How Much Does Minimum Legal Drinking Age Reduces the Probability of A Young Californian Being Arrested

Abstract:

In order to answer the question: “How much does minimum legal drinking age reduces the probability of a young Californian being arrested?” We need to access the true treatment effect of MLDA on arrests in terms of per person drinking. We set MLDA as the treatment of a Regression Discontinuity Design. Using the dataset: “National Health Interview Adult Files, 1997-2007”, we are able to estimate the First Stage, which is the complier proportion of the population in the case of MLDA. Using the official dataset “Arrest rates for California”, we are able to estimate the reduced form, which is the actual increase in number of crimes/arrests after MLDA. We can then further estimate the true treatment effect of MLDA on arrests using an IV estimator. Although only two of the three IV assumptions are met, we are statistically confident to conclude: first, MLDA reduces the proportion of the population that drinks; second, MLDA reduces crime and arrests numbers in young Californians; third, MLDA reduces the probability of a young Californian being arrested.

Introduction:

Does Minimum Legal Drinking Age helps the young population of California to stay out of trouble? In another word, does minimum legal drinking age reduces the probability of a young Californian being arrested? As we all know, the MLDA (Minimum Legal drinking Age) in the United States in twenty-one years old. However, alcohol is common in college, even high school parties. Some people might debate that MLDA effect anything since there are already a bunch of high-school and college kids drinking before MLDA. So, does MLDA reduce the drinking population? We access this question by setting the MLDA as the threshold of a Regression

Discontinuity Design, which can be a form of Instrumental Variable Approach. In this case, we have the instrumental variable Zi indicating if a person is over or under MLDA. We denote instrument variable Zi = 0 for people who are under MLDA; and we denote Zi,1 = 1for people who are over MLDA. We also have the variable Di indicating whether a person drinks or not. We denote Di,0 = 0 as people who does not drink; and Di,1 = 1 as people who actually drinks. Then, we can separate the total population into four sub-populations: always drinkers, never drinkers, compliers, and defiers. The always takers and never takers populations are the same groups of people before and after MLDA. Intuitively, these two sub-populations cancel each other out when estimating the effect of MLDA on the proportion of the population that drinks. Also, the second assumption of the Regression Discontinuity design does not allow defiers. Thus we are only estimating complier proportion of the population in the first stage of this regression discontinuity design. We use data from the dataset "National Health Interview Adult Files 1997-2007". Firstly, we run a density test to make sure the data is least manipulated; or the data represents an accurate reflection of the population characteristics. Secondly, we validate the first assumption of Regression Discontinuity Design by running regressions on the observable characteristics and check if there are any sharp changes in the covariates due to MLDA. If there are no sharp changes in covariates due to MLDA, we assume the continuity of covariates, which indicates similar potential outcomes between the population under MLDA and the population over MLDA. Thirdly, we use the parametric approach to regress the instrument Zi on Di to obtain a negative effect of MLDA on the proportion of the population that drinks, the complier population. We adjust the flexibility of the regression to get the most accurate estimation of the effect of MLDA on the complier population.

MLDA reduces the proportion of the population that drinks. We also need to know the effect in total numbers of arrests between MLDA and crime and arrests numbers, the reduced form. The additional data that we use is from the Arrest rates for California dataset that contains arrest rates per 10,000 by age all over and broken down by cause. Firstly, we regress age with all causes of arrests and each sub-category cause of arrests to obtain a negative effect of the MLDA on all arrest rates and each sub-category cause of arrests. Secondly, we adjust the flexibility of the regressions to obtain the most accurate estimation the negative effect of MLDA on crime/ arrest numbers. The reduced form estimates the difference in crime/arrest numbers of population over MLDA and arrest rates of population under MLDA. It is tempting to subtract the decrease in total number of arrests due to MLDA by the total population to get an estimation of the effect of MLDA on arrests in terms of per person drinking. However, knowing there is only partial compliance, the reduced form underestimate the actual effect of MLDA on arrest rates in terms of per person drinking. Thus, we divide the reduced form by the first stage to obtain the actual effect of MLDA on arrests in terms of per person drinking. In another word, we divide the change in arrest numbers in two populations due to MLDA by the complier proportion of the population. Thus, we are able to obtain the actual effect of MLDA on arrests in terms of per person drinking. Finally, we calculate the standard errors of this instrumental variable estimation with first stage and reduced form coming from different dataset using delta method, a linear approximation of the variance. The t-stats suggests that we are 99% confident that MLDA reduces arrests in terms of per person drinking.

