ARE 213 Problem Set 2B

Becky Cardinali, Yuen Ho, Sara Johns, and Jacob Lefler

Due 11/16/2020

Question 1

We first estimate an event study specification.

(a) First determine the minimum and maximum event time values that you can estimate in this data set. Code up a separate event time indicator for each possible value of event time in the data set. Estimate an event study regression using all the event time indicators. What happens?

Question 2

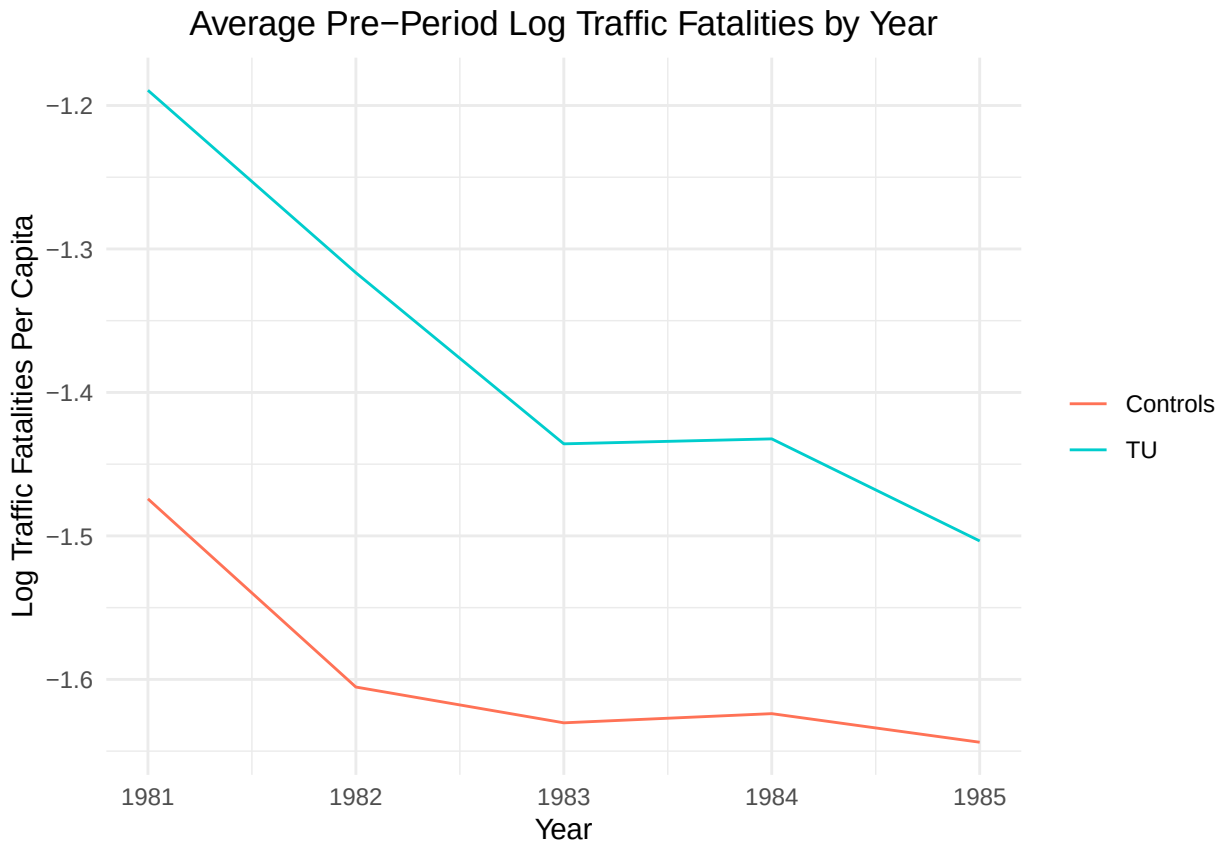
We now apply the synthetic control methods from Abadie et al (2010).

(a) Please use the aggregate "treatment" state (a population weighted average of the first 4 states to have a primary seatbelt law: CT, IA, NM, TX) as the treatment unit (TU) in the synthetic control analysis.

i. Compare the average pre-period log traffic fatalities per capita of the TU site to that of the average of all the "control" states. Next, graph the pre-period log traffic fatalities by year for the pre-period for both the TU and the average of the control group. Interpret.

```
#Create a treatment status variable that = 1 if state is CT, IA, NM, or TX and =0 otherwise.
traffic[,treat := ifelse(state_name == "CT" | state_name == "IA"|state_name == "NM" | state_name == "TX" |sta

controls <- traffic[primary == 0 & year == 2003,state] #select states that never pass a primary seat belt lau

traffic_SYM <- traffic[traffic$state %in% controls | state == 99,] #Create new data table with only the poten

#Set state_name for state 99 to TU
traffic_SYM[state == 99, state_name := "TU"]

#Change treatment variable to a factor variable
traffic_SYM$treat <- as.factor(traffic_SYM$treat)

#Create log fatalities per capita variable
traffic_SYM[, ln_fat_pc := log(fatalities/population)]

# Create log covariates
traffic_SYM[,ln_unemploy := log(unemploy)]
traffic_SYM[,ln_totalvmt := log(totalvmt)]
traffic_SYM[,ln_precip := log(precip)]
traffic_SYM[,ln_snow := log(snow32+0.01)] # to avoid NA from zeroes
```

```r
#Compare the average pre-period log traffic fatalities per capita between treatment and control
premeanT <- mean(traffic_SYM[treat == 1 & year<1986, ln_fat_pc]) #mean pre-period log traffic fatalities in t
premeanC <- mean(traffic_SYM[treat == 0 & year<1986, ln_fat_pc]) #mean pre-period log traffic fatalities in c

#Create variable of mean log traffic fatalities by treatment status by year
traffic_SYM[, mean_lnfat_treat := lapply(.SD, mean), .SDcols = c("ln_fat_pc"), by = c("treat","year")]

#Graph the mean pre-period log traffic fatalities by year for Treatment vs Control
traffic_SYM[year < 1986,] %>%
  ggplot(aes(x=year, y = mean_lnfat_treat, group = treat, color = treat)) +
  geom_line() +
  theme_minimal() +
  labs(title = "Average Pre-Period Log Traffic Fatalities by Year", x = "Year", y = "Log Traffic Fatalities P
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_color_manual(labels = c("Controls", "TU"), values = c("coral1", "cyan3"))
```



