(x)}{\mathbb{E}[D|X=x]}}{\mathbb{E}[D]} \\ &= \int_X \mathbb{E}[Y(0)|X=x]dF_{X|D=1}(x) \\ &= \mathbb{E}[Y(0)|D=1]\end{aligned}$$

So we have

$$\beta_1 = \mathbb{E}[Y(1)|D=1] - \beta_0 = \mathbb{E}[Y(1)|D=1] - \mathbb{E}[Y(0)|D=1] = \text{ATT}$$

which shows that the statement is true.

Problem 2

Let D be a binary treatment and let $Y(0)$ and $Y(1)$ be associated potential outcomes. Let X be another observables. Suppose that D is not independent of $(Y(0), Y(1))$ either unconditionally, or conditional on X. Let $\mu^* = \mathbb{E}[Y(1) - Y(0)]$ be the average treatment effect. Consider the quantities:

$$\mu_1 \equiv \mathbb{E}[Y \mid D = 1] - \mathbb{E}[Y \mid D = 0] \quad \text{(uncontrolled contrast)}$$
$$\mu_2 \equiv \mathbb{E}[\mathbb{E}[Y \mid D = 1, X] - \mathbb{E}[Y \mid D = 0, X]] \quad \text{(controlled imputation)}$$

Construct an economic story with a numerical example that produces $|\mu_2 - \mu^*| > |\mu_1 - \mu^*|$. Explain the relevance to selection on observables.

Solution
Intuitively, the first quantity μ_1 makes no attempt to control for selection, and the second quantity μ_2 attempts to control for potential selection in Y with observed selection in X. This can be seen more formally by noting that we can rewrite these expressions as

$$\mu_1 = \int_X \mathbb{E}[Y(1)|D = 1, X = x]dF_{X|D=1}(x) - \int_X \mathbb{E}[Y(0)|D = 0, X = x]dF_{X|D=0}(x)$$
$$\mu_2 = \int_X \mathbb{E}[Y(1)|D = 1, X = x]dF_X(x) - \int_X \mathbb{E}[Y(0)|D = 0, X = x]dF_X(x)$$

The first quantity only the conditional distribution of X given treatment, whereas the second quantity uses the unconditional distribution of X, to account for any selection into treatment based on X. So an economic story that produces $|\mu_2 - \mu^*| > |\mu_1 - \mu^*|$ will require that selection in X and selection in Y that counteract each other.

Consider the following story. We have two binary covariates, one observed X and one unobserved Z. Both are assigned by i.i.d. draws from the distribution

$$X \sim \text{Bernoulli}(1/2)$$
$$Z \sim \text{Bernoulli}(1/2)$$

Not receiving treatment results in a universal baseline outcome of zero, i.e., $Y(0) = 0$ for everyone. For those with $Z = 1$, the treatment is always harmful with $Y(1) = -1$. For those with $Z = 0$, their potential outcomes depend on X as specified in the table below:

	X=0	X=1
D=0	Y(0) = 0	Y(0) = 0
D=1	Y(1) = 0	Y(1) = 1

Suppose that people observe both their Z and their X, and they select into treatment based on two steps. First, they always choose no treatment if $Z = 1$. Next, if $Z = 0$, then they select into treatment with $1/4$ probability if $X = 1$ and $3/4$ probability if $X = 0$. So on average $1/2$ of those with $Z = 0$ select into treatment, unconditional on X.

This setup gives the following observed data: no treatment is selected by $3/4$ of the population (all of those with $Z = 1$ and one-half of those with $Z = 0$). Treatment is selected by the other $1/4$ of the population. By Bayes' rule, the conditional expectation of the observed covariate X given treatment status is:

$$\mathbb{E}[X|D = 1] = \frac{\Pr[D = 1|X = 1]\Pr[X = 1]}{\Pr[D = 1]} = \frac{[(0)(1/2) + (1/4)(1/2)](1/2)}{(1/4)} = \frac{1}{4}$$
$$\mathbb{E}[X|D = 0] = \frac{\Pr[D = 0|X = 1]\Pr[X = 1]}{\Pr[D = 0]} = \frac{[(1)(1/2) + (3/4)(1/2)](1/2)}{(3/4)} = \frac{7}{12}$$

The true ATE is given by

$$\begin{aligned}\mu^* &= \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] \\ &= \left(\frac{1}{2}(-1) + \frac{1}{4}(0) + \frac{1}{4}(1)\right) + \left(\frac{1}{2}(0) + \frac{1}{4}(0) + \frac{1}{4}(0)\right) \\ &= -\frac{1}{4}\end{aligned}$$

The uncontrolled contrast is calculated as

$$\begin{aligned}\mu_1 &= \mathbb{E}[Y(1)|D=1] - \mathbb{E}[Y(0)|D=0] \\ &= \left(\frac{1}{4}(1) + \frac{3}{4}(0)\right) - \left(\frac{7}{12}(0) + \frac{5}{12}(0)\right) \\ &= \frac{1}{4}\end{aligned}$$

The controlled imputation contrast is calculated as

$$\begin{aligned}\mu_2 &= \mathbb{E}[\mathbb{E}[Y(1)|D=1,X] - \mathbb{E}[Y(0)|D=0,X]] \\ &= \left(\frac{1}{2}(1) + \frac{1}{2}(0)\right) - \left(\frac{1}{2}(0) + \frac{1}{2}(0)\right) \\ &= \frac{1}{2}\end{aligned}$$

This means that the uncontrolled contrast is closer to the true ATE than the controlled imputation contrast, so we have $|\mu_2 - \mu^*| > |\mu_1 - \mu^*|$. The intuition is that selection on X was somewhat offsetting the selection on Z. By controlling for X, we controlled away this offset and made the estimator even more biased.

The lesson here is that the controlling that occurs in selection on observables is not always "harmless." In certain cases when selection on the observable variables offsets selection on unobservable variables, then methods like imputation can exacerbate the problem. So these methods must be used with caution.

Problem 3

Suppose that we have a sample of data $\{(Y_i, D_i, X_i)\}_{i=1}^n$. Let $\hat{\mu}_d(x)$ denote an estimator of $\mathbb{E}[Y_i \mid D_i = d, X_i = x]$. Compare two imputation estimators of the average treatment effect:

$$\hat{\alpha}_1 \equiv \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)$$

$$\hat{\alpha}_2 \equiv \frac{1}{n}\sum_{i=1}^{n} D_i\left(Y_i - \hat{\mu}_0(X_i)\right) + (1 - D_i)\left(\hat{\mu}_1(X_i) - Y_i\right)$$

Provide intuitive descriptions of both $\hat{\alpha}_1$ and $\hat{\alpha}_2$. When will they be equal if $\hat{\mu}_d(x)$ is computed from a linear regression(s)?

Solution

The first estimator of ATE, $\hat{\alpha}_1$, calculates the average difference between the two conditional potential outcome estimators for all observations in the sample. That is, it is the sample average difference between predicted outcome under treatment and predicted outcome under no treatment.

The second estimator of ATE, $\hat{\alpha}_2$, on the other hand, replaces one of the conditional potential outcome estimators with the observed outcome. So rather than estimating the ATE as the average difference in predicted potential outcomes, it uses the difference between the observed outcome and predicted unobserved outcome for each observation. For observations receiving treatment, this is $Y_i - \hat{\mu}_0(X_i)$, and for observations not receiving treatment, this is $\hat{\mu}_1(X_i) - Y_i$.

The ATE estimators $\hat{\alpha}_1$ and $\hat{\alpha}_2$ will be equal if and only if the following condition holds:

$$\left(\frac{1}{n}\sum_{i=1}^{n} D_i Y_i\right) + \left(\frac{1}{n}\sum_{i=1}^{n}(1 - D_i)Y_i\right) = \left(\frac{1}{n}\sum_{i=1}^{n} D_i\hat{\mu}_1(X_i)\right) + \left(\frac{1}{n}\sum_{i=1}^{n}(1 - D_i)\hat{\mu}_0(X_i)\right) \qquad (\mathbf{a})$$

Condition (a) holds is if the average predicted value for treated observations is equal to the average observed value for treated observations, and the average predicted value for untreated observations is equal to the average observed value for untreated observations. Showing that condition (a) implies $\hat{\alpha}_1 = \hat{\alpha}_0$ is algebraically straightforward:

$$\begin{aligned}
\hat{\alpha}_1 &= \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(D_i(\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)) + (1 - D_i)(\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i))\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(D_i(Y_i - \hat{\mu}_0(X_i)) + (1 - D_i)(\hat{\mu}_1(X_i) - Y_i)\right) \\
&= \hat{\alpha}_2
\end{aligned}$$

Now assume that $\hat{\mu}_d(x)$ is computed from a linear regression(s). As a best linear predictor of the conditional expectation, the regression decomposes each observed Y_i into some estimate of the conditional mean of Y_i given X_i and D_i, plus a residual term. We can rewrite this decomposition as:

$$\begin{aligned}
Y_i &= \mathbb{E}[\widehat{Y_i|X_i}, D_i] + \hat{\epsilon}_i \\
Y_i &= D_i\hat{\mu}_1(X_i) + (1 - D_i)\hat{\mu}_0(X_i) + \hat{\epsilon}_i \\
D_iY_i + (1 - D_i)Y_i &= D_i\hat{\mu}_1(X_i) + (1 - D_i)\hat{\mu}_0(X_i) + \hat{\epsilon}_i \\
\frac{1}{n}\sum_{i=1}^{n} D_iY_i + \frac{1}{n}\sum_{i=1}^{n}(1 - D_i)Y_i &= \frac{1}{n}\sum_{i=1}^{n} D_i\hat{\mu}_1(X_i) + \frac{1}{n}\sum_{i=1}^{n}(1 - D_i)\hat{\mu}_0(X_i) + \frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i
\end{aligned}$$

By construction, an OLS regression that allows for a nonzero intercept has $\sum_{i=1}^{n} \hat{\epsilon}_i = 0$, so the last term is equal to zero and condition (a) holds in this scenario, ensuring that $\hat{\alpha}_1 = \hat{\alpha}_2$. This remains true in the case in which $\mu_1(x)$ and $\mu_0(x)$ are computed from one linear regression on the full data, as well as the case in which they are computed separately from linear regressions on the treated and untreated observations. In the former case, we have $\sum_{i=1}^{n} \hat{\epsilon}_i = 0$, and in the latter case, we have $\sum_{i=1}^{n} D_i \hat{\epsilon}_i = 0$ and $\sum_{i=1}^{n} (1 - D_i) \hat{\epsilon}_i = 0$.

