Evaluation of the Educational Impacts of Recreational Marijuana Legalization in Nevada

Hao Tang, Jinyu Hu,
Karen D. Britt, Liuyang Xu and Yu Wang

APSTA-GE 2401
Statistical Consulting
Applied Statistics for Social Science Research
New York University

Fall 2022

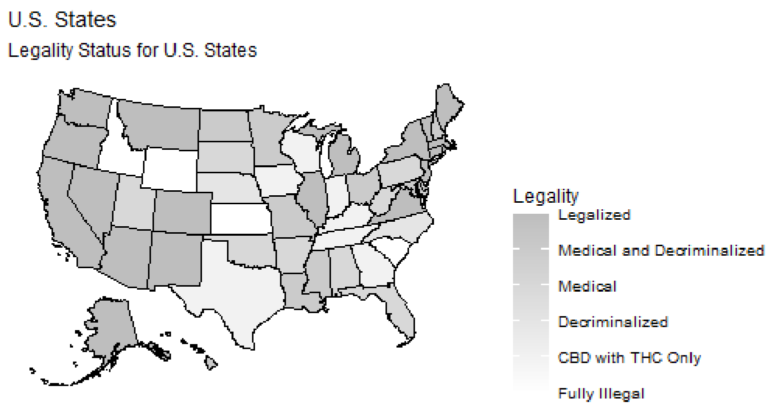
Table of Contents

Introduction

Since the passing of The Marihuana Tax Act of 1937 in the United States it has been illegal on the federal level, banning any use or distribution of the marijuana plant also known as cannabis. Over the decades following its passing, public opinion towards cannabis has changed from a completely negative substance to a more culturally accepted one akin to alcohol. Despite this each state has the ability to choose whether they want to legalize it for medical or recreational use. Reflecting this change in attitude, 20 states have legalized the medical use of cannabis while 19 more states have legalized the recreational use of marijuana in the past ten years. As of 2022, only 11 states have continued to keep cannabis use illegal. However, with the culture and stigma around cannabis changing, it is likely that the pattern of it becoming legal to certain degrees will continue. Nevada legalized medical marijuana in the early 2000s which allowed the use of cannabis for severely ill patients. In 2013 the state of Nevada passed a law expanding this to include those who were prescribed cannabis for less severe illnesses. In 2016 Nevada citizens voted for the full legalization of recreational marijuana for those over the age of 21. This study seeks to evaluate the educational impacts of recreational marijuana on high school graduation and rates and test scores in Nevada from 2009 through 2021. Although the state has legalized it only for those 21 and older, it is possible that minors have been able to obtain the substance illegally which could interfere with their graduation and test rates. This research is beneficial because there is still a lack of research diving into youth's experience and the possible interference cannabis may have on youths' educational outcomes even though cannabis is continuing to be legalized.

Literary Review

In the United States alcohol was made illegal in 1920 and then made legal again in 1933, four years prior to the prohibition of Cannabis. Due to its legal status in the last eighty nine years its acceptance in the country has aided researchers to be given grants and funds for researching various societal effects of alcohol. This has led to a growing body of scholarly knowledge in the topic in comparison to cannabis, which has not been afforded this because of its illegal status and taboo stigma in the United States for so long. This has led to an imbalance in scholarly knowledge between alcohol and cannabis. The Canadian study by Castellanos-Ryan et al. (2017) investigated the bidirectional associations between adolescent cannabis use and neurocognitive performance because according to Castellanos-Ryan et al, "There is evidence to suggest that the adolescent brain may be particularly vulnerable to the neurotoxic effects of substances, especially about neurocognitive functioning (Rubino et al., 2009; Schweinsburg, Brown, & Tapert, 2008). This is because adolescence represents a critical period of neurodevelopment, characterized by synaptic pruning and increased myelination, particularly in cortical and frontal areas of the brain."(2017:1253). The brain is continuing to develop in adolescents making them more vulnerable to substances and other chemicals that may alter the brain's state. The study took a sub-sample of 294 young men from the Montreal Longitude and Experimental Study of Low SES boys (MLES) who were studied annually from 10 to 17 and then again at 20 years of age. They underwent neuro testing at 20 years of age on average. The researchers found that the age of cannabis use initiation and the frequency of its use were linked to a decline in verbal IQ and executive function tasks like in a testing trial and error learning and reward processing by early adulthood. Evaluating the educational impacts of recreational marijuana in Nevada will help grow this body of knowledge by gathering district level data to see how graduation rates have changed each year. This will allow policy makers to determine how to better serve high school students in counties with legalized recreational cannabis.

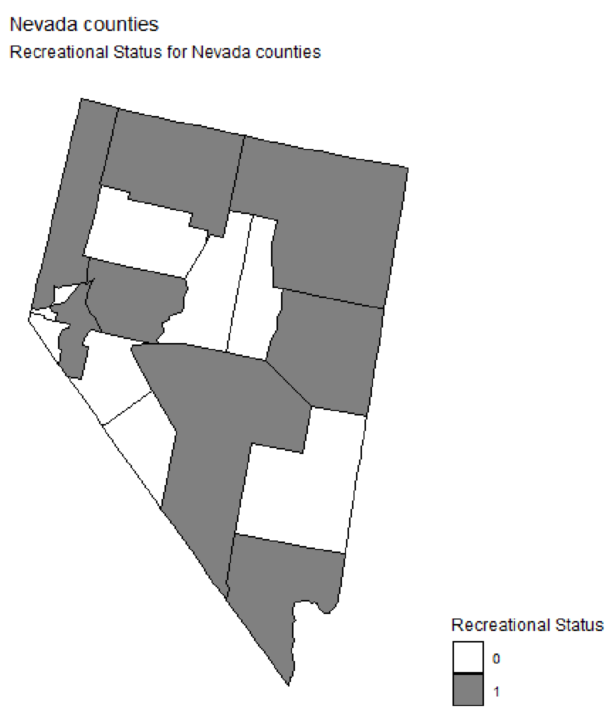


Background

The current study aims to address the educational impacts of recreational marijuana legalization in Nevada. Specifically, we will use the data from CI Graduation data, which contains Graduation information for each public and charter school district in Nevada. Non-Fiscal LEA Geographic data, which provides a directory of all local educational institutions or school districts in the United States. Nevada Accountability Portal data, which Provides detailed information on each school in Nevada, as well as two variables Geographic Relationship data.