The average pre-period log traffic fatalities per capita in our aggregate treatment unit is -1.38 compared with -1.6 in our control states. Graphically, we can see that while log traffic fatalities per capita are declining over time in both groups, treatment units have on average higher traffic fatalities per capita than control units in all pre-period years. This makes sense since states with higher traffic fatalities are more likely to take steps to reduce such fatalities, leading to a higher likelihood of implementing seat belt laws.

ii. Compare the dependent variable between the TU and each control state for the year before the treatment. Which control state best matches the TU? Now compare this state's covariates with the TU covariates. Do they appear similar? What might this imply in terms of using this state as the counterfactual state?

```r
#generate a variable that is the absolute value of the difference between the dependent variable of the TU si
TU_dep_1985 <- traffic_SYM[state == 99 & year == 1985, ln_fat_pc] #1985 log traffic fatalities per capita for
traffic_SYM[year == 1985 & state != 99,compare_dep := abs(ln_fat_pc - TU_dep_1985)]
traffic_SYM[is.na(compare_dep), compare_dep := 999]
control <- traffic_SYM[compare_dep == min(traffic_SYM$compare_dep),state] #WV best matches the TU

#Compare average of WV's covariates to TU's covariates in pre-period
avg_comparisons <- data.table(covariates = colnames(traffic_SYM[,c(3:4, 7:13, 17)]), WV_avg = colMeans(traffi
avg_comparisons[,c("WV_avg", "TU_avg")] <- round(avg_comparisons[,c("WV_avg", "TU_avg")], 2) #round numeric v

avg_comparisons
```

```
##       covariates   WV_avg   TU_avg
##  1:     college     0.12     0.22
##  2:        beer     1.11     1.62
##  3:  population  1936.66 11467.00
##  4:    unemploy    14.12     6.75
##  5:  fatalities   435.20  2903.17
##  6:    totalvmt 11680.60 95298.49
##  7:      precip     3.61     2.44
##  8:      snow32     0.41     0.18
##  9: rural_speed    55.00    55.00
## 10:   ln_fat_pc    -1.49    -1.38
```

When we compare log traffic fatalities per capita in 1985 (the year before treatment), we find that West Virginia is the state closest to the TU. We then compare the average pre-period values for various covariates between West Virginia and TU in the table above and find that the two are not very comparable. For example, West Virginia has on average a lower percentage of college grads, per capita beer consumption, population, and total vehicle miles traveled (VMT) than the TU and on average higher unemployment, precipitation, and snowfall. This suggests that West Virginia is not a good counterfactual state since it differs from the TU systematically in the pre-period.

(b) Apply the synthetic control method using the available covariates and pre-treatment outcomes to construct a synthetic control group.

i. Discuss the synthetic control method including its benefits and potential drawbacks.

Compared to a diff-in-diff estimator, the synthetic control estimator has the benefit of providing a more rigorous, less ad-hoc way of selecting control units from a large pool of potential controls. Unlike the diff-in-diff estimator, the synthetic control estimator also does not reply on the assumption of parallel preimplementation trends. A key advantage is that a synthetic control estimator controls for both observed and unobserved unit-by-time shocks, whereas a diff-in-diff estimator only controls for observed unit-by-time shocks. To see this intuitively, note that only units that are similar in both observed and unobserved determinants of the outcome variable and in the effect of those determinants on the outcome variable will produce similar trajectories of the outcome variable over extended periods of time (Abadie et al. 2010). With a synthetic controls estimator, we can also leverage the large pool of potential controls to conduct permutation-based inference.

One drawback, however, is that the credibility of the synthetic controls method relies on achieving a good preimplementation fit for the outcome of interest between treated unit and synthetic control, which is difficult if the treated unit is an outlier. This also necessitates having enough data on the outcome and covariate variables for the treated unit and a suitable pool of comparison units over a significant period of time. Furthermore, judging whether there is a good fit is more of an art than a science, as there is currently no consensus on what constitutes a "good fit." A related drawback is that the synthetic control units need to match the treated unit on both levels and trends. Thus, if there are control units that are only a good match for the trends but not the levels, or vice versa, then we may be discarding control units that

satisfy the parallel trends assumption because they do not match the baseline levels. We also need to assume that there are no shocks that affect the treated unit differentially than the potential control units and that there are no spillovers of the treatment effect from the treated unit into the control units, although these are also necessary assumptions for a diff-in-diff estimator.

ii. Use the software package provided by Abadie et al. to apply the synthetic control method. Please be sure to state precisely what the command is doing and how you determined your preferred specification.