To sum up, we will have $\hat{\alpha}_1 = \hat{\alpha}_2$ if condition (a) holds. This is necessarily true if the estimated conditional potential outcomes are computed with an OLS regression on either the full sample, or separately on treatment and control samples.

Problem 4

This question is about "Persecution Perpetuated: The Medieval Origins of Anti-Semitic Violence in Nazi Germany" by Nico Voigtländer and Hans-Joachim Voth, published in The Quarterly Journal of Economics in 2012. The paper, as well as the data and code used in the paper, are available on Canvas.

4(a)
Critically discuss the authors' empirical strategy for estimating the effect of medieval antisemitism on 20th century antisemitism. What approach do the authors take? What assumptions are needed to justify their strategy? What are the primary threats to their strategy? Do you find their results credible? Two paragraphs: one on approach, assumptions, threats, and one on evaluation.

Solution
The authors use three different versions of a selection on observables approach to estimate the effect of medieval antisemitism on six different measures of 20th century antisemitism in German cities. The first approach uses a linear regression imputation estimator controlling for log population, Jewish share of population, Protestant share of population, and log Jewish population. The second and third approaches are matching estimators. The second estimator uses propensity score matching, where propensity scores are estimated using the aforementioned covariates, and the third estimator is direct matching of cities by latitude and longitude. All three approaches require unconfoundedness and overlap assumptions. Unconfoundedness in the first two approaches means that the potential outcomes are independent of medieval pogroms (the treatment) given the four listed covariates. Unconfoundedness in the third approach means that the potential outcomes are independent of medieval pogroms given location. That is, conditional on city covariates or distance, the medieval pogroms were as good as random. The overlap assumption requires that the probability of a pogrom was between 0 and 1, exclusive, given the city covariates used. The main threat to identification is the potential failure of the unconfoundedness assumption. If some other unobserved factors are correlated with medieval pogroms and also impact 1920s antisemitic activities, then the parameters of interest will not be identified.

I do not find the methods particularly credible. One immediate problem is that, with the exception of location matching, all of the authors' control covariates are measured in the 1920s-1930s, well after the medieval pogroms. In a word, these covariates are not predetermined. The medieval pogroms may have had a causal effect on the covariates themselves, in which case the unconfoundedness assumption is invalid. Even if pogroms did not affect the control covariates (for example, the authors provide suggestive evidence that the share of Jews in the 20th century is unrelated to the medieval pogroms), then I would still assert that the selection on observables is untenable here. It will probably not be uncontroversial to suppose that many factors could be associated with medieval pogroms, outside of the four chosen demographic covariates from the 1920s and 1930s. If the pogroms themselves had such long-lasting impacts on anti-semitic activity, as the authors assert, then these other pogrom-associated factors likely would endure as well.

4(b)
Replicate column (1) of Table VI. Remember to read the table notes carefully and to consult the authors' Stata code. The nnmatch command in Stata is well-documented in a paper by Abadie, Drukker, Herr, and Imbens (2004, The Stata Journal). For the purposes of this problem you may use the bootstrap to compute standard errors (although the bootstrap is not valid.) For bonus points, try to implement the formulas given in the paper by Abadie et al.

Solution
I successfully replicated column (1) of Table VI. The output of my code is in the table below. For the two matching procedures, I use bootstrap standard errors (although I understand that it is not valid) with 4000 samples. For these, my bootstrapped standard errors are slightly larger than the standard errors from the "nnmatch" procedure done by the authors.

Replication of Table VI, Column 1 in Voigtlander and Voth (2012)

Panel	Variable Name	Estimate	Standard Error	Observations	Adjusted R2
A: Regression Imputation	Pogrom 1349	0.0607***	0.0226	320	0.0544
	Ln(Pop 1925)	0.039**	0.0152		
	Pct Jewish 1925	0.0135	0.0114		
	Pct Protestant 1925	0.0003	0.0004		
B: Demographic Matching	Pogrom 1349	0.0744***	0.0192	320	
C: Geographic Matching	Pogrom 1349	0.0819***	0.0267	320	

4(c)
Discuss any oddities or inconsistencies between how the authors describe their results in Table VI, column (1), and how they actually implemented them in their code.

Solution
I noticed two oddities in how the authors describe their results in Table VI, column (1), and how it's implemented in code. The most glaring inconsistency is that the authors describe their Panel B matching estimator as "propensity score matching", which it is definitely not. Rather, their code matches directly on the three demographic covariates, using the default inverse variance weighting matrix. No propensity scores are estimated.

The other inconsistency that I noticed is more subtle. For the geographic match in Panel C, the paper's exposition makes it seem as if a treated city would be matched to the closest untreated city. But this is not necessarily the case, because the authors use the default inverse variance weighting matrix in their code. Since there is more north/south variation in cities than east/west variation, longitude differentials count more than latitude differentials while matching.

4(d)
Evaluate the sensitivity of the authors' results to their choice of control variables.

Solution
I evaluated six alternative control variable specifications. For each specification, I refit the regression imputation model in Panel A and the covariate matching model in Panel B. The results are in the table below. The covariates used in these alternate specifications are listed in this footnote.[1]

I find that for minor tweaks to the covariates (e.g., switching to 1933 data, adding Catholic shares, or including labor market controls), the regression estimate shrinks somewhat. But my alternate specifications 5 and 6 show big increases in effect size for the regression imputation estimator. This may partly a result of sample size reduction: the authors tended to choose the controls that had less missing data. But it does suggest that there is some instability in estimates and significance depending on choice of control variables. There are many more potential specifications, and the authors have a large number of potential controls.

[1] I use the following alternate specifications

1. Add share of Catholic population to the original three controls.
2. Change population control and Jewish population share control to 1933 data, rather than the 1925 data the authors use.
3. Only use controls from 1925 religious data: percentage of Jewish population, percentage of Protestant population, and percentage of Catholic population.
4. Only use controls from 1925 labor market data: shares of population in agriculture, in industry, in blue collar work, and self-employed.
5. Only use controls that are predetermined, i.e., fixed before medieval pogroms. These are log 1300 population, whether the city is on a navigable river, and the ruggedness of the terrain.
6. Use all of the controls mentioned in previous five specifications together.

Evaluating Alternate Control Variable Specifications

Specification	Regression Estimate	Regression SE	Observations
1 - Include Catholic Share	0.0571**	0.0226	320
2 - Change to 1933 Demographics	0.0507**	0.0233	320
3 - Only Religion Variables	0.0624***	0.0232	320
4 - Only Labor Variables	0.0673*	0.0378	177
5 - Only Predetermined Variables	0.2565**	0.1079	46
6 - Fully Loaded Regression	0.381*	0.2126	42

Specification	Matching ATT Estimate	Matching ATT SE	Observations
1 - Include Catholic Share	0.0733***	0.0196	320
2 - Change to 1933 Demographics	0.0711***	0.0202	320
3 - Only Religion Variables	0.0657***	0.0194	320
4 - Only Labor Variables	0.0352	0.0268	177
5 - Only Predetermined Variables	0.0917	0.0902	46
6 - Fully Loaded Regression	0.0608	0.0936	42

Ultimately, they should have more thoroughly justified their three chosen controls. If it was due to missing due in many other potential controls, as I suspect, then they should say so.

4(e)
Implement propensity score matching estimators of the ATE, ATU, and ATT, using the same covariates as the authors do in panel B. I leave the specifics up to you, but you might consider nearest neighbor matching on the propensity score, and/or a blocking approach. Compare your estimates to the authors' estimates. For the purposes of this problem you may use the bootstrap to compute standard errors (although the bootstrap might not be valid, depending on your approach.)
Note: Optimization packages are not considered high-level commands for the purpose of this class, since they are not statistical in nature. You may (and should) use one to optimize a likelihood, e.g. for a logit estimator.

Solution
I fit a logistic regression to estimate propensity scores, and then I do nearest neighbor matching[2] to estimate ATT, ATU, and ATE. I bootstrap standard errors (although I understand this is not valid) with 4000 samples. The results are in the table below:

Propensity Score Matching Estimates

Parameter	Estimate	Standard Error
ATT	0.0722***	0.0208
ATU	0.0426*	0.0228
ATE	0.0641***	0.0108

The ATT result is approximately equal to the authors' ATT estimate in Table VI, Panel B, with a slightly larger standard error from the bootstrapping. The ATU is much smaller than the ATT and only significant at a 10% level. The ATE is a weighted average of the ATT and ATU.

[2]I somewhat arbitrarily choose the same matching specification as the authors. I match each observation to four neighbors of opposite treatment status using an inverse variance weighting matrix.

Problem 5

Consider the binary treatment potential outcomes model with $D \in \{0,1\}$ and $Y = DY(1)+(1-D)Y(0)$. For concreteness, suppose that D is whether one enrolls in a job training course, and Y is earnings at some point afterwards. Suppose that we also have a set of predetermined covariates, X. Our data consists of these variables for both workers who enrolled in the job training course, and those who did not.

5(a)
Suppose that the job training experiment accepts all applicants into the course. Using the definitions in the lecture notes, show that the selection on observables model is not falsifiable.