As the graph shows below, We use the data relationship with Nevada and separate into counties that allow recreation and counties not, 0 means the county does not allow recreational Marijuana and 1 means the county allows. After grabbing and cleaning the data, we have two methods to use. The first one is of the method we use is focus on using the Difference-in-Differences (DID) method and using r to find out the overall summary of the average treatment effect for the treated subpopulation (ATT), and the second one we use is methods developed by Callaway and Sant'Anna (2020). The authors provide a framework for estimating the "group-time" average treatment effect on the treated (ATT), defined as a unique ATT for a cohort of units treated at the same point in time.

This study will provide important insights into the potential risks and benefits of marijuana legalization in Nevada on student graduation rates and may inform the educational impact students will face in other states when marijuana is legalized.



Data

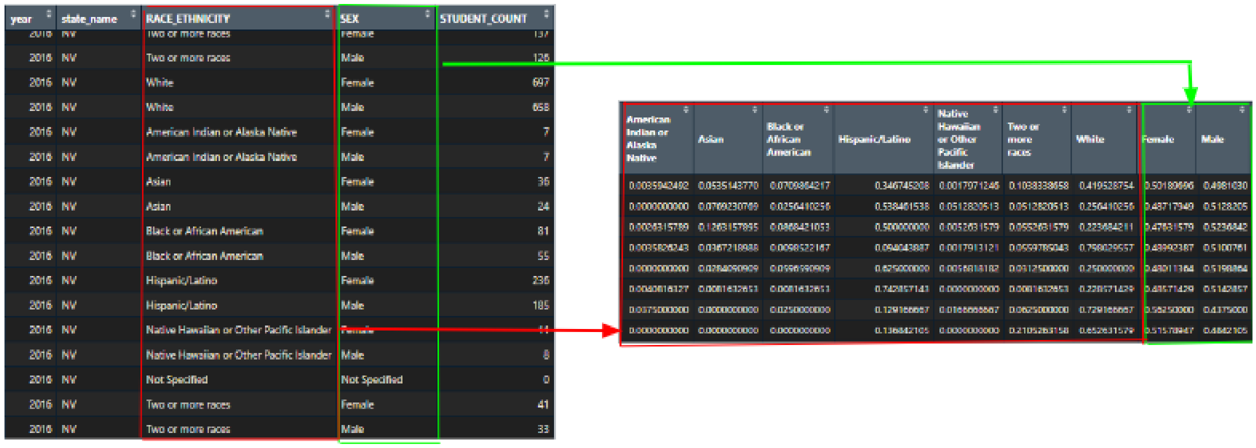
CI Graduation data

The CI Graduation data was pulled from the Nevada Accountability Portal website. The CI Graduation data contains Graduation information for each public and charter school district in Nevada. Information such as district, number of students, graduation rate, number of diplomas given and total college and higher education admittance. The data was cleaned with missing values removed. Only the county, school district, graduating class and graduation rate was kept for the analysis dataset. In addition to this, fips codes were added by district based on the governments fips website.

Non-Fiscal LEA Geographic data

The Non-fiscal LEA Geographic data from 2011 to 2022 was scraped from the NCES (National Center for Education Statistics) website by district level. The LEA data provides a directory of all local education agencies or school districts in the U.S, and contains multiple characteristics in each agency, such as year of data collection, mailing address, age and ethnicity composition, grade level etc. Through this data, each school district's gender ratio, race ratio, and the student enrollment in each year from 2011 to 2022 in Nevada will be generated at the end.

The LEA data from 2011 to 2015 are presented in the form of txt files, while the data from 2016 to 2022 are presented in csv. Although the two are data representing the same geographic information, their presentation forms have changed over time. Therefore, Selecting the variables needed for further modeling, and then mutating them with the same name is the first step. The LEA raw data only provides the specific quality for each variable, especially in gender and race, so the gender ratio and race ratio need to be calculated for each Local Education Agency in Nevada from 2011 to 2022.

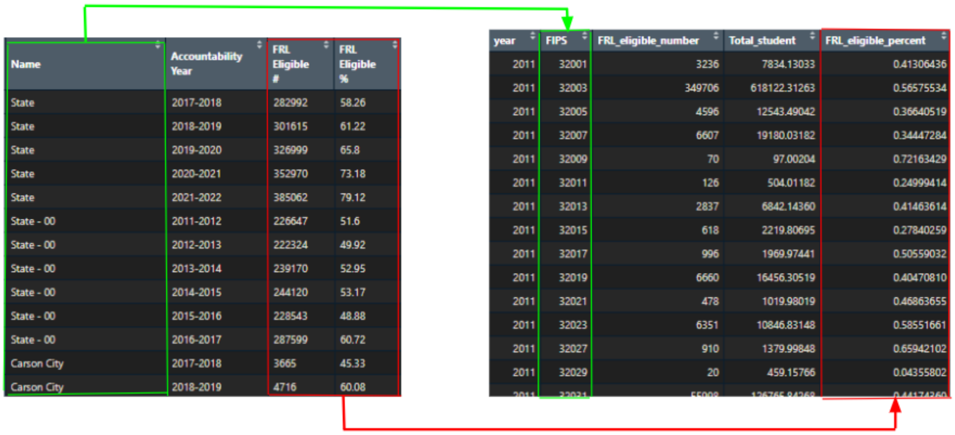


(Generating gender ratio and race ratio)

Nevada Accountability Portal data

The Nevada Accountability Portal data from 2011 to 2023 was downloaded from the NDE (Nevada Department of Education) website by school level. The data provides detailed information for each school in Nevada, and two variables (the population and population proportion of Free or Reduced Priced Lunch Eligible) will be adopted from this data set.

The Nevada Accountability Portal data is based on school level, but the LEA data is based on the district level; Therefore, to calculate weighted average for each school and shape the data to district level is the idea for cleaning up this dataset.



Name	Accountability Year	FRL Eligible #	FRL Eligible %
State	2017-2018	282992	58.26
State	2018-2019	301615	61.22
State	2019-2020	326999	65.8
State	2020-2021	352970	73.18
State	2021-2022	385062	79.12
State - 00	2011-2012	226647	51.6
State - 00	2012-2013	222324	49.92
State - 00	2013-2014	239170	52.95
State - 00	2014-2015	244120	53.17
State - 00	2015-2016	228543	48.88
State - 00	2016-2017	287599	60.72
Carson City	2017-2018	3665	45.33
Carson City	2018-2019	4716	60.08