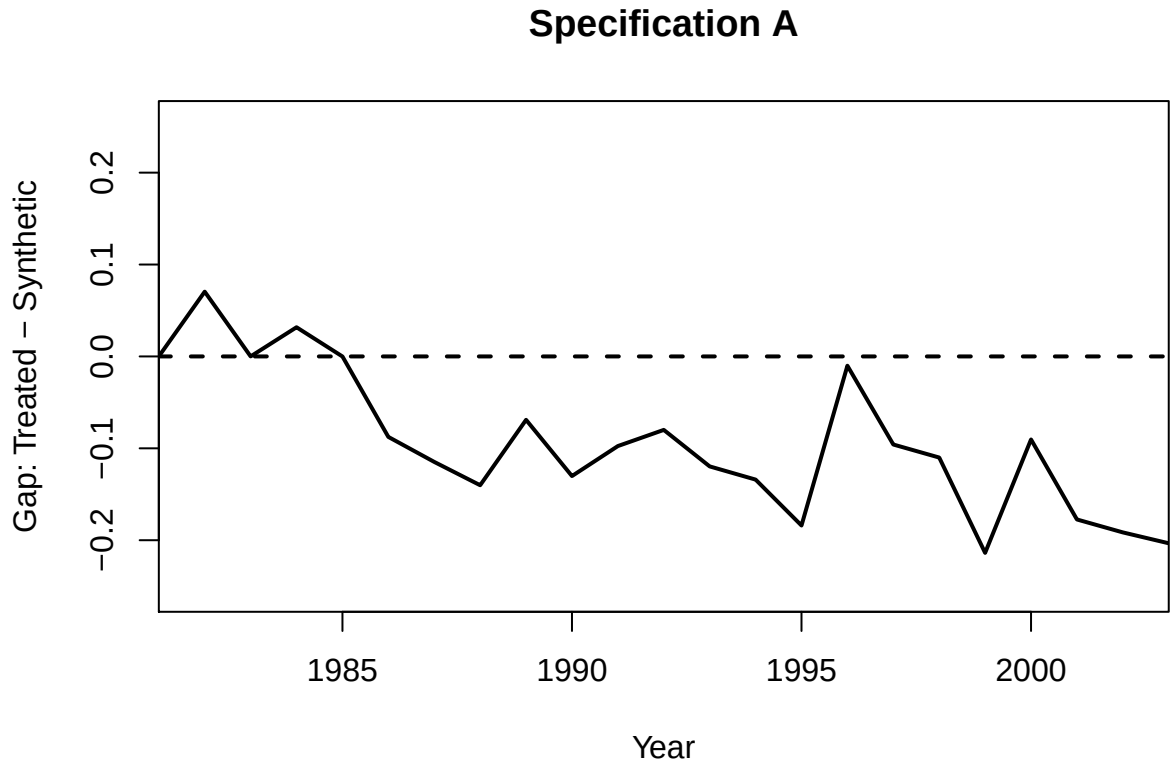
We use the R package "Synth" to implement the synthetic control method. Synth constructs a synthetic control group by searching for a weighted combination of control units chosen to approximate the treated unit (TU) in terms of characteristics that are predictive of the outcome. Synth requires us to supply four matrices as its main arguments, XO, X1, Z1, and Z0. X1 and X0 contain the predictor values for the TU and the control units respectively. Similarly, Z1 and Z0 contain the outcome variable for the pre-intervention period for the TU and the control units respectively. The command "dataprep" allows us to create the matrices X1, X0, Z1, and Z0. The command "synth" is then used to construct the synthetic control group, where weights are assigned to the control units to minimize the mean squared prediction error (MSPE) over the pre-intervention time period (in our case 1981-1985). We then use the command "gaps.plot" to plot the gaps in the trajectories of the outcome variable for the TU and the constructed synthetic control group.

```
## Specification A
dataprep.outA <- dataprep(foo = traffic_SYM,
                          predictors = c("college","unemploy","totalvmt","precip"),
                          predictors.op = "mean",
                          dependent = "ln_fat_pc",
                          unit.variable = "state",
                          time.variable = "year",
                          special.predictors = list(
                             list("ln_fat_pc", 1981, "mean"),
                             list("ln_fat_pc", 1983, "mean"),
                             list("ln_fat_pc", 1985, "mean")),
                          treatment.identifier = 99,
                          controls.identifier = controls,
                          time.predictors.prior = c(1981:1985),
                          time.optimize.ssr = c(1981:1985),
                          unit.names.variable = "state_name",
                          time.plot = 1981:2003
                          )
# synth command identifies weights to construct the synthetic control
synth.outA <- synth(dataprep.outA)
```

```
##
## X1, X0, Z1, Z0 all come directly from dataprep object.
##
##
## ****************
##  searching for synthetic control unit
##
##
## ****************
## ****************
## ****************
##
## MSPE (LOSS V): 0.001194251
##
## solution.v:
##  3.89403e-05 0.003843508 3.62368e-05 7.72858e-05 0.3894023 0.2917817 0.3148201
##
```

```
## solution.w:
##  0.009182045 0.01750278 0.04763439 0.3056906 0.01256928 0.007678048 0.02068668 0.008355919 0.008982594 0.0
```

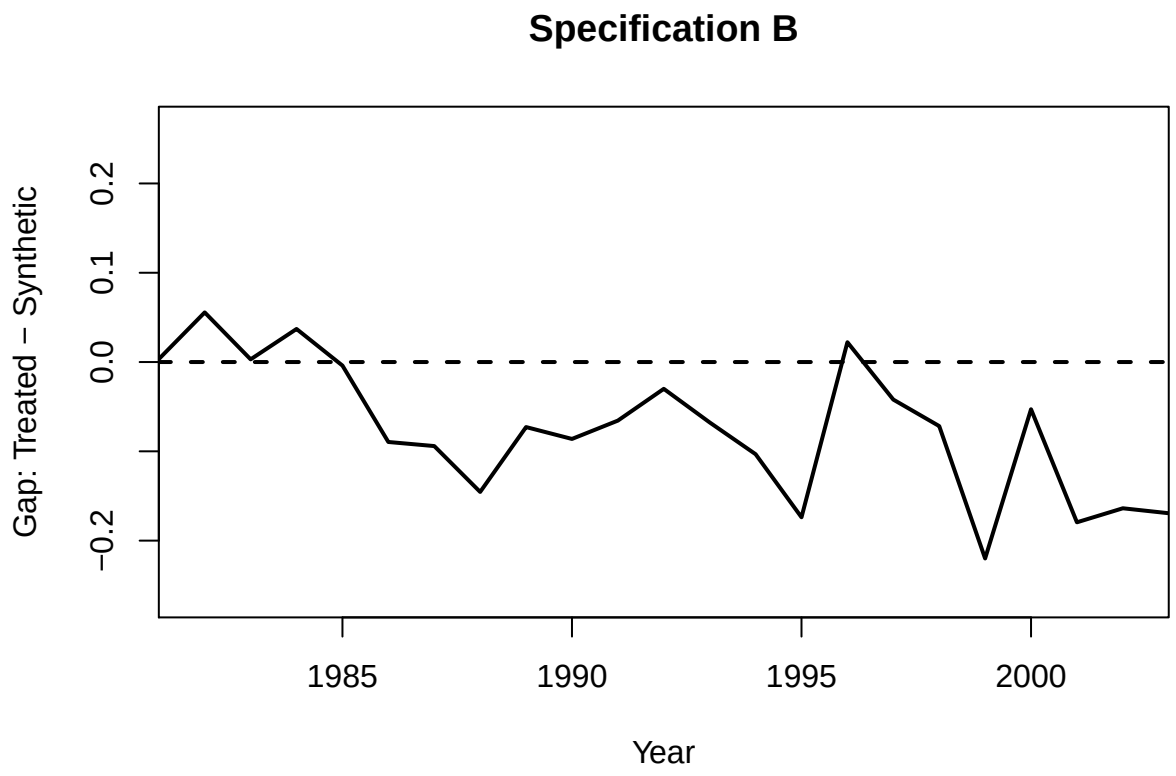
```
# plot the gaps (treated - synthetic)
gaps.plot(dataprep.res = dataprep.outA,
          synth.res = synth.outA,
          Ylab = "Gap: Treated - Synthetic",
          Xlab = "Year",
          Main = "Specification A")
```




```
## Specification B - Include more covariates and log of covariates
dataprep.outB <- dataprep(foo = traffic_SYM,
                          predictors = c("college","ln_unemploy","ln_totalvmt","ln_precip", "beer", "populatio
                          predictors.op = "mean",
                          dependent = "ln_fat_pc",
                          unit.variable = "state",
                          time.variable = "year",
                          special.predictors = list(
                             list("ln_fat_pc", 1981, "mean"),
                             list("ln_fat_pc", 1983, "mean"),
                             list("ln_fat_pc", 1985, "mean")),
                          treatment.identifier = 99,
                          controls.identifier = controls,
                          time.predictors.prior = c(1981:1985),
                          time.optimize.ssr = c(1981:1985),
                          unit.names.variable = "state_name",
                          time.plot = 1981:2003
                          )

# synth command identifies weights to construct the synthetic control
synth.outB <- synth(dataprep.outB)
```

```
##
## X1, X0, Z1, Z0 all come directly from dataprep object.
##
##
## ****************
##  searching for synthetic control unit
##
##
## ****************
## ****************
## ****************
##
## MSPE (LOSS V): 0.0008992903
##
## solution.v:
##  0.007032546 0.02793624 9.68e-08 0.02148516 0.01572332 0.0005958511 0.0305687 0.3547645 0.2498015 0.292092
##
## solution.w:
##  2.59908e-05 1.3498e-05 9.74753e-05 0.2486637 4.78782e-05 0.09215148 3.2912e-05 7.62018e-05 0.002666741 4.
```