Solution
We follow the notation of the lecture notes on identification. Denote A to be an indicator for applying to the job training program. I take the experiment accepting all applicants to mean that $A = 1 \implies D = 1$ and $A = 0 \implies D = 0$. These mean that $A = D$. So redefine D to be an indicator for applying and enrolling, which are always the same. The observed data is

$$W \equiv (Y, D, X) \sim G$$

Each $\theta = F$ implies a joint distribution of $(Y(0), Y(1), D, X)$. The selection on observables model restricts the parameter space of θ to

$$\Theta \equiv \{F \in \mathcal{F} : (Y(0), Y(1)) \perp\!\!\!\perp D|X \text{ under } F\}$$

From lecture, we know that a model is falsifiable if there exists a known function $\tau : \mathcal{G} \to \{0,1\}$ such that the following two statements are true:

- $\tau(G) = 1$ implies that $\Theta^*(G) = \emptyset$ (i.e., the model is misspecified)
- $\tau(G) = 1$ for at least one $G \in \mathcal{G}$

where $\Theta^*(G)$ is the identified set for θ given that the θ-parameterized distribution G_θ matches the distribution of G, i.e., $\Theta^*(G) = \{\theta \in \Theta : G_\theta = G\}$. So to show that the selection on observables model is not falsifiable, we must show that function τ does not exist, which is equivalent to showing that $\Theta^*(G) \neq \emptyset$ for all $G \in \mathcal{G}$. This, in turn, is equivalent to showing that there exists some $\tilde{\theta} \in \Theta$ such that $G_{\tilde{\theta}} = G$ for all $G \in \mathcal{G}$.

To show this, fix G. Denote G_0 and G_1 to be the conditional distributions of the data given treatment and nontreatment, i.e.,

$$G_0 \equiv (Y|D = 0, X) = (Y(0)|D = 0, X)$$
$$G_1 \equiv (Y|D = 1, X) = (Y(1)|D = 1, X)$$

We will choose distribution $\tilde{\theta} = (Y(0), Y(1), D, X)$ so that

$$Y(0)|D = 1, X \sim G_0$$
$$Y(1)|D = 0, X \sim G_1$$
$$(Y, D, X) \sim G$$

This gives us

$$(Y(0)|D = 1, X) \stackrel{d}{=} (Y(0)|D = 0, X) \implies Y(0) \perp\!\!\!\perp D|X$$
$$(Y(1)|D = 1, X) \stackrel{d}{=} (Y(1)|D = 0, X) \implies Y(1) \perp\!\!\!\perp D|X$$

Further restrict $\tilde{\theta}$ so that

$$Y(0) \perp\!\!\!\perp Y(1)|D, X$$

which means that the marginal conditional independence of potential outcomes implies joint conditional independence of the potential outcomes. Taken together, we have

$$(Y(0), Y(1)) \perp\!\!\!\perp D|X$$

which is the selection on observables assumption. So we have $\tilde{\theta} \in \Theta^*(G)$. Since we had fixed G to be arbitrary, this means that $\Theta^*(G)$ is never empty, and the function $\tau(G)$ does not exist. So the selection on observables model is not falsifiable.

5(b)
Suppose that the job training experiment has the following structure. First, we open the program to everyone and collect a list of workers who apply to take the program. Then, we offer the program to a random subset of these applicants, but do not provide job training for any of the applicants not in this random subset. We collect data on the outcomes for workers who took the program, workers who applied to take the program but were not randomized in, and other workers who did not even apply to the program.

Explain how this structure could be used to falsify selection on observables.

Solution
Now, we no longer have $A = D$. Rather, we have $A = 0 \implies D = 0$ and $A = 1$ could result in $D = 0$ or $D = 1$. The fact that $A = 0 \implies D = 0$ means that

$$(Y(0)|A = 0, X) \stackrel{d}{=} (Y(0)|A = 0, D = 0, X) \stackrel{d}{=} (Y|A = 0, D = 0, X) \tag{1}$$

Since enrollment upon applying is random, we have

$$\begin{aligned} &(Y(0), Y(1)) \perp\!\!\!\perp D|A = 1 \\ \implies &(Y(0), Y(1)) \perp\!\!\!\perp D|A = 1, X \end{aligned}$$

where the implication comes from the fact that including X in the conditioning set does not change the independence given $A = 1$. This implies that

$$(Y(0)|A = 1, X) \stackrel{d}{=} (Y(0)|A = 1, D = 0, X) \stackrel{d}{=} (Y|A = 1, D = 0, X) \tag{2}$$

Assume that selection on observables in this instance means that we are controlling for selection into application (since conditional on application, enrollment is random). Since joint independence implies marginal independence, the selection on observables assumption implies $Y(0) \perp\!\!\!\perp A|X$, which in turn implies $(Y(0)|A = 1, X) \stackrel{d}{=} (Y(1)|A = 0, X)$. Combining this with (1) and (2), we see that this is equivalent to

$$(Y|A = 1, D = 0, X) \stackrel{d}{=} (Y|A = 0, D = 0, X)$$

This is a testable comparison of two groups: the distribution of outcomes for those who applied and did not enroll, given covariates, and the distribution of outcomes for those who did not apply, given covariates. These are functions of the data distribution G. So define the function

$$\tau(G) = \mathbf{1}[(Y|A = 1, D = 0, X) \stackrel{d}{\neq} (Y|A = 0, D = 0, X)].$$

Here, $\tau(G) = 1$ implies misspecification, and this can generally occur without onerous restrictions on $\mathcal{G}$. So the model is falsifiable. In practice, one can test for the above misspecification with an equality of distribution test, like the Kolmogorov-Smirnov two-sample test.

Problem 6

Consider a balanced panel data setting with time periods $t = 1, \ldots, T$. Suppose that $D_{it} \in \{0, 1\}$ is a binary treatment with associated potential outcomes $Y_{it}(0)$ and $Y_{it}(1)$. Assume that treatment is absorbing, so that $D_{it} = 1$ implies $D_{is} = 1$ for $s \geq t$. Let $E_i = \min\{t : D_{it} = 1\}$ denote the event time or cohort. Suppose that $\mathbb{P}[E_i = e] > 0$ for all $e = 2, \ldots, T$, as well as for $e = +\infty$, so that there are always-untreated units. Assume that common trends holds, so that

$$\mathbb{E}[Y_{it}(0) - Y_{is}(0) \mid E_i = e] = \mathbb{E}[Y_{it}(0) - Y_{is}(0) \mid E_i = e']$$

for all t, s, e, and e'.

Consider the random variable $\tilde{Y}_{it}$ defined as follows:

- Regress Y_{it} on a full set of time and cohort indicators in the subset of observations (i, t) with $D_{it} = 0$.
- For all (i, t) (regardless of D_{it}), use the previous regression to construct fitted values $\dot{Y}_{it}$ based on the time period and cohort.
- Let $\tilde{Y}_{it} \equiv Y_{it} - \dot{Y}_{it}$.

Let $\text{ATT}_t(e) = \mathbb{E}[Y_{it}(1) - Y_{it}(0) \mid E_i = e]$.

6(a)
Show that

$$\mathbb{E}\left[\tilde{Y}_{it} \mid E_i = e\right] = \begin{cases} 0, & \text{if } t < e \\ \text{ATT}_t(e), & \text{if } t \geq e \end{cases}$$

Solution
Denote the fitted values of the regression of Y_{it} on the full set of time and cohort dummies in the subset of observations with $D_{it} = 0$ as:

$$\dot{Y}_{it} = \sum_{s=1}^{T} \delta_s \mathbf{1}_{t=s} + \sum_{e=1}^{E} \gamma_e \mathbf{1}_{E_i=e}$$

First consider the case for $t < e$, i.e., those not yet treated as of t. We have

$$\begin{aligned} \mathbb{E}[\tilde{Y}_{it} | E_i = e] &= \mathbb{E}[Y_{it} - \dot{Y}_{it} | E_i = e] \\ &= \mathbb{E}[Y_{it} - \delta_t - \gamma_e | E_i = e] \\ &= 0 \end{aligned}$$

where the last equality comes from the zero mean of residuals in a regression.

Now consider $t \geq e$, i.e., those treated as of t. Choose some e', s such that $e' > t \geq e > s$. In words, e' is a time period after the time of interest t, and s is a time period prior to the treatment commencement of the cohort of interest e. Doing some adding and subtracting, we have

$$\begin{aligned} \mathbb{E}[\tilde{Y}_{it} | E_i = e] &= \mathbb{E}[Y_{it} - \dot{Y}_{it} | E_i = e] \\ &= \mathbb{E}[Y_{it} - Y_{is} | E_i = e] + \mathbb{E}[Y_{is} - \dot{Y}_{it} | E_i = e] \\ &= \mathbb{E}[Y_{it} - Y_{is} | E_i = e] + \mathbb{E}[(Y_{is} - \dot{Y}_{is}) + (\dot{Y}_{is} - \dot{Y}_{it}) | E_i = e] \end{aligned}$$

Since $e > s$, using the same argument about regression residuals as before, we have $\mathbb{E}[Y_{is} - \dot{Y}_{is} | E_i = e] = 0$. So the second expectation becomes $\mathbb{E}[\dot{Y}_{is} - \dot{Y}_{it} | E_i = e]$. Given the cohort of i, we know treatment status at each time and can rewrite in terms of potential outcomes. Doing this and some more adding and subtracting,

we get

$$\begin{aligned}\mathbb{E}[\tilde{Y}_{it}|E_i = e] &= \mathbb{E}[Y_{it}(1) - Y_{is}(0)|E_i = e] + \mathbb{E}[\dot{Y}_{is} - \dot{Y}_{it}|E_i = e] \\ &= \mathbb{E}[Y_{it}(1) - Y_{it}(0)|E_i = e] + \mathbb{E}[Y_{it}(0) - Y_{is}(0)|E_i = e] + \mathbb{E}[\dot{Y}_{is} - \dot{Y}_{it}|E_i = e]\end{aligned}$$