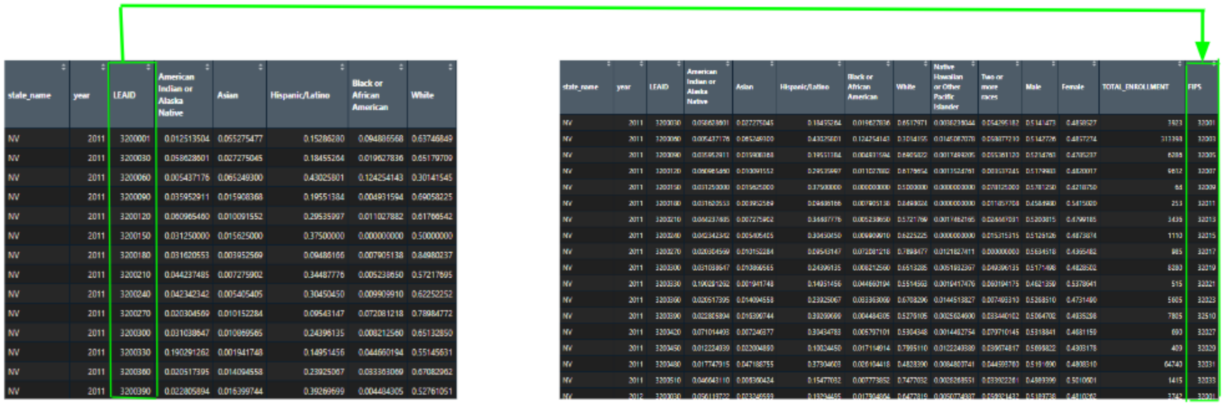
year	FIPS	FRL_eligible_number	Total_student	FRL_eligible_percent
2011	32001	3236	7834.13033	0.41306436
2011	32003	349706	618122.31263	0.56575534
2011	32005	4596	12543.49042	0.36640519
2011	32007	6607	19180.03182	0.34447284
2011	32009	70	97.00204	0.72163429
2011	32011	126	504.01182	0.24999414
2011	32013	2837	6842.14360	0.41463614
2011	32015	618	2219.80695	0.27840259
2011	32017	996	1969.97441	0.50559032
2011	32019	6660	16456.30519	0.40470810
2011	32021	478	1019.98019	0.46863655
2011	32023	6351	10846.83148	0.58551661
2011	32027	910	1379.99848	0.65942102
2011	32029	20	459.15766	0.04355802

(**Transfering school district to county level Geo-indicator, and re-calculating eligible percent in each county**)

Geographic Relationship data

As mentioned above, The Nevada Accountability Portal data and LEA data were calculated by different levels (county and LEA district). In order to map them together, the Geographic Relationship data will be introduced here.

The School District Geographic Relationship Files (GRF) were built to provide a complete set of geographic associations between school districts and other types of geographic areas including counties. More specifically, it provides the LEA ID and its corresponding county code which could help mapping the LEA data with Nevada Accountability Portal data, and merging them together.



(**Transfering LEA district to county level Geo-indicator**)

year	FIPS	FRL_eligible_percent	state_name	LEAID	American Indian or Alaska Native	Asian	Hispanic/Latino	Black or African American	White	Native Hawaiian or Other Pacific Islander	Two or more races	Male	Female	TOTAL_ENROLLMENT
2011	32001	0.41306436	NV	3200030	0.058628601	0.027275045	0.18455264	0.019627836	0.6517971	0.0038236044	0.054295182	0.5141473	0.4858527	3923
2011	32003	0.56575534	NV	3200060	0.005437176	0.065249300	0.43025801	0.124254143	0.3014155	0.0145087078	0.058877210	0.5142726	0.4857274	313398
2011	32005	0.36640519	NV	3200090	0.035952911	0.015908368	0.19551384	0.004931594	0.6905822	0.0017499205	0.055361120	0.5214763	0.4785237	6286
2011	32007	0.34447284	NV	3200120	0.060965460	0.010091552	0.29535997	0.011027882	0.6176654	0.0013524761	0.003537245	0.5179983	0.4820017	9612
2011	32009	0.72163429	NV	3200150	0.031250000	0.015625000	0.37500000	0.000000000	0.5000000	0.0000000000	0.078125000	0.5781250	0.4218750	64
2011	32011	0.24999414	NV	3200180	0.031620553	0.003952569	0.09486166	0.007905138	0.8498024	0.0000000000	0.011857708	0.4584980	0.5415020	253
2011	32013	0.41463614	NV	3200210	0.044237485	0.007275902	0.34487776	0.005238650	0.5721769	0.0017462165	0.024447031	0.5200815	0.4799185	3436
2011	32015	0.27840259	NV	3200240	0.042342342	0.005405405	0.30450450	0.009909910	0.6225225	0.0000000000	0.015315315	0.5126126	0.4873874	1110
2011	32017	0.50559032	NV	3200270	0.020304569	0.010152284	0.09543147	0.072081218	0.7898477	0.0121827411	0.000000000	0.5634518	0.4365482	985
2011	32019	0.40470810	NV	3200300	0.031038647	0.010869565	0.24396135	0.008212560	0.6513285	0.0051932367	0.049396135	0.5171498	0.4828502	8280
2011	32021	0.46863655	NV	3200330	0.190291262	0.001941748	0.14951456	0.044660194	0.5514563	0.0019417476	0.060194175	0.4621359	0.5378641	515
2011	32023	0.58551661	NV	3200360	0.020517395	0.014094558	0.23925067	0.033363069	0.6708296	0.0144513827	0.007493310	0.5268510	0.4731490	5605
2011	32027	0.65942102	NV	3200420	0.071014493	0.007246377	0.30434783	0.005797101	0.5304348	0.0014492754	0.079710145	0.5318841	0.4681159	690
2011	32029	0.04355802	NV	3200450	0.012224939	0.022004890	0.10024450	0.017114914	0.7995110	0.0122249389	0.036674817	0.5696822	0.4303178	409
2011	32031	0.44174360	NV	3200480	0.017747915	0.047188755	0.37304603	0.026104418	0.4828390	0.0084800741	0.044593760	0.5191690	0.4808310	64740
2011	32033	0.38275273	NV	3200510	0.046643110	0.006360424	0.15477032	0.007773852	0.7477032	0.0028268551	0.033922261	0.4989399	0.5010601	1415
2011	32510	0.50267762	NV	3200390	0.022805894	0.016399744	0.39269699	0.004484305	0.5276105	0.0025624600	0.033440102	0.5064702	0.4935298	7805
2012	32001	0.43460065	NV	3200030	0.056119722	0.023249599	0.19294495	0.017904864	0.6477819	0.0050774987	0.056921432	0.5189738	0.4810262	3742

(**Final data after the mapping procedure**)

Methods

The question of interest, “Does marijuana legalization affect the high school graduation rate of the districts in Nevada?” can be translated into finding the effect for the treatment group after 2017 when the legalization happened. In order to find the average treatment effect for treatment (ATT), we need to check what the data looks across different groups.