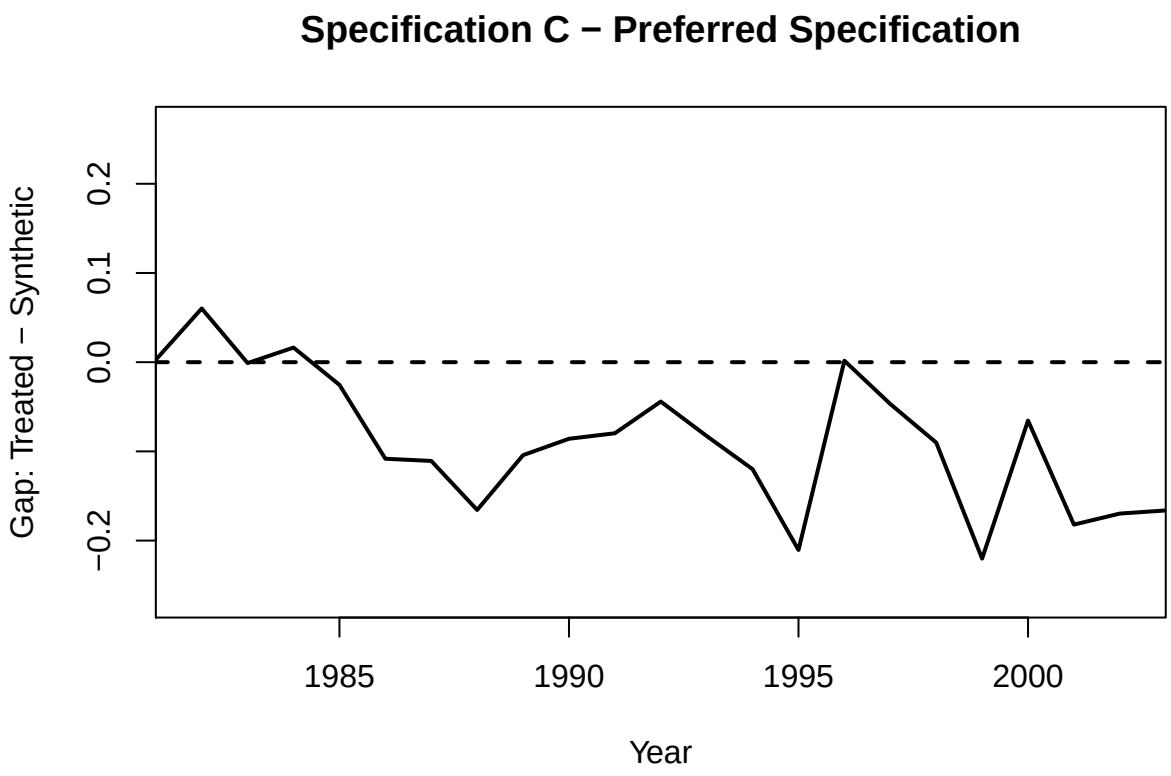
```
# plot the gaps (treated - synthetic)
gaps.plot(dataprep.res = dataprep.outB,
          synth.res = synth.outB,
          Ylab = "Gap: Treated - Synthetic",
          Xlab = "Year",
          Main = "Specification B")
```



```r
## Specification C - include more covariates and more "special predictors" (Preferred Specification)
dataprep.outC <- dataprep(foo = traffic_SYM,
                          predictors = c("college","ln_unemploy","ln_totalvmt","ln_precip", "beer", "populatio
                          predictors.op = "mean",
                          dependent = "ln_fat_pc",
                          unit.variable = "state",
                          time.variable = "year",
                          special.predictors = list(
                             list("ln_fat_pc", 1981, "mean"),
                             list("ln_fat_pc", 1983, "mean"),
                             list("ln_fat_pc", 1984, "mean"),
                             list("ln_fat_pc", 1985, "mean")),
                          treatment.identifier = 99,
                          controls.identifier = controls,
                          time.predictors.prior = c(1981:1985),
                          time.optimize.ssr = c(1981:1985),
                          unit.names.variable = "state_name",
                          time.plot = 1981:2003
                          )
# run the synth command to identify weights
synth.outC<- synth(dataprep.outC)
```

```
##
## X1, X0, Z1, Z0 all come directly from dataprep object.
##
##
## ****************
##  searching for synthetic control unit
##
##
## ****************
## ****************
## ****************
##
## MSPE (LOSS V): 0.0009086118
##
## solution.v:
##  0.0352303 0.02747538 4.736e-06 0.02476681 0.002578475 0.002176447 0.1163897 0.3337352 0.1786962 0.1879783
##
## solution.w:
##  3.1391e-06 6.3988e-06 2.253e-06 0.2422559 8.0645e-06 0.1258041 3.31925e-05 1.07805e-05 1.7978e-06 0.00036
```

```r
# plot the gaps (treated - synthetic)
gaps.plot(dataprep.res = dataprep.outC,
          synth.res = synth.outC,
          Ylab = "Gap: Treated - Synthetic",
          Xlab = "Year",
          Main = "Specification C - Preferred Specification")
```