The first term is $\text{ATT}_t(e)$, and we can use the common trends assumption to rewrite the second term as $\mathbb{E}[Y_{it}(0) - Y_{is}(0)|E_i = e']$. So the whole thing becomes

$$\begin{aligned}\mathbb{E}[\tilde{Y}_{it}|E_i = e] &= \text{ATT}_t(e) + \mathbb{E}[Y_{it}(0) - Y_{is}(0)|E_i = e'] + \mathbb{E}[\dot{Y}_{is} - \dot{Y}_{it}|E_i = e] \\ &= \text{ATT}_t(e) + \mathbb{E}[Y_{it} - Y_{is}|E_i = e'] + \mathbb{E}[\dot{Y}_{is} - \dot{Y}_{it}|E_i = e]\end{aligned}$$

If we can change the conditioning set in the third term to $E_i = e'$, then we can combine the second and third terms and use another residual argument to zero them out. To do so, consider the third term and write:

$$\begin{aligned}\mathbb{E}[\dot{Y}_{is} - \dot{Y}_{it}|E_i = e] &= \mathbb{E}\left[\sum_{\tau=1}^{T}(\delta_\tau \mathbf{1}_{\tau=s} - \delta_\tau \mathbf{1}_{\tau=t}) + \sum_{d=1}^{E}(\gamma_d \mathbf{1}_{E_i=d} - \gamma_d \mathbf{1}_{E_i=d})\middle| E_i = e\right] \\ &= \mathbb{E}\left[\sum_{\tau=1}^{T}(\delta_\tau \mathbf{1}_{\tau=s} - \delta_\tau \mathbf{1}_{\tau=t})\middle| E_i = e\right] \\ &= \mathbb{E}\left[\sum_{\tau=1}^{T}(\delta_\tau \mathbf{1}_{\tau=s} - \delta_\tau \mathbf{1}_{\tau=t})\middle| E_i = e'\right] \\ &= \mathbb{E}[\dot{Y}_{is} - \dot{Y}_{it}|E_i = e']\end{aligned}$$

where the switch in cohort comes from noticing that δ_τ are invariant across cohorts. Plugging this back in, we get

$$\begin{aligned}\mathbb{E}[\tilde{Y}_{it}|E_i = e] &= \text{ATT}_t(e) + \mathbb{E}[Y_{it} - \dot{Y}_{it}|E_i = e'] - \mathbb{E}[Y_{is} - \dot{Y}_{is}|E_i = e'] \\ &= \text{ATT}_t(e)\end{aligned}$$

since residuals of $\tilde{Y}_{it}$ and $\tilde{Y}_{is}$ are mean zero prior to treatment commencement e'.

6(b)
Consider a pooled regression of $\tilde{Y}_{it}$ onto D_{it} and a constant. Let δ denote the population regression coefficient on D_{it}. Show that

$$\delta = \sum_{s=1}^{T}\sum_{e=1}^{T}\omega_s(e)\,\text{ATT}_s(e)$$

for some weights $\omega_s(e)$ that are non-negative and sum to one: $\sum_{s=1}^{T}\sum_{e=1}^{T}\omega_s(e) = 1$.

Solution
The regression is saturated, so we have

$$\begin{aligned}\delta &= \frac{\text{Cov}[\tilde{Y}_{it}, D_{it}]}{\text{Var}[D_{it}]} \\ &= \frac{\mathbb{E}[\tilde{Y}_{it}D_{it}] - \mathbb{E}[\tilde{Y}_{it}]\mathbb{E}[D_{it}]}{\text{Var}[D_{it}]} \\ &= \frac{\mathbb{E}[\tilde{Y}_{it}|D_{it} = 1]\mathbb{E}[D_{it}] - (\mathbb{E}[\tilde{Y}_i t|D_{it} = 1]\mathbb{E}[D_{it}] - \mathbb{E}[\tilde{Y}_{it}|D_{it} = 0](1 - \mathbb{E}[D_{it}]))\mathbb{E}[D_{it}]}{\text{Var}[D_{it}]} \\ &= \frac{\mathbb{E}[D_{it}](1 - \mathbb{E}[D_{it}])(\mathbb{E}[\tilde{Y}_{it}|D_{it} = 1] - \mathbb{E}[\tilde{Y}_{it}|D_{it} = 0])}{\text{Var}[D_{it}]} \\ &= \mathbb{E}[\tilde{Y}_{it}|D_{it} = 1] - \mathbb{E}[\tilde{Y}_{it}|D_{it} = 0] \\ &= \mathbb{E}_e\left[\mathbb{E}[\tilde{Y}_{it}|D_{it} = 1, E_i = e]\middle| D_{it} = 1\right] - \mathbb{E}_e\left[\mathbb{E}[\tilde{Y}_{it}|D_{it} = 0, E_i = e]\middle| D_{it} = 0\right]\end{aligned}$$

Since the second term conditions on $D_{it} = 0$ and treatment is absorbing, from the same mean-zero residual argument as before we can say that this expectation is equal to zero. So we can focus on the first term. We write

$$\begin{aligned}
\delta &= \mathbb{E}_e\left[\mathbb{E}[\tilde{Y}_{it}|D_{it}=1, E_i=e]\middle| D_{it}=1\right] \\
&= \sum_{e=1}^{T} \mathbb{P}[E_i=e|D_{it}=1]\mathbb{E}[\tilde{Y}_{it}|D_{it}=1, E_i=e] \\
&= \sum_{e=1}^{T}\sum_{s=1}^{T} \mathbb{P}[s|D_{is}=1, E_i=e]\mathbb{P}[E_i=e|D_{is}=1]\mathbb{E}[\tilde{Y}_{is}|D_{is}=1, E_i=e] \\
&= \sum_{e=1}^{T}\sum_{s=1}^{T} \underbrace{\mathbb{P}[s|D_{is}=1, E_i=e]\mathbb{P}[E_i=e|D_{is}=1]}_{\omega_s(e)} \mathrm{ATT}_s(e)
\end{aligned}$$

Consider the two probability terms that make up $\omega_s(e)$. The second, the share of cohort e is

$$\mathbb{P}[E_i=e|D_{is}=1] = \frac{N_e}{N}$$

The first probability term is just the share of observations at time s given $D_{is}=1$ and $E_i=e$. Since the panel is balanced, this is

$$\mathbb{P}[s|D_{is}=1, E_i=e] = \begin{cases} 0 & \text{if } e > s \\ \frac{1}{T-e+1} & \text{if } e \leq s \end{cases}$$

This means we have the ATT weights are

$$\omega_s(e) = \begin{cases} 0 & \text{if } e > s \\ \frac{N_e}{N}\frac{1}{T-e+1} & \text{if } e \leq s \end{cases}$$

These are clearly positive, and we have

$$\sum_{e=1}^{T}\sum_{s=1}^{T}\omega_s(e) = \sum_{e=1}^{T}\frac{N_e}{N}\left(\sum_{s=e}^{T}\frac{1}{T-e+1}\right) = 1$$

6(c)
Let $R_{it} \equiv t - E_i$ denote relative time, and let $D_{it}^j \equiv \mathbf{1}[R_{it}=j]$ be relative time indicators. Consider a pooled no-constant regression of $\tilde{Y}_{it}$ onto all post-treatment relative time indicators (so $\left\{D_{it}^j\right\}_{j=0}^{T-\min E_i}$) together with an indicator for relative time strictly less than 0 . Show that the coefficients on D_{it}^j can also be written as non-negative weighted averages of $\mathrm{ATT}_t(e)$.

Solution
Denote δ^j to be the coefficient on D_{ij}^j. Since the regression is saturated and there is no constant, we have

$$\begin{aligned}
\delta^j &= \mathbb{E}[\tilde{Y}_{it}|D_{it}^j=1] \\
&= \mathbb{E}[\tilde{Y}_{it}|t-E_i=j] \\
&= \mathbb{E}[\mathbb{E}[\tilde{Y}_{it}|E_i=e, t-e=j]] \\
&= \mathbb{E}[\mathbf{1}_{e+j\leq T}\times \mathrm{ATT}_{e+j}(e)] \\
&= \sum_{e=1}^{T}\frac{\mathrm{ATT}_{e+j}(e)}{\sum_{e=1}^{T}\mathbf{1}_{e+j\leq T}}
\end{aligned}$$

so we have non-negative weights

$$\omega_j(e) = \frac{1}{\sum_{e=1}^{T}\mathbf{1}_{e+j\leq T}}$$

6(d)
Discuss the content and significance of these result in the context of event studies.

Solution
We know that in event studies, static two-way fixed effects regressions will necessarily incorporate some negative weights, and dynamic two-way fixed effects regressions can only be given a positive interpretation if the regression is saturated and homogeneity is assumed. These results show that an imputation approach can get around these issues to estimate $ATT_t(e)$ and then aggregated in an somewhat sensible way. While the weights do feel somewhat arbitrary, that is the price we may for wanting to just run regressions.