Although we believe we have an accurate estimation using the above two sample instrumental variable approach, only two of the three instrumental variable assumptions are met.

First, we confirmed that population under MLDA and population over MLDA have similar potential outcomes by comparing their observable characteristics. Second, there is indeed a negative effect of the MLDA on the complier proportion of the population. However, we cannot determine if MLDA is the only variable affecting the arrest rates. There are many other variables that can potentially contribute to the reduction of arrest rates. For example, the minimum legal age for cannabis is the same as the minimum legal drinking age.

Data:

The data we use is a collection of observable characteristics of individuals generated by the National Health Interview Survey (NHIS). NHIS is the principle source of information on the health of civilian non institutionalized population of the United States and serves as one of the major data collection programs of the National Center for Disease Control and Prevention. The data consists of individual's age, ethnicity, gender, married status, employment status, insurance status, and education status. NHIS has monitored the health of the nation since 1957 and it collects data through personal household interviews. From 1997 to 2007, the NHIS adopted geographically mustered sampling techniques to select the sample of dwelling units for the NHIS. The sample is designed in such a way that each month's sample is nationally representative. Nationally, about 750 interviewers (Field Representatives) are trained and directed by health survey supervisors of NHIS. The Arrest Rates in California found using official arrest certificates in California.

Minimum Legal Drinking Age (MLDA) in the United States is set at age 21. Comparing the population under and over MLDA requires similar potential outcomes between the two

groups. Since we cannot observe a person's behavior when he is a month away from his 21st birthday and his behavior when he is a month past his 21st birthday at the same time. The second best option is to compare population from different countries with similar age. However, the population around 21 in other countries can differ systematically from the population around 21 in the United States out of several reasons. One of the reasons can be different MLDAs across countries. For example, alcohol is accessible to teenagers under 18 in China, while you can only drink legally in U.S. after you turn 21. The young population in China and the United States is thus incomparable. Thus, the best option for us is a regression discontinuity design that requires continuity in covariates that are not affected by MLDA. We compare the observable characteristics from NHIS to check for covariates continuity. We regress age by proportion of people drinks to estimate first stage using NHIS. We regress age by arrest numbers to estimate reduced form using Arrest data. Finally, we use NHIS and Arrest data combined to estimate the standard error of the IV estimation using delta method.

Methods:

A randomized control trial (RCT) can provides us with unbiased estimation of the effect of MLDA on arrests in per-drinker terms. However, a RCT is unrealistic due to ethical reasons since giving the treatment group alcohol can be harmful and possibly increase their probability of being arrested. Adjusting for observable differences cannot get us unbiased estimation since the population that choose to drink is likely systematically different from population that choose not to drink. Thus we use a regression discontinuity design that can be regarded as a two sample

instrumental variable approach that compares the population affected by MLDA and the population not affected by MLDA. MLDA serves as the threshold in this case.

The continuity assumption of the regression discontinuity design requires no sharp changes in the potential outcomes at the threshold MLDA. In this case, valid comparison between the population over MLDA and the population under MLDA requires these two populations have similar potential outcomes. However, we cannot check that the untreated and treated potential outcomes are the same on both sides of the threshold. We can only observe largely treated outcomes on the left side of the threshold and largely untreated outcomes on the other side. The second option is a continuity check in potential relevant variables besides treatment variable and outcome variable at the threshold C. In another word, the observable and unobservable characteristics between the population over MLDA and the population under MLDA must be similar. Let f be a continuous function at threshold c. We denote x as a covariate that is not the outcome nor the instrument. The equation of continuity assumption follows:

$$f(c) = \lim_{x>c} f(x) = \lim_{x<c} f(x)$$

In our case, age is the running variable x; and f(x) is some covariate that may or may not change due to the increase in age. The threshold C is MLDA. To compare the values of each covariate as age approach MLDA: we run regression analysis on the observable characteristics of the population under MLDA and the population over MLDA and look for statistically significant variations due to MLDA.