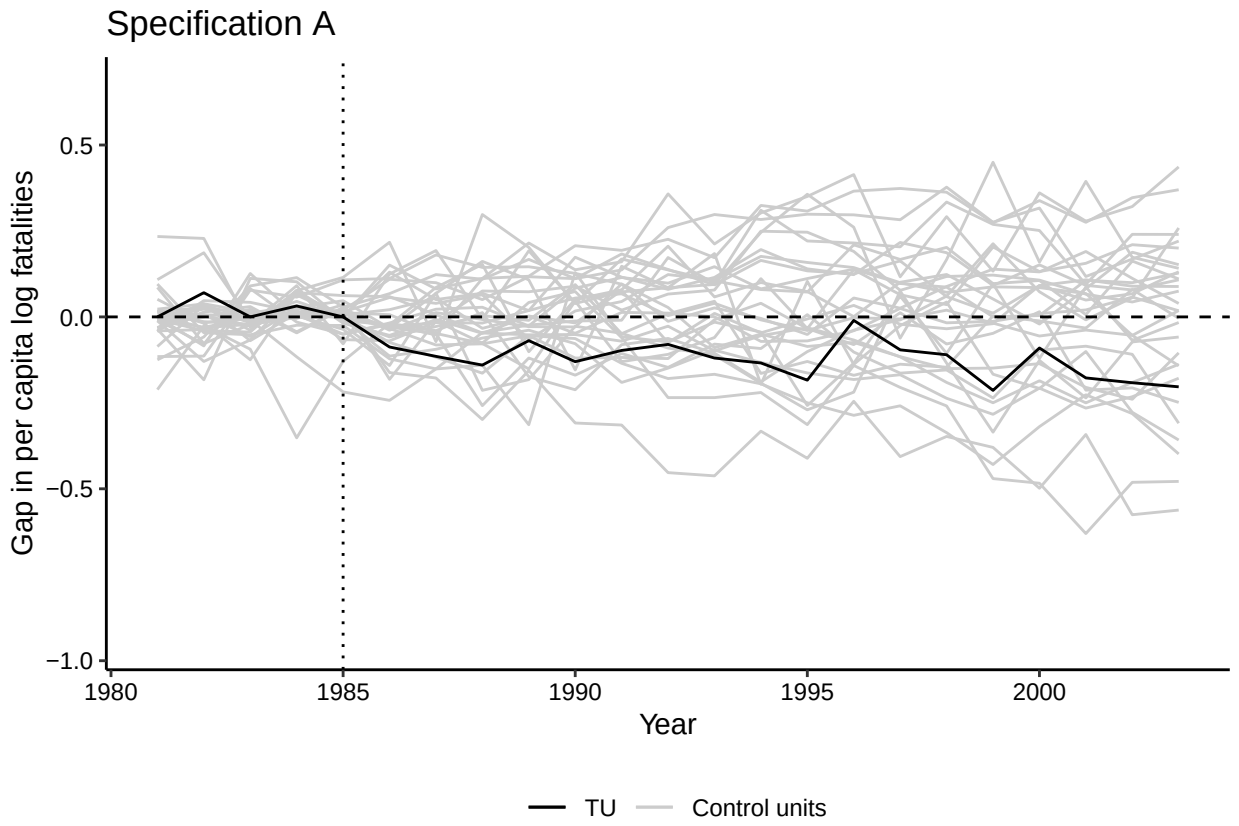


We run the synthetic control method for a variety of specifications, varying which variables we include as control variables and varying the number of "special predictors" (i.e. the mean of the outcome variable in chosen pre-intervention years) we include. Our preferred specification is specification "C", which includes the most covariates and "special predictors," since the gap between the TU and the synthetic control in the pre-intervention period is modestly closer to 0 with this specification, particularly in the years closest to the treatment year.

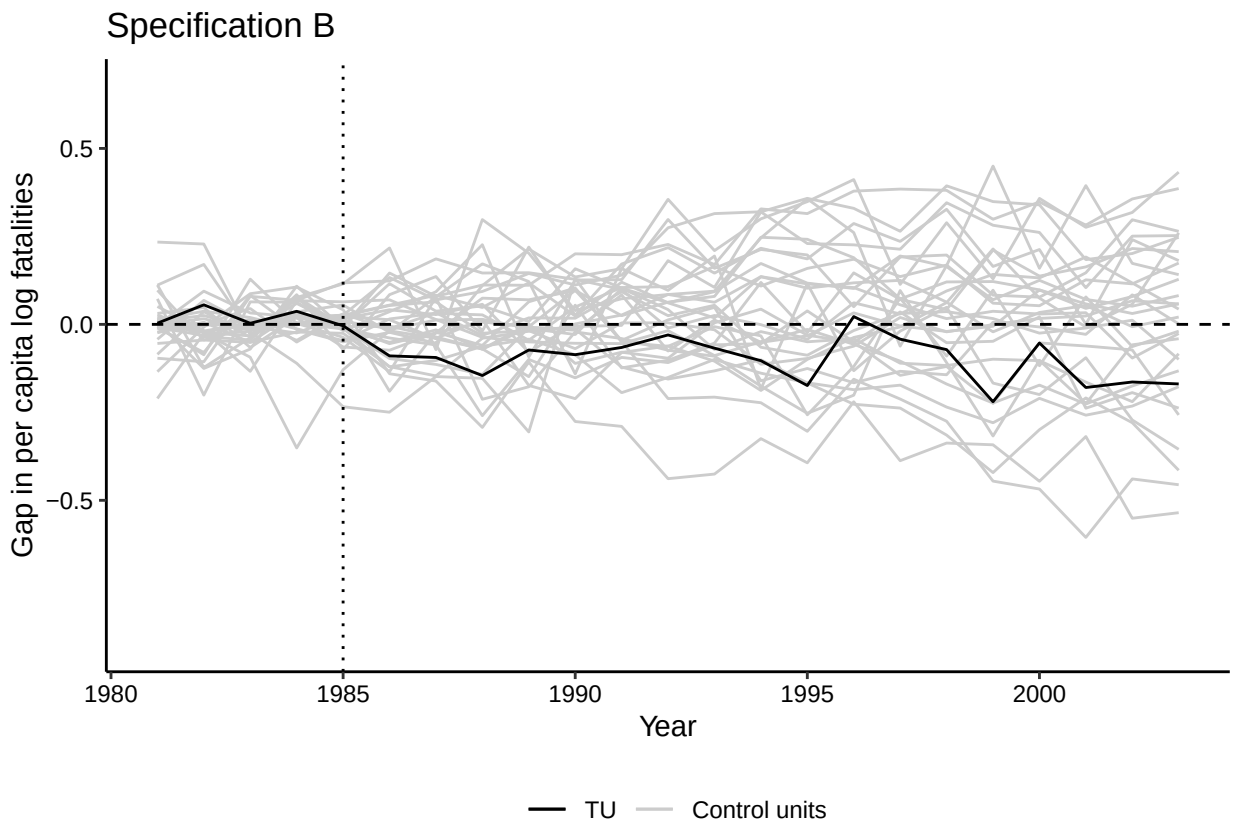
(c) Graphical interpretation and treatment significance