Problem 7

Design a Monte Carlo experiment for an event study, and use it to illustrate the following points:

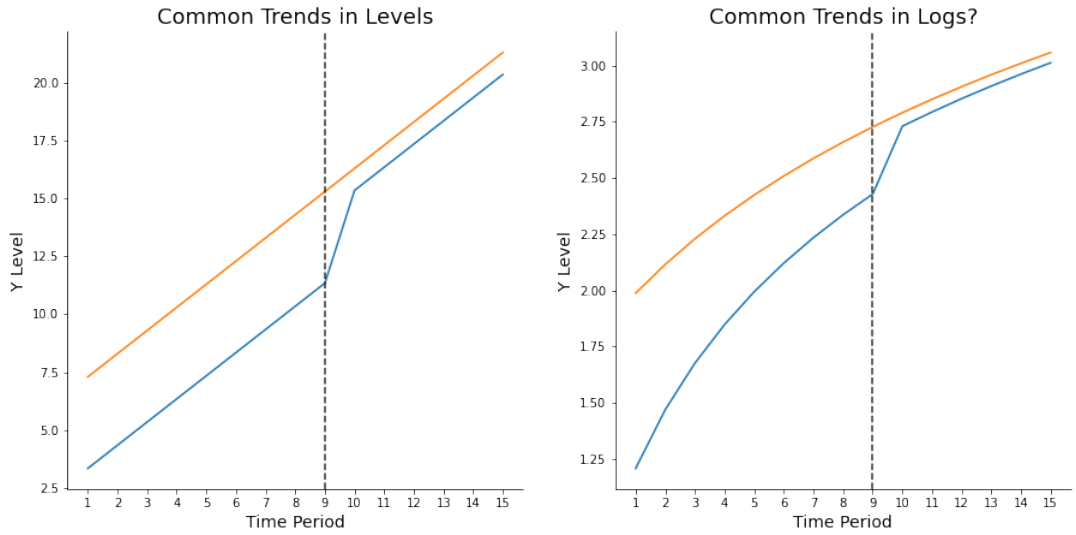
7(a)
Common trends will not generally hold for both Y_{it} and $\log(Y_{it})$ simultaneously.

Solution
For this and all subsequent simulations, potential outcomes are equal to

$$Y_{it}(D) = \gamma_t + \alpha_i + \beta_{it} D + \epsilon_{it}$$

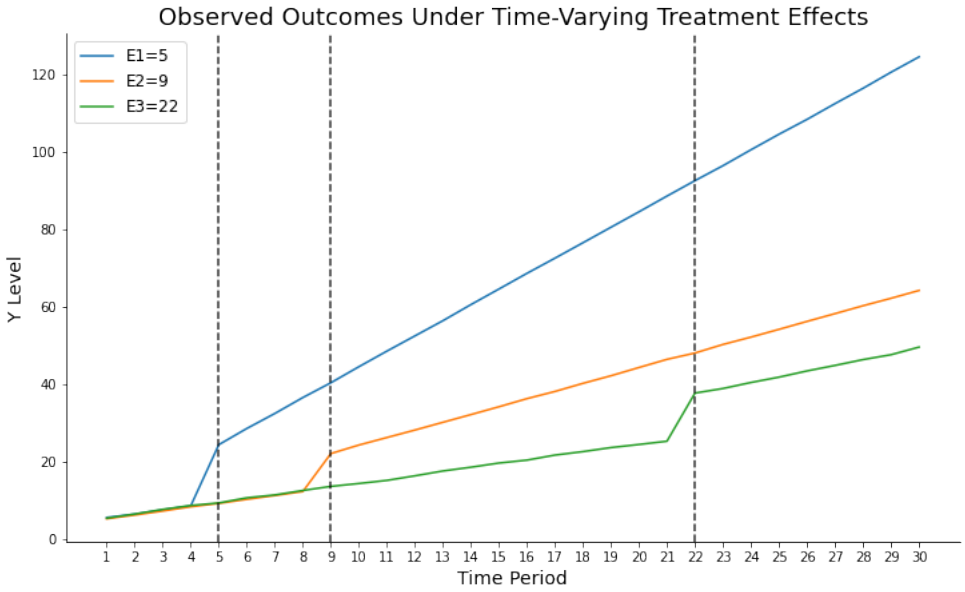
See the Q7 Jupyter notebook for exact simulation details.

We begin with a simple panel of $N = 2$ and $T = 15$. Unit 1 gets treated halfway through at time $t = 8$. The unit fixed effects are uniformly-distributed, and there is a linear time trend. (For some later simulations, I add noise to the time trend). The treatment effect is 3.0. For clarity, I suppress any noise ϵ_{it} in this simulation. One of the Monte Carlo samples is shown in the graph below. Common trends clearly holds for the linear case, but not for the log case.



7(b)
The static two-way fixed effects estimator will incorporate negatively-weighted treatment effects.

Solution
Consider a simulated event study with the following features. There is a balanced panel with $N = 300$ observed units over $T = 30$ time periods. There are three cohorts, one treated at time $E_1 = 5$, one treated at time $E_2 = 9$, and one treated at time $E_3 = 22$. These three cohorts consist of $N_1 = 150$, $N_2 = 100$, and $N_3 = 50$ units, respectively. There is a linear time trend with normally-distributed noise, and individual fixed effects distributed uniformly. Treatment effects depend on time and cohort, with later-treated cohorts having smaller treatment effects. A graph of average outcomes for units in the three cohorts is given below:



With this setup in mind, we run a regression with the static two-way fixed effects estimator, which uses unit dummies, time dummies, and one indicator for an observation having received treatment at that point in time. Despite all treatment effects set up to be positive, we find a negative estimate for the coefficient on the treatment indicator, shown in the table below. This means that there must be some negative weighting involved in this estimate

Evidence of Negative Weights in a Static TWFE Estimator

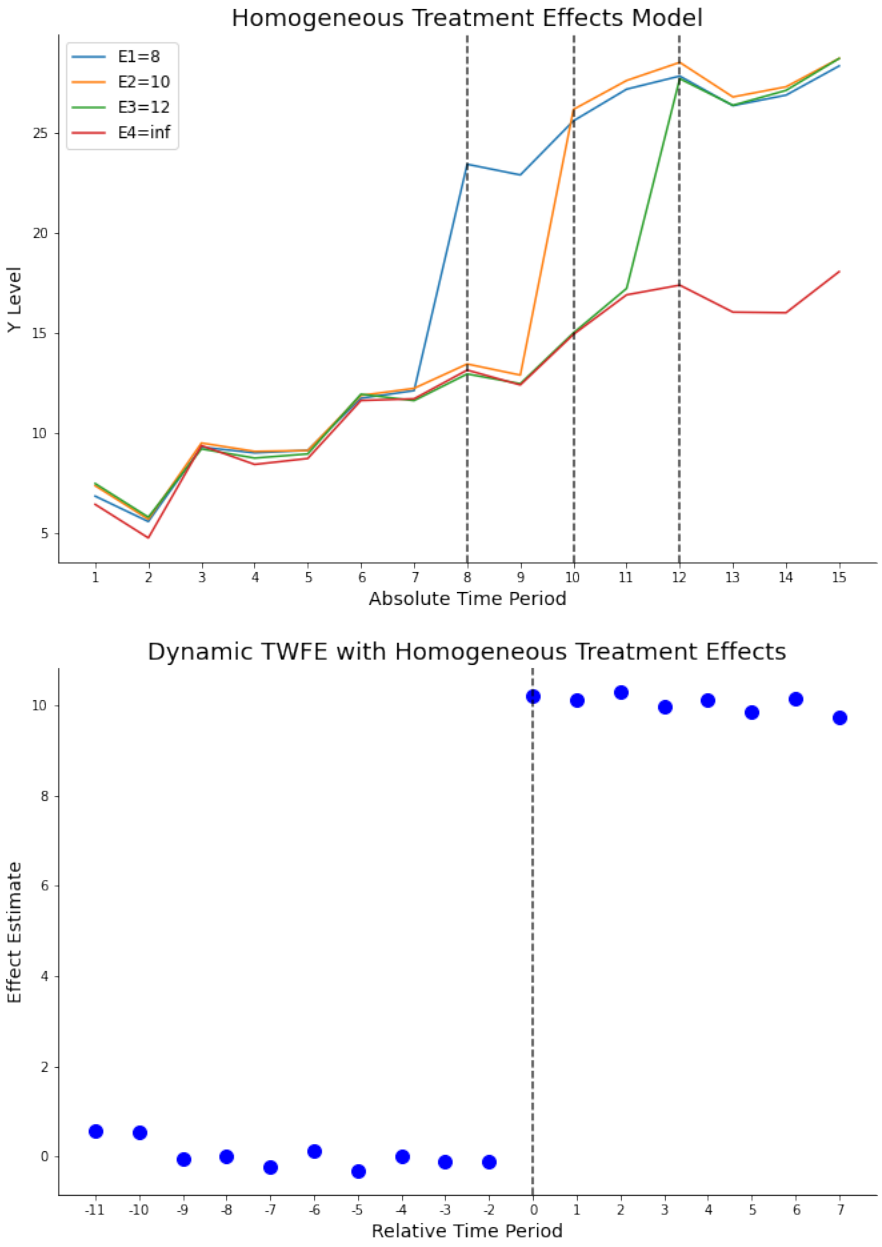
Coefficient	Value
ATT Estimate	-5.148
Intercept	72.607

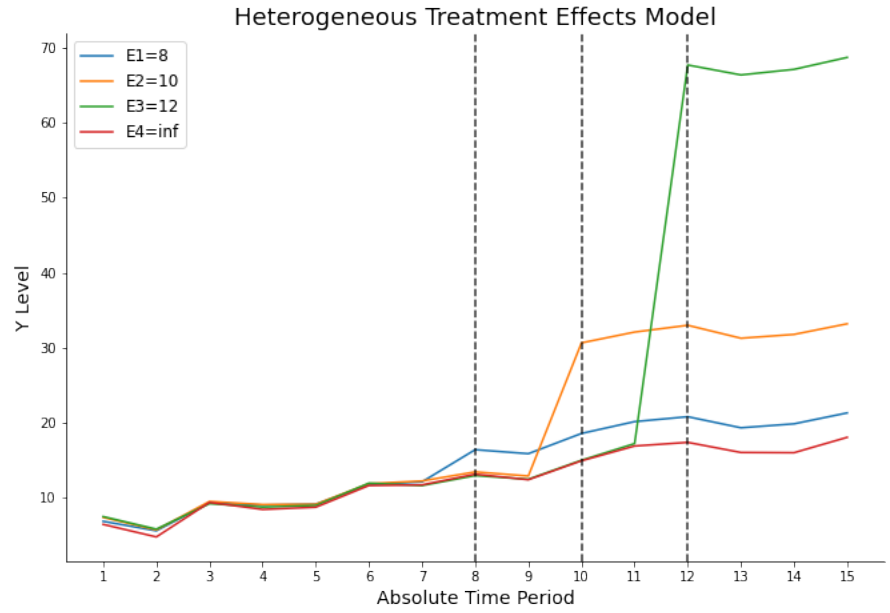
7(c)
The dynamic two-way fixed effects estimator will incorporate negatively-weighted treatment effects unless relative time treatment effects are homogeneous across cohorts. If there is such heterogeneity, then analyzing the coefficients on the leads may not be a good way to test for common trends.