In this case, we have the instrumental variable Zi indicating if a person is over or under MLDA. We denote instrument $Z_{i,0} = 0$, people who are under MLDA, as the control group; and $Z_{i,1} = 1$, people who are over MLDA, as the treatment group. We also have the variable Di

indicating whether a person drinks or not. We denote Di,0 = 0 as people who does not drink; and Di,1 = 1 as people who actually drinks. Then, we can separate the total population into four sub-populations: always drinkers, never drinkers, compliers, and defiers. The always takers and never takers populations are the same groups of people before and after MLDA. Intuitively, these two sub-populations cancel each other out when estimating the effect of MLDA on the proportion of the population that drinks. Also, the second assumption of the Regression Discontinuity design does not allow defiers. Thus we are only estimating the effect of MLDA on the proportion of compliers sub-population.

Intuitively, the fact that the complier proportion of the population in the MLDA case is smaller than one implies a fuzzy regression discontinuity design. To get an accurate estimation of the effect of MLDA on arrests in terms of per person drinking, we need to calculate the proportion of compliers in the total population, the first stage. The running variable Xi is age, the threshold MLDA is denoted as c, and the outcome variable Di is a binary variable indicating whether a person drinks or not. The equation for first stage follows:

$$F.S. = \lim_{x>c} E[Di|Xi = x] - \lim_{x<c} E[Di|Xi = c]$$

In plain words, this equation implies that the first stage is equal to the complier proportion of the population. We regress the age by Di, whether a person drinks or not, and look for the discontinuity that occurs at the threshold MLDA. Comparing both side of the threshold we are able to obtain the proportion of population truly affected by MLDA. A parametric approach estimates the first stage by best fitting regression lines on both side of the threshold:

$$yi = B0 + B1Zi + B2*Xi + B3*Xi*Zi + B4*Xi^2 + B5*Xi^2*Zi + B6*Xi^3 + B7*Xi^3*Zi + e$$

As we can interpret the above equation under the assumption as Xi approach the limit of the threshold c, the difference between the population with Zi = 0 and Zi = 1 simplifies to the estimated value of the coefficient B1. We then adjust the flexibility of above regression to get the most accurate estimation of the first stage. In this case, Yi represents the variable Di, whether a person drinks or not; Xi represents age centered at MLDA; and Zi is the whether a person is above or below MLDA.

A graphical presentation can best reflect this discontinuity. However, since Di, whether a person drinks or not, is a binary variable. Thus we take the means of Di in the range of bin size that we assign to the x-axis. A larger bin width represents a more precise estimation however can be biased; while a smaller bin width provides more accurate estimation with more noise in the graph. In this case, a 40-day bin width is preferred since it provides fairly unbiased estimation and reflects the discontinuity clearly. One of the preferred age range to best represent the effect of MLDA on the proportion of the population that drinks is from 19 to 23 years old and center it at 21, the MLDA. A larger age range may result in irrelevant discontinuities and less visual effect of the MLDA; while a smaller range blocks the whole visual of the regression on each side of the discontinuity. Finally, to fairly show the complier proportion of the population, the range of alcohol consumption can be chosen between 0.45 to 0.7. A larger range of alcohol consumption population proportion increases the noise of the graph; while a smaller range of alcohol consumption population proportion can exaggerate the effect of MLDA on the proportion of the population that drinks.

To estimate the actual decrease in arrests number due to MLDA, the reduced form, we use the following formula.

$$R.F. = \lim_{x>c} E[Yi|Xi = x] - \lim_{x<c} E[Yi|Xi = x]$$

In this case, Yi is the number of arrests in each state; and Xi is the age variable centered at c, MLDA. The reduced form is estimating the decrease in number of crimes/arrests due to MLDA. The parametric approach can be used again here for the reduced form when estimating the change in all cases of arrests and each sub-category of arrests due to MLDA.

The two sample instrumental variable approach suggests that the true treatment effect can be estimated by dividing the reduced form by the first stage. In the context of this paper, the total change in number of arrests due to MLDA over the population that is actually affected by MLDA implies that MLDA changes the crime/arrests rates by changing the proportion of population that drinks.