i. Generate graphs plotting the gap between the TU and the synthetic control group under both your preferred specification and a few other specifications you tried.

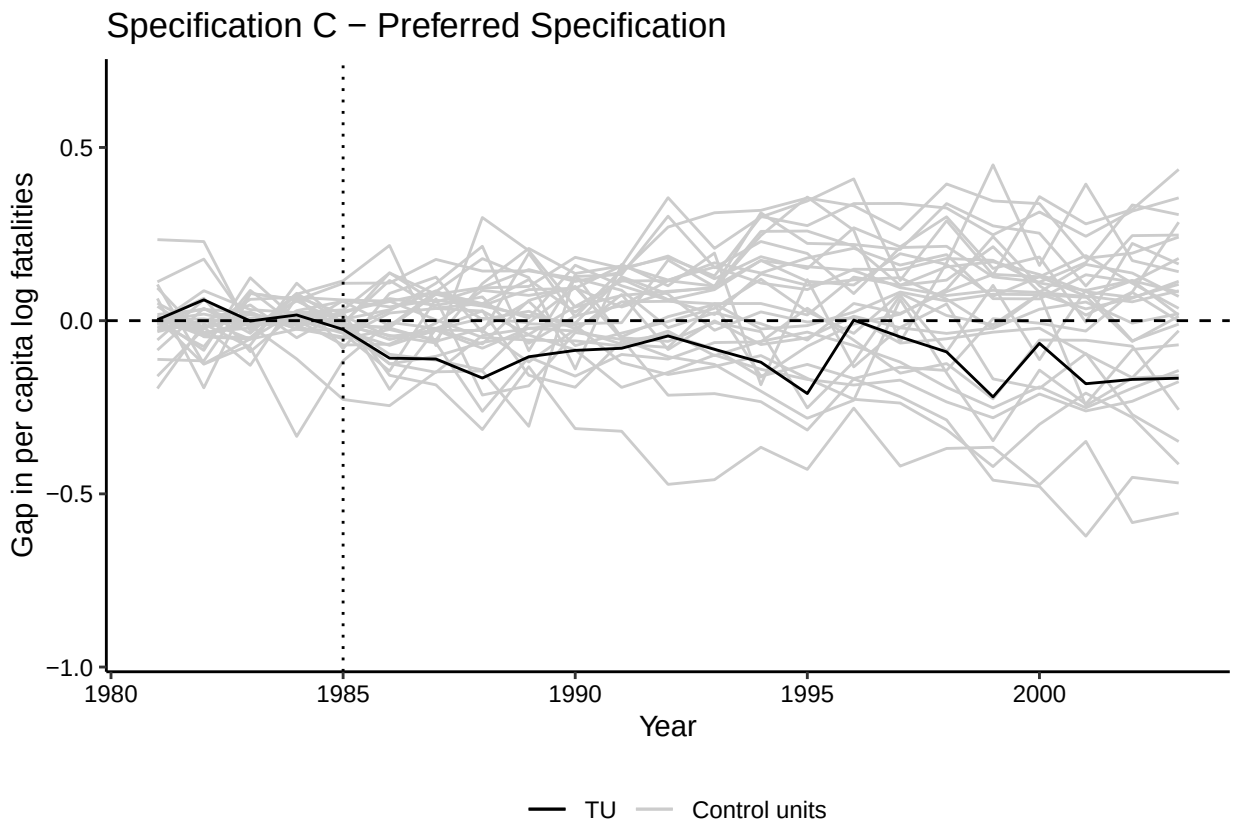
```
## Plot the gaps in outcome values over time of each unit --
# treated and placebos -- to their synthetic controls

#Specification A
pA <- plot_placebos(tdfA,
                    discard.extreme=FALSE,
                    mspe.limit=20,
                    title = "Specification A",
                    xlab='Year',
                    ylab='Gap in per capita log fatalities',
                    alpha.placebos = 1)
pA
```




```
#Specification B
pB <- plot_placebos(tdfB,
                    discard.extreme=FALSE,
                    mspe.limit=20,
                    title = "Specification B",
                    xlab='Year',
                    ylab='Gap in per capita log fatalities',
                    alpha.placebos = 1)
pB
```




```
#Specification C - Preferred Specification
pC <- plot_placebos(tdfC,
                    discard.extreme=FALSE,
                    mspe.limit=20,
                    title = "Specification C - Preferred Specification",
                    xlab='Year',
                    ylab='Gap in per capita log fatalities',
                    alpha.placebos = 1)
pC
```



