Marijuana Legalization and Violent Crime

In recent years, many US states have decided to legalize the use of marijuana.

When these ideas were first proposed, there were many theories about the relationship between crime and the "War on Drugs" (the term given to US efforts to arrest drug users and dealers over the past several decades).

In this exercise, we're going to test a few of those theories using drug arrest data from the state of California.

Though California has passed a number of laws lessening penalities for marijuana possession over the years, arguably the biggest changes were in 2010, when the state changed the penalty for possessing a small amount of marijuana from a criminal crime to a "civil" penality (meaning those found guilty only had to pay a fine, not go to jail), though possessing, selling, or producing larger quantities remained illegal. Then in 2016, the state fully legalized marijuana for recreational use, not only making possession of small amounts legal, but also creating a regulatory system for producing marijuana for sale.

Proponents of drug legalization have long argued that the war on drugs contributes to violent crime by creating an opportunity for drug dealers and organized crime to sell and distribute drugs, a business which tends to generate violence when gangs battle over territory. According to this theory, with drug legalization, we should see violent crime decrease after legalization in places where drug arrests had previously been common.

To be clear, this is far from the only argument for drug legalization! It is simply the argument we are well positioned to analyze today.

(Students from Practical Data Science: This should sound familiar! Last semester

we did this analysis in a very simple, crude manner; in this class we'll do it rigorously with your new found difference-in-differences skills!)

Exercise 1

Download and import California arrest data from https://www.github.com/nickeubank/MIDS_Data/UDS_arrest_data.csv [https://www.github.com/nickeubank/MIDS_Data/UDS_arrest_data.csv]. What is a unit of observation (a single row) in this data? What entities are being tracked, and over what time period? (This data is derived from raw California arrest data from the State Attorney General's office here [https://openjustice.doj.ca.gov/data], in the "Arrests" category.)

Note that `VIOLENT` is a count of arrests for violent offenses, and `F_DRUGOFF` is a count of felony drug arrests. `total_population` is total population.

Exercise 2

In this analysis, we will split our sample into "treated" and "control" on the basis of whether a given county had a high average drug arrest rate in the three years before California began drug legalization in 2010. Counties with high drug arrest rates, after all, will be more impacted by drug liberalization policies.

Calculate each county's average drug arrest *rate* for the period from 2007-2009. Then calculate the median value across counties, and create an indicator called `treated` for counties whose average drug arrest rate during this period was above the median average drug arrest rate. In other words, half your counties should be in the "treated" group, and half in "control".

Note that this indicator should be *time-invariant*—if a county is in the treated group, it should always be identified as being in the treated group.

Exercise 3

Our outcome in this analysis is the violent arrest rate – if drug liberalization reduces crime overall, we would expect to see this rate fall in counties with high drug arrest rates after liberalization; if not, we would not expect to see any changes. Create a `violent_rate` variable with is violent arrests per 100,000 people.

Exercise 4

Differences-in-differences get their name from the fact that the estimator, in its most basic implementation, is just the difference between:

- difference in the average change in outcome among eventually-treated units from before to after when treatment is applied, and
- difference in the average change in outcome among never-treated units from before to after when treatment (to the treated units).

(Obviously treatment is never a applied to the never-treated units – when we talk about pre / post, we refer to before and after the point in time in which treatment is applied to the treated units. So if treated units are treated in 2008, then for the never-treated units, we are also comparing outcomes before 2008 to after 2008, even though 2008 has no special significance for the never-treated units).

In its most basic implementation, therefore, calculating a difference-in-difference estimate requires calculating just 4 numbers:

- $\bar{y}_{T=1,Post}$ Avg for Treatment, Post-Treatment
- $\bar{y}_{T=0,Post}$ Avg for Control, Post-Treatment
- $\bar{y}_{T=1,Pre}$ Avg for Treatment, Pre-Treatment
- $\bar{y}_{T=0,Pre}$ Avg for Control, Pre-Treatment

The difference-in-differences estimator $\hat{\delta}$ is defined as

$$\hat{\delta} = (\bar{y}_{T=1,\, Post} - \bar{y}_{T=1,\, Pre}) - (\bar{y}_{T=0,\, Post} - \bar{y}_{T=0,\, Pre})$$

Calculate (a) the change in violent arrest rates for our treated groups from before legalization to after ($\bar{y}_{T=1,Post} - \bar{y}_{T=1,Pre}$), and (b) our difference in difference estimator $\hat{\delta}$ by calculating these four values. Does doing your difference-in-difference estimate tell you something different from what you'd learn if you had just done a pre-post comparison?

For the `Pre` period, consider the three years before liberalization begins in 2010 (e.g. 2007-2009). For the `Post` period, consider the three years after final legalization took place (2016-2018). We will ignore the middle period in which marijuana was decriminalized but not yet legal.

Exercise 5

Now calculate $\hat{\delta}$ using a regression with an indicator for post-2010, an indicator for treated, and an interaction of the two. Use only the same set of years you used above. How does your estimate compare to the estimate you calculated in Exercise 4?

What does this tell you about interpretation of interaction terms with two indicator variables?

Note: You need to cluster your standard errors by county, since we expect counties (over time) to be subject to common fluctuations.

Exercise 6

In the preceding exercise, we did a simple pre-post / treated-control comparison. But one important limitation of these designs is that they do not allow us to test for

parallel trends.

Plot a difference-in-difference model using data from 2000-2009 (inclusive) and from 2016-2018 (inclusive). Note this will have four different geometric components: a time trend for treated counties pre-2010, a time trend for control counties pre-2010, a time trend for treated counties post-2016 (include 2016), and a time trend for control counties post-2016 (include 2016).

Do you see evidence of parallel trends for these two datasets? Does that make you feel more or less confident in your diff-in-diff estimates?

Exercise 7

While we can estimate the model described above precisely as a regression, it's actually much easier to estimate a more flexible model by running the regression we ran in Exercise 5 but with both `county` and `year` fixed effects. Use `PanelOLS` (or `lfe` in R) to estimate this fixed effects regression.

With all these additional fixed effects, do you find evidence that marijuana legalization reduced violent crime?