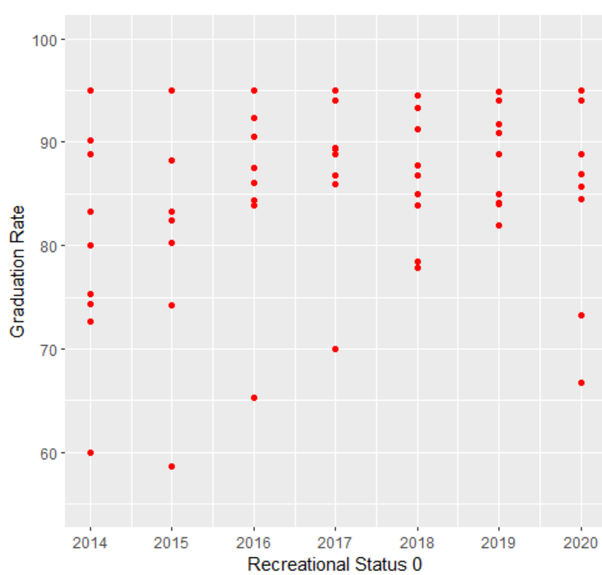
The counties that allow recreational sales after the legalization are Churchill County, Clark County, Elko County, Humboldt County, Lyon County, Nye County, Washoe County, and White Pine County. Before choosing our method for analyzing the data, we first took a closer look at the comparison of data distribution of all the outcomes and potential covariates between the groups of counties that allow and don't allow recreational sales. A summary of the data is shown in the table below. All values for both groups are under the format of Mean(SD).

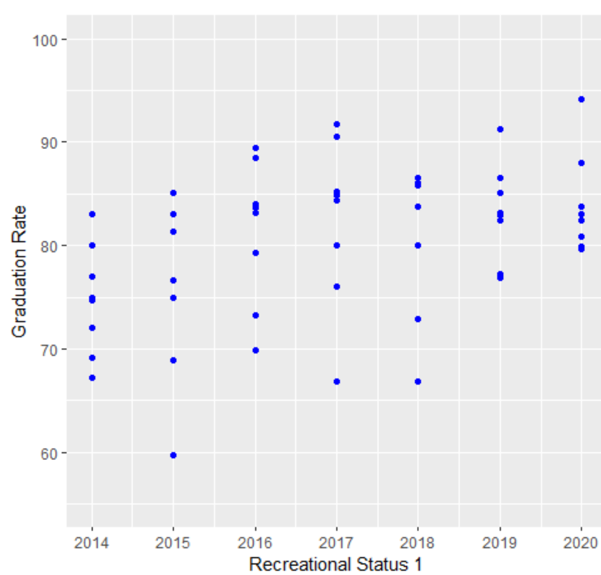
Stratified by Recreational Status	0	1	p-test
Count	63	56	
Graduation Rate	85.33 (8.84)	80.38 (7.14)	0.001
FRL eligible number	2616.03 (3251.01)	65934.21 (145356.25)	0.001
American Indian or Alaska Native	0.05 (0.04)	0.03 (0.02)	0.052
Asian	0.01 (0.01)	0.02 (0.02)	<0.001
Hispanic/Latino	0.25 (0.13)	0.31 (0.09)	0.003
Black or African American	0.02 (0.02)	0.03 (0.04)	0.003
White	0.64 (0.13)	0.55 (0.14)	<0.001
Native Hawaiian or Other Pacific Islander	0.00 (0.00)	0.01 (0.00)	<0.001
Two or more races	0.04 (0.02)	0.05 (0.02)	0.163
Male	0.52 (0.02)	0.51 (0.01)	0.036
Female	0.48 (0.02)	0.49 (0.01)	0.036
Total enrollment	10289.43 (15136.87)	190879.61 (467536.92)	0.003

More specifically, the standardized mean differences between 2 groups is shown in the table below.

Recreational Status	0 vs. 1
Graduation Rate	0.6162720
FRL eligible number	0.6158871
American Indian or Alaska Native	0.3670685
Asian	0.7464486
Hispanic/Latino	0.5688609
Black or African American	0.5438340
White	0.6654979
Native Hawaiian or Other Pacific Islander	1.1229159
Two or more races	0.2573643
Male	0.3978854
Female	0.3978854
Total enrollment	0.5459662

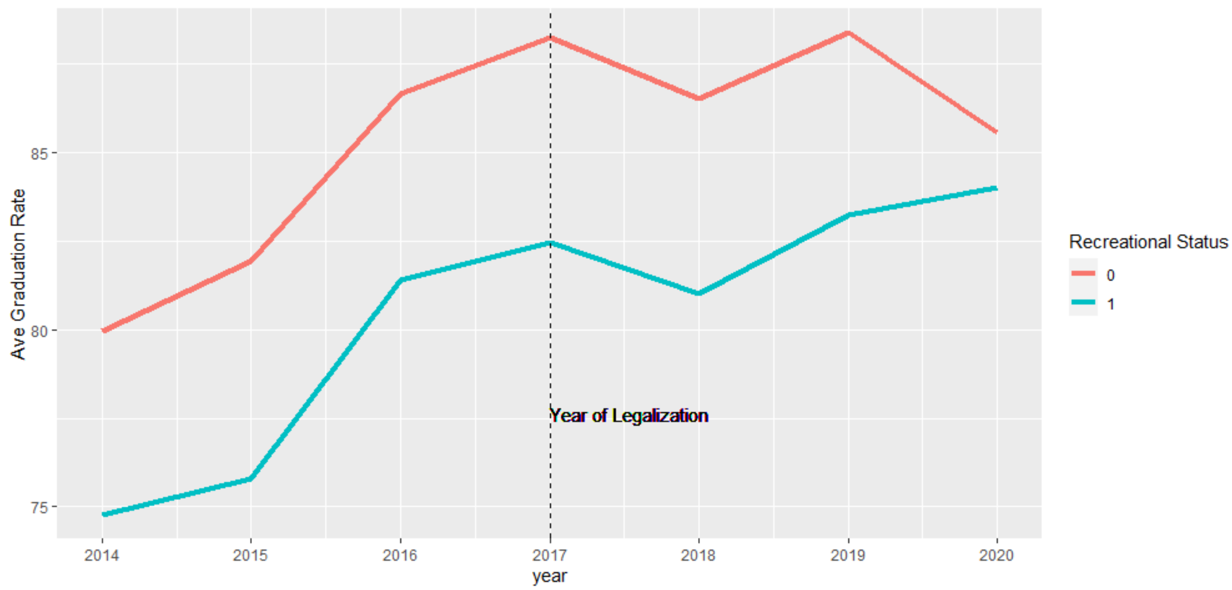
From the tables above we can tell that the data across different groups are quite similar. Since we particularly interested in the graduation rate as the outcome, we drew a scatter plot to show the graduation rate distribution for the control and treatment groups across different years.