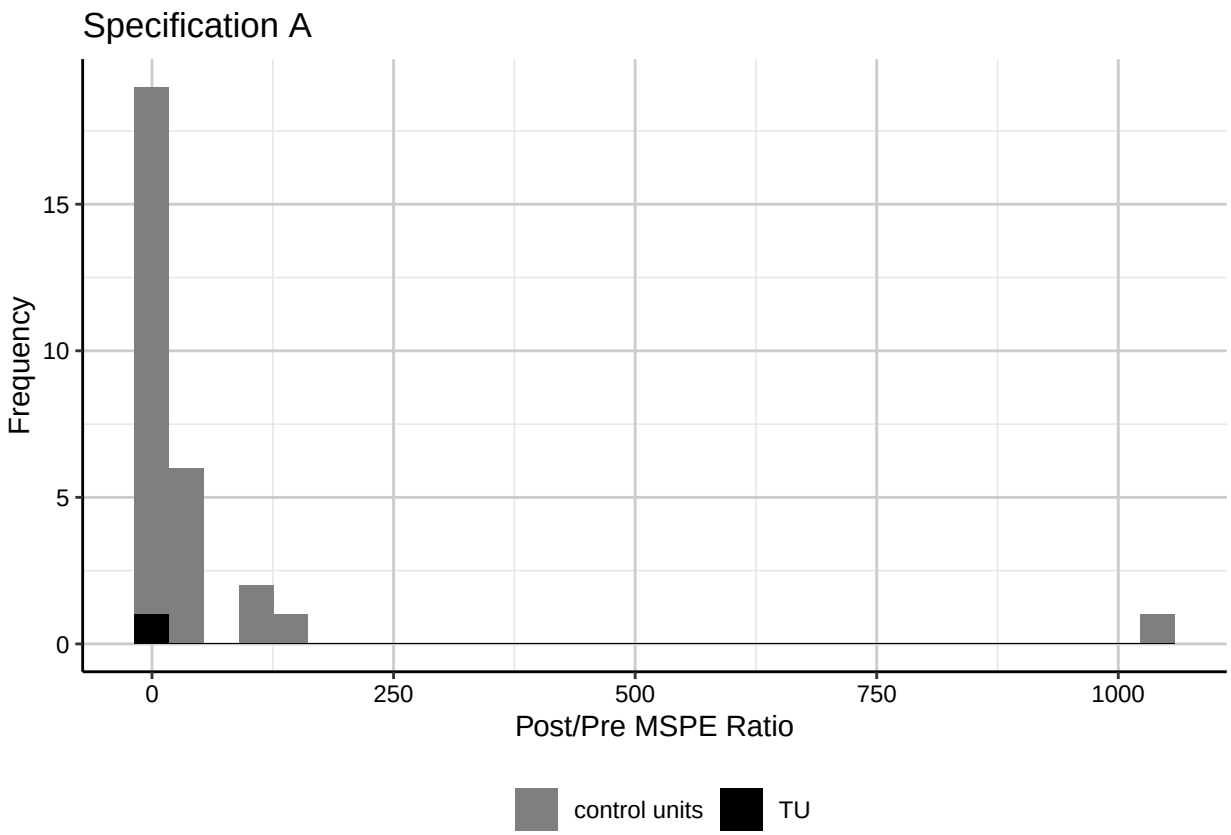
ii. Compare the graph plotting the gap between the TU and the synthetic control group under your preferred specification with the graphs plotting the gap between each control state and its "placebo" treatment. Do you conclude that the treatment was significant? Why or why not?

Comparing the above graphs, we conclude that the treatment was not significant. Since we constructed the synthetic control unit so that it tracked the TU closely in the pre-intervention period, we expect the two units to diverge in the post-intervention period more than in the pre-intervention period by construction, even if there is no treatment effect. When we compare the observed treatment effect to the "placebo" treatment effects for the control states, the treatment effect for the TU is near the middle of the distribution, suggesting that our measured treatment effect could be simply due to chance. If we had randomly picked an untreated state and implemented the same procedure, it is likely we would have found a post-intervention deviation of the observed magnitude or larger.

iii. Create a graph of the post-treatment/pre-treatment prediction ratios of the Mean Squared Prediction Errors (MSPE) for the actual and "placebo" treatment gaps in (ii). Do you conclude that the treatment was significant? Why or why not?