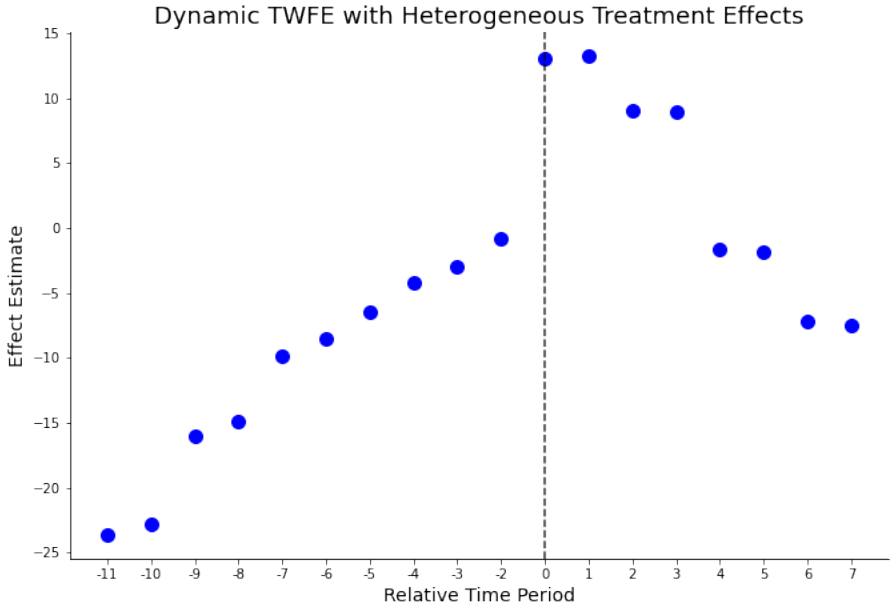
Solution
Consider a simulated event study with the following features. There is a balanced panel with $N = 600$ observed units over $T = 15$ time periods. There are four cohorts. The first three cohorts are treated at times $t = 8$, $t = 10$, and $t = 12$, respectively. One remains untreated. The sizes of these cohorts are $N_1 = 350$, $N_2 = 100$, $N_3 = 12$, and $N_4 = 100$. We keep a linear time trend with random normal noise, plus random uniform unit fixed effects. Under this setup, we consider two possible treatment effects. Under homogeneous treatment effects, we have a treatment effect equal to 10 for every unit at all times. Under heterogeneous treatment effects, earlier-treated cohorts have smaller treatment effects. Specifically, the treatment effect is equal to 3, 14.5, 50, and 10, respectively for the four cohorts. We run a dynamic TWFE regression with unit dummies, time dummies, and relative time indicators. The outcomes and estimated treatment effects in both the homogeneous and heterogeneous scenarios are shown in the charts below.

The charts show that the effect size of 10 is accurately estimated in the homogeneous treatment effect scenario. In the heterogeneous treatment effect scenario, some of the later relative time effects are estimated

to be negative, which is definitive evidence that negative weights are being used in its calculation. Also note that in the heterogeneous effect scenario, the leading estimates are not zero, even though common trends does hold here. This is a sign that if such hetereogeneity exists, the dynamic TWFE estimator should not be used to test for common trends in the pre-periods.







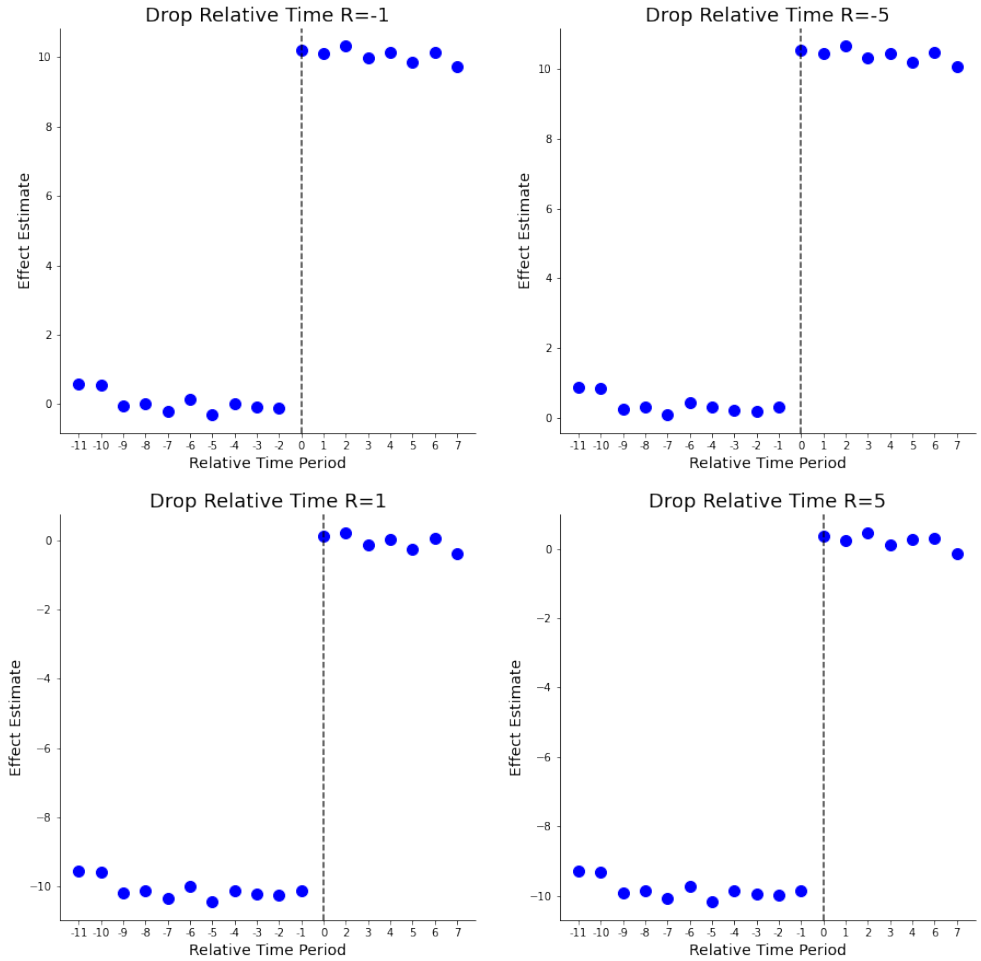
7(d)
In the dynamic two-way fixed effects estimator, one needs to leave out two relative time indicators. Unlike in the usual dummy variable trap, the choice of which two indicators can make a substantive difference.

Solution

We keep the exact same setup as in 7(c), but focus on the homogeneous treatment effect scenario. In the previous problem, I had run the dynamic TWFE estimator leaving out two relative time periods. These two left-out times were $R = -\infty$ (for the never-treated group) and $R = -1$ (as is tradition).

First, I attempt to run the dynamic two-way fixed effects regression without the $R = -1$ group left out. I find that the inversion of the least squares normal matrix does not work, implying that there are collinearity issues in the data. This makes sense, since the reason that we must leave out two relative time periods is because relative time is a linear function of cohort and real time.

Next, I compare four different pairs of left-out groups. I always leave out the $R = -\infty$ group, and I vary between leaving out the $R = -1, -5, 1$, and 5 groups. The dynamic TWFE estimates are shown below. I find that leaving out a group before or after treatment (i.e., $R < 0$ or $R > 0$) changes the level of the pre-treatment and post-treatment estimates. In this example, when we leave out $R = 1$ or $R = 5$, the pre-treatment estimates are roughly equal to -10, and the post-treatment estimates equal to 0. But when we leave out $R = -1$ or $R = -5$, we find the pre-treatment estimates roughly equal to 0 and the post-treatment estimates equal to 10. So the choice of indicator can make a substantive difference.



7(e)
Either a direct (Callaway and Sant'Anna) or imputation approach can be used to consistently estimate the cohort-weighted average treatment effect at any given relative time. Compare the finite-sample bias and standard deviations of the two estimators.

Solution
We keep the exact same setup as in 7(c), but this time we focus on the heterogeneous treatment effect scenario. I implement both the direct approach and imputation approach to estimate cohort-weighted average treatment effect at each relative time. A table of the effect estimates, the true cohort-weighted treatment effects, and the estimate standard errors is shown below.