The plot indicates that the graduation rate has a trend of increasing as time goes by, and it also shows that there might be a linear relationship between the year and graduation rate for the treatment group. The control group, however, always has some observations with a high graduation rate each year.

To check the general changing trends for both groups, we drew a line chart based on the average graduation rate for each group. The figure is shown below.



Ignoring all the other covariates and only focusing on how the average graduation rate changes, we can see that the lines for both the treatment group and control group have similar slopes before the year 2017, when recreational sales were first allowed. The slopes after 2017 are also quite similar, but there is a huge difference after 2019. So there seems to be a parallel trend between these two groups. Further parallel trends examination will be conducted later.

Since we can only observe the graduation rate when the treatment group got the treatment after 2017 and the graduation rate when the control group never got the treatment after 2017, and we couldn't have known any counterfactual data, based on this structure of the data, as one of the most efficient ways to find the causal effect for government policies, a classical Difference in Differences(DID) method is suitable for achieving our question of interest.

For our models, in general, the continuous outcome variable is the graduation rate(in percent), the binary treatment variable is the recreational sale legalization status in each county(legalized as 1), and the potential covariates are free/reduced lunch eligible student number, race(American Indian or Alaska Native, Asian, Hispanic/Latino, Black or African American, White, Native Hawaiian or Other Pacific Islander, two or more races), gender(male/female), total enrollment. All races and genders are treated as binary variables.

Unlike the normal data set with only two time periods when applying the DID, our data set has multiple years from 2014 to 2020. In order to obtain a more accurate estimate of the causal effect and allow the model to keep being useful for the potential future data, we applied three different types of DID for analysis.

(1) DID with the canonical format for 2016-2017

For canonical format DID, there are usually two time points, t0 and t1, in the dataset. At t0, both the treatment group and control group have not received the effect of treatment. That is, treatment has not yet occurred. t0 is followed by t1. The treatment group receives the effect of treatment, while the control group remains unchanged. We assume that the outcome of the two groups is equally affected by time. Thus we can use the change in the control group to estimate the counterfactual data of the treatment group to estimate our required ATT.

Let Y be the outcome variable and D be the treatment variable, we have

$$Y_{it0} = Y_{it0}(0)$$

$$Y_{it1} = D_i Y_{it1}(1) + (1 - D_i)Y_{it1}(0)$$

And the ATT can be given by

$$ATT = E[Y_i(1) - Y_i(0)|D = 1]$$

when we assume the parallel trends assumption

$$E[Y_{t1}(0) - Y_{t0}(0)|D = 1] = E[Y_{t1}(0) - Y_{t0}|D = 0]$$

The safest way to make the causal interpretation for the canonical format of DID is simply focusing on the effect between 2016 and 2017 when the legalization happened. We first tried the regression model only with the outcome variable and the treatment variable as such

$$Y \sim D$$

We then tried the regression model with all the covariates as such

$$Y \sim D + FRL_{eligible} + race + gender + enrollment_{total}$$

The race are binary variables of American Indian or Alaska Native, Asian, Hispanic/Latino, Black or African American, White, Native Hawaiian or Other Pacific Islander, two or more races and the gender are binary variables of male and female.

We also tried using stepAIC to simplify the model without impacting much on the performance.

(2) DID with all the years before and after legalization

Though only focusing on 2016 and 2017 is best for the causal interpretation of the canonical format of DID, after getting the estimation of the effect between 2016 and 2017, we still want to find out what the effect of marijuana is among all years.

In order to assess the ATT among all the years after legalization, we considered a second method which we see all the years after marijuana is legalized as one entire time period. Thus all the years before 2017 are all collapsed and considered as t0, and all the years after 2017 are all collapsed and considered as t1.

We applied this method to the 3 regression models described above.

(3) DID with Multiple Time Periods

Though the second method can assess the estimate of the effect of legalization among all years, it ignores the cases when there are different time effects in different years. We then rely on a third recent method called DID with Multiple Time Periods. We implemented this method by using a package called "did" developed by the authors of the paper. The package can give us a separate estimate for different years, an overall summary of ATT, the confidence interval for each result, and whether the parallel trends assumption is valid.

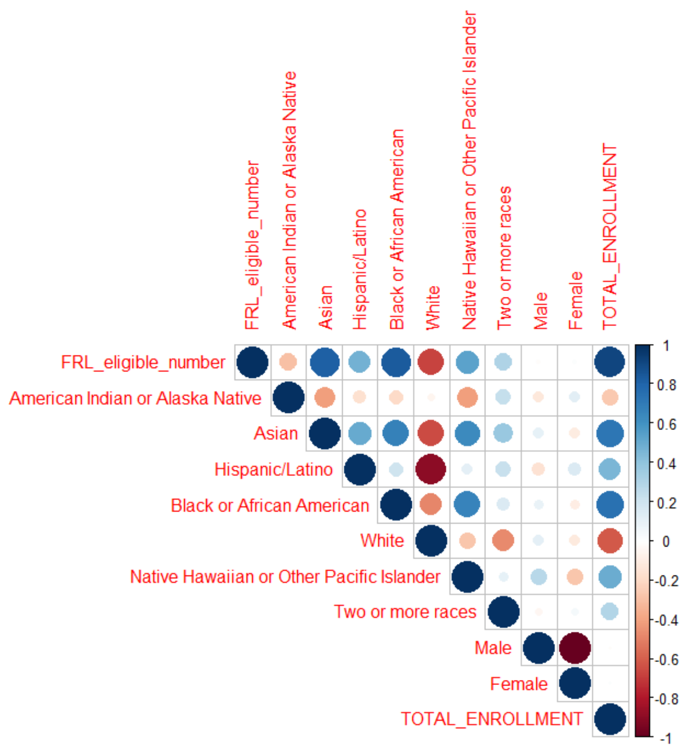
Since we are taking multiple time periods into consideration and we are mainly concerned with the time after legalization, we let the year 2017 of legalization as t, the year 2016 as t-1, the year 2018 as t+1, etc.

Similar to above, we first tried the regression model only with the outcome variable and the treatment variable as such

$$Y \sim D$$

There is a restriction between the amount of group and covariates in this method, so we could only include parts of the covariates in the regression model as a comparison.

We checked the correlation between the covariates. The correlation plot is shown below.