We use hypothesis tests and t-stat to check for statistical significance. We use the delta method when estimating the standard error for the IV estimator of data from two datasets.

Results:

Table 1 verifies the continuity assumption of the regression discontinuity design:

Table 1: balance check

VARIABLES	(1) hs_diploma	(2) hispanic	(3) white	(4) black	(5) uninsured	(6) working_lw	(7) going_school	(8) male
post1	-0.070***	-0.011	0.009	-0.005	-0.031***	-0.000	0.020**	0.014
	(0.010)	(0.010)	(0.012)	(0.009)	(0.011)	(0.011)	(0.009)	(0.012)
agec	0.130***	-0.002	0.007	-0.003	0.047***	0.073***	-0.076***	-0.018***
	(0.004)	(0.004)	(0.005)	(0.004)	(0.005)	(0.005)	(0.004)	(0.005)
agec_post	-0.137***	0.005	-0.014**	0.006	-0.041***	-0.042***	0.038***	0.018**
	(0.006)	(0.006)	(0.007)	(0.005)	(0.006)	(0.007)	(0.005)	(0.007)
Constant	0.911***	0.239***	0.564***	0.148***	0.334***	0.655***	0.151***	0.433***
	(0.007)	(0.007)	(0.009)	(0.006)	(0.008)	(0.008)	(0.006)	(0.009)
Observations	27,826	27,826	27,826	27,826	27,826	27,826	27,826	27,826
R-squared	0.049	0.000	0.000	0.000	0.006	0.034	0.052	0.001

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

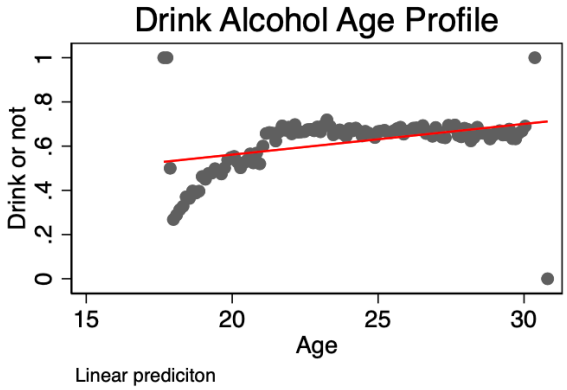
This is our balance table above. The variable of interest is our dummy variable that indicates whether the individual is 21 years old. Asterisks next to the number represents that the variation in this particular covariate is statistically significant. As we can see in the balance table, only his_diploma, working_lw, and going_school has variation above and below 21 that is statistically significant. But these are all understandable changes since as a person gets older, he is more likely to participate in the labor force rather than continuing to go to school. Also, as a person reach 21, he is more likely to have graduated from high school. These variations are all normal. Similar observable characteristics of the population under MLDA and the population over MLDA implies similar potential outcomes between two groups. We proceed by fitting regression line on both side of MLDA and adjusting for flexibility to obtain the best fitting line and most accurate estimation of the complier population.

Image 2: comparison of average drinking habits of individuals with Zi = 1 and Zi = 0

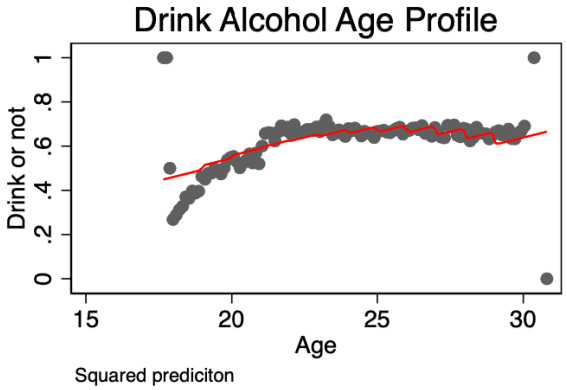
Table 2: regression estimates of different polynomial orders