Comparison of Direct and Imputation Methods for Event Studies

Rel. Time	N	Direct	Imputation	True Effect	Direct Std.Err	Imputation Std.Err
-11	50	2.438	0.893	0.00	0.3948	0.3275
-10	50	2.010	0.536	0.00	0.4727	0.3893
-9	150	1.107	0.201	0.00	0.2861	0.2313
-8	150	0.775	-0.046	0.00	0.2672	0.2079
-7	500	-0.085	-0.130	0.00	0.2419	0.1038
-6	500	0.467	0.094	0.00	0.2070	0.0925
-5	500	-0.087	-0.171	0.00	0.2037	0.0907
-4	500	0.376	0.005	0.00	0.1986	0.0966
-3	500	0.375	0.028	0.00	0.2155	0.0971
-2	500	-0.217	-0.088	0.00	0.2040	0.0835
-1	500	0.187	0.072	0.00	0.1870	0.0816
0	500	10.332	10.332	10.00	0.4935	0.6139
1	500	10.237	10.237	10.00	0.4724	0.6167
2	500	10.685	10.686	10.00	0.5133	0.6260
3	500	10.005	10.007	10.00	0.4960	0.6329
4	450	5.853	5.899	5.56	0.3716	0.3929
5	450	5.491	5.538	5.56	0.3518	0.4154
6	350	3.701	3.759	3.00	0.3664	0.3414
7	350	2.811	2.869	3.00	0.3473	0.3651

Both approaches come very close to the true cohort-weighted effects, suggesting that I correctly implemented them and that both are consistent estimators. In this case, the direct approach has slightly larger finite sample bias, with a mean absolute deviation from the true effect of 0.56, compared to 0.25 for the imputation approach. This appears to be primarily because the direct approach did worse for estimating the pre-period effects $R = -11$ and $R = -11$, which had a smaller sample size. The standard errors were smaller for the imputation approach in the relative pre-period, but larger in the relative post-period. Standard errors were calculated using a blocked bootstrap with 200 samples.

Problem 8

This paper is about "The War on Poverty's Experiment in Public Medicine: Community Health Centers and the Mortality of Older Americans," by Martha Bailey and Andrew Goodman-Bacon, which was published in The American Economic Review in 2015. The paper is posted on Canvas, along with the authors' replication package.

8(a)
Reproduce Table 2, column (1), panel A.

Solution
The coefficients and standard errors were successfully reproduced, shown in the table below:

Replication of Table 2, Column 1, Panel A in Bailey and Goodman-Bacon (2015)

Variable	Coefficient	Standard Error
Years -6 to -2	0.034	2.844
Years 0 to 4	-5.636	3.540
Years 5 to 9	-12.045***	4.569
Years 10 to 14	-9.384*	5.639

See the associated Jupyter notebook for further details of implementation.

8(b)
On pg. 1078, the authors state:
Our empirical strategy uses variation in when and where CHC programs were established to quantify their effects on mortality rates.
Do you agree with this statement? Why or why not?

Solution
I agree that this is what the authors thought they were doing. In reality, we know that these dynamic two-way fixed effects regressions can give negative weights to certain time-cohort treatment effects unless the regression is fully saturated and we believe in homogeneous treatment effects. The authors' regression specification is not fully saturated, and even so, homogeneous treatment effects is doubtful here. And even further, the author's use of urban group / year dummy interactions in their TWFE specification means that they are not just using variation in where and when CHC were established, but also how those establishments related to areas with varying levels of urbanization.

8(c)
Take the specification in Table 2, column (1), panel A, but simplify it so that it only includes year fixed effects that are not interacted with urban dummies. Use this simpler specification to repeat the estimates for Table 2, column (1), panel A. Provide a formal theoretical explanation of how interacting year fixed effects with urban dummies affects the required common trends assumption.

Solution
The estimates for Table 2, Column 1, Panel A under the simpler specification without year-urban interactions are shown in the table below. The estimated effects are somewhat larger in magnitude years 0 to 4 and much smaller in years 5 through 14.

See the associated Jupyter notebook for further details of implementation.

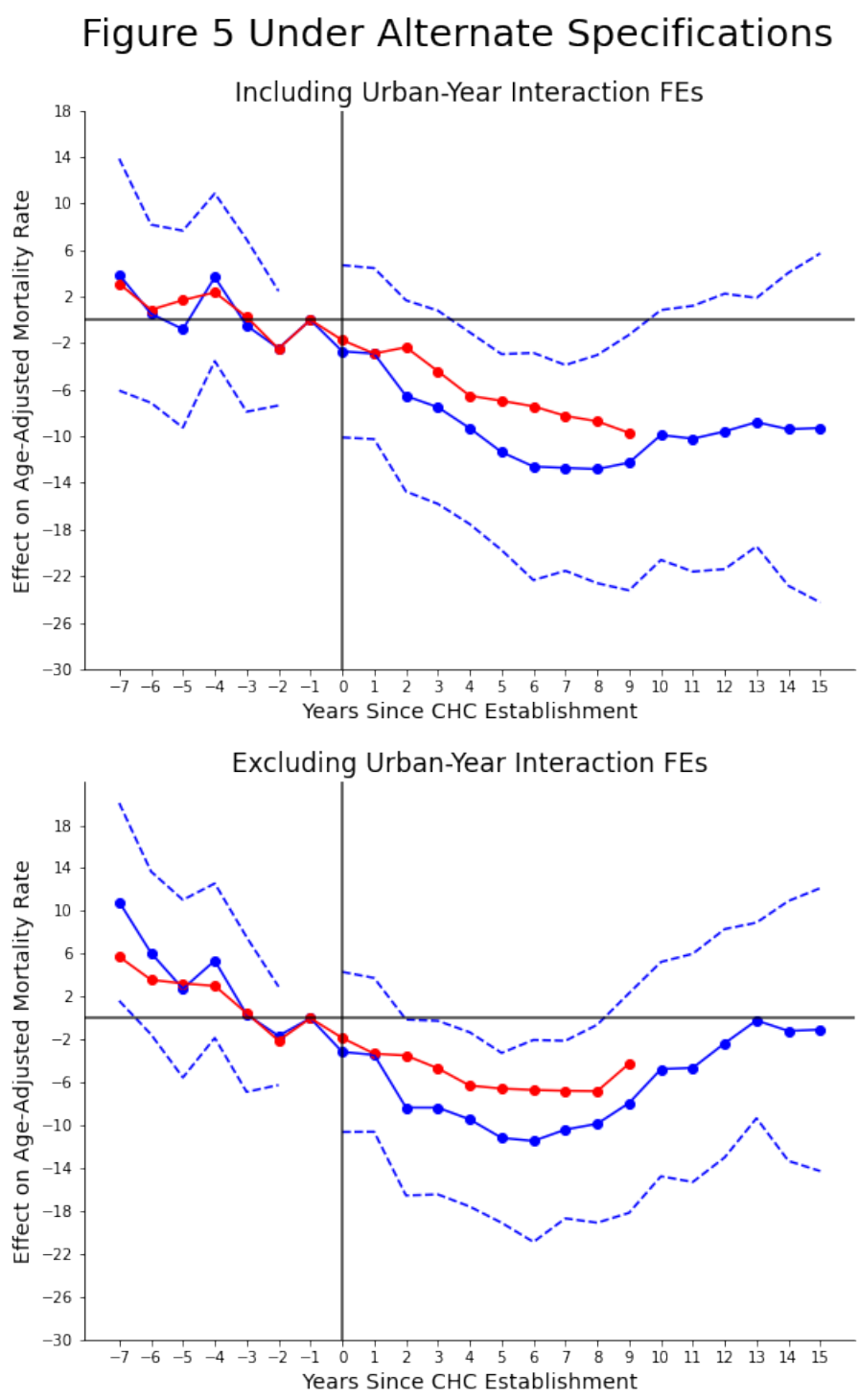
Simple Specification of Table 2, Column 1, Panel A in Bailey and Goodman-Bacon (2015)

Variable	Coefficient	Standard Error
Years -6 to -2	2.49	2.829
Years 0 to 4	-6.42*	3.460
Years 5 to 9	-10.022**	4.321
Years 10 to 14	-2.651	5.087

The fixed effects specification without urban-year interaction dummies demeans each county's time average of mortality rate and each year's cross-section average. This setup (also assuming no covariates) leaves differential trends in counties at any given time to be explained by CHC treatment. So the classic common trends assumption applies. For the specification with urban-year interaction dummies, on the other hand, further demeans yearly effects specific to areas with different urbanization levels. Now it requires that differential trends in counties *at the same level of urbanization* are explained by CHC treatment. That is, common trends is restricted to just be assumed among counties that are similarly-urbanized.

8(d)
Create two versions of Figure 5: one that uses the specification in Table 2, column (1), panel A, and another that uses the simpler specification in the previous part.

Solution
The chart below replicates two versions of Figure 5. The first figure uses the specification in Table 2, column (1), panel A, with the urban-year interaction fixed effects. The second figure uses the simpler specification removing these interaction fixed effects. The blue and red lines match the line colors used in Figure 5, corresponding to estimation sample. That is, the blue line is estimated using counties with early CHCs from 1959-1988, and the dashed blue lines are its 95% confidence intervals. The red line is the sample of counties with all CHCs. I ran into issues replicating these specifications for the sample of early CHCs observed through 1998 (the green line in Figure 5), so this is excluded.



8(e)
Use the direct (Callaway and Sant'Anna) approach to estimate cohort-averaged relative time effects for

the simpler no-urban specification, and produce counterparts to both Figure 5 and Table 2, column (1), panel A. Compute standard errors using a nonparametric block bootstrap.

Solution
The new versions of Figure 5 and Table 2 using the direct (Callaway and Sant'Anna) approach are shown below, keeping urban-interaction dummies removed from the specification. I bootstrap standard errors using a block bootstrap with 200 samples.