We excluded the covariates that have strong correlations and tested the regression model with different amounts of covariates. The regression model with only one covariate:

$$Y \sim D + FRL_{eligible}$$

The regression model with 2 covariates:

$$Y \sim D + FRL_{eligible} + Female$$

The regression model with 3 covariates:

$$Y \sim D + FRL_{eligible} + Female + races_{two/more}$$

The method allows us to both predict ATT at different times and to aggregate these different ATTs at different times into an overall summarized ATT among all four regression models.

Results

We get results from the DID under three different conditions, the first one is the canonical format DID in the two years before and after the legalization. The second is to evaluate DID in all years before and after the legalization, the third one is applying the DID with multiple periods, the overall summary of ATTs based on group.

(1) First, we apply the canonical format of DID in the two years before and after the legalization.

The model chosen by stepAIC is

$$Y \sim D + FRL_{eligible} + race_{American\ Indian/Alaska\ Native} + race_{Asian} + race_{Hispanic/Latino} + race_{white} + race_{Native\ Hawaiian/Other\ Pacific\ Islander} + enrollment_{total}$$

The estimated ATT are shown below:

	estimate	p-value
without covariate	-0.5389	0.923
with all covariates	-1.465	0.7190
with covariates chosen by stepAIC	-1.557	0.692885

(2) Then we apply the canonical format of DID in all years before and after the legalization.

The model chosen by stepAIC is

$$Y \sim D + FRL_{eligible} + race_{American\ Indian/Alaska\ Native} + race_{Hispanic/Latino} + race_{white} + race_{Black/African\ American} + Female$$

The estimated ATT are shown below:

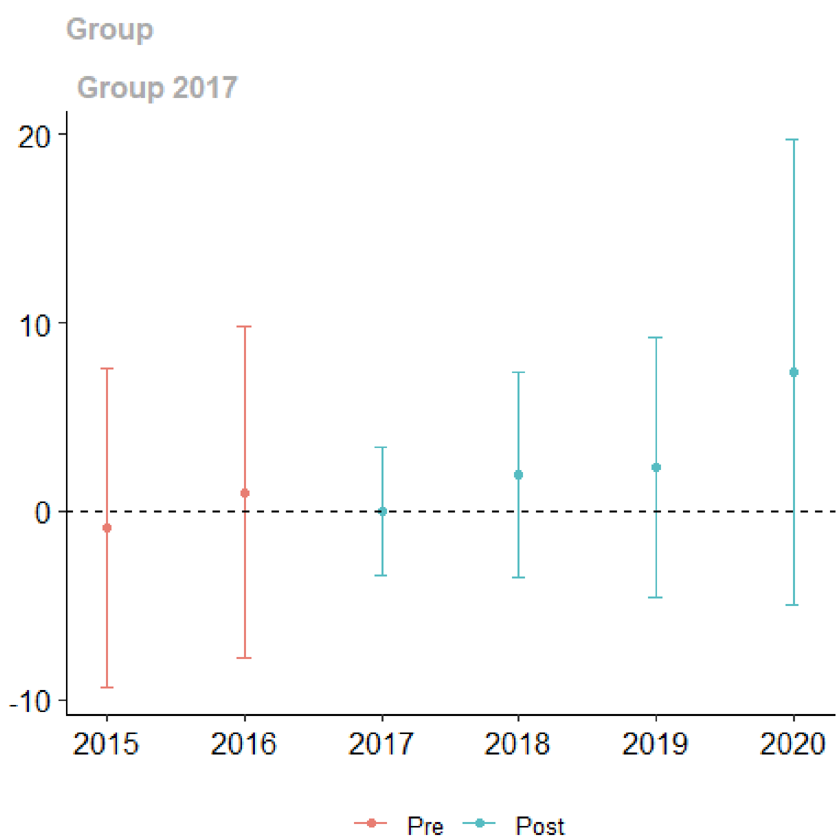
	estimate	p-value
without covariate	1.021	0.7244
with all covariates	1.147	0.6829
with covariates chosen by stepAIC	1.593	0.5568

(3) In the third method, we applied the DID with multiple periods.

(i) For the model without covariate

$$Y \sim D$$

The figure and table below showed the estimated ATT.

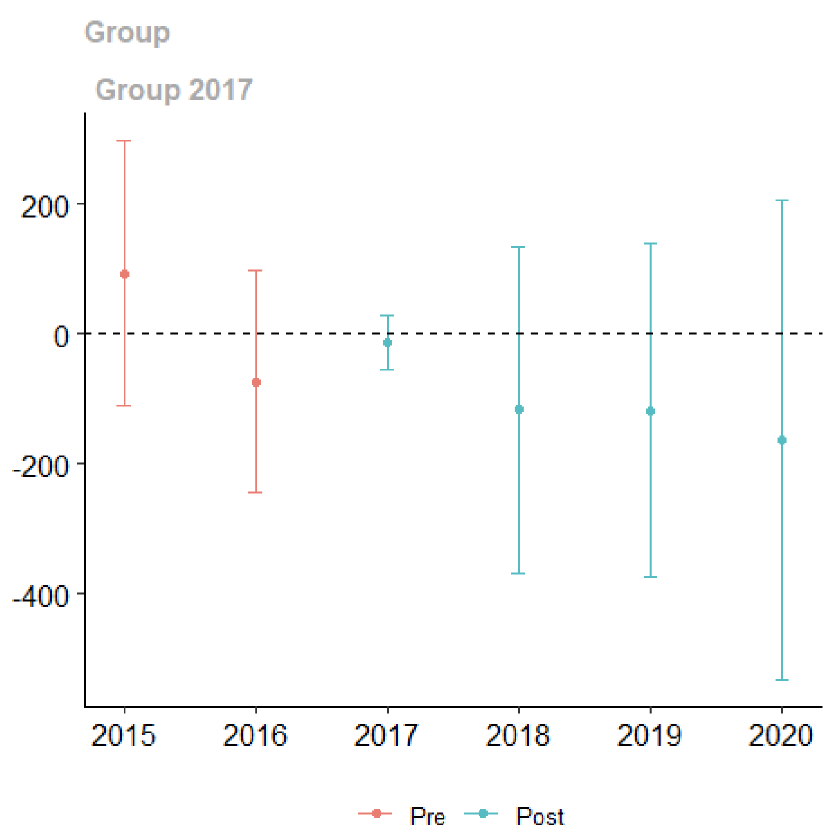


Group Time	estimate	standard error	95% CI
2015	-0.8857	4.1286	[-10.134, 8.363]
2016	1.0089	3.8064	[-7.518, 9.536]
2017	-0.0214	1.4575	[-3.286, 3.243]
2018	1.9732	2.6346	[-3.929, 7.875]
2019	2.3125	3.0761	[-4.578, 9.203]
2020	7.3714	5.3797	[-4.680, 19.422]

(ii) For the model with 1 covariate

$$Y \sim D + FRL_{eligible}$$

The figure and table below showed the estimated ATT.