VARIABLES	(1) drinks_alcohol	(2) drinks_alcohol	(3) drinks_alcohol	(4) drinks_alcohol
post1	0.032***	0.040***	0.168***	0.136***
	(0.012)	(0.012)	(0.033)	(0.027)
agec	0.086***	0.086***	-0.076**	-0.006
	(0.005)	(0.005)	(0.036)	(0.015)
o.agec_post				-
agec_sq			-0.041***	-0.003
			(0.009)	(0.013)
agec_sq_post			0.041***	0.004
			(0.009)	(0.016)
agec_cu				0.006*
				(0.004)
agec_cu_post				-0.006*
				(0.003)
birthday		-0.021**	-0.016	-0.022
		(0.009)	(0.011)	(0.018)
agec_post	-0.087***	-0.089***	0.077**	
	(0.005)	(0.005)	(0.036)	
Constant	0.637***	0.637***	0.503***	0.541***
	(0.011)	(0.011)	(0.031)	(0.015)
Observations	61,784	61,784	61,784	61,784
R-squared	0.031	0.031	0.031	0.031

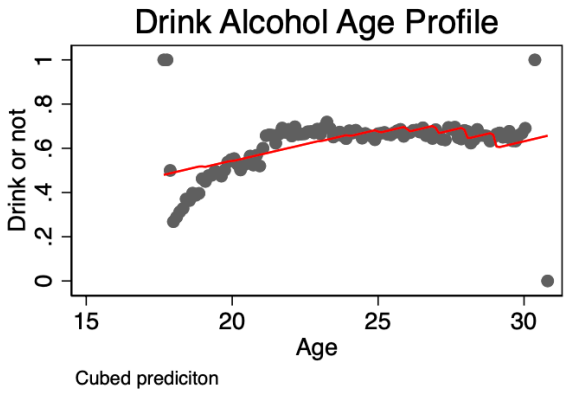
Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1