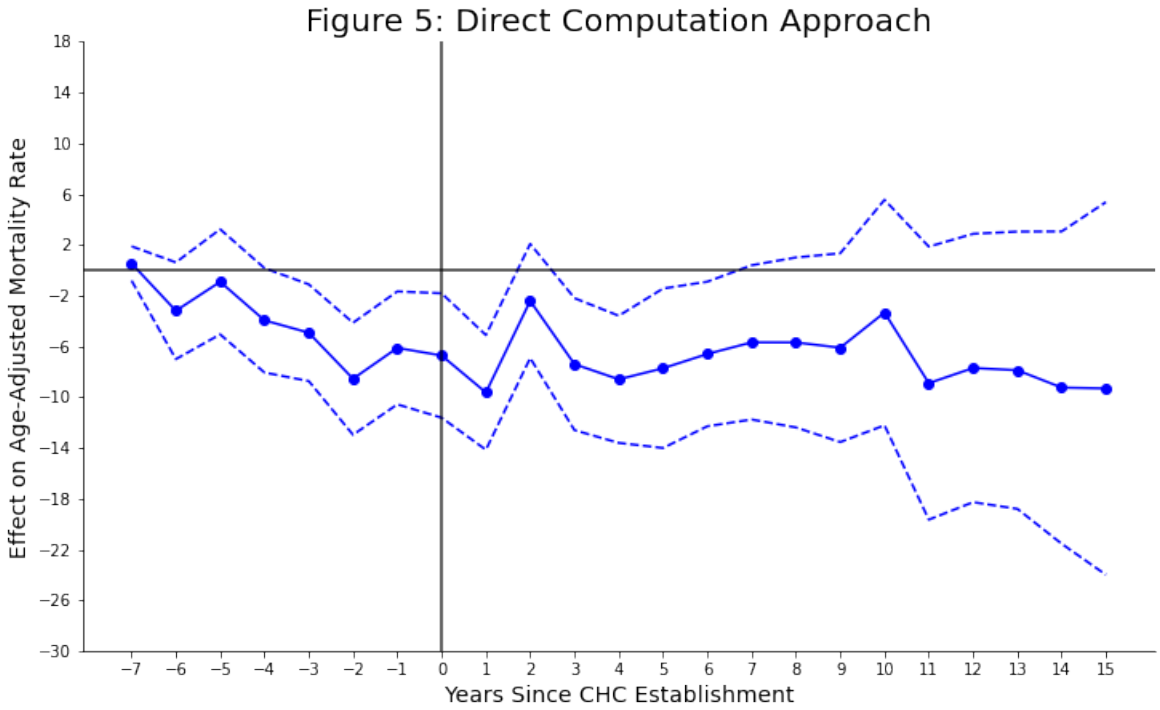


Table 2, Column 1, Panel A: Direct Approach

Time	Coefficient	Standard Error
Years -6 to -2	0.539	1.344
Years 0 to 4	-4.298***	2.270
Years 5 to 9	-6.112***	2.145
Years 10 to 14	-6.945***	2.971

8(f)
Use the imputation approach to estimate cohort-averaged relative time effects for the simpler no-urban specification, and produce counterparts to both Figure 5 and Table 2, column (1), panel A. Compute standard errors using a nonparametric block bootstrap.

Solution
The new versions of Figure 5 and Table 2 using the imputation approach are shown below, keeping urban-interaction dummies removed from the specification. I bootstrap standard errors using a block bootstrap with 200 samples. The effects are identical to the direct approach, which is reassuring since these two methods are, in theory, mechanically equal to each other. I prefer this imputation approach because my code ran much quicker with this method, compared to the direct method.

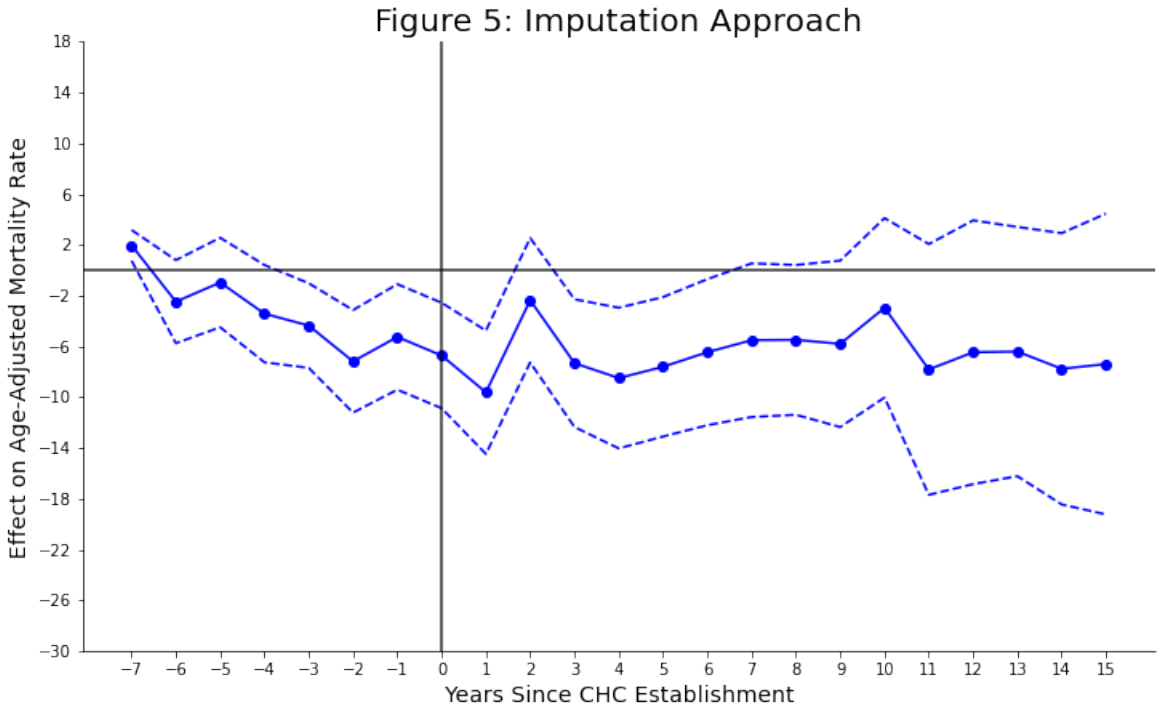


Table 2, Column 1, Panel A: Direct Approach

Time	Coefficient	Standard Error
Years -6 to -2	0.539	1.344
Years 0 to 4	-4.298***	2.270
Years 5 to 9	-6.112***	2.145
Years 10 to 14	-6.945***	2.971

8(g)
Now interact year fixed effects with urban dummies, as in the authors' original Table 2, column (1) specification, and attempt to use both the direct and imputation approaches. Discuss any difficulties or problems in doing so.

Solution
I implemented the imputation approach, including the year/urban status fixed effects in the regression. Interestingly, this did not appear to change the point estimates of effects in Figures 5 and Table 2 too much, except in later relative years.[3] Standard errors also changed somewhat in Figure 5. Standard errors seem to decrease for early post-treatment period, but slightly inflate at the end of the post-treatment period. See the Figure 5 and Table 2 alternative below.

[3]Perhaps this was a coding error, in which case I would add that this was one difficulty I found during implementation.

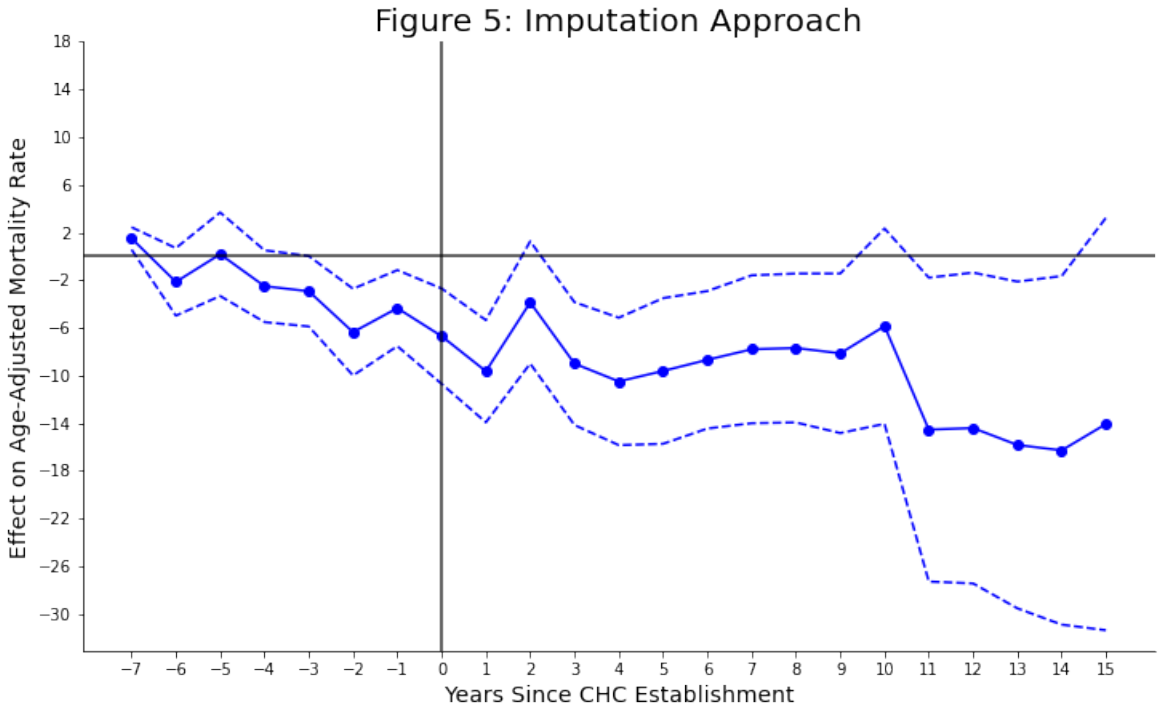


Table 2, Column 1, Panel A: Direct Approach

Time	Coefficient	Standard Error
Years -6 to -2	0.539	1.344
Years 0 to 4	-4.298***	2.270
Years 5 to 9	-6.112***	2.145
Years 10 to 14	-6.945***	2.971

I was not able to succesfully implement the direct approach with the year fixed effects and urban dummy interactions. Hypothetically, I think this is possible by making comparisons within urban status and years, but I could not get the right set of regression indicator functions for it to work, since this was very tricky to formalize into code. The result should be mechanically equal to the imputation approach.