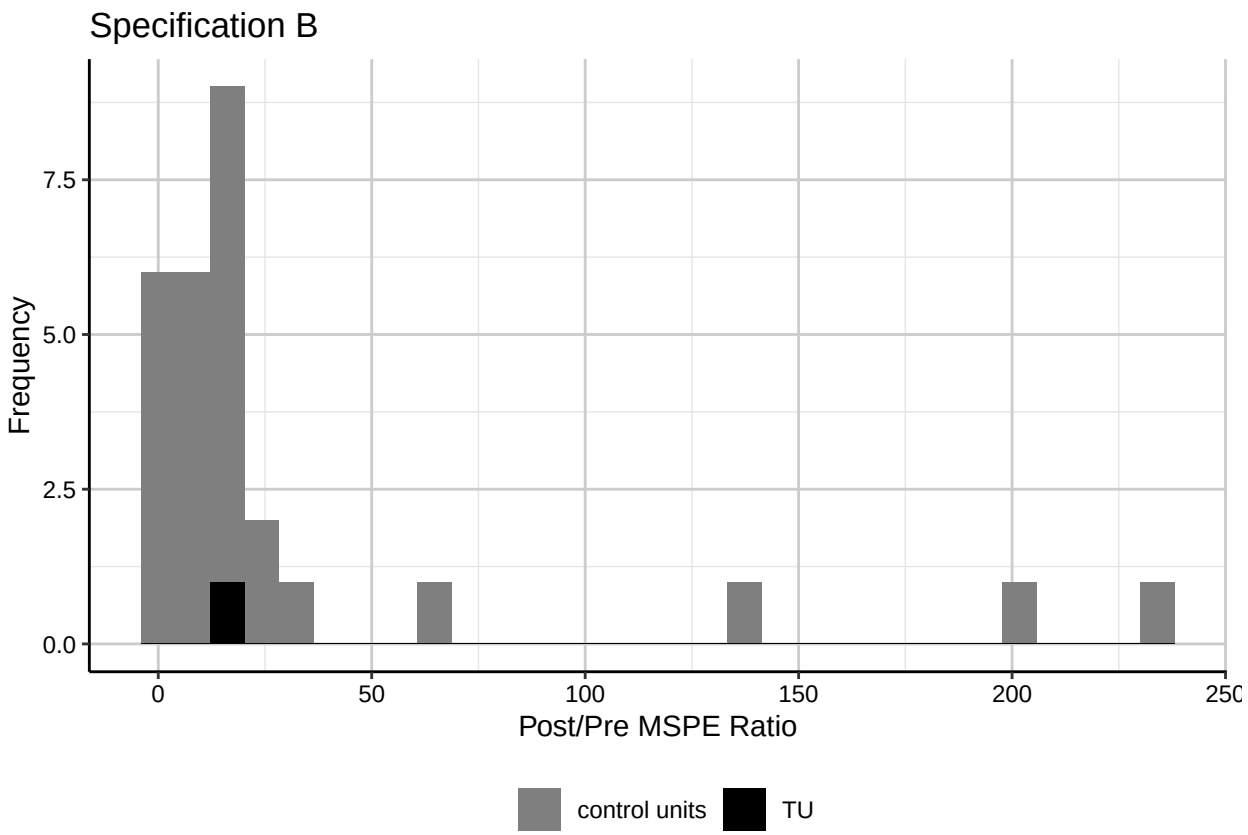
```
## Specification A
mspe.plot(tdfA,
          discard.extreme = TRUE,
          mspe.limit = 20,
          plot.hist = TRUE,
          title = "Specification A",
          xlab = "Post/Pre MSPE Ratio",
          ylab = "Frequency"
          )
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



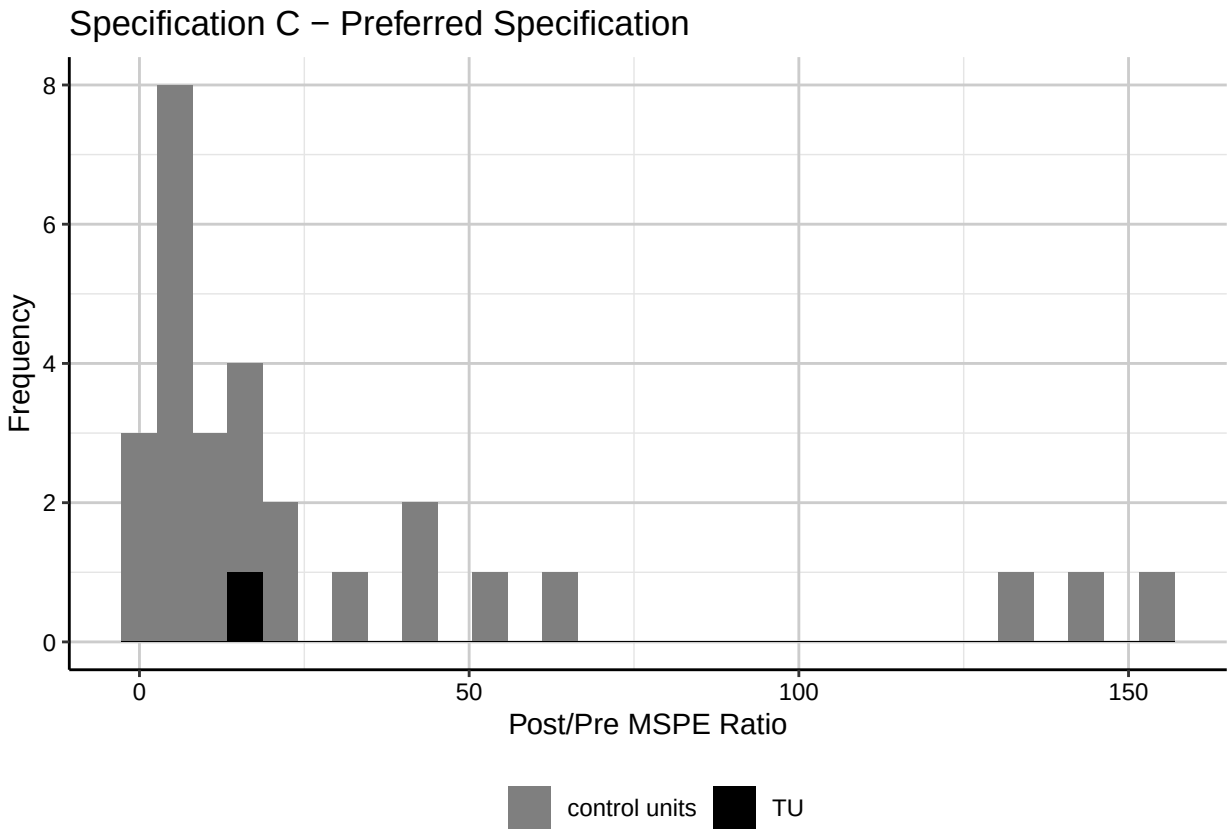

```
## Specification B
mspe.plot(tdfB,
          discard.extreme = TRUE,
          mspe.limit = 20,
          plot.hist = TRUE,
          title = "Specification B",
          xlab = "Post/Pre MSPE Ratio",
          ylab = "Frequency"
          )
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
## Specification C
mspe.plot(tdfC,
          discard.extreme = TRUE,
          mspe.limit = 20,
          plot.hist = TRUE,
          title = "Specification C - Preferred Specification",
          xlab = "Post/Pre MSPE Ratio",
          ylab = "Frequency"
          )
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Based on the above graphs of the Post/Pre MSPE Ratio for our three specifications, we conclude that the treatment was not significant. In our preferred specification (C), 13 control states obtain the same Post/Pre MSPE ratio as the TU or larger. Thus we calculate a p-value of $p = 0.43$ (13/30), that is if we were to assign the intervention at random in the data, the probability of obtaining and Post/Pre MSPE ratio as large as the TU's is 0.43. Based on these results, we cannot reject the null hypothesis of no treatment effect.

(d) How do your synthetic control results compare to your fixed effects results from Question (3) in the last problem set? Interpret any differences.

From the last problem set, Q3 Part f, the results from our FE estimator with all covariates indicated that primary seat belt laws are associated with a 8.98% decrease in log fatalities per capita, ceteris paribus, which is statistically significant at the 1% level. This contrasts with our findings using the synthetic control method, where we cannot reject the null hypothesis that primary seat belt laws have no effect on log fatalities per capita. The difference in results could arise from a few factors. One, the results from our synthetic control model may be less credible if we think our constructed synthetic control does not have a good enough pre-implementation fit. As we can see from part b above, across all of our specifications the gap between the TU and the synthetic control is small (less than 0.1) but not very close to 0, particularly for time periods further away from the implementation year. Two, it could be that our FE estimator suffers from omitted variable bias, whereas our synthetic controls estimator also controls for unobserved state-by-year shocks. If this is the case, then our synthetic controls estimator would be less biased than our FE estimator from the previous problem set.