Linear prediciton



Squared prediciton

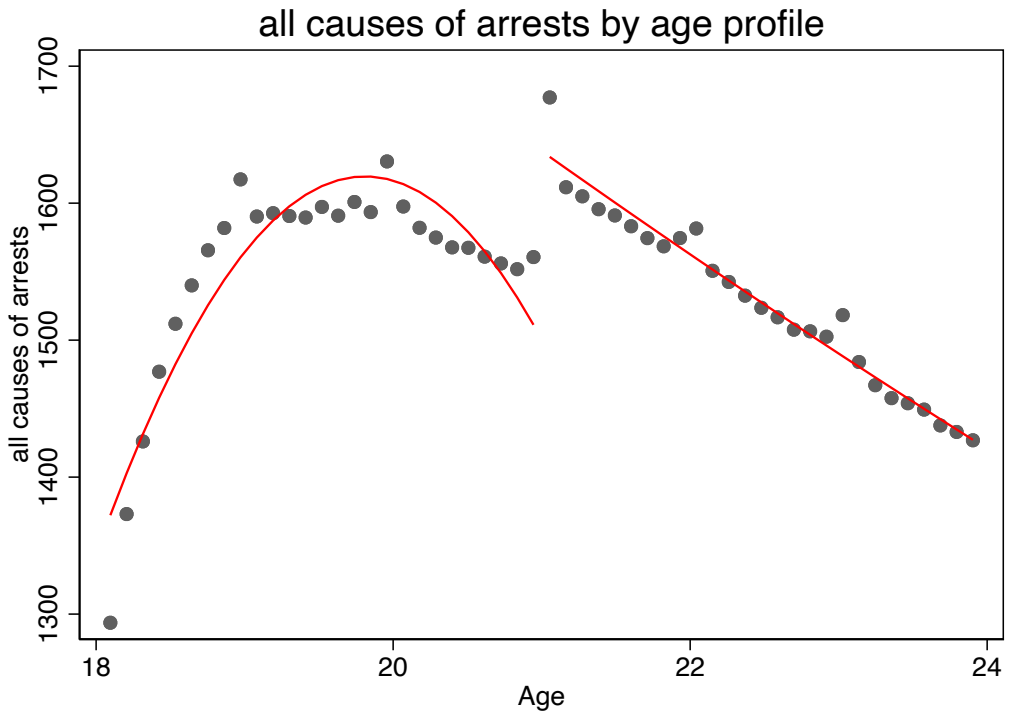


Cubed prediciton

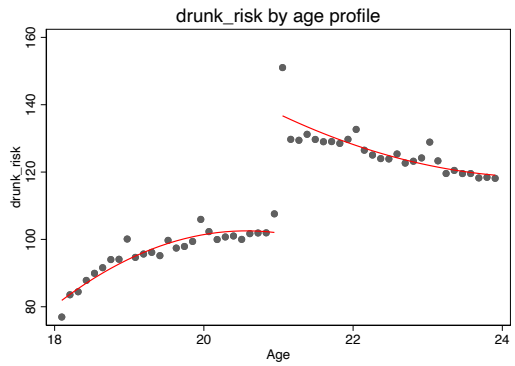
The above figures are the regression estimates with different polynomial orders. As we can see in the figure of the linear, squared, and cubic regression estimates, the linear is the worst fitting regression line. Both the squared and cubic gives us similar fitting regression estimations and they are fitting well. The reason that we choose to use the quadratic regression estimation can be seen in the regression table above. If we use the third order polynomial regression estimation, the results give us a reduction of 13.6% of the proportion of population that drinks with 2.7% standard error. The t-stat in this case is 5.037. Also, notice the statistical significance level differences between the second order polynomial and the third order polynomial, the square term of age centered to 21 of the third order polynomial is not statistically significant. While all the age terms and post term in the second order polynomial method are statistically significant. The second order polynomial regression estimation gives us better t stats. As a result, we are 99% confident that MLDA reduces the proportion of the population that drinks by 16.8% with a standard error of 3.3%. The t-stat is equal to 5.09. We are 99% confident there are 16.8% compliers in the population that starts to drink after they pass MLDA.

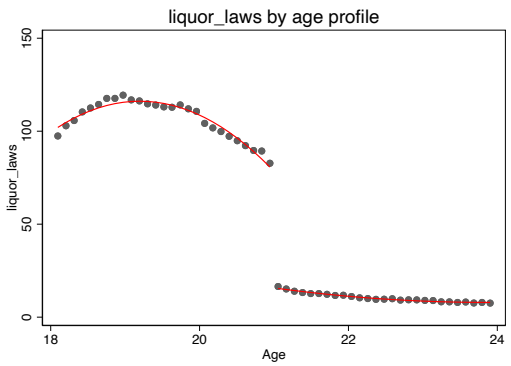
The reduced form requires us to estimate the total change in number of crimes and arrests. The results are shown below:

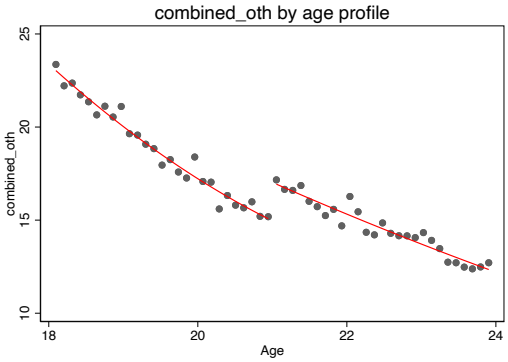
Image 3: all causes of arrests by age



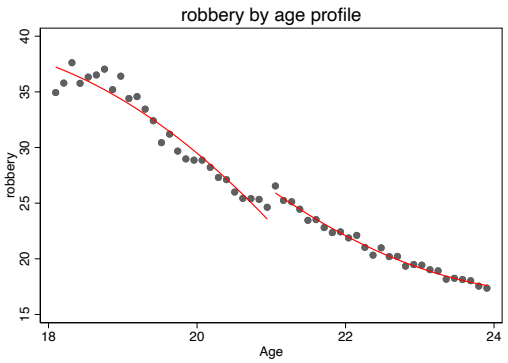
In the figure above, there is a clear increase in all causes of arrests due to MLDA. The regression discontinuity suggests that the population just turned 21 has approximately 80 more arrests compared to the population that is approaching their 21st birthday. However, there is only approximately 5% increase in the total number of arrests when compared to the 1570 arrests that is already happening in the population before they turn 21. Let us breakdown the increase in arrests in each sub-category of crimes in the following figures:

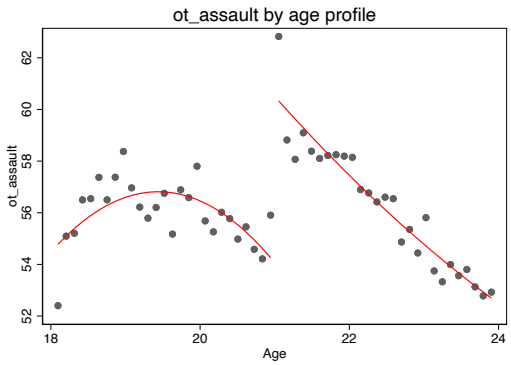

drunk_risk by age profile
drunk_risk
80
100
120
140
160
18
20
22
24
Age

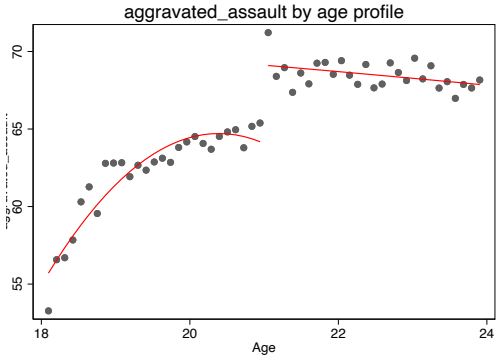

liquor_laws by age profile
liquor_laws
0
50
100
150
18
20
22
24
Age


combined_oth by age profile
combined_oth
10
15
20
25
18
20
22
24
Age


dui by age profile
dui
100
150
200
250
18
20
22
24
Age


robbery by age profile
robbery
15
20
25
30
35
40
18
20
22
24
Age


ot_assault by age profile
ot_assault
52
54
56
58
60
62
18
20
22
24
Age


aggravated_assault by age profile
55
60
65
70
18
20
22
24
Age

From the above figures, we can clearly see an increase in the number in every cause of arrests occurs due to MLDA except liquor law violations. There is a sudden decrease in liquor law violations since the population over 21 is not affected by the MLDA.

Then we regress age by all causes of arrests; and we regress age by each sub-category of arrests to observe the increase in numbers of due to MLDA.

Table 3: regression of all and each cause of arrests due to MLDA

VARIABLES	(1) all	(2) dui	(3) liquor_laws	(4) robbery	(5) aggravated_assault	(6) ot_assault	(7) drunk_risk	(8) combined_oth
post1	80.568***	53.771***	-66.582***	1.213**	3.912***	5.473***	35.641***	2.129***
	(6.526)	(1.837)	(0.778)	(0.475)	(0.780)	(0.738)	(1.806)	(0.392)
agec	83.977***	32.837***	-23.483***	1.038	5.358***	1.870	12.514***	-1.543*
	(13.352)	(3.758)	(1.593)	(0.972)	(1.596)	(1.509)	(3.694)	(0.802)
agec_post	-211.024***	-36.937***	15.421***	-6.620***	-8.696***	-6.418***	-38.996***	-0.856
	(18.832)	(5.299)	(2.246)	(1.370)	(2.251)	(2.128)	(5.210)	(1.131)
agec_sq	156.657***	7.389**	3.866***	6.056***	4.559***	3.358***	9.421***	0.631
	(10.331)	(2.907)	(1.232)	(0.752)	(1.235)	(1.168)	(2.858)	(0.620)
agec_sq_post	-114.383***	-0.023	-0.714	-4.757***	-2.058	-2.198	4.851	-0.090
	(14.580)	(4.103)	(1.739)	(1.061)	(1.743)	(1.648)	(4.034)	(0.875)
agec_cu	54.405***	3.259***	3.473***	1.532***	1.410***	0.996***	2.841***	0.078
	(2.262)	(0.636)	(0.270)	(0.165)	(0.270)	(0.256)	(0.626)	(0.136)
agec_cu_post	-63.321***	-5.402***	-3.968***	-1.678***	-1.985***	-1.194***	-5.603***	-0.179
	(3.193)	(0.899)	(0.381)	(0.232)	(0.382)	(0.361)	(0.883)	(0.192)
Constant	1,570.484***	195.455***	82.959***	25.189***	65.927***	55.420***	105.829***	15.080***
	(4.632)	(1.303)	(0.552)	(0.337)	(0.554)	(0.523)	(1.281)	(0.278)
Observations	2,191	2,191	2,191	2,191	2,191	2,191	2,191	2,191
R-squared	0.787	0.966	0.991	0.838	0.420	0.145	0.701	0.615