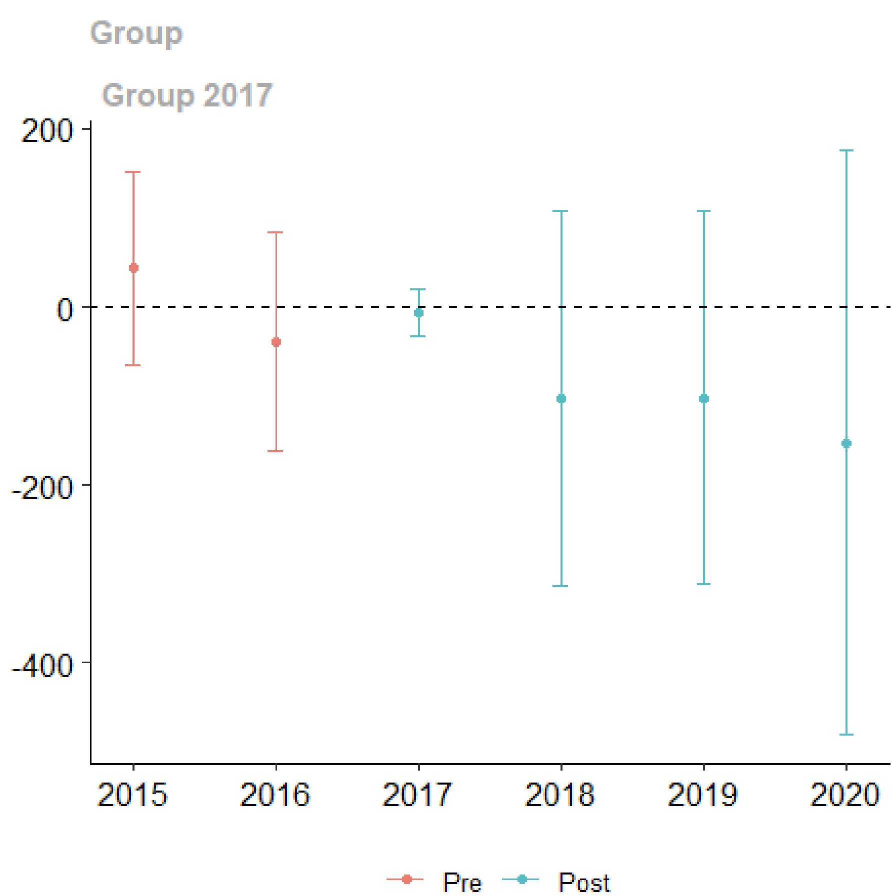


Group Time	estimate	standard error	95% CI
2015	43.0925	54.2902	[-64.308, 150.493]
2016	-39.4643	61.4347	[-160.999, 82.070]
2017	-6.0415	14.2843	[-34.300, 22.217]
2018	-103.4610	108.0483	[-317.210, 110.288]
2019	-102.1314	108.5551	[-316.883, 112.620]
2020	-153.3583	162.8908	[-475.601, 168.885]

(iii) For the model with 2 covariates

$$Y \sim D + FRL_{eligible} + Female$$

The figure and table below showed the estimated ATT.

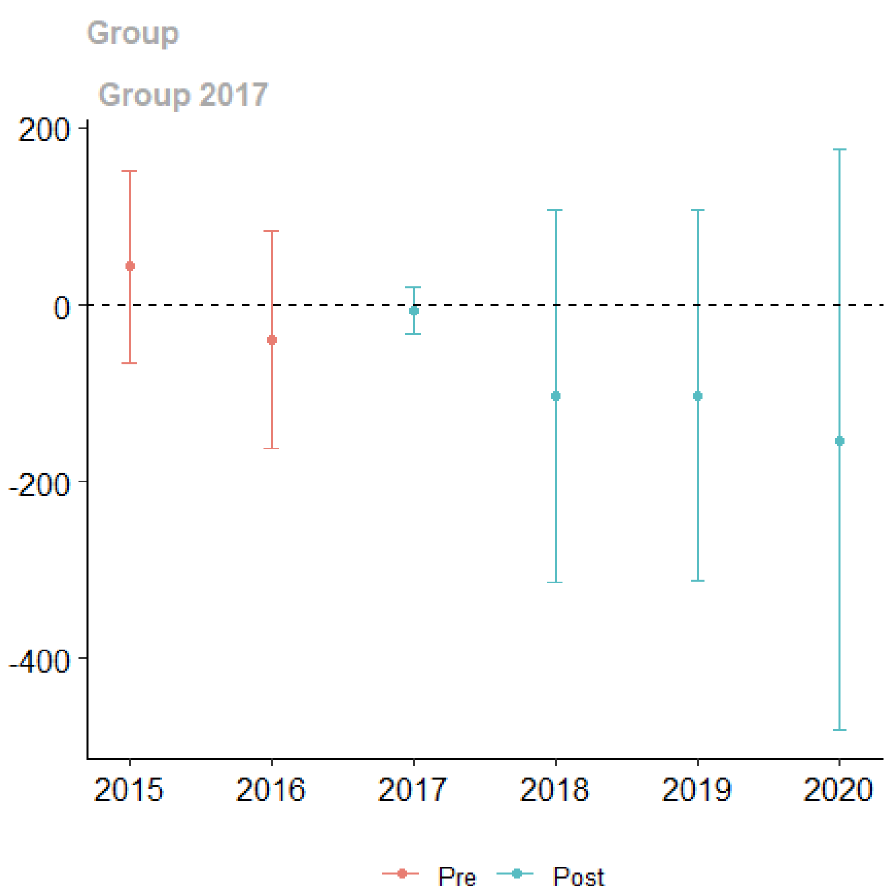


Group Time	estimate	standard error	95% CI
2015	93.3274	102.7470	[-119.150, 305.805]
2016	-73.7978	84.3425	[-248.215, 100.620]
2017	-117.0481	126.4681	[-378.580, 144.484]
2018	-103.4610	108.0483	[-317.210, 110.288]
2019	-118.0876	130.4917	[-387.940, 151.765]
2020	-164.8292	181.1274	[-539.395, 209.736]

(iv) For the model with 3 covariates

$$Y \sim D + FRL_{eligible} + Female + race_{two/more}$$

The figure and table below showed the estimated ATT.