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

As we can see in the regression table, almost all decrease in arrests due to MLDA is statistically significant. Except liquor laws violation, there is an increase in liquor law violation. But other than this, we are 99% confident that MLDA does reduce crimes and arrests. The variable post1 in this table is the Zi variable indicating whether a person is over 21. The variable agec represents age centered at 21; in another word, if a person is 22 years old, his/her agec is equal to 1. The other variables in the regression tables should explain itself intuitively. However, there does not seem to be a large effect of MLDA on the total number of arrests. MLDA seems to reduce Dui and drunk risk by over 20%; but MLDA only reduces sub-5% in all sub-categories of arrests.

Finally, we compute effect of MLDA on arrests in terms of per person drinking using an instrumental variable approach: as we computed the following:

IV estimates for All:1228.3673. Drunk risk: 420.63238. DUI: 634.03849. liquor laws: -753.55781. combined OTH: 25.600392. Robbery: 20.876471. Ot assault: 67.018468. Aggravated assault: 50.760259

Standard error All: 214.49339. Drunk risk: 72.180112. Dui: 106.03745. Liquor laws: 123.80272. Combined OTH: 5.9198793. Robbery: 6.0015353. Ot assault: 13.658825. Aggravated assault: 11.867391

The t-stats for each IV estimates and its standard errors suggests the IV estimation is statistically significant. Although the reduced form suggests a small effect of MLDA on overall arrests numbers, all the effects of MLDA on overall arrests are contributed by only 16.8% of the total population. In another word, the reduced form only shows 16.8% of the true treatment effect of MLDA on arrests. If we put the instrument variable estimation, the true treatment effect of MLDA, into percentage form: MLDA reduces all arrests by 44%; MLDA reduces dui by 76%; MLDA increases liquor law violations by 90%; MLDA reduces drunk risk by 80%; MLDA reduces combined OTH by 62.5%; MLDA reduces robbery by 45%; MLDA reduces Ot assaults by 54%; MLDA reduces aggravated assault by 44%. After we adjust the total number of decrease/increase in arrests to the complier proportion of the population, the IV estimation suggests that MLDA reduces arrests hugely.

Conclusion:

When accessing the question if MLDA reduces arrests in young Californians: the differences in total numbers of arrests between the population affected by MLDA and the population not affected by MLDA does not suggest that MLDA affect crime/arrests rates very much. However, the complier population only takes up 16.8% of the total population. In another word, all the changes in the number of arrests due to the effect of MLDA on the population are

contributed by only 16.8% of the total population, which implies that the reduced form underestimate the true treatment effect of MLDA by 83.2%. In order to get a true estimation for the treatment effect of MLDA on the young population in California. We adjust the total change in number of arrests to the complier population. Mathematically, the complier proportion of the population, the first stage, cancels out with the percentage coefficient of the true treatment effect of MLDA estimated by the reduced form. Theoretically, the IV estimation reflects the true treatment effect of MLDA; statistically, we are confident that people who drinks have much higher probability of being arrested compared to people who does not drink. Intuitively, MLDA reduces the probability of the young Californian population getting arrested.

The IV estimator requires three assumptions: first, the randomized selection assumption implies similar potential outcomes between the treatment and the control group; in this paper, we statistically proved that the population affected by MLDA and the population not affected by MLDA has similar potential outcomes by doing a balance check, in which we regress covariates, that are neither instrument nor outcomes, by age; then we compare their limits as age approach MLDA. Second, the instrument must have an affect on the outcome; in another word, the complier population can not be zero; which we proved in first stage estimation. We fail to prove the third assumption of the IV estimation, which states that the outcome is only affected by the instrument variable. Thus we do not obtain a perfectly unbiased estimation in this paper on how much MLDA reduces crime/arrests in terms of per person drinking. We are likely overestimating the effect of MLDA on arrests since there are many other factors that can increase crime/arrests rates after a person turns 21.