Group Time	estimate	standard error	95% CI
2015	21.8577	46.8718	[-75.128, 118.843]
2016	-27.4589	46.7786	[-124.252, 69.334]
2017	4.5045	10.1086	[-16.412, 25.421]
2018	-94.3746	99.3184	[-299.880, 111.132]
2019	-101.6302	110.4875	[-330.247, 126.986]
2020	-129.9604	149.0971	[-438.467, 178.546]

(v) The overall summary of ATTs based on group/cohort aggregation is shown below:

	estimate	standard error	95% CI
without covariate	2.9089	2.7181	[-2.418, 8.236]
with 1 covariate	-91.248	100.3685	[-287.967, 105.471]
with 2 covariates	-89.0468	96.9828	[-279.130, 101.036]
with 3 covariates	-80.3652	87.184	[-251.243, 90.512]

The result of our study shows that in the period of DID in 2016-2017, the negative influence is small. In the second condition before and after legalization, the positive effect is small. In the third condition, the multiple period, the DID didn't show a significant result. In conclusion, without confunder, the result is positive but small, and is close to estimate under the first and second condition. In situations with confunders, the result is negative and large. In the multiple period, in the condition without confunder, the year 2017 has little negative effect and increases by the year back to positive effect. If with confunders, 2017 starts with little negative effect and increases by the year. Also, after we review the p-value of the first and second condition, we can find out that the p-value is larger than 0.05, which is not significant.

Limitations

While Difference in Difference with Multiple Time Periods Design is a powerful statistical method in terms of estimating the causal effect of participating in treatment on some outcome between multiple time slots. According to Brantly Callaway and Pedro H.C. Sant'Anana, there are three main assumptions that must be declared and clarified priorly in order to draw a reliable causal inference from using this method. After carefully studied the logistics and mathematics behind this statistical method from Callaway and Sant'Anana's article, we list the three essential assumptions with detailed explanations below.

(1) Staggered Treatment Adoption Assumption

The first core assumption of DID with Multiple Periods is the Stagged Treatment Adoption. As the name described, this assumption implies that for those participants who receive the treatment assignment, they will remain treated during the rest of the research study period. In other words, their treatment assignment will be fixed and cannot be changed once it is assigned. In terms of mathematical interpretation, the formula for the Staggered Treatment Adoption Assumption can be written as

$$For\ t = 1, ..., T - 1,\ D_{it} = 1 \Rightarrow D_{it+1} = 1$$

In the above mathematical formula, we use D_{it} as an indicator variable for whether the unit i has been treated by time t. Roughly speaking, without assuming the treatment effect heterogeneity across time, groups, and treatment, it would be extremely hard for the researchers to measure the causal effect.

(2) Parallel Trends Assumption based on never-treated units

The second major assumption of DID with Multiple Periods is the Parallel Trends Assumption based on never-treated units, and the mathematical formula can be written as

$$For\ all\ g = 2,....,T\ ,\ t = 2,....,T\ with\ t \geq g$$

$$E[Y_t(0) - Y_{t-1}(0)|G = g] = E[Y_t(0) - Y_{t-1}(0)|C = 1]$$

In the above equation, $Y(0)$ is the untreated potential outcome, G is the time when units become treated, C is an indicator variable indicate units are in the never-treated group, T is the time period. This based on the never-treated units parallel assumption illustrates that without the intervention of treatment assignment, the potential outcome for the untreated groups first become treated in time g will parallel with the average untreated potential outcome for the never treated groups in all post-treatment periods, $t \geq g$. One thing that should be clarified is that this parallel assumption only relies on using the never-treated groups as a baseline comparison. In order for this assumption works appropriately, we also need to suppose that 1) The proportion for the never-treated group should be large enough in the dataset. 2) Units in the never-treated group should be similar enough to the units in the eventually treated group so that we can indeed make a comparison.

(3) Parallel Trends Assumption based on not yet treated units

If the large enough of never-treated and similarity assumptions cannot be satisfied, an alternative parallel assumption can be applied. The mathematical formula for this assumption can be written as

$$For\ g = 2,....,T,\ s,t = 2,....,T\ with\ t \geq g\ and\ s \geq t$$
$$E[Y_t(0) - Y_{t-1}(0)\,|G = g] = E[Y_t(0) - Y_{t-1}(0)|D_s = 0,\ G \neq g]$$

Here we set $Y(0)$ is the untreated potential outcome, G is the time when units become treated, C is an indicator variable indicates units are in the never-treated group, D is an indicator variable for whether the unit i has been treated by time t. This implies that to calculate the average treatment effect for the group that first becomes treated, we can use the not-yet-treated group by the time $s(s \geq t)$ as a baseline comparison.

As a result, when implementing the Difference in Difference with Multiple Periods, researchers should be very careful with all three assumptions listed above, such as examining the dataset to check the assumptions' plausibility and possible violations. Fortunately, in our study, we found that the assumptions are met, and we can confidently apply DID with Multiple Periods to answer our research question.

Conclusion

In summary, after we compared traditional Difference In Difference and Difference In Difference with Multiple time Periods, also including and excluding some confounding variables, we found that due to the shortage of the data collection and limited number of observations, it is too weak to reject the null hypothesis until this point which is "There are no effects of Marijuana legalization on high school graduation rate". This result somehow contradicts our intuition on the effects of Marijuana legalization program, but the reason could be due to shortage of data and limited observations in the state of Nevada. As a consequence, it is necessary to perform this study after we collect more data in the future.

Bibliography

Stinnesbeck, Jann. 2018."Fact Sheet Recreational Marijuana in Nevada". Retrieved December 6th 2022. (https://www.leg.state.nv.us/Division/Research/Publications/Factsheets/RecreationalMarijuana.pdf)

Brantly Callaway, Pedro H. C. Sant'Anna. "Difference-in-Differences Methods". Retrieved December 6th 2022. (Function reference • did (bcallaway11.github.io))

Castellanos-Ryan, Natalie; Pingault, Jean-Baptiste; Parent, Sophie; Vitaro, Frank; Tremblay, Richard E ; et al. (2017). Adolescent cannabis use, change in neurocognitive function, and high-school graduation: A longitudinal study from early adolescence to young adulthood. *Development and Psychopathology*; Cambridge Vol. 29, Iss. 4, (Oct 2017): 1253-1266.

FCC. "Federal Information Processing System (FIPS) Codes for States and Counties". Retrieved December 6th 2022. (https://transition.fcc.gov/oet/info/maps/census/fips/fips.txt)

Appendix

The source code of this paper is here.